PROJECT SUMMARY

Effect of Crohn's Disease Risk Alleles on Enteric Microbiota

Dr. Ellen Li, Stony Brook University, Stony Brook, NY and Washington University, St. Louis, MO
Dr. Erica Sodergren, Washington University, St. Louis, MO
Dr. George Weinstock, Washington University, St. Louis, MO
Dr. Thaddeus Stappenbeck, Washington University, St. Louis, MO
Dr. Daniel Frank, U. of Colorado, Denver, CO
Dr. Charles Robertson, U. of Colorado, Boulder, CO
Dr. Norman Pace, U. of Colorado, Boulder CO
Dr. Wei Zhu, Stony Brook University, Stony Brook, NY
Dr. Edgar Boedeker, U. of New Mexico, Albuquerque, NM
Dr. R. Balfour Sartor, U. of North Carolina, Chapel Hill, NC
Dr. Phillip Tarr, Washington University, St. Louis, MO
Dr. Lloyd Mayer, Mount Sinai School of Medicine, New York, NY
Dr. David Dietz, Cleveland Clinic, Cleveland, OH

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

Project ID: 46323
Publication Moratorium: 12 months.

This UH2/UH3 demonstration project entitled “Effects of Crohn’s disease risk alleles on enteric microbiota” is focused on characterizing intestinal associated microbiota in patients with ileal Crohn’s disease (ileal CD), ulcerative colitis (UC) and control patients without inflammatory

bowel diseases (non-IBD). We hypothesize that genetic factors that affect Paneth cell function, contribute to compositional changes in intestinal microbiota. These changes in microbiota may lead to reduction of protective commensal organisms and increased numbers of aggressive organisms that incite intestinal inflammation. This hypothesis is being tested by high throughput 16S rRNA sequence analysis of de-identified ileal and colonic tissues that have been archived at Washington University St. Louis, University of North Carolina, Mount Sinai Hospital and the Cleveland Clinic. Multivariate analysis of the metagenomic data will be conducted with genotyping metadata (highly reproducible CD risk alleles, including NOD2 and ATG16L1) and phenotyping metadata (e.g. age, gender, race, body mass index, medications and smoking). Shotgun sequencing will be performed on selected fecal specimens linked to ileal tissues to identify additional, or auxiliary, or synergistic pathogenic factors or other functional changes in the microbiome. Because members of this research team have observed that a chronic viral infection is required for the Paneth cell defect in Atg16l1 hypomorphic mice, a major focus of these studies will be towards identifying potential viral triggers for the defective Paneth cell phenotype in individuals harboring the ATG16L1 risk allele. Novel genetic probes for protective and aggressive organisms will be developed by mining bacterial genome and shotgun sequencing data. Genomic sequences will be produced for candidate protective and aggressive strains (e.g. adherent-invasive strains of E. coli) isolated from human intestinal tissues where there is limited existing genome information. Quantitative qPCR assays using the novel as well as established genetic probes will be conducted to test the hypothesis that an imbalance between protective and aggressive organisms is associated with genetic factors that affect Paneth cell function.

Our combined expertise in multiple disciplines across multiple institutions, our demonstrated ability to collect a large number of well-phenotyped samples with longitudinal clinical information that will be linked to host response and morphologic studies, and our consortium's capacity for high-throughput sequencing will be used to investigate how alterations in human microbiome relate to CD risk alleles and CD pathogenesis.

II. DATA QUALITY:

Genotype data: Quality of the genotype data was checked by reproducibility of 5% of the samples

16S sequence and WGS data: The quality of capillary sequencing data (Sanger sequencing on the AB3730 instrument) at the Genome Center at Washington University is measured by assessing the failure rate of each individual set of 96 lanes within one full run. Within each run these failures are samples with no data or samples that have fewer than 20 high quality bases. A high quality base is one with a quality score greater than phred Q20. The number of such failed samples is noted for each run. Successful runs are those with fewer than 20% failures, although this number is

often set more stringently.

In addition, the overall read length for all passing samples is measured across many different variables (e.g. high quality bases) to make sure that it says within the standard expected for this platform. The expected read length at this time is 700 bases of a quality score greater than phred Q20.

The Illumina production pipeline is evaluated by the number of passing reads that contain high quality data. The Illumina software on the instrument calculates the number of passing reads as well as the number of clusters (a cluster is formed from a single DNA fragment) that might produce data. The reports also offer information regarding the phasing or the ability of the instrument to stay in step with each base that is called. In addition to this, when possible, the error rate is evaluated by evaluation of an internal standard of known sequence or by alignment of the experimental sequences to known reference sequences, when available. Successful runs are those producing an expected full set of reads with a low error rate. The full set of reads depends on the sequencing conditions while error rates are typically < 1%, analogous to phred Q20.

Sequencing on the Roche-454 platform is similarly evaluated by the number of reads and bases produced per run, the read length distribution, as well as an error rate in base calling.

III. DATA ANALYSIS AND PUBLICATION PLANS:

Data Analysis:

A. Comparison of broad range 16S rRNA PCR amplification/Sanger (ABI platform) sequencing with amplification of the 16S rRNA hypervariable regions/ followed by sequencing on the 454 and/or the Illumina platforms. We initially began sequencing the Washington University samples using Sanger sequencing at a depth of 384 x 2, in order to directly compare with the sequencing results obtained with the Mount Sinai samples. 454 sequencing of a hypervariable region of the 16S rRNA gene can be conducted at a greatly reduced cost making it sequencing to a much greater depth (~10,000) affordable. Thus far we have paired results for 165 samples using the Sanger the 454 v1_v3 window and 175 samples using both Sanger and the 454 v3_v5 window. There are 135 samples for which we have Sanger and both v1_v3 and v3_v5 windows. Phylogenetic classification of the sequences was conducted using the Naïve Bayesian Classifier of the Ribosome Database Project (Version 2.1, http://sourceforge.net/projects/rdp-classifier/) and the taxa were binned initially at the phyla level: 1. Actinomycetes; 2. Bacteroidetes; 3. Firmicutes; 4. Proteobacteria; 5. Other phyla. The Firmicutes clade was further subdivided into the

following five subcategories, based on concordance between the RDP classifier and the Greengenes 16S rRNA phylogenetic schema]: 3A. Firmicutes/Clostridium Group XIVa, 3B. Firmicutes/Clostridium Group IV, 3C. Firmicutes/Clostridium/ Other, 3D. Firmicutes/Bacillus, and 3E. Firmicutes/Other classes. To analyze the overall microbial composition we have subjected the relative frequencies of eight of the nine categories (excluding Phyla/other) to logit transformation: log2{(x+0.01)/[100-(x+0.01)]}, where x is the relative frequency (%). The purpose of the logit transformation was to expand the original relative frequency values (0-100%) to the real space, in which the term 0.01 was added to avoid infinite value during the transformation. The Sanger results for each of the eight individual categories are compared with the results for each of the 454 windows by the Wilcoxon signed-rank test. The preliminary analysis indicates that the bacterial categories in the Firmicutes phyla are for the most part statistically equivalent. The correlation between Sanger results and each of the 454 windows will also be compared. Finally, we will use structural equation modeling (SEM) to construct a latent variable representing the "true" value for the relative frequency of each of the eight bacterial categories that is constructed from the measured Sanger, 454 v1_v3 and 454 v3_v5 results. This latent variable SEM method will allow us to determine which of the 454 windows most closely approximates the "true value" for the relative frequency of each bacterial category, particularly with respect to the Bacteroidetes and Proteobacteria phyla.

B. Integration of 16S sequence analysis data with clinical data, genotype data and whole human expression profiling data. 16S sequence data (Sanger) was integrated with clinical data and genotype data on 178 intestinal (ileal and colonic, disease affected and unaffected) samples that were collected at Mount Sinai Hospital after reducing the dimensions of the molecular phylogenetic data to eight bacterial categories at the phyla/ subphyla level as described in section A. We have also examined selected categories at the genera level that are of interest to inflammatory bowel diseases, such as Fecalibacterium and Escherichia. The effect of disease phenotype (CD, UC, Control), NOD2 genotype (NR/NR, R/NR, R/R), ATG16L1 genotype (NR/NR, R/NR, R/R), and patient age and gender on these eight categories was analyzed by nonparametric multivariate analysis of covariance (MANCOVA) using the R software package (Version 2.8.1). The effect of these independent variables on lower taxonomic ranks was analyzed by nonparametric analysis of covariance (ANCOVA). The same approach is being used to analyze 16S sequence data, clinical data (age, gender smoking, BMI, C. difficile) and genotype data on a second independent set of 276 ileal samples that were collected at Washington University and at the U. of North Carolina. Our preliminary analysis indicate that NOD2 genotype is significantly associated with a shift in overall bacterial composition by MANCOVA for Sanger and both 454 v1_v3 and v3_v5 windows. In addition we will integrate the molecular phylogenetic data obtained from the Washington University samples with whole human genome expression profiling data obtained from the same 84 samples. We will

reduce the dimensions of the whole human genome expression profiling data from 26,765 probe set (after pre-processing and normalization of the array data) to 39 clusters based on gene-gene correlations. The first principal component of each of the 39 clusters will be used as the independent variable in addition to the genotype and clinical variables listed above in a logistic regression analysis while applying the forward variable selection method. One of the 39 clusters is enriched for Paneth cell genes and we are particularly interested in examining whether the 1st PC of th cluster is associated with shifts in any of the eight bacterial categories. Finally we are in the process of scoring Paneth cell morphology in adjacent sections of the samples used for microarray and molecular phylogenetic studies. Once these scores are completed, we plan to incorporate this variable in our logistic regression analysis.

C. Comparative genomic analysis of inflammatory bowel disease associated adherent invasive Escherichia coli strains. We have sequenced and assembled the genomes of 27 *E. coli* strains cultured from human GI tissues with and without IBD, and have made automated gene predictions for these strains. These strains have been extensively characterized with respect to their adherence-invasive phenotype in vitro, including survival in macrophages. Pairwise comparisons off the E. coli isolates will be conducted to determine areas of diversity between the strains.

Accepted publications:

1. Frank DN. BARCRAWL and BARTAB: software tools for the design and implementation of barcoded primers for highly multiplexed DNA sequencing. BMC Bioinformatics. 2009; 10: 362.
2. Frank DN, Robertson CE, Hamm CM, Kpadeh Z, Zhang T, Chen H, Zhu W, Sartor RB, Boedeker EC, Harpaz N, Pace NR, Li E. Disease phenotype and genotype are associated with shifts in intestinal-associated microbiota in inflammatory bowel diseases. Inflammatory Bowel Diseases, in press.

Intended publications:

1. Logistic regression analysis of Crohn's disease genotype – disease location associations submitted June 2010
2. Human ileal whole genome expression profiling in inflammatory bowel diseases. Plan to submit September 2010.
3. The association between NOD2 genotype and altered human ileal-associated microbial composition is independent of its association with inflammatory bowel disease phenotype. Plan to submit October 2010
4. Ileal mucosal-microbial interactions in inflammatory bowel diseases plan to submit October 2010

5. Comparative genomic analysis of inflammatory bowel disease associated adherent-invasive Escherichia coli plan to submit May 2011

IV. DATA RELEASE PLAN:

The Genome Center at Washington University (GCWU) plans to release the sequence reads to the NCBI Trace Archive without delay and has a data submission pipeline in place that has been doing this for years. The Short Read Archive for nexgen data has only recently reopened after closing due to a data flood it was not equipped to handle. We look forward to working with the NCBI on the continued development of this repository. We will also release our metagenomic data to either the Trace Archive or the GenBank division that handles this type of sequencing. We will also release genome assemblies without delay, although the NIAID policy allows for a brief delay. Again, we have a pipeline in place that has routinely submitted assemblies and annotations to GenBank and worked with them to resolve discrepancies. We are submitting this grant fully aware of NIH policies regarding the dissemination and sharing of results and products that are derived from government funded research. Published data sets containing clinical information will be made available immediately to all qualified investigators both through the institutional platforms maintained by the Center for Biomedical Informatics. For unpublished data sets, qualified investigators will initiate requests for resources by providing basic information about their research project to the steering committee.

V. CONTACT PERSON:

Dr. Ellen Li, Stony Brook University, Ellen.li@stonybrook.edu, 1-631-632-5977