PROJECT SUMMARY

The Neonatal Microbiome and Necrotizing Enterocolitis.

Dr. Phillip I. Tarr, Washington University, PI
Dr. Barbara Warner, Washington University, co-PI
Collaborators: Drs. Erica Sodergren, William Shannon, Aaron Hamvas, Vincent Magrini, George Weinstock

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 Proceedings HMP summary page.

The data has been deposited in NCBI with the following project ID number 46337. There is a one year publication moratorium on all data other than DNA sequence data from this study starting from the time of data submission. There is no publication moratorium on DNA sequence data from this study.

Necrotizing enterocolitis (NEC) is a devastating disorder that affects approximately 10% of premature infants. Its mortality remains high (15-30%), and its cause remains unknown. About 80% of cases occur within 35 days of birth among hospitalized newborns of low birth weight. Probiotics diminish the incidence and severity of NEC, and NEC does not occur antepartum. NEC affects a readily identifiable at-risk group, has a tightly defined interval before its onset, occurs in an organ system that is intimately associated with a microbial population in flux, has a plausible association with the intestinal microbiota, and cohorts at risk have rarely been studied in large numbers, or prospectively. This disorder, therefore, provides a unique opportunity to explore the role of the human enteric microbiome in a devastating disease. Moreover, NEC epidemiology and age-incidence present an ability to enroll and study cohorts that are highly likely to provide valuable pathophysiologic and microbiologic insights.

In this project, we will identify and quantify the microbial components of stool and its products before and at the onset of NEC. In doing so, we will test the **overarching hypothesis that NEC is a direct or indirect consequence of the enteric biomass, its products, or both**. We will use

multicenter cohorts of premature infants at high risk of developing NEC, extend our research on this disease currently sponsored by the Washington University Institute of Clinical and Translational Sciences, and continue our longstanding collaborations with the Genome Center at Washington University and the Washington University Digestive Diseases Research Core Center (Informatics Core). The Aims of this proposal are to (1) conduct a case cohort study in which we compare clinical data and biological specimens from cases and well-matched controls; (2) determine if the kind and density of intestinal biomass, its gene content, and transcriptional activity are associated with, and potential determinants of, NEC; and (3) determine if host risk alleles for intestinal inflammation play a role in the development of NEC. These efforts will be accomplished using subjects from three collaborating neonatal intensive care units (NICUs), focusing on the critical, instructive, and understudied pre-NEC stage of illness, and formulating a data repository that will be a resource for investigators worldwide who wish to focus their efforts on NEC, its precipitants, and its prevention and cure.

II. DATA QUALITY:

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.

In addition, we have instituted a barcode specimen tracking system to increase our ability to monitor specimens as they are processed, stored, and distributed throughout the project.

III. DATA ANALYSIS AND PUBLICATION PLANS:

Our data analysis objectives are to report on the kinetics, and to characterize in depth, the enteric microbiota in the premature infant at risk for NEC. Publications concerning methodology and the descriptive ecology of this biomass are published, submitted, or in preparation. Additional publications, to be submitted at a rate of approximately two per annum, will focus on bacteria present, the genes they contain, the transcripts they contain, the host response to this biomass, and technological validations and methodologic assessments.

IV. DATA RELEASE PLAN:

The 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. We are releasing our 16S and whole genome shotgun metagenomic data to SRA. 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 releasing disease phenotype data linked to microbiome data to controlled access dbGaP. The PI should be contacted if there is an interest in accessing the dbGaP controlled access data.

V. CONTACT PERSON:

Dr. Phillip I. Tarr, Washington University, tarr@kids.wustl.edu