PROJECT SUMMARY

Metagenomic study of the human skin microbiome associated with acne

Dr. Huiying Li, University of California, Los Angeles

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

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: 46327.
Publication moratorium: One year.

The human microbiota contributes to our normal postnatal development and plays a significant role in defining our physiology. To understand the role of microbiota in human health and disease, we study the skin microbiome in pilosebaceous units (hair follicles) and its association with acne.

Acne is one of the most common skin diseases. Although its etiology still needs to be defined, a bacterial factor has been suggested in the development of the disease. In fact, antibiotic therapy targeting *Propionibacterium acnes* has been a mainstay treatment for more than 30 years.

Our preliminary study shows that the microcomedone, a specialized skin compartment where acne arises, has a tractable microbiome, with a single dominant species, *P. acnes*. This system offers a unique advantage allowing in-depth analysis of a human microbiome at the subspecies level by sequencing. Our preliminary study suggests that the microbiome associated with acne offers promise for understanding the correlation between the composition of the microbiome and human health and disease.

The goal of the project is to determine whether the microbiota in the pilosebaceous units contributes to acne. We plan to investigate the microbiome associated with acne in three directions. First, we plan to investigate the strain diversity of *P. acnes* in a disease cohort and a normal cohort and examine whether certain strains of *P. acnes* are correlated with the disease. Second, we plan to investigate the non-*P. acnes* microbes in microcomedones and disease lesions and examine whether they correlate with acne pathogenesis. Third, we will examine the interactions between the microbes and the host by transcriptional profiling of both the microbiota and the host.

During the first year of this project, two main questions were asked. 1. Are certain strains of *P. acnes* associated with acne, but rarely found in normal individuals? 2. If specific strains are

associated with acne, what are the differences in their genetic composition compared to other *P. acnes* strains that are not associated with acne? We collected microcomedone samples from more than 100 subjects, including acne patients and normal individuals. Genomic DNA was extracted from each sample, and 16S rDNA was amplified using universal primers (8F and 1510R), cloned and sequenced using Sanger method. Approximately 384 near full length 16S rDNA sequences were obtained for each sample. Some of the microcomedone samples were also cultured under anaerobic condition to isolate different *P. acnes* strains. Sixty-eight isolates were selected for whole genome shotgun sequencing using Solexa/Illumina platform. By the end of the first year of the project, we completed the sequencing of more than 40,000 16S rDNA clones and 68 genomes of *P. acnes* isolates.

II. DATA QUALITY:

Our data quality control follows the guidelines established for the Human Microbiome Project (HMP). All the library construction, Sanger sequencing and Solexa/Illumina sequencing were performed at the Genome Center at Washington University (GCWU), which is one of the four NIH-funded genome sequencing centers for the HMP.

The quality of capillary sequencing data (Sanger sequencing on the AB3730 instrument) at GCWU is measured by assessing the failure rate of each individual set of 96 lanes within one full run. Within each run the 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 stays 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 Solexa/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.

III. DATA ANALYSIS AND PUBLICATION PLANS:

Data Analysis Plan

We are currently analyzing the 16S rDNA data from acne patients and normal individuals to investigate whether there are differences at the species level and *P. acnes* subspecies level between the two groups. Bidirectional 16S rDNA reads are assembled and aligned to a core set of NAST-formatted sequences using AmosCmp16Spipeline and NAST-iEr from the Microbiome

Utilities Portal of the Broad Institute (http://microbiomeutil.sourceforge.net/). Suspected chimeras are identified using ChimeraSlayer and Wigeon (Microbiome Utilities Portal of the Broad Institute). Sequences with at least 90% bootstrap support for a chimeric breakpoint (ChimeraSlayer) or containing a region that varies at more than the 99% quantile of expected variation (WigeoN) are removed from further analysis.

For the *P. acnes* ribotype analysis, the resulting sequences are trimmed to include only positions 29 to 1483 (numbering based on the *E. coli* system of nomenclature). Sequences without full coverage over this region are excluded from further analysis. Low quality sequences with more than 50 bases between positions 79 to 1433 with phred quality scores of less than 15 are also excluded. Since the ribotype analysis compares highly similar sequences, the data are extensively manually edited. Chromatograms are visually inspected at all bases with a phred quality score < 30, and appropriate corrections are applied.

For the all species analysis, requirements of full length and high quality sequences are not used to screen sequences, since a high quality match is required at the species assignment step. Species assignments are transferred from BLASTN top scoring matches to the rRNA16S gold database (Microbiome Utilities Portal of the Broad Institute) only for matches with > 1000 nucleotides aligned at > 97% identity.

To determine the differences in genetic composition between the *P. acnes* strains found in acne and the ones found in normal individuals, we are working on genome assembly, annotation and comparison of the 68 *P. acnes* isolates using the whole genome shotgun sequence data. For genome comparison, we are analyzing the core regions, non-core regions and the pan-genome of *P. acnes*. The core regions are identified as sequences that are present in all *P. acnes* genomes and are mapped to the reference genome *P. acnes* KPA171202 using Nucmer. Single nucleotide polymorphisms are identified for each strain. The non-core regions are identified by subtracting the core regions from the *P. acnes* genomes. The pan-genome calculation is based on the method described in Tettelin H. et al, 2005.

Publication Plan

We plan to submit for publication the initial results of the 16S rDNA sequence analysis and the *P. acnes* genome analysis within one year.

IV. DATA RELEASE PLAN:

We will follow the data release and resource sharing policies specified as part of the NIH Roadmap HMP, as described on the HMP website at http://nihroadmap.nih.gov/hmp/datareleaseguidelines.asp. All types of data generated, including clinical data, will be released in accordance with the guiding principles stated for the HMP. In addition, we will, when appropriate, collaborate with the DACC to facilitate the standardization, transfer, exchange and dissemination of information.

The data release plan will include release of raw sequence data, data from next generation sequencing platforms as agreed by the sequence producers and NCBI, genome assemblies and their annotations, clinical data, and other metadata associated with all data types. Data of all

types (human sequence, clinical data, transcription data, etc.) that contain potentially identifying information will be submitted to NCBI's dbGaP (a controlled access database). Some of these types of data will be released only after screening by NCBI. Data release plans specifying how those types of data will be verified are developed in consultation with the other funded HMP projects to ensure consistency in quality and verification standards.

Genome and Metagenome Sequence Data

All raw genome and metagenome sequence (which includes 16S rDNA sequence) and next generation sequence data that are generated by HMP data production cooperative agreements are being submitted to the Trace Archive or to the Short Read Archive at NCBI/NLM/NIH. These data will also include information on templates, vectors, and quality values for each sequence.

A minimum metadata set associated with genome and metagenome sequence data will be submitted to NCBI along with the sequence data. All sequence data and metadata that are potentially identifying of the donor will be deposited in controlled access portion of dbGaP.

Genome, Metagenome Assembly and Annotation

Genome and metagenome full and partial assemblies and their annotations will be deposited in GenBank at NCBI after verification by the center. Deposited metadata associated with genome and metagenome assemblies and annotation will be connected to the genome, metagenome assemblies and annotation data files.

Clinical Data

Clinical data will be submitted to the controlled access database dbGaP, consistent with the protection of donor privacy.

Finally, the data release plan will be in line with the data release guidelines already developed by the International Human Microbiome Consortium (IHMC).

V. CONTACT PERSON:

Dr. Huiying Li, University of California, Los Angeles, huiying@mednet.ucla.edu.