PROJECT SUMMARY

The Microbial Ecology of Bacterial Vaginosis: A Fine Scale Resolution Metagenomic Analysis.

Dr. Jacques Ravel, University of Maryland School of Medicine
Dr. Larry Forney, University of Idaho

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: 46331

The vaginal microbiota play an important protective role in maintaining the health of women. Disruption of the mutualistic relationship that exists between bacterial communities in the vagina and their hosts can lead to bacterial vaginosis (BV), a condition in which lactic acid producing bacteria are supplanted by a diverse array of strictly anaerobic bacteria. BV has been shown to be an independent risk factor for adverse outcomes including preterm delivery and low infant birth weight, acquisition of sexually transmitted infections and HIV, and development of pelvic inflammatory disease. National surveys indicate the prevalence of BV among U.S. women is 29.2%, and yet, despite considerable effort, the etiology of BV remains unknown. Moreover, there are no broadly effective therapies for the treatment of BV, and reoccurrence is common. **In the proposed research we will test the overarching hypothesis that vaginal microbial community dynamics and activities are indicators of risk to BV**. To do this, we propose to conduct a high resolution prospective study in which samples collected daily from 200 reproductive-age women over two menstrual cycles are used to capture molecular events that take place before, during, and after the spontaneous remission of BV episodes. We will use modern genomic technologies to obtain the data needed to correlate shifts in vaginal microbial community composition and function, metabolomes, and epidemiological and behavioral metadata with the occurrence of BV to better define the syndrome itself and identify patterns that are predictive of BV. The three specific aims of the research are: (1) Evaluate the association between the dynamics of vaginal microbial communities and risk to BV by characterizing the community composition of vaginal specimens archived from a vaginal douching cessation study in which 33 women self-collected vaginal swabs twice-weekly for 16 weeks; (2) Enroll 135 women in a prospective study in which self-collected vaginal swab samples and secretions are collected daily along with data on the occurrence of BV, vaginal pH, and information on time

varying habits and practices; (3) Apply model-based statistical clustering and classification approaches to associate the microbial community composition and function, with metadata and clinical diagnoses of BV. The large body of information generated will facilitate understanding vaginal microbial community dynamics, the etiology of BV, and drive the development of better diagnostic tools for BV. Furthermore, the information will enable a more personalized and effective treatment of BV and ultimately help prevent adverse sequelae associated with this highly prevalent disruption of the vaginal microbiome.

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.

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.

Data quality control:
Clinical data: All clinical information is collected on teleforms that are scanned and the clinical information directly loaded to a custom-designed database. The data is doubled checked by manual examination by the study coordinator.

Sequence data: 454 pyrosequencing of barcoded 16S rRNA genes: The V1-V2 regions of 16S rRNA genes are being sequenced using Roche/454 FLX and Titanium chemistries. To pass quality control a sequence read must (a) include a perfect match to the sequence tag (barcode) and the 16S rRNA gene primer; (b) be at least 300 bp in length; (c) have no more than two undetermined bases; and (d) have at least 60% match to a previously determined 16S rRNA gene sequence after alignment with NAST. An average 2,000-4,000 reads are generated per sample. The depth of coverage for each community is sufficient to detect taxa that constituted ~0.1% of the community.

III. DATA ANALYSIS AND PUBLICATION PLANS:

Sequence processing/analysis

Each processed 16S rRNA gene sequence was classified at a genus level using the RDP Naïve Bayesian Classifier. Overall, 70% of all sequence reads generated in this study were taxonomically assigned to the genus *Lactobacillus*. The median RDP *Lactobacillus* reads score was 0.94. About 92% of the reads assigned to the genus *Lactobacillus* by the RDP classifier had a score of 0.8 or higher.

Species level taxonomic assignment of Lactobacillus *16S rRNA gene sequences*

Because of the short read lengths obtained by 454 pyrosequencing, the classification of 16S rRNA sequences using phylogenetic approaches is typically limited to the genus-level. However, in studies of vaginal microbiota it is essential to classify *Lactobacillus* at the species level to differentiate the four species of *Lactobacillus* sp. that distinguish kinds of vaginal communities, namely *L. crispatus, L. iners, L. jensenii* and *L. gasserii*.

We used a validated custom algorithm (speciateIT, freely available at speciateit.sourceforge.net) to achieve accurate and rapid species level assignments from short V2 16S rRNA genes sequences generated by 454 pyrosequencing. This algorithm integrates validated hidden Markov models (HMM) for each of the 42 known species of the genus *Lactobacillus*. The sequence reads not assigned to any HMM model were classified as OTUs within the genus Lactobacillus using the DBSCAN clustering algorithm on a three-dimensional projection of unclassified reads using HMM scores.

Validation of the HMM based species assignments for Lactobacillus *sp.*

To validate the HMM-based speciation algorithm we used the dataset of Zhou *et al*. (ISME J (2007) 1:121-33) who sequenced the 8-926 region of 1,892 cloned 16S rRNA genes of *Lactobacillus* sp. from the vagina and assigned them to species of *Lactobacillus* using phylogenetic algorithms. A total of 6 species of *Lactobacillus* were identified. The algorithm developed for the present study was able to correctly classify each species with 98.69-100% accuracy.

Under Aim 3 of the study, we will apply model-based statistical clustering and classification approaches to correlate/associate microbial community composition, functions, metadata, and BV prognostic. Model-based approaches will be used to analyze the data to define baselines of "normal" vaginal microbial community structure and function in individual women, and to assess whether deviations from these as well as habits and practices are predictive of BV.

Statistical analyses

Most studies done to characterize and classify microbial communities have relied on classical statistical tools for grouping and classification such as cluster analysis, principle components analysis, correspondence analysis, and multidimensional scaling, among others. These data analysis methods are dominated by distance-based approaches that have several limitations. First, distance-based methods require reducing the dimensionality of the data in making inferences, which results in loss of information. Second, they depend heavily on the type of distance measures used and the manner in which results are graphically represented. Third, these methods do not provide measures of confidence in the results obtained and it is not possible to objectively compare their performance. Finally, there is no straightforward way to incorporate

metadata associated with the samples analyzed, and this is essential for uncovering important correlates between clinical findings and differences in community species/gene composition and diversity.

Model-based approaches are alternatives that overcome the limitations of distance-based methods outlined above. Model-based approaches directly use the observed raw data, and minimize the loss of information by eliminating the data reduction steps associated with distance-based methods. In the original application we proposed to use Bayesian and likelihood statistical frameworks and model-based approaches to provide a probabilistic measure of confidence in the conclusions. This is done by comparison of different models and methods through the use of model-selection strategies such as Bayes factors, Akaike's information criterion, Bayesian information criterion and decision theory. Model-based approaches also provide a means for the integrated analysis of metadata and data to classify microbial communities and community members. Importantly, these model-based approaches provide a classification tool to assign newly sampled individuals to pre-existing, well-established groups that result from clustering approaches that take both metadata and microbial community composition into account. This provides a framework that is both diagnostic and predictive in terms of the ability to identify microbial communities that are prone to BV.

Ecological Modeling

The data will also be analyzed using a novel approach that integrates network-analysis and game theory to explore microbial community dynamics in the human vagina. The novelty of the approach lies in the innovative application of survival analysis and non-equilibrium statistical mechanics to build species interaction networks and community metabolomic networks. In addition, survival analysis (including multivariate survival analysis) will be used to model population abundance, species diversity distributions, lifetimes of "game" players, species interactions, and stability modeling. It offers flexibility in modeling time- and covariate-dependent frailty and the unique capability to handle incomplete information (censoring). We contend that these approaches carry advantages over existing ecological modeling approaches and constitutes an effective way to create a seamless interface between models used in ecological and biomedical research. We believe that our expertise in applying these models will also benefit the HMP as a whole, as we continue in providing the groups with bioinformatics analysis tools.

The work will be published rapidly. We anticipate that at least one or two publications describing the outcome of these analyses will be generated under this study.

IV. DATA RELEASE PLAN:

NOTE: Due to the nature of the clinical information collected under this study such as sexual behaviors, not all the clinical metadata will be released to dbGAP for this project. However, researchers could contact the PIs to request access to the entire clinical data.

Sharing of data generated by this project is an essential part of our proposed activities and will be carried out in several different ways. We will make the data available to the community of researchers and clinicians. We believe that the proposed work could impact the way scientists approach bacterial vaginosis, but also the way clinicians will treat the disease in the future. Access to the data generated in this proposal will be critical for these two groups to develop new

treatment and guidelines for the treatment of bacterial vaginosis. Furthermore, we worked with the Data Analysis and Coordination Center (DACC) that support the HMP efforts so that data (sequence, metadata and analysis) was available on a timely manner.

Release of clinical data to a controlled access site in NCBI.
Clinical data has been released to a controlled access site via dbGAP. 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. We have complied with HMP guidelines to ensure that consenting is appropriate and allow release of clinical data in a way that it does not compromise participants privacy.

Release of sequence read and trace data.
We have been deposited 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 have addressed this issue by including within the "source" feature an "isolation_source" descriptor that stores the string "Homo sapiens vaginal sample", along with a coded patient id, health or BV (Nugent score) 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 ensured inclusion of the sequences in the Ribosomal Database Project at http://rdp.cme.msu.edu and GreenGenes (http://greengenes.lbl.gov), both of which automatically compiles sequences from GenBank on a monthly basis.

Flow Files. All sequences and trace files (Flows) generated under this proposal have been submitted to the Short Read Archive at NCBI/NLM/NIH. These data will also include information quality values for each sequence.

Release of metadata associated with sequence traces or other types of data. Metadata for this project has been deposited into dbGAP.

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.

Presentations at national/international scientific meetings. Given the scope of the project, we are making presentations at national and international meetings to disseminate the analyses and results obtained in this study. We are representing the project at key national and international meetings so that the scientific community is award of the data generated and knows of our open-access data release plan.

Web-site (http://www.humvdb.org): The web-site is a portal for this project. We are asking visitors who are interested in receiving update on data release to sign up, so that E-mails can be sent to this community to announce new data releases and how to access the data.

Resources sharing plan
The study is generating several types of resources, which we are shared with the scientific and clinical research communities, those includes:

Software and analytical pipelines. We believe in open-access to software and bioinformatic pipeline. IGS has a track record for making all the source code of their software and pipelines available under open-access at sourceforge.org. We have already released speciateIT (speciateIT.sourceforge.net) and our visualization tool, inVUE (invue.sourceforge.net).

Technologies and protocols. The proposed project will make use of new technologies and protocols. These will be changing greatly over the course of the project, as technology evolved. We will first make all our detailed SOP available to the rest of the HMP awardees, and to the rest of the scientific and clinical research communities though the project web-site (http://www.humvdb.org) and publications. We will update our protocols as needed and make them available as soon as validation is completed.

V. CONTACT PERSONS:

Dr. Jacques Ravel, University of Maryland School of Medicine, jravel@som.umaryland.edu
Dr. Larry Forney, University of Idaho, lforney@uidaho.edu