PROJECT SUMMARY

The Vaginal Microbiome: Disease, Genetics and the Environment

Jennifer M. Fettweis[1], Joao P. Alves[1], Joseph F. Borzelleca[1], James P. Brooks[1], Christopher J. Friedline[1], Yuan Gao[1], Xi Gao[1], Philippe Girerd[1], Michael D. Harwich[1], Stephanie L. Hendricks[1], Kimberly K. Jefferson[1], Vladimir Lee[1], Huan Mo[1] Michael C. Neale[1], Federico A. Puma[1], Mark A. Reimers[1], Maria C. Rivera[1], Seth B. Roberts[1], Myrna G. Serrano[1], Nihar Sheth[1], Judy L. Silberg[1], Logan J. Voegtly[1], Elizabeth C Prom-Wormley[1], Bin Xie[1], Timothy P. York[1], Cynthia N. Cornelissen[1], Jerome F. Strauss III[1], Lindon J. Eaves[1], and Gregory A. Buck[1].

[1]Virginia Commonwealth University, Richmond, Virginia 23298-0678

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 IDs: 46877, 46311
Publication moratorium: 12 months

The vagina is an interactive interface between the host and the environment. Its surface is covered by a protective epithelium colonized by bacteria and other microorganisms. The ectocervix is nonsterile, whereas the endocervix and the upper genital tract are assumed to be sterile in healthy women. Therefore, the cervix serves a pivotal role as a gatekeeper to protect the upper genital tract from microbial invasion and subsequent reproductive pathology. Microorganisms that cross this barrier can cause preterm labor, pelvic inflammatory disease, and other gynecologic and reproductive disorders. Homeostasis of the microbiome in the vagina and ectocervix plays a paramount role in reproductive health. Depending on its composition, the microbiome may protect the vagina from infectious or non-infectious diseases, or it may enhance its susceptibility to them. Because of the nature of this organ, and the fact that it is continuously colonized by bacteria from birth to death, it is virtually certain that this rich environment evolved in concert with its microbial flora. Specific interactions dictated by the genetics of both the host and microbes are likely responsible for maintaining both the environment and the microbiome. However, the genetic basis of these interactions in both the host and the bacterial colonizers is currently unknown. *Lactobacillus* species are associated with vaginal health, but the role of these species in the maintenance of health is not yet well defined. Similarly, other species, including those representing minor components of the overall flora, undoubtedly influence the ability of potential pathogens to thrive and cause disease. Gross alterations in the vaginal microbiome are frequently observed in women with bacterial vaginosis, but the exact etiology of this disorder is still unknown. There are also implications for vaginal flora in non-infectious conditions such as pregnancy, pre-term labor and birth, and possibly fertility and other aspects of women's health. Conversely, the role of environmental factors in the maintenance of a healthy vaginal microbiome is largely unknown. To explore these issues, we have proposed to address the following questions:

1. **Do the genes of the host contribute to the composition of the vaginal microbiome?** We hypothesize that genes of both host and bacteria have important impacts on the vaginal microbiome. We are addressing this question by examining the vaginal microbiomes of mono- and dizygotic twin pairs selected from the over 170,000 twin pairs in the Mid-Atlantic Twin Registry (MATR). Subsequent studies, beyond the scope of the current project, may investigate which host genes impact the microbial flora and how they do so.
2. **What changes in the microbiome are associated with common non-infectious pathological states of the host?** We hypothesize that altered physiological (e.g., pregnancy) and pathologic (e.g., immune suppression) conditions, or environmental exposures (e.g., antibiotics) predictably alter the vaginal microbiome. Conversely, certain vaginal microbiome characteristics are thought to contribute to a woman's risk for outcomes such as preterm delivery. We are addressing this question by recruiting study participants from the ~40,000 annual clinical visits to women's clinics of the VCU Health System.
3. **What changes in the vaginal microbiome are associated with relevant infectious diseases and conditions?** We hypothesize that susceptibility to infectious disease (e.g. HPV, *Chlamydia* infection, vaginitis, vaginosis, etc.) is impacted by the vaginal microbiome. In turn, these infectious conditions clearly can affect the ability of other bacteria to colonize and cause pathology. Again, we are exploring these issues by recruiting participants from visitors to women's clinics in the VCU Health System.

Three kinds of sequence data are generated in this project: *i*) rDNA sequences from vaginal microbes; *ii*) whole metagenome shotgun sequences from vaginal samples; and *iii*) whole genome shotgun sequences of bacterial clones selected from vaginal samples. The study includes samples from three vaginal sites: mid-vaginal, cervical, and introital. The data sets also include buccal and perianal samples from all twin participants. Samples from these additional sites are used to test the hypothesis of a per continuum spread of bacteria in relation to vaginal health. An extended set of clinical metadata associated with these sequences are deposited with dbGAP. We have currently collected over 4,400 samples from ~100 twins and over 450 clinical participants. We have analyzed and deposited data for 480 rDNA samples, eight whole metagenome shotgun samples, and over 50 complete bacterial genomes. These data are available to accredited investigators according to NIH and Human Microbiome

Project (HMP) guidelines. The bacterial clones are deposited in the Biodefense and Emerging Infections Research Resources Repository (http://www.beiresources.org/).

In addition to the extensive sequence data obtained in this study, we are collecting metadata associated with each of the study participants. Thus, participants are asked to complete an extensive health history questionnaire at the time samples are collected. Selected clinical data associated with the visit are also obtained, and relevant information is collected from the medical records when available. This data is maintained securely in a HIPAA-compliant data system as required by VCU's Institutional Review Board (IRB). The preponderance of these data (i.e., that judged appropriate by NIH staff and VCU's IRB are deposited at dbGAP (http://www.ncbi.nlm.nih.gov/gap). Selected fields of this data have been identified by NIH staff as 'too sensitive' and are not available in dbGAP. Individuals requiring access to these data fields are asked to contact the PI of this project or NIH Program Staff.

II. DATA QUALITY:

Metagenomic 16S rDNA Sequencing. Our metagenomic rDNA sequences are generated from the PCR amplified V1-V3 (*E. coli* rDNA coordinates: 27-534) region of the bacterial small subunit rRNA gene from DNA isolated from swab samples taken from the mid-vaginal wall, cervix, introitus, perianum and buccal cavity of twin participants from the Mid-Atlantic Twin Registry, or from visitors to women's clinics at the VCU Health Center. The amplified products are sequenced using Roche 454 Titanium technology. The 16S rDNA post-sequencing analysis pipeline is designed to filter the reads based on quality, to trim the barcode (Multiplex Identifier) and PCR amplification primer sequences, and to assign taxonomy to the high quality reads obtained. Our analysis pipeline retains those reads for which: *i)* a valid PCR primer and barcode is detected; *ii)* less than 10% of base calls have a quality score below Q10 *iii)*; average quality of the bases in the read exceeds Q20; and *iv)* the read includes more than 200 but less than 540 bases. Approximately 80% of the reads are retained after the application of these quality control metrics. These high quality reads are assigned taxonomy using the RDP Classifier (*1*) with a minimum bootstrap confidence of 80%. Additional analyses are also applied to specific subsets of reads to identify operational taxonomic units (OTUs).

Whole Metagenome Shotgun Sequencing. Whole metagenome shotgun sequencing; i.e., shotgun sequencing of unamplified total DNA in a sample, is applied to selected vaginal samples from twin and clinical participants as described above. Reads for data that pass minimum quality standards for 454 Titanium sequencing are subjected to a preliminary screen (BLASTN with 1E-20 E-value cutoff and at least 90% sequence identity to the National Center for Biotechnology Information reference human genome build 37) to identify and filter reads from human DNA prior to submission to National Center for Biotechnology Information SRA Protected Database. Taxonomic content and metabolic potential of these samples is assessed as described below.

Bacterial Clone Sequencing. Bacteria of interest are cultured and colony cloned from vaginal samples showing distinctive microbiome profiles. DNA is isolated using standard laboratory protocols and submitted for sequencing on the Illumina GAIIX or Roche 454 Titanium Sequencer. Sequences are assembled using Velvet (*2*), Newbler (Roche Diagnostics) or other appropriate assembler. Sequences of these bacterial clones will meet or exceed 'Standard Draft' quality as defined in Chain et al. (*3*). We have compared the complete genome sequences from a panel of *Gardnerella vaginalis* (*4*), *Lactobacillus sp., Sneathia*, and other isolates (in preparation), taken from our samples.

III. DATA ANALYSIS AND PUBLICATION PLANS:

Metagenomic 16S rDNA taxonomic analysis. The 16S rDNA reads are binned using both taxonomic approaches and operational taxonomic units (OTUs) as described above. These data are then analyzed using a variety of methods. First, an assessment of experimental variation through the analysis of technical replicates is conducted so that true differences in the samples can be correctly confirmed. We then use visualization techniques such as variations of principal component analysis (PCA) (*4*) and stacked bar plots to compare microbiome profiles in samples based on abundant taxa in samples. We are developing statistical techniques for confirming that groups of samples arise from different distributions. Machine learning techniques and techniques grounded in measurement theory are used to identify the organisms/functional groups in microbiome profiles that correlate with clinical metadata. Longitudinal analyses are also being conducted to assess effects of the time in cycle on the microbiome. Experiments are planned to quantify any bias introduced in the 16S rDNA analysis pipeline.

Analysis of whole metagenome shotgun sequence data. After removal of 'contaminating' human sequence (see above), analysis of data from whole metagenome shotgun sequencing is performed in two different dimensions: taxonomic content and metabolic potential. Taxonomic content, which is intended to be compared to that derived from 16S studies, is analyzed by similarity searches (BLASTX [*5*], 1E-6 cutoff) of the filtered reads against the NR database at the National Center for Biotechnology Information, followed by analysis with MEGAN (*6*). Metabolic potential is being analyzed using ASGARD (*7*) in combination with UniProt (*8*) and KEGG (*9*) data to generate lists of functional entities and putatively present pathways, as well as graphical metabolic maps comparing different samples.

Analysis of bacterial genome sequences. Interesting bacterial clones are sequenced and assembled as described above. Genes are called using a variety of gene calling packages. Metabolic reconstructions and genome comparisons are performed using ASGARD in combination with UniProt or KEGG data. Genomes are compared to existing reference genomes, when available, and to genomes of related bacterial isolated from relevant samples; e.g., mono- or dizygotic twins, microbiomes with unexpected or unusual profiles, etc.

Twin Studies. The VCU twin microbiome study is targeting 250-350 pairs of adult female mono- and dizygotic twins ascertained through the population-based Mid-Atlantic Twin

Registry ("MATR"). This sample, unselected for clinical history, will be supplemented by twin pairs ascertained to enrich the representation of significant clinical outcomes. Microbial samples from multiple body sites are obtained and characterized for twin participants, supplemented by metadata gathered by supervised self-report questionnaire obtained at the time of clinic visit. A variety of univariate and multivariate techniques for genetic analysis of twin data have been developed and will be applied to the microbial data to: *i*) estimate the contributions of genetic and environmental factors to differences in prevalence of individual taxa at specific sites; *ii*) identify clusters of bacterial taxa whose co-occurrence depends on shared genetic or environmental risk factors in the host; *iii*) test for genetic and environmental factors that have a general effect on the composition of the microbiome across different sites; *iv*) identify specific environmental factors in the metadata that account for non-genetic differences with monozygotic twin pairs; *v*) test for the interaction of genetic effects in the host with specific environmental exposure on aspects of the human microbiome (genotype-environment interaction); *vi*) identify intermediate behavioral pathways (e.g. sexual behavior, substance use) that mediate the influence of the host genotype through differential exposure to environmental risk (genotype-environment correlation).

Publication Plans. Consortium participants anticipate publication of papers describing the general approaches and methodology employed in this project, general data relevant to the specific aims of the project, and specific findings relevant to the specific milestones and objectives of the participants. Our analyses of these data will be published as soon as reasonably possible following its generation, in general, prior to the publication moratorium end dates. In some cases; e.g., variance components analysis of mono- and dizygotic twin data, analyses will be delayed until sufficient numbers of twin pairs have been recruited for the study. Thus, whereas preliminary data could be published with a suboptimal number of twin pairs, we will adopt a conservative approach to ensure that the conclusions are meaningful and valid. Therefore, to avoid compromising the final product and thus doing compromising the long term objectives of NIH in funding this project, we request that individuals planning to use our data in publications first consult with the PI of the project before beginning their analysis.

IV. DATA RELEASE PLAN:

The VCU Vaginal Microbiome Consortium ascribes to NIH standard data release polices for rapid release of sequencing data such as that generated in this project. The policies are summarized in several web sites; e.g., NIAID policy is found at: http://www3.niaid.nih.gov/research/resources/mscs/data.htm; General NIH guidelines: http://grants.nih.gov/grants/policy/data_sharing/; http://www.genome.gov/10506376. Metagenomic rDNA sequence and whole metagenome sequence data from vaginal samples, and whole genome shotgun data from selected bacterial clones will be deposited in the short read archives at National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi). Metadata associated with these samples will be released to appropriate investigators through dbGAP (http://www.ncbi.nlm.nih.gov/gap). As described above, NIH has requested that certain particularly sensitive fields of the metadata not be released with the standard metadata files. Individuals requiring access to these data are encouraged to contact NIH or the PI of the project.

V. CONTACT PERSON:

Dr. Gregory A. Buck, Virginia Commonwealth University, Richmond, Virginia; email: gabuck@vcu.edu; phone: (804) 828-2318.

REFERENCES CITED

1. Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy, *Applied and Environmental Microbiology 73*, 5261-5267.
2. Zerbino, D.R. and Birney, E. (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*. 2008. 18: 821-829.
3. Chain, P.S.G., Grafham, D.V., Fulton, R.S., et al. (2009) Genome Project Standards in a New Era of Sequencing. *Science* 326: 236-237.
4. Harwich, M., Alves, J., Buck, G.A., Strauss, J.F., Patterson, J.L., Oki, A.T., Girerd, P.H., and Jefferson, K. Drawing the line between commensal and pathogenic *Gardnerella vaginalis* through genome analysis and virulence studies. BMC Genomics. 2010; 11: 375-386.
5. Brooks, JP, Dula, JH, Boone, EL. *(submitted)* "A Pure L1-Norm Principal Component Analysis," in review. Preprint available at http://optimization-online.org/DB_HTML/2010/01/2513.html.
6. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol*. 215(3):403-10.
7. Huson, D.H.,, Auch, A.F.,, Qi, J., and Schuster, S.C. (2007) MEGAN analysis of metagenomic data. *Genome Res*. 17(3):377-86.
8. Alves, J.M.P, and Buck, G.A. (2007) Automated System for Gene Annotation and Metabolic Pathway Reconstruction Using General Sequence Databases. *Chem. Biodivers*. 4: 2593-2602.
9. The UniProt Consortium; The Universal Protein Resource (UniProt) in 2010. (2010) *Nucleic Acids Res.* 38, D142-D148.
10. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004) The KEGG resources for deciphering the genome*. Nucleic Acids Res*. 32, D277-D280.