# SF 424 R&R and PHS-398
# SpecificTable of Contents

| | | |
|---|---|---|
| PI: **STRICKLER, HOWARD D** | Title: Next Generation of HPV and Cervical Cancer Research in HIV+ Women | |
| | FOA: PA16-160 | |
| | FOA Title: NIH Research Project Grant (Parent R01) | |
| | Organization: ALBERT EINSTEIN COLLEGE OF MEDICINE, INC | |
| | Department: Epidemiology & Pop Health | |
| | | |
| *Senior/Key Personnel:* | *Organization:* | *Role Category:* |
| Howard Strickler MD | Albert Einstein College of Medicine | PD/PI |
| Robert Burk MD | Albert Einstein College of Medicine | MPI |
| Joel Palefsky | University of California, San Francisco | Co-Investigator |
| Philip Castle | Albert Einstein College of Medicine | Co-Investigator |
| Xiaonan Xue PhD | Albert Einstein College of Medicine | Co-Investigator |

OMB Number: 4040-0001
Expiration Date: 10/31/2019

# RESEARCH & RELATED OTHER PROJECT INFORMATION

**1. Are Human Subjects Involved?*** ● **Yes** ❍ **No**

1.a. If YES to Human Subjects

Is the Project Exempt from Federal regulations? ❍ Yes ● No

If YES, check appropriate exemption number: 1 __ 2 __ 3 __ 4 __ 5 __ 6If

NO, is the IRB review Pending? ● **Yes** ❍ **No**

IRB Approval Date:

Human Subject Assurance Number 00023382

**2. Are Vertebrate Animals Used?*** ❍ **Yes** ● **No**

2.a. If YES to Vertebrate Animals

Is the IACUC review Pending? ❍ Yes ❍ No

IACUC Approval Date:

Animal Welfare Assurance Number

**3. Is proprietary/privileged information included in the application?*** ❍ **Yes** ● **No**

**4.a. Does this project have an actual or potential impact - positive or negative - on the environment?*** ❍ **Yes** ● **No**

4.b. If yes, please explain:

4.c. If this project has an actual or potential impact on the environment, has an exemption been authorized or an environmental assessment (EA) or environmental impact statement (EIS) been performed? ❍ Yes ❍ No

4.d. If yes, please explain:

**5. Is the research performance site designated, or eligible to be designated, as a historic place?*** ❍ **Yes** ● **No**

5.a. If yes, please explain:

**6. Does this project involve activities outside the United States or partnership with international collaborators?*** ❍ **Yes** ● **No**

6.a. If yes, identify countries:

6.b. Optional Explanation:

**ABSTRACT**

HIV(+) women have high risk of cervical precancer and cancer as well as infection with human papillomavirus (HPV), the viral cause of cervical precancer/cancer. Recent advances in genetic/epigenetic methods provide previously unachievable opportunities to study the HPV viral factors and other local infectious influences on the natural history of HPV and development of cervical disease in HIV(+) women. Under this proposal, we will use next generation (next-gen) sequencing to conduct ***precision HPV genomic analysis*** able to determine whether a given HPV type detected at two or more time points is the same exact viral infection versus different HPVs of the same type. These data will be used to comprehensively study: type-specific differences in HPV persistence and their relation with precancer; the occurrence of HPV reactivation and how often reactivated HPV persists and leads to precancer; the impact of immune status on each of these steps in HPV natural history. If cervical HPV can reactivate and progress to precancer, it would preclude screening cessation at age 65 years (which is done in the general population). Furthermore, ***HPV DNA methylation*** will be studied, as we found methylation in the HPV L1/L2 region very strongly associated with precancer/cancer, and pilot data suggest similar associations between methylation and precancer may exist for HIV(+) women. This research therefore will provide insight into the epigenetic modifications related to the development of cervical disease, and could possibly be used to improve the specificity and positive predictive value of HPV testing. Importantly, the ***cervicovaginal microbiome*** may influence both the HPV natural history and HPV DNA methylation and will also be studied. A pilot study by our group showed reduced HPV prevalence with high relative abundance of *Lactobacillus crispatus* but not other *Lacobacillus species*, and transition to a *L. crispatus* community state type was associated with reduced incident HPV. The *L. crispatus* results were especially promising since the protective effects were observed even with low CD4. However, the microbiota and HPV relationship needs to be studied in appropriately designed HIV(+) cohort studies of adequate size before any future probiotic intervention studies may be considered. There are currently little data regarding the impact of the microbiota on HPV natural history/progression (building on Aim 1). We will also study the microbiota and HPV methylation (building on Aim 2), as local microbiota was shown to alter methylation in neighboring tissue. Overall, these integrated studies address critical aspects of HPV natural history/progression, with significant implications to cancer prevention. This study will utilize semiannually collected specimens from the WIHS, the largest, long term cohort of HIV(+) (N=2793) and high risk HIV(-) (N=975) women. ***The Specific Aims are, in HIV(+) women, to study:*** (i) ***The role of reactivation in HPV natural history and precancer risk, using next-gen "precision" HPV genomic assays; (ii) HPV DNA methylation and its relation with precancer risk; The microbiome's impact on HPV natural history, HPV methylation, and HPV progression.***

## PROJECT NARRATIVE

HIV(+) women have high risk of cervical precancer and cancer as well as infection with human papillomavirus (HPV), the viral cause of cervical precancer/cancer. Advances in genetic/epigenetic methods provide previously unachievable opportunities to study HPV viral risk factors and other local infectious influences on HPV natural history and precancer risk in HIV(+) women. Specifically, under this application we will use recently developed next generation (next-gen) sequencing assays to study, in HIV(+) women: the role of HPV reactivation in cervical disease, the effects of HPV DNA methylation on precancer risk, and the impact of the cervicovaginal microbiome on both HPV methylation as well as cervical precancer.

## FACILITIES AND RESOURCES

### Resources at the Albert Einstein College of Medicine

**Laboratory:**
Dr. Burk has a fully equipped laboratory on the 5th floor of the Ullmann Bldg. (U515) (1,200 sq. ft.), on the main campus at Albert Einstein College of Medicine (EINSTEIN). This laboratory is equipped for modern biochemistry, molecular and cell biology. He also has a dedicated sample-receiving lab (U503) separated from his main laboratory, a dedicated microbiome lab (U508) and a tissue culture room physically separated from the main laboratory with a separate entrance (U513). Within his laboratory he has a fume hood, Sorval RC5B Refrigerated Centrifuge, IEC Centra-7R refrigerated table top centrifuge, 4 Eppendorf microcentrifuges (2 refrigerated), Savant Speedvac Evaporator, 8 DNA thermal cyclers, 24 cu. ft. refrigerator/freezer, 16 cu. ft. -20°C freezer, electrophoresis apparatuses and power supplies, 2 Hybaid Minioven MKIIs for hybridization, Dynatech Ultrawash Plus microtiter plate washer, Dynatech MRX automated plate reader, Digital Capture System attached to PC with a video copy processor, and an Eppendorf Ep96 robotic pipettor. Within the tissue culture room there is a Biohazard hood, tissue culture incubator, inverted microscope, refrigerator, Eppendorf 5415 microcentrifuge and 2 waterbaths. In the sample-receiving lab there are 2 UV workflow hoods as well as an Eppendorf liquid handling robot and multiple centrifuges.

**Animal:**
No animals will be used in this project.

### Resources for Biostatistical Analyses

**Computer:**
Dr. Burk has 2 Macintosh computers, a color monitor and Xerox color printer. Dr. Burk's laboratory has 10 minimacs and 4 laptop computers, monitors and 4 printers; in addition, all computers are connected to the Einstein Internet with wifi access.

*Data Releases, Security and Backups:* Analytical data for this resides on separate severs behind firewalls, and access to these servers is restricted with password protection. All subject identifiers are removed prior to merging of the study data and linked using a unique study identification number. Unique access permissions will be assigned to project personnel based on roles. Data safety and security will be addressed by redundant approaches to data recovery. Cold standby systems are configured to ensure minimal downtime in the event of hardware system failure. For example, Cisco firewalls are used to protect all mission critical servers at EINSTEIN, and Symantec Backup Exec is implemented as an enterprise backup/restore system. Data is backed up onto disk, by way of a storage area network, in addition to backup tapes or discs. The implemented backup schedule includes daily and weekly backups with scheduled offsite storage. Study data is stored in its own SQL database with unique access permissions assigned to study personnel. The web servers are behind a firewall and access is restricted by username and password and IP address. Secure Socket Layer with 128-bit encryption is implemented on the web servers to ensure security during the transmission of data. A Virtual Private Network is available for remote user access to resources via the Internet. The system employs a defense in depth model to safeguard data ensuring that only authorized users can access the network resources.

### Shared Facilities

As a faculty member of the Albert Einstein Cancer Center, the Center for AIDS Research and the Liver Center, Dr. Burk has access to common equipment for Center investigators including DNA synthesizers, automated sequencers, cell sorters, ultracentrifuges, rotors, liquid N2 tanks, densitometer, histology equipment and additional tissue culture and microbiology incubators, HPLC apparatus, Perkin-Elmer spectrophotometer, immunochemical apparatus, liquid scintillation counter, gamma counter, and image analysis core with confocal microscopy. The Einstein-Montefiore CFAR oversees four shared laboratory core facilities and a shared clinical core facility, providing technologies and services in support of research activities at the center, including a biosafety level 3/clinical virology facility. The Einstein Human Genome Program also offers a variety of genomic facilities. These resources and expertise of the members of these groups are freely available and may be utilized as needed.

The Albert Einstein College of Medicine supports a broad array of Shared Scientific Facilities and Cores designed to advance the research efforts of Einstein investigators. It is an institution that is renowned for its commitment to sharing resources among investigators. These Shared Facilities and Cores are operated by

experienced personnel to provide researchers with access to a wide array of techniques and services requiring expensive instrumentation, specialized facilities and/or a high degree of expertise to perform. They are administered by various Einstein Centers and Departments. Apart from an institutional ethos for sharing resources there are numerous shared facilities that are supported by institutional, fee-for-service, and center funds (CFAR, Cancer, Liver, and Diabetes Centers). Shared facilities provide an excellent environment for research and for learning complicated technologies. Some facilities are organized to teach the technology and require the trainee to participate fully so as to become an expert. Others, such as the peptide synthesis facility, provide products in a timely fashion without investigator participation. To facilitate rapid access to these technologies, Shared Facilities and Cores maintain web sites describing the services they provide and how to access them.

Below are samples of the shared facilities that are available for this study.

**Bioinformatics -** The Bioinformatics Shared Resource (BISR) provides a range of services to Einstein investigators. The primary focus is on the management and analysis of genomic and epigenomic data using its WASP (Wiki-based Automated Sequence Processing) System software, which combines web-based sample submission, laboratory information management, administrative, and automated data analysis. This system also facilitates interactions with the Biostatistics, Epidemiology Informatics and CPDMU shared resources to allow integration of clinical and genomic information, and integrative analysis of data.

**Genomics -** The Genomics Shared Resource provides services utilizing current and emerging nucleic acid-based technologies for human and model organism studies. The facility provides a number of technologies for genotyping DNA for known SNPs (single nucleotide polymorphisms). Low throughput genotyping (a few SNPs) is routinely performed using Pyrosequencing, while medium throughput genotyping (30-500 SNPs) is done with the Sequenom iPLEX or Fluidigm BioMark instruments. Genome-wide SNP arrays (Affymetrix 6.0 chip) are performed on the Affymetrix Fluidics workstation. Genomic deletions or duplications are also detected using the Affymetrix 6.0 array. The same workstation is used for gene expression profiling using Affymetrix GeneST 1.0 arrays. Low to medium throughput gene expression profiling or measures of DNA copy number are performed using the ABI 7900 or Fluidigm BioMark instrument. Studies of DNA methylation utilizing a low throughput system to examine a few sites are available with the PSQ96 Pyrosequencing system (Qiagen). Genome-wide assays are done using two Illumina HiSeq2000 instruments. The facility routinely performs traditional Sanger (ABI 3730) as well as next-generation sequencing of DNA and RNA (same 2 Illumina HiSeq2000 instruments; 1 Illumina MiSeq for smaller projects).

**Computational Genomics Core Facility -** The Computational Genomics Core Facility is led by Dr. Fabien Delahaye, Ph.D., who supervises bioinformaticians Robert Dubin and Xusheng Zhang, providing data analysis services to investigators on a fee for service basis. They also develop genomics applications within WASP and are contributing to the development of the new spring version of the Wasp System (see below). Dr. Greally acts as faculty supervisor for this core facility.

**The Computational Epigenomics Group**

The Computational Epigenomics Group consists of several researchers with complementary interests who combine to give us expertise in all levels of data analysis. The primary stages of data analysis consists of transforming raw sequence or fluorescence data into biological information, using the tools provided by the companies making MPS platforms as well as those developed by researchers, such as those of the ENCODE consortium. The secondary stage of analysis involves comparing groups of assays, integrating with other sources of data and visualizing the resulting information, allowing the original and new hypotheses to be explored. The third stage of analysis is statistical, testing whether the observed differences are in fact statistically significant.

The facility's dual computational mission of providing a pipeline level infrastructure (WASP) and an analytical component (IMBAS) is served by a multilayered collaboration with the Computational Genomics and HPC (High Performance Computing) Cores locally and with fellow institutions (for example mskcc.org, nyu.edu, ucsc.edu) and the XSEDE Science Gateway development team.

The Integrated Team Software Development Environment (ITSDE) is currently hosted on a series of physical and VM servers at Einstein. In addition to our locally hosted environments, Amazon Cloud instances had been

initiated and configured, as well as local, private instances using Eucalyptus. The hardware support infrastructure currently consist of two Dell Poweredge R810 servers in highly available configuration which host the development environment, and two Sunfire x4600 servers in highly available configuration to serve the production environment. Central storage for the VM services is provided by Oracle 7410 units configured as NSPF (no single point of failure).

We have three dedicated programmers who develop the software required for the Center. One programmer is focused on database development and maintenance (Dr. Robert Dubin), a second on pipeline construction (Qiang Jing) and a third on the development of the algorithmic components of the pipelines (Dr. Andrew McLellan). Two of these programmers have PhD degrees in biology, a valuable combination of expertise for the Center. Their activities are supervised by Dr. Greally.

**Epigenomics Shared Facility**

The high-throughput molecular technological resources include both microarray and massively-parallel sequencing platforms, although the microarray service is now almost completely phased out. The massively-parallel sequencing (MPS) platforms in the core facility include the Illumina HiSeq 2000, MiSeq and Roche FLX technologies. The core facility space is dedicated and customized, with MPS library preparation performed in a positive-pressure room isolated from the separate 'dirty' room in which tubes containing amplified libraries are opened within a fume hood. The MPS machines are connected by high-bandwidth networking to dedicated computing equipment located in a server room one floor below, thus keeping the computers separate from the molecular biology space. The core facility is staffed by three dedicated and one shared personnel. The faculty supervisor, Dr. Shahina Maqbool, has a Ph.D. and was recruited from SUNY Syracuse. The two-fulltime technicians have Masters degrees and were hired to senior positions based on extensive experience. The part-time technician is shared with the Genomics Core Facility and performs assays that are more focused on genetic rather than epigenetic assays. Sample submission is coordinated through a laboratory information management system (LIMS). This web- based system requires that the user enter data about the sample that will subsequently be used when uploading results to public data repositories. The sample information is linked to a barcode on a tube used for sample submission, allowing tracking of samples through the assay process. Resulting data are linked to the barcode identifier for integration within the supporting relational database.

Quality control and assurance is a critical component of the functions of the Epigenomics Shared Facility. The MPS algorithms include error frequency testing as a function of read length and base composition. Sequence capture analysis includes off-target read frequency measurement. Data analysis is performed as described below, and made available to the investigators through the same LIMS interface (part of the WASP system) used at the time of submission so that the original sample data are linked to the results. The data are not moved from the computing cluster, and the analytical algorithms are performed within the database environment to eliminate data transfer bottlenecks. All primary data analyses are performed automatically on completion of data transfer from the Epigenomics Shared Facility, who assess data quality and release the results to the investigator by email contact.

**High-Performance Computing Core Facility**

The hardware, administrative and database resources of the Center for Epigenomics are described as an informatics core facility. The core facility's sequencing output per week is around 4-6 TB, with an additional weekly ~6 TB of data growth due to secondary analytics. We have dedicated systems administrators and programmers, relational database services and high-performance computing resources. We use the Illumina and Roche clusters that are part of their equipment for primary data processing, linked with an integrated 144 node local cluster (with nodes offering 32 GB-256 GB of RAM) for secondary stages of data analysis and a 500 TB RAID storage facility using a 10 GB storage network. The storage has growth capacity to 1 PB and offers PIs a simplified integration model. For large memory computation the facility installed a 1 TB memory server and an SGI UV1000 NUMA machine with 4 TB of memory. The UV1000 also offers 4 NVIDIA GPU units for FPGA programming, whose capability is used/shared with the Structural Biology Department and the Systems Biology Department. Monitoring and use allocation of the high-performance computing resources is under the supervision of the dedicated systems administrator who also maintains the software and security of the system. The system is based on Sun Grid Engine (UNIVA) software for resource management, and Rocks 5.4 (UCSD) for cluster installation and management. Over hundred and fifty commonly used sequence analysis packages are installed on the cluster, many of which are locally developed tools. We have also segregated a component of the cluster with high-speed RAID drives to host a local mirror of the UCSC Genome Browser,

Blat, Hub, Local Track services and our local MySQL databases, including the database supporting our WASP system. Cluster-integrated Galaxy services are also available for classroom support (Computational Genomics & Epigenomics, Fall 2011) and for small data analysis and visualization. All computational and hardware systems are designed in a modular fashion to ease new equipment acquisition and integration for PIs within and outside the facility. The facility's staff is making an active effort to collaborate with other core facilities and departments to maximize return on hardware investment.

**Grid resources: the Einstein Genome Gateway**

The Einstein Genome Gateway is an NSF funded incubation project, TG-MCB120070, which aims to provide a logical and physical framework for large-scale genome data analysis. The gateway will provide an automated pipeline for data movement from sequencing centers to XSEDE resources, an application layer supporting serial swarms, MPI, OMP, Hybrid, experimental (FPGA) and deep visualization techniques, a data curation and lifecycle management agent, and a federated analytical framework that can tap into multiple XSEDE computational resources and international genomics and epigenetic databases. The current incubation environment is developed on server resources provided by Indiana University via XSEDE and has access to computational and storage resources on Blacklight, Lonestar, Steele, Trestles and Ranger. GridFTP and Aspera services provide large-scale data movement while Einstein's WASP-engine administers policy and permissions for private and collaborative efforts. Pre-productions runs allowed our community to use XSEDE allocations via the gateway for R statistical swarms to scale to 512-1200 CPUs and test runs of SOM engine (self-organizing maps) to appraise parallel scaling and memory use. The portal aims to provide layered and customizable access for investigators who need to transfer and stage data to XSEDE resources and to perform single step or pipeline type multistep analysis. The intelligent provisioning engine, developed by Einstein Genomics Core and Einstein HPC Core provides automated workflow project/job submission based on understanding of investigator identity, XSEDE allocations, computational resource scheduling policies and resource availability on a queue, CPU/core, memory and interconnect basis. The Einstein Genome Gateway core components, the WASP engine, Apache Rave and Apache Airavata are open source development and provisioning tools.

**New York Area-wide Resources**

**New York Genome Center**

The New York Genome Center (NYGC) was founded by 12 Institutional Founding Members (IFMs) of which Einstein in one, and includes three Associate Members (with more institutions in the process of joining). The NYGC has a sequencing operation with several next-generation sequencers, extensive bioinformatics support and a large IT infrastructure for data storage and high-performance computation. NYGC has all of the necessary expertise to take a biological sample and complete the required steps to create a fully assembled and interpreted sequence (including clinical interpretations). NYGC provides sequencing services for the IFMs and Associate Members as well as other scientific, clinical, and pharmaceutical organizations. This organizational, intellectual, and technological infrastructure will support advanced development and enhancement of emerging informatics technologies. NYGC has established formal structures to foster collaboration among member institutions and among researchers at these institutions as well as to disseminate knowledge among collaborators.

Dr. Burk and this project will have access through the unique arrangement of Einstein as one of the founding members of this center.

**Howard Strickler, MD, MPH**
**Einstein/Montefiore Facilities and Other Resources.**

1. **Epidemiology Facilities and Other Resources**

Laboratory: Not Applicable

Clinical: Not Applicable

Animal: Not Applicable

Computer: Dr. Strickler and his staff all have advanced PC computers with word processing and statistical software (e.g., SAS, StatCalc, SPSS, etc.) and Laser printers. They have e-mail accounts, access to an FTP system through their PCs, and internet access through Alnet.

Office: The Department of Epidemiology & Population Health is located on the 13th floor of the Belfer Building at AECOM. The department has more than 40 faculty members with primary appointments (and another 51 with status-only appointments). All key personnel/coinvestigators have private offices in their departments. The staff engage in epidemiologic and statistical research on a wide range of topics, and there is an active Master's level program in clinical research methods, which currently involves 25 Master's level students, as well as a PhD program.

*Epidemiology Study Management and Informatics Resource (ESMIR)* - provides expert support to investigators in operationalizing their studies and managing them once they are in the field. The ESMIR currently serves 42 different epidemiology studies in the Department of Epidemiology and Population Health. It provides assistance with data collection instrument design, data entry, creation of data dictionaries, participant enrollment and scheduling systems, specimen and data tracking. Quality Assurance Systems include error reporting, the creation of standardized progress and status reports to monitor data quality, and adverse event reporting. The EISR also plays a major role in additional field services, such as the creation of the Manual of Operations, direct field supervision of data collection operations, Laboratory Information Systems, Multi-Center and Consortium Web-Based Collaboration Systems, as well as the creation of web-based automated QA systems, automated querying and reporting services, and web-based image annotation. Furthermore, database design and custom programming solutions for data manipulation and management are implemented upon request. An extensive web-based Data Collection and Management System (DCAMS) has been designed by the ESMIR.

Dr. Strickler is Leader of the Cancer Epidemiology Program of the Einstein / Montefiore Cancer Center and has access to all cores and facilities of the center. This includes the biostatical core, should there be a need for additional expert assistance, as well as bioinformatics services.

**2. Other Resources**

Montefiore Medical Center is a major New York City institution and acts as the primary clinical affiliate of AECOM. Other major Einstein affiliates include Jacobi Medical Center, and Long Island Jewish Medical Center, and White Plains Hospital.

OMB Number: 0925-0001
Expiration Date: 10/31/2018

## 1. Human Subjects Section

Clinical Trial? ❍ Yes ● No

*Agency-Defined Phase III Clinical Trial? ❍ Yes ❍ No

## 2. Vertebrate Animals Section

Are vertebrate animals euthanized? ❍ Yes ❍ No

If "Yes" to euthanasia

Is the method consistent with American Veterinary Medical Association (AVMA) guidelines?

❍ Yes ❍ No

If "No" to AVMA guidelines, describe method and proved scientific justification

## 3. *Program Income Section

*Is program income anticipated during the periods for which the grant support is requested?

❍ Yes ● No

If you checked "yes" above (indicating that program income is anticipated), then use the format below to reflect the amount and source(s). Otherwise, leave this section blank.

| *Budget Period | *Anticipated Amount ($) | *Source(s) |
|---|---|---|

## PHS 398 COVER PAGE SUPPLEMENT

4. Human Embryonic Stem Cells Section

*Does the proposed project involve human embryonic stem cells? ❍ Yes ● No

If the proposed project involves human embryonic stem cells, list below the registration number of the specific cell line(s) from the following list: http://grants.nih.gov/stem_cells/registry/current.htm. Or, if a specific stem cell line cannot be referenced at this time, please check the box indicating that one from the registry will be used:

Specific stem cell line cannot be referenced at this time. One from the registry will be used.

Cell Line(s) (Example: 0004):

5. Inventions and Patents Section (RENEWAL)

*Inventions and Patents: ❍ Yes ❍ No

If the answer is "Yes" then please answer the following:

*Previously Reported: ❍ Yes ❍ No

6. Change of Investigator / Change of Institution Section

❑ Change of Project Director / Principal Investigator

Name of former Project Director / Principal Investigator

Prefix:

*First Name:

Middle Name:

*Last Name:

Suffix:

❑ Change of Grantee Institution

*Name of former institution:

# PHS 398 RESEARCH PLAN

OMB Number: 0925-0001
Expiration Date: 10/31/2018

| | |
|---|---|
| **Introduction** | |
| 1. Introduction to Application (Resubmission and Revision) | |
| **Research Plan Section** | |
| 2. Specific Aims | Specific_Aims_strickler__10_4_17aa1027374542.pdf |
| 3. Research Strategy* | WIHS_HPV_Research_Strategy1027374732.pdf |
| 4. Progress Report Publication List | |
| **Human Subjects Section** | |
| 5. Protection of Human Subjects | PROTECTION_OF_HUMAN_SUBJECTS1027333118.pdf |
| 6. Data Safety Monitoring Plan | |
| 7. Inclusion of Women and Minorities | Inclusion_of_Women_and_Minorities1027374727.pdf |
| 8. Inclusion of Children | Inclusion_of_Children1027333099.pdf |
| **Other Research Plan Section** | |
| 9. Vertebrate Animals | |
| 10. Select Agent Research | |
| 11. Multiple PD/PI Leadership Plan | Multiple_PI_Leadership_Plan1027333101.pdf |
| 12. Consortium/Contractual Arrangements | Consortium_Arrangement1027374428.pdf |
| 13. Letters of Support | WIHS_LOS2_Strickler_W170271027333138.pdf |
| 14. Resource Sharing Plan(s) | Data_Sharing_Protocol1027333108.pdf |
| 15. Authentication of Key Biological and/or Chemical Resources | Authentification2b1027374746.pdf |
| **Appendix** | |
| 16. Appendix | |

**RESEARCH STRATEGY**

HIV(+) women have high risk of cervical precancer and cancer as well as infection with oncogenic human papillomavirus, the viral cause of cervical precancer/cancer. Recent advances in genetic/epigenetic methods provide previously unachievable opportunities to study the HPV viral factors and other local infectious influences on the natural history of HPV and development of cervical disease in HIV(+) women. Under this proposal, we will use next generation (next-gen) sequencing to conduct ***precision HPV genomic analysis*** able to determine whether a given HPV type detected at two or more time points is the same exact viral infection versus different HPVs of the same type. These data will be used to comprehensively study: type-specific differences in HPV persistence and their relation with precancer development; the occurrence of HPV reactivation and how often reactivated HPV persists and progresses to precancer (i.e., the role of reactivation in clinically relevant disease); the impact of immune status on each of these steps in HPV natural history; and whether HIV impacts precancer risk beyond its effects on HPV. If cervical oncogenic HPV can reactivate and progress to precancer, it would preclude screening cessation at age 65 years (which is done in the general population). Furthermore, ***HPV DNA methylation*** will be studied, as we found methylation in the HPV L1/L2 region very strongly associated with precancer/cancer (e.g., ORs>20; AUC>0.85), and pilot data suggest similar associations between methylation and precancer may exist for HIV(+) women. This research could potentially help improve the specificity and positive predictive value (PPV) of HPV testing, as well as provide insight into the epigenetic modifications related to the development of cervical disease. Importantly, the ***cervicovaginal microbiome*** may influence both the natural history of HPV and HPV DNA methylation. A pilot study by our group in only N=40 HIV(+) and 20 HIV(-) women showed significantly reduced HPV prevalence with high relative abundance of *Lactobacillus crispatus* but not other *Lactobacillus* species. In addition, transition to a *L. crispatus* community state type was associated with reduced incident detection of HPV. The results were especially promising for HIV(+) women as the protective effects of *L crispatus* were observed even in those with a low CD4 count. However, the impact of the cervicovaginal microbiota on HPV in HIV(+) women needs to be studied comprehensively in appropriately designed prospective/longitudinal studies of adequate size before any future probiotic intervention studies may be considered. There are currently little data regarding the impact of the microbiota on HPV persistence, reactivation, and risk of precancer (building upon Aim1) in any population. We will also study the influence of the cervicovaginal microbiota on HPV methylation (building upon Aim 2), as multiple reports show that the local microbiota affects the local milieu and can alter methylation patterns in neighboring tissue. Overall, these planned integrated studies address critical aspects of HPV natural history/progression, with significant implications to cervical cancer prevention in HIV(+) women.

This proposed research will utilize semiannually collected specimens and data from the Women's Interagency HIV Study (WIHS) cohort of HIV(+) (N=2793) and high risk HIV(-) (N=975) women, the largest cohort of US HIV(+) women. We note that it is unlikely that any other HIV(+) cohort has the extensive specimens/data and long term follow-up necessary to conduct the proposed comprehensive studies. We have conducted considerable HPV research in the WIHS, and these studies have significantly influenced CDC/NIH cancer screening guidelines in HIV(+) women.

This study will address the following Specific Aims:

1. ***To use next-generation (next-gen) "precision" HPV genomic analysis to study HPV natural history (incident detection, reactivation, persistence), and the development of cervical precancer (see definition, C.3.) in HIV(+) women –*** Through the use of next-gen HPV assays able to determine whether an HPV type detected at two or more time points is the same exact HPV infection, this is intended to be the definitive study of HPV natural history/progression in HIV(+) women, with important clinical implications; e.g., if reactivation is common and can lead to precancer, then cervical cancer screening cannot be stopped after age 65 years (as in the general population).

2. ***To study HPV CpG site methylation and its relation with incident cervical precancer in HIV(+) women –*** Longitudinal studies of appropriate size are needed to determine (i) temporality; e.g., was the methylation present when the HPV was first detected (perhaps a baseline characteristic of the virus and potentially useful in early patient risk stratification), vs detectable only at the time of precancer (but useful in detecting current precancer), and (ii) are these associations strong enough to enhance cancer screening?

3. ***To study the cervicovaginal microbiome and its impact on HPV and precancer risk in HIV(+) women*** – This will be one of the first prospective studies with serial microbiome testing and its association with precancer in *any population*. We will study the relation of the microbiome with HPV incident detection, persistence, and reactivation (building on Aim1), and with development of HPV methylation (building on Aim 2). Our pilot data raised the possibility probiotics may have clinical use in control of HPV infection.

## A. SIGNFICANCE / SCIENTIFIC PREMISE

The incidence of invasive cervical cancer is increased several-fold in women with HIV/AIDS relative to the general population,[1-3] as is the prevalence, incident detection, and persistence of both cervical precancer and oncogenic human papillomavirus (oncHPV) infection, the viral cause of cervical precancer/cancer. Each of these associations increases significantly with diminishing CD4+ T-cell count.[1-15] Nonetheless, most oncHPV infection, even in HIV(+) women, eventually resolves. Beyond immune status and oncHPV, there are few well established risk factors for cervical precancer/cancer; e.g., mainly smoking,[16-18] high number of live births, and possibly oral contraceptives and chlamydia.[18-20] Why a certain subset of oncHPV infections persist and progress to precancer/cancer is a major gap in our knowledge. It is also an important clinical concern with implications for screening and risk stratification, especially as most HIV(+) women in North America and Europe are beyond the age to receive HPV vaccine, and generations of HIV(+) women in limited resource settings have yet to be vaccinated. Under this proposal, we will use next-generation (nex-gen) assays to conduct novel studies of oncHPV natural history in HIV(+) women, including (i) oncHPV reactivation and its role in cervical precancer; (ii) the oncHPV DNA epigenetic (methylation) changes that may increase precancer risk; and (iii) the cervicovaginal microbiome which may impact both HPV methylation and HPV natural history/progression in HIV(+) women. These integrated studies address critical aspects of HPV infection and cervical disease, with significant implications to cancer prevention in HIV(+) women.

**HPV Type-Specific Differences in the Impact of HIV.** *In the general population*, one HPV type, HPV16, by itself accounts for approximately half of all cervical cancers, and HPV18 another 10-15%,[21-23] and together with other oncogenic HPV types (oncHPV; i.e., HPV31, 33, 39, 45, 51, 52, 56, 58, 59, 68) account for nearly all cervical cancer and precancer.[21,22] Prior data from our group based in the Women's Interagency HIV Study (WIHS) have shown that there are type-specific differences in the effects of host immune status on the natural history of HPV. In particular, the prevalence of HPV16, was found to be the least affected by HIV-status and CD4 count of any oncHPV.[3] This "relative independence" from host immune status has been interpreted as evidence that HPV16 may have an innate ability to avoid the effects of immune surveillance, which could partly explain HPV16's high prevalence in the general population and its unique oncogenicity. Recently, we reported on the clinical significance of this, namely, that the prevalence of HPV16 was significantly lower in HIV(+) than HIV(-) cervical precancers.[4] Meta-analyses by other groups also found that HIV(+) women had reduced HPV16 in cervical precancer (using data from around the world)[5] and in cervical cancer (using data from Africa).[6]

**Importance of HPV Redetection (Reactivation and Reinfection).** Prior studies in WIHS showed evidence of HPV reactivation and its relation with host immune status.[4] Specifically, we evaluated women who were currently celibate for at least 18 months and a comparison group of sexually active women. While the rate of incident detection (i.e., HPV types not present at earlier visits) was higher among sexually active women, it was also high among the women who had been sexually inactive for at least 18 months (Table, above); a rate more than half that observed in the sexually active group, and which correlated strongly with host immune status. Thus, this study showed that a source (or sources) of HPV unrelated to recent sexual activity, such as ***reactivation*** of previously acquired HPV, may account for a substantial fraction of "incident detection" of HPV in HIV(+) women. Other research groups have since published similar data, although it remains unclear whether these reactivation / redetection events represent reactivation from a quiescent state, or some formal latency analogous to that of EBV.[24-26] Furthermore, ***reinfection*** through sex with the same partner over time, may also represent an important part of redetection and the burden of HPV in HIV(+) women, given their high susceptibility to infection. Relevant prospective data in HIV(+) women remain limited. One prospective study by the HERS cohort (N=898),[27] found significantly higher redetection among HIV(+) than HIV(-) women that increased with diminishing CD4. However, HERS did not differentiate redetection due to reactivation vs reinfection. Studies of transplant patients have provided additional evidence of redetection and its relation with immune status, including several reports of higher type-specific HPV redetection in patients with chronic renal disease who did vs did not receive a transplant and immunosuppressive medication.[28] There have also been several studies of reactivation in the general population,[29-35] which elegantly demonstrated reactivation following serial negative HPV tests, using high frequency testing, but had short (e.g., 6 month) follow-up or used HPV antibodies as indicators of prior infection (serology has ~50% sensitivity and antibodies may wane with age). ***We hypothesize that reactivation is a major (albeit, little studied) source of HPV and precancer risk in HIV(+) women, in addition to other forms of redetection (e.g., reinfection). The***

Incident detection of HPV in sexually active and inactive women

| Sexual Activity | HIV-Positive Women | | | HIV-Negative Women |
|---|---|---|---|---|
| | <200 | 200-500 | >500 | |
| Sexually active for ≥18mo | 13/42 (31%) | 25/141 (18%) | 14/112 (13%) | 14/174 (8%) |
| No sexual activity for ≥18mo | 7/32 (22%) | 6/65 (9%) | 3/46 (7%) | 2/43 (5%) |

***proposed research will be the first large, long term study of HPV reactivation in any population (distinguishing reactivation from reinfection), assessing the rate and repetition of reactivation and reinfection events, the persistence of these HPV, as well as their role, if any, in cervical precancer, and the impact of host immune status on these endpoints.*** Characterizing reactivation and reinfection in HIV(+) women would provide new insight into HPV biology. Moreover, if cervical oncHPV can reactivate and cause cervical precancer, it would preclude screening cessation at age 65 (which is done in the general population).

**HPV DNA Methylation.** Viral genome methylation can influence viral protein expression, replication, susceptibility to immune surveillance, host cell proliferation, and risk of virus-related disease. DNA methylation occurs preferentially in regions where a cytosine occurs next to a guanine nucleotide linked to one another by a phosphate bond (termed a CpG site). As we recently reviewed,[36] although there are no classical CpG islands within the HPV genome, regions of high density and conservation of CpG sites across HPV types suggest the potential for a functional role. In earlier studies, we examined 32 sites in the L1 ORF and URR and found an association between increased DNA methylation and CIN3+ (N=23) in the L1 region.[37] An independent replication study of a larger number of CIN-2+ samples (n = 273), [38,39] was confirmatory, including a CpG site in L1, with sensitivity=91% / specificity=60%, and we subsequently,[40] reported similar results for HPV31 (AUC=0.81), HPV18 (AUC=0.85) and HPV45 (AUC=0.98). Recent studies published by other groups further supported the role of L1 and L2 methylation as a potentially useful biomarker for screening.[41-46] In particular, a series of reports by Lorincz and colleagues showed similar associations as our group for HPV16, 18, and 31.[41,42] Our most recent study used a new next-gen sequencing approach for measuring HPV methylation levels, which is more practical for high throughput laboratory use, and which we showed has high correlation with more labor intensive pyrosequencing test results (e.g., r=0.92 for HPV16 methylation)**.**[47]

However, in HIV(+) women, precancer is caused by a significantly greater diversity of oncHPV types. Therefore, in preparation for the current proposal we recently developed next-gen methylation assays for all oncHPV types focusing on CpG sites homologous to those previously associated with precancer**.** Our cross-sectional pilot data in (N=72 cases) HIV(+) WIHS women showed precancer associations across these CpG sites consistent with our prior studies (but involving a broader range of HPV types). Controls were individually matched 1:1 to cases on HPV type, CD4, race. Significant case-control differences (P<0.05 unadjusted) in L1/L2 methylation were observed for individual alpha-9 HPVs (HPV16/31/ 35/58) and alpha-7 HPVs (18/39/45), based on comparison of medians and tertiles (T3 vs. T1 and T2). In addition, to study the biologic relevance of homology across L1/L2 CpGs of phylogenetically-related HPV types, we combined tertile data of homologous sites of alpha-9 and of alpha-7 species types, and showed strong highly significant associations with precancer (Table, right). **Herein we propose the first prospective study to determine the HPV methylation patterns associated with incident cervical disease in HIV(+) women; whether these methylation patterns predict precancer years in advance (useful for early risk stratification) or mainly reflect current disease (but useful for detecting current precancer).** Given the several strong precancer cancer associations we have observed this research could aid efforts to improve the specificity and positive predictive value (PPV) of HPV testing in HIV(+) women.

| HPV Type / Region / CpG Site | OR (95% CI) | P value* |
|---|---|---|
| HPV16L2_1:4240 // HPV35L2_1:4214 | 5.84 (1.82-18.76) | 0.0030 |
| HPV16L2_1:4270 // HPV35L2_1:4244 // HPV52L2_1:4292 | 4.85 (1.68-14.02) | 0.0036 |
| HPV16L1_1:5608 // HPV31L2_1:5521 // HPV35L2L1_2:5570 | 7.48 (2.85-19.60) | 4.27E-05 |
| HPV16L1_1:5611 // HPV31L2_1:5524 // HPV35L2L1_2:5573 | 4.29 (1.74-10.60) | 0.0016 |
| HPV16L1_1:5617 // HPV31L2_1:5530 // HPV35L2L1_2:5579 | 5.27 (2.10-13.25) | 0.0004 |
| | | |
| HPV18L1_2:7041 // HPV45L1_2:7045 | 5.50 (1.15-26.41) | 0.0332 |
| HPV18L1_2:7062 // HPV45L1_2:7066 | 8.00 (1.52-42.04) | 0.0140 |
| HPV18L1_2:7068 // HPV45L1_2:7072 | 5.50 (1.15-26.41) | 0.0332 |
| OR, Odds Ratio; CI, Confidence Interval; *, univariate regression (T3 vs. T1+T2) | | |

**Cervicovaginal Microbiome.** There has long been evidence that the cervicovagional microbiome affects HPV natural history. For example, bacterial vaginosis (BV) is characterized by low abundance of *Lactobacillus* spp., high pH, immune cell infiltration, and has been associated with increased risk of HPV infection in the general population. A prior study by our group showed that BV was also associated with HPV in HIV(+) women, even after accounting for immune status (e.g., CD4 count) and other cofactors.[48] *Lactobacillus* spp. produce lactic acid, and vaginal pH significantly decreases as the relative abundance (RA) of *Lactobacillus* spp. (as a group) increases.[49] This may be important as low vaginal pH has been associated with low HPV detection.[50] There is however a paucity of studies of HPV in HIV(+) women that utilized next-gen sequencing to comprehensively characterize the cervicovaginal microbiome. We therefore conducted a small cross-sectional pilot study in the WIHS[51] Briefly, 16S ribosomal RNA gene amplicon pyrosequencing and HPV DNA testing were conducted annually in serial cervicovaginal lavages obtained over 8–10 years from African American WIHS women, of whom 22 were HIV(-), 22 were HIV(+) with a stable CD4+ > 500 cells/mm$^3$, and 20 were HIV(+) with

progressive immunosuppression. The RA of *L. crispatus* and other *Lactobacillus* species were inversely associated with vaginal pH (all P<0.001). High (vs low) *L. crispatus* RA was associated with decreased HPV (odds ratio [OR]=0.48; 0.24–0.96; Ptrend=0.03) after adjustment for repeated observations and multiple covariates, including pH and study group. However, there were no associations between HPV and the RA of *Lactobacillus* species as a group, nor with *L gasseri, L iners*, or *L jensenii* individually. Overall, the data suggested *L. crispatus* may have a beneficial effect on the burden of HPV in both HIV(-) and HIV(+) women, including those with low CD4. Other studies in the general population have also found inverse associations between *L. crispatus* and HPV[52] or other STIs[53-55]; although it must be noted that low rates of active STIs (e.g., Chlamydia, GC, etc.) in WIHS make it difficult to study STIs other than HPV as an endpoint or covariate.

**Few studies of the cervicovagional microbiome and HPV *in any population* have involved serial / longitudinal microbiome testing. Moreover, none to our knowledge, prospectively studied clinically relevant endpoints (e.g., risk of incident precancer) (building upon Aim 1). Further, multiple studies report that the microbiota effects the local milieu and tissue methylation. We will therefore also study the influence if any of the cervicovaginal microbiota on HPV methylation (building upon Aim 2).[56-59]** If cohort data suggest that *L. crispatus* or other bacteria are protective against HPV/precancer in HIV+ women, probiotic clinical trials (e.g., *L. crispatus* in a vaginal suppository) would be warranted, especially as the effects of *L. crispatus* may be independent of immune status.

**B. INNOVATION**

Recent advances in genetic/epigenetic methods provide previously unachievable opportunities to study the HPV viral factors and other local infectious influences on the natural history of HPV and development of cervical disease in HIV(+) women. We will use recently developed next-gen sequencing methods to conduct ***precision HPV genomic analysis*** able to differentiate HPV at the isolate level to comprehensively characterize the full natural history of HPV from initial infection through to the development of cervical precancer. Increasing evidence from our group and others suggest that reactivation may be a major component of HPV natural history and the burden of HPV in HIV(+) women. However, little is known regarding the frequency of HPV reactivation, its persistence, and whether it plays a significant role in precancer. While all sources of redetection in HIV(+) women, whether due to reactivation or reinfection are important, our approach will allow us to estimate the proportion due to each of these two forms of redetection. Our prior data in the WIHS has played a significant role in current CDC/NCI cervical cancer screening guidelines in HIV(+) women.[60-64] If as we hypothesize cervical oncHPV can reoccur and cause cervical precancer it would provide new biologic understanding of HPV and, most importantly, preclude cessation of screening at age 65 (which is done in the general population); a major area of uncertainty for the CDC/NIH HIV(+) guidelines committee. Furthermore, ***HPV DNA methylation*** will be studied using recently developed assays, as we found methylation in the HPV L1/L2 region very strongly associated with precancer/cancer, and pilot data suggest similar associations between methylation and precancer may exist for HIV(+) women. This research could potentially help improve the specificity and positive predictive value (PPV) of HPV testing, as well as provide insight into the epigenetic modifications related to the development of cervical disease. Importantly, the **cervicovaginal microbiome** may influence both HPV natural history/progression, including HPV reactivation (building upon Aim 1), and HPV methylation (building upon Aim 2). If as in our pilot study, *L. crispatus* or other bacteria are shown to be significantly protective against oncHPV infection in large prospective HIV+ cohort studies, randomized clinical trials of these bacteria (e.g., a probiotic suppository) would be warranted. The clinical implications could be especially important for HIV-infected women, as the effects of *L. crispatus* on oncHPV appeared to be independent of host immune status. *In summary, the proposed research will provide novel data regarding the HPV viral factors and local bacterial influences on the natural history of HPV and development of cervical precancer in HIV(+) women, with implications to our understanding of HPV biology, cancer risk, and cervical cancer screening/prevention.* ***Both enhancing scientific understanding and/or clinical implications are established criteria for significance in NIH review****. This study will exploit the unique longitudinally collected cervical specimens and long term follow-up in the WIHS, as few if any other cohorts have the large sample size, detailed clinical data, and follow-up to conduct this research.*

**C. APPROACH**

***C.1. PRELIMINARY DATA*** - *The Preliminary Data are divided into sections: (C.1.1.) The WIHS Cohort; (C.1.2.) Selected Studies of HPV/Cervical Dysplasia in WIHS; (C.1.3.) Selected Studies of HPV methylation; and (C.1.4) Selected cervicovaginal microbiome studies.*

### C.1.1. Description of WIHS Cohort (see C.2. for WIHS methods)

**Baseline Characteristics of the WIHS Cohort**

| | Enrolled 1994/95 | Enrolled 2001/02 |
|---|---|---|
| HIV + | n = 2055 | n = 738 |
| HIV - | n = 569 | n = 406 |
| Age (med) | 36 | 31 |
| Race | | |
| Black | 55% | 56% |
| Hispanic | 25% | 30% |
| White | 18% | 10% |
| Other | 2% | 4% |
| Consortium | n | n |
| Bronx/NYC | 539 | 234 |
| Brooklyn | 397 | 214 |
| Wash, DC | 395 | 170 |
| Los Angeles | 538 | 226 |
| San Fran | 425 | 159 |
| Chicago | 334 | 141 |
| *HIV(+) Subjects* | | |
| HAART | NA | 42% |
| CD4+ | | |
| Median | 329 | 492 |
| IQR | 163, 515 | 332, 696 |

The WIHS is a large, geographically and ethnically diverse population of HIV(+) (n=2,793) and HIV(-) (n=975) women, who are followed every 6 months with questionnaires, general physical and pelvic examination with Pap smear, and collection of exfoliated cervical cells for HPV testing. The HIV(+) and HIV(-) subjects were enrolled from similar clinical and outreach sources at each of six clinical consortia (Table, right). Two-thirds of subjects were enrolled during 1994-1995, and long term follow-up of these women has been achieved (e.g., 80% of HIV(+) subjects who were alive at 7 years). HIV(+) and HIV(-) WIHS subjects were shown to be comparable in age, race/ethnicity, education, income, number of sexual partners in the last 6 months, and substance abuse. Moreover, the resulting HIV(+) cohort was found to be similar in terms of risk behavior, race/ethnicity, and other demographic factors to national AIDS cases among U.S. women.[65] Thus, the WIHS has truly been a representative sample of HIV/AIDS cases among women in this country, and the continued follow-up of women in the WIHS will reflect the experience of HIV(+) women as they age. Subjects enrolled in 2001/02 (to expand the WIHS) share similar characteristics. The Table (right) also shows a comparison of those enrolled in 1994/95 and 2001/02. Widespread use of HAART in the US began in 1996-97, and this was reflected in the WIHS. The earliest four visits (1994 plus 3 follow-up visits = 1½ yrs) of WIHS were during the pre-HAART era, providing data across a range of host immune status, important to our analysis of the effects of HIV, and HAART on the natural history of HPV. All other visits and the 2001/02 cohort baseline were during the HAART era.

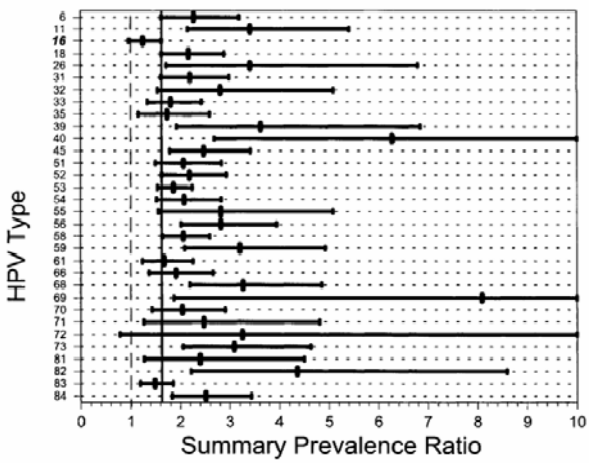
The WIHS has recently added several new clinical sites, which are not included as they have not yet had significant follow-up time or many precancer cases.

### C.1.2. Selected Prior Studies of HPV and Cervical Dysplasia in WIHS

***(i) Risk of cervical precancer in HIV+ women who test oncHPV+ despite a normal Pap*** **(Keller et al, CID, 2015)[60]-** These data played an significant role in the decision of the CDC/NIH HPV guidelines task force to change the guidelines; i.e., to state, that in HIV(+) women aged>30 years, *"Those who are Pap test normal but positive for HPV should have repeat co-testing in one year. If the initial HPV results identify HPV16 or HPV16/18, then referral to colpscopy is recommended."*[66] Briefly, the 5-year CIN-2+ cumulative risk in the HIV(+) oncHPV+ women was 22%, 12%, and 14% among those with CD4 counts <350, 350–499, and ≥500 cells/μL, respectively, whereas it was 10% in those without HIV. For CIN-3+, the cumulative risk averaged 4% in HIV(+) women positive for "any oncHPV", and 10% among those positive for HPV type 16. In HIV(+) women with LSIL, CIN-3+ risk was 7% (similar to HPV16). In multivariate analysis, HIV(+) HPV16(+) women had 13-fold (P = .001) greater CIN-3+ risk than oncHPV-negative women (referent), and HIV(+) women with LSIL had 9-fold (P < .0001) greater risk. Thus, while the prevalence of HPV16 is less altered than other oncHPV by low CD4, it still remains extremely oncogenic among those who do become infected with HPV16.

***(ii) Low Risk of Cervical Pre-cancer in OncHPV(-), Cytologically Normal HIV(+) Women*** **(Keller et al, JAMA, 2012)**[61] We examined whether a 3-year or 5-year interval until re-screening might be safe in HIV(+) women who have a normal Pap who test negative for oncHPV, similar to current guidelines for HIV(-) women. Among the 738 HIV(+) and 406 HIV(-) WIHS women enrolled in 2002, 50% of (369) HIV(+) and 63% (255) of HIV(-) women had a normal Pap and were oncHPV(-). Over 5 years, the cumulative incidence of CIN-2+ (histology) was low, and only one HIV(+) and one HIV(-) woman had CIN-3. None had cancer. These data played an important role in the decision of the CDC/NIH task force to change the guidelines to state *"Co-test negative HIV(+) women (i.e., a normal Pap and negative HPV test) can have their next cervical cancer screening in 3 years"*[66] Likewise, in another study, we assessed HPV data in HIV(+) women with borderline cytologic abnormalities (ASC-US), and showed that *HPV co-testing had 94% sensitivity and 64% specificity for CIN-2+, leading to the addition of HPV triage for ASC-US cytology to the guidelines* **(D'Souza, AIDS, 2014)**.[62]

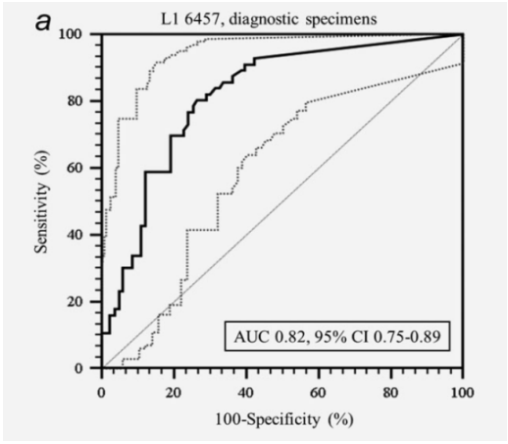
***(iii). The Weak Association of HPV 16 with Immune Status in HIV+ Women***.[67] Using data from multiple follow-up visits we assessed the possible weak association of HPV16 with immune status **(Strickler, JNCI, 2005; Xue, CEBP, 2010).**[67,68] We used prevalence ratios (PRs) to compare the prevalence of each HPV type in HIV(+) women with CD4 <200 to those with >500 T-cells/mm$^3$. The PR for HPV16 in WIHS was low compared with that of most other HPV types at every study visit. Because of the potential importance of this observation, we then confirmed these findings in the HIV Epidemiology Research Study (HERS) before reporting our results. As in the WIHS, the PR for HPV16 was low compared with that of most other HPV types. The GEE



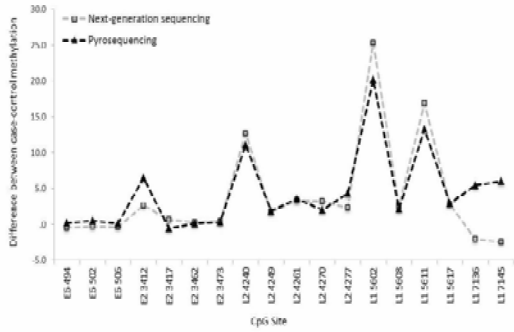

Summary PRs across visits and cohorts (rectangles) and 95% CIs (horizontal lines) for each HPV type are shown (Figure, left). The Summary PR for HPV16 (PR=1.25; 0.97-1.62) was the lowest measured, and it was significantly different (p=0.01) than that of all other oncHPV as a group. This relative independence from immune status has since been interpreted as evidence that HPV16 may have an innate ability to avoid the effects of host immune surveillance, contributing to its high prevalence and unique oncogenicity. Moreover, a study published 10 years later (after years of further follow-up in WIHS) demonstrated the clinical significance of these results **(Massad, Am J Obstet Gynecol, 2016).** [4] Specifically, the prevalence of HPV16 in HIV(+) women was significantly lower (half that) in HIV(-) cervical precancer. Two recent meta-analyses had similar results.[5,6]

### *C.1.3. Studies of OncHPV DNA CpG Methylation*

Our initial study **(Sun et al, Gynecol Oncol, 2011)[37]** used CVLs from HPV16(+) women in the *general population* with <CIN-2 (n=46), CIN2 (n=12), and CIN3+ (n=27) with CpG methylation quantified using EpiTYPER and pyrosequencing, and showed greater methylation at 14 CpG sites in cases than controls. We subsequently studied (**Mirabello et al, Int J Ca, 2012)[38]** CVLs from 273 additional patients (1) 92 with HPV16 clearance (controls), (2) 72 with HPV16 persistence (no CIN2+) and (3) 109 with CIN2+, using pyrosequencing assays at 66 CpGs across the whole HPV genome. A number of highly significant findings were observed. For example, the ROC curve (AUC – see Fig right) was 0.82 for a CpG site in L1; sensitivity = 91% // specificity = 60% for CIN2+. Of these 17% of CpG sites had high methylation in *pre-diagnostic* CIN2+ specimens (a median of 3 years before diagnosis). Furthermore, high methylation at three specific CpG sites was associated with a very high risk of CIN-2+ when combined compared with low methylation at these sites **(OR=216; 95% CI:21->999) (Mirabello et al, JNCI, 2012)**.[39] We then studied additional HPV types. **Wentzensen et al (JNCI, 2012)[40]** studied 92 women with precancer (N=45 HPV31; N=40 HPV18+; N=23 HPV45+) and a similar number of controls with these HPV. Strong associations between CpG methylation and CIN-2+ were found in L1/ L2. The highest AUC were 0.81 for HPV31, 0.85 for HPV18, 0.98 for HPV45. ***To our knowledge, these are among the strongest associations of HPV methylation with CIN-2+ reported.***




**Mirabello et al (Int J Ca, 2015)** assessed a novel **next-gen sequencing method**, and showed high correlation (ICC=0.61) and similar methylation levels using both pyrosequencing and next-gen methods in all but two HPV16 CpG sites (Fig, left), and similar associations with CIN2+ OR=14.3 and 22.4, respectively.




***Methylation of oncHPV types and its association with cervical precancer in HIV+ WIHS*** – As mentioned, precancer is caused by a greater diversity of oncHPV types in HIV(+) than HIV(-) women. Therefore, we developed Next-Gen methylation assays for all oncHPV types, focusing on CpG sites homologous to those previously associated with precancer. We then conducted a pilot study in HIV(+) precancer cases (N=72) selected from WIHS. Controls were individually matched 1:1 to cases on HPV type, CD4, race. Cases with ≥2 oncHPV (N=22; 23%) had a separate control for each type. Overall, 115 CpG sites in L1, L2, and E2 were assessed for methylation on a 300bp paired-end Illumina® MiSeq platform. Significant case-control differences (P<0.05 unadjusted) in L1/L2 methylation were observed for individual alpha-9 HPVs (HPV16/31/35/58) and alpha-7 HPVs (18/39/45), but not for HPV51/52/59, based on comparison of medians (Mann-Whitney U test) and tertiles (T3 vs. T1 and T2). To study case-control associations across homologous L1/L2 CpGs, we combined their tertile data **(see Table on pg 2)**. For example, the case-control association for homologous alpha-9 CpGs HPV16-L1-5608//HPV31-L1-5524//HPV35-L1-5579 was OR=7.5 (2.9-20; $p<4.3x10^{-5}$), and for the homologous alpha-7 CpGs HPV18-L1-7068//HPV45-L1-7066 the OR=5.5 (1.2-26; p=0.03). Thus, cervical precancer was associated with elevated L1/L2 methylation in HIV+ women and, as hypothesized, several homologous CpG sites in phylogenetically-related HPV types have similar precancer associations.

### *C.1.4. Studies of Cervicovaginal Microbiome*

***Pilot study of cervicovaginal microbiota and HPV in HIV(+) women (Reimers et al, JID, 2016)*** Bacterial vaginosis (BV) is characterized by low abundance of *Lactobacillus* species, high pH, and immune cell infiltration, and has been associated with an increased risk of HPV. In this pilot study, we molecularly assessed the cervicovaginal microbiota over time in HIV(+) and HIV(-) women to more carefully study the HPV-microbiota association, controlling for immune status. 16S ribosomal RNA gene amplicon pyrosequencing and HPV DNA testing were conducted annually in serial CVLs obtained over 8–10 years in 22 HIV(-), 22 HIV(+) with a stable CD4+ > 500 cells/mm3, and 20 were HIV(+) with progressive immunosuppression. Vaginal pH was serially measured. To cluster pyrosequencing 16S rRNA results into corresponding community state types (CSTs), we used hierarchical clustering based on relative abundance (RA) with Euclidean distance and Ward linkage. The number of clusters was determined by the Milligan point-biserial criterion. The Shannon diversity index (SDI) was calculated for the microbiome at each patient visit in the data set. In addition to CSTs, we conducted a priori-planned analyses of the RA of Lactobacilli at the genus and species level, modeled as ordinal data. The relative abundances of *L. crispatus* and other *Lactobacillus* species were inversely associated with vaginal pH (all P<0.001). High (vs low) *L. crispatus* RA was associated with decreased HPV detection (OR=0.48; 0.24-0.96; Ptrend=0.03) after adjustment for multiple covariates, including pH and study group. However, there were no associations between HPV and the RA of *Lactobacillus* species as a group, nor with *L. gasseri, L. iners*, and *L. jensenii* individually**.** *In conclusion, L. crispatus may have a beneficial effect on the burden of HPV in both HIV(+) and HIV(-) women.*

**Selected Additional Cervicovaginal Microbiome Studies** – In a series of studies, our research group reported on the variation in the cervicovaginal microbiome over time (i.e., every 200 days for up to 7 years) and showed that while there was substantial intra-individual variation this was significantly less than the inter-individual variation, with 6 community state types (Smith et al, PLoS One, 2012).[69] In a laboratory based study, filtered supernatants from CVLs with a predominance of *L. crispatus* but not *L. iners* or *Gardnerella vaginalis* were shown to have E.coli inhibitory activity (Gharety, PLoS One, 2014).[70] A study in high risk adolescents focused on three tissues commonly infected by HPV, namely, the anal, oral, and cervical mucosa, and found significant differences in bacterial community composition and diversity between these sites. Overall anal samples were dominated with Prevotella and Bacteriodes, oral with Streptococcus and Prevotella, and cervical with Lactobacillus. Cervical samples had the lowest alpha diversity. Thus, our results showed distinct microbial communities across three body sites; each site susceptible to HPV infection but with different risks of cancer.[71]

## *C.2. METHODS*

**C.2.1. The WIHS Cohort.** In this ongoing cohort, 2,793 HIV(+)and 975 HIV(-) women were first enrolled in between October, 1994, and November, 1995, and a second enrollment to expand the WIHS was conducted in 2001/02. As reported, subjects were enrolled from similar clinical and outreach sources at each of six WIHS study consortium.[125] These included HIV primary care clinics, hospital-based programs, research programs, community outreach sites, women's support groups, drug rehabilitation programs, HIV testing sites, and referrals from previously enrolled participants. The detailed characteristics of the cohort are described in the Preliminary Data (C.1.). *The NIH has determined in previous reviews that the primary importance of the WIHS cohort lies in the study of women-specific issues as they relate to HIV infection, and that these studies should specifically emphasize gynecologic manifestations and outcomes.* **However, support for the relevant scientific studies, including HPV testing, was never part of the core WIHS grant, and must come from standard extramural grant mechanisms**. Also please note that the WIHS has recently added several new clinical sites, which are not included since they have not yet had long follow-up and time to accrue many precancer cases.

**C.2.2. Overview of WIHS Data and Specimen Collection.** The baseline visit involves a 1-1.5 hour structured interview addressing sociodemographic factors, medical and health history, obstetrical / gynecological and contraceptive history, as well as alcohol/drug use and sexual history. Subjects undergo a physical and gynecologic examination, involving collection of laboratory specimens (see below). At each follow-up visit, most activities are repeated except that the questionnaire is shortened to ½ hour. The full gynecologic examination, however, is completed at each visit. Detailed information regarding current medications is obtained, and patients are asked to bring all medications with them for review. Specimen collection includes gynecologic specimens (see below), blood, CD4+/CD8+ flow cytometry, HIV serostatus or HIV plasma RNA level, pregnancy test, and urinalysis.

**C.2.3. Gynecologic Examination.** After the general physical examination, women undergo vaginal speculum examination. First, a cervical swab is obtained for HIV RNA quantification, followed by a cervicovaginal lavage (CVL). **Pap smears** are obtained by using an Ayre spatula to sample the ectocervix, and then rotating a cytological brush in the cervical os. If the cervix is not present a pap is obtained from the vaginal

cuff using a spatula. Cervical swabs are also now being collected but were not obtained at most earlier visits and therefore not useful for this study. Prior data suggest ~10% higher HPV detection with swabs vs CVLs.[72]

**C.2.5 Cytology.** All pap smears are centrally reviewed using the Bethesda System criteria for cytologic diagnosis. All smears are screened by two independent cytotechnologists, with 10% of all negative smears and all abnormal smears read by a cytopathlogist. Cytopathologic review is conducted masked to HIV status, as well as other patient characteristics, but not to the cytotechnologists' interpretations.

**C.2.6 Follow-up of Subjects with Abnormal Cytology: Colposcopy, Histology and Treatment.** By protocol (for research purposes), all subjects with dysplastic cytology, including borderline morphologic changes, termed atypical squamous cells of undetermined significance (ASC-US) and atypical glandular cells (AGC), are referred for colposcopy. In brief, colposcopy is performed as soon as possible following initial detection of an abnormal pap smear. During colposcopy, acetowhite lesions observed under magnification are biopsied and fixed in formaldehyde for histopathologic evaluation. If no significant lesion is found at colposcopy the subject returns to every 6 month pap smear follow-up. If a low grade lesion is detected, treatment is at the discretion of the physician, but colposcopy must be repeated every 6 months until both colposcopy and pap smear findings are normal. If a high grade lesion is detected the patient receives *definitive treatment, and then is followed with colposcopy every 6 months until both colposcopy and pap smear* findings are normal. All clinical, diagnostic and treatment information related to the patient's care are collected using standardized WIHS reporting forms.

***Rigor (of key clinical data)*** – WIHS uses q6mo follow-up with centralized cytologic review, and colposcopy / histology when indicated (with repeat screening as above) making it unlikely that precancer/cancer will go undetected, and q6mo testing helps the precision of time-to-event estimates. CD4+ and HIV RNA tests are conducted in labs in the DAIDS QA Program. Questionnaire data are obtained using validated forms/methods.

***Biologic Variables*** – Only women are included in the current proposal as it focuses on cervical HPV and cervical precancer. Variables such as age, health status (e.g., CD4 count, HIV viral load, gynecologic examination) are incorporated in the data analysis plan, along with other sociodemographic, behavioral, and risk factor information, as described below. Covariates for the major analyses are listed in Analysis of Aim1.

## *C.3. STUDY DESIGN AND STATISTICAL METHODS (by Specific Aim)*

***Rigor (of key statistical methods)*** – we have published several methodologic papers regarding the analysis of HPV and cervical disease data, including methods for prospective cohort studies as used in this study.[68,73-77] Additional statistical methods are referenced when appropriate, and our statistical assumptions are discussed.

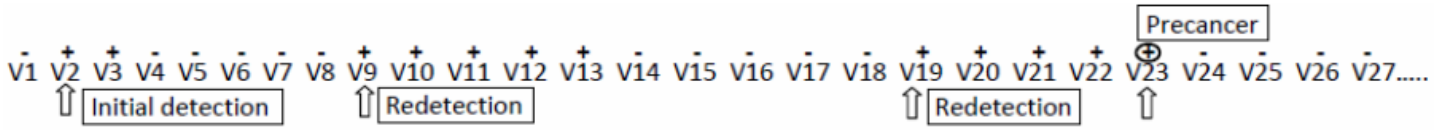
### *Aim 1. To use next-gen "precision" HPV genomic analysis to study HPV natural history (incident detection, reactivation, persistence), and cervical precancer in HIV(+) and HIV (-) women*

**a) Important Definitions**:[68]

- **HPV Prevalence** - the percentage of women with a positive HPV test for a given HPV type among women with adequate HPV test results (i.e., as reflected by human β-globin amplification).
- **Incident detection** - a positive PCR result for a given HPV ***type-specific / isolate*** in a participant who was negative for that isolate in all earlier CVL specimens, even if they were previously positive for HPV of the same type (but of another isolate).
- **HPV clearance** - following prevalent or incident detection of a given HPV type/isolate, clearance will be defined by the stringent criterion of at least two sequential negative results, to minimize concerns that a single false-negative result might affect the findings.
- **Reactivation/redetection** will be defined as detection of a specific HPV (characterized at the type and isolate level) that is then interrupted by at least two sequential visits with negative results before being detected again (e.g., + + - - ++).
- **Cervical Precancer** will be defined as a histologic diagnosis of CIN-3+ and those CIN-2 concurrent with a cytologic diagnosis of HSIL subcategorized as "severe". Limiting precancer to only those CIN-2 with concurrent severe HSIL is done to reduce the risk of misclassification due to over-reading of histology.

### b) Study Design for Aim 1

**Standard HPV DNA testing** has to date been completed for the first 8 years (e.g., V17) in all WIHS women (2793 HIV+, 975 HIV-). Under this proposal we will expand this to 15 years (e.g., V31) in a subgroup 1000 women (see below**). Furthermore, per the Figure (below), Next-Gen **"Precision HPV Genomic Analysis"** (up arrows) will be conducted at the time of (i) *initial HPV detection* (whether prevalent at baseline or detected thereafter), (ii) presumed HPV reactivation / other redetection (e.g., reinfection), and (iii) *precancer diagnosis*.




**Note: only the first sample in any series of samples positive for a given HPV type will be tested at the isolate level.** For example, with initial detection at V2 we would do Precision Genomics at v2 but not at V3.

****As mentioned, a subgroup 1000 women will be selected for extended follow-up HPV testing.** These women will be selected using stratified random sample amongst those 1994/95 enrollees who attended the year 8 visit (V17 or V18) and 2001/02 enrollees who attended their year 8 visit. Further, we will **"oversample" by including *all women with ever HPV16, 18, or 58*** at any visit (based on current data), as well as ***all cases* of precancer** through 15 years of follow-up. Nearly all precancers (i.e., 252 (94%) of 267 cases) occurred at or before 15 years follow-up. **All analyses account for over-sampling** of HPV16,18,58 and precancer in the 1000 woman subgroup (as below). The Table (right) shows the expected number of individuals with reactivation/redetection events after including the extended follow-up of 1000 women (estimated based on current data). With this extended testing the number of 1st reactivations / redetections increases 30% and 2nd and 3rd reactivations nearly 40% (time is needed to allow initial detection then clearance then redetection). Without this added testing there would be insufficient events to complete the planned analyses.

Table Estimated number of women with redetection with proposed extended follow-up

| N | 1st Redetection | | | 2nd or 3rd Redetection | | |
|---|---|---|---|---|---|---|
| | HIV+ | HIV- | Total | HIV+ | HIV- | Total |
| Any HPV | 674 | 44 | 718 | 339 | 22 | 354 |
| Any oncHPV | 298 | 24 | 332 | 109 | 14 | 129 |
| HPV16 | 81 | 17 | 98 | 35 | 5 | 40 |
| HPV18 | 41 | 8 | 49 | 12 | 2 | 14 |
| HPV58 | 54 | 7 | 61 | 20 | 3 | 23 |

***Statistical Analysis of HPV Natural History***

Several of the methods for analysis of HPV natural history and cervical disease reported by our group are now well established in the field, and will be applied in the analyses below,[68,73-77] including methods to study type-specific HPV prevalence, incidence, clearance / persistence, and the development of disease.

**- HPV Prevalence -** will be studied (across visits) using multivariate logistic regression models that incorporate generalized estimating equations (GEE), as previously reported.[16] These models estimate summary odds ratios (ORs) for each HPV type, while accounting for repeated observations of the same women over time and the possibility of multiple concurrent HPV.
**- Incident HPV detection –** on a HPV type/isolate basis will be analyzed using Wei-Lin-Weissfeld (WLW) marginal model approach to account for the possibility of the incident detection of multiple different HPV types/isolates in the same woman.[17] Participants who have missing data for two consecutive visits will be censored at the time of their last visit with complete data.
**- HPV clearance –** will be analyzed using a similar Cox model approach, with clearance defined as two sequential negative results for a given HPV type/isolate.
**- Reactivation/Redetection -** will also use WLW Cox models, and among those with a first reactivation we will also assess time to **second (and possibly third) redetection,** whereas to study associations ***across multiple redetection events*** for a given HPV type we will use established methods for analysis of recurrent events.[78]

For each of the above models we will **account for oversampling** of women with HPV16, 18, 58, and cervical precancer in the stratified random sample of 1000 women (e.g., after year 8); i.e., using time-dependent weights equal to the inverse of the sampling fraction. These weights will be incorporated in the pseudo-likelihood estimating equation using the methods of Chen and Lo under the framework of generalized case-cohort and case-control study[79]. *Statistical significance will be determined based on the two-sided Wald test or the likelihood ratio test*.

***Covariates of interest*** include *HIV status, CD4 count, HIV RNA level, HPV viral load, use of HAART, age, race/ethnicity, smoking, injection drug use, the lifetime number of male sexual partner at baseline, the number of male sexual partner in past 6 months, condom use, and cervical treatment in past 6 months. All variables will be time-updated except the lifetime number of male sexual partners*. Note that: because rates of active STIs (e.g., Chlamydia, GC, etc.) in WIHS are low in both HIV(+) and HIV(-), it is difficult to study STIs other than HPV either as an endpoint or covariate, and douching is uncommon. **These same covariates will be addressed in the models for all Aims (Aims 1, 2, and 3)**

| # of women | >3 yrs | >5 yrs | >10 yrs |
|---|---|---|---|
| Celibate | 1103 | 695 | 285 |

**- Reactivation vs Reinfection -** Each source of redetection in HIV(+) women, whether due to reactivation or reinfection is important, and substantial number of WIHS women reported >3 or >5 years of constant **celibacy** (Table, right ). Celibate is defined as no *vaginal or anal penetration or oral sex with a male or female partner for >6months;* i.e., the period starts the visit after 6

months of celibacy and the period ends at the visit prior to the visit when sex is next reported. Here we make a simplifying assumption that all redetection of the same HPV isolate in celibate women is "**reactivation**". Risk factors for reactivation (in celibate women) will be studied using multivariate Cox models. Descriptive data will include the frequency of reactivation, the average duration of infection following reactivation, etc. Analysis of the relation between reactivation and precancer is described separately below (see Precancer Analysis, below).

In contrast, redetection in noncelibate women includes both reactivation and reinfection. In an exploratory analysis, we will use redetection rates in noncelibate women and their HRs (e.g., associations with HIV/CD4 and convariates), along with reactivation rates in celibate women and their HRs (with HIV/CD4 and covariates), to calculate reinfection rates and their associations with HIV/CD4 and covariates, assuming that the redetection rate is the summation of reactivation and reinfection.

**Statistical Power** for analysis of reactivation/redetection is lower than the power to study HPV prevalent, incident detection, clearance, since only a subset of prevalent and incident HPV reoccurs. To save space, therefore, we present minimal detectable effects only for reactivation / redetection. As shown in the Table (right), using HIV/CD4 as an example exposure (assuming a linear trend across all HIV/CD4 strata), the results suggest good power to study **1st reactivation** for major HPV types, and adequate at ≥2nd reactivation. Analyses stratified rather than adjusting for celibacy as a time-dependent covariate will only have adequate power for studying HPV16, any oncHPV, and Any HPV. The statistical power for HPV **persistence following reactivation** is similar.

| Minimum Detectable HR for Time to Redetection* | | |
|---|---|---|
| HPV type | 1st Redetection | 2nd or Later Redetection |
| 16 | 1.64** | 1.97 |
| 18 | 2.04 | 2.53 |
| 58 | 1.90 | 2.11 |
| Any oncHPV | 1.30 | 1.46 |
| Any HPV | 1.19 | 1.25 |

*Based on linear trend HIV(-), HIV(+) CD4<200, 200-500, >500
**Contrast shown throughout table is HIV(+) CD4>500 vs, <200

***Statistical Analysis of Precancer***

i) Here we define precancer by individual oncHPV type *(after confirming that the same HPV type-specific isolate is found prior to and at precancer diagnosis, more than one oncHPV types may be detected)*. Briefly, precancer positive for a particular oncHPV type will be statistically treated as a "subtype of all precancer" (see (ii) below), and analyzed using methods to examine exposure and disease subtype associations, as we have previously reported (joint Cox models for evaluating multiple disease subtypes).[80] We will study type-specific oncHPV natural history in relation to the risk of precancer, where the natural history of oncHPV is characterized using several time-dependent variables concurrently using for example, the % of all visits positive for this oncHPV type, the total number of reactivations, etc. Because these variables may be correlated with each other, we will use a ridge regression method that our group developed [81] to address these correlations, as well as a lasso method to select the most informative natural history variables.[82] **Covariates** include variables mentioned (above) for HPV natural history, as well as number of live births, prior treatment for CIN, use of oral contraceptives, history of chlamydia. In addition, we will also assess whether or not HIV/CD4 impacts precancer risk *independent of its effects on HPV*, by assessing whether in our optimal models best controlling for oncHPV natural history and other covariates, HIV/CD4 remains a significant factor.

ii) One complexity in studying precancer in HIV(+) women is that more than one oncHPV type may be detected, without knowing which is the causal HPV. Approximately 1/3rd of HIV(+) precancers contain >1 oncHPV. Therefore, we will statistically address this uncertainty. For example, if two oncHPV (A, B) are concurrently detected with pre-cancer we will treat time to type A-precancer and time to type B-precancer as separate events (initial detection of each HPV may occur on different dates even if the date of precancer is fixed). Specifically, using inverse probability weighting methods, let $P_A$ denote the proportion of samples of precancer cases that are positive for HPV type A, and $P_B$ the proportion positive for type B. Then the weight=$P_A/(P_A+P_B)$ will be assigned to time to type A-precancer, and weight=$P_B/(P_A+P_B)$ will be assigned to time to type B-precancer[83-85]These weights are in addition to the sampling weights for oversampling of HPV16, 18 and 58 infections and precancer cases. The final weight of a particular observation is the product of uncertainty weight and oversampling weight. These weights will first be addressed analytically in the Cox model and then empirically by generating 10,000 datasets, and in each dataset assigning each oncHPV type according to its weight (*if only one HPV infection was detected, then the weight=1*) and averaging the results over the 10,000 data sets, providing weighted estimates of HRs and 95% CI. ***Note: similar methods for addressing this uncertainty will be used for analysis of precancer in Aims 2 and 3.***

***Statistical Power:*** Events: N=207 (82%) incident precancer cases (of a total of N=252 precancers) .[86] The Table (right) shows minimum detectable HRs for type-specific

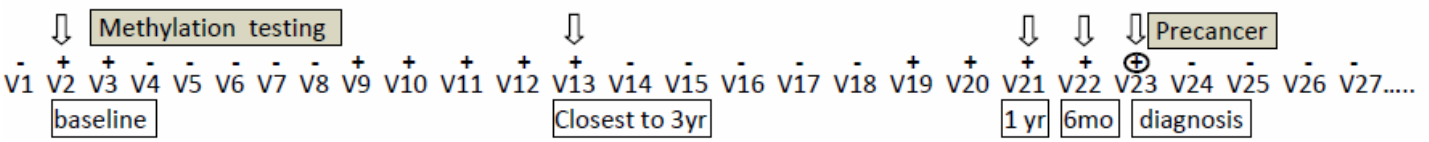
| Minimum detectable HR associated with type-specific pre-cancer development | | | | |
|---|---|---|---|---|
| Natural history Variable | HPV16 pre-cancer | HPV18 pre-cancer | HPV 58 pre-cancer | Any oncHPV pre-cancer |
| Prevalent vs Incident HPV (initial detection) | HR=1.45 | HR=1.84 | HR=1.87 | HR=1.22 |
| Per 10%⇧ of visits positive for HPV type | HR=1.08 | HR=1.10 | HR=1.13 | HR=1.04 |
| # of HPV reactivation/redetection events | HR=1.78 | HR=2.21 | HR=2.29 | HR=1.26 |

oncHPV associations with precancer risk. We assume 80% power and a two-sided type I error of 5%. Power is good for most planned analyses.

*For all Cox models in this proposal, proportional hazards will be examined graphically and by goodness of fit test. Violations of proportionality will be addressed using standard methods*.[87,88]

***Aim 2. To study HPV CpG site methylation and its relation with cervical precancer***

Each incident precancer case (N=207) will be matched to 3 separate controls for each oncHPV present in the case (if there are two oncHPV in the precancer there will be two sets of 3 controls each) based on HPV at the diagnosis of precancer. Additional matching variables are HIV status and above/below 350 CD4, and race. *The case-control differences at precancer diagnosis will be assessed using conditional logistic regression.* As in prior studies, our major analyses will treat methylation as a continuous variable. For each type of HPV infection, there are multiple CpG sites in each of the E2, L2 and L1 regions. Thus, multiple comparisons needs to be addressed. We will use Bonferroni correction to define a stringent type I error for our power calculations. However, Bonferroni is overly conservative and in our main analyses we will use the Benjamini and Hochberg's approach to control the false discovery rate[89]. Further, as in Aim 1 some precancer cases will be positive for more than one oncHPV; uncertainty regarding the causal HPV type which will be statistically addressed using inverse probability weighting methods (as discussed in Aim 1). Among those CpG sites with a significant case-control difference, we then want to assess temporality; i.e., determine when the precancer-related methylation pattern first appeared. Thus, methylation status for the cases will be determined at the diagnosis, 6mo, 1yr, 3yrs prior, and initial detection (baseline) and then compared with each other and with the methylation status for the control(s) assessed at the time of precancer (as shown in figure below).




We will also determine whether the level of methylation at these CpG sites in cases increased overtime. To accomplish this, methylation levels among cases over time will be analyzed simultaneously using a linear mixed effects model for continuous levels of methylation. Lastly, we will analyze case-control associations across homologous CpG sites for phylogenetically related HPV (as in our pilot study); combining data across alpha-9 HPV types (HPV16, 31, 33, 35, 52, 58) and across alpha-7 HPV (HPV 18, 39, 45, 59). For these analyses we will determine tertile values at each CpG and combine the homologous tertile data. For example, in the pilot study the case-control association for homologous alpha-9 CpGs HPV16-L1-5608//HPV31-L1-5524//HPV35-L1-5579 was OR=7.5 (2.9-20; $p<4.3x10^{-5}$).

| Minimal detectable OR for methylation and precancer with (and without) Bonferroni correction (based on 80% power, two-sided test with alpha=0.05). | |
|---|---|
| Proportion in pre-cancer | OR Per SD⇧ in methylation |
| Common ≥15%, (HPV16, 18, 58) | 1.61 (1.42) |
| Less-Common 5-15% (HPV31, 33, 35, 39, 51,52 56,68) | 2.00 (1.71) |
| Uncommon <5% (e.g., HPV59,45) | 3.37 (2.54) |

***Statistical Power:*** The Table (right) shows the minimum detectable case-control OR for methylation at a particular CPG site for a specific oncHPV. Our previous studies have shown ORs greater than those reflected here,[90,91] suggesting that our study will have power to detect a significant associations between methylation and precancer for "common" and "less-common" oncHPV.

***Aim 3. To study the cervicovaginal microbiome and its impact on HPV natural history and precancer risk in HIV(+) women and HIV(-) women.*** This will be one of the first prospective studies of the cervicovaginal microbiome and its relation with HPV natural history and precancer risk in any population.

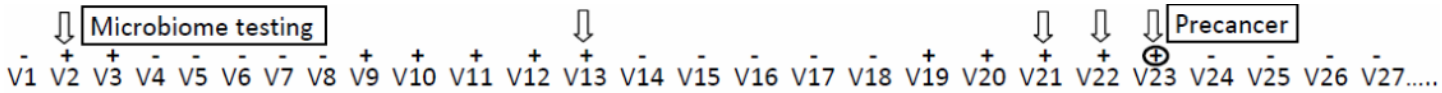
**T*o study HPV natural history we will use standard cohort design***, with the cervicovaginal microbiome characterized at 10 serial visits over a 5 year period in a stratified random sample of N=200 women; selected by HIV / CD4 level category, with oversampling women with prevalent or incident HPV16,18,58, as well as oversampling of *women with at least one oncHPV reactivation (~50 women)*. Sampling weights will be incorporated in all analyses (as discussed in Aim 1).

***To study incident precancer (N=207) we will use nested case-control design*,** with 3 controls per case matched on the oncHPV type(s) in pre-cancer sample (as in Aim 2), HIV status and CD4 above / below 350.

**Analysis of oncHPV incident and prevalent detection, persistence and reactivation/redetection** (Also see Microbiome BioInformatics; Laboratory Methods).- As metioned, at each of the 10 visits the microbiome will be characterized by community state (CST) group, the relative abundance (RA) of each individual bacteria at the genus or species level (e.g., *Lactobacillus or L. crispatus*), and Shannon Diversity Index (SDI). The SDI

is a measure of alpha-diversity (e.g. the within woman/visit diversity of the cervicovaginal microbiome). Vaginal pH level has already been measured at each WIHS visit. Our previous analyses showed 6 CST groups, and found RA for *L. crispatus* was associated with detection of oncHPV. As previously reported,[51] analyses will be similar to the GEE logistic regression models (for oncHPV prevalence) and WLW Cox models (for time to event and persistence / clearance) as reported in Aim 1. Our primary analyses will focus on RA for *Lactobacillus* at the genus and species levels, particularly *L. crispatus.* However, we will also study CSTs, and SDI as exposure variables. For each HPV event (incident detection, clearance, reactivation) we will analyze microbiome data at time t (time of event), and also at time t-0.5 and at time t-1 in separate models, to examine at which time point the relation of the cervicovaginal microbiome with a given HPV endpoint was greatest. In secondary analyses we will also examine temporal associations related to changes in CSTs and HPV detection over time (transition states), using continuous-time multistate Markov models fit using maximum likelihood, as reported.[51,52] Covariates include those mentioned above (see Aim 1).

**Analysis of incident precancer** - Precancer will be studied with microbiome testing at time of diagnosis, 6mo, 1yr, and 3-5yrs before (or the closest visit at which the HPV was detected) as well as at initial detection in both cases and matched controls (Fig, below). The analysis will use conditional logistic regression as in Aim 2, and address the multiple HPV type(s) using weighted analysis as in Aim 1.




***Statistical Power:*** As mentioned, our primary analyses will focus on RA for Lactobacillus at the genus and species levels. Minimal detectable effects (based on 80% power, and two-sided alpha = 0.05) are smaller than effect sizes in prior reports including our microbiome pilot study for most endpoints and moderate for reactivation, suggesting the study has adequate power.

**Analysis of microbiome and HPV methylation** – precancer cases will have both microbiome and methylation data at multiple time points (as above), and precancer controls at the time of case diagnosis. In separate models for cases and controls, the methylation/CPG site with the strongest precancer association for each HPV type (based on results of Aim 2) will be selected for analysis using multivariate linear regression. HPV methylation level will be treated as a continuous outcome. Microbiome will be characteristic first using RA of (i) *L. crispatus* or (ii) *Lactobacillus spp.* at the genus level (continuous), and then (iii) high, medium and low risk CST groups (ordinal). We only consider HPV that are relatively common as in Aim 2 (>15%; including HPV 16, 18, and 58). Overall, with a two-sided type I error rate of 5%, among **cases** our study will have 80% power to detect at least an increase of 0.47SD (moderate effect size) in HPV DNA methylation per SD increase in RA of *L. crispactus,* but good power for *Lactobacillus spp.* (Genus level analysis) Power will also be much greater among **controls** given the larger number of controls than cases.

***Laboratory Testing***

***Rigor (of lab tests)*** – all methods have been previously reported by our group and are referenced. All testing is completed in masked fashion, without knowledge of HIV-status or the presence of cervical disease. QC is discussed below for each assay and in the attached Authentication of Key Biological/Chemical Resources.

*Standard HPV DNA typing* - HPV DNA will be detected using a well-established degenerate primer MY09/MY11/HMBO1 polymerase chain reaction (PCR) assay.[92-94] Primer set PC04/GH20, which amplifies a cellular β-globin DNA fragment, will be used as an internal control to assess the adequacy of amplification. The amplification products will then be probed for individual HPV types using filters hybridized with type-specific biotinylated oligonucleotides for >40 individual HPV DNA types. β-globin–negative specimens will be excluded. Consistent with recommendations from the International Agency for Research on Cancer, HPV types 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 68 are considered oncogenic.[95] (see attached Authentication).

*Next-Gen "Precision HPV Genomic Analyses"*[96,97]**–** Our next-gen assay uses a custom HPV type-specific panel of multiplexed primers that will be sequenced on an Illumina platform. All BAM files are processed through a custom quality control and inhouse analysis pipeline. The HPV sequences are genotyped by the GATK UnifiedGenotyper. SNP calls are made by the GATK UnifiedGenotyper (SNP discovery mode). Low-quality SNP calls and SNP clusters are filtered and masked. Concordance vs non-concordance between serial samples in the same woman over time will be determined by multiple sequence alignment and manually inspecting each woman's individual sequence reads at each variable nucleotide position to determine if there were more than one isolate present. Regions with a high density of variants in close proximity are inspected for the presence of shared SNPs unique to a specific HPV.

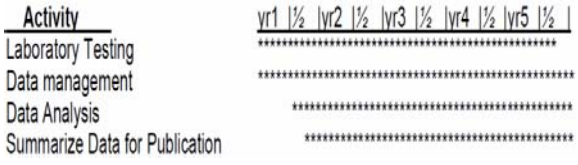
*Methylation Assay* – HPV type-specific barcoded primers will target HPV CpG-rich L1 and/or L2 regions as these are the best segments in the viral genome for disease association with methylation levels.[91,98,99] Bisulfite

treated DNA will be amplified with barcoded primers, amplicons multiplexed with each sample identified by it's unique barcode in downstream bioinformatics analyses. The purified pooled DNA will be combined in libraries prepared using KAPA kits (Kapa Biosystmens, Boston, MA) and submitted for paired-end 250bp Illumina HiSeq2500 sequencing. Illumina reads will be demultiplexed based on the unique Golay barcodes using Novobarcode software package (Novocraft Technologies Sdn Bhd) and prinseq to quality trim reads based on the run PHRED quality scores.[100] In-house Burk lab generated scripts will process the demultiplexed reads to align to HPV reference genomes by using bowtie2 v2.2.3. Methylation status of each CpG site will be determined by Bismark v0.16.3 [101] to assemble site-specific CpG methylation percentages for each assayed cytosine per sample by comparing the ratio of methylated C's to the total number of methylated and unmethylated (C+T) at each CpG site screened. The ratio of "*C/(C + T)*" indicates the proportion of methylated cytosines at each CpG site for the assayed sample. In-house Burk lab generated Python 2.7 (Python Software Foundation) scripts using matplotlib package will cluster samples in the form of a heatmap to determine similar patterns of methylation levels between groups.

*Microbiome assays*: The bacterial microbiome will be characterized using a unique barcoded primer for each sample that amplifies the 16S rRNA V4 region, respectively. PCR products will be pooled and the DNA amplicons will be purified as previously described.[102,103] Barcoded DNA amplicons will be sequenced on an Illumina MiSeq (Illumina Inc., San Diego, CA) by the Epigenomics and Genomics Core Facility at Einstein using 2x300 paired-end reads. Short reads will be deposited in the NCBI Sequence Read Archive (SRA). *Bioinformatics:* To process the Illumina reads, novobarcode (Novocraft Technologies Sdn Bhd) will be used to demultiplex reads, and prinseq to quality trim reads based on the run PHRED quality scores .[100] VSEARCH will be used to perform chimera detection and quality filtering.[104] OTUs will be created using closed reference selection with VSEARCH call by QIIME1.9.[105] using a custom Burk lab database that contains reference sequences from Green-Genes 13.8, HOMD 14, and Vaginal Microbiome Consortium (for 16S only). Representative sequences will be aligned using PyNAST [106] and phylogenic analysis will be performed using FastTree 2.0.[107] The set of OTUs will be available for analyses at different taxonomic levels, from phylum down to species. General community clustering will be performed on the most abundant genera and species using ward.D2 hierarchical clustering based on the samples' Euclidian distance matrix. Visualization will be accomplished with R 3.2.1 and the *ggplot* package. β-diversity will be assessed using weighted and unweighted unifrac distances. Significance will be calculated using PERMANOVA using the adonis function from the *vegan* package. α-diversity will be analyzed based on the Shannon, Simpson, Observed and Chao1 metrics and significance will be determined using the Kruskal-Wallis test.

***C.4. Limitations and Future Plans*** – There are potential limitations to the proposed study that warrant discussion First, although study subjects are followed at 6-month intervals it is possible that some incident HPV infections and cervical neoplasia (our major endpoints) will be missed due to their short natural history. Therefore, detection of HPV/neoplasia at any given visit is likely to be biased towards observance of infections/lesions with long duration; an issue that effects all studies of HPV regardless of HIV-status. As this will affect all groups, any significant differences found in the detection of HPV/neoplasia between risk factor groups will be conservative estimates. An additional complexity is that precancer cases may sometimes test positive for more than one oncHPV. This is addressed in the statistical methods using established methods, but we cannot rule out the possibility that this may still affect some findings, most likely a bias towards the null. Nonetheless, our prior data suggest strong associations despite this possible limitation. Future studies will need to translate the results of this study for clinical applications; for example, by operationalizing assays to measure methylation at the most promising HPV DNA CpG sites and conduct appropriate screening studies among a broader range of HIV(+) women. Similarly, promising cervicovaginal microbiome results will require validation and, if confirmed, translation into appropriate probiotic interventions. Future research is also warranted to study the cervicovaginal mycobiome ("fungal microbiome") and virome ("viral microbiome"). However, the use of these methods to study the cervicovaginal mucosa requires further development before large molecular epidemiology studies can be conducted.

***C.5. Timeline*** - An approximate time table is shown. Since the HPV laboratories have all the necessary reagents, and WIHS specimens will be made available upon request, this testing will begin almost immediately. Therefore, data management will also begin shortly after the grant begins, followed soon thereafter by initial analyses and summation of new findings for publication. The laboratories will complete testing in the final year, but leaving several months to summarize the last of the study data for publication.



**PROTECTION OF HUMAN SUBJECTS**

1) This proposed study will utilize clinical samples and questionnaire data that had been/are being collected in an NIH funded multicenter cohort investigation designed to understand the natural history and pathogenesis of HIV/AIDS and its complications in women, the WIHS. In WIHS, HIV+ and HIV- women have been followed every 6 months since their enrollment in either 1994/95 or 2001/2002.

2) At each visit, questionnaire data, blood, and cervicovaginal cells for HPV DNA assays are collected. The proposed study will analyze these routinely collected questionnaire data, cervical samples, and blood samples.

3) The WIHS study has been approved by NIH IRB, as well as by human subjects review committees at all participating local institutions. All subjects signed an informed consent, which stated that the blood and cervical cell samples would be stored and used in future investigations, before they were recruited into the study.

4) This proposed study will use data and specimens already being collected as part of the WIHS. No risk is involved.

5) All data files and laboratory results are kept confidential by the WIHS.

6) Women and minorities are included in this study. Men and children are not included since this study is focused on cervical HPV and risk factors for cervical dysplasia.

## INCLUSION OF WOMEN AND MINORITIES

Women and minorities are included in this study. Only women are included in the current proposal as it focuses on cervical HPV and cervical precancer. Variables such as age, health status (e.g., CD4 count, HIV viral load, gynecologic examination) are incorporated in the data analysis plan, along with other sociodemographic, behavioral, and risk factor information. The study cohort is majority Black and Hispanic, and race/ethnicity is a particularly important variable in this research.

# PHS INCLUSION ENROLLMENT REPORT

**This report format should NOT be used for collecting data from study participants.**

OMB Number:0925-0001 and 0925-0002
Expiration Date: 10/31/2018

***Study Title:** Next Generation of HPV and Cervical Cancer Research in HIV+ Women

***Delayed Onset Study?** ❑ Yes ☑ No

**If study is not delayed onset, the following selections are required:**

**Enrollment Type** ❑ Planned ☑ Cumulative (Actual)

**Using an Existing Dataset or Resource** ☑ Yes ❑ No

**Enrollment Location** ☑ Domestic ❑ Foreign

**Clinical Trial** ❑ Yes ☑ No

**NIH-Defined Phase III Clinical Trial** ❑ Yes ☑ No

**Comments:** This study will use existing specimens and data from the Women's Interagency HIV Study (WIHS)

| Racial Categories | Ethnic Categories | | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| | Not Hispanic or Latino | | | Hispanic or Latino | | | Unknown/Not Reported Ethnicity | | | |
| | Female | Male | Unknown/ Not Reported | Female | Male | Unknown/ Not Reported | Female | Male | Unknown/ Not Reported | |
| American Indian/Alaska Native | 70 | 0 | 0 | 30 | 0 | 0 | 0 | 0 | 0 | 100 |
| Asian | 148 | 0 | 0 | 12 | 0 | 0 | 0 | 0 | 0 | 160 |
| Native Hawaiian or Other Pacific Islander | 15 | 0 | 0 | 10 | 0 | 0 | 0 | 0 | 0 | 25 |
| Black or African American | 1904 | 0 | 0 | 442 | 0 | 0 | 0 | 0 | 0 | 2346 |
| White | 627 | 0 | 0 | 374 | 0 | 0 | 0 | 0 | 0 | 1001 |
| More than One Race | 85 | 0 | 0 | 46 | 0 | 0 | 0 | 0 | 0 | 131 |
| Unknown or Not Reported | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 5 |
| **Total** | 2849 | 0 | 0 | 914 | 0 | 0 | 0 | 5 | 0 | 3768 |

## INCLUSION OF CHILDREN

Children are not included, since this study focuses on HPV disease progression and cervical neoplasia in women, and is based in a women's HIV cohort.

## DATA SHARING PROTOCOL

All data generated under this grant will be part of a national resource, made available to outside researchers through the WIHS.

## AUTHENTICATION OF KEY BIOLOGICAL AND/OR CHEMICAL RESOURCES

**Procedures for assuring reproducibility and standardization of HPV DNA data.** In each of the two HPV laboratories, every 40th sample will be a replicate control specimen, and 5% of all test specimens will be retested in a masked fashion. Any samples with strong signals (i.e., 3+ - 5+ on gel) by PCR, but not typing by dot blot will have the HPV PCR product isolated and sequenced to identify potential novel types or variants not detected by our probes. New probes will be designed and the blots reprobed. To monitor inter-laboratory agreement, each year we repeat 50 randomly chosen to be tested in both laboratories, and the results assessed with respect to the presence or absence of HPV, and among positive samples for the HPV types present. When more than one type is present we will determine the number and specific types detected in both laboratories. In all comparisons, we require better than 80% concordance, and better than 90% concordance after excluding results that were identified as equivocally positive. If more substantial discordance is observed at any point, HPV testing will be halted until the problem is resolved.

**Next-Gen Precision HPV Genomic Analyses.** Methods for these assays were developed collaboratively and in parallel at the Burk laboratory and the NCI's Cancer Genomics Research Laboratory, Frederick National Laboratory for Cancer. For example, as an initial step the two laboratories designed a custom HPV16 AmpliSeq™ panel that generated 47 overlapping amplicons covering 99% of the genome sequenced on the Ion Torrent Proton platform. After validating with Sanger, the current "gold standard" of sequencing, in 89 specimens with concordance of 99.9%, the NGS method and custom annotation pipeline was used to sequence 796 HPV16-positive exfoliated cervical cell specimens. The median completion rate per sample was 98.0%. The deep coverage, often surpassing 500 ×, enabled the identification of frequent HPV16 variant lineages with a threshold of 1% for the less abundant variant in an HPV16 mixed infection (Cullen et al, Papillomavirus Res. 2015). These methods were then applied to 3200 precancers and invasive cancers. A4 sublineage was associated with an increased risk of adenocarcinoma (OR=9.81, 2.02-47/ P=4.7x10$^{-03}$). D2 were strongly associated with an increased risk of cancer (OR=28.48, 9.27-88.55, P=5.0x10$^{-09}$) and adenocarcinomas (OR=137.34, 37.21-506.88, P=1.5x10$^{-13}$) (Mirabello et al, JNCI, 2016).

**HPV DNA Methylation Assays.** Current next-gen sequencing methods were quality assessed and compared to previously applied well established bisulfate and site-specific pyrosequencing methods. Briefly, cervical DNA was isolated and treated with bisulfite and HPV16 methylation was quantified by (1) amplification with barcoded primers and massively parallel single molecule sequencing and (2) site-specific pyrosequencing. The analysis focused on 18 CpG sites in E6, E2, L2 and L1 ORF regions, using specimens and data from a case-control study of precancer in 99 women. Assays were evaluated for agreement using intraclass correlation coefficients (ICC). Odds ratios (OR) for high methylation vs. low methylation were calculated. Single site pyrosequencing and NGS data were correlated (ICC=0.61) and both indicated hypermethylation was associated with precancer (see fig). There was strong agreement among CpG sites in L1 (median ICC = 0.74), the region with the greatest case-control differences in prior studies. Within the L1 region, the ORs for CIN2-3 were 14.3 and 22.4 using pyrosequencing and NGS, respectively (Mirabello et al, Gyn Onc, 2015)

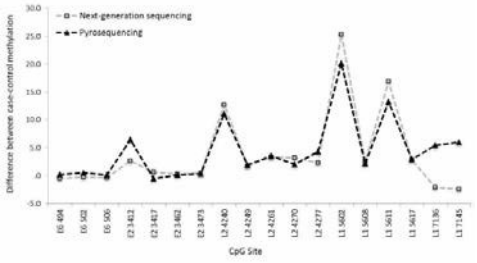



Figure 1.
The difference between the median percent methylation in the cases and controls detected with pyrosequencing (black dashed lines) and NGS (grey dashed lines) methylation assays at each CpG site.

**Cervicovaginal Microbiome Assay.** Adequate DNA quality and amplification is assessed in each assay as follows. Extracted DNA are PCR amplified using primers to an approximately 145 bp region spanning the V6 region of the bacterial 16S ribosomal RNA gene, using "universal" bacterial/archael primers. A unique 8 bp Hamming DNA barcode is introduced to the PCR amplicons from each sample by the forward PCR primers. Successful amplification of the predicted fragment size is confirmed and amplicon concentration estimated by relative band brightness against a control using gel electrophoresis. Additional quality control occurs throughout data bioinformatics. Illumina reads are demultiplexed and the 3' end of the demultiplexed reads trimmed with PrinSeq- lite V0.20.4 to remove bases that had a PHRED quality score below 25. Reads are then quality filtered by using the usearch quality-filtering pipeline in QIIME 1.9: reads are sorted by length, de-replicated to ensure that only unique sequences are analyzed, sorted by abundance, and chimeras are removed. Once created the database is pruned for redundancy in order to ensure that there are no sequences with sequence similarity of 97 percent or greater.

Laboratory chemicals: Chemicals will be purchased from commercial vendors. The vendors provide documents reporting the results of batch analysis.