REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
18-04-2017	Final Report	30-Jun-2014 - 30-Dec-2016

4. TITLE AND SUBTITLE
Final Report: Multi-Omics Approach to Identify Metabolic Biomarkers of Vaginal Microbiome Health and Disease

5a. CONTRACT NUMBER
W911NF-14-1-0311

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHORS
Brenda A. Wilson, Mengfei Ho, Barbara L. McFarlin, Melissa Pires-Alves, Samadh F. Ravangard, Patrick D. Thornton

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of Illinois - Urbana - Champaign
c/o Office of Sponsored Programs
1901 S. First Street, Suite A
Champaign, IL 61820 -7406

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
66096-CH-DRP.1

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
Vaginal infections characterized by microbial dysbiosis lead to inflammatory responses that are believed to be responsible for serious reproductive complications, including spontaneous preterm birth (SPTB). The etiology of SPTB remains unclear and to date the clinical and scientific communities lack an understanding of the vaginal environment with regard to the microbiome, inflammatory and hormonal factors, and the metabolome. The proposed pilot study will significantly advance the field by exploring the use of metabolomics for detection and analysis of microbiomes, and developing the analytical tools needed for identifying metabolomic and microbial

15. SUBJECT TERMS
metabolomics, metagenomics, preterm birth, vaginal microbiome, biomarkers

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Brenda Wilson
UU	UU	UU	UU		19b. TELEPHONE NUMBER 217-444-9631

Report Title

Final Report: Multi-Omics Approach to Identify Metabolic Biomarkers of Vaginal Microbiome Health and Disease

ABSTRACT

Vaginal infections characterized by microbial dysbiosis lead to inflammatory responses that are believed to be responsible for serious reproductive complications, including spontaneous preterm birth (SPTB). The etiology of SPTB remains unclear and to date the clinical and scientific communities lack an understanding of the vaginal environment with regard to the microbiome, inflammatory and hormonal factors, and the metabolome. The proposed pilot study will significantly advance the field by exploring the use of metabolomics for detection and analysis of microbiomes, and developing the analytical tools needed for identifying metabolomic and microbial biomarkers that could be used to discriminate healthy and diseased states indicative or predictive of SPTB. In this pilot study, vaginal specimens will be collected at 16-20 weeks gestation from two cohorts of healthy pregnant women: 6 normal women with no prior history of PTB (low risk of presenting with SPTB) versus 6 women with prior history of PTB (high risk of presenting with SPTB). Using these specimens, next-generation metagenomic sequencing and multiplex-immunologic profiling (Aim 1) and metabolomic profiling (Aim 2) will be used to characterize and compare the vaginal microbiomes and their associated immunologic states and metabolomes to identify early biomarkers (Aim 3) of vaginal health during pregnancy.

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in non peer-reviewed journals:

(c) Presentations

Number of Presentations: 0.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received Paper

TOTAL:

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received Paper

TOTAL:

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received Paper

TOTAL:

Number of Manuscripts:

Books

Received Book

TOTAL:

Received	Book Chapter

TOTAL:

Patents Submitted

Patents Awarded

Awards

Brenda A. Wilson - 2015 YWCA Leadership in STEM Award
Brenda A. Wilson - Scientific Teaching Fellow 2016-2017

Graduate Students

NAME	PERCENT_SUPPORTED

FTE Equivalent:
Total Number:

Names of Post Doctorates

NAME	PERCENT_SUPPORTED

FTE Equivalent:
Total Number:

Names of Faculty Supported

NAME	PERCENT_SUPPORTED

FTE Equivalent:
Total Number:

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED

FTE Equivalent:
Total Number:

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period: 0.00

The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:..... 0.00

Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):..... 0.00

Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense 0.00

The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields: 0.00

Names of Personnel receiving masters degrees

NAME

Total Number:

Names of personnel receiving PHDs

NAME

Total Number:

Names of other research staff

NAME	PERCENT_SUPPORTED

FTE Equivalent:

Total Number:

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

See Attachment.

Technology Transfer

Final Report

RDRL-ROS-I Proposal Number: 66096-CH-DRP, Agreement # W911NF-14-1-0311

Period Covered: July 31, 2015 – July 31, 2016

Title: "Multi-Omics Approach Toward Identification of Metabolic Biomarkers of Vaginal Microbiome Health and Disease"

PI: Brenda A. Wilson, Department of Microbiology, University of Illinois at Urbana-Champaign (UIUC)

Co-Is: Other key personnel included co-investigators Dr. Mengfei Ho (Research Assistant Professor of Microbiology, UIUC) and Dr. Melissa Pires-Alves (Research Scientist, Microbiology, UIUC) who helped performed the sample processing and metagenomic and metabolomic analyses, and collaborators Dr. Barbara L. McFarlin (Associate Professor and Head of the Department of Women, Children and Family Health Science, University of Illinois at Chicago, UIC), Dr. Samadh Ravengard (DO, Fellow, Maternal Fetal Medicine, UIC), and Dr. Patrick D. Thornton (Nurse-Midwife, UIC), who performed the sample collection and delivery outcome and demographic information collection.

Statement of the Problem Studied:

Impact of Proposed Study – In this pilot study we critically evaluated existing GC-MS technologies and metabolomics tools for diagnostic applications. We also evaluated several different methods of analyzing the metabolomic data. We found that metabolite profiling using GC-MS has the potential to discriminate vaginal samples that correlate with adverse pregnancy outcome, spontaneous preterm birth (sPTB). We also performed metagenomic sequence analysis to provide a microbial context of the vaginal heath state. This multi-omic approach could provide additional verification for each of these diagnostic methods.

Initial Aims of Proposed Study – Vaginal specimens would collected at 16-20 weeks gestation age (GA) from two cohorts of healthy pregnant women: 6 normal women with no prior history of PTB (low risk of presenting with sPTB) versus 6 women with prior history of PTB (high risk of presenting with sPTB). Next-generation metagenomic sequencing and multiplex-immunologic profiling (Aim 1) and metabolomic profiling (Aim 2) of this specimens would be used to characterize and compare the vaginal microbiomes and their associated immunologic states and metabolomes to identify early biomarkers (Aim 3) of vaginal health during pregnancy.

Revised Aims and Budget Redirect to Collect Additional Samples – Considering the promising preliminary GC-MS and 16S rRNA gene sequencing results obtained from our studies with the first 12 specimen sets, we requested and were granted a no-cost extension and budget redirect to increase the number of enrolled subjects from 12 to 42 (30 additional subjects). We subsequently received IRB approval for collection of the 30 additional sets of samples from 2 cohorts: 15 pregnant women with low-risk for sPTB and 15 pregnant women with high-risk for sPTB. We proposed to process the samples and perform multiple GC-MS metabolomics as well as 16S rRNA gene and metagenomic sequencing analyses on these samples, similar to that described for the first 12 sample sets. As a consequence of the budget redirect and preliminary findings from the LC-MS analysis, which indicated that the LC-MS approach was not suitable, we dropped both the proposed LC-MS analysis and the proposed immunologic studies. The released funds enabled us to collect and process the additional 30 sets of samples and to perform sequencing and metabolomic analyses on all 42 sets of samples.

Summary of the Most Important Results from Work Performed during Project Period 06/30/2014 – 12/29/2016:

Overview – *Key findings from this pilot project to date (detailed in next section below)*

For considering metabolite profiling:

(1) For metabolic profiling, application of LC-MS analysis for metabolite profiling is premature for predictive risk assessment of sPTB.
(2) For metabolic profiling was able to identify the top metabolites with the strongest correlation to gestational age (GA) at delivery.
(3) GC-MS analysis revealed inconsistencies. We found that it is best to perform repeat analyses of each sample in replicates within short periods of time to identify any inconsistencies.

(4) To minimize inconsistencies in GC-MS analyses, all samples should be in the same batch to reduce variation due to uncontrollable factors, such as instrument and reagent conditions, GC column, etc.
(5) For GC-MS data analysis, only metabolites shared among the majority of samples should be used for subsequent clustering analysis to provide better discriminatory power.
(6) GC-MS analysis can be used to discriminate among vaginal samples with different health states (normal versus sPTB), provided a scaling factor is included in the clustering analysis of samples with wide ranges in compound abundances.
(7) Unusually abundant compounds found only in a limited number of samples must be identified and isolated in PCA analyses.

For considering microbial profiling:

(8) For microbial sequencing analysis, we introduced an alternative BLAST-based binning or clustering method that enabled more rapid and efficient clustering of sequence reads for OTU assignment.
(9) For microbial sequencing analysis, when there is a high percentage of chimeric sequence reads present in a dataset, those sequences must be considered for analysis. We introduced an alternative method for counting chimeras.
(10) The occurrence of chimeric sequence reads is attributed to the composition of the microbial community of the sample, and not simply due to sample processing or PCR bias.

Detailed Summary of Results:

Subject Recruitment, Sample Collection and Processing – We completed the collection and processing of the first 12 sets of vaginal specimens (12 sets of 1 lavage + 2 swabs) at 16-20 weeks GA, as proposed. Demographics data and delivery outcome, whether full term (38-40 weeks) or sPTB (≤37 weeks), were also obtained from all of the 12 subjects. After obtaining program approval for budget redirect and IRB approval for increasing the subject enrollment to 42 total, we also completed the collection and processing of the additional 30 sets of samples (1 lavage + 2 swabs each) from 2 cohorts: 15 pregnant women with low-risk for sPTB and 15 pregnant women with high-risk for sPTB. While we received the final delivery outcome and demographic data for the first 12 subjects, we are still awaiting final delivery outcome and demographic data for some of the second set of 30 subjects.

Metabolic Profiling – We prepared and subjected the initial 12 vaginal lavage samples for metabolic content determination using liquid chromatography-mass spectrometry (LC-MS) and gas chromatography-mass spectrometry (GC-MS). We conducted two preliminary metabolite-profiling analyses on the 12 vaginal lavage samples, comparing the LC-MS and GC-MS technologies. We also conducted the metabolomic experiments on the swab samples to compare results with that from the lavage samples to determine whether data from lavage and swab samples correlated with each other.

For the LC-MS analysis, the assignment of metabolite profiles was difficult due to insufficient standard reference compound profiles in the available databases. While we will retain the LC-MS profile data for the 12 lavage samples and may return to this approach at a later time once more updated and comprehensive compound reference databases are available, we decided not to pursue this approach any further, and instead focused on exploring GC-MS analysis as a means for profiling content metabolites.

For the GC-MS analysis, the assignment of metabolite profiles of the initial 12 vaginal lavage samples was performed similarly to what we previously reported [Yeoman et al. *PLoS One* (2013), PMID 23405259]. From our initial single run of GC-MS analysis, we identified 149 compounds. The relative abundances of each compound across the 12 samples were compared, according to their Euclidean distances to identify which compounds have similar relative abundances. From this process, we identified hexadecanoic acid (C_{16}) and octadecanoic acid (C_{18}) as two abundant compounds that showed a constant ratio in all samples. The abundances of other metabolites were normalized to these two compounds as: $2X_i / (C_{16}+C_{18})$, where X_i, C_{16}, and C_{18} are the readout values from GC-MS analysis for a given compound X_i and the two references, C_{16} and C_{18}. The corrected abundances were further normalized, according to the average value for each metabolite among the 12 samples. The corrected normalized values of metabolite abundance were used for further cluster analysis comparing the 149 compounds with GA (**Figure 1**). We identified a few metabolites that clustered together with GA, indicating that these compounds might be correlated with GA.

The corrected normalized abundances were also used for principle component analysis (PCA) (**Figure 2**) and clustering analyses (**Figure 3**), where the metabolites were ranked according to the Pearson correlation of their normalized, corrected abundances and GA. Preliminary data showed that content analysis using GC-MS was able to identify the top 30, 50, 75 or 100 metabolite abundances with the strongest correlation to GA at delivery (**Figure 2**). These preliminary results indicated that metabolites present in the vaginal lavage samples correlated well with GA, i.e. the abundance signatures of metabolites could discriminate among normal (full-term birth) and diseased (sPTB) samples.

Clustering analysis (**Figure 3**) also clearly separated the sPTB samples (#5, #7, #9 and #12 with GA of 19, 34, 23 and 35 weeks, respectively) from the other 8 non-sPTB samples (GA of 37-40 weeks), even though one of the sPTB samples (#7) was from a subject in the low-risk cohort without prior history of PTB. Sample #5, where outcome was fetal demise, was further distinguished from samples #7, #9 and #12, which tightly clustered together. These preliminary results supported the feasibility that metabolic profiling using GC-MS for metabolite content analysis could be used to provide discriminating signatures indicative of low versus high risk for sPTB. We further determined that there appears to be an optional range of compounds (30-75) that provides maximal discriminatory power.

Critical Examination of Metabolomics Methodology – To further verify our initial findings obtained from the lavage, we repeated the GC-MS analysis three independent times with one performed in duplicate. From these repeat analyses we identified several limitations of using the GC-MS technology that suggest a more comprehensive evaluation of the performance of the GC-MS protocol must be conducted for obtaining reproducible and robust metabolite profiles from samples. One of the major challenges that we encountered was the

Figure 1. Cluster Analysis of Metabolites Identified by GC-MS. The metabolite profiles of the initial 12 samples were first normalized to the two compounds C16 and C18. Shown is a portion of the dendrogram for the cluster analysis of 149 metabolites using Euclidean Distance Matrix calculated from their normalized abundances. Enlargements show metabolites clustered near GA (red) and two reference metabolites (green) that were the most abundant among those present in all samples and at a constant ratio.

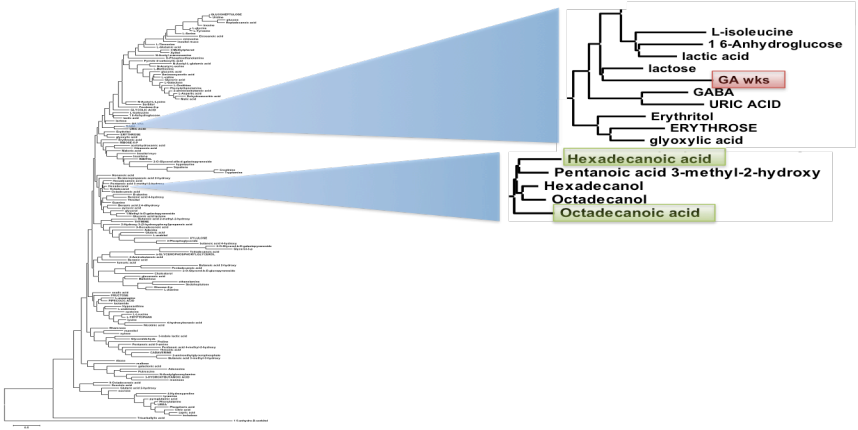


Figure 2. Principal Component Analysis of Top Metabolites Correlated with Gestational Age. PCA plot using metabolite abundance with the strongest correlation to GA (weeks). Samples are indicated as black numbers. Blue circles indicate samples with sPTB delivery outcome.

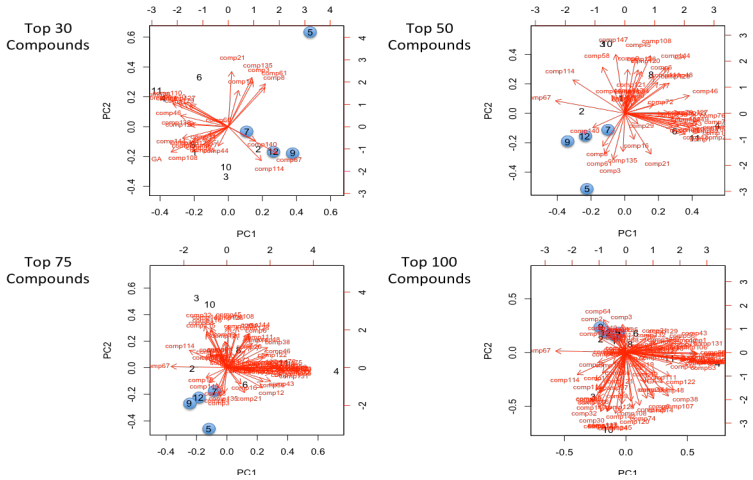
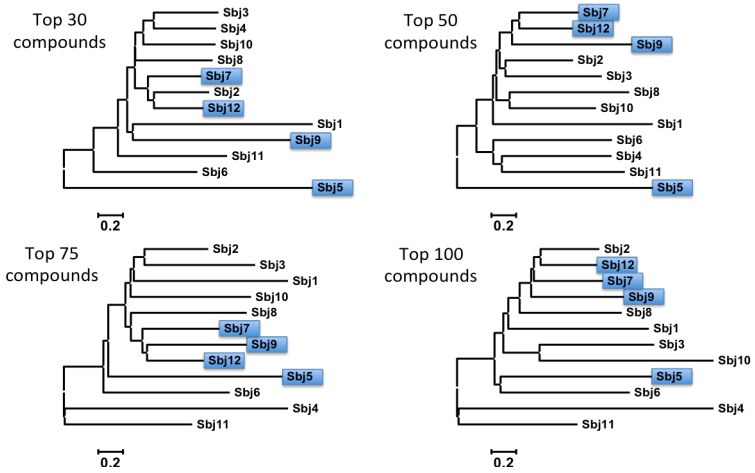
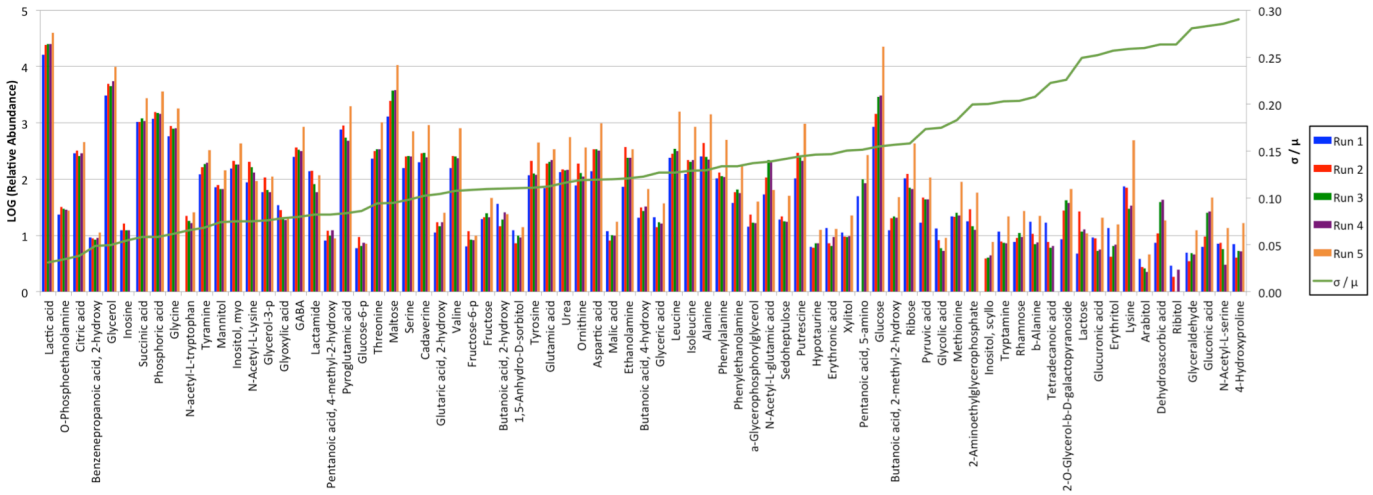


Figure 3. Cluster Analysis of Top Metabolites Correlated with Gestational Age. Shown are cluster analyses of 12 vaginal lavage samples using 30, 50, 75 or 100 top metabolites with corrected abundances most correlated with GA. Blue denotes samples from subjects (Sbj) with sPTB delivery outcome.



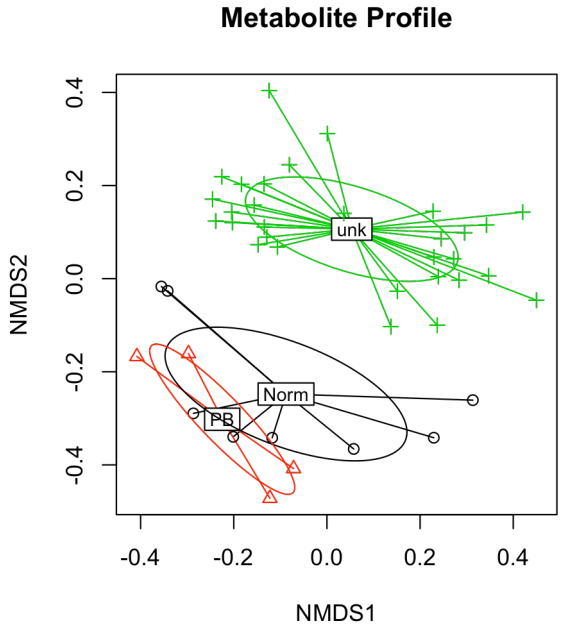
variation in metabolite output between experiments and between GC-MS runs, even on the same sample. The sums of the relative abundances for each compound from the initial 12 samples were compared among five analyses, two performed at the same time. Although 149 compounds were detected during the initial run, there were only 115 compounds that could be detected in at least 4 out of all 5 times. As shown in **Figure 4**, comparison was possible only when the abundances were compared at less stringent levels using their log values. The variability among the analyses was defined as the ratio of the standard deviation over the mean (σ/μ). Of the 115 compounds, 77 compounds had a σ/μ values of <0.3, 51 <0.2, and 24 <0.1. We concluded that it was necessary to perform multiple GC-MS runs on each sample within short periods of time, where the instrument and reagent conditions were most likely to be stable and consistent. Variations in GC-MS data output have also been report before [Roessner U, Wagner C, Kopka J, Trethewey RN, Willmitzer L "Technical advance: simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry" *The Plant Journal* (2000) 23(1):131-42, PMID: 10929108].

Figure 4. Reproducibility of Metabolite Detection using GC-MS. Shown is a plot of the relative metabolite abundance (log-scale, left y-axis) profiles from five independent rounds of GC-MS analysis. Metabolites are ordered along the x-axis based on the standard deviation (σ) over the mean (μ) on the right y-axis.



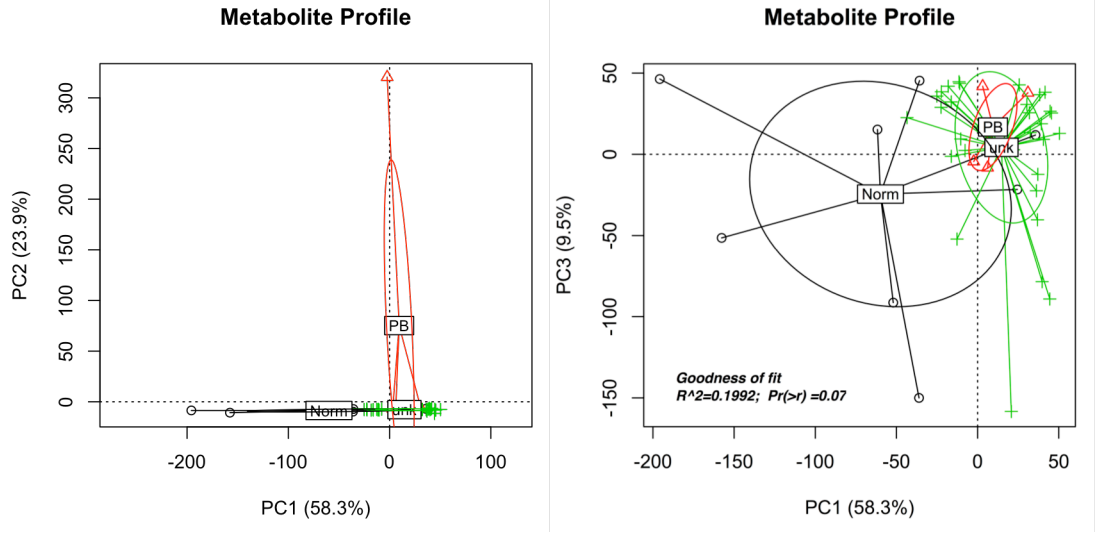
The limited reproducibility of the GC-MS analysis could be attributed to a number of experimental factors, including instrument performance, GC columns, reagents, sample preparation and derivatization conditions. These uncontrollable factors make GC-MS difficult to use as an affordable and rapid detection method to achieve reliable and reproducible metabolite profiling. This conclusion was further supported, when we performed GC-MS analyses in triplicate on the additional 30 lavage samples. In this case, the triplicates were performed at the same time. GC-MS derived metabolite profiles of the previous 12 samples were combined with that of the new 30 samples for non-metric multidimensional scaling (NMDS) analysis (**Figure 5**). Out of 226 compounds detected in the 12-sample group and the 30-sample group, only 124 compounds were detected in both groups. Considering the groupings, two with the known delivery outcomes from the first 12 samples (normal versus sPTB) and all of the 30 new samples (denoted as unknown, unk), the NMDS analysis showed weak separation between normal (Norm) and sPTB (PB), but the unknown (unk) group clearly clustered separately from both of the other groups. These results suggest that different batches of sample analysis using GC-MS method, and possibly collection

Figure 5. Non-metric Multidimensional Scaling (NMDS) Analysis. An NMDS ordination plot of all 42 vaginal lavage samples was generated based on Bray-Curtis distances using the metaMDS() function in the Vegan package.



and processing of clinical samples, contribute to the separation of clusters. NMDS analysis was based on the distance matrix generated by Bray-Curtis method on the relative abundance of metabolites without normalization or correction for the wide range of observed abundance that could be as much as 5 orders of magnitude across the compounds within a given sample, as well as the varying dilution factors that may occur among different sample collections.

Figure 6. PCA analysis of selected 124 compounds. PCA plots of metabolite profiles for all 42 lavage samples: 8 normal (Norm), 4 sPTB (PB), and 30 unknown (unk). Analysis was performed using function *rda()* of Vegan package in *R*. The variance associated with PC2 (24.9%) in the PC1 versus PC2 plot (left) was attributed primarily to one sample due to unusual enrichment of one compound (glycerol). All other samples were not distinguishable by this component. Better separation could be seen in the PC1 versus PC3 plot (right). Goodness of fit for the three groupings (p=0.07) was determined using function *envfit()* of Vegan package.

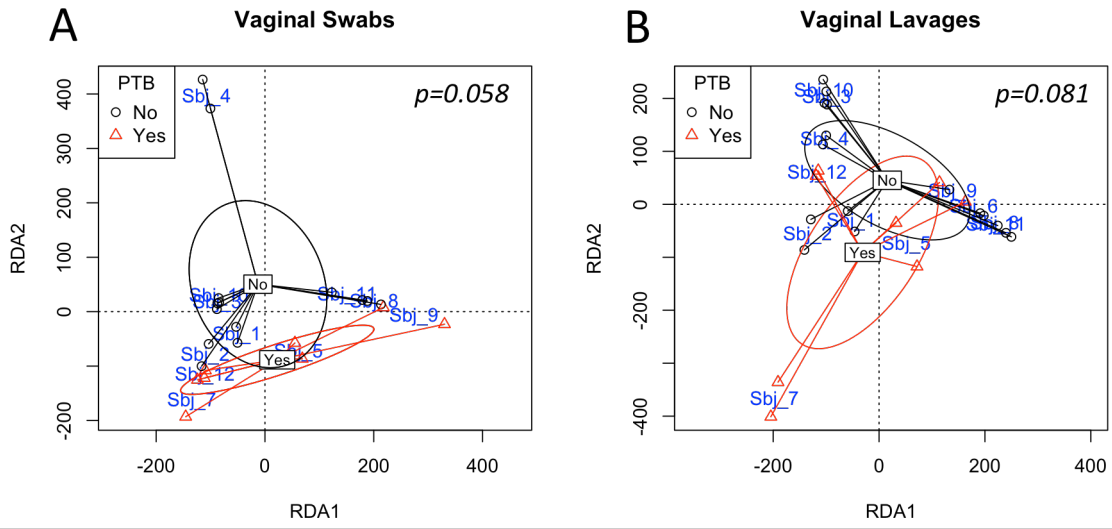


We also analyzed the same selected set of 124 compounds using PCA analysis, which has the advantage that it is less sensitive to the magnitude of the variables. As seen in **Figure 6**, we found that PC2 contributes to 24.9% of the total variance of the dataset, and this was due to unusual enrichment of a single compound (glycerol) in one sample. A better spread of the data was seen in the PC1 versus PC3 plot. Although we do not have the delivery outcomes for the 30 additional samples yet, the discrimination of the groups show less influenced by batch-to-batch variation. We conclude that normalization, scaling and magnitude are important to consider in the analyses. We will further analyze this metabolite profiling data once we receive the delivery outcome for remaining samples, and we will also include a few corrections for the varying magnitudes of abundances among the compounds.

16S rRNA Gene-based Microbial Profiling for the 12 samples – We prepared genomic DNA from all 12 vaginal lavage samples and 11 of the 12 swab samples (1 swab sample did not yield sufficient genomic DNA for inclusion in the analysis) for DNA sequencing to determine vaginal microbial content. We used Illumina MiSeq 250nt paired-end sequencing with the Fluidigm platform of the V3V4 and V4V5 regions of the 16S rRNA gene. This sequencing run produced over 17M paired-end reads of excellent data. Application of PCA based on 16S rRNA gene sequences from swabs or lavage samples showed discriminatory clustering of the full-term versus sPTB samples (**Figure 7**). OTUs were assigned according to a reference 16S rRNA collection from NCBI at 97% identities binning. Chimera or unassigned reads were not included in the analysis.

Critical Examination of In-depth Short-read DNA Sequencing Analysis – While compelling, the pilot data identified several potential problems of using the current high-throughput sequencing technologies, in particular, data analysis of sequencing output, which must be addressed constructively in future research. The first issue that we encountered was the massive computational complexity of the sequence read processing and taxonomical assignment and clustering of the sequence reads for comparative phylogenetic analysis. Current pipelines for microbial content profiling analysis include clustering and assignment of the 16S rRNA gene sequence reads into operational taxonomic unit (OTUs) based on sequence similarities to reference sequences of previously characterized and annotated microbes in the available sequence databases. Clustering into OTUs usually involves pre-alignment or pairwise alignment of the entire collection of reads, and consequently this method is computationally demanding, particularly for large dataset sequence read outputs.

Figure 7. PCA plots of 16S rRNA gene sequencing from 12 subjects. PCA plot for 16S rRNA gene sequencing profiles of 11 swabs (panel A) and 12 lavages (panel B) was calculated using function *rda()* of Vegan package in *R*. Goodness of fit for PTB status (p=0.058 and p=0.081) was determined using function *envfit()* of Vegan package.



The NAST method partially solves this problem by performing alignment of each read to a set of reference sequences, such as Greengenes' Core Set sequences. To address this issue, we adopted an alternative BLAST-based binning or clustering method, whereby randomly selected reads served as seed sequences to recruit similar sequences (at thresholds of 99%, 98%, or 97% identity) into the cluster (or bin). This improvement will save the computational time by a factor equal to the ratio between the size of the core set and the actual number of clusters in the dataset of an experiment. To test the feasibility of our clustering approach, we generated a mock community comprised of a defined mixture of 80 bacterial strains and subjected the sample to Illumina MiSeq 250nt paired-end sequencing with the Fluidigm platform of the V3V4 and V4V5 regions of the 16S rRNA gene. As shown in **Table 1**, using our method allows for efficient clustering of sequencing reads without pre-alignment or pairwise alignment of the entire collection of reads. This approach allowed binning, and hence OTU assignment of greater than 98% of the sequence reads at a threshold of 97% identity.

During the analysis of our mock community data, we found that that the majority of the OTUs could be attributed to chimera formation of the expected sequences for the community (**Table 2**). Among the reads matched to the reference rRNA database, 99.6% were from the expected strains in the community. In addition, there were a few sequences resulting from possible contamination during processing. However, 24% and 35% of the reads from V3V4 and V4V5, respectively, were from chimeras. The current strategy for handling chimeras is to identify them and remove them from the final analysis. When there is a high percentage of chimera reads present in a dataset, such as that from our mock community, it may be problematic to simply ignore the chimera reads during sequencing data analysis. We used a UNIX shell script to assign those OTUs suspected to be chimeras as composite sequences of the known identified reference sequences based on the BLAST output. Our script could assign up to 4 possible components for each chimera sequence. A fractional count could be assigned for each component in the chimera read. For a chimera sequence composed of two sequences, each component would be counted as half the count of this chimera. Similarly, a chimera of three or four components would receive one-third or one-forth of the count for each component. For any segment that could not be assigned, a corresponding fraction would be counted as unknown.

The mock community was generated from equal aliquots of bacterial culture. However, the microbial profiles of the resulting sequence datasets did not reflect the expected composition of bacteria. Inclusion of chimera counts allowed for increased counting of *Klebsiella*, *Morganella*, *Enterococcus* and *Pseudomonas* by 2- to 5-fold, and most strikingly *E. coli* by up to 100-fold. Thus, by using our method for counting sequence reads that include the information hidden in the chimeras, we can account for greater than 98% and 99% of the total sequencing reads for V3V4 and V4V5, respectively.

Table 1. Clustering profiles of V3V4 versus V4V5 reads at various identity thresholds using our BLAST-based clustering approach

	99 % Identities			98 % Identities			97 % Identities		
	#Reads	%	#clusters	#Reads	%	#clusters	#Reads	%	#clusters
Total V3V4 reads:	159967			159967			159967		
Reads in clusters >1000:	99617	62.3%	15	134289	83.9%	20	146030	91.3%	17
Reads in clusters >100:	122914	76.8%	91	146517	91.6%	66	153320	95.8%	46
Reads in clusters >10:	134474	84.1%	468	151962	95.0%	238	156424	97.8%	140
Reads in clusters <=5:	23691	14.8%	19218	7259	4.5%	6179	3185	2.0%	2814
Reads in clusters <=4:	23021	14.4%	19084	7029	4.4%	6133	3130	2.0%	2803
Reads in clusters <=3:	22209	13.9%	18881	6765	4.2%	6067	3010	1.9%	2773
Reads in clusters <=2:	20454	12.8%	18296	6300	3.9%	5912	2842	1.8%	2717
Reads in clusters =1:	16138	10.1%	16138	5524	3.5%	5524	2592	1.6%	2592
Total V4V5 reads:	151041			151041			151041		
Reads in clusters >1000:	111960	74.1%	14	129782	85.9%	13	137897	91.3%	12
Reads in clusters >100:	129469	85.7%	74	142923	94.6%	52	150004	97.9%	45
Reads in clusters >10:	139561	92.4%	377	148100	98.1%	204	150004	99.3%	114
Reads in clusters <=5:	10094	6.7%	8070	2456	1.6%	1860	811	0.5%	582
Reads in clusters <=4:	9744	6.5%	8000	2306	1.5%	1830	751	0.5%	570
Reads in clusters <=3:	9324	6.2%	7895	2142	1.4%	1789	687	0.5%	554
Reads in clusters <=2:	8556	5.7%	7639	1899	1.3%	1708	585	0.4%	520
Reads in clusters =1:	6722	4.5%	6722	1517	1.0%	1517	455	0.3%	455

Table 2. Microbial profile of a mock community*

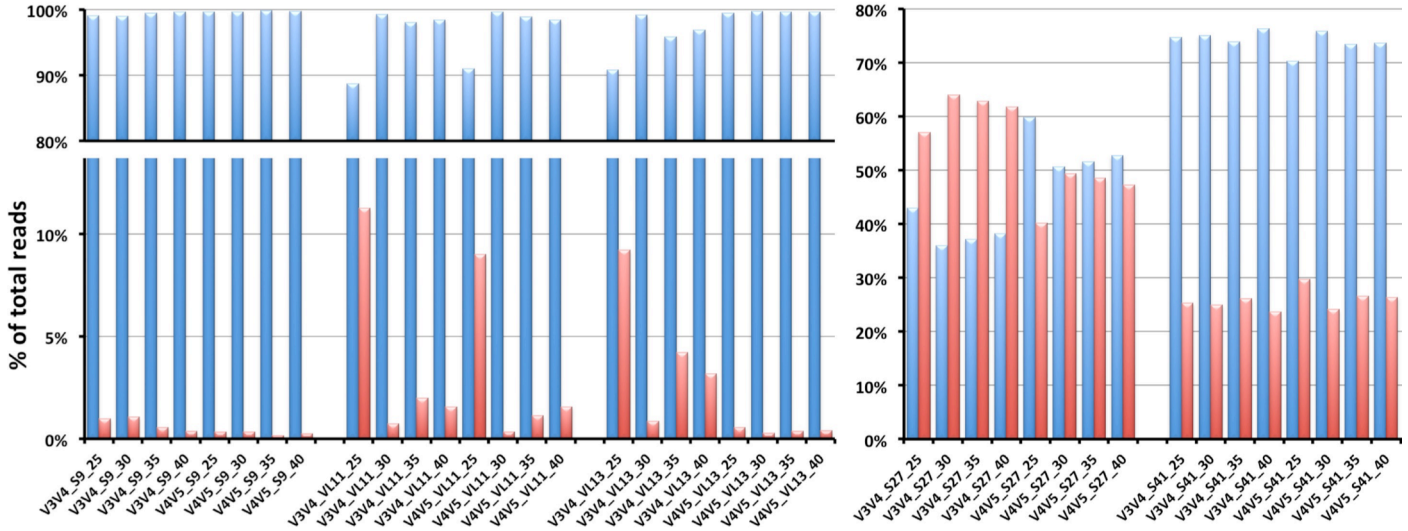
Name	Number of bacteria strains	Expected %	V3V4 w/o chimera %	V3V4 w/ chimera %		V4V5 w/o chimera %	V4V5 w/ chimera %	
Proteus mirabilis	50	63%	45%	39%		58%	48%	
Escherichia coli	16	20%	0.1%	5%		0.1%	9%	
Providencia vermicola	4	5%	10%	10%		8%	8%	
Pseudomonas aeruginosa	4	5%	2%	4%		1%	2%	
Morganella morganii	2	3%	8%	10%		2%	5%	
Enterococcus hirae	2	3%	1.4%	2%		0.3%	2%	
Staphylococcus cohnii	1	1%	33%	29%		30%	25%	
Klebsiella pneumoniae	1	1%	0.5%	1%		0.2%	1%	
Lactobacillus crispatus	none		0.03%	0.1%		0.1%	0.04%	
Lactobacillus iners	none		0.01%	0.05%		0.1%	0.1%	
Bacillus cereus	none			1%		0.1%	0.2%	
Stenotrophomonas pavanii	none			0.01%		0.01%	0.02%	
Number of total reads and Sum of total assigned count			121259 (76%)	156722 [a] (98%)	159967 [b]	98636 (65%)	148814 [a] (99%)	151041 [b]

[a] The total count of read assignments including matched and chimeras. [b] Total reads used for analysis.

Effect of PCR conditions on microbial profiles of selected communities – It has been suggested that PCR conditions can affect the extent of chimera formation during amplification. Contrary to intuitive reasoning that fewer cycles of PCR should minimize the extent of chimera formation, our results show that this is not the case, the makeup of the microbial content appears to be responsible for the resulting chimera formation. As shown in **Figure 7**, sequencing data for communities VL11 and VL13 showed that 25 cycles of PCR amplification actually results in more chimeras than when more PCR cycles are performed. And, for

communities S27 and S41, the same percentage of chimeras was found under all PCR conditions used. Although the choice of primer pairs for community S27 affected the amount of chimeras by about 10%, no consistent trend was observed for other communities.

Figure 7. PCR conditions do not consistently alter chimera occurrence in sequencing results. Shown are the in-depth sequencing results from Illumina sequencing run using the Fluidigm platform for 16S rRNA gene libraries from five selected microbial communities, amplified using 27F and 1492R universal primer pairs under different PCR conditions, number of cycles (25-40) and/or length of extension time (60-90 sec). Blue bars denote the % of reads matched to known reference sequences, and red bars denote the % of chimeras in a sample. V3V4 and V4V5 indicate the two different primer pairs used for PCR amplification.



Metagenomic Sequencing Analysis – The 11 swab samples were also subjected to Illumina HiSeq2500 160nt paired-end metagenomic sequencing using the Shotgun TruSeq platform for genomic DNA library preparation. This sequencing run produced 2.4B reads (384 Gbases). As soon as we have completed the 16S rRNA gene analysis of all samples from all 42 subjects and have worked out some technical issues identified with read clustering analysis (described above), we will apply these strategies toward analysis of the metagenomic data and will then integrate the resulting information into our multi-omic profiles.