**Sample Collection**

Info about IDing patients and metadata survey. Sampling procedure (I should be able to write it up and double check with Dr. Moore, etc.).

Two swabs were collected from each site (bilaterally from the vaginal and oral sites, sequentially for the rectal swabs) due to the incompatibility of performing genomic and metabolomics analysis on the same sample – each method requires use of the entire swab. The swabs from the patient’s left side, and second rectal swab, were selected for genomic analysis. Swabs from the right side, and first rectal swab, were used for metabolomic analysis.

**Laboratory Methods**

*DNA-based methods*

DNA was extracted from swabs using the MoBio PowerSoil DNA Isolation Kit (now Qiagen DNeasy PowerSoil Kit), following manufacturer’s protocols with the addition of a 10-minute incubation at 60 °C prior to the initial vortexing. Applied Biosystems’ SYBR Green PCR Master Mix and *Escherichia coli* standards of known 16S gene copy number were used to quantify each extract using qPCR with the 515F (GTGCCAGCMGCCGCGGTAA) and 806R (GGACTACHVGGGTWTCTAAT) Caporaso primers (Caporaso et al. 2011). Extracts were grouped together based on 16S rRNA gene copy number and diluted to the same copy number within each group. Following dilution, extracts were PCR amplified in triplicate with a unique, error-correcting Golay barcode for each sample and primers 515F and 806R (Caporaso et al. 2011). Purified PCR products were sequenced on an Illumina MiSeq with 2x250 kits.

*Metabolomics-based methods*

**Bioinformatic Processing**

The resulting sequencing reads were filtered to remove reads with ambiguous bases and quality scores below 30 with AdapterRemoval2. Following quality filtering, sequences were demultiplexed in Qiime using the unique Golay barcode for each sample. Sequences occurring less than 5 times were removed, as they are unlikely to be true biological sequences, and the remaining sequences were used to cluster sequences at 97% into *de novo* Operational Taxonomic Units (OTUs). The demultiplexed sequences were used to pick OTUs with USEARCH, which assigned each sequence to a *de novo* OTU.

Because samples were sequenced across a number of MiSeq sequencing runs (rarefaction 9000).

**Statistical Analysis**

For the BJOG abstract the stats come from looking at the samples where we had enough sequence depth to analyze, ended up being 95 individuals. I called any sample that had greater than 50% abundance of a specific taxa as being dominated by that taxa (basically just Lactobacillus and Escherichia). For further reference, Parse 2 is L. iners, Parse 26 is L. jensenii, Parse 6 is L. crispastus, and Parse 4 is L. gasesri. For vaginal Eschericha dominance, I compared PFI < 6 vs PFI. 24 months using Fisher’s exact test (5 of 18 dominant PFI < 6, 0 of 28 PFI > 24 months) resulting in a p-value of 0.0063. There wasn’t any difference in Lactobacillus dominance by group (haven’t done stats test yet but it’s clear visually that there isn’t a difference). But compared all samples to see if our study had lower than expected lactobacillus dominated communities (not including benign) – compared to Brotman 2014 which had the most comparable methods (age, 16S amplicon). Our samples had significantly lower than expected lactobacillus dominated communites (25 out of 88, vs 15 out of 28, Fisher’s exact p-value = 0.0218) Also need to double check that vaginal communities don’t change over time with chemotherapy.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, and Knight R. 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proceedings of the national academy of sciences 108(Supplement 1):4516-4522.