Markov-Based Model of Cervicovaginal Bacterial Dynamics Predicts Community Equilibrium States in Young South African Women

1 Abstract

Commensal cervicovaginal bacteria modulate female genital tract (FGT) inflammation and HIV acquisition risk. High diversity communities with low *Lactobacillus* abundance are associated with more activated cervical HIV target cells and an increased risk of HIV acquisition. While *Lactobacillus* colonization of the genital tract is believed to be beneficial for reproductive health, the dynamics of cervicovaginal bacteria are not well understood. In this thesis, I construct a Markov-based model of bacterial community transitions in a South African cohort of young healthy women living in a region with high HIV incidence and prevalence. I found that while *Lactobacillus crispatus* colonization was relatively stable, *Lactobacillus iners* colonization was unstable with high frequency transitions to *Prevotella*-rich, high diversity communities. Because of this, the cohort experienced numerous transitions toward more diverse communities that are associated with increased rates of HIV acquisition. Based upon the model, I predict that blocking the FGT colonization transition from *L. crispatus* to *L. iners* will most effectively increase the amount of *L. crispatus* dominant, low-inflammation, bacterial communities that are associated with lower risk of HIV acquisition. These observations may be used to develop more effective methods to prevent HIV acquisition in young women living in sub-Saharan Africa.

2 Introduction

Of the 1.8 million new HIV infections annually, 1.2 million occur in sub-Saharan Africa [25], illustrating the need for the development of additional safe and efficacious methods to reduce HIV risk in this region. Globally, 90% of new HIV acquisitions occur via heterosexual intercourse, and women are twice as likely to acquire HIV as compared to men after sex with an infected partner. Therefore, mucosal tissues in the female genital tract (FGT) are the initial site of HIV primary infection in most new transmission cases. It has become increasingly clear that the mucosal environment of the FGT can have a significant impact on HIV acquisition risk [19]. For example, inflammation in the FGT, as measured by the presence of increased pro-inflammatory cytokines[i] in genital secretions, is associated with a three-fold increased rate of HIV acquisition [29] [31] [33]. Therefore, understanding the factors which modulate genital inflammation is important for reducing HIV risk.

There are 10^8 bacteria per mL of vaginal secretion. The structure and stability of these bacterial communities vary dramatically between disease states, populations, women, and over major life events such as menopause and pregnancy [13] [54] [24] [55] [23]. These cervicovaginal bacterial communities can modulate markers of genital health, including FGT inflammation.

[i] A cytokine is a molecule released by immune cells with the purpose of sending messages to other cells. Pro-inflammatory cytokines cause some recipient cells to induce the inflammatory response. The inflammatory response can be destructive to mucosal tissue, removing an important barrier between the host and the commensal and pathogenic microbes inhabiting the cervicovaginal compartment. Additionally, the inflammatory response recruits other immune cells to the tissue, such as CD4+ T cells, which are targets for HIV virus infection. Thus, the inflammatory response at the genital tissue is expected to increase the probability of HIV acquisition upon genital exposure to the virus by bringing the cell that HIV infects directly in contact with the genital epithelium (the location where HIV often first enters the body).

For example, high-diversity bacterial communities, especially *Prevotella*-rich communities, are highly correlated with genital pro-inflammatory cytokine concentrations [1]. In addition, these *Prevotella*-rich high diversity communities are associated with more activated cervical CD4+ T cells, which are thought to be the initial targets of HIV infection ("HIV target cells"), and are associated with a 4-fold increased risk of HIV acquisition [20].

Previous studies in the South African province Kwazulu-Natal have demonstrated that cervicovaginal bacterial communities distinctly cluster into four "cervicotypes" after sequencing of the variable region 4 (V4) of bacterial ribosomal (r)RNA to determine relative bacterial community abundances[ii] [20]. These CTs differ in the relative abundances of bacterial taxa they contain. For example, some samples are represented by a single driver species while others are very diverse with multiple highly-represented bacterial species.

Definition 1 *Define the dominant taxa to be the most abundant bacterial taxa in a sample.*

From here on, I use the Shannon index (referred to as Shannon alpha diversity) as a metric for the magnitude of this dominance:

$$H = -\sum_{i=1}^{R} p_i \ln p_i \tag{1}$$

where H is the Shannon alpha diversity, R is the total number of observed taxa, and p_i is the proportion of reads that mapped to taxa i. Note that

[ii]The variable region 4 (V4) of the 16S rRNA gene of bacterial DNA varies between bacterial species. Thus sequencing of this region and aligning reads to species-labeled reference sequences allow determination of bacterial species with high precision without sequencing the full genome. Sequencing of the whole bacterial genome and then aligning reads to a labeled full-genome reference yields similar results to sequencing only the V4 region in the FGT compartment [1], illustrating that sequencing V4 is an accurate metric for determining bacterial composition in the FGT at a species level.

the metric for diversity calculates the entropy over the abundances. Using log rules, we see:

$$H = -\sum_{i=1}^{R} \ln p_i^{p_i} \tag{2}$$

$$H = -\ln \left(\prod_{i=1}^{R} p_i^{p_i} \right) \tag{3}$$

If the distribution of bacterial abundances is discrete uniform over the support of taxa (i.e. all taxa have an abundance of $\frac{1}{R}$):

$$H_{\text{unif}} = -\ln \left(\prod_{i=1}^{R} \left(\frac{1}{R} \right)^{\left(\frac{1}{R}\right)} \right) = -\ln \frac{1}{R} = \ln R \tag{4}$$

And now if the bacterial abundance distribution is completely dominated (i.e. all reads mapped to a single taxa):

$$H_{\text{dominated}} = -\ln \left(\prod_{i=1}^{1} p_i^{p_i} \right) = -\ln \left(p_1^{p_1} \right) = -\ln(1) = 0 \tag{5}$$

By equation 4, we see that a discrete uniform distribution yields a Shannon alpha diversity of $\ln R$, while a completely dominated bacterial abundance distribution yields a Shannon alpha diversity of 0. It is easy to see that these are the bounds of Shannon alpha diversity and that the metric will scale based on the number of bacteria sequenced between these two bounds as with entropy.

The clustered cervictotypes in the FGT separate via Shannon alpha diversity. Cervicotype 1 (CT1) has the lowest Shannon alpha diversity and is dominated by *Lactobacillus crispatus*, while cervicotype 2 (CT2) has low Shannon alpha diversity and is dominated by *Lactobacillus iners*. These low-diversity, *Lactobacillus*-dominant CTs are believed to be beneficial for vaginal health. While 90% of white women in developed countries belong to these *Lactobacillus*-dominated CTs [39], the minority of black, sub-Saharan African women belong to CT1 or CT2. Instead, sub-Saharan African women

tend to have more diverse communities, including the *Gardnerella vaginalis* dominated cervicotype 3 (CT3) and the highly diverse and generally *Prevotella*-rich cervicotype 4 (CT4).

Because the *Prevotella*-rich CT4 is associated with increased HIV acquisition risk, targeting the pro-inflammatory *Prevotella* with an antibiotic in the female genital tract would logically seem to reduce host HIV acquisition risk and subsequent transmission rates. However, the longitudinal dynamics of FGT bacteria are not well-modeled, and clinical trials attempting to displace high diversity communities with *Lactobacillus* probiotics or eliminate non-*Lactobacillus* species by using the antibiotic Metronidazole[iii] show recurrence in most cases [8] [57]. To further elucidate these microbial dynamics, I studied the longitudinal FGT microbiota in the FRESH (Females Rising through Education, Support, and Health) cohort of healthy, black, HIV-uninfected young South African women aged 18-23. Following 16S rRNA V4 gene sequencing and determination of microbial community state, I constructed a Markov-based model of bacterial community transitions in the cohort. I then performed parameter sensitivity analyses to predict interventions that effectively shift CT stationary distributions to low-inflammation and low HIV risk microbial communities.

The Markov-based model accurately predicted the CT stationary distribution of bacterial communities compared to empirical data. Furthermore, I demonstrated that while CT1 is relatively stable, CT2 appears to be a transient community type, often transitioning to the more diverse cervicotypes

[iii] Oral Metronidazole and topical Metronidazole are antibiotics that specifically targets anaerobic bacteria. It is often used in the treatment of bacterial vaginosis (BV), a vaginal disease characterized by excessive growth of anaerobic bacteria. Symptoms can include vaginal discharge, foul smell, burning with urination, or itching. Often, it is asymptomatic. Bacterial vaginosis doubles the probability of acquiring HIV upon vaginal exposure to the virus [28] [7]. However, recurrence to anaerobic bacterial populations occurs in over 50% of cases after treatment with oral Metronidazole.

CT3 and CT4. I additionally show that targeting the CT1 to CT2 transition is the most potent intervention to increase the number of women with low inflammation and decreased HIV acquisition risk. These observations may be used to develop more effective methods to prevent HIV acquistion among young women in sub-Saharan Africa.

3 Cohort and DNA sequencing

Study participants were 18- to 23-year-old, HIV-uninfected women recruited through the Females Rising through Education, Support, and Health (FRESH) prospective observational study in Umlazi, South Africa. Longitudinal samples from 50 women in the cohort were collected at 3-month intervals, each with between 2 and 7 longitudinal samples. The study protocol was approved by the Biomedical Research Ethics Committee of the University of KwaZulu-Natal and the Massachusetts General Hospital Institutional Review Board (2012P001812/MGH). Informed consent was obtained after explaining the nature and possible consequences of the study.

Participants had a finger prick blood draw for HIV RNA viral load testing twice weekly. Every 3 months, participants had a peripheral blood draw, a pelvic exam, and a counselor administered an HIV risk questionnaire. During the pelvic exam, midvaginal and ectocervical swabs were collected (Catch-All Epicenter). STI testing was performed on drawn blood by Global Labs, South Africa. Participants included in this study remained HIV negative.

After isolating nucleic acids from the swabs, the V4 region of the 16S rRNA gene was sequenced [1] [10]. Sequence reads were clustered into groups called operational taxonomic units (OTUs) that were 97% similar. This cut-off is taken to reflect the genetic diversity found within the 16S gene of a bacterial species complex. Following clustering, each cluster was assigned to a bacterial taxon after alignment of the consensus sequence. OTU clustering and OTU taxa assignment were performed using the QIIME analysis pipeline [9].

Following OTU assignment, I assigned each sample to a CT following a classification schema described by Anahtar, et al [20]. Specifically, I classified each sample into one of the 4 CTs using the following pseudocode:

```
def ct_classify (sample):
    ''' classifies an abundance sample into a cervicotype '''
    if (sample is dominated by Lactobacillus) \
       and (sample is not dominated by L. iners):
        return CT1
    elif sample is dominated by L. iners:
        return CT2
    elif sample is dominated by G. vaginalis:
        return CT3
    else:
        return CT4
```

where "dominated" is as defined in Definition 1. Note that though our pseudocode for `ct_classify` assigns all non-*L. iners Lactobacillus*-dominated samples to CT1, this consists almost entirely of *L. crispatus* dominated samples. However, any other *Lactobacillus* dominated edge-case samples are also classified into CT1.

Since the `ct_classify` classification schema is heavily supervised and largely qualitative, I decided to verify these groupings using an unsupervised method. Specifically, I performed hierarchical clustering[iv] with Ward's linkage (Figure 1). By clustering into 4 groups which were similar to those found by `ct_classify`, unsupervised hierarchical dendrogram formation of the samples closely resembled Anahtar et al.'s classification schema. A few samples were hierarchically classified separately from their `ct_classify` classification; however, upon inspection, these samples were in the process of transitioning between the two CTs.

[iv] Note that, as will be discussed later, typical linkage methods cannot be applied to microbial abundance data because of inherent sparsity. Instead of performing hierarchical clustering on the raw abundance data, I instead performed hierarchical clustering on the Bray-Curtis beta diversity matrix D (later defined in Equation 41).

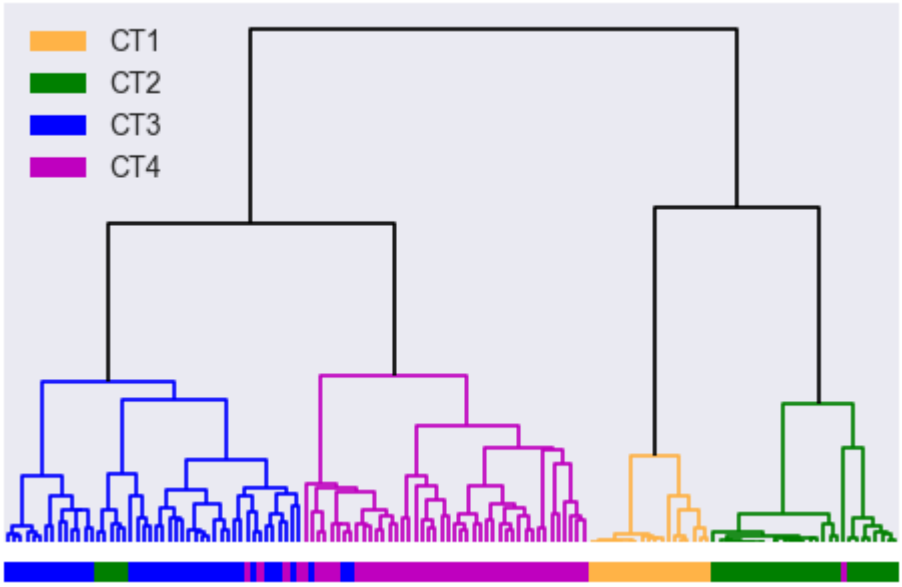

Figure 1: Hierarchical clustering dendrogram of samples with Ward's linkage. Samples are additionally labeled below the dendrogram with `ct_classify`. The unsupervised clustering closely resembled the classification schema from Anahtar et al. The few outlying samples are, in fact, in the process of transitioning between CTs.

4 Markov-based modelling

While previous studies have illustrated the cross-sectional distribution of FGT microbiota CTs [1] [20], the longitudinal dynamics of this distribution contain crucial information about underlying drivers of population-level microbial community transition. To model these longitudinal dynamics, I needed a discrete-time stochastic model for probabilistic forecasting over the CT state space. To force tractability of the model, I assume memorylessness; namely, that conditional on the present state, the future and past states are independent.

Definition 2 *Let $A_{n,t}$ be the abundance vector (a vector of length R where the ith element represents the abundance of bacterial taxa i) for person n at time-point t.*

I assume:

$$\forall n, t_p < t_c, t_c < t_f : A_{n,t_p} \perp\!\!\!\perp A_{n,t_f} | A_{n,t_c} \tag{6}$$

This is a fair assumption given the underlying biology: I assume that two people with the same FGT microbial CT have approximately the same probability of shifting to a different CT. This memorylessness property leads us directly to the Markov property, that the conditional probability distribution of future states depends only on the present state. Via Equation 6:

$$P(A_{n,t} | A_{n,t-1}, A_{n,t-2}, ..., A_{n,1}) = P(A_{n,t} | A_{n,t-1}) \tag{7}$$

I thus chose to apply a finite-state-space, time-homogeneous, discrete-time Markov chain model to the data (Figure 2). Each state in the Markov chain is one of the four CTs, and each discrete timestep represents the 3-month interval between cervical swab collections. To fully define the Markov chain, I must now estimate the transition probabilities.

Definition 3 *Define Q to be the transition matrix for the Markov chain, such that $Q_{i,j}$ represents the probability of transitioning from CT i to CT j.*

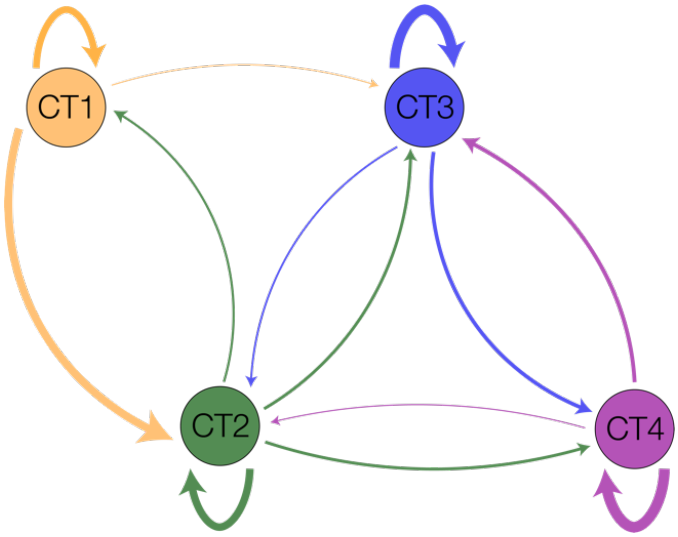

Figure 2: Markov model diagram over the CT state space where arrow thickness is proportional to the probability of that transition occurring $Q_{i,j}$.

Definition 4 *Define $S_{n,t}$ to be the CT of sample n at time-point t.*

Note that the pseudocode `ct_classify` is a function mapping a sequenced and observed $A_{n,t}$ to an $S_{n,t}, \forall n, t$.

I now investigate 2 separate methods of estimating Q.

4.1 Maximum likelihood estimates

To solve analytically for the MLE of the model, I first calculate the likelihood of a single sample n over T timesteps:

$$\mathcal{L}(Q_n) = P(S_{n,1}) \prod_{t=2}^{T} P(S_{n,t}|S_{n,t-1}) \tag{8}$$

Thus, the likelihood for all N samples is:

$$\mathcal{L}(Q) = \prod_{n=1}^{N} \left(P(S_{n,1}) \prod_{t=2}^{T} P(S_{n,t}|S_{n,t-1}) \right) \tag{9}$$

Substituting for Q:

$$\mathcal{L}(Q) = \prod_{n=1}^{N} \left(P(S_{n,1}) \prod_{t=2}^{T} Q_{S_{n,t-1},S_{n,t}} \right) \tag{10}$$

$$\mathcal{L}(Q) = \left(\prod_{n=1}^{N} P(S_{n,1}) \right) \left(\prod_{n=1}^{N} \prod_{t=2}^{T} Q_{S_{n,t-1},S_{n,t}} \right) \tag{11}$$

Note that the above equation actually multiplies the probability $Q_{i,j}$ by itself the number of times that it appears in transitions across all samples and time-points.

Definition 5 *Let $R_{i,j}$ be the total number of transitions from CT i to CT j across all N samples.*

We now rearrange our product to multiply over the 4-by-4 transition matrix Q:

$$\mathcal{L}(Q) = \left(\prod_{n=1}^{N} P(S_{n,1}) \right) \left(\prod_{i=1}^{4} \prod_{j=1}^{4} Q_{i,j}^{R_{i,j}} \right) \tag{12}$$

Calculating the log-likelihood:

$$\ln \mathcal{L}(Q) = \left(\sum_{n=1}^{N} \ln P(S_{n,1}) \right) + \left(\sum_{i=1}^{4} \sum_{j=1}^{4} R_{i,j} \ln Q_{i,j} \right) \tag{13}$$

We now want to optimize the log-likelihood $\ln \mathcal{L}(Q)$, while maintaining the constraint that the sum of probabilities leaving a CT is 1, i.e.:

$$\forall i : \sum_{j=1}^{4} Q_{i,j} = 1 \tag{14}$$

To optimize the log-likelihood for the MLE of Q, while maintaining the 4 constraints, I use 4 Lagrange multipliers [34] λ_i to form a new expression to optimize $\mathcal{K}(Q)$:

$$\mathcal{K}(Q) = \ln \mathcal{L}(Q) - \sum_{i=1}^{4} \lambda_i \left(\sum_{j=1}^{4} Q_{i,j} - 1 \right) \tag{15}$$

$$\mathcal{K}(Q) = \left(\sum_{n=1}^{N} \ln P(S_{n,1}) \right) + \left(\sum_{i=1}^{4} \sum_{j=1}^{4} R_{i,j} \ln Q_{i,j} \right) - \sum_{i=1}^{4} \lambda_i \left(\sum_{j=1}^{4} Q_{i,j} - 1 \right) \tag{16}$$

Now, to optimize our Lagrangian expression, I take derivatives of $\mathcal{K}(Q)$ with respect to parameters $Q_{i,j}$ and Lagrangian multipliers λ_i:

$$\frac{d\mathcal{K}(Q)}{dQ_{i,j}} = \frac{R_{i,j}}{Q_{i,j}} - \lambda_i \tag{17}$$

$$\frac{d\mathcal{K}(Q)}{d\lambda_i} = 1 - \sum_{j=1}^{4} Q_{i,j} \tag{18}$$

We can now set the derivatives of $\mathcal{K}(Q)$ above equal to zero and solve for $Q_{i,j}$ to optimize for the MLE of our model:

$$\frac{R_{i,j}}{Q_{i,j}} - \lambda_i = 0 \tag{19}$$

$$1 - \sum_{j=1}^{4} Q_{i,j} = 0 \tag{20}$$

Solving for $Q_{i,j}$ in Equation 19 then plugging into Equation 20, we see:

$$Q_{i,j} = \frac{R_{i,j}}{\lambda_i} \tag{21}$$

$$1 - \sum_{j=1}^{4} \frac{R_{i,j}}{\lambda_i} = 0 \tag{22}$$

$$\lambda_i = \sum_{j=1}^{4} R_{i,j} \tag{23}$$

We plug in our expression for λ_i from Equation 23 into Equation 21, to attain our expression for the MLE of the model:

$$\hat{Q}_{i,j} = \frac{R_{i,j}}{\sum_{j=1}^{4} R_{i,j}} \tag{24}$$

Note that our MLE for $\hat{Q}_{i,j}$ makes intuitive sense as it is the number of transitions observed from CT i to CT j divided by the total number of transitions from CT i to any CT.

4.2 Maximum *a posteriori* estimates

Besides the frequentist MLE derivation shown above, estimates for $Q_{i,j}$ transition probabilities can also be derived in a Bayesian paradigm.

Definition 6 *Let X_i be a random variable vector of length 4 to represent the next CT after starting from CT i. Let X_i be one-hot encoded for that CT such that all elements in the vector equal zero except the jth element equals 1 if the value of the next CT crystallizes to j.*

Conditional on starting from CT i, we know that the probabilities for the next transition are independent from the rest of the past states via Equation 7. Thus, X_i crystallizes into one of the 4 possible vectors in the support with probabilities equal to the ith row of matrix Q. Since the random variable crystallizes into one of these 4 vectors with certain probabilities, we can express each transition starting from CT i as a categorical multinoulli trial.

Definition 7 *Let p_i be a vector of length 4, where the jth element is the probability of transitioning from CT i to CT j. Note that p_i is the ith row of matrix Q.*

Via the Markov property, we thus know the distribution of X_i conditional on the probability vector p_i is a categorical multinoulli:

$$X_i|p_i \sim \text{Cat}_4(p_i) \tag{25}$$

A categorical multinoulli trial with output vector of length 4 will correctly yield a one-hot encoded vector for X_i. We further know that if we set a Dirichlet prior on p_i, via Dirichlet-categorical conjugacy [49], the posterior distribution on p_i will also be Dirichlet distributed. Since little prior information is known about the transitions between CTs, I chose to use a flat Dirichlet prior hyperparameter α, making the prior congruent to the uniform distribution over the 4 possible transitions.

$$p_i \sim \text{Dir}_4(\alpha) \tag{26}$$

$$\alpha =< 1, 1, 1, 1 > \tag{27}$$

By Dirichlet-categorical conjugacy, we know that the posterior distribution of p_i will be Dirichlet distributed with concentration parameter equal to the sum of the hyperparameter and the vector of counts transitioning from CT i.

Definition 8 *Let R_i be a vector of length 4, where the jth element represents the counts of the number of transitions starting from CT i and going to CT j. In other words, $R_i =< R_{i,1}, R_{i,2}, R_{i,3}, R_{i,4} >$.*

Conditional on observing all data, we know that the posterior distribution on p_i is Dirichlet distributed as follows:

$$p_i|R_i \sim \text{Dir}_4(\alpha + R_i) \tag{28}$$

The flat Dirichlet prior hyperparameter α essentially adds a pseudocount of 1 to each transition.

Now that we have determined the posterior distribution of $p_i|R_i$, we can calculate the Maximum *a posteriori* (MAP) estimate, the mean, and the variance of the posterior distribution.

The MAP estimate is simply the mode of the Dirichlet posterior. As previously derived, the mode of a Dirichlet [4] is:

$$\hat{Q}_{i,j} = \frac{(\alpha_j + R_{i,j}) - 1}{\sum_{j=1}^{4} (\alpha_j + R_{i,j} - 4)} \tag{29}$$

$$\hat{Q}_{i,j} = \frac{R_{i,j}}{\sum_{j=1}^{4} R_{i,j}} \tag{30}$$

We note that the MAP estimate equals the MLE estimate for our model, since a flat, uniform prior was used. However, the Bayesian framework allows us to additionally easily calculate the mean and variance of the posterior distribution and thus to construct a credible interval over each $\hat{Q}_{i,j}$. The

mean and variance of the Dirichlet are again previously derived in Bishop [4]:

$$E\big[Q_{i,j}|R_i\big] = \frac{\alpha_j + R_{i,j}}{\sum_{j=1}^{4}(\alpha_j + R_{i,j})} \tag{31}$$

$$E\big[Q_{i,j}|R_i\big] = \frac{1 + R_{i,j}}{4 + \sum_{j=1}^{4} R_{i,j}} \tag{32}$$

$$Var\big[Q_{i,j}|R_i\big] = \frac{(\alpha_j + R_{i,j})\left(\left(\sum_{j=1}^{4}(\alpha_j + R_{i,j})\right) - (\alpha_j + R_{i,j})\right)}{\left(\sum_{j=1}^{4}(\alpha_j + R_{i,j})\right)^2 \left(1 + \sum_{j=1}^{4}(\alpha_j + R_{i,j})\right)} \tag{33}$$

$$Var\big[Q_{i,j}|R_i\big] = \frac{(1 + R_{i,j})\left(3 - R_{i,j} + \sum_{j=1}^{4} R_{i,j}\right)}{\left(4 + \sum_{j=1}^{4} R_{i,j}\right)^2 \left(5 + \sum_{j=1}^{4} R_{i,j}\right)} \tag{34}$$

While credible intervals from the posterior Dirichlet distribution on Q demonstrate the range of likely values for Q, I decided to continue downstream Markov analyses with a point estimate for each $Q_{i,j}$. Since the MAP and MLE estimates both predict the same estimate, I proceeded with downstream analyses using Equations 24 and 30: $\hat{Q}_{i,j} = \frac{R_{i,j}}{\sum_{j=1}^{4} R_{i,j}}$.

4.3 Estimates implemented on data

After sequencing samples, classifying CTs, and counting CT transitions $R_{i,j}$, I used the derived estimator for $\hat{Q}_{i,j}$ to fill in the matrix $\hat{Q}$ (Fig. 3).

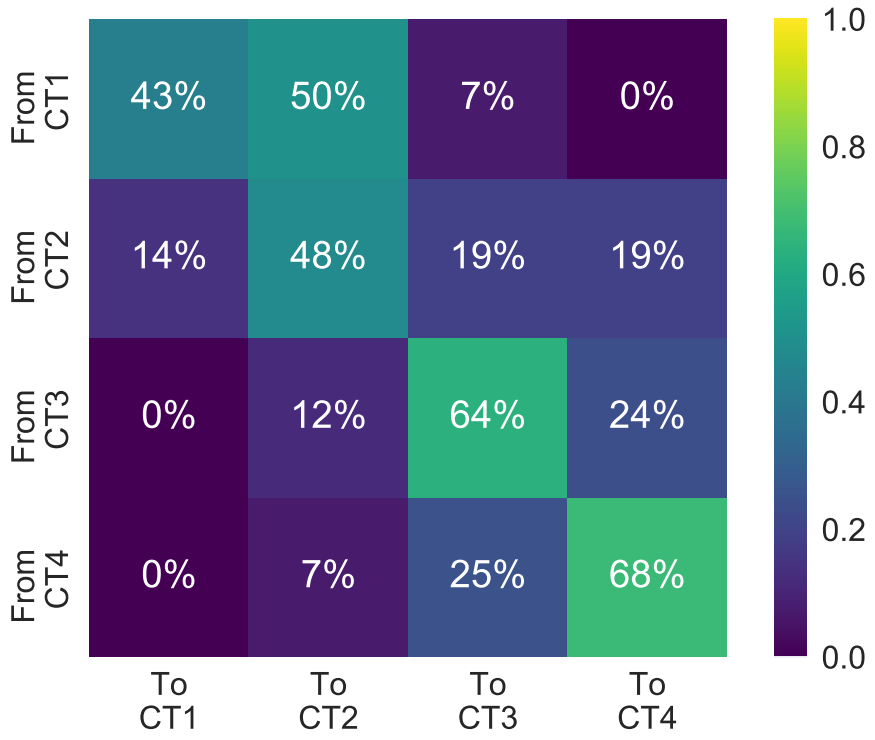

Figure 3: Markov transition probabilities $\hat{Q}$ estimated as expressed in the MLE and MAP derivations for the model.

Examining the estimated transition probability matrix $\hat{Q}$, we can see that starting in CT1, there is only a 7% probability of transitioning to one of the more inflammatory states CT3 or CT4. However, interestingly, these CT1 participants have a high probability of transitioning to the *L. iners* dominated CT2 (50% probability of moving from CT1 to CT2 compared with 43% probability of remaining in CT1). Furthermore, after arriving in

CT2, the women had an elevated probability of moving to the inflammatory CTs 3 and 4 (38% probability of moving from CT2 to CT3 or CT4 vs 7% probability of moving from CT1 to CT3 or CT4). Compared to CT2, CT3 and CT4 are "stickier," in that samples have increased probabilities of remaining in those CTs for a long period of time and not transitioning after arriving. As a result, women who transition from CT1 to CT2 in this South African cohort have a high probability of subsequently shifting to one of the pro-inflammatory CTs 3 or 4 and remaining there for long timelengths.

4.4 Stationary distributions

We note that our Markov chain is aperiodic, irreducible, and ergodic and will thus have a unique stationary distribution vector [36].

Definition 9 *Let π be the unique stationary distribution vector for the Markov chain.*

The stationary distribution of a Markov chain is an equilibrium probability distribution vector over the state space (where all entries are non-negative probabilities that sum to 1), such that the probability distribution over the state space no longer changes with time.

$$\pi Q = \pi \tag{35}$$

From Equation 35, we note that π is the left eigenvector corresponding to matrix Q's eigenvalue of 1. We can recast equation 1 to form the homogeneous linear equation:

$$\pi(I - Q) = 0 \tag{36}$$

where I is a 4-by-4 identity matrix. This homogenous linear equation can be solved using the inverse iterative method as previously described [48] [45].

Employing this technique, I can solve for the unique equilibria of Markov transition matrices. To assess the robustness of the Markov chain model applied to CT transitions, I calculated the stationary distribution of our calculated transition probability matrix and found that the distribution closely resembled an empirical [1] CT distribution from the same cohort (Figure 4).

Definition 10 *Define $\bar{\pi}$ to be a larger sample size cross-sectional CT distribution from the same cohort ($n = 146$) [1]. Assume $\bar{\pi}$ is fixed and true.*

To generate error bars and calculate significant difference, I performed bootstrapping; namely, I subsampled the transitions with replacement 10,000 times, recalculated the CT stationary distribution with the inverse iterative

method each time, then determined standard deviations over the samples. Furthermore, a chi-square goodness-of-fit test to compare the two categorical probability distributions π and $\bar{\pi}$ was insignificant (χ^2 test statistic= $7.35, p > .05$).

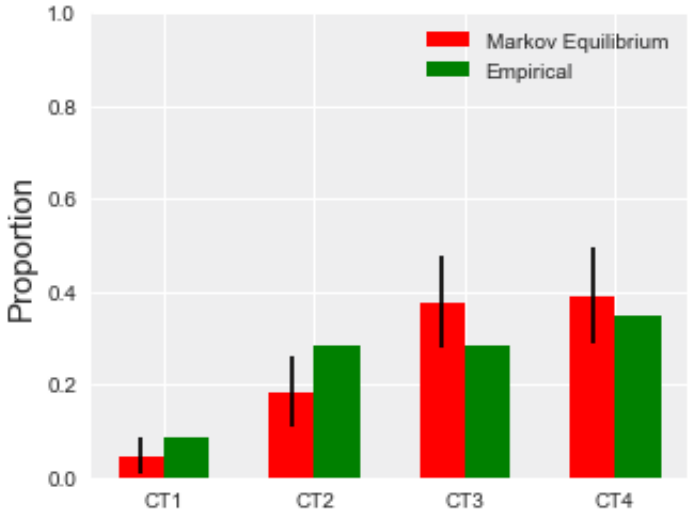

Figure 4: Markov transition probabilities $\hat{Q}$ estimated as expressed in the MLE and MAP derivations for the model. Differences between stationary distributions are statistically insignificant via a chi-square goodness-of-fit test (χ^2 test statistic= $7.35, p > .05$, power= .8840). Error bars are bootstrapped estimates of standard error.

The power of the chi-square goodness of fit test for a moderate effect size of $\omega = 0.3$ was 0.8840, where ω is Cohen's chi-square effect size, defined as:

$$\omega = \sqrt{\sum_{i=1}^{4} \frac{(\bar{\pi}_i - \pi_i)^2}{\bar{\pi}_i}} \tag{37}$$

To visualize the statistical power to detect various differences in probability between the two distributions, I calculated ω for various values of $\bar{\pi} - \pi$ and proceeded to calculate the statistical power to detect this difference

(Figure 5). Note that we can detect a 7% probability difference between the two distributions with a statistical power of 95%. An insignificant chi-square goodness of fit test ergo demonstrates that the probability difference between the two distributions $\bar{\pi} - \pi$ is likely less than 7%. Given the underlying randomness of the population, a 7% probability difference is small enough to confidently state that the Markov model accurately converged to the empirical cross-sectional distribution.

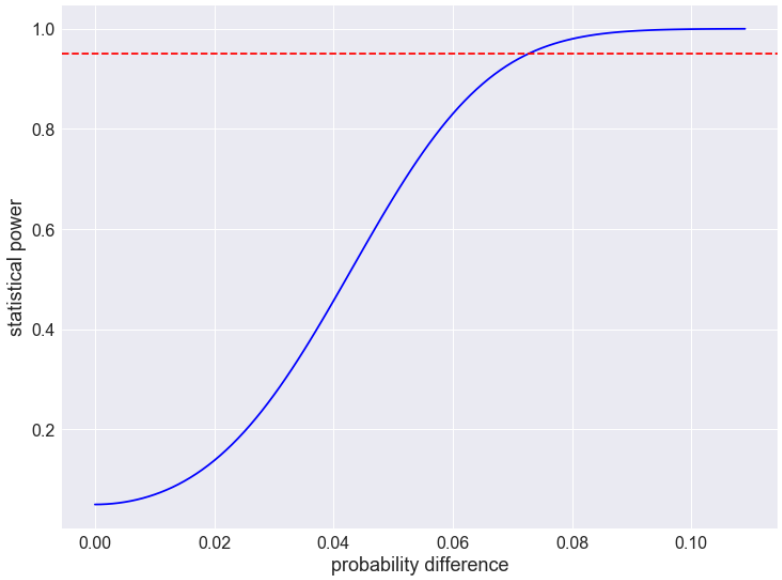

Figure 5: Power curve showing the statistical power to detect a probability difference $\bar{\pi} - \pi$ via a chi-squared goodness of fit test. The red horizontal line represents a statistical power of 95%. We can detect a 7% probability difference between the two distributions with 95% statistical power.

The similarity between my calculated stationary from solely transition data and a larger empirical cross-sectional sample from the same cohort demonstrates the robustness of my model to the conditionally independent memorylessness Markov assumption.

5 Probability preference toward high-diversity CTs

We note that by the reversibility property of Markov chains with stationary distributions, the detailed balance condition holds [16], where:

$$\forall i, j : \pi_i Q_{i,j} = \pi_j Q_{j,i} \tag{38}$$

By Equation 38, we see that there is no flux between CTs once the stationary distribution has been reached (hence the term equilibrium distribution to describe a stationary distribution). However, we can determine the "preference" of the model by finding the difference in transition probabilities between two states.

Definition 11 *Define $P_{i,j}$ to be the preference from CT i to CT j; namely, the difference in transition probabilities from CT i to CT j and CT j to CT i.*

Expressing $P_{i,j}$ mathematically, we see:

$$P_{i,j} = Q_{j,i} - Q_{i,j} \tag{39}$$

We clearly note by symmetry that:

$$P_{i,j} = -P_{j,i} \tag{40}$$

We further note that a larger value of $P_{i,j}$ implies a tendency to "prefer" state i when transitioning between states i and j. We can use the preference values between all states to determine the direction of CT movement conditional on a discrete uniform population density probability distribution over the CT states (Figure 6). I thus demonstrate that after ignoring any population probability differences over the CT state space, the transition dynamics of the system prefer the high inflammatory states CT3 and CT4.

Conditional on a uniform population density over the CTs, preferences $P_{i,j}$

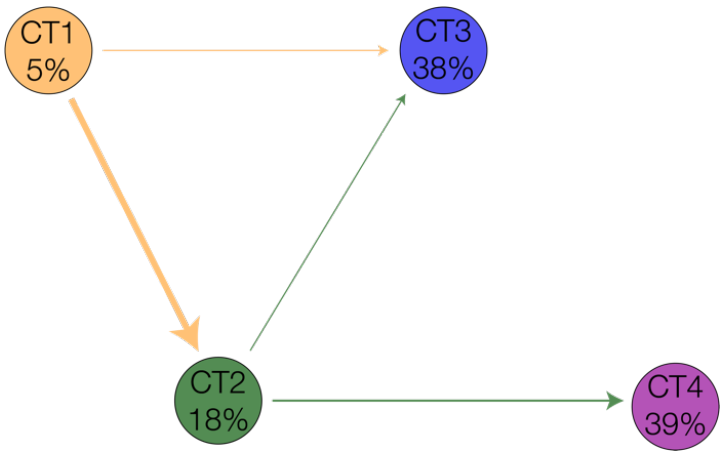

Figure 6: Arrows between CT states have thicknesses proportional to preference $P_{i,j}$. Conditional on a uniform population density over the states, these arrows would represent flux. However, the population density at equilibrium (stationary probability labeled on each state) balances this preference to make the flux between any two states equal to zero.

would represent flux. However, the stationary state probability distribution balances this preference to make flux zero. Thus, preferred states CT3 and CT4 have increased probability in the stationary distribution. These $P_{i,j}$ preferences ergo push women toward more diverse and HIV-predisposed CTs.

5.1 Non-Euclidean dimensionality reduction

Dimensionality reduction can be used to further demonstrate the impetus in the cohort toward higher CTs due to the $P_{i,j}$ preferences. However, the Euclidean distance used by simple linear methods of dimensionality reduction such as Principal Component Analysis (PCA) fail to deal with double zeros inherent in sparse data like bacterial abundances. PCA fails to properly attain uncorrelated components because of this inherent sparsity.

Instead of PCA on sparse abundance data, we perform dimensionality reduction on a dissimilarity matrix:

Definition 12 *Let D be a dissimilarity matrix, where the element in the nth row and mth column of the matrix represents the dissimilarity between samples n and m.*

We now have a choice on which metric to use for the dissimilarity between samples. Multiple dissimilarity indices are used with ecological data including Bray-Curtis (count-based), Jensen-Shannon (entropy-based), and Unifrac (phylogenetic tree-based). After analyzing literature using the metrics, I determined that the Bray-Curtis metric was the most appropriate for discrete data like 16S rRNA sequencing read counts and was the most consistent for sample sizes in the hundreds [12] [30] [15] [42] [46] [17] [18] [26] [53] [32] [47].

Definition 13 *Define $\mathcal{D}_{n,m}$ to be the Bray-Curtis dissimilarity index (also referred to as Bray-Curtis beta diversity) calculated between sample n and sample m. Note that $\mathcal{D}_{n,m}$ are entries in matrix D.*

We define read counts, since Bray-Curtis dissimilarity is a count-based metric:

Definition 14 *$\forall n, k$ let $C_{n,k}$ be the number of reads assigned to bacterial OTU k from sample n.*

Recall R as the number of bacterial OTU's sequenced. We calculate Bray-Curtis beta diversity as follows:

$$\mathcal{D}_{n,m} = 1 - 2\frac{\sum_{k=1}^{R} \min(C_{n,k}, C_{m,k})}{\sum_{k=1}^{R} (C_{n,k} + C_{m,k})} \tag{41}$$

Looking at the above calculation for Bray-Curtis beta diversity, we see that if the two samples are identical, then:

$$\forall n, k : C_{n,k} = C_{m,k} \tag{42}$$

Thus calculating the Bray-Curtis beta diversity for two identical samples we see:

$$\mathcal{D}_{\text{identical}} = 1 - 2\frac{\sum_{k=1}^{R} C_{n,k}}{\sum_{k=1}^{R} (2C_{n,k})} \tag{43}$$

$$\mathcal{D}_{\text{identical}} = 0 \tag{44}$$

Thus the Bray-Curtis beta diversity between two identical samples is zero. If we have two "opposite" samples, such that at least one of the two samples has zero counts for every OTU, then we calculate the Bray-Curtis beta diversity as:

$$\mathcal{D}_{\text{opposite}} = 1 - 2\frac{0}{\sum_{k=1}^{R} (C_{n,k} + C_{m,k})} \tag{45}$$

$$\mathcal{D}_{\text{opposite}} = 1 \tag{46}$$

The Bray-Curtis beta diversity between two opposite samples is therefore one. It is now easy to see that all values of Bray-Curtis beta diversity must lie in $[0, 1]$. More "similar" samples will have a $\mathcal{D}_{n,m}$ value close to zero, while dissimilar samples will have a $\mathcal{D}_{n,m}$ value close to one.

Furthermore, note that the Bray-Curtis beta diversity index does not follow the triangle inequality [22]. As a result, we must use a non-linear dimensionality reduction (unlike PCA) to maintain samples with low Bray-Curtis beta diversity close together and separate samples with high Bray-Curtis beta diversity. Principal Coordinate Analysis (PCoA), a method of multi-dimensional scaling (MDS), accomplishes this by minimizing a loss function

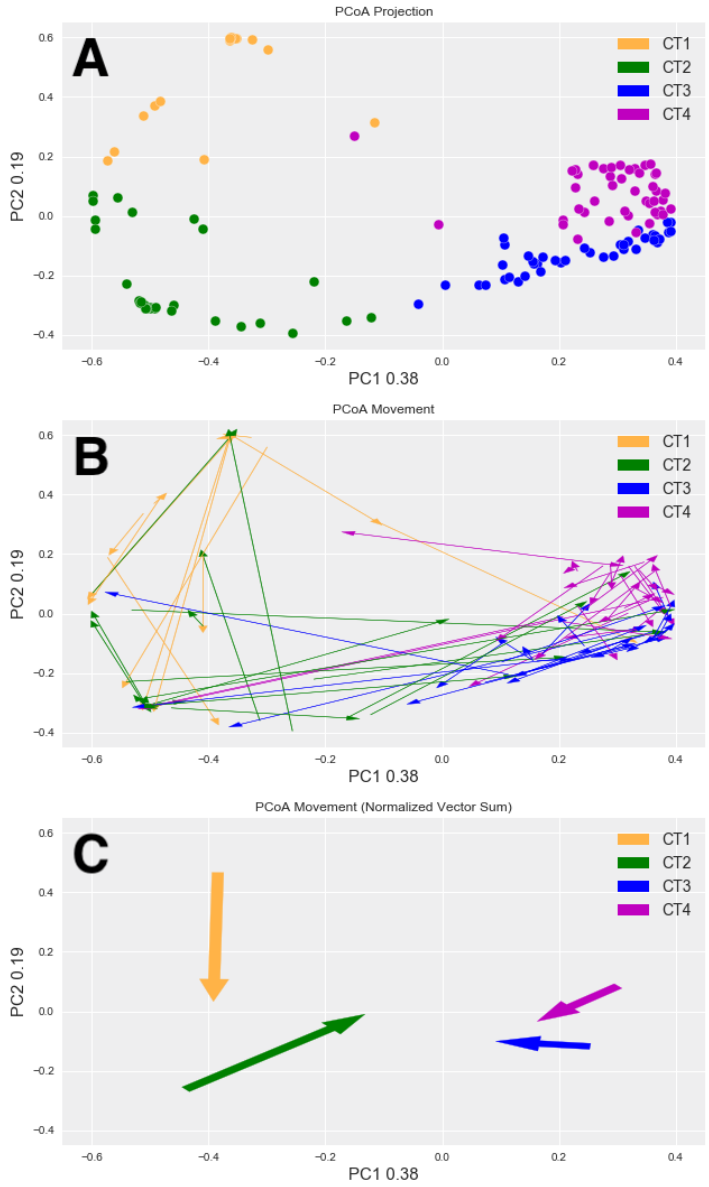

Figure 7: A: PCoA projection of samples colored by CT. B: Adjacent time-point PCoA projected samples connected by arrows colored by starting CT. C: Vector sum of colored arrows. Dimensionality reduction allows visualization of the path from CT1 to CT3/4 going through CT2.

of distance between similar samples, as previously described [21].

I used PCoA to project the samples onto a non-Euclidean principle coordinate space (Figure 7A). Next, I connected adjacent time-points in PCoA space with an arrow, colored by the starting CT (Figure 7B). Note that long arrows represent large changes in PCoA space, and thus, large shifts in underlying microbiota are predominantly going to or coming from CT2. Furthermore, by taking the vector sum of arrows leaving a given CT, I determined the net movement in PCoA state starting from a given CT (Figure 7C). I hereby demonstrate that the largest expected movements in PCoA space occur starting from CT2 (note that the x-axis represents twice the variance as the y-axis in these projections). This trend further illustrates the instability of *L. iners* as an FGT colonizer, for its colonization often causes large-scale movements in PCoA space.

5.2 Taxon-level differences

To further investigate the differences between participants that return to their CT and participants that are pushed toward CT3/4, I chose to examine the data at a more specific taxon-based level. I first plotted the abundances of bacterial OTUs after subdividing the CT clusters into samples that remained in the same CT at the next measured time-point ("Will Return") or samples that shifted to a different CT at the next measured time-point ("Won't Return"). After ordering samples within groupings by their PCoA1 component, this taxon-level visualization revealed similar CT clustering as found in previous studies [20] (Figure 8).

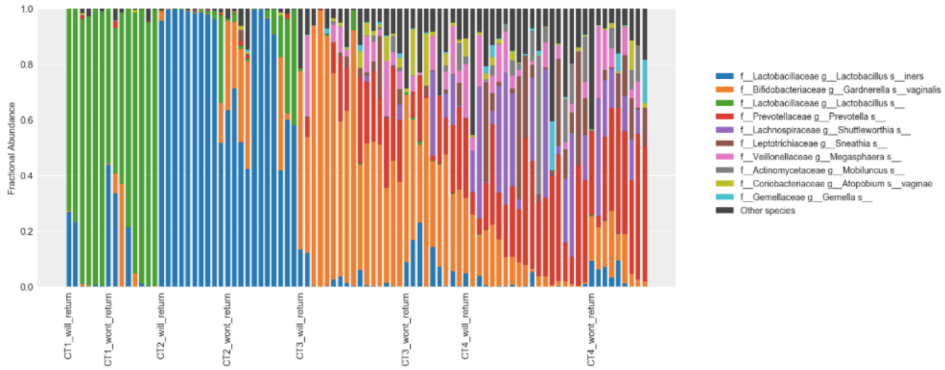

Figure 8: OTU abundances of samples separated by CT and subdivided by samples that return to the same CT at the next time-point and samples that move to a different CT at the next time-point. Samples are ordered within groupings by their PCoA1 component.

To statistically test for differences between the returning and non-returning groups, I first looked at Shannon alpha diversity H (previously defined). I determined that Shannon alpha diversity in the FGT significantly increased the probability of shifting CT conditional on being in CT2 or CT3 (Figure 9).

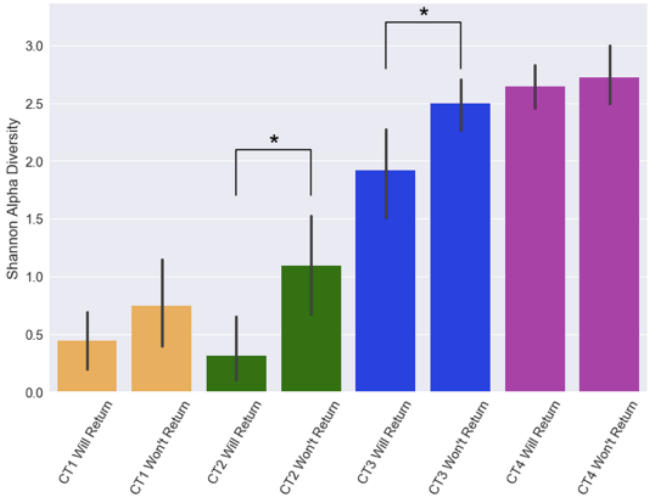

Figure 9: Shannon alpha diversity significantly increased the probability of shifting CTs conditional on a sample being in CT2 or CT3. Significant difference was calculated using non-parametric two-sample Monte Carlo t-tests. * $p < .05$

Interestingly, having more *L. iners* specifically increased the probability of a transition, a trend which can be observed by directly comparing taxa abundances in samples that return vs don't return (Figure 10). The *L. iners* taxa was thus most associated with community instability and CT transitions (the only other taxa significantly associated with CT transitions was *G. vaginalis*, which was associated with shifting from CT2 to CT3).

In conclusion, FGT microbial dynamics display a preference in transitioning to CT's 3 and 4 in the cohort, while elevated Shannon alpha diversity or the presence of common higher CT bacterial taxa increase the probability of these transitions.

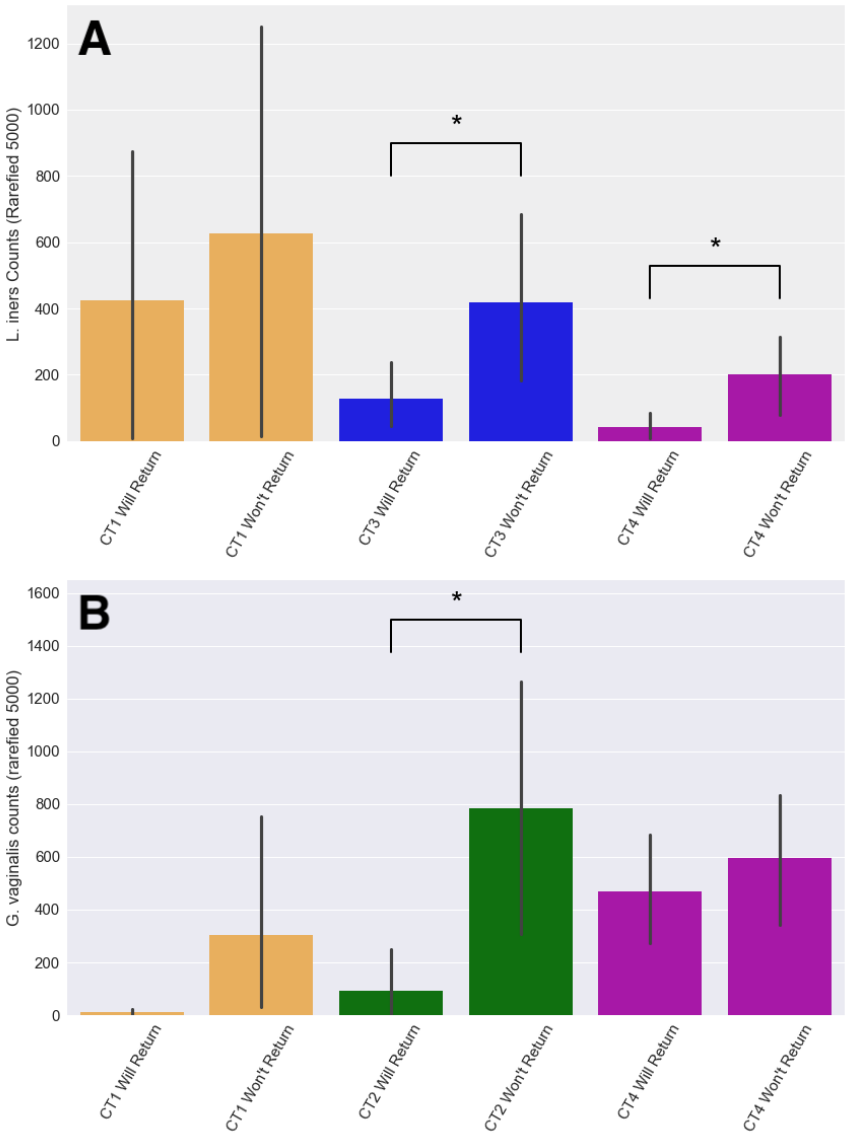

Figure 10: *L. iners* (A) and *G. vaginalis* (B) counts by CT subdivided by samples that return to the same CT at the next time-point and samples that move to a different CT at the next time-point. Significant differences were calculated using Kruskal-Wallis non-parametric ANOVA. * $p < .05$

6 Cross-sectional population CT shifts revert

After demonstrating the internal cohort tendency to push the population toward a predominantly diverse, pro-inflammatory CT distribution, I decided to investigate the effect of potential interventions. One potential intervention is the use of *Lactobacillus* probiotics or an anaerobic antibiotic such as oral Metronidazole, yet these methodologies have interestingly demonstrated recurrence to the original microbial distribution following treatment in most cases [8] [57].

To further explore this recurrence, I simulated maintaining our transition dynamics following a cross-sectional intervention such as probiotics or oral Metronidazole.

Definition 15 *Let $\mathcal{T}_k$ be a probability mass function vector of length 4 where the ith element of the vector represents the probability of being in CT i after k time-points.*

Note that $\mathcal{T}_k|Q, \mathcal{T}_0$ is deterministic and not a random variable.

Since treating with probiotics or oral Metronidazole causes a patient to be dominated by *L. iners* or *L. crispatus* in the cohort, we now assume 100% of the population is cross-sectionally treated and *Lactobacillus*-dominated for simulation:

$$\mathcal{T}_0(\text{L. crisp}) =< 1, 0, 0, 0 > \tag{47}$$

$$\mathcal{T}_0(\text{L. iners}) =< 0, 1, 0, 0 > \tag{48}$$

Given our definition of Q, we know that matrix-multiplying a probability distribution $\mathcal{T}_k$ by this value yields the next timestep's probability mass function:

$$\forall k : \mathcal{T}_{k+1} = \mathcal{T}_k Q \tag{49}$$

And iterating this multiplication:

$$\mathcal{T}_k|Q, \mathcal{T}_0 = \mathcal{T}_0 Q^k \tag{50}$$

We know that any initialization on the Markov chain will converge to π by definition of the stationary distribution:

$$\forall \mathcal{T}_0 : \lim_{k \to \infty} \mathcal{T}_k | \mathcal{T}_0 = \pi \tag{51}$$

However the rate of this convergence to the stationary π varies greatly based on both the transition probabilities Q and on the initialization of the chain $\mathcal{T}_0$. To investigate the rate of the convergence to π in our cohort following probiotic or oral Metronidazole treatment, I modeled the probability dispersion using Equation 50 after initializing to $\mathcal{T}_0$(L. crisp) and $\mathcal{T}_0$(L. crisp) from Equations 47 and 48 (Figure 11).

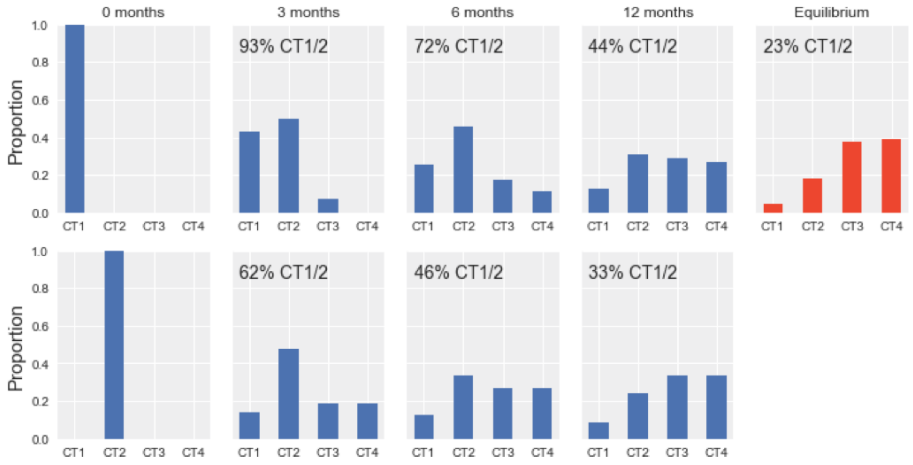

Figure 11: Simulations of Markov chain transitions after initializing to an entirely CT1 or entirely CT2 population quickly revert to the stationary.

Shifting a population to an entirely CT2 distribution quickly rebounds to the previous equilibrium distribution; and while shifting to an entirely CT1 distribution slows the recurrence to the equilibrium, approximate recurrence still occurs within just a year with the cross-sectional intervention.

7 Interventional targeting of CT1-CT2 transition

After demonstrating the ineffectiveness of cross-sectional interventions in causing chronic shifts in FGT microbial population distributions, I chose to investigate longitudinal interventions that change the underlying transition probabilities of the chain. Specifically, I investigate two types of interventions: "edge-boosting" and "edge-blocking."

7.1 Edge interventions

Definition 16 *Define the edge-boosting intervention between CT i and CT j by rate $\alpha \in [0, 1]$ as decreasing the probability that CT j returns by α and instead transitioning to CT i.*

More specifically, the edge-boost intervention shifts probability density in the transition matrix from returning to CT j to instead transitioning to CT i as can be seen in the pseudocode below:

```
def edge_boost (i, j, alpha, Q):
    ''' edge-boosts the CT i to CT j transition by alpha '''
    Q_new = copy.deepcopy(Q)
    Q_new[j,j] = Q[j,j] - alpha * Q[j,j]
    Q_new[j,i] = Q[j,i] + alpha * Q[j,j]
    return Q_new
```

Similarly I define the edge-blocking intervention:

Definition 17 *Define the edge-blocking intervention between CT i and CT j by rate $\beta \in [0, 1]$ as decreasing the probability that CT i transitions to CT j by β and instead returns to CT i.*

The edge-block intervention shifts probability density in the transition matrix from transitioning to CT j to instead returning to CT i as can be again seen in the pseudocode below:

```
def edge_block (i, j, beta, Q):
    ''' edge-blocks the CT i to CT j transition by beta '''
    Q_new = copy.deepcopy(Q)
    Q_new[i,j] = Q[i,j] - beta * Q[i,j]
    Q_new[i,i] = Q[i,i] + beta * Q[i,j]
    return Q_new
```

By changing the transition probabilities of the chain, we fundamentally change the underlying dynamics of the network with the intervention and cause long-term changes to the population's CT distribution.

Definition 18 *Define an edge to be one of the* $\binom{4}{2} = 6$ *unique pairs of CTs, i.e.* $\forall i \neq j : (CT\ i, CT\ j)$.

I simulated the edge-blocking and edge-boosting interventions on all 6 edges for a range of $< \alpha, \beta >$ values, calculating the new CT stationary distribution with the updated transition probability matrix at each iteration (Figure 12).

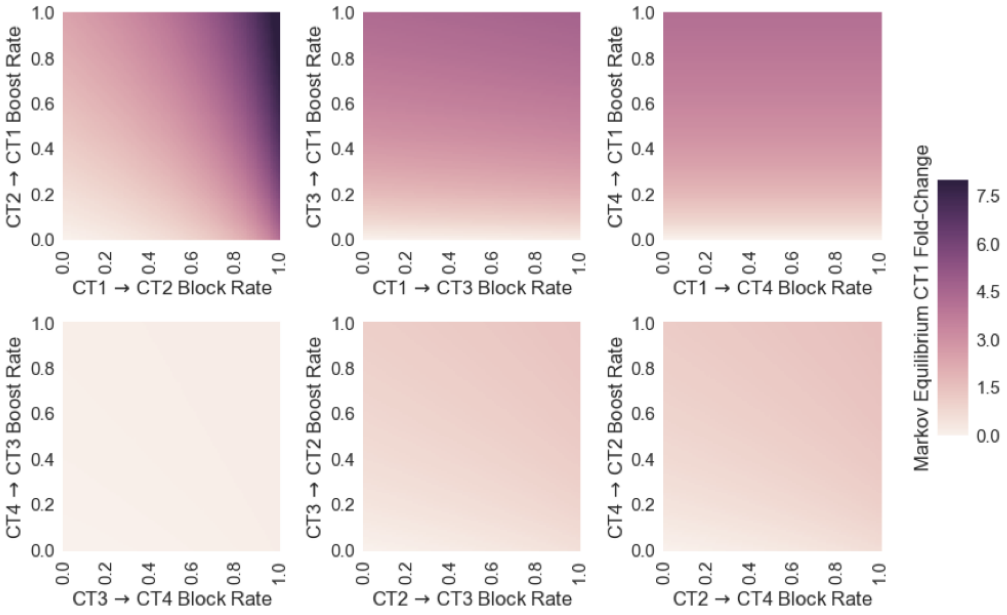

Figure 12: Fold-change in proportion of population in CT 1 at stationary distribution, following edge-boost and edge-block interventions along an individual edge with various $< \alpha, \beta >$ values.

Corresponding to my previous findings about the transient nature of *L. in-*

ers dominance, edge-blocking or edge-boosting along CT2-CT3 or CT2-CT4 edges has a negligible effect on the underlying population. This negligible fold-change in *L. crispatus* at equilibrium following shifting dynamics to prefer *L. iners* further explains the recurrence following oral Metronidazole or probiotics. The most efficacious interventions were the CT edges that specifically edge-boosted *L. crispatus*. And interestingly, of all the CT edges, the most potent intervention was the CT1-CT2 edge, which resulted in dramatic shifts in stationary distributions. This finding illustrates the importance of carefully differentiating between *L. crispatus* and *L. iners* when considering interventions.

7.2 Combining edges

Noting the importance of focusing interventions on *L. crispatus*, I next explored an intervention combining multiple effective edges: edge-boosting and edge-blocking the CT1-CT2, CT1-CT3, and CT1-CT4 edges simultaneously.

Definition 19 *Define the CT1-boost-block intervention by rate* $< \alpha, \beta >$ *s.t.* $\alpha \in [0,1], \beta \in [0,1]$ *to edge-block all edges leaving CT1 by* α *and edge-boost all edges entering CT1 by* β*.*

This intervention focuses on *L. crispatus* instead of failing to differentiate between *L. crispatus* and *L. iners*, as can be more specifically seen through the pseudocode below:

```
def ct1_boost_block(alpha, beta, Q):
    ''' performs CT1-boost-block by <alpha,beta> '''
    Q_new = zeros((4,4)) #4-by-4 matrix of zeros
    Q_new[0,1:] = Q[0,1:] - beta * Q[0,1:]
    Q_new[0,0] = Q[0,0] + sum(beta * Q[0,1:])
    Q_new[1:,1:] = Q[1:,1:] - alpha * Q[1:,1:]
    Q_new[1:,0] = Q[1:,0] + sum(alpha * Q[1:,1:], axis=1)
    return Q_new
```

Note that this pseudocode is equivalent to calling `edge-block` and `edge-boost` iteratively:

```
def ct1_boost_block(alpha, beta, Q):
    ''' performs CT1-boost-block by <alpha,beta> '''
    Q_new = copy.deepcopy(Q)
    for all (i,j) with i<j:
        if i == 0:
            Q_new = edge_block(i, j, beta, Q_new)
        else:
            Q_new = edge_boost(j, i, alpha, Q_new)
    return Q_new
```

I performed CT1-boost-block for various $< \alpha, \beta >$ pair values on the data (Figure 13).

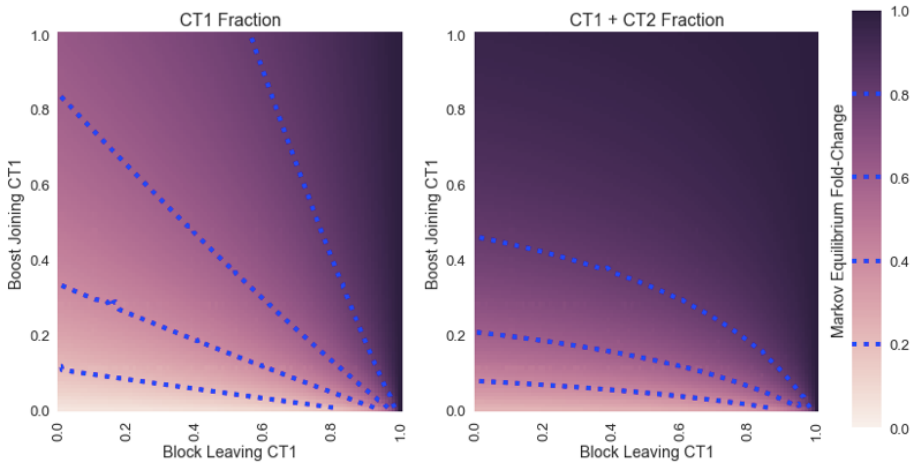

Figure 13: Fold-change in proportion of population in CT 1 and CT 1 or CT2 at stationary distribution, following a CT1-boost-block intervention along an individual edge with various $< \alpha, \beta >$ values.

Through this potent intervention, a 50% α boost rate and 50% β block rate results in a stationary CT distribution with 60% of the population in CT1 and more than 80% of the population in a low-inflammation CT1 or CT2.

In conclusion, focusing on longitudinal interventions that specifically shift to *L. crispatus* dominance demonstrated the greatest impact on increasing the low-inflammatory population.

8 Conclusion

8.1 Challenging *Lactobacillus* FGT literature

Previous studies have cross-sectionally examined HIV incidence with FGT microbiota and have found that women with CT3 or CT4 cervicovaginal microbial communities acquire HIV at significantly higher rates. Furthermore, women in CT3 or CT4 are also more likely to have bacterial vaginosis (BV) and sexually transmitted infections other than HIV such as *Neisseria gonorrhoeae* (NGI) and *Chlamydia trachomatis* (CTI) infection [14] [35] [52]. As a result of the cross-sectional differences in vaginal health between CT1/CT2 and CT3/CT4, the field's literature is dominated by the opinion that *Lactobacillus* species are beneficial for vaginal health, since women in CT1 or CT2 are less likely to acquire HIV, BV, NGI, or CTI [5] [2] [37] [50] [40] [6] [44] [56] [41] [43] [3].

However, I hereby respond to this literature by providing the first examination of FGT microbiota in sub-Saharan African women using next-generation sequencing from a longitudinal perspective. While women are not likely to acquire HIV while *Lactobacillus*-dominant, that does not guarantee that all species of *Lactobacillus* are safe. Via modeling FGT microbiota from a longitudinal perspective, I show that in fact, one species of *Lactobacillus*, *L. iners*, may not be as safe as the field's literature indicates.

While it is indeed true that women with *L. iners*-dominated vaginal flora are much less likely to acquire HIV in the immediate future, I created a longitudinal model to predict the dynamics of these bacterial communities in the long-term. Interestingly, my longitudinal model demonstrated that *L. iners* may actually serve as a "stepping stone" between the healthy *L. crispatus* dominance and the high-HIV risk CT3 and CT4. Many participants demonstrated a longitudinal trend of passing through CT2 on the way to

more high-diversity communities. This observation demonstrates that while *L. iners* itself may not increase a woman's HIV acquisition risk, it favors transitions toward bacterial communities that increase risk in the long-term.

Furthermore, the *L. iners* dominance of the FGT appears to be significantly less stable than the dominance by other bacterial species. *L. iners* is a less "absorbing" dominator and has a very high probability of transitioning to CT3 or CT4, whereas CT1 has a low probability of transitioning to CT3 or CT4. Because *L. iners* is transient and tends to "prefer" the high-diversity CTs 3 and 4 when transitioning, women in CT2 often change CTs and this transition is likely to a more HIV-risky CT. These HIV-risky CTs are more absorbing with high return probabilities, so women tend to stay at these CTs for long times after arriving.

My observations demonstrate the importance of carefully differentiating between *L. crispatus* and *L. iners* in the FGT, and will hopefully direct future research toward this differentiation.

8.2 Simulating interventions

The transient nature and long-term riskiness of *L. iners* may explain the surprising ineffectiveness of single-time-point interventions. The clinical observation that most women revert to their previous microbial distribution following treatment with oral/topical Metranidazole (an antibiotic that targets anaerobic bacteria, leaving aerobic *Lactobacilli*) or *Lactobacillus* probiotics echoes the ineffectiveness I showed of single time-point interventions in my longitudinal model. After single time-point interventions, simulated samples reverted to equilibrium at a fast rate, and the simulation was very close to the steady state π again after a single year.

However, I showed that when interventions tuned the underlying dynamics of the chain, long-term population-level effects were possible. Specifically, the most potent intervention target was in fact not focusing on the CT1-CT4 transition, but rather the CT1-CT2 transition. Interventions that focused on decreasing the probability that bacterial communities leave CT1 and go to CT2 ("edge-blocking") or on increasing the probability that communities transition from CT2 to CT1 ("edge-boosting") most increased the number of women in low-HIV-risk CTs at stationary.

These observations will hopefully direct future research to examine interventions that specifically affect the transition between *L. crispatus* dominance and *L. iners* dominance. Treatment with a drug that prevents CT1 to CT2 transitions, for example, may be substantially more potent than oral/topical Metranidazole or *Lactobacillus* probiotics, decreasing the risk of HIV infection for the many women in high-incidence areas of sub-Saharan Africa.

9 Supplemental information

Differences in group Shannon alpha diversity were calculated using non-parametric two-sample Monte Carlo t-tests. Significantly different taxa between CTs were calculated using Kruskal-Wallis non-parametric ANOVA. P-values are two-sided and are not adjusted for multiple hypothesis testing. Statistical analyses were performed in QIIME or Python with SciPy [27] and NumPy [51]. Power calculations were performed in R [38] with `pwr` [11].

This research was conducted at the Ragon Institute of MGH, MIT, and Harvard.

10 Acknowledgements

I would like to acknowledge my group's Principal Investigator, Dr. Douglas Kwon, who helped plan the project, met with me throughout the year that I've been working in the lab, and provided both the freedom to investigate the data as well as guidance on noteworthy avenues of investigation and biological background. I'd like to thank postdoc Matt Hayward for helping edit the thesis, graduate student David Gootenberg for helping formulate computational paradigms, research technician Mara Farcasanu for helping with data organization, and postdoc Seth Bloom for explaining the underlying biological phenomena. I'd also like to thank everyone else in the Kwon Lab for making the lab such a welcoming and special place.

Thank you to Prof. Sean Eddy from the Harvard Applied Mathematics and Molecular & Cellular Biology Departments for evaluating this thesis.

Additional thanks to Kristina Li for the edits and support for this thesis, couldn't have done it without the snacks and coffees.

Last but not least, I'd like to acknowledge my parents who have always supported my mathematical interests and who fueled a deep-rooted desire to help people, guiding me to fill the interdisciplinary, impactful, and truly fascinating field of computational biology.

11 References

[1] Melis N Anahtar et al. "Cervicovaginal bacteria are a major modulator of host inflammatory responses in the female genital tract". In: *Immunity* 42.5 (2015), pp. 965–976.

[2] Kingsley C Anukam et al. "Lactobacillus vaginal microbiota of women attending a reproductive health care service in Benin city, Nigeria". In: *Sexually transmitted diseases* 33.1 (2006), pp. 59–62.

[3] Alla Aroutcheva et al. "Defense factors of vaginal lactobacilli". In: *American Journal of Obstetrics & Gynecology* 185.2 (2001), pp. 375–379.

[4] Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[5] Sandra Borges, Joana Silva, and Paula Teixeira. "The role of lactobacilli and probiotics in maintaining vaginal health". In: *Archives of gynecology and obstetrics* 289.3 (2014), pp. 479–489.

[6] Soledad Boris and Covadonga Barbés. "Role played by lactobacilli in controlling the population of vaginal pathogens". In: *Microbes and infection* 2.5 (2000), pp. 543–546.

[7] Catriona S Bradshaw and Rebecca M Brotman. "Making inroads into improving treatment of bacterial vaginosis–striving for long-term cure". In: *BMC infectious diseases* 15.1 (2015), p. 292.

[8] Catriona S Bradshaw et al. "Efficacy of oral metronidazole with vaginal clindamycin or vaginal probiotic for bacterial vaginosis: randomised placebo-controlled double-blind trial". In: *PLoS One* 7.4 (2012), e34540.

[9] J Gregory Caporaso et al. "QIIME allows analysis of high-throughput community sequencing data". In: *Nature methods* 7.5 (2010), p. 335.

[10] J Gregory Caporaso et al. "Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms". In: *The ISME journal* 6.8 (2012), p. 1621.

[11] Stephane Champely. "pwr: Basic functions for power analysis. R package version 1.1. 1". In: *The R Foundation: Vienna* (2009).

[12] K Robert Clarke, Paul J Somerfield, and M Gee Chapman. "On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray–Curtis coefficient for denuded assemblages". In: *Journal of Experimental Marine Biology and Ecology* 330.1 (2006), pp. 55–80.

[13] Mary L Delaney et al. "Nugent score related to vaginal culture in pregnant women1". In: *Obstetrics & Gynecology* 98.1 (2001), pp. 79–84.

[14] David A Eschenbach et al. "Prevalence of hydrogen peroxide-producing Lactobacillus species in normal women and women with bacterial vaginosis." In: *Journal of clinical microbiology* 27.2 (1989), pp. 251–256.

[15] Karoline Faust et al. "Microbial co-occurrence relationships in the human microbiome". In: *PLoS computational biology* 8.7 (2012), e1002606.

[16] Crispin Gardiner. *Stochastic methods*. Vol. 4. springer Berlin, 2009.

[17] Jean-François Ghiglione et al. "Pole-to-pole biogeography of surface and deep marine bacterial communities". In: *Proceedings of the National Academy of Sciences* 109.43 (2012), pp. 17633–17638.

[18] Jean-François Ghiglione et al. "Pole-to-pole biogeography of surface and deep marine bacterial communities". In: *Proceedings of the National Academy of Sciences* 109.43 (2012), pp. 17633–17638.

[19] David B Gootenberg, Caroline M Mitchell, Douglas S Kwon, et al. "Cervicovaginal Microbiota and Reproductive Health: The Virtue of Simplicity". In: *Cell host & microbe* 23.2 (2018), pp. 159–168.

[20] Christina Gosmann et al. “Lactobacillus-deficient cervicovaginal bacterial communities are associated with increased HIV acquisition in young South African women”. In: *Immunity* 46.1 (2017), pp. 29–37.

[21] John C Gower. “Principal coordinates analysis”. In: *Encyclopedia of biostatistics* (2005).

[22] John C Gower and Pierre Legendre. “Metric and Euclidean properties of dissimilarity coefficients”. In: *Journal of classification* 3.1 (1986), pp. 5–48.

[23] Gale B Hill. “The microbiology of bacterial vaginosis”. In: *American journal of obstetrics and gynecology* 169.2 (1993), pp. 450–454.

[24] Sharon L Hillier et al. “The normal vaginal flora, H2O2-producing lactobacilli, and bacterial vaginosis in pregnant women”. In: *Clinical Infectious Diseases* 16.Supplement_4 (1993), S273–S281.

[25] Joint United Nations Programme on HIV/AIDS (UNAIDS) et al. “UNAIDS data 2017”. In: *UNAIDS: Geneva, Switzerland* (2017).

[26] Curtis Huttenhower et al. “Structure, function and diversity of the healthy human microbiome”. In: *Nature* 486.7402 (2012), p. 207.

[27] Eric Jones, Travis Oliphant, and Pearu Peterson. “{SciPy}: open source scientific tools for {Python}”. In: (2014).

[28] Chris Kenyon, Robert Colebunders, and Tania Crucitti. “The global epidemiology of bacterial vaginosis: a systematic review”. In: *American Journal of Obstetrics & Gynecology* 209.6 (2013), pp. 505–523.

[29] J Lajoie et al. “A distinct cytokine and chemokine profile at the genital mucosa is associated with HIV-1 protection among HIV-exposed seronegative commercial sex workers”. In: *Mucosal immunology* 5.3 (2012), p. 277.

[30] Morgan GI Langille et al. “Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences”. In: *Nature biotechnology* 31.9 (2013), p. 814.

[31] Lindi Masson et al. “Genital inflammation and the risk of HIV acquisition in women”. In: *Clinical Infectious Diseases* 61.2 (2015), pp. 260–269.

[32] Michael G Michie. “Use of the Bray-Curtis similarity measure in cluster analysis of foraminiferal data”. In: *Journal of the International Association for Mathematical Geology* 14.6 (1982), pp. 661–667.

[33] Charles Morrison et al. “Cervical inflammation and immunity associated with hormonal contraception, pregnancy, and HIV-1 seroconversion”. In: *JAIDS Journal of Acquired Immune Deficiency Syndromes* 66.2 (2014), pp. 109–117.

[34] “Note: Maximum Likelihood Estimation for Markov Chains”. In: *Carnegie Mellon University Statistics* (2009).

[35] Robert P Nugent, Marijane A Krohn, and Sharon L Hillier. “Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation.” In: *Journal of clinical microbiology* 29.2 (1991), pp. 297–301.

[36] Dianne P O’leary. “Iterative methods for finding the stationary vector for Markov chains”. In: *Linear Algebra, Markov Chains, and Queueing Models*. Springer, 1993, pp. 125–136.

[37] Mariya I Petrova et al. “Lactobacillus species as biomarkers and agents that can promote various aspects of vaginal health”. In: *Frontiers in physiology* 6 (2015), p. 81.

[38] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2016. URL: `https://www.R-project.org/`.

[39] Jacques Ravel et al. “Vaginal microbiome of reproductive-age women”. In: *Proceedings of the National Academy of Sciences* 108.Supplement 1 (2011), pp. 4680–4687.

[40] Jacques Ravel et al. “Vaginal microbiome of reproductive-age women”. In: *Proceedings of the National Academy of Sciences* 108.Supplement 1 (2011), pp. 4680–4687.

[41] Vicente Redondo-Lopez, Roger L Cook, and Jack D Sobel. “Emerging role of lactobacilli in the control and maintenance of the vaginal bacterial microflora”. In: *Reviews of infectious diseases* 12.5 (1990), pp. 856–872.

[42] Gavin N Rees et al. “Ordination and significance testing of microbial community composition derived from terminal restriction fragment length polymorphisms: application of multivariate statistics”. In: *Antonie Van Leeuwenhoek* 86.4 (2004), pp. 339–347.

[43] Gregor Reid and Jeremy Burton. “Use of Lactobacillus to prevent infection by pathogenic bacteria”. In: *Microbes and infection* 4.3 (2002), pp. 319–324.

[44] Gregor Reid et al. “Probiotic Lactobacillus dose required to restore and maintain a normal vaginal flora”. In: *FEMS Immunology & Medical Microbiology* 32.1 (2001), pp. 37–41.

[45] Riccardo Scalco. “Pykov, a Python module on finite regular Markov chains.” In: *Github Repository* (2017).

[46] Nicola Segata et al. “Metagenomic microbial community profiling using unique clade-specific marker genes”. In: *Nature methods* 9.8 (2012), p. 811.

[47] Lucas Sinclair et al. “Microbial community composition and diversity via 16S rRNA gene amplicons: evaluating the illumina platform”. In: *PloS one* 10.2 (2015), e0116955.

[48] William J Stewart. *Introduction to the numerical solution of Markov chains*. Princeton University Press, 1994.

[49] Stephen Tu. "The dirichlet-multinomial and dirichlet-categorical models for bayesian inference". In: *Computer Science Division, UC Berkeley* (2014).

[50] Alejandra Vásquez et al. "Vaginal Lactobacillus flora of healthy Swedish women". In: *Journal of clinical microbiology* 40.8 (2002), pp. 2746–2749.

[51] Stéfan van der Walt, S Chris Colbert, and Gael Varoquaux. "The NumPy array: a structure for efficient numerical computation". In: *Computing in Science & Engineering* 13.2 (2011), pp. 22–30.

[52] Harold C Wiesenfeld et al. "Bacterial vaginosis is a strong predictor of Neisseria gonorrhoeae and Chlamydia trachomatis infection". In: *Clinical Infectious Diseases* 36.5 (2003), pp. 663–668.

[53] Ben P Willing et al. "A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes". In: *Gastroenterology* 139.6 (2010), pp. 1844–1854.

[54] Steven S Witkin, Iara Moreno Linhares, and Paulo Giraldo. "Bacterial flora of the female genital tract: function and immune regulation". In: *Best Practice & Research Clinical Obstetrics & Gynaecology* 21.3 (2007), pp. 347–354.

[55] Steven S Witkin, Iara Moreno Linhares, and Paulo Giraldo. "Bacterial flora of the female genital tract: function and immune regulation". In: *Best Practice & Research Clinical Obstetrics & Gynaecology* 21.3 (2007), pp. 347–354.

[56] Steven S Witkin, Iara Moreno Linhares, and Paulo Giraldo. "Bacterial flora of the female genital tract: function and immune regulation". In: *Best Practice & Research Clinical Obstetrics & Gynaecology* 21.3 (2007), pp. 347–354.

[57] Zenda Woodman. “Can one size fit all? Approach to bacterial vaginosis in sub-Saharan Africa”. In: *Annals of clinical microbiology and antimicrobials* 15.1 (2016), p. 16.