Untangling the dynamics of persistence and colonization in microbial communities

Sylvia L. Ranjeva[1,5], Joseph R. Mihaljevic*[1,4,5], Maxwell B. Joseph[2], Anna R. Giuliano[3], and Greg Dwyer[1]

[1]Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637

[2]Earth Lab, University of Colorado, Boulder, CO 80303

[3]Center for Immunization and Infection in Cancer Research (CIIRC), Moffitt Cancer Center & Research Institute, Tampa, FL 33612

[4]Current Address: School of Informatics, Computing, and Cyber Systems, Northern Arizona University, Flagstaff, AZ 86011

[5]These authors contributed equally to this work.

Running title: Community interactions through time

*Corresponding Author: Joseph R. Mihaljevic; School of Informatics, Computing, and Cyber Systems, 1295 S. Knoles Drive, Flagstaff, AZ 86011; joseph.mihaljevic@nau.edu; 928-523-5125

Abstract

A central goal of community ecology is to infer biotic interactions from observed distributions of co-occurring species. Evidence for biotic interactions, however, can be obscured by shared environmental requirements, posing a challenge for statistical inference. Here we introduce a dynamic statistical model that quantifies the effects of spatial and temporal covariance in longitudinal co-occurrence data. We separate the fixed pairwise effects of species occurrences on persistence and colonization rates, a potential signal of direct interactions, from latent pairwise correlations in occurrence, a potential signal of shared environmental responses. We apply our modeling approach to a pressing epidemiological question by examining how human papillomavirus (HPV) types coexist. Our results suggest that while HPV types respond similarly to common host traits, direct interactions are sparse and weak, so that HPV type diversity depends largely on shared environmental drivers. Our modeling approach is widely applicable to microbial communities and provides valuable insights that should lead to more directed hypothesis testing and mechanistic modeling.

Introduction

A fundamental goal of community ecology is to understand how interactions between species in a shared environment shape observed patterns of diversity over time. A key challenge in understanding community turnover is to disentangle effects of environmental drivers of species co-occurrence from inter-species interactions, especially when the goal is to infer these mechanisms from observational data [1, 2]. This challenge is also found in epidemiology, in which a major goal is to understand the factors that allow pathogens to coexist [3]. As is the case with free-living species, when determinants of environmental niches are shared among pathogen types, inferring interactions is difficult [4]. Understanding the mechanisms of microbial community turnover thus presents an ecological, statistical, and computational challenge, especially considering the

size of microbial and pathogen data sets [5, 6]. Ecological models of community turnover that account for shared environmental drivers are thus important for understanding mechanisms that underlie pathogen diversity.

For macroscopic organisms, null model analysis has historically been used to infer potential species interactions from observational data sets, through the identification of statistically non-random aggregations of species across multiple habitats [7, 8, 1, 9]. Similar approaches have been used to develop computationally efficient algorithms that make it possible to infer large correlation networks from microbial sequence data [5, 10]. Disentangling the simultaneous effects of species interactions and environmental filters from survey data is nevertheless a challenge for analyses of both macroscopic and microscopic communities [11, 2]. For example, highly mobile, competing species should transiently aggregate in habitats with shared resources, even if competitive exclusion is expected at equilibrium. Snap-shot surveys of co-occurrence can therefore lead to biased interpretations of species interactions, but time-series data can help overcome this problem.

In the microbial ecology literature, network inference models have only rarely been adapted to incorporate time-series data from multiple localities. Available methods include local similarity analysis [12, 11, 13] and generalized Lotka-Volterra modeling [14, 15]. While local similarity analysis can be used with incidence data, Lotka-Volterra modeling requires measures of abundance, which are notoriously difficult to infer from sequence data, whereas relative abundances can bias statistical analyses [16]. Local similarity analysis can infer microbial networks from observations of time-delays and temporal correlations between microbes and environmental covariates, but it relies on multiple, independent tests with p-value corrections, instead of an integrated analysis [12, 13]. Joint species distribution models provide a more comprehensive method for identifying putatively interacting species from static ecological survey data, while accounting for shared environmental drivers [17, 18, 19, 20, 21, 22]. These models use logistic regression to estimate how environmental covariates affect species occupancy probabilities across a heterogeneous landscape. Species interactions are then inferred from residual correlations between

species occurrences. While joint-species models can generate hypotheses about static community assemblages, most methods fail to capture important drivers of co-occurrence that emerge from dynamic properties of the community dynamics [2]. For example, species co-occurrence may be positively correlated across heterogeneous habitats, because of shared resources, but negatively correlated across time, because of negative species interactions within sites (i.e. Simpson's paradox, fig. 1).

Here we extend the joint-species modeling framework to infer more complex, biologically realistic dynamics in a way that is computationally tractable for large microbial data sets. We develop a statistical model of a dynamic, multi-species metacommunity in which species are affected by each other's persistence and colonization probabilities, and by shared environmental drivers. This approach can be readily applied to pathogenic microbe populations, in which distinct pathogen types represent species coexisting within a heterogeneous landscape of host organisms. In our method, we model correlations in species occupancy across habitats and across time, resolving Simpson's paradox and accounting for latent environmental covariates. We also estimate pairwise species effects on rates of colonization and persistence. Using synthetic data, we demonstrate the ability of our model to accurately and precisely infer dynamics consistent with Simpson's paradox, even with sparse occurrences. We then apply our model to data on human papillomavirus (HPV), a pathogen of significant public health concern.

Human papillomavirus (HPV) is the most common sexually transmitted infection and a major cause of cervical, genital, and oropharyngeal cancers, and it consists of over 200 types [23]. Uncertainty about the mechanisms underlying HPV type coexistence, and particularly about potential HPV type interactions, reflects a crucial unknown. Four HPV types cause most disease symptoms [24, 23, 25] and quadrivalent vaccination has demonstrated high efficacy in reducing rates of cervical dysplasia and genital warts [26, 27]. A recent 9-valent HPV vaccine targets additional oncogenic types [28]. Because the HPV vaccine is multivalent, it is possible that type replacement will occur, in which non-vaccine types increase in frequency due to population-level removal of vaccine-targeted types [29]. Type replacement following vaccination depends on inter-

actions between HPV types during natural infection, and particularly on inter-type competition through cross-immunity [30]. Understanding the ecological mechanisms that underlie HPV type diversity could therefore inform strategies for disease management and prevention. It has thus far been difficult to distinguish HPV type interactions from the effects of shared host-specific risk factors. Our dynamical community model allows us to investigate how type interactions and risk factors together structure the HPV viral community.

In this study we address two questions, which differ in their scope. First, we use our full model to ask which interactions between specific HPV types warrant future investigation? Second, we ask a more ecological question: what are the dominant drivers of community composition across space and time? To address this second question, we build models of increasing complexity, and we use model selection to determine whether HPV community patterns are determined by putative interactions between HPV types, by host-level factors that determine HPV distributions, or both. Our full model identified several interactions that warrant further experimental investigation, including negative pairwise effects on persistence and colonization probabilities. In addition, there is a strong signal of shared environmental drivers among HPV types, highlighting the importance of host-specific risk factors in supporting coexistence. By comparing models of varying complexity, however, we show that the dynamics of the HPV community are most parsimoniously explained by shared environmental drivers, rather than by strong pairwise interactions between HPV types. Pairwise species interactions thus do not appear to drive community-wide patterns of co-occurrence in the HPV community. Our study demonstrates the ability of our joint-species models to quickly and efficiently infer properties of a large, real-world viral community, and the model could therefore be of broad usefulness in understanding microbial communities.

Materials and Methods

HPV natural history

HPV types are classified based on the L1 viral capsid protein. A distinct HPV type is a variant whose L1 gene sequence is at least 10% dissimilar from any other HPV type [31]. The transmission and coexistence of individual HPV types depend on traits and risk factors of individual hosts [32, 33, 34, 35]. These include determinants of sexual behavior, including frequency of condom use, number of new and steady sexual partners, and sexual orientation; demography, including race and ethnicity; and non-sexual behavior, including smoking and alcohol consumption.

Interactions between HPV types could determine HPV diversity, though conclusive evidence of HPV type interactions is lacking [36, 37, 30].As in any species, HPV type interactions may be synergistic, neutral, or competitive. Synergism occurs when one type facilitates infection by another, while competition occurs when one type prevents infection by another. Under competitive interactions, removal of one HPV type should lead to an increase in prevalence of the competing type in the host population, resulting in type replacement. Natural history surveys reporting elevated odds ratios for multiple to single infections with HPV have suggested that cross-immunity among HPV types is unlikely [38, 39, 40, 41]. Additionally, the genetic stability of HPV as a double-stranded DNA virus has been used to support arguments against the possibility of type replacement [42], on the grounds that rapid emergence of antigenic variants is unlikely [27]. Nevertheless, a recent increase in prevalence of non-vaccine types was found in young women following vaccination and in the United States [36], suggesting that type replacement may be occurring. Indeed, several models of HPV type interactions indicate that competition between HPV types is plausible under observed patterns of coinfections [43, 30] and have demonstrated the possibility of type-replacement after vaccination [43, 30, 44, 45].

Data

We fit models of HPV type dynamics to data from the HPV Infection in Men (HIM) study [32, 33, 46], a multinational cohort study of HPV infection in men with no prior diagnosis of genital cancer or other sexually transmitted infections. The HIM study enrolled over 4000 men between 2005 and 2009 from three cities: Tampa, Florida, USA; Cuernavaca, Mexico; and Sao Paulo, Brazil. Detailed study methods are described elsewhere [32]. Briefly, the HIM study tracked PCR-confirmed infections with 37 types of HPV in men over a mean of 5 years of follow-up, recording behavioral and demographic information for all participants. The data for each individual consist of binary time series describing infection status with respect to each type over a median of 10 clinic visits, at median intervals of 6.0 months (variance = 0.7 months).

For the present analysis, we included the 3656 participants with no reported diagnosis of HIV and PCR samples for each HPV type at all clinic visits (see Appendix). We limited our analysis to ten of the HPV types available in the HIM dataset: the nine HPV types included in the most recent HPV vaccine [28]) and HPV84, a type that has shown high prevalence in several studies among men [23, 47]. Of the ten types analyzed, seven oncogenic or high-risk types - HPV16, HPV18, HPV31, HPV33, HPV45, HPV52, and HPV58 - have a demonstrated connection to cervical cancer, while three nononcogenic or low-risk types - HPV6, HPV11, and HPV84 - have been implicated in benign anogenital lesions [48]. Overall, our study includes 30,525 data points: one point per patient per virus type per visit.

Statistical Model

Our goal is to extend current joint-species modeling techniques to biological processes that may be needed to understand community dynamics. Currently, only a limited number of joint-species modeling techniques are available for longitudinal survey data. Sebastian-Gonzalez et al. [20] extended the joint-species modeling framework to allow for multiple community surveys through time by modeling the fixed, pairwise effects of species co-occurrence between subsequent time

points. Dorazio [49] introduced a model that separately estimated rates of species colonization and persistence from sequential community surveys. Although this latter model specifies the processes of extinction and colonization that can explain occupancy dynamics over time, it does not account for the residual dependence among species that can result from species interactions. Here we describe a statistical model that is tailored to the repeated surveys of patients in the HIM dataset, thereby combining the methods of Sebastian-Gonzalez et al. [20] and Dorazio [49] in a computationally tractable way.

Our data consist of observations made in I patients, who can harbor up to J HPV types (in our case limited to 10 types), sampled over a maximum of T sequential visits to the clinic. Observations of the HPV dataset are therefore aggregated as binary presence/absence data in the $I \times J \times T$ incidence array $\mathbf{Y}$, such that $Y_{i,j,t}$ indicates the presence or absence of HPV type j in patient i at visit t. Importantly, however, this model generalizes to metacommunities sampled repeatedly through time. Specifically, the model structure is the same as considering a metacommunity made up of I discrete habitats or sites, which harbor up to J species from the regional species pool, and that are surveyed over a maximum of T time points.

We fit a multivariate probit regression model to the binary presence/absence data in $\mathbf{Y}$, which has been used in other joint-species modeling approaches [21]. Probit regression relates a linear predictor to occupancy probabilities using a standard normal cumulative distribution function. In this model, the probability that a binary random variable is equal to one (i.e. $P(Y = 1)$) is equal to the probability that the latent variable z is greater than zero. The linear predictor μ completely determines the latent variable z and can be a function of one or more covariates and their effects. As part of the probit definition, the residual variance of z is equal to one. In general then, we are interested in understanding how linear predictors influence the probability that an HPV type occurs in a given patient. A generalized probit model with a single covariate x is

formulated for the ith sample as:

$$\begin{aligned} Y &\in \{0,1\}, \\ P(Y_i = 1) &= P(z_i > 0), \\ z_i &\sim N(\mu_i, 1), \\ \mu_i &= \beta x_i. \end{aligned} \tag{1}$$

Our model extends the generalized probit model by assuming that occurrence probabilities are affected by both patient-level effects and potential interactions between HPV types. We therefore build upon the general case of the probit model (Eq. 1) to model observations of the dynamic HPV metacommunity. To account for temporal dynamics, we assume that the linear predictor $\mu_{i,j,t}$ for each observation depends on observation-specific probabilities of persistence and colonization:

$$\mu_{i,j,t} = \alpha_j + \mathbf{Y}_{i,1:J,t-1}\mathbf{B}_j^{(\phi)'}(Y_{i,j,t-1}) + \mathbf{Y}_{i,1:J,t-1}\mathbf{B}_j^{(\gamma)'}(1 - Y_{i,j,t-1}) + \epsilon_{patient_{i,j}} + \epsilon_{visit_{i,j,t}} \tag{2}$$

Here, α_j is an adjustment to account for among-type variation in commonness. The presence of a given HPV type can affect the probability of persistence or colonization of other types, with a one time-step lag. If HPV type j was present in patient i on the previous clinic visit $(t-1)$, then persistence effects are represented by the product $\mathbf{Y}_{i,1:J,t-1}\mathbf{B}_j^{(\phi)'}$, where $\mathbf{Y}_{i,1:J,t-1}$ is a row vector of length J containing the presence/absence states of strains $j = 1, ..., J$ in patient i on the previous visit $(t-1)$, and $\mathbf{B}_j^{(\phi)'}$ is a column vector of length J containing pairwise interaction coefficients. These coefficients thus specify how HPV type composition at the previous visit affects persistence (ϕ) of type j. If type j was absent in patient i on visit $t-1$, colonization effects are represented by the product $\mathbf{Y}_{i,1:J,t-1}\mathbf{B}_j^{(\gamma)'}(1 - Y_{i,j,t-1})$, where $\mathbf{B}_j^{(\gamma)'}(1 - Y_{i,j,t-1})$ is a column vector of length J, again containing pairwise interaction coefficients. These coefficients thus specify how HPV type composition at the previous visit affects the colonization (γ) of type j. Both interaction matrices $(\mathbf{B}^{(\phi)}$ and $\mathbf{B}^{(\gamma)})$ are $J \times J$ dimensional, and $\mathbf{B}_j^{(\phi)}$ and $\mathbf{B}_j^{(\gamma)}$ represent the row vectors acquired by extracting row j.

Lastly, patient-level and visit-level adjustments are specified as $\epsilon_{patient_{i,j}}$ and $\epsilon_{visit_{i,j,t}}$, respectively. The multivariate patient-level random effect $\epsilon_{patient}$ allows pairwise correlations in HPV type occurrence across patients, thereby describing pairwise similarities in environmental requirements. In the case of the HIM data, $\epsilon_{patient}$ therefore controls for shared determinants of host risk, such as host behavioral covariates, that could confound estimates of HPV type interactions. The random visit-level effect ϵ_{visit} allows for pairwise correlations in HPV type occurrence across clinic visits that are not explained by the fixed temporal effects. $\epsilon_{patient}$ and ϵ_{visit} allow for residual pairwise correlations in co-occurrence that are not explained by the fixed, pairwise effects. Following the definition of the multivariate probit density, $\epsilon_{patient}$ and ϵ_{visit} are nested effects, such that the same $\epsilon_{patient}$ is added to to all of that patient's visits, such that the variances of $\epsilon_{patient}$ and ϵ_{visit} must sum to one (i.e. $z \sim N(\mu, 1)$). These random effects are therefore structured as follows:

$$\begin{aligned} \epsilon_{patient} &\sim N(0, \Sigma_{patient}) \\ \epsilon_{visit} &\sim N(0, \Sigma_{visit}) \\ \sigma^2_{patient_j} + \sigma^2_{visit_j} &= 1 \end{aligned} \tag{3}$$

where $\Sigma_{patient}$ and Σ_{visit} are $J \times J$ variance-covariance matrices, constrained so that the jth variance parameters from the two matrices sum to one, for $j = 1, ...J$. Therefore, $\rho_{patient_{p,q}}$ represents the pairwise correlation between HPV types that is measured among patients, which is derived from the variance-covariance matrix $\Sigma_{patient}$. Then, $\rho_{visit_{p,q}}$ represents the pairwise correlation between HPV types that is measured between visits and within patients (i.e. longitudinally), which is derived from the variance-covariance matrix Σ_{visit}.

We also model fixed effects of the time between visits (TBV) on persistence and colonization, to allow for the variability in when patients visited the clinic. The median TBV was 6.0 months with variance = 0.7 months, which we centered and scaled for use in the model. We allowed for fixed effects of TBV on the HPV type-specific probability of persisting ($\beta_j^{(\text{TBV},\phi)}$) and the probability of colonizing ($\beta_j^{(\text{TBV},\gamma)}$). We hypothesized that the probability that an HPV type colonizes a patient increases with TBV, due to a longer period of risk, while the probability that

a HPV type persists in the patient decreases with TBV, due to a longer time in which clearance may occur. The structure of these fixed effects is:

$$\mathbf{Z}_{i,j,t-1}\beta_j^{(\mathrm{TBV},\phi)}(Y_{i,j,t-1}) + \mathbf{Z}_{i,j,t-1}\beta_j^{(\mathrm{TBV},\gamma)}(1 - Y_{i,j,t-1}) \tag{4}$$

In this formula, $\mathbf{Z}$ is an $I \times T$ matrix that holds the centered and scaled values of TBV for each patient. This formula is added to μ_{ijt}.

Model inference

We coded our Bayesian model in *Stan* [50], an efficient, generalizable, statistical programming language, which employs adaptive Hamiltonian Monte Carlo (HMC) for model inference. We used vague priors for all parameters, although as mentioned earlier, we constrained the patient- and visit-level standard deviations to sum to one, to conform to the definition of the multivariate probit. We also included priors on the HPV type-specific, baseline probabilities of occurrence, α_j, that allowed us to assume that all types are rare across patients and clinic visits. Indeed, the most common type, HPV84, was still only present in 8.3% of all observations.

Testing the model with synthetic data

Using synthetic data, we tested the ability of our model to: (1) infer dynamics consistent with Simpson's Paradox, meaning opposite correlations in among-patient effects versus among-visit effects, (2) infer dynamics given observations of rare species, reflective of the HIM data, and (3) infer weak inter-species interactions, as are likely in nature. We generated a synthetic data set roughly half the size of the HIM data set to demonstrate the ability of our model to correctly estimate model parameters from a sparser data set. We therefore simulated data for a community of ten hypothetical pathogen strains sampled in 1500 patients, in which each patient was sampled 10 times. We assumed low but variable baseline probabilities of occurrence for each strain, with the baseline occurrence set to the baseline prevalence of the ten least prevalent HPV types in the

HIM dataset. We further assumed positive patient-level correlations and negative observation-level correlations, such that correlations were equal across pathogen strain pairs ($\rho_{patient_{p,q}}$ = 0.5, $\rho_{visit_{p,q}}$ = -0.1). Pairwise effects on persistence and colonization $\beta^{\phi}_{p,q}$ and $\beta^{\gamma}_{p,q}$ were drawn from normal distributions. All of our code for generating the synthetic data, as well as the data set itself, is available in our open-source repository `https://bitbucket.org/jrmihalj/hpv_jsdm`.

Fitting the model to the HIM data

Our first goal was to use our model to identify any interactions between HPV types that might warrant future epidemiological investigations. We therefore fit our full model and quantified the posterior distributions of the pairwise effects of HPV types on colonization and persistence rates. Our second goal was to understand the relative contributions of environmental effects, such as host-specific risk factors, and pairwise inter-type interactions to HPV community dynamics. We therefore fit four nested models of varying complexity. Model 1 has fixed, pairwise effects between HPV types, model 2 has residual correlations that account for environmental effects, and model 3, our full model, has both. Model 4 includes only baseline occurrence probabilities α_j, and is therefore a type of null model. All of these models include the effects of the time between visits (TBV). We then compared the models' out-of-sample predictive abilities using the leave-one-out information criterion (LOO-IC), estimated using Pareto-smoothed importance sampling in the *R* package "*loo*" [51]. Compared to the Watanabe-Akaike information criterion (WAIC), which is asymptotically equal to LOO-IC, the LOO-IC has been found to be more robust when using vague priors [52], as in our models. We considered two models to be substantially different if their LOO-IC values differed by 3, which is the common convention [53]. In practice, for a data set this large, small changes in overall goodness-of-fit could lead to very large changes in the likelihood when integrated across the many data points, and thus large differences in LOO-IC. We therefore emphasize that we use this model selection procedure as a heuristic to guide our understanding of community dynamics, rather than as a robust hypothesis test.

Results

Model validation with synthetic data

When we tested our model with synthetic data, it accurately and precisely inferred dynamics consistent with Simpson's Paradox, even when the data were sparse (Fig. 2). The model correctly inferred the low baseline probabilities of species occurrence (Fig. 2 A) and all patient-level correlations (Fig. 2 B). It also accurately estimated the majority of negative correlations at the observation level, although some inferred pairwise correlations were indistinguishable from zero (Fig. 2 C). This latter effect was not surprising, because we assumed a weak negative correlation (ρ_{visit} = -0.1). Importantly, although the model's estimates of the magnitude of simulated correlations were sometimes incorrect, the model was unbiased with respect to the direction of the simulated correlations. The model also correctly estimated persistence ($\beta^{\phi}_{p,q}$) and colonization ($\beta^{\gamma}_{p,q}$) under both strong and weak interactions (Fig. 2 D,E). Finally, the model accurately recovered the effects of the time between visits on both persistence and colonization probabilities, which we assumed were the same for all pathogen strains (Fig. S2).

Metacommunity dynamics of HPV and model comparisons

In our full model, there were only a few interactions between HPV types that were worthy of future investigation, including several weakly negative effects on colonization probability (Fig. 3). Importantly, including these fixed effects and the random effects of patient-level and observation-level correlations led to a substantial improvement relative to a null model that accounts only for type-specific baseline occurrence probabilities, suggesting that the biology added to our model helps explain HPV community composition relative to the null model (Table 1). Based on LOO-IC selection, however, the most parsimonious model included only the random effects of patient-level and observation-level correlations, without pairwise interactions between the HPV types (Table 1). Pairwise inter-type interactions can thus be identified by our model, but the effect of

these interactions is not strong enough to substantially mediate the overall community composition in this subset of 10 HPV types. The best model, which did not include these pairwise interactions, gives qualitatively similar insights for the random effects, meaning the patient-level and observation-level correlations, as our full model (Fig. S4).

The best model captured important qualitative aspects of the HPV dynamics, as well. The inferred baseline occurrence probability recovered the observed rank order of prevalence of the ten HPV types (Fig. 3A). The model confirmed that increasing values of TBV had positive effects on colonization probabilities ($\beta_j^{(TBV,\gamma)} > 0$) for all HPV types, but it had negative effects on persistence probabilities ($\beta_j^{(TBV,\phi)} < 0$) for all but two HPV types (Figs. S3, S4).

Patient-level correlations were positive for all but one pair of HPV types (Fig. 3B). These positive correlations suggest that there are shared environmental drivers across human hosts, in the form of risk factors. In the case of HPV52 and HPV58 (Fig. 3C), there are both positive patient-level and negative observation-level correlations. Positive observation-level correlations, or correlations within individuals over time, likely signal affinity for co-transmission, because in the models these effects are in addition to the pairwise effects on persistence and colonization. Negative observation-level correlations thus signal reduced affinity for co-transmission. However, the negative observation-level correlations between HPV52 and HPV58 must be interpreted with caution, as they could reflect the masking of HPV58 detection by HPV52, a problem that has been documented in the linear array genotyping test used in the HIM study [42].

Discussion

Our results suggest that HPV type coexistence is strongly driven by shared environmental characteristics. While the full model is able to estimate even sparse and weak (putative) interactions between HPV types, our model selection procedure suggests that these interactions are not important for explaining overall patterns of community turnover in HPV. The influence of patient-level correlations on HPV community dynamics suggests that HPV types segregate among hosts

with shared traits. It is therefore likely that human subpopulations exist that could promote HPV type coexistence across space and time. This finding is consistent with epidemiological evidence of type-specific differences in the risk factors that promote HPV transmission [54, 55], and with another recent modeling study that characterized subtle differences in the profile of host-specific risk factors that affect infection with each type [56].

Model selection shows that pairwise inter-type interactions that affect colonization and persistence probabilities do not influence overall patterns of community turnover in this HPV data set. However, the full model identified several putative interactions worthy of future epidemiological investigations. In particular, it is possible that interactions could mediate the occurrence patterns of specific pairs of HPV types, even though model selection suggests that pairwise interaction effects have no meaningful effects on the HPV community dynamics as a whole. In other words, the community-level patterns could swamp out the patterns of specific HPV pairs. Further, by limiting our analysis to a subset of ten HPV types, it is possible that we by chance did not include HPV types that have larger effects on the community. Also, our model only estimates pairwise effects, and future studies could account for higher order interactions, which have been shown to be important in diverse competitive networks [57].

The results of our analysis complement the results of a previous, mechanistic model of HPV dynamics fitted to 6 HPV types of the HIM dataset [56]. The authors of this previous work formulated an epidemiological model that allowed for homologous immunity, a form of within-species competition, as well as the effects of 11 host-specific covariates. The best-fit version of this model included no homologous immunity for any of the six HPV types (HPV84, HPV62, HPV89, HPV16, HPV51, and HPV6), finding instead that previous infection with any type significantly increases the risk of re-infection with the same type. In our statistical model, this effect is further confirmed by the positive baseline persistence probabilities ($\beta^{\phi}_{p,p}$) across all ten HPV types analyzed. That study [56] also detected no pairwise interaction between two taxonomically similar types, HPV16 and HPV31, which had been hypothesized to compete through cross-immunity [58, 59]. Furthermore, the risk of initial infection with any HPV type was concentrated among

high-risk subpopulations, which were linked to host-specific covariates. Taken together, the results of this previous analysis [56] suggest that both intra-specific and inter-specific competition are weak or absent in the HPV viral community, such that stabilizing competitive mechanisms cannot explain HPV diversity. Instead, diversity may depend on sustained infection within high-risk subpopulations specific to each HPV type. These findings are consistent with our finding that inter-type interactions have little effect on HPV community dynamics (Table 1). Furthermore, by showing how host-specific traits define niches that are used by different HPV types, the previous work [56] supports the importance of shared among-patient traits to explain patterns of co-occurrence.

While the different quantitative approaches between the previous study [56] and our study provide complementary results, there are important differences in the methods, applications, and conclusions. Ranjeva et al. [56] tested mechanistic biological models about type-specific HPV dynamics, whereas our approach allowed for the identification of statistical patterns in the community dynamics of multiple types. Also, our method can be generalized to any metacommunity that is sampled through time, rather than being specific to a pathogen community that interacts via cross-immunity, as modeled by Ranjeva et al. [56]. Indeed, our statistical framework is agnostic to the specific mechanisms of interactions. Instead our model specifies latent mechanisms that affect probabilities of persistence and colonization, which are estimated from the occurrence data.

We have shown that a relatively simple statistical model can be used to infer community dynamics, even in a system with rare species occurrences. Sparsity of observational data in real-world metacommunities generally limits the power of statistical models to correctly infer ecological effects [49, 60, 61]. We showed that our model can be used to infer opposing environmental and temporal dynamics from communities of rare species, and to detect weak interactions among rare species, which are the most common types of interactions in nature [62]. Inferring residual correlations with rare species requires a substantial amount of data, but, in the age of affordable, high-throughput sequencing technologies, such data can often be obtained easily. Moreover, our

model accounts for the effects of unobserved environmental drivers, specifically host-specific risk-factors in the case of the HPV data, without having to specify covariates explicitly. This may be useful for analyzing large microbial communities, such as microbiome communities, in which the environmental drivers are unknown.

In classical joint-species distribution models, residual correlations in species occurrence are used to infer species interactions, but such residual correlations can arise instead from shared covariate responses that are not explicitly included in the model structure [21, 2]. Our model, however, does not rely on residual correlations to infer interspecies interactions *per se*. We use species occupancy at the previous time step to estimate lagged, pairwise effects of species' occurrences on the probabilities of persistence and colonization of cohabitating species. Residual correlations in our models instead account for latent environmental covariates, such as unmeasured host-specific traits. Although our statistical modeling approach can thus identify important signatures of species interactions, mechanistic models and experimentation are nevertheless required to rigorously test hypotheses about species interactions. Furthermore, we estimate interspecies effects on persistence and colonization using a one-timestep lag, which requires that the timescale of the species interactions be equal to the timescale of observations. This assumption may not always hold. Our method is therefore best used to refine testable hypotheses from observed dynamics of large community assemblages, such as microbiome assemblages, in a computationally-feasible manner, rather than as a final step in inferring interactions.

A final caveat is that our models do not allow for dynamics that occur between observations. Given two consecutive observations of a species, our models instead assume that there is either persistence over the entire interval, or that at most one extinction or colonization has occurred. This assumption may result in bias in communities that are poorly sampled relative to the timescale of the dynamics. Indeed, recent evidence shows that standard joint-species distribution modeling approaches cannot accurately capture simulated predator-prey dynamics, especially if habitats are relatively homogeneous, probably because of non-linear dynamics [2]. This problem is likely to be important for non-linear host-pathogen dynamics as well, and should be a subject

of future simulation efforts. Our dataset however spans a wide diversity of patients, and includes the effects of the time between visits, which should limit this type of bias.

Competing Interests

The authors declare no competing financial interests

Literature Cited

[1] Gotelli NJ, Graves GR. Null Models in Ecology. Washington, D.C.: Smithsonian Institution Press; 1996.

[2] Zurell D, Pollock LJ, Thuiller W. Do joint species distribution models reliably detect interspecific interactions from co-occurrence data in homogenous environments? Ecography. InPress;.

[3] Keeling MJ, Pejman R. Formulating the Deterministic SIR Model. In: Modeling Infectious Diseases in Humans and Animals. 2nd ed. Princeton, NJ: Princeton University Press; 2008. p. 15–52. Available from: https://books.google.com/books?hl=en&lr=&id=G8enmS23c6YC&oi=fnd&pg=PP2&dq=force+of+infection+keeling+and+rohani&ots=rFKSDv9toM&sig=hHVhDvYUpjrd4lkCgLe3bkH8m_4#v=onepage&q=forceofinfectionkeelingandrohani&f=false.

[4] Godsoe W, Franklin J, Blanchet FG. Effects of biotic interactions on modeled species' distribution can be masked by environmental gradients. Ecology and evolution. 2017;7(2):654–664.

[5] Faust K, Raes J. Microbial interactions: From networks to models. Nature Reviews Microbiology. 2012;10(8):538–550. Available from: http://dx.doi.org/10.1038/nrmicro2832.

[6] Seabloom EW, Borer ET, Gross K, Kendig AE, Lacroix C, Mitchell CE, et al. The community ecology of pathogens: coinfection, coexistence and community composition. Ecology Letters. 2015;18:401–415. Available from: http://doi.wiley.com/10.1111/ele.12418.

[7] Diamond JM. Assembly of species communities. Ecology and evolution of communities. 1975;p. 342–444.

[8] Connor EF, Simberloff D. The Assembly of Species Communities: Chance or Competition? Ecology. 1979;60(6):1132. Available from: http://www.jstor.org/stable/1936961?origin=crossref.

[9] Gotelli NJ, McCabe DJ. Species co-occurrence: A meta-analysis of JM Diamond's assembly rules model. Ecology. 2002;83(8):2091–2096.

[10] Fisher CK, Mehta P. Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. PLoS ONE. 2014;9(7):1–10.

[11] Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. ISME Journal. 2016;10(7):1669–1681. Available from: http://dx.doi.org/10.1038/ismej.2015.235.

[12] Ruan Q, Dutta D, Schwalbach MS, Steele JA, Fuhrman JA, Sun F. Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. Bioinformatics. 2006;22(20):2532–2538.

[13] Xia LC, Ai D, Cram JA, Liang X, Fuhrman JA, Sun F. Statistical significance approximation in local trend analysis of high-throughput time-series data using the theory of Markov chains. BMC Bioinformatics. 2015;16(1):1–14. Available from: http://dx.doi.org/10.1186/s12859-015-0732-8.

[14] Stein RR, Bucci V, Toussaint NC, Buffie CG, Rätsch G, Pamer EG, et al. Ecological Modeling from Time-Series Inference: Insight into Dynamics and Stability of Intestinal Microbiota. PLoS Computational Biology. 2013;9(12):31–36.

[15] Carrara F, Giometto A, Seymour M, Rinaldo A, Altermatt F. Inferring species interactions in ecological communities: A comparison of methods at different levels of complexity. Methods in Ecology and Evolution. 2015;6(8):895–906.

[16] Cardona C, Weisenhorn P, Henry C, Gilbert JA. Network-based metabolic analysis and microbial community modeling. Current Opinion in Microbiology. 2016;31:124–131. Available from: http://dx.doi.org/10.1016/j.mib.2016.03.008.

[17] Ovaskainen O, Hottola J, Siitonen J. Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. Ecology. 2010;91(9):2514–2521. Available from: http://www.esajournals.org/doi/abs/10.1890/10-0173.1.

[18] Ovaskainen O, Tikhonov G, Norberg A, Blanchet FG, Duan L, Dunson D, et al. How to make more out of community data? A conceptual framework and its implementation as models and software. Ecology Letters. 2017;20:561–576.

[19] Ovaskainen O, Abrego N, Halme P, Dunson D. Using latent variable models to identify large networks of species-to-species associations at different spatial scales. Methods in ecology and evolution. 2015;.

[20] Sebastián-González E, Sánchez-Zapata JA, Botella F, Ovaskainen O. Testing the heterospecific attraction hypothesis with time-series data on species co-occurrence. Proceedings of the Royal Society B: Biological Sciences. 2010;277:2983–2990.

[21] Pollock LJ, Tingley R, Morris WK, Golding N, O'Hara RB, Parris KM, et al. Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). Methods in Ecology and Evolution. 2014 may;5(5):397–406. Available from: http://doi.wiley.com/10.1111/2041-210X.12180.

[22] Joseph MB, Preston DL, Johnson PTJ. Integrating occupancy models and structural equation models to understand species occurrence. Ecology. 2016;97(3):765–775.

[23] Ma Y, Madupu R, Karaoz U, Nossa CW, Yang L, Yooseph S, et al. Human papillomavirus community in healthy persons, defined by metagenomics analysis of human microbiome project shotgun sequencing data sets. Journal of Virology. 2014 5;88(9):4786–97. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24522917.

[24] Frazer IH. Interaction of human papillomaviruses with the host immune system: A well evolved relationship. Virology. 2009;384(2):410–414. Available from: http://dx.doi.org/10.1016/j.virol.2008.10.004.

[25] Todd RW, Roberts S, Mann CH, Luesley DM, Gallimore PH, Steele JC. Human papillomavirus (HPV) type 16-specific CD8+ T cell responses in women with high grade vulvar intraepithelial neoplasia. International Journal of Cancer. 2004;108(6):857–862.

[26] Dunne EF, Unger ER, Mcquillan G, Swan DC, Patel SS, Markowitz LE. Prevalence of HPV Infection. 2014;297(8):813–819.

[27] Markowitz LE, Hariri S, Lin C, Dunne EF, Steinau M, McQuillan G, et al. Reduction in human papillomavirus (HPV) prevalence among young women following HPV vaccine introduction in the United States, National Health and Nutrition Examination Surveys, 2003-2010. The Journal of infectious diseases. 2013 8;208(3):385–93. Available from: http://jid.oxfordjournals.org/cgi/content/long/jit192v1.

[28] Joura EA, Giuliano AR, Iversen OE, Bouchard C, Mao C, Mehlsen J, et al. A 9-Valent HPV Vaccine against Infection and Intraepithelial Neoplasia in Women. New England Journal of Medicine. 2015 2;372(8):711–723. Available from: http://www.nejm.org/doi/10.1056/NEJMoa1405044.

[29] Bernard E, Pons-Salort M, Favre M, Heard I, Delarocque-Astagneau E, Guillemot D, et al. Comparing human papillomavirus prevalences in women with normal cytology or invasive cervical cancer to rank genotypes according to their oncogenic potential: a meta-analysis of observational studies. BMC infectious diseases. 2013;13(1):373. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23941096%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3751808%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3751808&tool=pmcentrez&rendertype=abstract.

[30] Durham DP, Poolman EM, Ibuka Y, Townsend JP, Galvani AP. Reevaluation of epidemiological data demonstrates that it is consistent with cross-immunity among human papillomavirus types. Journal of Infectious Diseases. 2012;206(8):1291–1298.

[31] Bernard HU, Burk RD, Chen Z, van Doorslaer K, Hausen Hz, de Villiers EM. Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. Virology. 2010 5;401(1):70–79. Available from: https://www.sciencedirect.com/science/article/pii/S0042682210001005.

[32] Giuliano AR, Lazcano-Ponce E, Villa LL, Flores R, Salmeron J, Lee JH, et al. The human papillomavirus infection in men study: human papillomavirus prevalence and type distribution among men residing in Brazil, Mexico, and the United States. Cancer Epidemiology, Biomarkers & Prevention. 2008 8;17(8):2036–43. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3471778&tool=pmcentrez&rendertype=abstract.

[33] Giuliano AR, Lazcano E, Villa LL, Flores R, Salmeron J, Lee JH, et al. Circumcision and sexual behavior: factors independently associated with human papillomavirus detection among men in the HIM study. International Journal of Cancer. 2009 3;124(6):1251–7. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3466048&tool=pmcentrez&rendertype=abstract.

[34] Nyitray A, Carvalho R, Baggio M, Beibei L, Abrahamsen M, Papenfuss M, et al. Age-Specific Prevalence of and Risk Factors for Anal Human Papillomavirus (HPV) among Men Who Have Sex with Women and Men Who Have Sex with Men : The HPV in Men (HIM) Study. Journal of Infectious Diseases. 2011;203(1):49–57.

[35] Nyitray AG, Carvalho Da Silva RJ, Baggio ML, Smith D, Abrahamsen M, Papenfuss M, et al. Six-month incidence, persistence, and factors associated with persistence of anal human papillomavirus in men: The HPV in men study. Journal of Infectious Diseases. 2011;204(11):1711–1722.

[36] Kahn Ja, Brown DR, Ding L, Widdice LE, Shew ML, Glynn S, et al. Vaccine-type human papillomavirus and evidence of herd protection after vaccine introduction. Pediatrics. 2012

8;130(2):249–56. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3408690&tool=pmcentrez&rendertype=abstract.

[37] Wheeler CM, Castellsagué X, Garland SM, Szarewski A, Paavonen J, Naud P, et al. Cross-protective efficacy of HPV-16/18 AS04-adjuvanted vaccine against cervical infection and precancer caused by non-vaccine oncogenic HPV types: 4-year end-of-study analysis of the randomised, double-blind PATRICIA trial. Lancet Oncology. 2012;13(1):100–110.

[38] Rousseau MC, Pereira JS, Prado JC, Villa LL, Rohan TE, Franco EL. Cervical coinfection with human papillomavirus (HPV) types as a predictor of acquisition and persistence of HPV infection. Journal of infectious diseases. 2001 12;184(12):1508–17. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11740725.

[39] Liaw KL. A prospective study of human papillomavirus (HPV) type 16 DNA detection by polymerase chain reaction and its association with acquisition and persistence of other HPV types. Journal of Infectious Diseases. 2001;183(1):8–15. Available from: http://dx.doi.org/10.1086/317638.

[40] Chaturvedi AK. Prevalence and Clustering Patterns of Human Papillomavirus Genotypes in Multiple Infections. Cancer Epidemiology, Biomarkers & Prevention. 2005;14(10):2439–2445. Available from: http://cebp.aacrjournals.org/cgi/doi/10.1158/1055-9965.EPI-05-0465.

[41] Chaturvedi AK, Engels EA, Pfeiffer RM, Hernandez BY, Xiao W, Kim E, et al. Human Papillomavirus and Rising Oropharyngeal Cancer Incidence in the United States. Journal of Clinical Oncology. 2011 11;29(32):4294–4301. Available from: http://ascopubs.org/doi/10.1200/JCO.2011.36.4596.

[42] Tota JE, Ramanakumar AV, Jiang M, Dillner J, Walter SD, Kaufman JS, et al. Epidemiologic approaches to evaluating the potential for human papillomavirus type replacement

postvaccination. American journal of epidemiology. 2013 8;178(4):625–34. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23660798.

[43] Elbasha EH, Galvani AP. Vaccination against multiple HPV types. Mathematical Biosciences. 2005;197(1):88–117.

[44] Murall CL, McCann KS, Bauch CT. Revising ecological assumptions about Human papillomavirus interactions and type replacement. Journal of Theoretical Biology. 2014;350:98–109. Available from: http://dx.doi.org/10.1016/j.jtbi.2013.12.028.

[45] Poolman EM, Elbasha EH, Galvani AP. Vaccination and the evolutionary ecology of human papillomavirus. Vaccine. 2008;26(Supplement 3):25–30.

[46] Giuliano AR, Lee JH, Fulp W, Villa LL, Lazcano E, Papenfuss MR, et al. Incidence and clearance of genital human papillomavirus infection in men (HIM): a cohort study. Lancet. 2011 3;377(9769):932–40. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3231998&tool=pmcentrez&rendertype=abstract.

[47] Han JJ, Beltran TH, Song JW, Klaric J, Choi YS. Prevalence of Genital Human Papillomavirus Infection and Human Papillomavirus Vaccination Rates Among US Adult Men. JAMA Oncology. 2017;3(6):810–816. Available from: http://oncology.jamanetwork.com/article.aspx?doi=10.1001/jamaoncol.2016.6192.

[48] Giuliano AR, Nyitray AG, Kreimer AR, Pierce Campbell CM, Goodman MT, Sudenga SL, et al. EUROGIN 2014 roadmap: Differences in human papillomavirus infection natural history, transmission and human papillomavirus-related cancer incidence by gender and anatomic site of infection. International Journal of Cancer. 2015;136(12):2752–2760.

[49] Dorazio RM, Kéry M, Royle JA, Plattner M. Models for inference in dynamic metacommunity systems. Ecology. 2010 8;91(8):2466–75. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20836468.

[50] Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan : A Probabilistic Programming Language. Journal of Statistical Software. 2017;76(1). Available from: http://www.jstatsoft.org/v76/i01/.

[51] Vehtari A, Gelman A, Gabry J. loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models; 2016. R package version 1.1.0. Available from: https://CRAN.R-project.org/package=loo.

[52] Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing. 2016;27(5):1–20.

[53] Gelman A, Carlin J, Stern H, Dunson D. Bayesian data analysis. 2nd ed. Chapman and Hall; 2013. Available from: http://books.google.com/books?hl=en{&}lr={&}id=ZXL6AQAAQBAJ{&}oi=fnd{&}pg=PP1{&}dq=Bayesian+data+analysis{&}ots=uNYiu1bGW5{&}sig=eZoeglTTl{_}60ozzh4Hrf5YQDvwE.

[54] Albero G, Castellsagué X, Lin HY, Fulp W, Villa LL, Lazcano-Ponce E, et al. Male circumcision and the incidence and clearance of genital human papillomavirus (HPV) infection in men: the HPV Infection in men (HIM) cohort study. BMC Infectious Diseases. 2014;14(1):75. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-84893644201&partnerID=tZOtx3y1.

[55] Nyitray AG, Chang M, Villa LL, Carvalho RJ, Baggio ML. us c t Ac ce p te d Ac ce p te d us. 2015;p. 1–26.

[56] Ranjeva SL, Baskerville EB, Dukic V, Villa LL, Lazcano-Ponce E, Giuliano AR, et al. Recurring infection with ecologically distinct HPV types can explain high prevalence and diversity. Proceedings of the National Academy of Sciences of the United States of America. 2017 12;114(51):13573–13578. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29208707http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5754802.

[57] Levine JM, Bascompte J, Adler PB, Allesina S. Beyond pairwise mechanisms of species coexistence in complex communities. Nature. 2017 5;546(7656):56–64. Available from: http://www.nature.com/doifinder/10.1038/nature22898.

[58] Draper E, Bissett S, Howell-Jones R, Edwards D. Neutralization of non-vaccine human papillomavirus pseudoviruses from the A7 and A9 species groups by bivalent HPV vaccine sera. Vaccine. 2011;29(47):8585–8590. Available from: http://www.sciencedirect.com/science/article/pii/S0264410X11014277.

[59] Kemp TJ, Hildesheim A, Safaeian M, Dauner JG, Pan Y, Porras C, et al. HPV16/18 L1 VLP vaccine induces cross-neutralizing antibodies that may mediate cross-protection. Vaccine. 2011;29(11):2011–2014. Available from: http://www.sciencedirect.com/science/article/pii/S0264410X1100017X.

[60] Mihaljevic JR, Joseph MB, Johnson PT. Using multispecies occupancy models to improve the characterization and understanding of metacommunity structure. Ecology. 2015;96(7):1783–1792.

[61] Warton DI, Blanchet FG, O'Hara RB, Ovaskainen O, Taskinen S, Walker SC, et al.. So Many Variables: Joint Modeling in Community Ecology. Elsevier Ltd; 2015. Available from: http://dx.doi.org/10.1016/j.tree.2015.09.007.

[62] Chesson P. Mechanisms of Maintenance of Species Diversity. Annual Review of Ecology and Systematics. 2000 11;31(1):343–366. Available from: http://www.annualreviews.org/doi/10.1146/annurev.ecolsys.31.1.343.

[63] Gelman A. Inference and monitoring convergence. In: Gilks W, Richardson S, Spiegelhalter D, editors. Markov Chain Monte Carlo in Practice. CRC Press; 1996. p. 131–143.

Figure Legends

Figure 1: Simpson's paradox demonstrated for two species that are sampled across ten habitat sites, with each site surveyed fifteen times. **A** Species covary positively across sites (over space), indicating response to similar habitat requirements. **B** Species covary negatively within sites over time, indicating inter-specific competition. Probabilities of occurrence are on the probit scale.

Figure 2: Inference of model parameters from synthetic data simulated for 10 pathogen strains across 1500 patients, where each patient was tested 10 times. **A** Recovery of baseline occurrence probability for the 10 strains. Red vertical line gives the true value. **B** Recovery of positive, pairwise correlations in among-patient random effects. Dashed line represents zero effect, while dotted line represents the true value (0.5). **C** Recovery of weakly negative, pairwise correlations in within-patient, observation-level random effects. Dashed line represents zero effect, while dotted line represents the true value (-0.11). **D** Recovery of fixed inter-strain effects on probability of strain persistence. **E** Recovery of fixed inter-strain effects on probability of strain colonization.

Figure 3: Inference of model parameters from the HIM data. **A** Estimate of the baseline occurrence probability for each HPV type. **B** Inferred correlations in among-patient random effects. **C** Inferred correlations in within-patient, observation-level random effects. **D** Recovery of fixed inter-type effects on the probability of type persistence. **E** Recovery of fixed inter-type effects on probability of type colonization.

Tables

Table 1: Comparison of candidate models using leave-one-out cross-validation. The table shows whether fixed and/or random effects were included, the log-likelihood of the model fit (i.e. $\mathcal{L}(\theta|D)$), and the LOO-IC. The standard error (SE) in the LOO-IC is shown to emphasize that the LOO-IC is an estimated statistic with error, but also that none of our LOO-IC values overlap within $\pm$ 2SE.

HPV Interactions	Among-patient and Among-visit Correlations	Log-Likelihood	LOO-IC	SE LOO-IC
	✓	-220310.1	708323.6	373.8
✓	✓	-279574.9	825439.5	329.8
✓		-433036.9	1109717.0	465.3
		-432977.8	1130039.0	681.3

Figure 1:

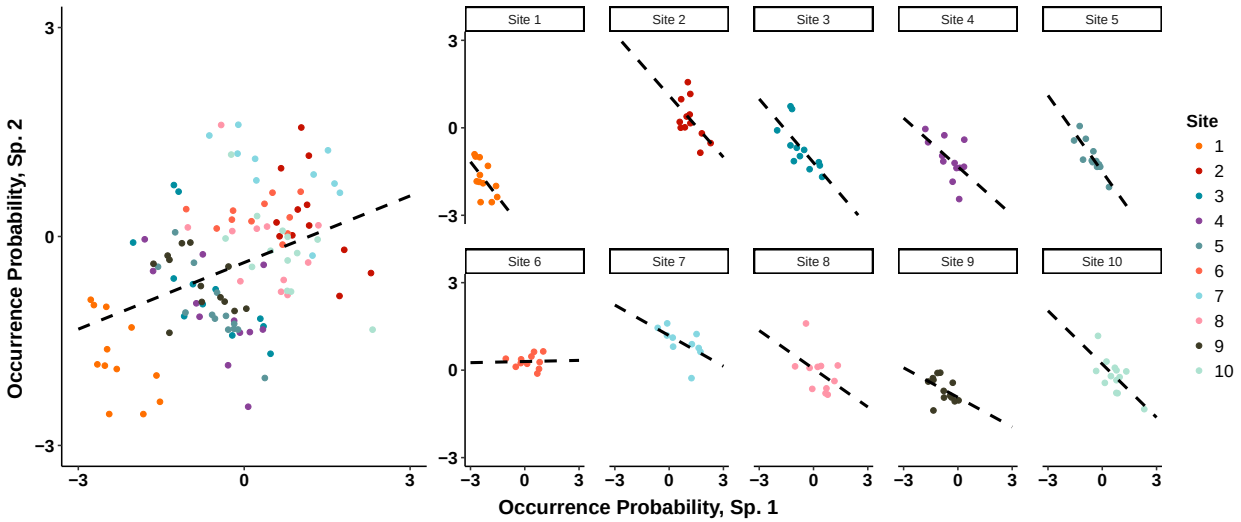

Occurrence Probability, Sp. 2
Occurrence Probability, Sp. 1
Site 1
Site 2
Site 3
Site 4
Site 5
Site 6
Site 7
Site 8
Site 9
Site 10
Site
1
2
3
4
5
6
7
8
9
10

Figure 2:

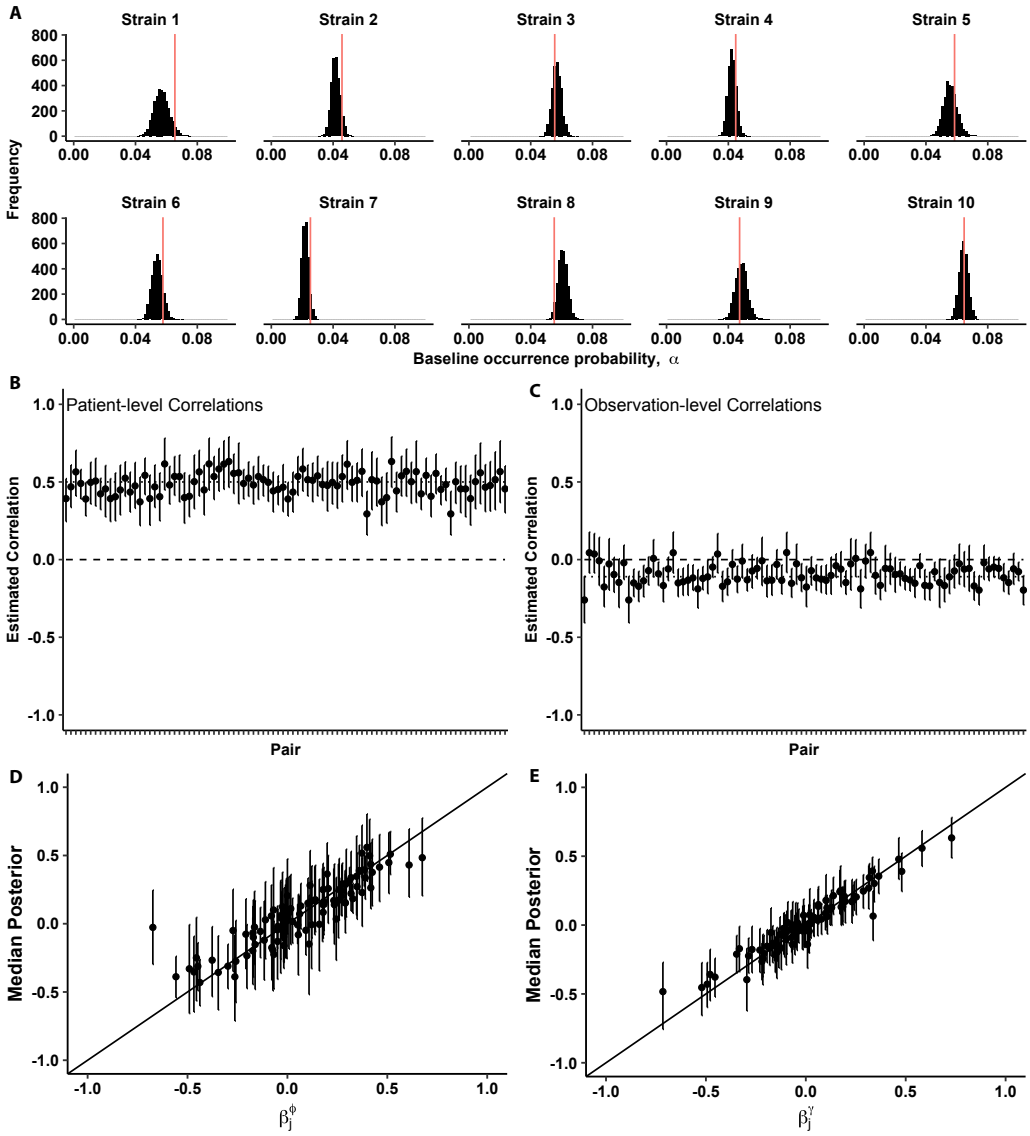
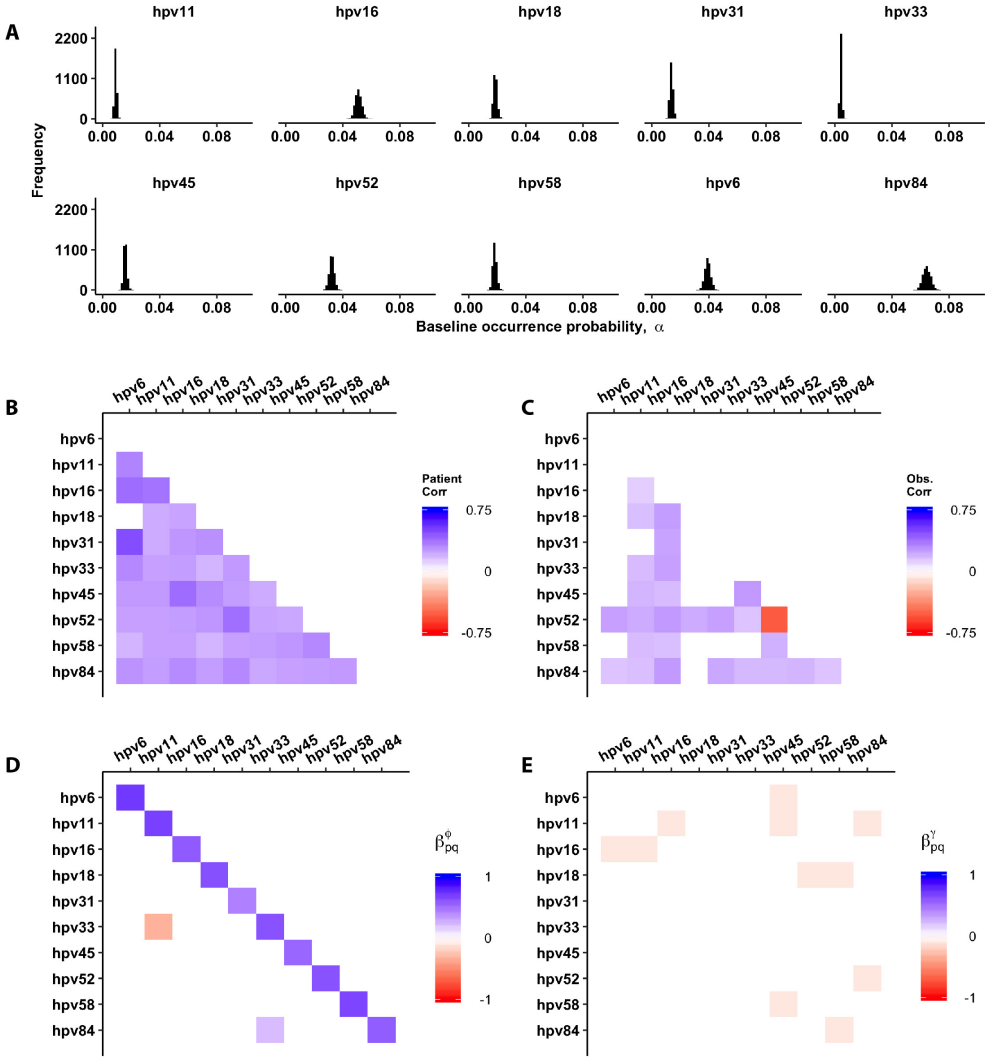

A
Strain 1
Strain 2
Strain 3
Strain 4
Strain 5
Strain 6
Strain 7
Strain 8
Strain 9
Strain 10
Frequency
Baseline occurrence probability, α
B
Patient-level Correlations
Estimated Correlation
Pair
C
Observation-level Correlations
Estimated Correlation
Pair
D
Median Posterior
β_j^{ϕ}
E
Median Posterior
β_j^{γ}

Figure 3:


A
hpv11
hpv16
hpv18
hpv31
hpv33
hpv45
hpv52
hpv58
hpv6
hpv84
Frequency
2200
1100
0
0.00
0.04
0.08
Baseline occurrence probability, α
B
Patient Corr
0.75
0
-0.75
C
Obs. Corr
0.75
0
-0.75
D
β^{ϕ}_{pq}
1
0
-1
E
β^{γ}_{pq}
1
0
-1

Supporting Information (SI)

Subset of HIM data included in the analysis

We excluded individuals that failed to meet the full eligibility criteria described by the HIM study [32]. The criteria included: ages 18 to 70 years; residents of one of three sites — Sao Paulo, Brazil; Morelos, Mexico; or southern Florida, United States; no prior diagnosis of penile or anal cancers; no prior history of genital or anal warts; no symptoms of a sexually transmitted infection at baseline or recent treatment for a sexually transmitted infection; no history of participation in an HPV vaccine study; and no history of HIV or AIDS.

We identified 3,656 eligible participants from the 4,123 men enrolled in the HIM study as of October 2014. For each of the 10 HPV types that we analyzed, we include in our data the binary infection status of each man at each clinic visit. We also include the length of time between consecutive clinic visits.

Type-specific HPV prevalence over follow-up

We calculated the prevalence of the 10 HPV types included in the analysis at each visit (Fig. S1). Note that, because individuals varied in their visit dates, the prevalence at each visit is a time-averaged estimate. The data show that the expected distribution of HPV types in the metacommunity is consistent across visits.

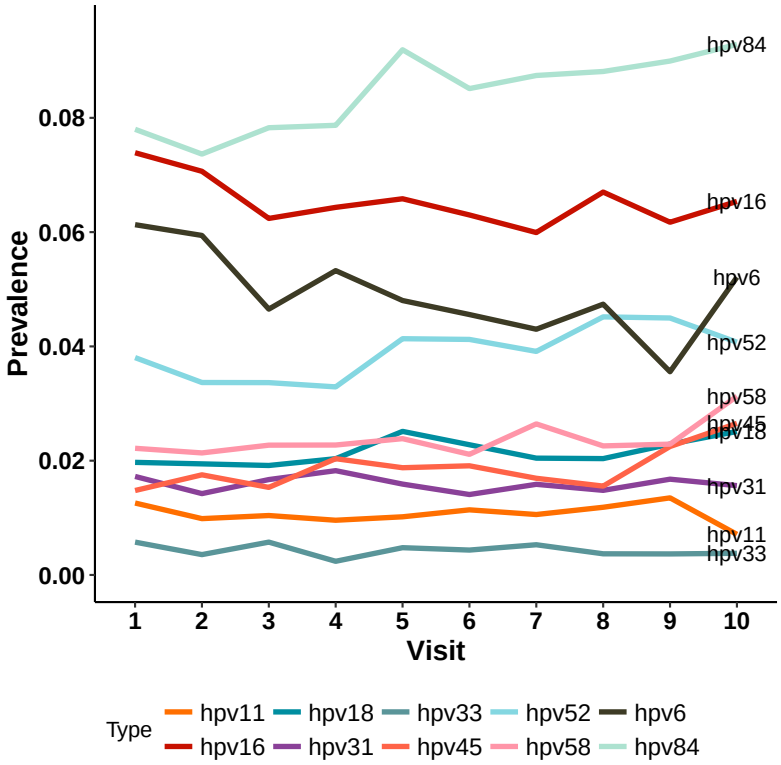


Figure S1: Observed visit-level prevalence of each of the 10 HPV types included in this analysis.

Stan model details

All of our code to run the Stan model is provided in our open-source repository **LINK**, but we will briefly describe the fitting routine here. For each nested model, we ran three MCMC chains in parallel on the Gardner high performance computing (HPC) cluster at the University of Chicago (Center for Research Informatics). Each chain ran for 5000 iterations with a 2000 iteration warm-up period, and we thinned the samples by three, giving us a total of 1000 posterior samples from each chain. Parameter samples were stored as tables in a SQLite database for later processing. Due to the large number of columns of the log-likelihood table, we split this table into sub-components before storage. We monitored convergence with the Gelman-Rubin ($\hat{R}$) statistic, and we conducted several standard visual diagnostics to check MCMC chain performance [63, 53]. All models converged after 5000 iterations, and no problems were observed in the MCMC chains.

Time between visits

Here we display the effects of time between visit (TBV) on persistence and colonization probabilities for the synthetic data (Fig. S2) and for the HIM dataset, using the full model that includes both correlations and fixed, pairwise interactions (Fig. S3).

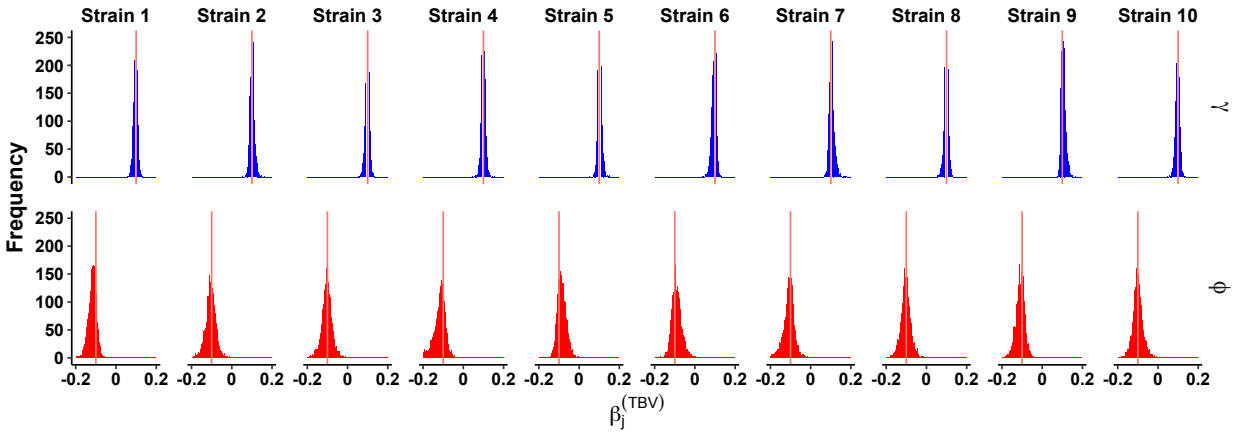


Figure S2: Effects of time between visit (TBV) on colonization (top row) and persistence (bottom row) probabilities for each of 10 simulated pathogen strains, from the synthetic data. These results are generated from the full model, which has both fixed effects of pairwise interactions, as well as patient-level and observation-level correlations among residuals. Blue histograms are effects greater than zero, while red histograms are effects less than zero, based on a 95% credibal intervals (CI) that does not overlap zero. The true, simulated values are shown as red vertical lines

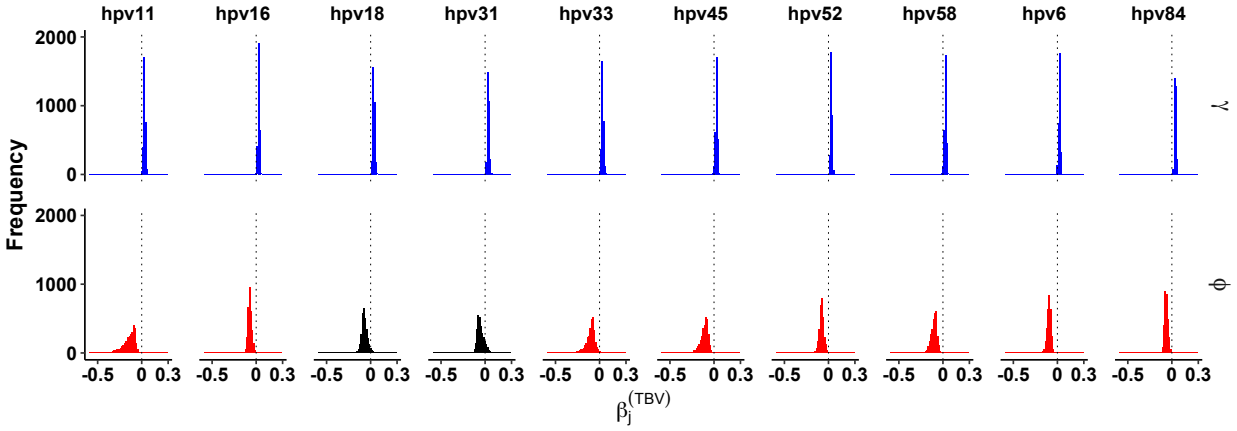


Figure S3: Effects of time between visit (TBV) on colonization (top row) and persistence (bottom row) probabilities for each HPV type. These results are generated from the full model, which has both fixed effects of pairwise interactions, as well as patient-level and observation-level correlations among residuals. Blue histograms are effects greater than zero, while red histograms are effects less than zero (marked as the vertical dotted line), based on a 95% credibal intervals (CI) that does not overlap zero. Histograms with black bars have effects with 95% credibal intervals (CI) that overlap zero.

Results from "best" model, with no pairwise interaction effects

The figure below displays the results from the most preferred model, which includes the random effects (i.e. patient-level and observation-level correlations among HPV types), but does not include pairwise effects on persistence and colonization probabilities (Fig. S4). Notably, this model is nearly identical to the full model in terms of baseline probabilities of occurrence (Fig. S4 A), the random effects (Fig. S4 B,C), and the effects of time between visit (TBV) (Fig. S4 D).

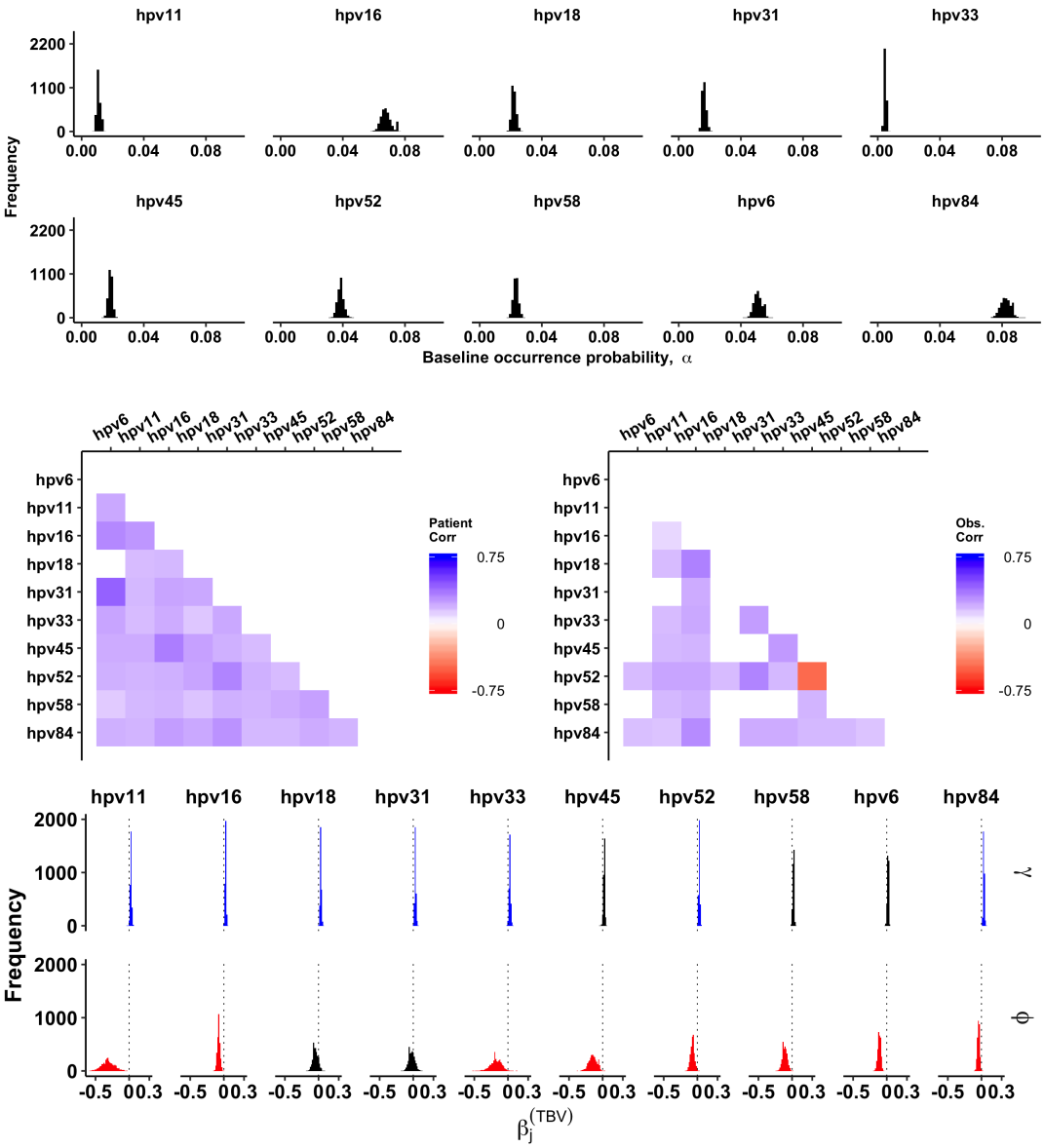


Figure S4: Inference of model parameters from the HIM data, using the "best" model, which only has patient-level and observation-level correlations among types, but does not have fixed effects on persistence and colonization. **A** Estimate of the baseline occurrence probability for each HPV type. **B** Inferred correlations in among-patient random effects. **C** Inferred correlations in within-patient, observation-level random effects. **D** Estimates of the effects of time between visit (TBV) on persistence and colonization probabilities. Colors and vertical lines in **D** are the same as in Fig. S3.