Supporting Information: Recurring infection with ecologically distinct HPV types can explain high prevalence and diversity

Ranjeva et al.

Type	Model (n parameters)	Log likelihood	SE	ΔAICc
HPV62	Memoryless (13)	-2026.1	1.4	389.6
	Homologous immunity (15)	-2027.4	1.5	396.3
	Additional risk (17)	-1827.2	0.4	0
HPV16	Memoryless (13)	-1663.2	1.1	117.9
	Homologous immunity (15)	-1663.7	0.7	123.1
	Additional risk (17)	-1600.1	0.1	0
HPV89	Memoryless (13)	-1902.3	1.3	137.4
	Homologous immunity (15)	-1902.2	1.5	141.2
	Additional risk (17)	-1829.5	0.1	0
HPV51	Memoryless (13)	-1670.8	0.6	63.9
	Homologous immunity (15)	-1670.4	1.7	67.3
	Additional risk (17)	-1634.7	0.7	0
HPV84	Memoryless (13)	-1979.6	1.7	270.3
	Homologous immunity (15)	-1980.5	1.1	276.3
	Additional risk (17)	-1840.3	0.4	0
HPV6	Memoryless (13)	-1461.0	1.0	121.7
	Homologous immunity (15)	-1461.5	1.3	126.8
	Additional risk (17)	-1396.1	0.6	0

Table S1. Comparison of Candidate Models. The ΔAICc gives the AICc score of each candidate model relative to the best-fit model, such that the best-fit model has ΔAICc = 0.

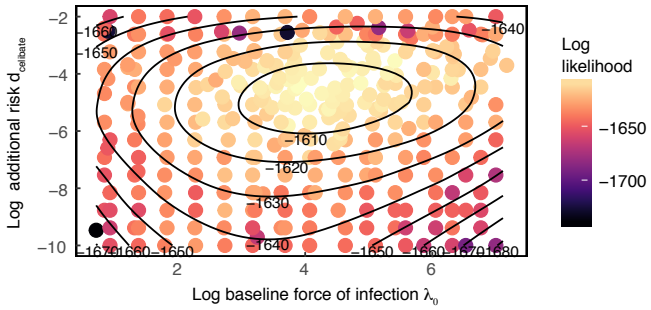

Fig. S1. Bivariate likelihood profile for the additional risk d_{celibate} and the baseline infection risk λ_0 for HPV16.

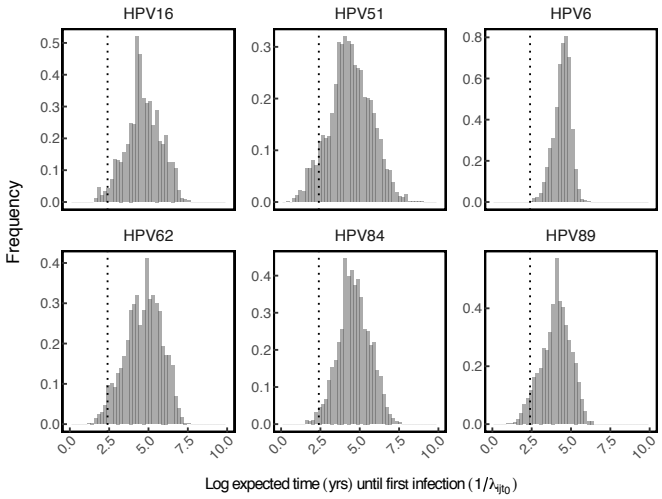

Fig. S2. Distribution of the (log) baseline expected time to infection ($t_{\text{expected}} = 1/\lambda_{ij}$) in uninfected individuals assuming no prior infection, such that $\lambda_{ij} = \lambda_{0_j} f(\overrightarrow{\alpha_j} \boldsymbol{C_i 0})$. For each individual, λ_{ij} was calculated using the maximum likelihood estimate for each element in $\overrightarrow{\alpha}$ and the individual-specific covariates $\boldsymbol{C_i 0}$, which were reported at the baseline visit ($t = 0$). The y axis reports frequency, while the vertical dashed line in each panel marks an expected time to infection of 10 years. Thus, only the portion of the distribution to the left of the dashed line in each panel represents individuals for whom the the expected infection time falls within the next ten years.

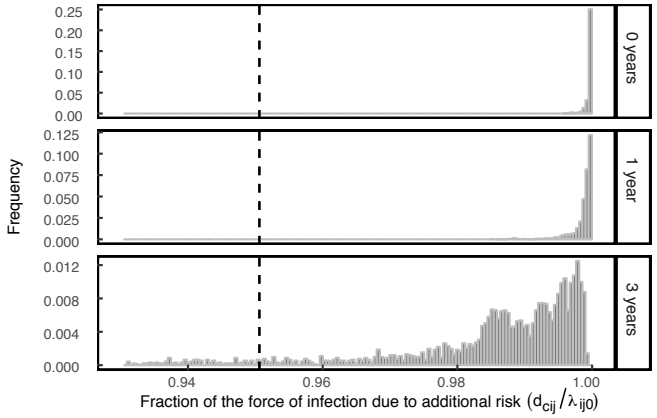

Fig. S3. Contribution of the effect of previous infection to overall infection risk in individuals with respect to HPV16. Histograms show the distribution of the fraction of the overall force of infection λ_{ijt} made up by the additional risk d in the population at various times post-clearance of the precedent infection. The mean value for the population-level distribution at each time point is given by the vertical dotted line.

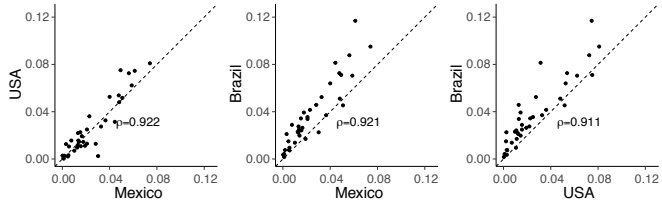

Fig. S4. Comparison of the time-averaged prevalence of HPV types in the HIM Dataset. The black dotted lines indicates 1:1 correlation, and ρ denotes Pearson's correlation coefficient ($p < 10^{-15}$ with 35 degrees of freedom for each comparison).

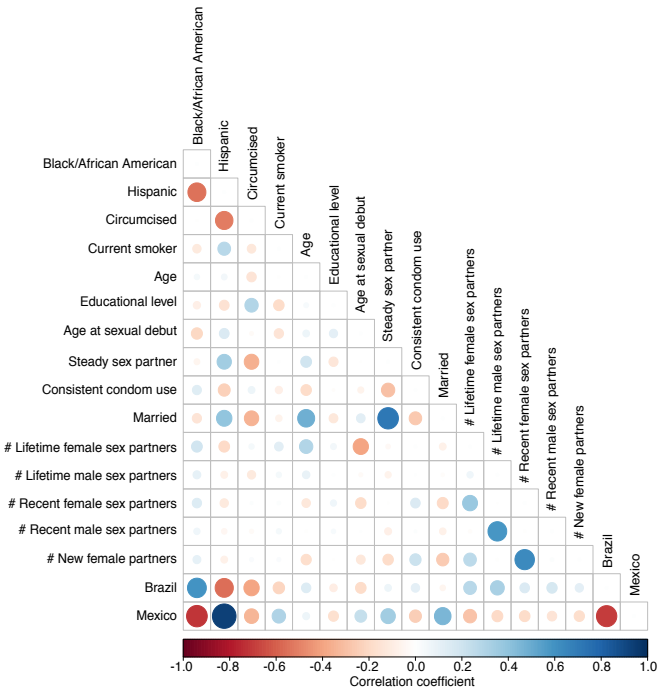

Fig. S5. Correlations between the full candidate set of self-reported covariates at the baseline visit. Pearson product-moment correlations were calculated between continuous variables, polyserial correlations (inferred latent correlations between continuous and categorical variables) were calculated between numeric and binary variables, and polychoric correlations (inferred latent correlations between categorical variables) were calculated between binary variables. All correlations shown were significant at the $\alpha = .05$ significance level based on tests for bivariate normality.

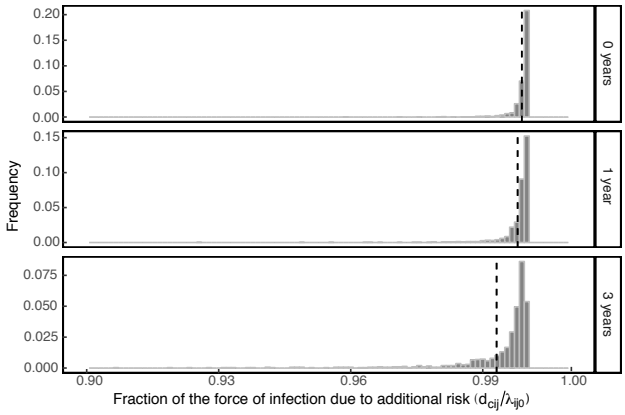

Fig. S6. Contribution of the effect of previous infection to the overall infection risk among individuals for HPV16, using a model with 17 covariates. Histograms show the distribution of the fraction of the overall force of infection λ_{ijt} made up by the additional risk d in the population at various times post-clearance of the precedent infection. The mean value for each distribution is given by the vertical dotted line.

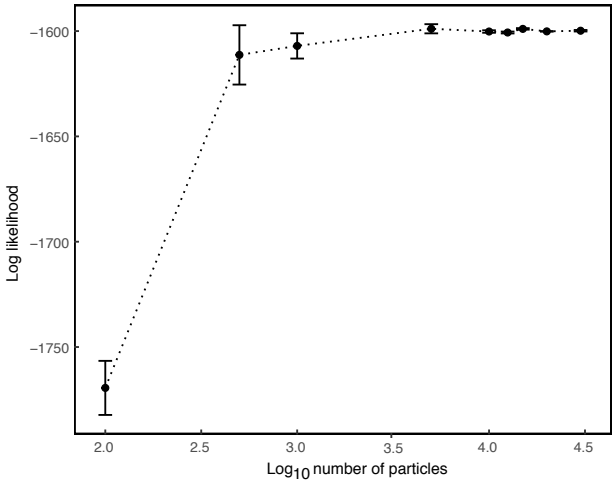

Fig. S7. Likelihood of the maximum likelihood parameter set for HPV16 calculated at increasing particle sizes. Point estimates and error bars represent the mean and standard error, respectively, of 10 particle filter replicates.

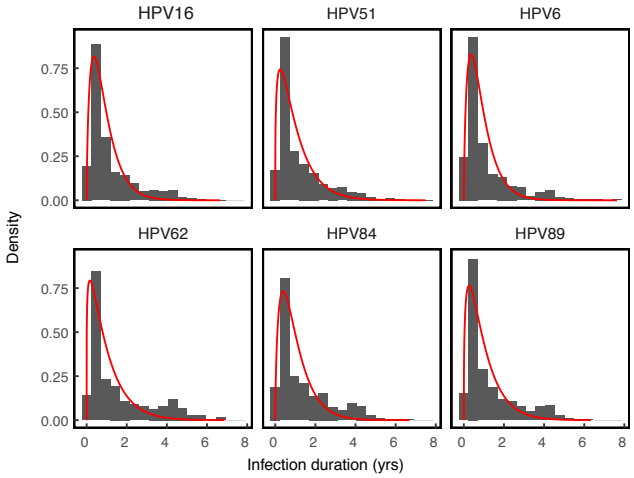

Fig. S8. Empirical distribution of infection durations for each HPV type (grey histogram) with an overlying gamma distribution having the same mean and coefficient of variation as in the data (red).

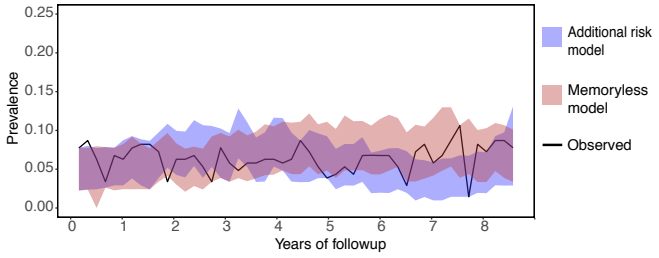

Fig. S9. Comparison of the dynamics under candidate models with the observed HPV16 infection data. Fraction of infected individuals at various time-points during followup. The shaded areas are bounded by the 2.5% and 97.5% quantiles from 1,000 simulations of each model. The solid line gives the observed fraction infected among the individuals used to simulate the data.

1. Details of the HPV in Men data and covariate variables

1.1. Participants included in the analysis. We excluded individuals that failed to meet the full eligibility criteria described by the HIM study (1), which included no prior diagnosis of genital cancer, warts, HIV, or other STIs. We identified 3,656 eligible participants from the 4,123 men enrolled in the HIM study as of October 2014. Of the eligible participants, we excluded $n = 575$ participants that had missing data at the time of enrollment for the 11 covariates that we included in our three

candidate models. We then divided the remaining 3,081 participants into three sexual subclasses based on the number of recent sex partners. For all covariates, "recent" activity indicates activity in the past six months, or since the last clinic visit if the last clinic visit was more than six months prior, as reported at each clinic visit. The three sexual subclasses were celibate individuals, individuals with one recent sex partner, and individuals with two or more recent sex partners. To account for the effects of sexual subclass on the infection risk, we restricted our analysis to include only the $n = 1,099$ individuals who remained in the same subclass for at least three years over the course of their participation in the study.

1.2. Covariate variables. Covariates, including those used to determine sexual subclass were derived from a risk factor questionnaire that was administered at each visit, which was described and validated previously (1, 2). The survey was conducted via computer-assisted self-interviewing (CASI) to preserve patient anonymity during the response process. Questions covered sociodemographic characteristics, sexual behaviors, sex partnerships, and condom use. Participants were asked to recall their recent behavior, such as recent number of male or female sexual partners, where "recent" referred to behavior over the past six months. Participants had the option of refusing to answer any question, and refusals were treated as missing values as in Giuliano et al. (1), such that a missing covariate was assigned its value at the closest visit. Covariates were selected for inclusion in the model based on known risk factors identified in the literature for HPV in men (1, 3–7). We also included country of residence as a covariate. The full set of candidate covariates (Fig. S5) included race (black/African American or other), ethnicity (Hispanic or other), age, age at sexual debut, lifetime numbers of male and female sexual partners, numbers of recent male and female sexual partners, numbers of new male and female sexual partners, presence of a steady sexual partner, marital status, level of education, circumcision status (confirmed by a clinician at the baseline visit), whether or not the participant was a current smoker, whether or not the participant used condoms for the majority of recent sexual encounters, Brazilian nationality, and Mexican nationality, with US nationality used as the baseline.

To decrease statistical non-identifiability and to increase computational feasibility, we reduced the full candidate set of covariates to a subset based on observed correlations (Fig. S5). Among highly correlated pairs of similar covariates, we heuristically selected one representative. For example, we chose to include recent numbers of sexual partners instead of lifetime numbers of sexual partners and the presence of a steady sexual partner instead of marital status. The variables describing race and ethnicity were strongly correlated with country of origin, so we excluded them. Table S2 gives the final subset of covariates included in the model.

Through this selection process, we identified five continuous covariates that we included as linear predictors (Table S2). The continuous and ordinal covariates that we modeled as linear predictors were age, educational level, age at sexual debut, number of recent female sex partners, and number of recent male sex partners. Following previous analyses from the HIM study group (3, 5), we specified these covariates according to a previous study (2) that evaluated the test-retest reliability of the risk factor questionnaire across the three study sites and languages. Nyitray et al. identified the covariate specifications that maximized the calculated reliability coefficient of the response between geographical sites. For recent sexual partnerships, for example, they found that across-site and across-time reliability was maximized when recent sexual partnerships were grouped as 0, 1, 2, and ≥ 3 sexual partners of either gender over the previous six months. We preserved this variable grouping and found that less than 4% of the individuals in our data reported >3 recent sexual partnerships at any visit. Binning of ≥ 3 sexual partnerships thus likely did not affect the estimated covariate effect. For educational level, the HIM study group identified a binned ordinal set of educational levels (0-4) that were similar across cultures. Nyitray et al. found that covariate responses had substantial reliability regardless of participant age, and the previous analyses of the HIM dataset found no association of age with HPV incidence across different binnings of age. Furthermore, previous analyses identified a simple positive relationship between HPV incidence and younger age at sexual debut (3). We therefore used age and age at sexual debut as continuous predictors without binning.

To model the effect of the covariates on the instantaneous risk of infection λ_{ijt}, we used a log-linear model (main text Eq. 2), such that each covariate effect α is directly interpretable as a multiplicative effect on λ_{ijt}. Time-varying covariates (age, recent numbers of male and female sex partners, the presence of a steady sex partner, smoking status, educational level, and condom use) were updated at each visit.

	Name	Structure	Notes
1	Age	Continuous	
2	Educational Level	Ordinal	Levels: <12 years (0), 12 years completed (1), 13-15 years (2), 16 years completed (3), $\geq$ 17 years (4)
3	Age at sexual debut	Continuous	
4	Recent female sexual partners	Continuous	0,1,2, $\geq$ 3 different partners within the last six months
5	Recent male sexual partners	Continuous	0,1,2, $\geq$ 3 different partners within the last six months
6	Circumcision status	Binary	
7	Smoking status	Binary	Current smoker
8	Steady sex partner	Binary	Presence of a steady sexual partner
9	Consistent condom use	Binary	Use of a condom more than 50% of the time during intercourse
10	Brazil	Binary	Country of residence
11	Mexico	Binary	Country of residence

Table S2. Covariates included in the analyses.

1.3. Identification of celibate individuals. At each visit, celibate individuals were identified as those reporting no recent receptive or insertional vaginal sex, no recent receptive or insertional anal sex, and no recent receptive or insertional oral sex. Participants were asked separately at each visit about insertional and receptive sexual practices and were asked separately about all sexual practices with male or female partners. Vaginal sex was defined explicitly as "your penis in your partner's vagina," anal sex was defined explicitly as "your penis in your partner's anus or your partner's penis in your anus," and oral sex was defined explicitly as "your penis in your partner's mouth or your partner's penis in your mouth."

1.4. Potential effects of covariate specification on the inference results. While we based our selection and specification of the covariates on previous analyses, we cannot ignore the possibility of residual bias. For example, for recent male sex partners, if any sexual activity with male partners significantly affected risk, but the distinction between two and three recent partners was negligible, then we could have underestimated the effect of male sexual partnership activity by inferring a linear coefficient. Furthermore, we inferred a linear coefficient to estimate the general effect of increasing educational level on HPV risk, but changes between discrete educational levels may correspond imperfectly to proportional changes in risk, which may have limited our power to detect a true effect. Dichotomizing condom use into more or less than 50%, a necessary measure given the small fraction of individuals that reported 100% condom use, could have introduced bias. Particularly, we may underestimate the protective effect of condoms if they are only strongly effective at 100% usage. However, at least one previous study identified a significant reduction in the risk of HPV infection with condom use when condom use was measured as use more than 50% of the time (8).

2. Likelihood-based inference

Inference by maximum likelihood was carried out using multiple iterated filtering (MIF) (9). Briefly, iterated filtering is an algorithm that uses sequential Monte Carlo (SMC) to approximate maximum likelihood estimates of parameters from POMP models. SMC uses a population of particles drawn from the parameters of a given model to generate Monte Carlo samples of the latent dynamic variables and evaluate the likelihood of observed time series (10). Iterated filtering successively filters the particle population, perturbing the parameters between iterations. The perturbations decrease in amplitude over time, allowing convergence at the maximum likelihood estimate.

The extension of SMC and iterated filtering to longitudinal panel data has been previously described (11), and we extended longitudinal POMP methods to binary data. The data for each HPV type is a set of binary time series, or panel units, describing the observed infection trajectory for each individual. The panel POMP contains a POMP model for each individual, and individuals share parameters. To evaluate the likelihood of a shared parameter set, SMC is carried out over the time series for each individual to generate a log-likelihood for the corresponding panel unit. The log likelihood of the panel POMP object is the sum of the individuals' log likelihoods. All optimization routines were carried out using 20,000 particles to overcome high Monte Carlo error (SI 6).

For each model, we initialized the iterated filtering with 100 random parameter combinations. Optimization involved series of successive MIF searches, with the output of each search serving as the initial conditions for the subsequent search. The likelihood of the output for each search was calculated by averaging the likelihood from ten passes through the particle filter, each using 50,000 particles. The optimization was repeated until additional operations did not arrive at a higher maximum likelihood (SI 6).

For model selection, we used the corrected Akaike Information Criterion (AICc) (12). We obtained maximum likelihood estimates for each parameter and associated 95% confidence intervals by constructing likelihood profiles. We used Monte Carlo Adjusted Profile methods (13) to obtain a smoothed estimate of the profile that accounts for the increased Monte Carlo error associated with longitudinal data. The lower and upper limits of the 95% confidence interval were the points that lay 1.92 log-likelihood units below the maximum likelihood estimate on the smoothed profile curve. These points correspond to one-half the 95% critical value for a χ^2 distribution with one degree of freedom.

3. Comparison of our results to previous analyses of the HIM dataset

Our estimates of the effects of type-specific risk factors are generally consistent with previous studies, but there are some exceptions. First, a previous analysis of the HIM data showed no difference in HPV16 incidence between men who have sex with men and men who have sex with women (5), but our results instead show that risk of HPV16 infection declines with an increase in the number of male sex partners, suggesting that HPV16 infection depends largely on heterosexual transmission. Meanwhile, our results show that the risk of infection with HPV51, HPV89, HPV6, and HPV84 increases with the number of male or female partners, consistent with previous work, although the effect was statistically significant only for HPV89. Finally, the previous analysis concluded that circumcision increases the risk of HPV51 (7), but our results suggest that circumcision instead reduces the risk of HPV51. Our modeling approach, however, may provide greater statistical power than previous analyses. First, our models account for variation in risk per unit time, and they describe the dynamics in a way that is unconstrained and unbiased by the frequency of observations. Second, we infer the contribution of host-specific risk factors to infection risk separately from the dynamics of previous infection. We therefore have reason to believe that our results are robust and that differences with previous analyses arise from differences in model structure.

4. Additional analyses and alternative models

4.1. Testing for homologous immunity by excluding individuals with prevalent type-specific infection. Including individuals infected with HPV at baseline could have enriched the data for individuals with less immunity to HPV, causing the additional risk model to be favored over the homologous immunity model. Therefore, to help validate the results of the model selection, we fitted each of the three candidate models (the memoryless model, main text Eq. 1, the homologous immunity model, main text Eq. 3, and the additional risk model, main text Eq. 4) to the subset of individuals that were uninfected for at least two visits at baseline. We found strong support for the additional risk model in this subpopulation (Table S3, consistent with our major findings. For each type, the homologous immunity model again reduced to the memoryless model (the maximum likelihood was achieved at $d = 1$ for all types), affirming the lack of homologous immunity in any type.

Type	Model (n parameters)	Log likelihood	SE	ΔAICc
HPV62	Memoryless (13)	-974.9	0.3	52.5
	Homologous immunity (15)	-974.2	0.6	55.3
	Additional risk (17)	-944.5	0.3	0
HPV16	Memoryless (13)	-879.9	1.1	17.8
	Homologous immunity (15)	-879.7	0.5	21.5
	Additional risk (17)	-866.9	0.4	0
HPV89	Memoryless (13)	-1081.8	0.2	22.0
	Homologous immunity (15)	-1081.4	0.9	25.3
	Additional risk (17)	-1066.7	0.4	0
HPV51	Memoryless (13)	-991.5	0.8	33.1
	Homologous immunity (15)	-992.1	0.7	38.4
	Additional risk (17)	-970.8	1.0	0
HPV84	Memoryless (13)	-1042.1	0.3	47.9
	Homologous immunity (15)	-1041.6	0.5	51.1
	Additional risk (17)	-1014.0	0.9	0
HPV6	Memoryless (13)	-797.7	0.8	30.6
	Homologous immunity (15)	-797.4	0.6	34.1
	Additional risk (17)	-778.3	0.3	0

Table S3. Comparison of the candidate models fit to the subset of the study of population without prevalent infection at baseline. The ΔAICc gives the AICc score of each candidate model relative to the best-fit model, such that the best-fit model has ΔAICc = 0.

4.2. Additional risk model fit to celibate individuals only. We fitted the additional risk model to data from the celibate individuals ($n = 237$). Celibate individuals were those who reported no receptive or insertional vaginal, anal, or oral sex between visits. For all analyses, we included only individuals that remained celibate for a minimum of three years, and we estimated the magnitude of the additional risk dcelibate only during the period of celibacy. Consistent with our main findings, when fitting the model to only celibate individuals, we recover the strong effect of previous infection with each HPV type on the risk of subsequent infection with the same type (Table S4). The effect of previous infection on the total risk of subsequent infection in the celibate population accounts on average for over 96% of the infection risk across types, even at three years after infection clearance.

Type	d_{celibate} (MLE)	95% Confidence Interval
HPV62	1.6	[0.1, 3.4]
HPV84	1.5	[0.4,9.8]
HPV89	5.8	[1.1,10.1]
HPV16	3.4	[1.3,5.4]
HPV51	3.0	[0.9,6.8]
HPV6	2.5	[0.8, 5.2]

Table S4. Maximum likelihood estimate of the additional risk d_{celibate} inferred using data from only celibate individuals. All values are reported on a log scale.

4.3. Additional risk only model. In this model, the force of infection λ_{ijt} for individual i with type j at time t was determined only by the baseline infection risk for type j and by the effect of previous infection, so that the behavioral and demographic risk factors had no effect. The force of infection is then:

$$\lambda_{ijt} = \lambda_{0_j} + \mathrm{I}_{\text{prev}} d_{jc_i} e^{-w_j(t-t_{\text{clr}})} \qquad [S1]$$

For each HPV type, this model out-performed the memoryless model but it performed much worse than the model that also took into account the covariates (Table S5).

Type	Log likelihood	SE	ΔAICc
HPV16	-1,636.5	0.6	50.6
HPV51	-1,651.5	1.8	11.7
HPV6	-1,403.6	0.3	8.3
HPV62	-1,877.0	0.4	66.8
HPV84	-1,870.0	0.1	37.3
HPV89	-1,843.4	0.8	22.5

Table S5. Performance of the additional risk only model for each HPV type relative to the full additional risk model. The ΔAICc gives the AICc score of each candidate model relative to the best-fit model, such that the best-fit model has ΔAICc = 0.

4.4. HPV16/HPV31 interaction model. To test whether infection with HPV31 affects the risk of infection with HPV16, we fitted a model in which the force of infection of HPV16 depends on whether an individual has ever been infected with HPV31. In this model, we included a covariate variable $\mathrm{I}_{\mathrm{HPV31}}$, updated at each visit, that indicated whether individual i was currently or previously infected with HPV31. The force of infection λ_{ijt} was thus:

$$\lambda_{ijt} = \lambda_{0_j} f(\overrightarrow{\alpha_j}\boldsymbol{C_{it}})\alpha_{\mathrm{HPV31}}\mathrm{I}_{\mathrm{HPV31}} + \mathrm{I}_{\mathrm{prev}} d_{jc_i} e^{-w_j(t-t_{\mathrm{clr}})}, \quad [S2]$$

where α_{HPV31} gives the effect of previous or current infection with HPV31 on the risk of infection with HPV16. Our goal was to estimate the direction and strength of interaction between HPV16 and HPV31. Our estimate of the interaction parameter was centered around 1 (MLE 1.3, CI[0.5,2.0]), indicating no significant interaction.

4.5. A model with additional covariates. To estimate the contribution of the additional risk to the force of infection when allowing for a large number of host risk factors, we added additional covariates to the additional risk model for HPV16, using the full set of 17 original candidate covariates (Fig. S5). We used the maximum likelihood estimates of the model parameters to estimate the fraction of the force of infection made up by the additional risk d among individuals. This model involved inference for n=1019 individuals that had complete covariate information at baseline for the full set of covariates. To compare this model to the best-fit additional risk model (with 11 covariates), we fitted the original additional risk model for the same n=1019 individuals and again calculated the maximum likelihood. The additional risk d still accounts on average for over 90% of the risk for several years after infection clearance (Fig. S6). The maximum likelihood and AICc for the model with additional covariates were -1468.4 (SE .63) and 2983.9, respectively, whereas the maximum log likelihood and AICc for the model without additional covariates, fitted to the data from the same individuals, were -1472.8 (SE .86) and 2979.6, respectively. Thus, the ΔAICc for the additional covariates model is 4.3, showing that the improvement in the likelihood from the additional complexity of the additional covariates does not provide a better explanation of the data than the additional risk model.

5. Model validation

To assess the ability of the best-fit model to capture the observed dynamics, we simulated infection data for the n = 1,099 individuals that we included when making inferences, preserving the visit dates and covariate data for each individual. We compared the results of 1,000 simulations from both the memoryless model and the additional risk model (the best-fit model) to the data. In the additional risk model (main text Eq. 4), the initial infection risk is low across the population, and an individual's risk of infection is sharply increased by previous infections. The prevalence of any HPV type should thus arise from repeated infections within a small number of individuals. In the memoryless model (main text Eq. 1), the initial infection risk is higher across the population, and previous infection has no effect on subsequent risk, such that the prevalence of any type arises from fewer infections across more individuals. Both the memoryless model and the additional risk model captured the observed prevalence over patient followup (Fig. S9), but the additional risk model accurately predicted the total number of unique individuals that were infected at any point during the study. The memoryless model, in contrast, overestimated the number of unique infected individuals (main text Fig. 3B). The best-fit model thus captures important qualitative aspects of the dynamics, supporting the results of model selection.

The success of the additional-risk model provides strong support for our use of a Bernoulli likelihood function. An implicit assumption of the Bernoulli likelihood function is that the error in the model's predictions is due only to measurement error rather than to systematic bias in the model's predictions. The ability of the best model to reproduce both the prevalence data, and the number of unique individuals infected, strongly suggests that any such systematic bias is reasonably weak.

6. Monte Carlo error in inference from binary panel data

Advances in simulation-based Monte Carlo methods have made it possible to fit complex models to large datasets. We took advantage of extensions of multiple iterated filtering (MIF) (9) to the case of panel data. Iterated filtering uses sequential Monte Carlo, also known as particle filtering, to estimate the likelihood of partially observed Markov process (POMP) models. Sequential Monte Carlo uses stochastic simulations of dynamical models to produce successive populations of weighted particles. Each particle represents a Monte Carlo sample from the probability density of the latent process, conditional on the parameters and the previous observations. As the particle population is propagated along the time series, the particles are weighted and resampled at each data point, and the likelihood of each observation is estimated as the weighted average of the particles.

Large data sets and complex models can result in non-negligible Monte Carlo error in estimated likelihoods. The structure of panel data, a collection of time series that are dynamically independent apart from shared model parameters, yields high Monte Carlo error that often makes it infeasible to calculate the likelihood with an error of less than one log likelihood unit (13). This is important because a standard approach to calculating 95% confidence intervals relies on the observation that parameter values with log likelihood scores that are within 1.92 units of the maximum log likelihood fall within the 95% confidence interval (14). Because high levels of Monte Carlo error can make it difficult to accurately estimate likelihoods, high Monte Carlo error rates also make it difficult to estimate 95% confidence intervals. Ionides et al. (13) show that one solution to this problem is to approximate the likelihood in the region of the maximum likelihood by fitting a quadratic to to likelihood scores from a large sample of parameter values, which can in turn be used to directly estimate the confidence bounds. Ionides et al. validated this approach with panel data (13), and here we apply it to the case of binary panel data.

Before carrying out the model fitting, we tested our approach by quantifying the effect of particle size on the likelihood for a given set of parameters. As is often the case in simulation-based approaches, the Monte Carlo error in our simulations is high enough that most particles have very low likelihoods. As a result, our likelihood estimates at first improve rapidly with increases in the number of particles (Fig. S7). For particle numbers above about 5000, however, further increases in particle numbers have at most weak effects. We therefore used 20,000 particles for the iterated filtering and 50,000 particles for evaluations of the likelihood by particle filtering. We also accounted for Monte Carlo error in maximum likelihood estimation by initiating a large number of independent MIF searches ($n = 100$) at random parameter values for any given model. Each of the 100 searches began with 200 MIF iterations and was continued successively (100 iterations) until the maximum likelihood for that particular search was stationary within one log likelihood unit. To identify the MLE for a given model, we required that three searches independently arrive within two log likelihood units of the maximum likelihood value.

7. Model parameters

Model	Parameter type	Parameter name	Fixed/estimated/nuisance parameter
All Models	Baseline infection risk	λ_{0j}	Estimated
All Models	Infection duration	Shape of gamma distribution (k_j)	Fixed
All Models	Infection duration	Scale of gamma distribution (θ_j)	Fixed
All models	Covariate effects (α)	Age	Estimated
All Models	Covariate effects (α)	Educational status	Estimated
All Models	Covariate effects (α)	Age at sexual debut	Estimated
All Models	Covariate effects (α)	Circumcision status	Estimated
All Models	Covariate effects (α)	Number of recent female sexual partners	Estimated
All Models	Covariate effects (α)	Number of recent male sexual partners	Estimated
All Models	Covariate effects (α)	Steady sexual partner	Estimated
All Models	Covariate effects (α)	Consistent condom use	Estimated
All Models	Covariate effects (α)	Current smoking status	Estimated
All Models	Covariate effects (α)	Mexican nationality (α_{MX})	Estimated
All Models	Covariate effects (α)	Brazilian nationality (α_{BZ})	Estimated
All Models	Initial conditions	Probability of initial infection (P_{initial_j})	Estimated
All Models	Initial conditions	Duration of remaining infection if initially infected ($F_{\text{remaining}}$)	Nuisance parameter for initial conditions
All Models	Initial conditions	Probability of previous infection if initially uninfected (p_{past})	Nuisance parameter for initial conditions
All Models	Measurement model	False positive rate (P_{FP})	Fixed
All Models	Measurement model	False negative rate (P_{FN})	Fixed
Homologous immunity model	Immunity dynamics	Magnitude of scaled risk (d)	Estimated
Homologous immunity model	Immunity dynamics	Rate of waning of immunity (w)	Estimated
Additional risk model	Additional risk (d)	d_{celibate}	Estimated
Additional risk model	Additional risk (d)	d_{1partner}	Estimated
Additional risk model	Additional risk (d)	d_{multiple}	Estimated

Table S6. Description of model parameters

Fixed parameters

The false positive and false negative rates of HPV detection were set equal to the high sensitivity (96%) and specificity (99%) of the Roche Linear Array HPV genotyping test reported by the manufacturer (Roche Diagnostics), which has been confirmed by other analyses (15). The duration of each simulated infection with each HPV type j was drawn from a $\Gamma(k_j, \theta_j)$ distribution, where k_j and θ_j were fixed according to the empirical distribution of infection durations in the data for type j (Fig. S8, Table S7). Following Giuliano et al. (1), we required two consecutive negative visits following a positive visit for any HPV type to constitute an observed clearance. Of the consecutive negative tests, the first was assumed to be the first true negative observation. Thus, we treated 1-0-1 transitions as false negatives, thereby adjusting data a priori based on the assumption that infection in such cases was actually continuous. Our justification was two-fold: first, this approach allowed us to account for false negatives in HPV sampling beyond the laboratory specifications of the HPV test that were included in the observation

model. Second, the approach ensured that the high rates of reinfection estimated by the best-fit model were not the result of failing to account for false clearances, ensuring in turn that our estimates of reinfection rates would be conservative.

Type	k_j	θ_j
HPV62	1.20	0.81
HPV84	1.67	0.59
HPV89	1.37	0.70
HPV16	1.71	0.52
HPV51	1.33	0.75
HPV6	1.71	0.51

Table S7. Values of the parameters describing infection durations, as estimated from observed infection and clearance events in the data.

Nuisance parameters

Two nuisance parameters were used to generate the initial conditions for model simulations but were not estimated from the data. For individuals that were initially infected during any simulation, the fraction of an infection duration that they had experienced prior to time $t = 0$ was given by the nuisance parameter $F_{\text{remaining}}$, which was drawn from a uniform (0,1) distribution for each model realization. For individuals that were initially uninfected during any simulation, the probability that the individual had previously been infected at some point in time was given by p_{past}, which was drawn from a uniform (0,1) distribution for each model realization.

Estimated parameters for the best-fit model

Parameter	MLE	Confidence interval	Type
λ_{0j}	-4.6	[-5.2,-4]	HPV62
	-4.6	[-5.2,-4.3]	HPV84
	-4.3	[-4.7,-3.6]	HPV89
	-4.2	[-5.1,-3.5]	HPV16
	-4.3	[-5,-3.8]	HPV51
	-5.1	[-5.7,-4.5]	HPV6
p_{initial} (logit scale)	-2.6	[-2.8,-2.5]	HPV62
	-2.6	[-2.7,-2.5]	HPV84
	-2.8	[-3.1,-2.6]	HPV89
	-2.9	[-3.2,-2.8]	HPV16
	-2.9	[-3.1,-2.8]	HPV51
	-3.2	[-3.4,-2.8]	HPV6
α_{age}	0.0	[-0.3,0.3]	HPV62
	-0.2	[-0.5,0.0]	HPV84
	-0.3	[-0.4,-0.2]	HPV89
	0.0	[-0.2,0.2]	HPV16
	-0.3	[-0.4,0.0]	HPV51
	-0.2	[-0.5,0.0]	HPV6
$\alpha_{\text{Ageatsexualdebut}}$	0.2	[-0.1,0.3]	HPV62
	-0.2	[-0.4,0.1]	HPV84
	-0.3	[-0.4,-0.1]	HPV89
	-0.1	[-0.3,0.1]	HPV16
	-0.3	[-0.5,-0.1]	HPV51
	-0.1	[-0.4,0.1]	HPV6
α_{BZ}	0.9	[0.1,1.2]	HPV62
	0.3	[-0.3,1.0]	HPV84
	0.3	[-0.3,0.7]	HPV89
	0.3	[-0.5,0.6]	HPV16
	-0.3	[-0.5,0.3]	HPV51
	-0.6	[-1.2,0.1]	HPV6
$\alpha_{\text{Circumcised}}$	-0.3	[-0.7,0.1]	HPV62
	0.6	[0.1,0.9]	HPV84
	0.0	[-0.3,0.2]	HPV89
	0.2	[-0.3,0.5]	HPV16

	-0.3	[-0.5,-0.1]	HPV51
	-0.5	[-0.8,0.2]	HPV6
$\alpha_{\text{Consistent condom use}}$	-0.2	[-0.8,0.1]	HPV62
	-0.7	[-1.4,-0.3]	HPV84
	-0.1	[-0.8,0.1]	HPV89
	-0.3	[-0.8,0.1]	HPV16
	-0.2	[-0.5,0.1]	HPV51
	0.3	[-0.1,0.5]	HPV6
$\alpha_{\text{Current smoker}}$	0.5	[0.2,0.8]	HPV62
	0.8	[0.5,1.2]	HPV84
	0.1	[-0.1,0.4]	HPV89
	-0.2	[-0.6,0.3]	HPV16
	0.4	[0.2,0.7]	HPV51
	0.6	[0.2,1.0]	HPV6
$\alpha_{\#\text{Recent female sex partners}}$	0.6	[0.5,0.8]	HPV62
	0.6	[0.5,0.9]	HPV84
	0.6	[0.5,0.8]	HPV89
	0.8	[0.6,0.9]	HPV16
	0.6	[0.4,0.8]	HPV51
	0.6	[0.5,1.8]	HPV6
$\alpha_{\#\text{Recent male sex partners}}$	0.0	[-0.5,0.2]	HPV62
	0.2	[0.0,0.4]	HPV84
	0.5	[0.3,0.7]	HPV89
	-0.8	[-1.9,-0.3]	HPV16
	0.1	[-0.1,0.3]	HPV51
	0.2	[-0.1,0.6]	HPV6
$\alpha_{\text{Educational level}}$	-0.2	[-0.4,0.0]	HPV62
	0.0	[-0.3,0.1]	HPV84
	0.1	[0.0,0.2]	HPV89
	0.0	[-0.2,0.1]	HPV16
	0.0	[-0.2,0.2]	HPV51
	0.0	[-0.2,0.1]	HPV6
α_{MX}	-1.3	[-1.9,-0.6]	HPV62
	-0.1	[-0.8,0.4]	HPV84
	-0.5	[-1.0,0.0]	HPV89
	-1.1	[-1.4,-0.7]	HPV16
	-1.0	[-1.4,-0.7]	HPV51
	-0.4	[-0.8,0.3]	HPV6
$\alpha_{\text{Steady sex partner}}$	-0.7	[-1.0,-0.3]	HPV62
	-0.9	[-1.2,-0.3]	HPV84
	-0.5	[-0.7,-0.3]	HPV89
	-0.9	[-1.2,-0.5]	HPV16
	-0.2	[-0.5,0.2]	HPV51
	-0.4	[-1.0,0.0]	HPV6
d_{celibate}	2.7	[2.2,3.3]	HPV62
	1.0	[0.1,1.6]	HPV84
	5.2	[4.3,6.2]	HPV89
	4.0	[3.5,4.9]	HPV16
	3.9	[2.4,4.2]	HPV51
	2.1	[1.3,2.9]	HPV6
$d_{1\text{ partner}}$	3.3	[2.9,3.7]	HPV62
	1.5	[-0.3,3]	HPV84
	4.2	[3.5,4.8]	HPV89
	3.8	[3.2,4.7]	HPV16
	3.6	[2.5,4.1]	HPV51
	2.8	[2.2,3.3]	HPV6
$d_{>1\text{ partner}}$	3.1	[2.7,3.6]	HPV62
	0.7	[0.3,3.4]	HPV84
	3.9	[3.1,4.5]	HPV89
	4.4	[3.8,5.2]	HPV16
	3	[2.7,3.9]	HPV51

	2.3	[1.7,2.9]	HPV6
w	-0.3	[-0.5,-0.1]	HPV62
	-0.5	[-0.8,-0.3]	HPV84
	0	[-0.3,0.2]	HPV89
	0.3	[0.1,0.5]	HPV16
	-1.3	[-1.6,-0.8]	HPV51
	0	[-0.2,0.3]	HPV6

Table S8. Values of the estimated parameters. Estimates are on a log scale unless otherwise indicated.

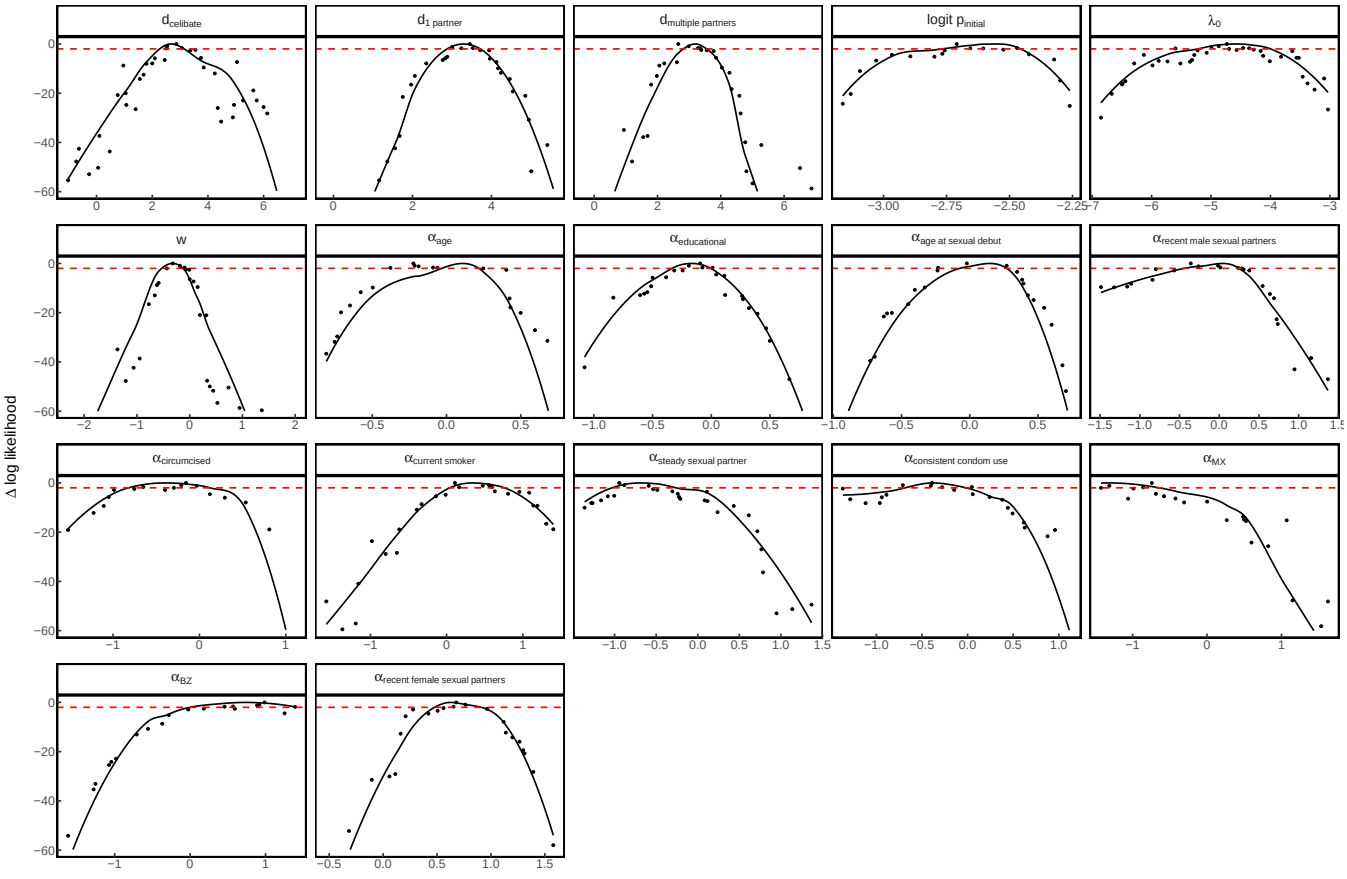

Fig. S10. HPV62: Likelihood profiles for each estimated parameter. The x axis gives the log of the profile parameter value unless the logit distribution is specified. The y axis gives the likelihood relative to the maximum likelihood for the additional risk model for HPV62. The red dashed horizontal line indicates the cutoff of 1.92 log likelihood units used to determine the confidence interval.

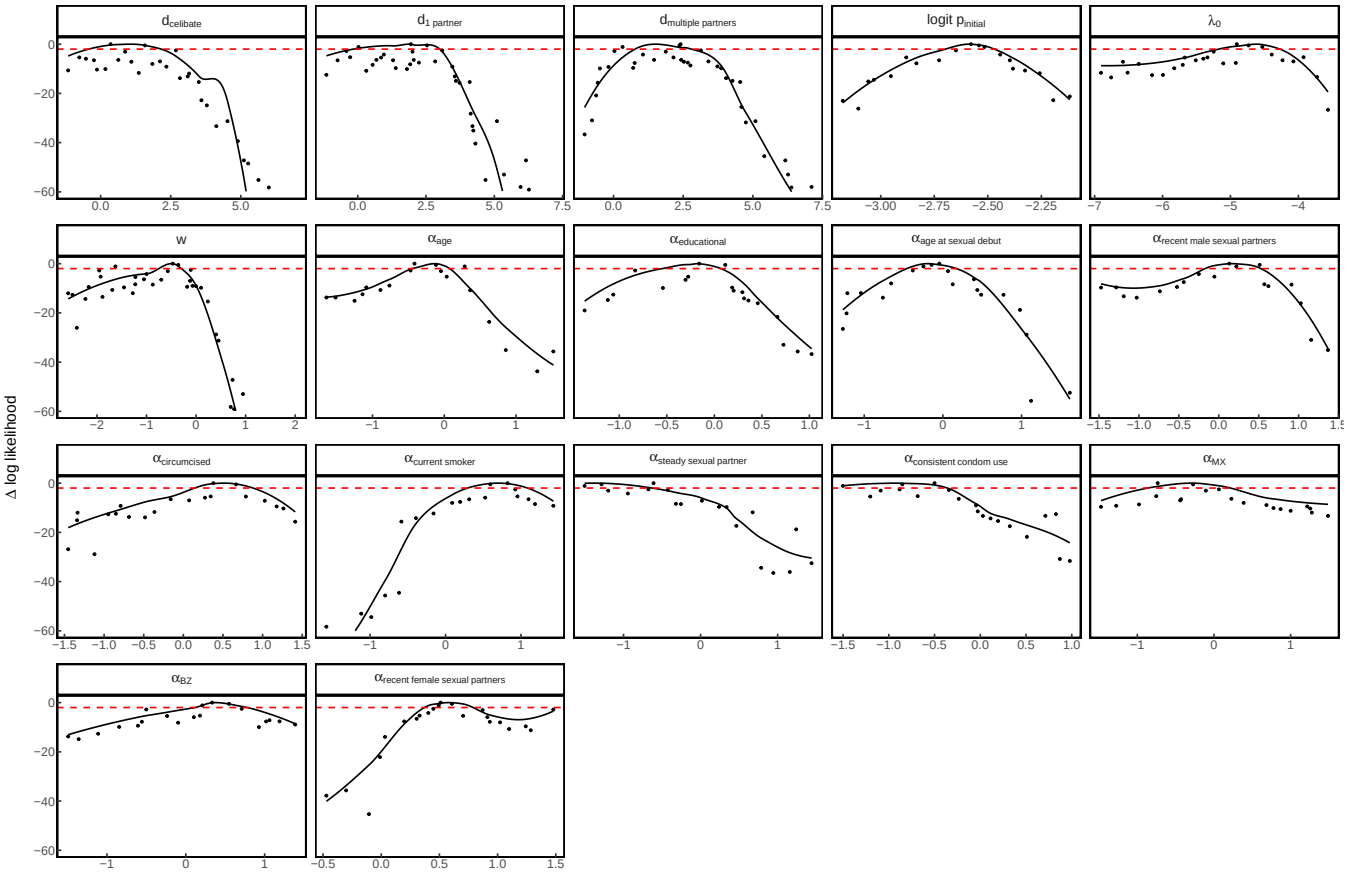

Fig. S11. HPV84: Likelihood profiles for each estimated parameter, as in Fig. S10.

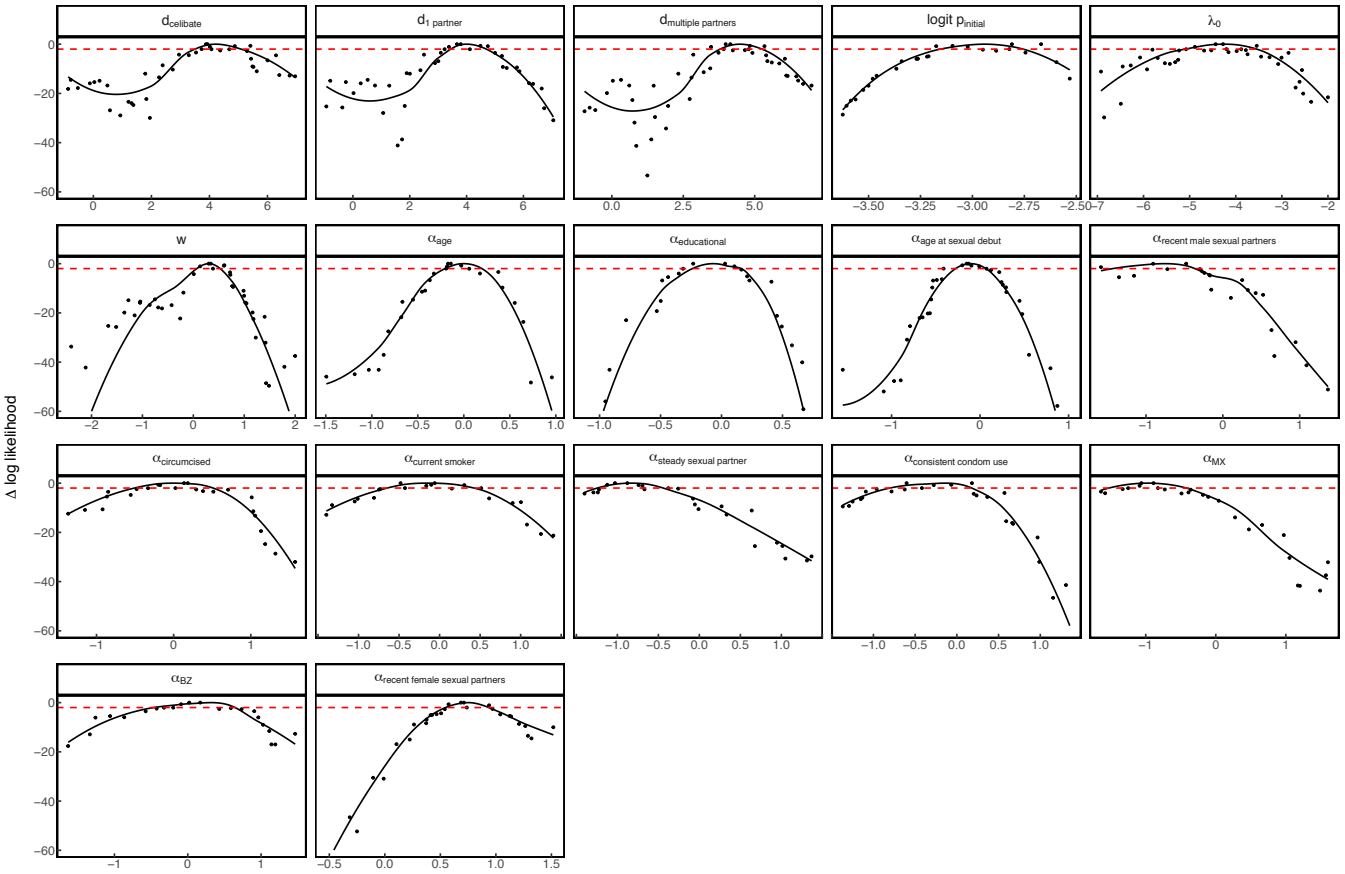

Fig. S12. HPV16: Likelihood profiles for each estimated parameter, as in Fig. S10.

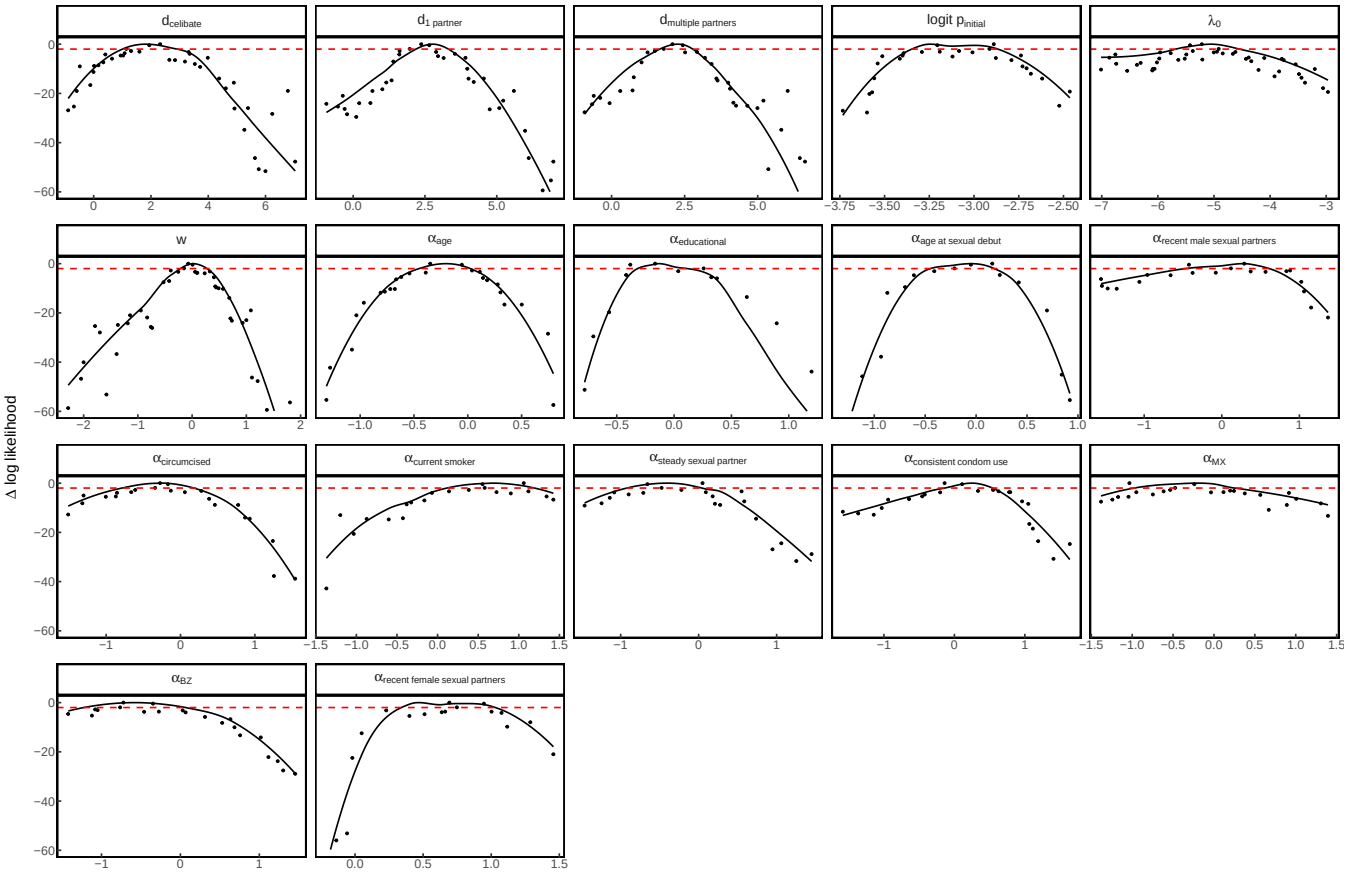

Fig. S13. HPV6: Likelihood profiles for each estimated parameter, as in Fig. S10.

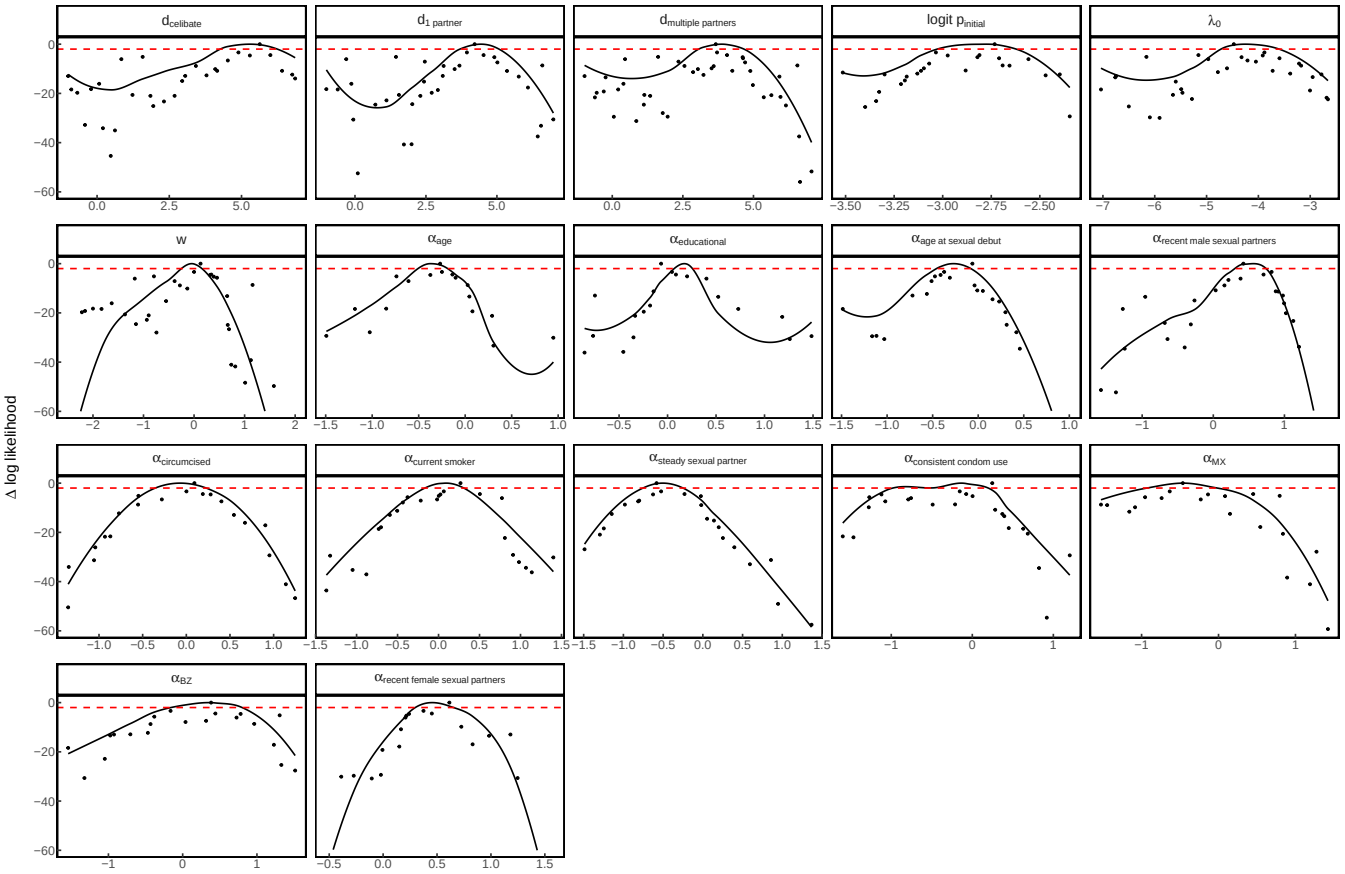

Fig. S14. HPV89: Likelihood profiles for each estimated parameter, as in Fig. S10.

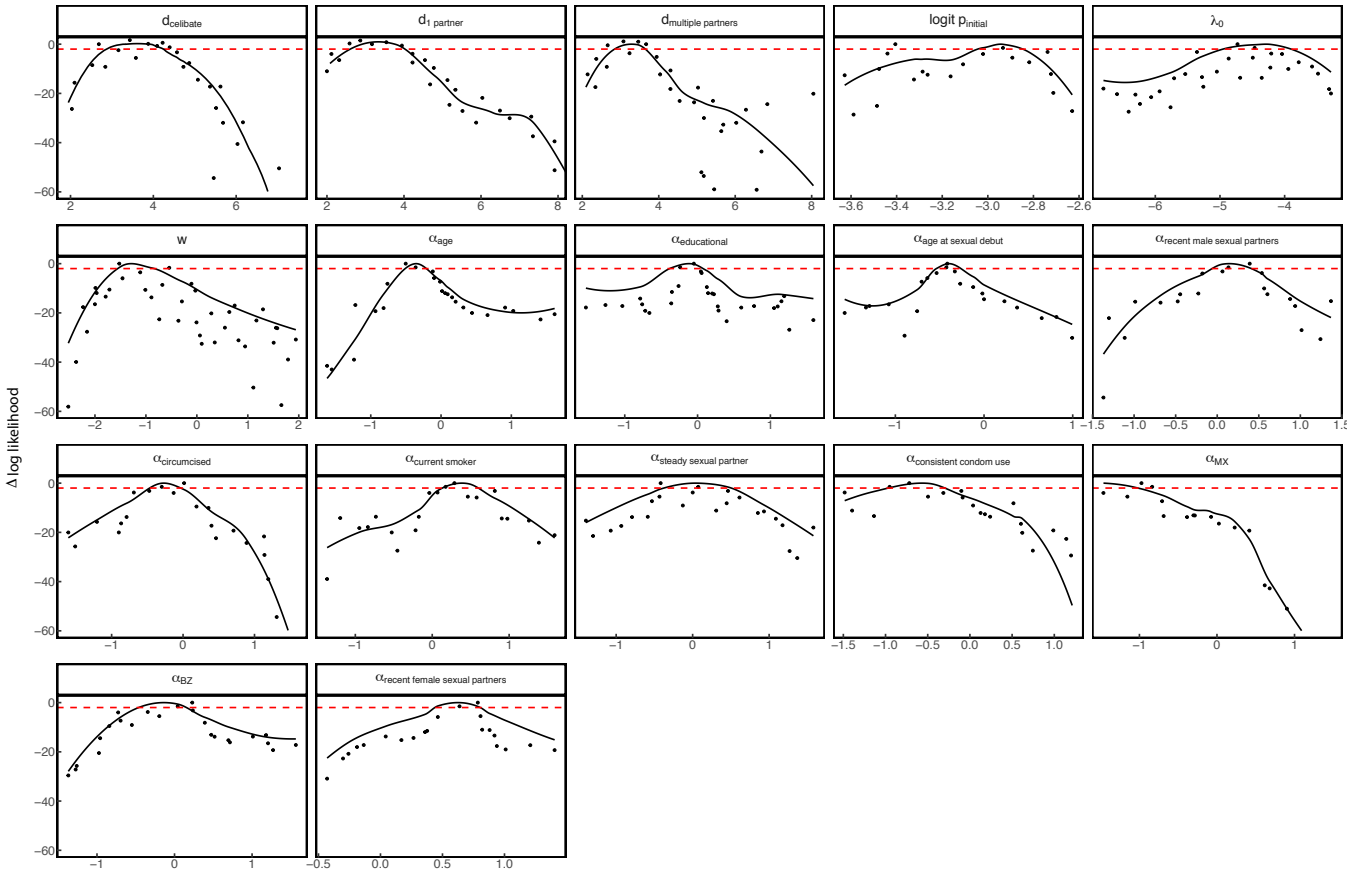

Fig. S15. HPV51: Likelihood profiles for each estimated parameter, as in Fig. S10.

1. Giuliano AR, et al. (2008) The human papillomavirus infection in men study: human papillomavirus prevalence and type distribution among men residing in Brazil, Mexico, and the United States. *Cancer Epidemiol Biomarkers Prev* 17(8):2036–43.
2. Nyitray AG, et al. (2009) Test-retest reliability of a sexual behavior interview for men residing in Brazil, Mexico, and the United States. *A J Epidem* 170(8):965–974.
3. Giuliano AR, et al. (2009) Circumcision and sexual behavior: factors independently associated with human papillomavirus detection among men in the HIM study. *Int J Cancer* 124(6):1251–7.
4. Giuliano AR, et al. (2015) EUROGIN 2014 roadmap: Differences in human papillomavirus infection natural history, transmission and human papillomavirus-related cancer incidence by gender and anatomic site of infection. *Int J Cancer* 136(12):2752–2760.
5. Nyitray AG, et al. (2015) The Natural History of Genital Human Papillomavirus Among HIV-Negative Men Having Sex With Men and Men Having Sex With Women. *J Infect Dis* 212(2):202–212.
6. Han JJ, et al. (2017) Prevalence of Genital Human Papillomavirus Infection and Human Papillomavirus Vaccination Rates Among US Adult Men. *JAMA Oncol* 3(6):810–816.
7. Albero G, Castellsagué X, Lin H (2014) Male circumcision and the incidence and clearance of genital human papillomavirus (HPV) infection in men: the HPV Infection in men (HIM) cohort study. *BMC Infect Dis*.
8. Winer RL, et al. (2006) Condom Use and the Risk of Genital Human Papillomavirus Infection in Young Women. *New Eng J Med* 354(25):2645–2654.
9. Ionides EL, Bretó C, King AA (2006) Inference for nonlinear dynamical systems. *Proc Natl Acad Sci* 103(49):18438–18443.
10. Doucet A, de Freitas N, Gordon N (2001) Sequential Monte Carlo Methods in Practice. Series Statistics For Engineering and Information Science.
11. Romero-Severson E, Volz E, Koopman J, Leitner T, Ionides E (2015) Dynamic Variation in Sexual Contact Rates in a Cohort of HIV-Negative Gay Men. *Am J Epidemiol* 182(3):255–62.
12. Hurvich CM, Tsai CL (1989) Regression and Time Series Model Selection in Small Samples. *Biometrika* 76(2):297.
13. Ionides EL, Breto C, Park J, Smith RA, King AA (2017) Monte Carlo profile confidence intervals for dynamic systems. *J R Soc Interface* 14(132):20170126.
14. Bolker B (2007) Likelihood and all that in *Ecol Model Data R*. (Princeton University Press), 508 edition, pp. 169–221.
15. Gravitt PE, Peyton CL, Apple RJ, Wheeler CM (1998) Genotyping of 27 human papillomavirus types by using L1 consensus PCR products by a single-hybridization, reverse line blot detection method. *J Clin Microbiol* 36(10):3020–7.