Modelling coinfections to detect within-host interactions from genotype combination prevalences

Samuel Alizon[1,*], Carmen Lía Murall[1], Emma Saulnier[1] and Mircea Sofonea[1,2]

1 Laboratoire MIVEGEC (UMR CNRS 5290, IRD 224, UM), Montpellier, France

2 CEFE, CNRS, Université de Montpellier, Université Paul-Valéry Montpellier, EPHE, Montpellier Cedex 5, France

* samuel.alizon@cnrs.fr

Abstract

Parasite genetic diversity has been argued to be informative about the way infectious diseases spread and interact within their hosts. However, most methods developed to detect such interactions rely on infection ranks (i.e. number of genotypes per host) and the few that do use all the combinations of genotypes lack an underlying epidemiological setting. To overcome this limitation, we take advantage of a recent model that captures the dynamics of an arbitrary number of strains with coinfections and cotransmission. Using genital infections by different types of Human Papillomaviruses (HPVs) as a test case, we show that regression Approximate Bayesian Computing (ABC) has the power to detect interactions between high-risk and low-risk HPV types. We also show that contrary to existing method, this detection is not affected by another source of host heterogeneity (here the number of sexual partners). Overall, combining mathematical modelling and sophisticated inference techniques allows us to use new types of data to extract relevant epidemiological information.

Introduction

With the advent of next generation sequencing, an increasing number of infections turn out to be coinfections by multiple genotypes [1]. Of course for some systems, such as genital infections by Human Papillomaviruses (HPVs), this was already known to be the case [2, 3].

Multiple infections, that is the circulation of several parasite genotypes in a host population [4], raise questions at three levels. At the infection level, the virulence expressed in coinfected hosts (or 'overall virulence') can be different from the virulence in single infections. At the epidemiological level, allowing for parasites to infect already infected hosts may affect the way parasites spread. For example, coinfection by malaria and HIV may speed the spread of both parasites [5]. Finally, multiple infections create an additional level of selection that may impact the way parasite traits evolve [6].

Here, we investigate how equilibrium prevalence data can help us infer potential interactions between parasite genotypes. Although these methods can be applied to many systems, we focus in particular on genital HPV infections for three reasons. First, HPV multiple infections are well described thanks to screening for HPV-induced cancers [7–9] and prevalences are relatively stable through time [10]. Second, HPV evolutionary rates are generally slow, which limits within-host evolution and facilitates detection [11]. Third, the existence of within-host interactions between types is strongly debated, especially in the context of vaccination, given that they may affect a potential parasite evolutionary response [12].

The clearest source of within-host interaction between HPV genotypes is the apparent competition mediated by the immune system. Indeed, pre-vaccine and vaccine studies have shown that there is limited natural cross-reactivity between phylogenetically related HPV types and that the vaccines confer some cross-immunity against non-target types [13–15]. Evidence for other kinds of interactions is limited. Within-cell interactions are possible since different HPVs can coinfect the same cell [16]. For some types, virus loads also seem to be differ in single and in coinfections [17], which could impact transmission or recovery rates. There is also indirect epidemiological evidence. First, infection by HPV is known to affect the risk of contracting another infection [3, 18, 19] and to decrease type recovery rate [20]. Second,

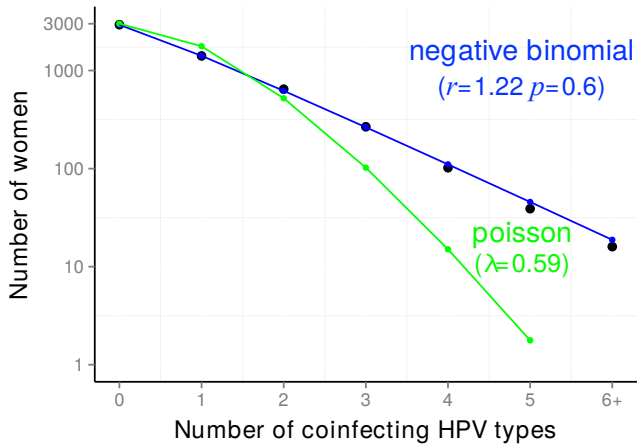

Fig 1. Rank distribution for HPV infections. Black dots show data from 5412 sexually active women in the Costa Rica Vaccine Trial reported by [8]. Lines show maximum likelihood fits performed using the `bbmle` package in `R` [31].

HPV coinfections may interfere with chronic infection and cancer. For example, when high-risk HPV types coinfect with low-risk types, time to diagnosis is longer and the risk of progression to cancer is lower [21]. To summarise, we do know that HPV types may interact within hosts but it is unclear whether these interactions are sufficiently strong to be detected at the population level.

Binary or rank models

Most epidemiological models that allow for parasite genotypes to coexist within a host only allow for up to two genotypes per host and do not allow for cotransmission, although there are exception for both [22–25]. In spite of these simplifications, these models have been instrumental in epidemiology [26].

Studies on macro-parasites have long been focusing on high multiplicity of host infection [27]. They showed that the distribution of the number of macro-parasites per host, which we here refer to as the 'rank' of an infection, can provide information regarding the contact structure within the host population. In absence of heterogeneity of any kind, one would expect rank distributions to follow a Poisson distribution. Interestingly, in many populations, the number of macro-parasites per host tends to follow a negative-binomial distribution, which is often interpreted as evidence for some sort of host population structure or a specific functional response [28–30].

For microparasites, similar studies have been developed, where the rank of the

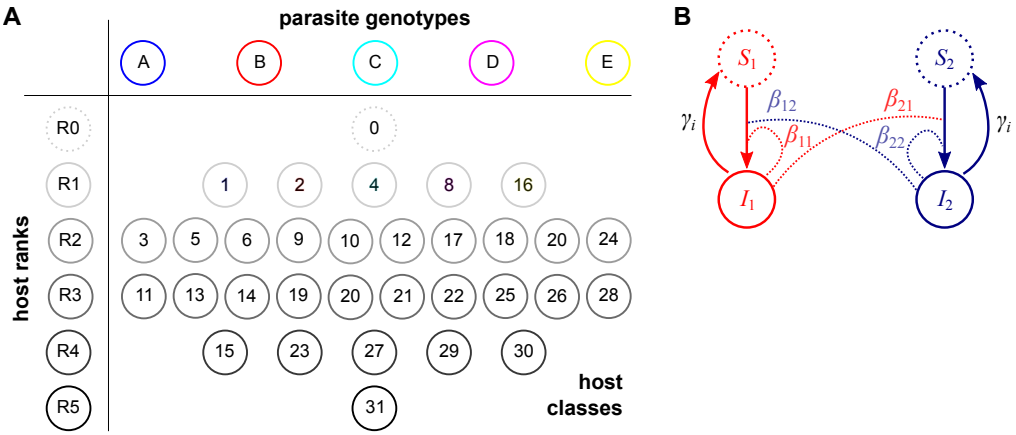

Fig 2. The coinfection epidemiological setting. A) The different prevalences that can be used for $n = 5$ genotypes (per genotype, per rank or per combination). B) Flow diagram showing the population structure with 'normal-spreader' hosts (1 in red) and 'super-spreader' hosts (2 in dark blue). The β and γ indicate transmission and recovery rates.

infection corresponds to the number of genotypes detected in a host. For example, Chaturvedi *et alii* [8] showed that a Poisson distribution can be rejected for HPV coinfections suggesting that there is an excess of coinfections compared to what would be expected in a standard Susceptible-Infected (SI) model. Additional analyses show that a negative binomial distribution provides an excellent fit to the data (Figure 1). This is consistent with the result of the study that identifies the 'number of lifetime sex partners' as the cofactor the most strongly associated with being infected by multiple types instead of a single type [8].

Parasite combination prevalences

Intuitively, there should be more information in the prevalence of each combination of genotypes than in the rank prevalence. With 5 circulating genotypes, there are only 6 host ranks whereas there are 32 combinations (Figure 2A). Some studies have therefore used combination prevalence data to detect interactions. Their approach was to compare the observed prevalence of each combination to an expected value derived from the total prevalence of each genotype.

Fenton *et alii* [32] compared several techniques using a dataset involving 2 species for which the real interactions were known from laboratory experiments. They concluded that correlation techniques performed worse and that the best method

required time series and not just cross-sectional data (see [33] on how to infer interaction parameters from time series using particle filtering techniques). However, the restricted number of strain they used also potentially limited the power of their conclusion (3 ranks and 2 total prevalences versus 4 combinations).

Although longitudinal data is generally richer for epidemiological inference [34], it is not always available and we often need to deal with equilibrium prevalences. To analyse such data, the study by Vaumourin *et alii* [35] considered systems with a larger number of genotypes using a variety of existing techniques (generalised chi-square, network, and multinomial GLM approaches) and developed a new association screening approach that has the advantage to identify and rank combinations based on their deviation from the expectation (see the Methods). To test the power and accuracy of each method, they used simulated distributions but without an explicit epidemiological model.

Inference using explicit modelling

We wish to assess whether, in a setting where n prevalent parasite genotypes or species are circulating, the prevalence of the 2^n coinfected host classes gives us more information about the way parasites spread and interact within their hosts than the $n+1$ rank prevalences. More precisely, our hypothesis is that modelling epidemiological dynamics explicitly can allow us to distinguish between within-host interactions and other types of heterogeneities generated from the host contact structure. Indeed, it is known that for many infectious diseases, especially sexually-transmitted ones [36], some hosts may act as 'super-spreaders' [37]. Intuitively, these hosts should be more exposed and therefore have higher infection ranks independently of any features of the parasites themselves (as mentioned in the case of HPV above [8]).

HPV offers an ideal setting to test these questions because coinfections are frequent and rich data exists. Based on the literature, we use our model to evaluate our ability to test the hypothesis that oncogenic HPV types, also called 'high-risk' (HR) types, have a competitive advantage (or disadvantage) when competing with non-oncogenic types or 'low-risk' (LR) types that tend to cause warts. Given that the probability of HPV transmission per sexual contact is high [38], we assume that any interaction between HR and LR types takes place through the recovery rate.

To test these hypotheses, we adopt mechanistic approach and simulate epidemiological dynamics. This is made possible by a recent analytical framework that can handle an arbitrary number of types in a Susceptible-Infected-Susceptible (SIS) model [25]. In order to assess the ability to infer interactions from the observed coinfection classes, we use a regression-based Approximate Bayesian Computing (ABC) approach [39, 40]. We show that our method performs well on simulated data and that existing methods that lack an explicit epidemiological setting cannot distinguish genotype interaction from general host heterogeneity.

Results

Associations and interaction strength

First we use existing methods developed to detect significant associations between parasites from coinfection data. These have been tested by generating distributions but without any epidemiological model.

The chi-square approach exhibits a slightly positive correlation between the probability that the test is significant and the intensity of interaction between types (estimated by fitting the data using a logistic regression model, Fig 3A). However, even with only 1,000 individuals sampled (in black), most of the observed prevalence distributions tend to deviate from the expected one. With 5,000 hosts sampled or more (in gray), most combinations lead to significant tests (Fig 3A).

The GLM approach seems to be more robust to sample size (Fig 3B) and the positive association between interaction intensity and test significance only occurs if 5,000 or 10,000 individuals are sampled. As for the chi-spare approach, most of the associations remain significant.

Vaumourin et al. [35] cleverly proposed to analyse coinfection combination data using network-based approaches. For the combination network, we found that non-significant runs exhibited higher interaction intensity than significant runs, which was unexpected (Fig S3A). We also found a slight decrease in connectance with increasing interaction intensity, which could be consistent with some combinations being removed due to genotype interaction (Fig 3C).

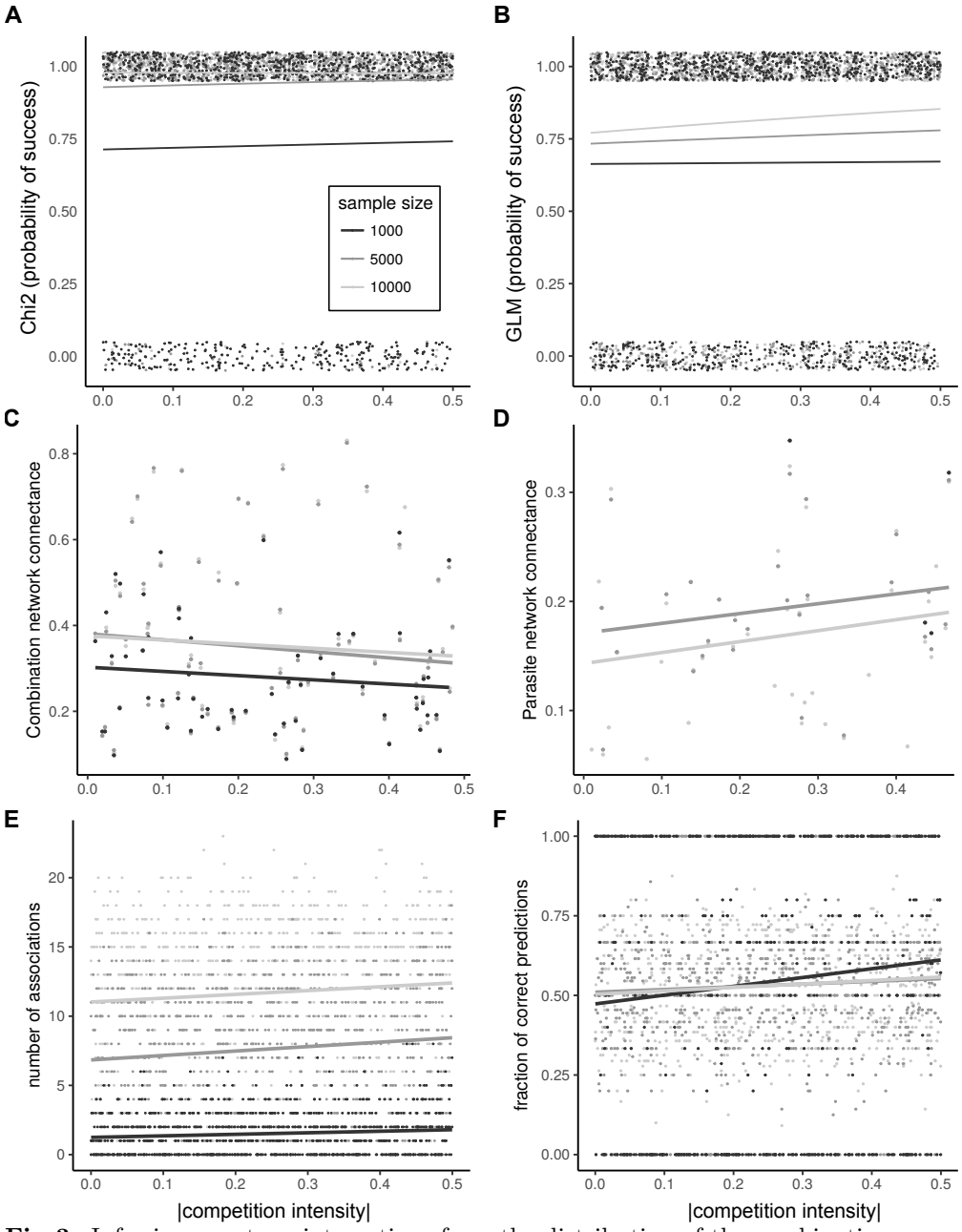

Fig 3. Inferring genotype interactions from the distribution of the combination prevalences using the chi-square (A), the GLM (B), the network (C and D) and the association screening (E and F) approaches. The grayscale indicates the size of the target dataset (100 targets for the network approach and 1000 for the others). Lines show a generalised linear model fit. In A and B the data was scattered vertically for clarity. C and D show the combination and parasite network connectances only when significant. E shows the number of significant interactions and F the fraction of correct predictions based on the correlations from the learning dataset (see Fig S1). Parameter values are drawn in the same prior as the ABC (see Fig S3).

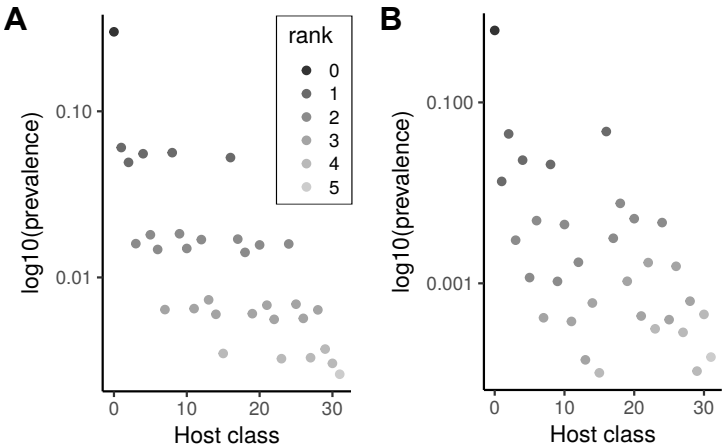

Fig 4. Combination prevalence for a scenario with weak (A) and strong interaction (B). The shading indicates the infection rank (or number of coinfecting genotypes) and the class is a binary code indicating the genotypes present.

For the parasite network, when only 1,000 hosts were sampled significant runs exhibited strikingly high interaction strengths (Fig S3B). We also find an increase in connectance with interaction strength, but only when sampling 5,000 or 10,000 hosts (Fig 3D). This result should be interpreted with caution since parasite network connectance was rarely significant (2, 10 and 15 of the 100 test runs were significant for 1,000, 5,000 and 10,000 hosts sampled respectively). In comparison, combination connectance was significant for 21, 31, 32 of the 100 runs depending on sampling intensity.

Finally, the association screening approach reports an increase in the number of significant associations (i.e. more or less than expected) with host sample size (Fig 3E). By computing equilibrium prevalences for 1,000 parameter values, we estimated the correlation between interaction intensity and the prevalence of each host combination (Fig S1). This allowed us to determine whether the prediction made by the association screening algorithm was correct or not. The fraction of predictions that match our prediction is generally close to 50% with a slight increasing trend with interaction strength for small sample sizes (Fig 3F) This suggests that the other source of heterogeneity (namely contact structure) is sufficient to blur the effect of within-host interactions on the equilibrium prevalences.

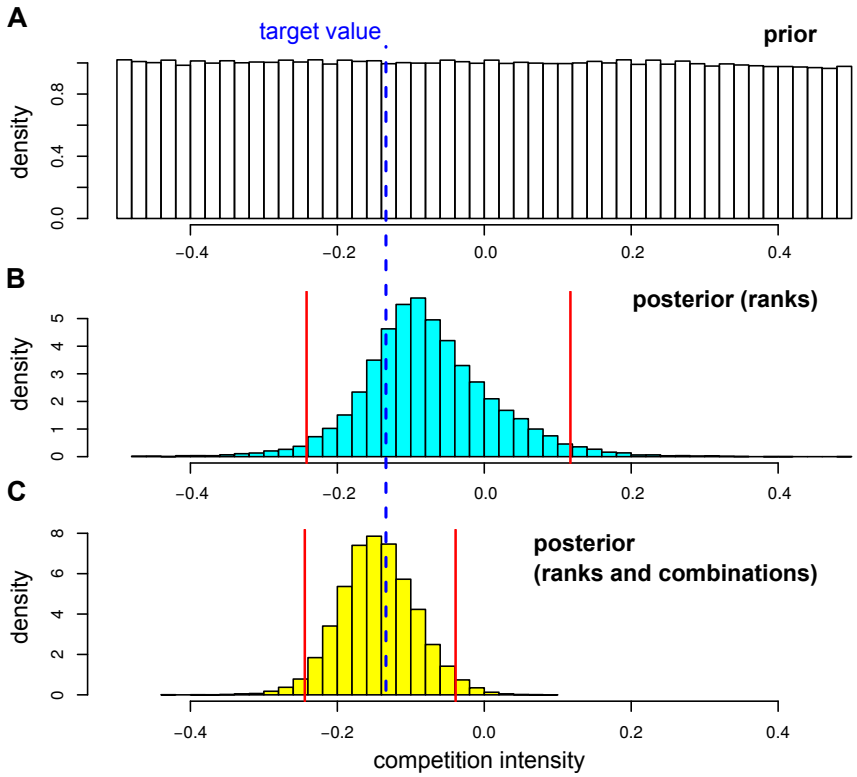

Fig 5. Inferring interaction strength (k). Prior (A) and posterior distributions using only the ranks (B) or the ranks and the combinations (C) as summary statistics. The dashed blue line shows the target value ($k \approx -0.13$) and the red lines show the 95% Highest Posterior Density (HPD).

Epidemiological model: single runs

We first show the fraction of each host combination for two scenarios, one with moderate interactions (parameter set #2 with $k \approx 0.02$, Fig 4A) and another with strong interactions (parameter set #7 with $k \approx 0.25$, Fig 4B). When the interactions are weak, we clearly see the different ranks with uninfected hosts on the top, then a row with the five singly infected host types, etc. When interaction strength increases, these ranks become impossible to distinguish. Fig 4A also illustrates that each parasite genotype in this model has its own infection duration, since they do not all have the same prevalence in single infection. Importantly, we only show the total prevalence of each combination but these may differ among each of the two host types (prevalence is higher in the high rank combinations in the 'superspreader' population). Our goal is to infer the intensity and sign of the interaction between HR and LR genotypes (parameter k) in a heterogeneous host population.

To this end, we applied an ABC approach. As any bayesian method, this means searching a prior distribution in the parameter space. This distribution is shown for all the key parameters in Fig S2. We drew 50,001 parameter sets in this prior, used them to simulate equilibrium densities (as shown in Fig 4). We assessed the performances of the ABC approach following a leave-one-out cross-validation procedure, where we treated one simulation as observed data and the remaining as learning data.

Figure 5 shows the results for parameter set #3 and illustrates how using more summary statistics helps to narrow the distribution from the prior for a dataset with 10,000 individuals. If we only use the ranks, we do narrow the prior distribution but its width remains large enough such that 0 (no interaction) cannot be ruled out from the 95% Highest Posterior Density (HPD), which can be seen as a credibility interval 5B). Using the combinations in addition to the ranks as summary statistics for the ABC allows us to narrow this interval and to exclude 0 from the 95% confidence interval (5C). Using additional information, for example being able to distinguish between the two host types, would narrow it even more as we will see below.

Epidemiological model: cross-validation

The previous analysis was based on a single run but all parameters may vary in a relatively large prior distribution (Fig S2). We therefore repeated the analysis for 100 different target runs. We varied the number of sampled individuals (included the deterministic prevalence value as a proxy for an infinite sample size). Furthermore, we report here a third set of summary statistics involving the rank and combinations for the two hosts subpopulations (see the Methods).

Logically, the width of the 95% HPD for the estimate of interaction intensity decreased with the number of host sampled (Fig 6A). On the same figure, we see that including more summary statistics also decreased the width of this interval, especially for an infinite sample size.

In terms of the relative error regarding the interaction parameter (k), we found a similar effect with a lower error when more host were sampled or more summary statistics were involved (Fig 6B). However, using the combinations in addition to the ranks only improved the analysis if enough hosts were sampled (5,000 or 10,000). In

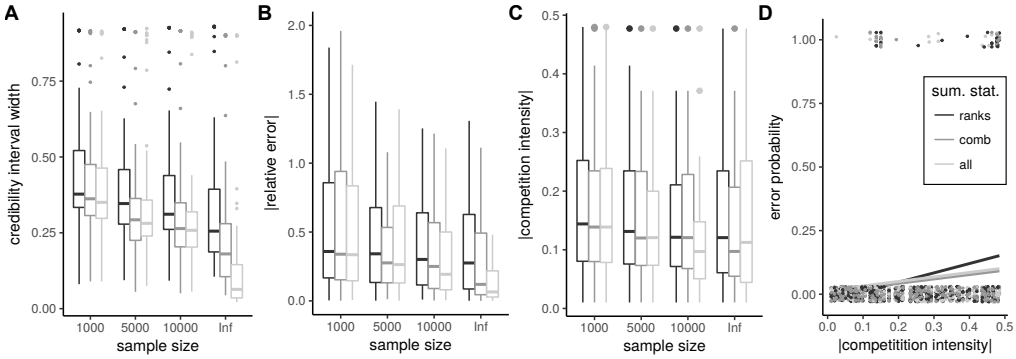

Fig 6. ABC inference precision over 100 runs. A) 95% Highest Posterior Density (HPD), B) absolute value of the relative error, C) average of the absolute value of interaction intensity in runs where 0 is in the 95% HPD and D) runs for which the target value lies outside the 95% HPD. Grayscales indicate the summary statistics used for the ABC. In D, the lines show the result of a generalised linear model.

general, the relative error decreased with interaction strength (figure not shown).

If we focus on the runs for which we could not exclude an absence of interaction (i.e. 0 lied within the 95% HPD), we see that the number of such runs decreased as the number of summary statistics increased (Fig S6). We also see that, in these runs, interaction strength decreased with the sample size and with the number of summary statistics involved (Fig 6C). Notice that for large sample sizes, 95% HPD are narrower, which means that absence of interaction can usually be excluded, making it more difficult to draw conclusions regarding interaction strength because other parameters vary.

Finally, the proportion of errors, that is when the target value was outside the 95% HPD was close to the expected 5% (6.25% with the ranks and 5% with both the ranks and the combinations) but it slightly increased with interaction strength (Fig 6D).

Discussion

Multiple infections are known to affect the virulence of an infection [41], the spread of infectious diseases [5] and their evolution [6]. This is due to the fact that when sharing a host, parasites can interact in various ways [42]. The goal of this study was to determine to what extent the prevalence of parasite combinations can inform us on such interactions.

By generating prevalence data from an mechanistic epidemiological model, we were able to first test the power of existing heuristic methods based on the distribution of classes. Overall, these results show that these methods are limited. This is largely due to the fact that we introduced host heterogeneity in the model, which affects the distribution of host classes in a way that cannot be distinguished from interaction between parasite genotypes. This therefore corroborates a limitation often mentioned in such studies, which is that departures from expected distributions need not be due to interaction between genotypes.

We then used an ABC approach to infer parameters from the model. We show that this yields more consistent results than existing heuristic methods. Quite expectedly, the accuracy of the method increases with the number of hosts sampled. We also show that using the prevalence of all the combinations of host classes tends to decrease the error made compared to using only the prevalence of infection ranks. Finally, adding knowledge about host type ('super-spreader' or 'normal-spreader') can further improve the power of the inference.

The fact that decent results can be obtained by only using the rank of the infections may seem surprising considering the difficulty from existing models to infer interactions. One reason for this could be that we have a mechanistic model, which limits the range of rank distributions that can be explored. Another reason is that we here use the same model to generate the target dataset and the learning datasets, which facilitates the ABC inference.

We do not report it here but the accuracy of the inference varied widely across parameters. For the interaction parameter (k), the inference reduced the initial 95% HPD of the prior by 66%. In comparison, this was less than for the transmission probability (β, 75%), but much better than for the assortativity parameter (a, 45%),

host heterogeneity (h, 38%) or the individual recovery rates (γ_i, 13%).

There are several ways to extend this framework. One would be to use more powerful regression techniques, such as neural networks. However, these may be more difficult to parameterise. Furthermore, even though it contains several parameters, our model remains relatively simple compared to the power of these algorithms. One possibility to address this could be to use a agent-based model with sophisticated agent behaviours to generate a richer dataset. This would be useful in itself to generate test runs with known parameter values to further test the power of our method on more noisy data. It would also allow to control for biases related to the contact network structure between hosts and the dynamical aspect of sexual partnerships that have been shown to interfere with the detection of coinfection interactions [43].

Finally, the next step is, of course, to test this model using actual epidemiological data. We here used HPV as a case study but it would be possible to study coinfections between different parasite species, although this might require substantial modifications in the model to capture the life-history of each parasite. Even in the case of HPV, analysing real data will require to add several processes we chose to ignore here. First, HPV detection tests may exhibit cross-reactivity between HPV types, thus inflating the prevalence of some combinations. This effect if well described and can be handled for each detection test. Second, when hosts are infected by many HPV types, some of these may not be detected, thus decreasing the prevalence of high-rank infections. This effect is more subtle and would require to be inferred in the model.

Overall, ABC and machine learning allow us to extract the information from the equilibrium prevalence of all the combinations of genotype prevalences. Therefore, combining coinfection modelling with epidemiological data can bring new elements to the controversy regarding the importance of interactions between HPV types.

Methods

The epidemiological model

The model is based on the deterministic ODE-based framework introduced by Sofonea et al. [25] that allows for an arbitrary number of parasite genotypes to circulate in a host population without assuming any particular infection pattern (see [4] for the importance of this relaxation). Furthermore, the framework enables cotransmission in the sense that infected hosts can simultaneously transmit any subset of genotypes they are infected with.

Multiple infections Let us consider that hosts can be potentially infected by any combination of n parasite genotypes and sort them in classes according to the genotypes present (we use a binary code to map the presence/absence of the genotypes the hosts class labels). For computational reasons, we assumed that $n \leq 5$, as the number of classes increases geometrically with the number of genotypes.

Epidemiological dynamics follow a classical susceptible-infected-susceptible (SIS) framework, where upon contact with an infected host, a 'recipient' host can acquire any subset of the genotypes carried by this 'donor' host (cotransmission). In terms of recovery, we assume that genotypes can be cleared independently. Importantly, each genotype g is cleared at a specific rate $\gamma_g \geq 1\text{year}^{-1}$. This sets the average infection duration to a year [20, 44]. Given that we focus on HPV infections in young adults, we neglect infection-induced mortality.

Mathematically, the dynamics can be captured in a compact form using the master equation [25]:

$$\mathrm{d}\mathbf{y}/\mathrm{d}t = \beta\mathbf{\Phi}(\mathbf{y} \otimes \mathbf{y}) - \beta(\mathbf{\Psi}\mathbf{y}) \odot \mathbf{y} + (\mathbf{\Xi} - \mathbf{\Theta})\,\mathbf{y} \tag{1}$$

where $\mathbf{y}$ is the vector of densities of the 2^n host classes, $\odot$ denotes the Hadamard (element-wise) matrix product, $\otimes$ the Kronecker (outer) product, $\mathbf{\Phi}$ is the infection input flow matrix, $\mathbf{\Psi}$ is the infection output flow matrix, $\mathbf{\Xi}$ is the recovery input flow matrix and $\mathbf{\Theta}$ is the recovery output flow matrix and β is the (constant) probability of transmission per contact that scales all infection processes. Equation system 1 allow us

to track all the flows going in and out of host compartments through time. For simplicity, we neglect host demography (births and deaths) and assume that the host population size is constant. Given that infected hosts do not always sero-convert and that natural immunity is much lower than vaccine-induced immunity [15], we neglect immunisation in the model.

Population structure The model was enhanced by splitting the host population into two sub-populations that differ in their contact rates ('super-spreader' versus 'normal-spreader' hosts) as shown in Figure 2B. Contact between the two sub-populations follows a classical pattern based on the assortment (a) between host types, the proportion of each host type ($p_1 = p$ and $p_2 = 1 - p$) and their activity rates (equal to $c_1 = 1$ and $c_2 = h$, with $h \geq 1$). Overall, the contact rate between a 'recipient' individual from sub-population j and a 'donor' individual from sub-population i is

$$c_{ji} = (1-a)\frac{c_i\ c_j}{p + (1-p)\ h} + \delta_{ij}\ a\ c_i \tag{2}$$

where δ_{ij} is the Kronecker delta and h is the difference in activity between the two host classes.

This population structure implies that we have two vectors of host classes ($\mathbf{y_1}$ and $\mathbf{y_2}$). If we denote the combined vector $\mathbf{y_\bullet} = (\mathbf{y_1}, \mathbf{y_2})$, the master equation can be written similarly to 1 by updating the matrices in the following way:

$$\mathbf{A_\bullet} = \mathrm{diag}\,(\mathbf{A}, \mathbf{A}) \ \text{ for } \mathbf{A} \equiv \boldsymbol{\Delta}, \boldsymbol{\Theta}, \boldsymbol{\Xi}, \qquad \boldsymbol{\Psi_\bullet} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \otimes \boldsymbol{\Psi}$$

$$\text{and} \quad \boldsymbol{\Phi_\bullet} = \begin{bmatrix} \left(\mathbf{1}\mathbf{1}^{\mathrm{T}} \otimes (c_{11}, c_{12}) \otimes \mathbf{1}^{\mathrm{T}}\right) \odot \boldsymbol{\Phi}' & \mathbf{0} \\ \mathbf{0} & \left(\mathbf{1}\mathbf{1}^{\mathrm{T}} \otimes (c_{21}, c_{22}) \otimes \mathbf{1}^{\mathrm{T}}\right) \odot \boldsymbol{\Phi}' \end{bmatrix},$$

where $\mathbf{1}$ denotes the 2^n-dimensional column vector with unit elements, and $\boldsymbol{\Phi}'$ is obtained by repeating each $2^n \times 2^n$ block $\boldsymbol{\Phi}^{[i]}$ of the original $2^n \times 2^{2n}$ matrix $\boldsymbol{\Phi} = \left(\boldsymbol{\Phi}^{[i]}\right)_{i=1,\ldots,2^n}$ as $\boldsymbol{\Phi}' = \left(\boldsymbol{\Phi}^{[i]}, \boldsymbol{\Phi}^{[i]}\right)_{i=1,\ldots,2^n}$.

Model simulations The model was implemented and simulated in R. The script is already available upon request and will be published on a repository along with the raw

data (simulated prevalences).

The equilibrium prevalences from the deterministic model were used to generate datasets in finite populations of 1000, 5000 and 10,000 hosts assuming a multinomial distribution where the probability to draw a host with a given genotype combination was equal to this combination's prevalence.

HPV interactions For simplicity, within-host dynamics were neglected here and the effect of genotype diversity on the infection parameters was modelled in the following way. First, we assumed that genotype transmission was unaffected by the presence of other genotypes in the host. This was motivated by the very high transmission probability of HPV per contact [38]. Second, we assumed that interactions between HPV types take place through the recovery rates.

Even with 5 genotypes, this could mean 20 interaction parameters (e.g. how the presence of genotype A affect the clearance rate of genotype B). To reduce this complexity, we assumed that genotypes could be sorted into two groups. Biologically, these groups can correspond to high-risk (i.e. carcinogenic) and low-risk HPV types, or to any other binary classification. Whenever a genotype from the second group coinfects a host with a genotype from the other group, its individual recovery rate is multiplied by a factor $1 + k$, with $k \in [-0.5, 0.5]$). Genotypes from the first group are assumed to be unaffected by the presence of other genotypes (otherwise we would need an additional parameter and assumptions as to the interaction between the two parameters). Depending on whether k is greater or lower than 1, we expect host classes containing genotypes from the second group to be under- or over-represented respectively. We assumed that one of the groups contains 3 genotypes and the other 2 but a different partitioning would lead to similar results and should eventually be decided based on the data.

Inference from distributions

In order to compare our framework to existing methods, we use 4 techniques used by Vaumourin *et al.* [35], who implemented them in R. These are briefly described here but readers interested in more detailed should refer to the original publication.

Assocation screening This approach involves simulating datasets of occurrence count of each combination based on the genotype prevalences [35]. From these simulations, a 95% confidence envelope is calculated for each combination, thus allowing to detect deviation from the expected distribution in the dataset.

Multinomial GLM This model consists in calculating the deviance from a statistical distribution obtained with a Generalised Linear Model and a multinomial family. Practically, the multinomial logistic regression model was performed using the *vglm* function from the `VGAM` package in R [45].

Generalised chi-square This test does not involve any simulations and is based on the expected chi-square distribution of combinations given the total prevalence of each parasite strain. Note that combinations with 5 hosts or less were grouped together.

Network connectance Another possibility is to represent the parasite combinations as a network and to study the connectance, that is the proportion of observed edges relative to the number of edges. Here, individuals are connected if they share the same parasite (parasite network) or the same combination of parasites (combination network). Connectance was computed using the `igraph` R package. These scripts are available upon request and will be published on a repository.

Regression-ABC

The methods used here follow that developed in phylodynamics [40] and apply them to different summary statistics. In short, Approximate Bayesian Computation (ABC) is a likelihood-free method to infer parameter values from a given dataset [46]. It consists in simulating many datasets, for which by definition the underlying parameters are known, and comparing them to the target dataset the parameters of which we want to estimate. This comparison is often done by breaking the datasets into summary statistics. We use regression-ABC [39], which is divided into two steps. First, a rejection step, where only the simulated runs that are close enough from the target are kept. Second, a regression model is learnt on the remaining runs. Once we know how to map summary statistics to the parameter space, we can infer the parameters from any target dataset from which

the same summary statistics can be extracts.

Here, using model (1) and following [25], we calculated the equilibrium prevalences of each of the 64 host classes (32 classes for each host type) for 50,001 parameter combinations. We used large and uninformative priors for the varied parameters (Figure S2). More specifically, we varied the interaction strength (our parameter of interest, $k \in [-0.5, 0.5]$) the transmission rate ($\beta \in [0.5, 1.5]$), the assortativity ($a \in [0, 1]$), the activity difference between host types ($h \in [1, 20]$) and the specific infection duration modified ($d_i \in [0.6, 1]$).

We report three sets of summary statistics:

- the RANKS set: the rank prevalences and the total prevalence of each genotype, that is 10 summary statistics
- the COMB set: the rank set combined with all the combination prevalences, that is 42 summary statistics
- the ALL set: the comb set for each of the two types of hosts (84 summary statistics) plus all the differences between the combination prevalence and the corresponding rank prevalence (64 summary statistics), that is 148 summary statistics.

The first set is intended to be compared to classical methods that ignore combinations of genotypes, the second is based on the type of data that could be easily accessed and the third is for a very optimistic scenario in which we would know which group every host belongs to.

We compared several levels of tolerance using a preliminary run of the model (with narrower priors) and identified 50% as an optimal cut-off for the rejection: lowering the tolerance did not improve the inference (measured via the fraction of runs where the target value ended up in the 95% HPD), whereas increasing it decreased the inference quality.

Following an earlier study [40], we used a LASSO regression to learn the model. Although it performs a linear regression, it has the advantage to be less prone to overlearning than more elaborate non-linear regressions, such as Support Vector Machines. The LASSO adjustment was implemented using the `glmnet` R package and the ABC itself was performed using the `abc` package. In practice, one of the 50,001 runs

was removed and used as a target, whereas the remaining runs were used to learn the regression model (after performing a rejection step). We repeated the operation 100 times to generate 100 target datasets.

Supporting information

Supplementary Figures.

Acknowledgments

We thank Elise Vaumourin for sharing her R script and helping with its implementation.

Funding

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 648963) with additional funding from the CNRS and the IRD.

Conflict of Interests

All authors declare no conflicts of interests.

References

1. Juliano JJ, Porter K, Mwapasa V, Sem R, Rogers WO, Ariey F, et al. Exposing malaria in-host diversity and estimating population diversity by capture-recapture using massively parallel pyrosequencing. Proc Natl Acad Sci USA. 2010;107(46):20138–43.

2. Thomas KK, Hughes JP, Kuypers JM, Kiviat NB, Lee SK, Adam DE, et al. Concurrent and Sequential Acquisition of Different Genital Human Papillomavirus Types. J Infect Dis. 2000;182(4):1097–1102.

3. Rousseau MC, Pereira JS, Prado JCM, Villa LL, Rohan TE, Franco EL. Cervical Coinfection with Human Papillomavirus (HPV) Types as a Predictor of Acquisition and Persistence of HPV Infection. J Infect Dis. 2001;184(12):1508–1517.

4. Sofonea M, Alizon S, Michalakis Y. Exposing the diversity of multiple infection patterns. J Theor Biol. 2017;419:278–289.

5. Abu-Raddad LJ, Patnaik P, Kublin JG. Dual infection with HIV and malaria fuels the spread of both diseases in sub-Saharan Africa. Science. 2006;314(5805):1603–6.

6. Alizon S, de Roode JC, Michalakis Y. Multiple infections and the evolution of virulence. Ecol Lett. 2013;16(4):556–567.

7. Vaccarella S, Franceschi S, Snijders PJF, Herrero R, Meijer CJLM, Plummer M, et al. Concurrent Infection with Multiple Human Papillomavirus Types: Pooled Analysis of the IARC HPV Prevalence Surveys. Cancer Epidemiol Biomarkers Prev. 2010;19(2):503–510.

8. Chaturvedi AK, Katki HA, Hildesheim A, Rodríguez AC, Quint W, Schiffman M, et al. Human Papillomavirus Infection with Multiple Types: Pattern of Coinfection and Risk of Cervical Disease. J Infect Dis. 2011;203:910–920.

9. Dickson EL, Vogel RI, Bliss RL, Downs LS. Multiple-type Human Papillomavirus (HPV) Infections: A Cross-sectional Analysis of the Prevalence of Specific Types in 309,000 Women Referred for HPV Testing at the Time of Cervical Cytology. Int J Gynecol Cancer. 2013;23(7):1295–1302.

10. Alemany L, de Sanjosé S, Tous S, Quint W, Vallejos C, Shin HR, et al. Time trends of human papillomavirus types in invasive cervical cancer, from 1940 to 2007. Int J Cancer. 2014;135(1):88–95.

11. Bravo IG, de Sanjosé S, Gottschling M. The clinical importance of understanding the evolution of papillomaviruses. Trends Microbiol. 2010;18(10):432–8.

12. Murall CL, Bauch CT, Day T. Could the human papillomavirus vaccines drive virulence evolution? Proc Biol Sci. 2015;282:20141069.

13. Herrero R. Human papillomavirus (HPV) vaccines: limited cross-protection against additional HPV types. J Infect Dis. 2009;199(7):919–922.

14. 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 Oncol. 2012;13(1):100–110.

15. Beachler DC, Jenkins G, Safaeian M, Kreimer AR, Wentzensen N. Natural Acquired Immunity Against Subsequent Genital Human Papillomavirus Infection: A Systematic Review and Meta-analysis. J Infect Dis. 2016;213(9):1444–1454.

16. McLaughlin-Drubin ME, Meyers C. Evidence for the coexistence of two genital HPV types within the same host cell in vitro. Virology. 2004;321(2):173–180.

17. Xi LF, Edelstein ZR, Meyers C, Ho J, Cherne SL, Schiffman M. Human Papillomavirus Types 16 and 18 DNA Load in Relation to Coexistence of Other Types, Particularly Those in the Same Species. Cancer Epidemiol Biomarkers Prev. 2009;18(9):2507–2512.

18. Méndez F, Muñoz N, Posso H, Molano M, Moreno V, van den Brule AJC, et al. Cervical Coinfection with Human Papillomavirus (HPV) Types and Possible Implications for the Prevention of Cervical Cancer by HPV Vaccines:. J Infect Dis. 2005;192(7):1158–1165.

19. Tota JE, Ramanakumar AV, Villa LL, Richardson H, Burchell AN, Coutlée F, et al. Cervical Infection With Vaccine-Associated Human Papillomavirus (HPV) Genotypes as a Predictor of Acquisition and Clearance of Other HPV Infections. J Infect Dis. 2016;.

20. Trottier H, Mahmud S, Prado JCM, Sobrinho JS, Costa MC, Rohan TE, et al. Type-Specific Duration of Human Papillomavirus Infection: Implications for Human Papillomavirus Screening and Vaccination. J Infect Dis. 2008;197(10):1436–1447.

21. Sundström K, Ploner A, Arnheim-Dahlström L, Eloranta S, Palmgren J, Adami HO, et al. Interactions Between High- and Low-Risk HPV Types Reduce the Risk of Squamous Cervical Cancer. J Natl Cancer Inst. 2015;107(10):djv185.

22. May RM, Nowak MA. Coinfection and the evolution of parasite virulence. Proc R Soc Lond B. 1995;261:209–215.

23. Lion S. Multiple infections, kin selection and the evolutionary epidemiology of parasite traits. J Evol Biol. 2013;26(10):2107–22.

24. Alizon S. Co-infection and super-infection models in evolutionary epidemiology. Interface Focus. 2013;3:20130031.

25. Sofonea M, Alizon S, Michalakis Y. From within-host interactions to epidemiological competition: a general model for multiple infections. Philos Trans R Soc Lond B. 2015;370:20140303.

26. Keeling MJ, Rohani P. Modeling infectious diseases in humans and animals. Princeton University Press; 2008.

27. Anderson RM, May RM. Infectious Diseases of Humans. Dynamics and Control. Oxford: Oxford University Press; 1991.

28. Grafen A, Woolhouse MEJ. Does the negative binomial distribution add up? Parasitology Today. 1993;9(12):475–477.

29. Shaw DJ, Dobson AP. Patterns of macroparasite abundance and aggregation in wildlife populations: a quantitative review. Parasitology. 1995;111(S1):S111–S133.

30. Wilber MQ, Johnson PTJ, Briggs CJ. When can we infer mechanism from parasite aggregation? A constraint-based approach to disease ecology. Ecology. 2017;98(3):688–702.

31. Bolker BM. Ecological models and data in R. Princeton, NJ: Princeton University Press; 2008.

32. Fenton A, Knowles SCL, Petchey OL, Pedersen AB. The reliability of observational approaches for detecting interspecific parasite interactions: comparison with experimental results. Int J Parasitol. 2014;44(7):437–45.

33. Shrestha S, King AA, Rohani P. Statistical inference for multi-pathogen systems. PLoS Comput Biol. 2011;7(8):e1002135.

34. Rohani P, King AA. Never mind the length, feel the quality: the impact of long-term epidemiological data sets on theory, application and policy. Trends Ecol Evol. 2010;25(10):611–618.

35. Vaumourin E, Vourc'h G, Telfer S, Lambin X, Salih D, Seitzer U, et al. To be or not to be associated: power study of four statistical modeling approaches to identify parasite associations in cross-sectional studies. Front Cell Infect Microbiol. 2014;4.

36. Liljeros F, Edling CR, Amaral LA, Stanley HE, Aberg Y. The web of human sexual contacts. Nature. 2001;411(6840):907–8.

37. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature. 2005;438(7066):355–9.

38. Winer RL, Hughes JP, Feng Q, O'Reilly S, Kiviat NB, Holmes KK, et al. Condom use and the risk of genital human papillomavirus infection in young women. N Engl J Med. 2006;354(25):2645–54.

39. Csilléry K, Olivier F, Blum MGB. abc: an R package for approximate Bayesian computation (ABC). Method Ecol Evol. 2012;3(3):475–479.

40. Saulnier E, Gascuel O, Alizon S. Inferring epidemiological parameters from phylogenies using regression-ABC: A comparative study. PLoS Comput Biol. 2017;13(3):e1005416.

41. Balmer O, Tanner M. Prevalence and implications of multiple-strain infections. Lancet Infect Dis. 2011;11(11):868–878.

42. Mideo N. Parasite adaptations to within-host competition. Trends Parasitol. 2009;25(6):261–8.

43. Malagón T, Lemieux-Mellouki P, Laprise JF, Brisson M. Bias Due to Correlation Between Times-at-Risk for Infection in Epidemiologic Studies Measuring Biological Interactions Between Sexually Transmitted Infections: A Case Study

Using Human Papillomavirus Type Interactions. Am J Epidemiol. 2016;184(12):873–883.

44. Insinga RP, Dasbach EJ, Elbasha EH, Liaw KL, Barr E. Incidence and Duration of Cervical Human Papillomavirus 6, 11, 16, and 18 Infections in Young Women: An Evaluation from Multiple Analytic Perspectives. Cancer Epidemiol Biomarkers Prev. 2007;16(4):709–715.

45. Yee TW. Vector Generalized Linear and Additive Models. Springer, New York, NY; 2015.

46. Beaumont MA. Approximate Bayesian Computation in Evolution and Ecology. Annu Rev Ecol Evol Syst. 2010;41:379–406.

A Supplementary Figures

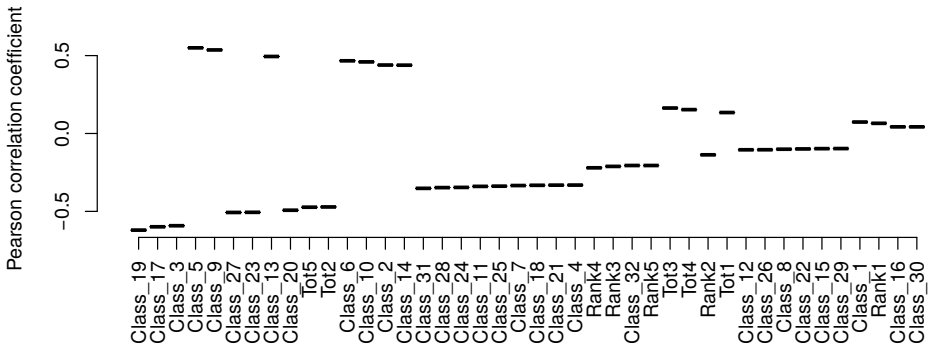

Fig S1. Correlation between interaction intensity and combination, rank or genotype prevalence. The values show the Pearson correlation coefficient obtained using 1,000 parameter sets from the ABC training dataset (priors in Figure S2).

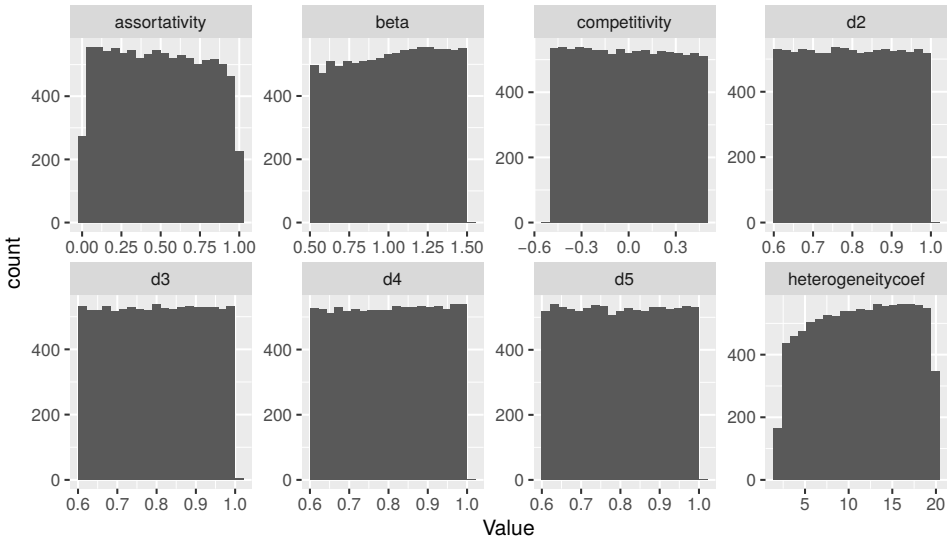

Fig S2. Prior distributions for all the parameters. The same priors are used to generate target datasets and training datasets.

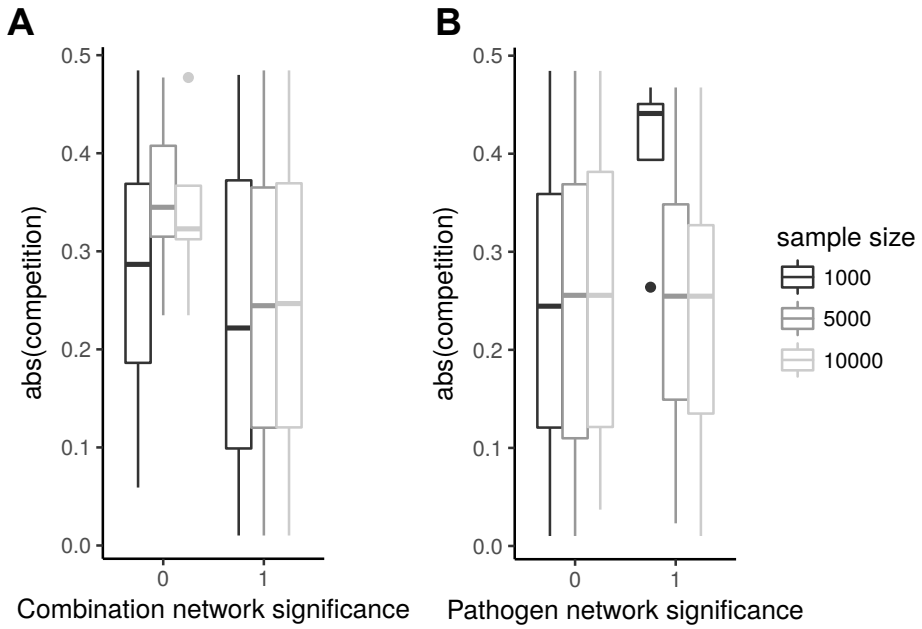

Fig S3. Difference in interaction intensity depending on the p-value of the network-based test. A) If the combination network test is non significant, the interaction is likely to be strong. B) The difference for the pathogen network in the small sample size scenario is explained by the rarity of significant tests.

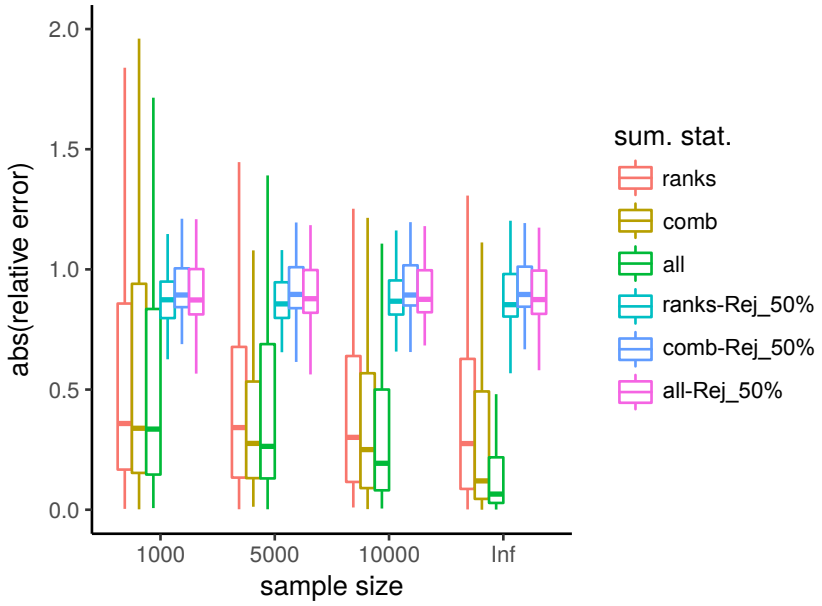

Fig S4. Relative error depending on the summary statistics and the methods used. The regression part of the ABC improves the inference compared to the rejection step alone.

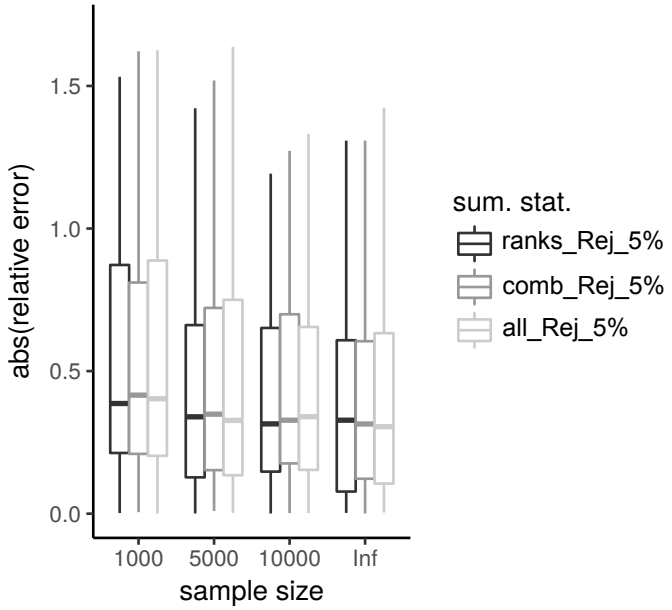

Fig S5. Rejection-based inference. When ignoring the regression part of the ABC, the set of summary statistics has little effect on the quality of the fit.

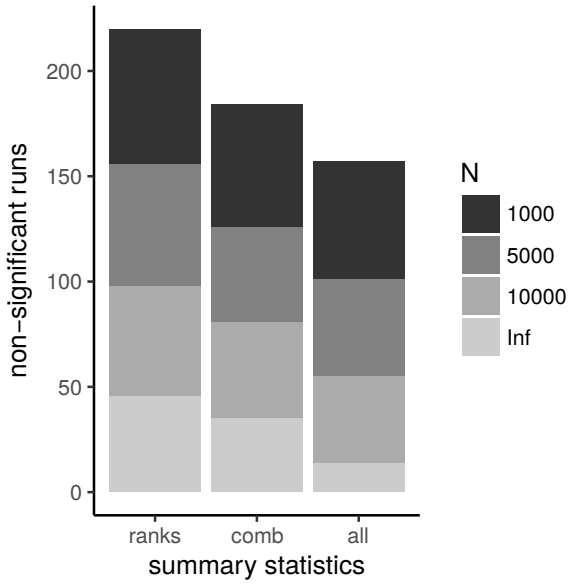

Fig S6. Number of runs where 0 cannot be excluded from the 95% HPD. Increasing the sample size and the number of summary statistics decreases the number of such non-significant runs.