bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Inference of coevolutionary dynamics and parameters from host and parasite polymorphism data of repeated experiments

Hanna Märkle[a,*], Aurélien Tellier[a]

[a]*Section of Population Genetics, TUM School of Life Sciences Weihenstephan, Technical University of Munich, Liesel-Beckmann-Str. 2, 85354 Freising, Germany*

Abstract

There is a long-standing interest in understanding host-parasite coevolutionary dynamics and associated fitness effects. Increasing amounts of genomic data offer new ways to understand the past coevolutionary history. To extract such information, it is crucial to understand the link between allele frequency dynamics and polymorphism data. We couple coevolutionary models, which include costs of resistance, infectivity and infection, with coalescent simulations to generate polymorphism data at the involved loci. We show that under trench-warfare dynamics the allele frequencies at the internal equilibrium point determine the strength of the resulting balancing selection signatures.

As a proof-of-principle, we then apply an Approximate Bayesian Computation approach to infer the cost values using host and parasite polymorphism data from repeated experiments. First, we demonstrate that the cost of infection and host and parasite population sizes can be inferred knowing the costs of resistance and infectivity. Second, we can infer simultaneously all three costs when population

*Corresponding author

Email addresses: hanna.maerkle@tum.de (Hanna Märkle), tellier@wzw.tum.de (Aurélien Tellier)

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

sizes are known. Third, the polymorphism data in the host are informative about the cost of infectivity (parasite cost), while the signatures in the parasite inform about the cost of resistance and infection (host costs). We discuss the implications of our results for genomic based inference of host-parasite coevolution.

Keywords: Host-parasite coevolution, polymorphism data, ABC, fitness cost

1. Introduction

Host-parasite coevolution is an ubiquitous process and has been demonstrated in terrestrial (Thrall et al., 2012), limnological (Decaestecker et al., 2007) and marine environments (Martiny et al., 2014). It describes the process of parasites and hosts excerting reciprocal selective pressures on one another. A necessary condition for coevolution to take place is some underlying heritable variation for the traits being involved into the coevolutionary interaction, such as resistance or infectivity. Irrespectively of the underlying genomic architecture, coevolution and the associated reciprocal selective pressures result in allele frequency changes at the underlying genes. One central question in host-parasite coevolution is how allele frequencies change over time (directional changes, regular fluctuations, chaotic fluctuations) and the resulting short and long-term patterns of allelic polymorphism (transient or stable polymorphism). Allele frequency dynamics are usually classified based on a continuum ranging from arms-race dynamics on the one extreme and trench-warfare dynamics on the other extreme (Brown and Tellier, 2011; Woolhouse et al., 2002). In arms race dynamics frequencies of new beneficial alleles (such as new resistance alleles, infectivity alleles) arising by *de novo* mutations increase towards fixation in both inteacting partners. Accordingly, alleles are re-

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

currently replaced and thus, are short lived and polymorphism is only transient. Such dynamics have been e.g. demonstrated in the intial phase of *Pseudomonas fluorescens* coevolving with a phage (Hall et al., 2011). Opposed to arms-race are trench-warfare dynamics (Stahl et al., 1999; Woolhouse et al., 2002), where several alleles segregate simultaneously and are stably maintained in the host and the parasite. Here, allele frequencies either converge towards a stable equilbrium or they fluctuate persistently over time, both dynamics resulting in long-term maintenance of allelic polymorphism.

The type of occuring dynamics depends on the type and strength of various forms of selection (negative indirect frequency dependent selection, negative direct frequency dependent selection, overdominant selection) and their interplay with genetic drift. In frequency dependent selection the strength of selection depends on the frequency of particular alleles. Negative indirect frequency dependent selection (niFDS) takes place when the fitness of a particular host alleles increases with decreasing frequencies of a particular allele in the parasite and vice-versa (Seger, 1988; Tellier and Brown, 2007). When the fitness of a host or parasite allele decreases with its own frequency negative direct frequency-dependent selection (ndFDS) is acting. ndFDS can be promoted by factors such as asynchrony between host and parasite life-cycles (overlapping parasite generations, several parasite generation per host generation) or epidemiological feedback due to density dependent disease transmission (Brown and Tellier, 2011). Overdominant selection or some form of ndFDS are a necessary but not sufficient condition for trench-warfare dynamics to take place in single locus host-parasite coevolutionary interactions (Tellier and Brown, 2007). Even with some form of ndFDS acting, arms-race dynamics can take place if either the strength of ndFDS compared to

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

niFDS is weak or genetic drift is causing random loss of alleles.

The exact nature of the dynamics, such as the equilibrium frequencies of alleles and the period and amplitude of coevolutionary cycles, is further affected by the way host and parasite genotypes interact at the molecular level and the fitness costs associated with the coevolutionary interaction. The interaction at the molecular level is captured by the infection matrix which stores the level of infection in all possible pairwise interactions between host and parasite genotypes. One well studied type of interaction is the gene-for-gene (GFG) interaction which presents one endpoint of a continuum of infection matrices (Agrawal and Lively, 2002). GFG-interactions are characterized by one universally infective parasite genotype and one universally susceptible host type. Such interactions have been found for example in the Flax-*Melampsora lini* system (Flor, 1971).

A fitness cost which has been shown to crucially affect the coevolutionary dynamics is the loss in fitness due to infection (Tellier and Brown, 2007; Tellier et al., 2014). In addition, costs of resistance such as reduced competitive ability or fertility in absence of the parasite (Kraaijeveld and Godfray, 1997; Bergelson and Purrington, 1996; Lenski, 1988) and costs of infectivity such as reduced spore production of infective pathogens (Thrall and Burdon, 2003) can further alter the dynamics. These costs also determine the equilibrium frequencies of the coevolutionary system (Leonard, 1994; Frank, 1992).

Understanding the link between allele frequency dynamics at the coevolving loci under the joint influence of coevolution and genetic drift and the resulting genomic signatures at linked neutral sites is important for the application and development of inference methods. The classic expectation is that arms-race dynamics result in selective sweep signatures which are characterized by lower genetic diversity

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

at neutral sites being linked to the coevolving locus compared to the genome-wide average and increased levels in linkage disequilibrium (Maynard Smith and Haigh, 1974). Trench-warfare dynamcics on the other hand are expected to result in signatures of balancing selection being characterized by higher than average diversity due to the maintenance of several alleles for a long period of time (Charlesworth et al., 1997). However, trench-warfare dynamics do not necessarily result in observable genomic signatures in both interacting partners (Tellier et al., 2014).

Therefore, we aim to investigate explicitely the link between expected dynamics in infinite population size and the resulting signatures around the coevolving loci in the host and the parasite. We can show that the resulting coevolutionary signatures in both interacting partners change with the equilibrium frequencies of alleles.

Accordingly, as a proof-of principle, we aim to infer information about key parameters from polymorphism data at the host and parasite coevolving loci. Therefore, we apply Approximate Bayesian Computation (Csillery et al., 2010; Beaumont et al., 2002; Sunnaker et al., 2013) based on a coevolutionary model and the assumption that data are available from repeated coevolutionary experiments. We test this approach on pseudo-observed data sets which we obtained by repeatedly simulating a coevolutionary experiment for $r = 200$ or $r = 10$ times. We are able to extract information about the cost of infection, a parameter substantially shifting the equilibrium frequencies, with good accuracy. The inference works best when polymorphism data of both the host and the parasite are available. However, this requires that some information about other parameters such as population sizes or cost of resistance are available.

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

2. Methods and Material

2.1. Simulation of polymorphism data

We model coevolution between a single host and a single parasite species. The coevolutionary interaction is driven by a single bi-allelic major gene in each haploid species. Hosts are either resistant (*RES*) or susceptible (*res*) and parasites are either non-infective (*ninf*) or infective (*INF*). Thus, the model follows a gene-for-gene interaction with the following infection matrix:

$$\begin{array}{cc} & \begin{array}{cc} ninf & INF \end{array} \\ \begin{array}{c} RES \\ res \end{array} & \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}. \end{array} \tag{1}$$

A 1-entry in the infection matrix indicates that the parasite is able to infect the host and a 0-entry indicates that the host is fully resistant towards the parasite. To obtain polymorphism data at these major genes we combine a forward-in-time coevolutionary model (Tellier and Brown, 2007) including genetic drift and recurrent mutations with backward-in-time coalescent simulations (Tellier et al., 2014).

2.1.1. Forward in time coevolution model

In the forward part, we obtain the frequencies of the different alleles at the beginning of each discrete host generation g in three steps:

1. Using a discrete-time gene-for-gene coevolution model (shown below) we

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

compute the expected allele frequencies in the next generation (in infinite population size)

2. We incorporate genetic drift by performing a binomial sampling based on the frequency of the *RES*-allele (*INF*-allele) after selection and the finite and fixed haploid host population size N_H (parasite population size N_P).
3. We allow for recurrent allele mutations to take place and change genotypes from *RES* to *res* at rate μ_{Rtor} or *res* to *RES* at rate μ_{rtoR} in the host and from *ninf* to *INF* at rate μ_{ntoI} and from *INF* to *ninf* at rate μ_{Iton} in the parasite. Henceforward, we call such mutations as functional mutations. We set all functional mutation rates to $\mu_{Rtor} = \mu_{ntoI} = \mu_{rtoR} = \mu_{Iton} = 10^{-5}$.

Repeating this procedure for g_{max} host generations, we obtain the so called frequency path, which summarizes the allele frequencies at both loci forward in time.

The coevolution model (henceforward termed **model A**) is based on the polycyclic auto-infection model in Tellier and Brown (2007). It incorporates fitness costs for the *RES*-allele (c_H) and the *INF*-allele (c_P). There are $T = 2$ discrete parasite generations per discrete host generation g and disease transmission is frequency-dependent. Successive parasite generations t ($t > 1$) within host generation g infect the same host as their parent at rate $\psi = 1$ (auto-infection rate). Once a host becomes infected in parasite generation t it stays infected until it dies from natural death at the end of generation g. The total amount of fitness loss for a host which has been infected in parasite generation t within host generation g is s_t. s_t is a function of the number of parasite generations a host is infected and the cost of being infected s (cost of infection) for a whole host generation:

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

$$s_t = s \cdot \left(\frac{T - (t-1)}{T} \right) \tag{2}$$

The allele frequencies of resistant hosts (R_g), susceptible hosts (r_g), non-infective parasites ($A_{g,t}$) and infective parasites ($a_{g,t}$) are given by the following recursive equations (for the corresponding fitness matrices see Tab. S1):

$$a_{g,2} = \frac{a_{g,1} \cdot (1 - c_P)}{a_{g,1} \cdot (1 - c_P) + A_{g,1} \cdot r_g} \tag{3a}$$

$$a_{g+1,1} = \frac{(1 - c_P) \cdot \left[R_g \left(A_{g,1} a_{g,2} + a_{g,1} \right) + r_g a_{g,1} \right]}{(1 - c_P) \cdot \left[R_g \left(A_{g,1} a_{g,2} + a_{g,1} \right) + r_g a_{g,1} \right] + r_g A_{g,1}} \tag{3b}$$

$$R_{g+1} = \frac{R_g \cdot (1 - c_H)(A_{g,1} A_{g,2} + A_{g,1} a_{g,2}(1 - s_2) + a_{g,1}(1 - s_1))}{R_g \cdot (1 - c_H)(A_{g,1} A_{g,2} + A_{g,1} a_{g,2}(1 - s_2) + a_{g,1}(1 - s_1) + r_g(1 - s_1)} \tag{3c}$$

with $A_{g,t} = 1 - a_{g,t}$ and $r_g = 1 - R_g$. The equilibrium frequencies $\hat{a}$, $\hat{R}$ (Tellier and Brown, 2007) at the internal, non-trivial equilibrium point are approximately given by:

$$\begin{aligned} \hat{a} &\approx \frac{s_2 + s_1 - \sqrt{(s_2 + s_1)^2 - 4 s_2 (s_1 - c_H)}}{2 s_2 (1 - c_H)} \\ \hat{R} &\approx \frac{c_P}{2 - c_P - \hat{a}} \\ &\approx \frac{2 c_P \cdot s_2 \cdot (1 - c_H)}{s_2(3 - 4c_H - 2c_P(1 - c_H)) - s_1 + \sqrt{(s_2 + s_1)^2 - 4 s_2 (s_1 - c_H)}} \end{aligned} \tag{4}$$

For investigating the link between coevolutionary dynamics in infinite population size and genomic signatures in finite population size, we further use these recursion equation to simulate allele frequency trajectories in infinite population size for $g_{max} = 30,000$ generations (without genetic drift and recurrent mutations) for all pairwise combinations of $s = \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$, $c_P = \{0.1, 0.3\}$ and $c_H = \{0.05, 0.1\}$.

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Additionally, we use two extensions, **B** and **C** respectively, of the basic model to check for the generality of our results. In model **B**, we extend model **A** to more than two parasite ($T > 2$) generations per host generation g (Eq. S5). Model **C** extends model **A** (keeping $T = 2$) by allowing for allo-infections to take place at rate $(1 - \psi)$ in the second parasite generation ($t = 2$) within host generation g (Eq. S6).

2.1.2. Backward in time coalescent

To obtain polymorphism data for neutral sites being linked to the coevolving loci we combine the obtained frequency paths which include genetic drift and recurrent mutations with coalescent simulations separately for the host and the parasite. Therefore, we first rescale time for the frequencies paths appropriately (for more information see S1.1.1). Based on these time-rescaled frequency paths we launch a modified version of msms (Ewing and Hermisson, 2010; Tellier et al., 2014) once for each species. We set the sample size to $n_H = 50$ for the host ($n_P = 50$ for the parasite). For both species we assume a non-recombining locus of length 2500 bp and a per site neutral mutation rate of 10^{-7}. Accordingly, the neutral population mutation rate is $\theta_H = 2 \cdot N_H \cdot 2500 \cdot 10^{-7}$ for the host ($\theta_P = 2 \cdot N_P \cdot 2500 \cdot 10^{-7}$ for the parasite). Based on the respective msms-output we calculate eight summary statistics for each species which are based on the site frequency spectrum (SFS) (Tab. S5). In addition to these 16 summary statistics we calculate an additional summary statistic (**P**airwise **M**anhattan **D**istance) which is combining information from host and parasite polymorphism data (see S2).

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

2.2. *Generating pseudo-observed data sets (PODs)*

To understand the link between coevolutionary dynamics in finite population size and genomic signatures we simulate pseudo-observed data sets (PODs) for various costs of infection (s). This parameter strongly influences the coevolutionary dynamics, namely the allele frequencies at the non-trivial equilibrium point and the stability of the internal equilibrium point (Fig. 1 and Tab. S3, Tellier and Brown, 2007). Therefore by changing s, we can investigate a continuum between arms-race and trench-warfare dynamics with varying internal equilibrium frequencies. We vary s from 0.2 to 0.8 in steps of 0.1. We simulate $r = 200$ repetitions for each value of s using the above mentioned forward-backward approach, and afterwards average the summary statistics across the $r = 200$ repetitions/per parameter combination/per model. For the so called standard case we fix the parameters as follows: $c_H = 0.05$, $c_P = 0.1$, $N_H = N_P = 10,000$, $n_H = n_P = 50$. We extend this standard case in two ways. First, we simulate data for various combinations of host population size $N_H = (5,000;\ 10,000;\ 15,000)$ and parasite population size $N_P = (5,000;\ 10,000;\ 15,000)$. Second, we assess the signatures for combinations of $c_H = (0.05,\ 0.1)$ and $c_P = (0.1,\ 0.3)$ while fixing the population sizes to their standard values.

2.3. *Performing ABC on the pseudo-observed data sets for model A*

As a proof of principle we aim to infer different parameters determining the coevolutionary dynamics in Model **A**. In scenario 1, we aim to infer simultaneously the cost of infection (s), the host population size (N_H) and the parasite population size (N_P) assuming that we know the true cost of resistance c_H and the true cost

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

of infectivity c_P. In scenario 2, our goal is to infer simultaneously the cost of infection (s), the cost of infectivity (c_P) and the cost of resistance (c_H) assuming we know the true host (N_H) and parasite population sizes (N_P). In doing so we further test for the effect of the number of repetitions on the inference results. Thus, we base our inference on the average summary statistics of $r = 200$ and $r = 10$ repetitions. Besides the effect of the number of repetitions, we access how the type of polymorphism data available affects the accuracy of inference. Therefore, we perform inference based on a) polymorphism data of both, the host and the parasite, b) polymorphism data of the host and c) polymorphism data of the parasite. For the sampling step of the ABC we use the ABCsampler from ABCtoolbox (Wegmann et al., 2010) and perform 100,000 simulations using the standard sampler. The chosen priors, complex parameters and fixed parameters can be found in Tab. S6. For the estimation step we retain the 1% best simulations and apply the post-sampling adjustment (generalized linear model) as implemented in ABCestimator (Wegmann et al., 2010). All codes and pipelines used are available upon request (and will be placed on a Dryad repository).

3. Results

3.1. Link between coevolutionary dynamics and sequence data

The internal equilibrium frequency of the *RES*-allele mainly increases with increasing cost of infectivity (c_P) (Fig. 1 a+b vs. Fig. 1 c+d), increases very slightly with increasing cost of infection (s) and remains almost unaffected by changing costs of resistance (c_H) (Fig. 1 a+c vs. Fig. 1 b+d). The opposite is true for

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

the parasite. Here, the equilibrium frequency of the infective (*INF*)-parasite rises mainly with increasing cost of infection (s) (Fig. 1). Higher costs of resistance (c_H) decrease the equilibrium frequency of *INF*-parasites (Fig. 1 a+c vs. Fig. 1 b+d) for a given value of s. In contrast to the host, the equilibrium frequencies in the parasite are almost unaffected by changes in the cost of infectivity (c_P). For high costs of infection s the dynamics are always switching to arms-race dynamics, irrespectively of the underlying costs of resistance and infectivity (Fig. 1, Fig. S7).

The changes in equilibrium frequencies with changing cost of infection (s), cost of resistance (c_H) and changing cost of infectivity (c_P) are reflected by the resulting genomic signatures at the coevolving genes (Fig. 2). Generally, the strongest signatures of balancing selection (here indicated by high Tajima's D values) can be observed when the equilibrium frequencies of *INF*-parasites or *RES*-hosts are close to 0.5 (see Fig. 1, Tab. S3, Fig. 2). The strength of the signatures declines the further the equilibrium frequencies move away from 0.5.

The genomic signature in the parasite changes strongly with changing cost of infection (s), irrespectively of c_H and c_P. Further, the resulting genomic signatures in the parasites for a given cost of infection s are distinguishable for different costs of resistance but not for different costs of infectivity.

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

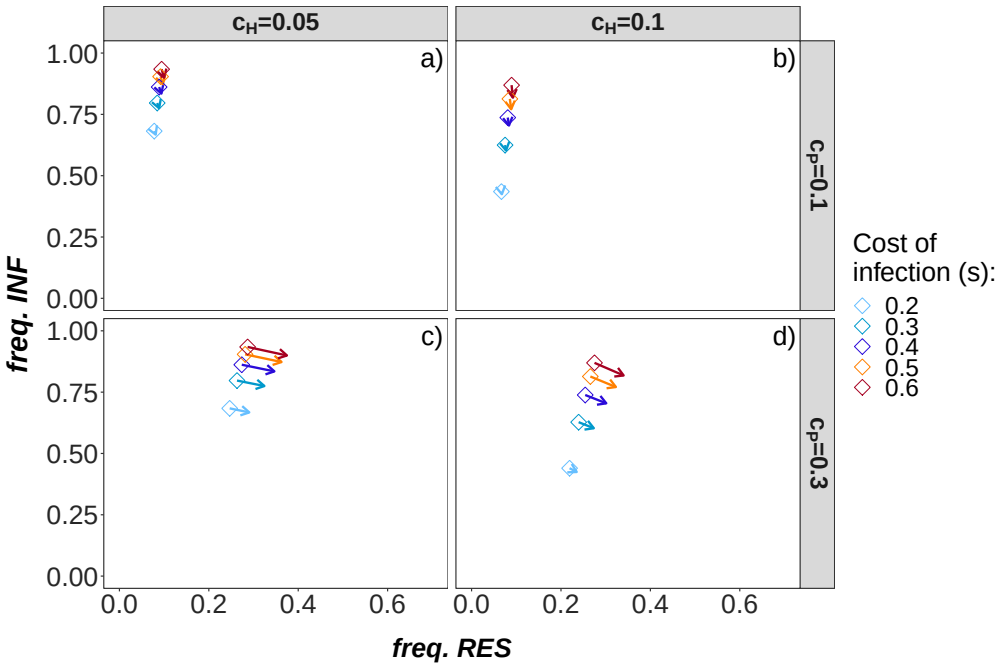


Figure 1: Deterministic equilibrium frequencies for model A (pure autoinfection model with $T = 2$ parasite generations) for different combinations of cost of resistance $c_H = (0.05, 0.1)$ (columns), cost of infectivity $c_P = (0.1, 0.3)$ (rows) and cost of infection $s = (0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8)$ (color of the squares). Only combinations with trench-warfare dynamics are shown. Centres of the squares represent the equilbrium frequencies obtained by simulating numerically the recursion equations Eq. 3 for 30,000 generations starting with an initial frequency of $R_0 = 0.2$ resistant hosts and $a_0 = 0.2$ infective parasites. Heads of the arrows represent the equilibrium frequencies based on Eq. 4 which slightly differ from the numerical computations due to analytical approximations.

The genomic signature in the host changes very slightly with increasing cost of infection (s) as the resepective equilbrium frequencies are strongly affected by c_P. Thus, the strongest balancing selection signature for the host is found for an increased costs of infectivity ($c_P = 0.3$) and intermediate costs of infection.

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

The combination of host and parasite signatures holds the highest information content about the fitness parameters guiding the coevolutionary dynamics. This is due to the fact that the equilibrium frequencies in the host and parasite are differentially affected by these fitness parameters. The genomic signature in the host is mainly indicative about the cost of infectivity (c_P), a cost which is affecting the parasite fitness, whereas the signature in the parasite is mainly informative about the costs of resistance (c_H) and infection (s), parameters with a direct fitness effect in the host (Fig. 2). This results from the action of niFDS. Increasing costs of resistance (c_H) disfavor resistant hosts. Thus, the frequency of infective parasites is decreasing which results in lower equilibrium frequencies of *INF*-parasites. The opposite is true for increasing costs of infectivity (c_P). This cost reduces the fitness of *INF*-parasites which in turn favors *RES*-hosts compared to *res*-hosts.

The qualitative changes of the genomic signatures for changing costs of infection in the standard case remain similar even when population sizes differ in both interacting partners (Fig. 3). However, the strength of genomic signatures is affected by the population sizes. The strongest signature of balancing selection in the parasite is found when the parasite population size is small compared to the host population size (Fig. 3c). Here, the large host population size reduces the amount of genetic drift in the host. Thus, there are less allele frequency fluctuations in the host around the internal equilibrium point. This in turn, also reduces allele frequency fluctuations in the parasite.

Overall, there is a strong link between the equilibrium frequencies under trench-warfare dynamics and the resulting genomic signatures. We obtain similar results when we slightly modify the assumptions about the coevolutionary interaction by either a) extending the model to more than two parasite generations per host gen-

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

eration (**Model B**, Fig. S1, Fig. S3cd) or b) allowing for allo-infections at rate $1 - \psi$ in the second parasite generation within host generation g (**Model C**, Fig. S2, Fig. S3ab). Increasing the number of parasite generations extends the parameter space in which trench-warfare dynamics occur. Here, the strongest signatures of balancing selection are found for intermediate costs of infection (Fig. S3). Allowing for allo-infections, decreases the parameter space in which trench-warfare dynamics take place and thus, also the range in which balancing selection signatures can be observed in both interacting partners (Fig. S3ab).

The strong link between equilibrium frequencies and resulting genomic signatures can be explained in terms of a structured coalescent tree. The coalescent tree in both coevolving species consists of two demes (*RES* and *res* for the host and *INF* and *ninf* for the parasite). As we assume that the neutral sites are fully linked to the coevolving locus neutral mutations are usually linked to the allele in which they arose, unless a functional mutation is taking place and accordingly, one lineage is migrating from one deme to the other. When the frequencies of both allles (*ninf* and *INF* in the parasite or *RES* and *res* in the host) are fairly similar they have equal contributions to the sample. Thus, the underlying coalescent tree is well balanced. Accordingly, we observe an excess of intermediate frequency variants in the SFS. As the equilibrium frequencies move away from 0.5, the average sample configuration changes and the coalescent tree becomes less balanced (see Fig. S7 a-c and p-r). Therefore, the number of SNPs at intermediate frequencies drops and Tajima's D decreases (Fig. 2).

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

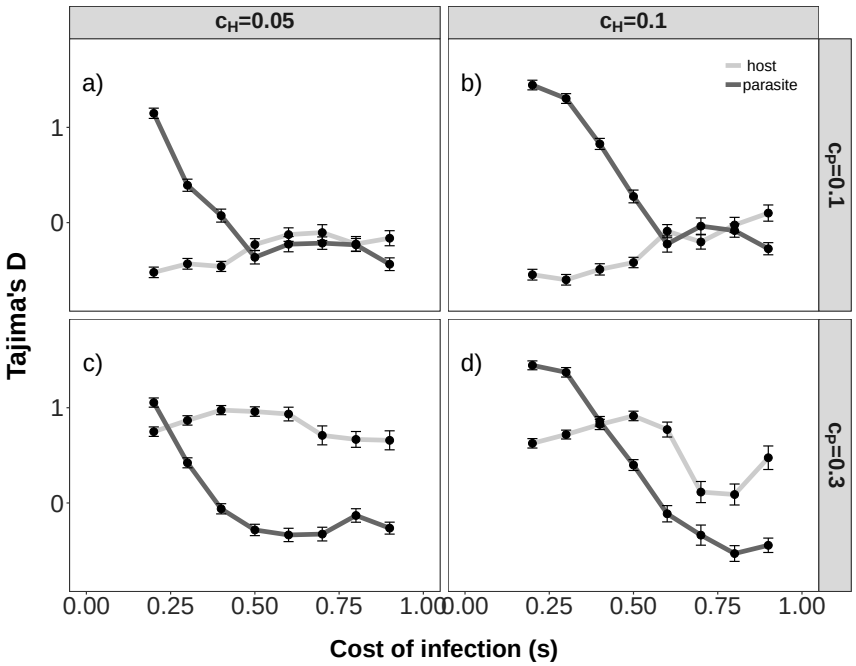


Figure 2: Tajima's D (y-axis) for model A for various cost of infection s (x-axis). The results are shown for different combinations of c_P ($c_P = 0.1$ top, $c_P = 0.3$ bottom) and c_H ($c_H = 0.05$ left, $c_H = 0.1$ right). The mean and standard error of Tajima's D of the parasite population (dark grey) and of the host population (light grey) are plotted for $r = 200$ repetitions. The other parameters are fixed to: $N_H = N_P = 10,000$, $n_H = n_P = 50$, $\theta_H = \theta_P = 5$, $\mu_{Rtor} = \mu_{rtoR} = \mu_{ntoI} = \mu_{Iton} = 10^{-5}$.

3.2. *Inference of coevolutionary dynamics from polymorphism data*

Our results clearly indicate that it is possible to infer the cost of infection using polymorphism data from the host and parasite (Fig. 4, Fig. 5). The accuracy of inference mainly depends on four factors being 1) the true value of the cost of infection, 2) the type of polymorphism data being used (host and parasite together,

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

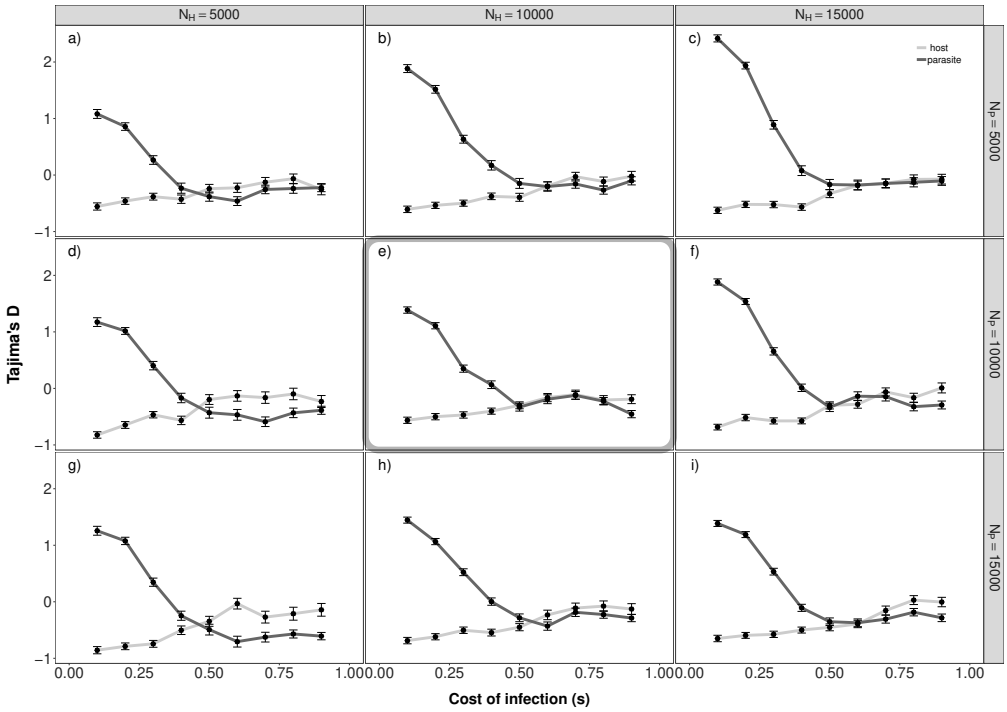


Figure 3: Tajima's D (y-axis) for **Model A** for various cost of infection s (x-axis) and different combinations of N_P ($N_P = 5,000$ top, $N_P = 10,000$ middle, $N_P = 15,000$ bottom) and N_H ($N_H = 5,000$ left, $N_H = 10,000$ middle, $N_H = 15,000$ right). The mean and standard error of Tajima's D of the parasite population (dark grey) and of the host population (light grey) are plotted for $r = 200$ repetitions. Note that subfigure e corresponds to Fig. 2a. The other parameters are fixed to: $c_H = 0.05$, $c_P = 0.1$, $\theta_H = N_H/2000$, $\theta_P = N_P/2000$, $n_H = n_P = 50$, $\mu_{Rtor} = \mu_{rtoR} = \mu_{ntoI} = \mu_{Iton} = 10^{-5}$.

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

only host or only parasite), 3) the number of available repetitions and 4) the type of known parameters.

Inferences of the cost of infection and of the population sizes are the most accurate if the number of repetitions is high ($r = 200$), and host and parasite polymorphism data are both available (Fig. 4, Fig. S4, Fig. S5). Generally, inference for Scenario 1 works best if host and parasite data are used together, irrespectively of the number of repetitions available (compare Fig. 4a to Fig. 4b+c; Fig. 4d to Fig. 4e+f). Using parasite polymorphism data only is also quite powerful for small to intermediate values of the cost of infection ($s < 0.6$) (Fig. 4c+f) where trench-warfare dynamics take place and SFS of the parasite changes pronouncedly with s (Fig. S7). In contrast, using only host polymorphism shows markedly less power in the same parameter range (Fig. 4b+e), especially if the number of available repetitions is low. For low costs of infection the respective equilibrium frequencies of the *RES*-genotype are close to zero (< 0.1) and increase only very slightly when s increases (Tab. S3). Thus, the host sample mostly consists of polymorphism data from *res*-hosts. Accordingly, the coalescent tree consists of a very large subtree containing the *res*-samples and a very small subtree containing the *RES*-samples and the overall tree looks almost neutral (Fig. 2). The power of estimating the cost of infection using only host information diminishes in the transition between trench-warfare and arms-race dynamics (around $s \approx 0.6$), especially if the number of repetitions is low ($r = 10$). In this range, fixation of alleles in both species can either happen due to genetic drift or due to the inherent dynamics of the coevolutionary interaction. This effect decreases the accuracy of parameter estimation even if host and parasite polymorphism data are available (Fig. 4a+d, Fig. 2).

The results clearly indicate that the availability of more repetitions increases the

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

accuracy of inference (compare Fig. 4a-c to Fig. 4d-f). There are three sources of stochasticity affecting the neutral polymorphism around the coevolutionary loci: 1) The effect of genetic drift on the allele frequency trajectory under coevolution, 2) the stochasticity in the coalescent process for a given allele frequency trajectory and 3) the stochasticity in the neutral mutation process on top of the coalescent process. As the first type of stochasticity affects the 'population' sizes of the functional alleles in the host (in the parasite) over time it also has a subsequent effect on the other two sources of stochasticity. Using data from several repetitions allows to better handle and to average out the effect of genetic drift on the variability of the allele frequency path and its subsequent effect on the observed summary statistics. This is especially helpful in the range of parameter values where the dynamics switch from arms-race to trench-warfare.

Like in scenario 1 inference for scenario 2 works best if data from both the host and the parasite are available for a large amount of repetitions $r = 200$. However, the accuracy of inference for the cost of infection s is generally not as accurate as in scenario 1. Simultaneous inference of all three parameters in scenario 2 is most accurate for intermediate costs of infection and if both, host and parasite polymorphism data, are available. This is due to the fact that signatures in the host and the parasite are differentially affected by the various costs (Fig. 2).

Inference of the cost of resistance (c_H) works reasonably well if polymorphism data only from the parasite are available. However, this comes at the cost of less accurate inference of the cost of infection (s) as both paramters are affecting the equilibrium frequency in the parasite (Fig. 4, Fig. 5). While the equilibrium frequency of INF-parasites increases with increasing cost of infection, an increase in the cost of resistance for a fixed cost of infection (s) decreases the respective

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

equilibrium frequency. Thus, overestimating the cost of infection (s) can be compensated by overestimating the cost of resistance (c_H) simultaneously. This effect can be seen for low costs of infection (s) if only the information from the parasite polymorphism data is used in Scenario 2 (see Fig. 5 c+f). In contrast, inference of the cost of infectivity (c_P) works reasonably well if polymorphism data only from the host are available. This is due to the fact that changing costs of infectivity (c_P) mainly affect the equilibrium frequencies in the host but not in the parasite (Fig. 1). Therefore, inference of this parameter does not work if only parasite polymorphism data are available. All the above mentioned effects explain why the simultaneous inference of several cost becomes less accurate with less ($r = 10$) repetitions (Fig. S6).

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

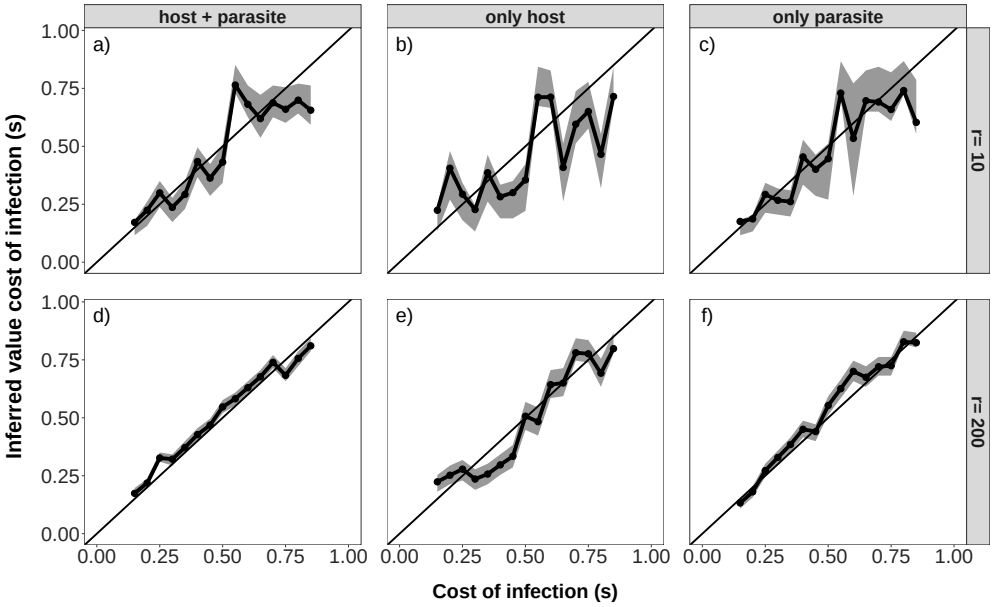


Figure 4: Median of the posterior distribution (y-axis) for the cost of infection s compared to the true value (x-axis) for $r = 10$ (top, a-c) and $r = 200$ (bottom, d-f). The inference results for scenario 1 based on host and parasite polymorphism data (left, a+d), host polymorphism data only (middle, b+e) and parasite polymorphism data only (right, c+f) are shown. The chosen parameters are: $c_H = 0.05$, $c_P = 0.1$, functional mutation rate 10^{-5}, $c_H = 0.05$, $c_P = 0.1$, $g_{max} = 30,000$, $\theta_H = 5$, $\theta_P = 5$, $n_H = 50$ and $n_P = 50$. Priors have been chosen as follows: N_H log uniform$(2000, 40000)$, N_P log uniform$(2000, 40000)$, s uniform $(0.1, 0.9)$. s, N_H and N_P are inferred simultaneously for all plots.

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

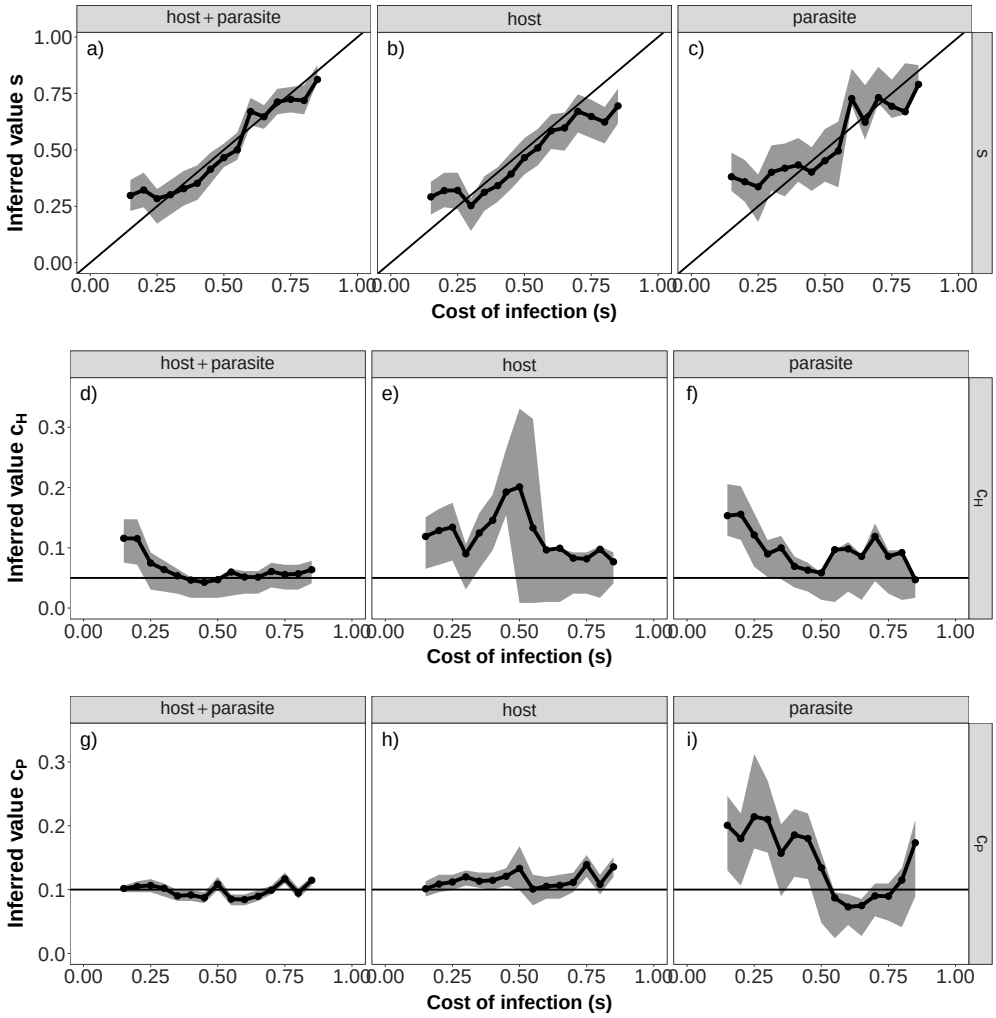


Figure 5: Median of the posterior distribution (y-axis) for the cost of infection s (a-c), cost of resistance c_H (d-f) and cost of infectivity c_P (g-i) compared to the true value (x-axis) for $r = 200$. Inference results for scenario 2 are based on host and parasite polymorphism data (left, a+d+g), host polymorphism data only (middle, b+e+h) and parasite polymorphism data only (right, c+f+i). The chosen parameters are: $N_H = N_P = 10,000$, functional mutation rate 10^{-5}, $g_{max} = 30,000$, $\theta_H = 5$, $\theta_P = 5$, $n_H = 50$ and $n_P = 50$. Priors have been chosen as follows: s uniform$(0.1, 0.9)$, c_H uniform $(0.01, 0.35)$, c_P uniform$(0.01, 0.35)$. s, c_H and c_P are inferred simultaneously for all plots.

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

4. Discussion

We could establish the link between coevolutionary dynamics (Fig. 1), the resulting genomic signatures (Fig. 2, Fig. 3) and subsequentially the amount of information about the underlying coevolutionary dynamics which can be extracted from genomic signatures at the coevolving loci (Fig. 4, Fig. 5). Our results indicate that under trench-warfare dynamics the allele frequencies at the non-trivial internal equilibrium point affect the strength of genomic signatures at the coevolving genes in both, the host and parasite. We further could show as a proof of principle that it is possible to infer information about parameters underlying the coevolutionary interaction from polymorphism data at the genes under coevolution if some relevant parameters such as diverse costs (Fig. 4) or population sizes (Fig. 5) are known. This is due to the fact that various parameter combinations can give rise to similar equilibrium frequencies and thus, result in undistinguishable genomic signatures. In general, inference works best if polymorphism data from both the host and the parasite are available from repeated experiments.

As already shown in Tellier et al. (2014) the link between coevolutionary dynamics in finite population and resulting signatures at the coevolving loci is not always black (arms-race result in selective sweep signatures) and white (trench-warfare dynamics in balancing selection signatures) but follows a continuum of outcomes. The strength of the genomic signatures under trench-warfare dynamics is a result of the internal equilibrium frequencies, the fluctuations around these equilbrium frequencies, the amount of genetic drift in both partners and the proximity of these equilibrium frequencies to the fixation boundaries. When equilibrium frequencies are close to boundaries, alleles can be easily lost by drift and thus, arms-race dy-

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

namics take place although trench-warfare dynamics would be predicted based on the model. In such cases signatures of balancing selection cannot be observed.

The found links between dynamics in infinite population size and genomic signatures in finite population size have several implications. Model based inference of parameters governing the coevolutionary dynamics is possible if they substantially shift the equilibrium frequencies of the dynamics and thus, the resulting genomic signatures. In cases where different parameters shift the equilibrium frequencies along the same axis, three different inference scenarios are possible. First, it is only possible to infer a compound parameter if there is no *a priori* information available. This is illustrated by the inference results for scenario 2 when only parasite polymorphism data are available (Fig. 5). Here, overestimating the cost of infection s compensates for overestimating the cost of resistance c_H. Second, if some of these parameters are known *a priori*, the other parameters can be inferred conditional on this information. Third, the parameters have different effects on the equilibrium frequencies in the host and parasite and, thus, combining host and parasite polymorphism data allows to infer the different parameters simulataneously.

For many host-parasite models (including the one used here) it has been shown that the equilibrium frequencies in the host are substantially or exclusively affected by fitness penalties applying to the parasite and vice-versa. Thus generally speaking, the strength of genomic signatures in either species are presumably most indicative about processes affecting the coevolving partner. We therefore speculate, that the balancing selection signatures which have been found at R-genes in *Arabidopsis thaliana* (Stahl et al., 1999; Bakker et al., 2006) (Karasov et al., 2014), *Solanum sp.* (Rose et al., 2007; Hoerger et al., 2012; Caicedo and

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Schaal, 2004), *Phaseolus vulgaris* (De Meaux et al., 2003), *Capsella* (Gos et al., 2012), are indicative about the selective pressure in the coevolving parasite or parasite community. Conversely, the long term maintenance of strains in *P. syringae* (Karasov et al., 2018) could reflect fitness costs in *A. thaliana*.

Further, we have shown that the genomic signatures might be rather weak and almost undistinguishable from neutral signatures if the the internal equilibrium frequencies are close to fixation. In such cases it is very likely that genes under coevolution are missed when applying classic outlier scan methods.

In general, our results should not be restricted to the used coevolution model (see Appendix). We acknowledge that we assumed the most simple type of coevolutionary interaction possible. However, understanding the link between dynamics, signatures and resulting accuracy of inference is a useful starting point to develop a further and deeper understanding when several major genes are involved into the coevolutionary interaction. There are various other coevolution models with respect to the biology of the coevolving species or the ecology of the disease which have been shown to result in trench-warfare dynamics. Nevertheless as long as the coevolutionary interaction is driven by a single bi-allelic locus in each species, the resulting equilbrium frequencies will be always confined to the 2-dimensional plane and a limited amount of possible genomic signatures (see Fig. S1, Fig. S1 and Fig. S3). Therefore, our findings should also apply to coevolutionary epidemiology models such as in Ashby and Boots (2017); Gokhale et al. (2013).

So far we did not take population size changes and the resulting temporal variation in the amount of genetic drift into account. In host-parasite coevolution, population size changes can be due to two different sources: 1) Population size changes which are independent of the coevolutionary interaction and 2) popula-

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

tion size changes which arise as an immediate result of coevolutionary interaction, e.g. from epidemiological feedback or any other form of eco-evolutionary feedback. Independently of the particular source, demographic changes always affect all loci in the genome simultaneously. Therefore, existing methods to estimate the demography based on whole-genome data may offer the possibility to approximate the demographic history of both species, especially if time sampled data are available. However, the resolution of the demography depends on the amplitude and time-scales on which such population size fluctuations take place. Živković et al. (2019) could show that fluctuations in population size arising from host-parasite coevolution only leave a signature in the genome-wide parasite site frequency spectrum if they happen at a slow enough time scale. Irrespectively of whether the demographic changes can be resolved from genome-wide data or not, the resulting genomic signatures at the coevolving loci will be always the result of the allele frequency path at the coevolving locus itself. Therefore, further studies should focus on the specific effect of eco-evolutionary feedback on the variability of the allele frequency path and the resulting effect of the population size changes on mutation supply at the coevolving loci.

We could show that of our ABC-approach is suited to infer the cost of infection with very good accuracy by jointly using host and parasite polymorphism data from repeated experiments. Thus, we could demonstrate as a proof-of-principle that there is enough information contained in the site frequency spectra of the loci under coevolution to infer information about the past coevolutionary history. So far, our approach relies on data from repeated experiments and it is probably best met by data from microcosm experiments (e.g. Hall et al., 2011; Frickel et al., 2016) where coevolutionary interactions can be tracked across several replicates

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

for a reasonable amount of generations. Using data from repeated experiments is one possible attempt to deal with the variability in allele frequency trajectories resulting from the interaction between genetic drift and coevolution. The usage of data from several independent populations or the usage of time-sampled might be possible alternatives. Time samples offer an at least partially time-resolved view on changes in allele frequencies and accordingly, can help to better capture the coevolutionary dynamics.

Our results further show that analyzing both interacting partners in a joint framework rather than analyzing them separately helps to better recover information about the coevolutionary history. This is in line with recent method developments (MacPherson et al., 2018; Nuismer et al., 2017; Wang et al., 2018) which show the value of analyzing hosts and parasite in a joint framework. Additionally, these methods can be promising approaches to identify candidate genes being involved into the coevolutionary interaction on which our approach is based on.

Overall, we investigated the link between coevolutionary dynamics and resulting genomic signatures and quantify the amount of information available in polymorphism data. Although, we started from a very simple coevolutionary interaction we could show that model-based inference is possible. With growing availability of highly resolved genome data, even of non-model species, it is important to gain a differentiated and deep understanding of the continuum of possible links between coevolutionary dynamics without or with eco-evolutionary feedbacks and their effect on polymorphism data. Such as thorough understanding is the basis for devising appropriate sampling schemes, for using optimal combinations of diverse sources of information and for developing model-based refined inference methods.

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Agrawal, A., Lively, C., 2002. Infection genetics: gene-for-gene versus matching-alleles models and all points in between. EVOL ECOL RES 4 (1), 79–90.

Ashby, B., Boots, M., 2017. Multi-mode fluctuating selection in host-parasite coevolution. ECOL LETT 20 (3), 357–365.
URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/ele.12734`

Bakker, E. G., Toomajian, C., Kreitman, M., Bergelson, J., 2006. A genome-wide survey of R gene polymorphisms in *Arabidopsis*. PLANT CELL 18 (8), 1803–1818.

Beaumont, M. A., Zhang, W. Y., Balding, D. J., 2002. Approximate bayesian computation in population genetics. Genetics 162 (4), 2025–2035.

Bergelson, J., Purrington, C., 1996. Surveying patterns in the cost of resistance in plants. AM NAT 148 (3), 536–558.

Brown, J. K. M., Tellier, A., 2011. Plant-parasite coevolution: Bridging the gap between genetics and ecology. ANNU REV PHYTOPATHOL 49, 345–367.

Caicedo, A., Schaal, B., 2004. Heterogeneous evolutionary processes affect R gene diversity in natural populations of *Solanum pimpinellifolium*. P NATL ACAD SCI USA 101 (50), 17444–17449.

Charlesworth, B., Nordborg, M., Charlesworth, D., 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. GENET RES 70 (2), 155–174.

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Csillery, K., Blum, M. G. B., Gaggiotti, O. E., Francois, O., 2010. Approximate Bayesian Computation (ABC) in practice. TRENDS ECOL EVOL 25 (7), 410–418.

De Meaux, J., Cattan-Toupance, I., Lavigne, C., Langin, T., Neema, C., 2003. Polymorphism of a complex resistance gene candidate family in wild populations of common bean (*Phaseolus vulgaris*) in Argentina: comparison with phenotypic resistance polymorphism. MOL ECOL 12 (1), 263–273.

Decaestecker, E., Gaba, S., Raeymaekers, J. A. M., Stoks, R., Van Kerckhoven, L., Ebert, D., De Meester, L., 2007. Host-parasite 'Red Queen' dynamics archived in pond sediment. NATURE 450 (7171), 870–U16.

Ewing, G., Hermisson, J., 2010. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. BIOINFORMATICS 26 (16), 2064–2065.

Fay, J. C., Wu, C.-I., 2000. Hitchhiking under positive darwinian selection. GENETICS 55 (1), 1405–1413.

Flor, H. H., 1971. Current status of the gene for gene concept. ANNU REV PHYTOPATHOL 9, 275–296.

Frank, S., 1992. Models of plant pathogen coevolution. TRENDS GENET 8 (6), 213–219.

Frickel, J., Sieber, M., Becks, L., 2016. Eco-evolutionary dynamics in a coevolving host-virus system. ECOL LETT 19 (4), 450–459.

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Fu, Y.-X., Li, W.-H., 1993. Statistical tests of neutrality of mutations. GENETICS 133, 693–709.

Gokhale, C. S., Papkou, A., Traulsen, A., Schulenburg, H., 2013. Lotka-Volterra dynamics kills the Red Queen: population size fluctuations and associated stochasticity dramatically change host-parasite coevolution. BMC EVOL BIOL 13.

Gos, G., Slotte, T., Wright, S. I., 2012. Signatures of balancing selection are maintained at disease resistance loci following mating system evolution and a population bottleneck in the genus *Capsella*. BMC EVOL BIOL 12.

Hall, A. R., Scanlan, P. D., Morgan, A. D., Buckling, A., 2011. Host-parasite coevolutionary arms races give way to fluctuating selection. ECOL LETT 14, 635–642.

Hoerger, A. C., Ilyas, M., Stephan, W., Tellier, A., van der Hoorn, R. A. L., Rose, L. E., 2012. Balancing selection at the Tomato RCR3 guardee gene family maintains variation in strength of pathogen defense. PLOS GENET 8 (7).

Karasov, T. L., Almario, J., Friedemann, C., Ding, W., Giolai, M., Heavens, D., Kersten, S., Lundberg, D. S., Neumann, M., Regalado, J., Neher, R. A., Kemen, E., Weigel, D., 2018. *Arabidopsis thaliana* and *Pseudomonas* pathogens exhibit stable associations over evolutionary timescales. CELL HOST MICROBE 24 (1), 168+.

Karasov, T. L., Kniskern, J. M., Gao, L., DeYoung, B. J., Ding, J., Dubiella, U., Lastra, R. O., Nallu, S., Roux, F., Innes, R. W., Barrett, L. G., Hudson, R. R.,

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Bergelson, J., 2014. The long-term maintenance of a resistance polymorphism through diffuse interactions. NATURE 512 (7515), 436–U472.

Kraaijeveld, A., Godfray, H., 1997. Trade-off between parasitoid resistance and larval competitive ability in *Drosophila melanogaster*. NATURE 389 (6648), 278–280.

Lenski, R., 1988. Experimental studies of pleiotropy and epistasis in *Escherichia-Coli* .1. Variation in competitive fitness among mutants resistant to Virus-T4. EVOLUTION 42 (3), 425–432.

Leonard, K. J., 1994. Stability of equilibria in a gene-for-gene coevolution model of host-parasite interactions. PHYTOPATHOLOGY 84 (1), 70–77.

MacPherson, A., Otto, S. P., Nuismer, S. L., 2018. Keeping pace with the Red Queen: Identifying the genetic basis of susceptibility to infectious disease. GENETICS 208, 779–789.

Martiny, J. B. H., Riemann, L., Marston, M. F., Middelboe, M., 2014. Antagonistic coevolution of marine planktonic viruses and their hosts. Vol. 6 of ANNU REV MAR SCI. pp. 393–414.

Maynard Smith, J., Haigh, J., 1974. The hitch-hiking effect of a favourable gene. GENET RES 23 (1), 23–35.

Nei, M., Tajima, F., 1981. DNA polymorphism detectable by restriction endonucleases. GENETICS 97, 145–165.

Nuismer, S. L., Jenkins, C. E., Dybdahl, M. F., 2017. Identifying coevolving loci using interspecific genetic correlations. ECOL EVOL 7 (17), 6894–6903.

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Rose, L. E., Michelmore, R. W., Langley, C. H., 2007. Natural variation in the Pto disease resistance gene within species of wild tomato (*Lycopersicon*). II. Population genetics of Pto. GENETICS 175 (3), 1307–1319.

Seger, J., 1988. Dynamics of some simple host-parasite models with more than 2 genotypes in each species. P ROY SOC B-BIOL SCI 319 (1196), 541–555.

Stahl, E. A., Dwyer, G., Mauricio, R., Kreitman, M., Bergelson, J., 1999. Dynamics of disease resistance polymorphism at the Rpm1 locus of *Arabidopsis*. NATURE 400 (6745), 667–671.

Sunnaker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., Dessimoz, C., 2013. Approximate Bayesian Computation. PLOS COMPUT BIOL 9 (1).

Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. GENETICS 123, 585–595.

Tellier, A., Brown, J. K. M., 2007. Stability of genetic polymorphism in host-parasite interactions. P ROY SOC B-BIOL SCI 274 (1611), 809–817.

Tellier, A., Moreno-Gamez, S., Stephan, W., 2014. Speed of adaptation and genomic footprints of host-parasite coevolution under arms race and trench warfare dynamics. EVOLUTION 68 (8), 2211–2224.

Thrall, P., Burdon, J., 2003. Evolution of virulence in a plant host-pathogen metapopulation. SCIENCE 299 (5613), 1735–1737.

Thrall, P. H., Laine, A.-L., Ravensdale, M., Nemri, A., Dodds, P. N., Barrett, L. G., Burdon, J. J., 2012. Rapid genetic change underpins antagonistic coevolution in a natural host-pathogen metapopulation. ECOL LETT 15 (5), 425–435.

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Wang, M., Roux, F., Bartoli, C., Huard-Chauveau, C., Meyer, C., Lee, H., Roby, D., McPeek, M. S., Bergelson, J., 2018. Two-way mixed-effects methods for joint association analysis using both host and pathogen genomes. P NATL ACAD SCI USA.
URL http://www.pnas.org/content/early/2018/05/29/1710980115

Watterson, G. A., 1975. Number of segregating sites in genetic models without recombination. THEOR POPUL BIOL 7 (2), 256–276.

Wegmann, D., Leuenberger, C., Neuenschwander, S., Excoffier, L., 2010. ABC-toolbox: a versatile toolkit for approximate Bayesian computations. BMC BIOINFORMATICS 11.

Woolhouse, M., Webster, J., Domingo, E., Charlesworth, B., Levin, B., 2002. Biological and biomedical implications of the co-evolution of pathogens and their hosts. NAT GENET 32 (4), 569–577.

Zeng, K., Fu, Y.-X., Shi, S., Wu, C.-I., 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. GENETICS 174 (3), 1431–1439.

Živković, D., John, S., Verin, M., Stephan, W., Tellier, A., 2019. Neutral genomic signatures of host-parasite coevolution. bioRxiv.
URL https://www.biorxiv.org/content/early/2019/03/25/588202

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplementary information:

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

S1. Supplementary information models

S1.1. Model A

S1.1.1. Detailed description how the allele frequency path is obtained

In order to obtain the frequency of a given allele in the next generation, we perform the following steps:

- We compute the allele frequency after selection using the difference equations Eq. 3.
- We incorporate genetic drift by performing a binomial sampling based on the frequency after selection and the finite and fixed haploid population size (N_H for the host and N_P for the parasite).
- We allow for recurrent allele mutations (functional mutations) to take place and change genotypes from *RES* to *res* at rate μ_{Rtor} or *res* to *RES* at rate μ_{rtoR} in the host and from *ninf* to *INF* at rate μ_{ntoI} and from *INF* to *ninf* at rate $\mu_I ton$ in the parasite. We set all functional mutation rates to $\mu_{Rtor} = \mu_{ntoI} = \mu_{rtoR} = \mu_{Iton} = 10^{-5}$.

Note that the above mentioned steps are repeated twice for the parasite as there are two parasite generation per host generation. Once when going from parasite generation $g, 1$ to $g, 2$ and once when going from parasite generation $g, 2$ to $g+1, 1$. Accordingly, the detailed calculations for each parasite generation are as follows:

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

1. The expected frequency of *INF*-parasites after selection a_x (x=g,2 or x=g+1,1) is obtained by using the respective recursion equation in Eq. 3. The corresponding frequency of *ninf*-parasites is calculated as $A_x = 1 - a_x$.
2. The number of *INF*-parasite individuals after drift N_I is sampled from a Binomial distribution $N_I \sim \mathcal{B}(N_P, a_x)$. Thus, the number of *ninf*-parasites after drift ist equal to $N_n = N_P - N_I$.
3. In order to include the functional mutations the following two samplings are performed:
   - the number of mutants M_{In} from *INF* to *ninf* is obtained by sampling from a Poisson distribution with rate $\lambda = \mu_{Iton} \cdot N_I$.
   - the number of mutants M_{nI} from *ninf* to *INF* is obtained by sampling from a Poisson distribution with rate $\lambda = \mu_{ntoI} \cdot N_n$.

Thus, the number of *INF*-parasites in generation x is given by:

$$N_{x,I} = N_I - M_{In} + M_{nI} \tag{S1}$$

And the frequency of *INF*-parasites at the beginning of generation x is equal to:

$$\frac{N_{x,I}}{N_P} \tag{S2}$$

The corresponding steps for the host population are as follows.

1. The expected frequency of *RES*-hosts after selection R_{g+1} is obtained by using difference equation Eq. 3. The frequency of *res*-hosts is calculated as

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

$r_{g+1} = 1 - R_{g+1}$.

2. The number of *RES*-host individuals after drift is sampled from a Binomial distribution $N_R \sim \mathcal{B}(N_H, R_{g+1})$. Thus, the number of *res*-host after drift ist equal to $N_r = N_H - N_R$.
3. In order to include the functional mutations the following two samplings are performed:
   - the number of mutants from *RES* to *res* M_{Rr} is obtained by sampling from a Poisson distribution with rate $\lambda = \mu_{Rtor} \cdot N_R$.
   - the number of mutants from *res* to *RES* M_{rR} is obtained by sampling from a Poisson distribution with rate $\lambda = \mu_{rtoR} \cdot N_r$.

Thus, the number of *RES*-individuals in generation $g + 1$ is given by:

$$N_{g+1,R} = N_R - M_{Rr} + M_{rR} \tag{S3}$$

And the frequency of *RES*-hosts at the beginning of generation $g + 1$ is equal to:

$$\frac{N_{g+1,R}}{N_H} \tag{S4}$$

By repeating this procedure for $g_{max} = 3 \cdot \max(N_H, N_P)$ generations we obtain the so called frequency path which consists of the frequencies of all four alleles at the beginning of each generation g. In order to constrain a modified version of msms (Ewing and Hermisson, 2010; Tellier et al., 2014) by this frequency path we rescale the generations g in the host to $g_H^* = g/(2N_H)$ and in the parasite to $g_P^* = g/(2N_P)$. Note that msms is in diploid size. As N_H and N_P are haploid

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

population sizes this rescaling is equivalent to rescale time in units of $4 \cdot N_{H,diploid}$ and $4 \cdot N_{P,diploid}$. Based on this time rescaled frequency path we launch msms once for the host and once for the parasite.

S1.1.2. Fitness matrix

Table S1: Fitness matrix for **Model A** capturing the fitness effects of different interactions between host genotypes and parasites genotypes within a single host generation g. R_g (r_g) denotes the frequency of resistant (susceptible) hosts in generation g. $A_{g,1}$ ($A_{g,2}$) denotes the frequency of non-infective parasites and $a_{g,1}$ ($a_{g,2}$) denotes the frequency of infective parasites at the beginning of the first (second) parasite generation $t = 1$ ($t = 2$) within host generation g. The costs are: c_H=cost of resistance, c_P=cost of infectivity, s_1, s_2=cost of infection.

	first generation	fitness $g, 1$	second generation	fitness $g, 2$	host fitness
host genotype *RES* (R_g)	*ninf* ($A_{g,1}$)	0	*ninf* ($A_{g,2}$)	0	$1 - c_H$
	ninf ($A_{g,1}$)	0	*INF* ($a_{g,2}$)	$1 - c_P$	$(1 - c_H)(1 - s_2)$
	INF ($a_{g,1}$)	$1 - c_P$	*INF* ($a_{g,2}$)	$1 - c_P$	$(1 - c_H)(1 - s_1)$
host genotype *res* (r_g)	*ninf* ($A_{g,1}$)	1	*ninf* ($A_{g,2}$)	1	$1 - s_1$
	INF ($a_{g,1}$)	$1 - c_P$	*INF* ($a_{g,2}$)	$1 - c_P$	$1 - s_1$

S1.2. Model B

S1.2.1. Model description

Model B extends the basic model to $T > 2$ parasite generations per host generation. As in the basic model the cost of infection s_t is a function of the parasite

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

generation t in which the host became infected and the maximum cost of infection s, which correspond to the cost of being infected at the first parasite generation t=1 within host generation g. Upon infection a host stays infected until it reproduces and dies from natural death (at the end of the host generation g). An infected host is reinfected by the offspring of the particular parasite for all subsequent parasite generations within host generation g (100% auto-infection). Hosts which have not been infected so far can be attacked by the offspring of any parasite type at the beginning of each parasite generation t. Whether this interaction subsequently results in an infection depends on the infection matrix. The recursion equations for this model are given by:

$$a_{g,t+1} = \frac{(1-c_P)\left[a_{g,1} + \sum_{l=2}^{t} a_{g,l} R_g \prod_{m=1}^{l-1} A_{g,m}\right]}{(1-c_P)\left[a_{g,1} + \sum_{l=2}^{t} a_{g,l} R_g \prod_{m=1}^{l-1} A_{g,m}\right] + A_{g,1} r_g} \tag{S5a}$$

$$a_{g+1,1} = \frac{(1-c_P)\left[a_{g,1} + \sum_{l=2}^{T} a_{g,l} R_g \prod_{m=1}^{l-1} A_{g,m}\right]}{(1-c_P)\left[a_{g,1} + \sum_{l=2}^{T} a_{g,l} R_g \prod_{m=1}^{l-1} A_{g,m}\right] + A_{g,1} r_g} \tag{S5b}$$

$$a_{g,2} = \frac{(1-c_P)\cdot a_{g,1}}{(1-c_P) a_{g,1} + A_{g,1} r_g} \tag{S5c}$$

$$R_{g+1} = \frac{R_g \cdot (1-c_H)\left((1-s_1)\, a_{g,1} + \sum_{t=2}^{T}\left((1-s_t) a_{g,t} \prod_{l=1}^{t-1} A_{g,l}\right) + \prod_{t=1}^{T} A_{g,t}\right)}{R_g \cdot (1-c_H)\left((1-s_1)\, a_{g,1} + \sum_{t=2}^{T}\left((1-s_t) a_{g,t} \prod_{l=1}^{t-1} A_{g,l}\right) + \prod_{t=1}^{T} A_{g,t}\right) + r_g (1-s_1)} \tag{S5d}$$

Note that in this side analysis, genetic drift and functional mutations are only taken into account when going from host generation g to host generation $g+1$ in both, the host and the parasite. The frequency path in the parasite which is used to launch msms consists of the frequencies at the first parasite generation within host generation g. Time is rescaled as $g_P^* = g/(2N_P)$ in the parasite.

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

S1.3. Model C

S1.3.1. Model description

Model C is based on Model C in Tellier and Brown (2007). As in model A, we assume $T = 2$ discrete parasite generations per discrete host generation g and frequency-dependent disease transmission. Parasites of the second ($t = 2$) generation within host generation g infect the same host individual as there parent at rate ψ (auto-infection) or a different host at rate $1 - \psi$ (allo-infection). A host which is infected throughout the whole host generation g looses the amount $s_1 = s$ (cost of infection) of its fitness. If it is only infected during a single parasite generation the cost of infection reduces to $s_2 = \frac{s}{2}$. The respective fitness matrix is shown in table S2 with $A_{g,t}$ ($a_{g,t}$) denoting the frequency of *ninf* (*INF*)-parasites in the t-th parasite generation within host generation g and R_g (r_g) denoting the frequency of *RES* (*res*)-hosts in host generation g.

$$a_{g,2} = \frac{a_{g,1} \cdot (1 - c_P)}{a_{g,1} \cdot (1 - c_P) + A_{g,1} \cdot r_g} \tag{S6a}$$

$$a_{g+1,1} = \frac{(1 - c_P) \cdot \Big(R_g A_{g,1} a_{g,2} + r_g A_{g,1} a_{g,2}(1 - \psi) + a_{g,1}[\psi + a_{g,2}(1 - \psi)]\Big)}{r_g \cdot (\psi A_{g,1} + A_{g,2}(1 - \psi)) + (1 - c_P) \cdot \Big(R_g A_{g,1} a_{g,2} + r_g A_{g,1} a_{g,2}(1 - \psi) + a_{g,1}[\psi + a_{g,2}(1 - \psi)]\Big)} \tag{S6b}$$

$$R_{g+1} = \frac{R_g \cdot (1 - c_H)(A_{g,1}A_{g,2} + (1 - s_2)(A_{g,1}a_{g,2} + a_{g,1}A_{g,2}(1 - \psi)) + (1 - s_1)(a_{g,1}\psi + a_{g,1}a_{g,2}(1 - \psi)))}{R_g \cdot (1 - c_H)\Big(A_{g,1}A_{g,2} + (1 - s_2)(A_{g,1}a_{g,2} + a_{g,1}A_{g,2}(1 - \psi)) + (1 - s_1)(a_{g,1}\psi + a_{g,1}a_{g,2}(1 - \psi))\Big) + r_g(1 - s_1)} \tag{S6c}$$

The allele frequency path for this model is obtained in the same way as in **Model A**.

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

S1.3.2. Fitness matrix

Table S2: Fitness matrix for **Model C** capturing the fitness effects of different interactions between hosts and parasites within a single host generation g. R_g (r_g) denotes the frequency of resistant (susceptible) hosts in generation g. $A_{g,1}$ ($A_{g,2}$) denotes the frequency of non-infective parasites and $a_{g,1}$ ($a_{g,2}$) denotes the frequency of infective parasites at the beginning of the first (second) parasite generation $t = 1$ ($t = 2$) within host generation g. n/a indicates that these hosts have not been infected as *ninf*-parasites fail to infect *RES*-hosts. The costs are: c_H=cost of resistance, c_P=cost of infectivity, s_1, s_2=costs of infection.

	first generation	auto-infection (ψ) allo-infection ($1-\psi$)	second generation	fitness of second parasite generation	host fitness
host genotype *RES* (R_g)	*ninf* ($A_{g,1}$)	n/a	*ninf* ($A_{g,2}$)	0	$1-c_H$
			INF ($a_{g,2}$)	$1-c_P$	$(1-c_H)(1-s_2)$
	INF ($a_{g,1}$)	ψ	*INF* ($a_{g,2}$)	$1-c_P$	$(1-c_H)(1-s_1)$
		$1-\psi$	*INF* ($a_{g,2}$)	$1-c_P$	$(1-c_H)(1-s_1)$
		$1-\psi$	*ninf* ($A_{g,2}$)	0	$(1-c_H)(1-s_2)$
host genotype *res* (r_g)	*ninf* ($A_{g,1}$)	ψ	*ninf* ($A_{g,2}$)	1	$1-s_1$
		$1-\psi$	*ninf* ($A_{g,2}$)	1	$1-s_1$
		$1-\psi$	*INF* ($a_{g,2}$)	$1-c_P$	$1-s_1$
	INF ($a_{g,1}$)	ψ	*INF* ($a_{g,2}$)	$1-c_P$	$1-s_1$
		$1-\psi$	*INF* ($a_{g,2}$)	$1-c_P$	$1-s_1$
		$1-\psi$	*ninf* ($A_{g,2}$)	1	$1-s_1$

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

S2. Pairwise Manhattan Distance (PMD)

PMD is calculated as the sum of manhattan distances between class i in the host site frequency spectrum and class i in the parasite site frequency spectrum. It is calculated as:

$$PMD = \sum_{i=1}^{n-1} |\xi_{H,i} - \xi_{P,i}| \tag{S1}$$

with $\xi_{H,i}$ ($\xi_{P,i}$) being the total number of neutral SNPs linked to the coevolving locus which are in frequency class i of the unfolded site frequency spectrum of the host (parasite). Note that in the current formulation the summary statistic relies on the sample size of the host (n_H) and the parasite (n_P) being the same. However it is possible to adjust this summary statistic by downsampling the site frequency spectrum of the species with the higher sample size.

S3. Supplementary tables

Table S3: Approximate frequencies of resistant hosts ($\hat{R}$) and infective parasites ($\hat{a}$) at the non-trivial interal equilbrium point in **Model A** (Eq. 4) for various combinations of cost of resistance (c_H), cost of infectivity (c_P) and cost of infection (s) as plotted in Fig. 1.

panel of Fig. 1	c_H	c_P	s_1	s_2	$\hat{a}$	$\hat{R}$
a	0.05	0.10	0.20	0.10	0.667	0.081
a	0.05	0.10	0.30	0.15	0.775	0.089
a	0.05	0.10	0.40	0.20	0.835	0.094
a	0.05	0.10	0.50	0.25	0.873	0.097
a	0.05	0.10	0.60	0.30	0.899	0.100
a	0.05	0.10	0.70	0.35	0.919	0.102
a	0.05	0.10	0.80	0.40	0.934	0.104

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

b	0.10	0.10	0.20	0.10	0.424	0.068
b	0.10	0.10	0.30	0.15	0.603	0.077
b	0.10	0.10	0.40	0.20	0.704	0.084
b	0.10	0.10	0.50	0.25	0.771	0.089
b	0.10	0.10	0.60	0.30	0.818	0.092
b	0.10	0.10	0.70	0.35	0.853	0.096
b	0.10	0.10	0.80	0.40	0.881	0.098
c	0.05	0.30	0.20	0.10	0.667	0.291
c	0.05	0.30	0.30	0.15	0.775	0.324
c	0.05	0.30	0.40	0.20	0.835	0.347
c	0.05	0.30	0.50	0.25	0.873	0.363
c	0.05	0.30	0.60	0.30	0.899	0.375
c	0.05	0.30	0.70	0.35	0.919	0.384
c	0.05	0.30	0.80	0.40	0.934	0.392
d	0.10	0.30	0.20	0.10	0.424	0.235
d	0.10	0.30	0.30	0.15	0.603	0.273
d	0.10	0.30	0.40	0.20	0.704	0.301
d	0.10	0.30	0.50	0.25	0.771	0.323
d	0.10	0.30	0.60	0.30	0.818	0.340
d	0.10	0.30	0.70	0.35	0.853	0.354
d	0.10	0.30	0.80	0.40	0.881	0.366

Table S4: Overview of all parameters and variables used in this paper.

name	standard value	Description
c_H	0.05	cost of resistance

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Tab. S4 Continued:

s	0.3	cost of infection
c_P	0.10	cost of virulence
ψ	1	auto-infection rate
R_0	0.20	initial frequency of *RES*-hosts
a_0	0.20	initial frequency of *INF*-parasites
R_g		frequency of *RES*-hosts in generation g
r_g		frequency of *res*-hosts in generation g
$A_{g,t}$		frequency of *ninf*-parasites at the beginning of the t-th parasite generation within host generation g
$a_{g,t}$		frequency of *INF*-parasites at the beginning of the t-th parasite generation within host generation g
g_{max}	$30,000$	number of host generations to simulate
T	2	number of parasite generations within host generation g
g	-	counter for host generations
t	-	counter for parasite generations within host generation g
n_H	50	sample size host
n_P	50	sample size parasite
N_H	$10,000$	haploid host population size
N_P	$10,000$	haploid parasite population size
θ_H	5	neutral population mutation rate host
θ_P	5	neutral population mutation rate parasite
μ_{Rtor}	10^{-5}	mutation rate *RES* to *res*
μ_{rtoR}	10^{-5}	mutation rate *res* to *RES*
μ_{ntoI}	10^{-5}	mutation rate *ninf* to *INF*
μ_{Iton}	10^{-5}	mutation rate *INF* to *ninf*
$\mu_{neutral}$	10^{-7}	neutral mutation rate per bp
N_R		number of *RES*-hosts after selection and genetic drift
N_r		number of *res*-hosts after selection and genetic drift
N_n		number of *ninf*-parasites after selection and genetic drift

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Tab. S4 Continued:

N_I	number of *INF*-parasites after selection and genetic drift
M_{Rr}	current number of mutants from *RES* to *res*
M_{rR}	current number of mutants from *res* to *RES*
M_{nI}	current number of mutants from *ninf* to *INF*
M_{In}	current number of mutants from *INF* to *ninf*
$N_{g+1,R}$	number of *RES*-hosts in host generation $g + 1$
$N_{x,I}$	number of *INF*-parasites in parasite generation x

Table S5: Summary statistics calculated for the pseudo-observed data sets

Summary statistic	reference
number of segregating sites S	(Watterson, 1975)
θ_W	(Watterson, 1975)
nucleotide diversity π	(Nei and Tajima, 1981)
Tajimas' D	(Tajima, 1989)
Fu and Li's D	(Fu and Li, 1993)
Fu and Li's F	(Fu and Li, 1993)
θ_H	(Fay and Wu, 2000)
Hprime	(Zeng et al., 2006)
PMD	

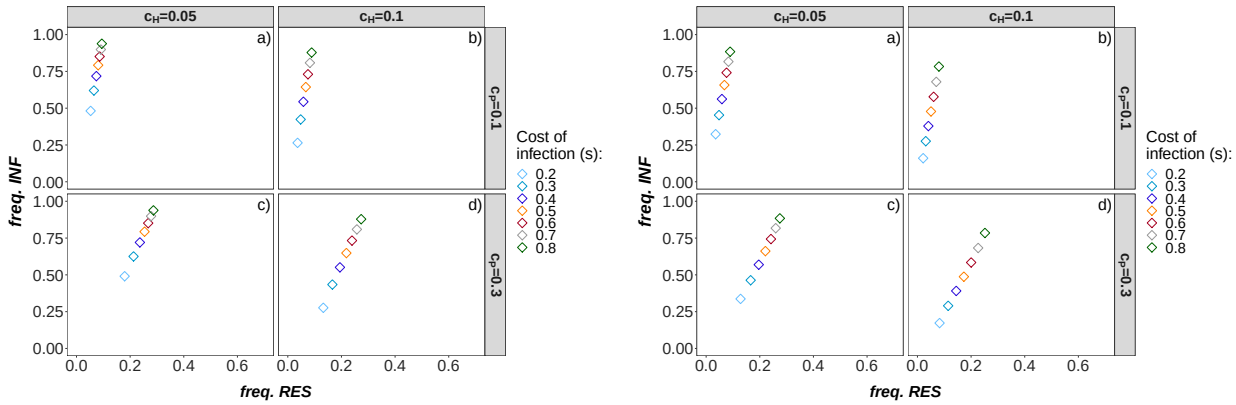
bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Table S6: Settings which have been used for running Approximate Bayesian Computation for the different scenarios and number of repetitions.

Scenario S	nb repetitions r	information used	'known' parameters	inferred parameters	prior distribution	complex parametes
$S = 1$	200	host + parasite	model A, $c_H = 0.05$, $c_P = 0.10$, $T = 2$, $\psi = 1$	s	unif$(0.1, 0.9)$	$\theta_H = N_H/1000$
				N_H	logunif$(2,000, 20000)$	$\theta_P = N_P/1000$
				N_P	logunif$(2,000, 20000)$	$g_{max} = 3 \cdot \max(N_H, N_P)$
$S = 1$	200	host	model A, $c_H = 0.05$, $c_P = 0.10$, $T = 2$	s	unif$(0.1, 0.9)$	$\theta_H = N_H/1000$
				N_H	logunif$(2,000, 20000)$	$\theta_P = N_P/1000$
				N_P	logunif$(2,000, 20000)$	$g_{max} = 3 \cdot \max(N_H, N_P)$
$S = 1$	200	parasite	model A, $c_H = 0.05$, $c_P = 0.10$, $T = 2$	s	unif$(0.1, 0.9)$	$\theta_H = N_H/1000$
				N_H	logunif$(2,000, 20000)$	$\theta_P = N_P/1000$
				N_P	logunif$(2,000, 20000)$	$g_{max} = 3 \cdot \max(N_H, N_P)$
$S = 1$	10	host + parasite	model A, $c_H = 0.05$, $c_P = 0.10$, $T = 2$	s	unif$(0.1, 0.9)$	$\theta_H = N_H/1000$
				N_H	logunif$(2,000, 20000)$	$\theta_P = N_P/1000$
				N_P	logunif$(2,000, 20000)$	$g_{max} = 3 \cdot \max(N_H, N_P)$
$S = 1$	10	host	model A, $c_H = 0.05$, $c_P = 0.10$, $T = 2$	s	unif$(0.1, 0.9)$	$\theta_H = N_H/1000$
				N_H	logunif$(2,000, 20000)$	$\theta_P = N_P/1000$
				N_P	logunif$(2,000, 20000)$	$g_{max} = 3 \cdot \max(N_H, N_P)$
$S = 1$	10	parasite	model A, $c_H = 0.05$, $c_P = 0.10$, $T = 2$	s	unif$(0.1, 0.9)$	$\theta_H = N_H/1000$
				N_H	logunif$(2,000, 20000)$	$\theta_P = N_P/1000$
				N_P	logunif$(2,000, 20000)$	$g_{max} = 3 \cdot \max(N_H, N_P)$
$S = 2$	200	host + parasite	model A, $N_H = 10,000$, $N_P = 10,000$, $T = 2$, $g_{max} = 30,000$	s	unif$(0.1, 0.9)$	-
				c_H	unif$(0.01, 0.35)$	-
				c_P	unif$(0.01, 0.35)$	-
$S = 2$	200	host	model A, $N_H = 10,000$, $N_P = 10,000$, $T = 2$, $g_{max} = 30,000$	s	unif$(0.1, 0.9)$	-
				c_H	unif$(0.01, 0.35)$	-
				c_P	unif$(0.01, 0.35)$	-
$S = 2$	200	parasite	model A, $N_H = 10,000$, $N_P = 10,000$, $T = 2$, $g_{max} = 30,000$	s	unif$(0.1, 0.9)$	-
				c_H	unif$(0.01, 0.35)$	-
				c_P	unif$(0.01, 0.35)$	-
$S = 2$	10	host + parasite	model A, $N_H = 10,000$, $N_P = 10,000$, $T = 2$, $g_{max} = 30,000$	s	unif$(0.1, 0.9)$	-
				c_H	unif$(0.01, 0.35)$	-
				c_P	unif$(0.01, 0.35)$	-
$S = 2$	10	host	model A, $N_H = 10,000$, $N_P = 10,000$, $T = 2$, $g_{max} = 30,000$	s	unif$(0.1, 0.9)$	-
				c_H	unif$(0.01, 0.35)$	-
				c_P	unif$(0.01, 0.35)$	-
$S = 2$	10	parasite	model A, $N_H = 10,000$, $N_P = 10,000$, $T = 2$, $g_{max} = 30,000$	s	unif$(0.1, 0.9)$	-
				c_H	unif$(0.01, 0.35)$	-
				c_P	unif$(0.01, 0.35)$	-

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

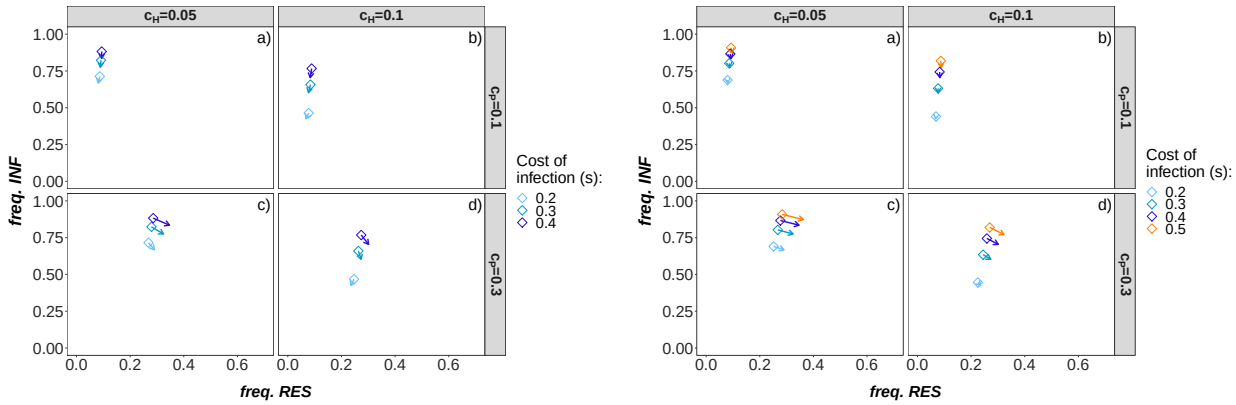
S4. Supplementary figures



(a) $T = 5$ **(b)** $T = 10$

Figure S1: Deterministic equilibrium frequencies for **Model B** for a) $T = 5$ parasite generations (left) and b) $T = 10$ parasite generations (right) per host generation. The equilibrium frequencies for different combinations of cost of resistance $c_H = (0.05, 0.1)$ (columns), cost of infectivity $c_P = (0.1, 0.3)$ (rows) and cost of infection $s = (0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8)$ (color of the squares) are shown. Only combinations with trench-warfare dynamics are shown. Centres of the squares represent the equilbrium frequencies obtained by simulating numerically the recursion equations in Eq. S5 for $g_{max} = 30,000$ host generations starting with an initial frequency of $R_0 = 0.2$ resistant hosts and $a_0 = 0.2$ infective parasites.

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



(a) $\psi = 0.75$ **(b)** $\psi = 0.95$

Figure S2: Deterministic equilibrium frequencies for **Model C** (auto-allo-infection model) with $T = 2$ parasite generations per host generation for two different autoinfection rates $\psi = 0.75$ (left) and $\psi = 0.95$ (right). The equilibrium frequencies for different combinations of cost of resistance $c_H = (0.05, 0.1)$ (columns), cost of infectivity $c_P = (0.1, 0.3)$ (rows) and cost of infection $s = (0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8)$ (color of the squares) are shown. Only combinations which result in trench-warfare dynamics are plotted. Centres of the squares represent the equilbrium frequencies obtained by simulating numerically the recursion equations in Eq. S6 for $g_{max} = 30,000$ host generations starting with an initial frequency of $R_0 = 0.2$ resistant hosts and $a_0 = 0.2$ infective parasites. Heads of the arrows represent the equilibrium frequencies based on Eq. 4 which corresponds to the case $\psi = 1$ (Tellier and Brown, 2007).

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

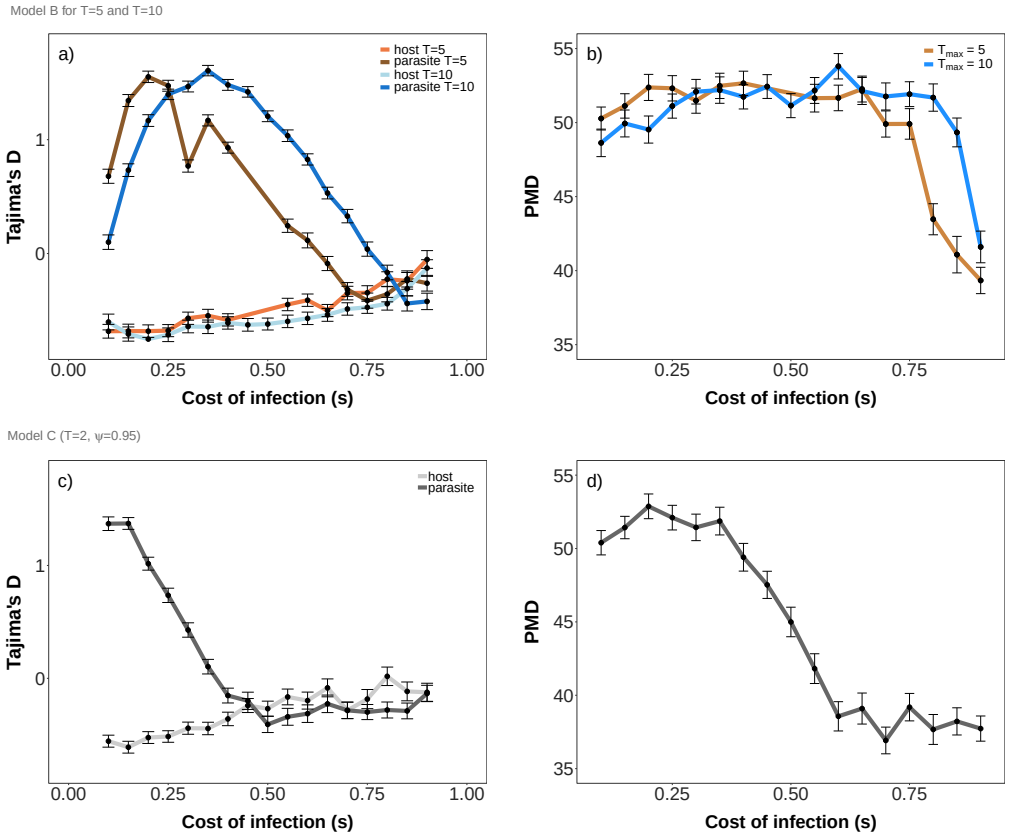


Figure S3: Mean and standard error of Tajima's D (a+c) and pairwise manhattan distance (PMD) (b+d) for various costs of infection s (x-axis) and $r = 200$ repetitions. Results for **Model B** (pure autoinfection model with $T = 5$ and $T = 10$) are shown at the top, results for **Model C** (auto-allo-infection model with $\psi = 0.95$) are shown at the bottom. The other parameters are fixed to: $c_H = 0.05$ and $c_P = 0.1$. Initial frequencies R_0 and a_0 in a and b are chosen randomly from a uniform distribution between 0 and 1 while $R_0 = a_0 = 0.2$ in c and d.

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

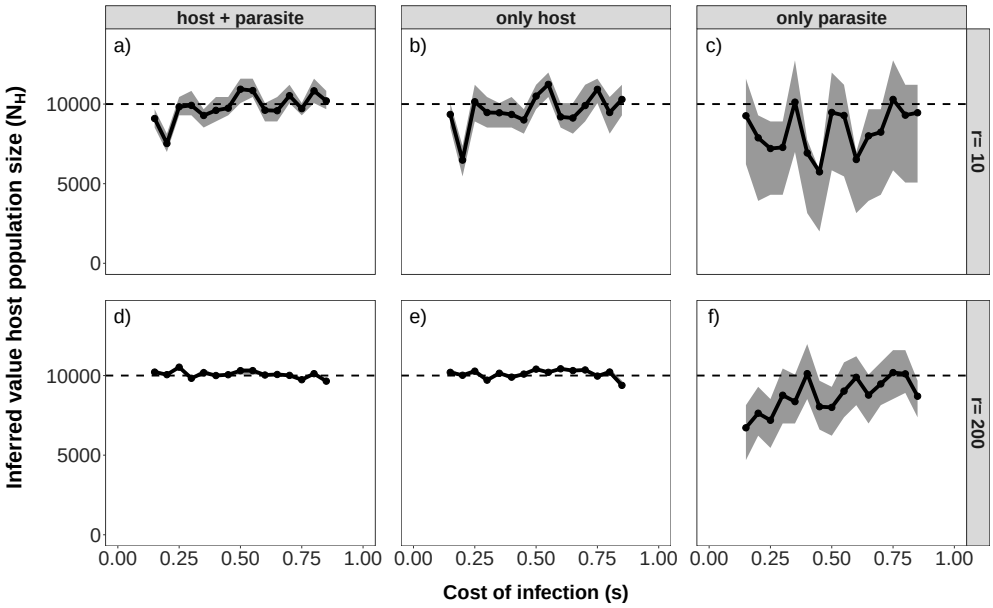


Figure S4: Median of the posterior distribution (y-axis) for the host population size N_H against the true cost of infection s for $r = 10$ (top, a-c) and $r = 200$ (bottom, d-f). The inference results for scenario 1 based on host and parasite polymorphism data (left, a+d), host polymorphism data only (middle, b+e) and parasite polymorphism data only (right, c+f) are shown. The true host population size is always $N_H = 10,000$ as indicated by the dashed horizontal line. The chosen parameters are: $c_H = 0.05$, $c_P = 0.1$, functional mutation rates $= 10^{-5}$, $c_H = 0.05$, $c_P = 0.1$, $g_{max} = 30,000$, $\theta_H = 5$, $\theta_P = 5$, $n_H = 50$ and $n_P = 50$. Priors have been chosen as follows: N_H log uniform$(2000, 40000)$, N_P log uniform$(2000, 40000)$, s uniform$(0.1, 0.9)$. s, N_H and N_P are inferred simultaneously for all plots.

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

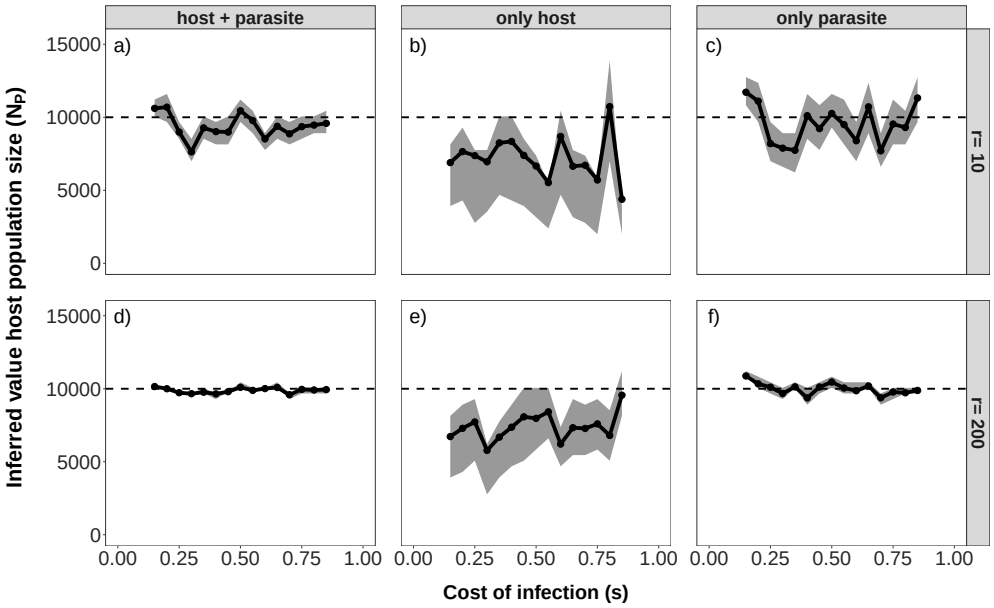


Figure S5: Median of the posterior distribution (y-axis) for the parasite population size N_P against the true cost of infection s for $r = 10$ (top, a-c) and $r = 200$ (bottom, d-f). The inference results for scenario 1 based on host and parasite polymorphism data (left, a+d), host polymorphism data only (middle, b+e) and parasite polymorphism data only (right, c+f) are shown. The true parasite population size is always $N_P = 10,000$ as indicated by the dashed horizontal line. The chosen parameters are: $c_H = 0.05$, $c_P = 0.1$, functional mutation rates $= 10^{-5}$, $c_H = 0.05$, $c_P = 0.1$, $g_{max} = 30,000$, $\theta_H = 5$, $\theta_P = 5$, $n_H = 50$ and $n_P = 50$. Priors have been chosen as follows: N_H log uniform$(2000, 40000)$, N_P log uniform$(2000, 40000)$, s uniform$(0.1, 0.9)$. s, N_H and N_P are inferred simultaneously for all plots.

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

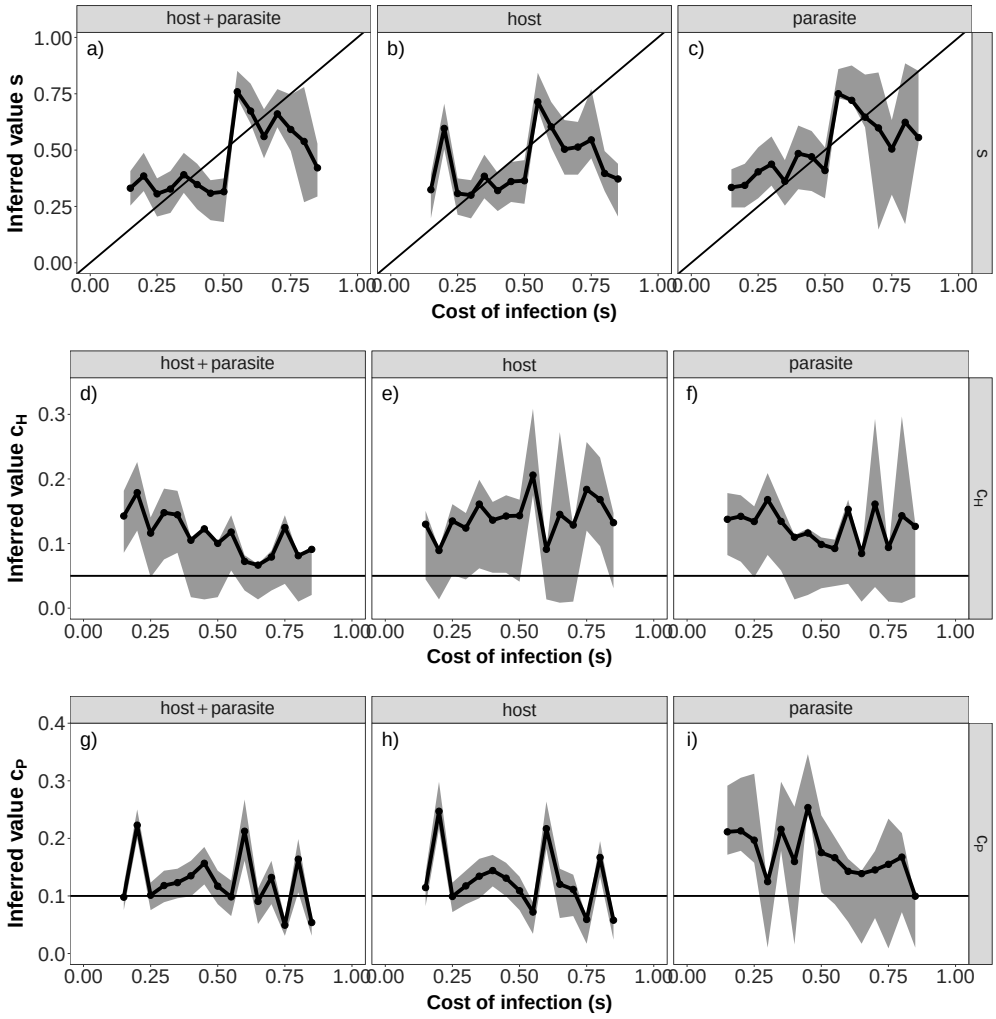
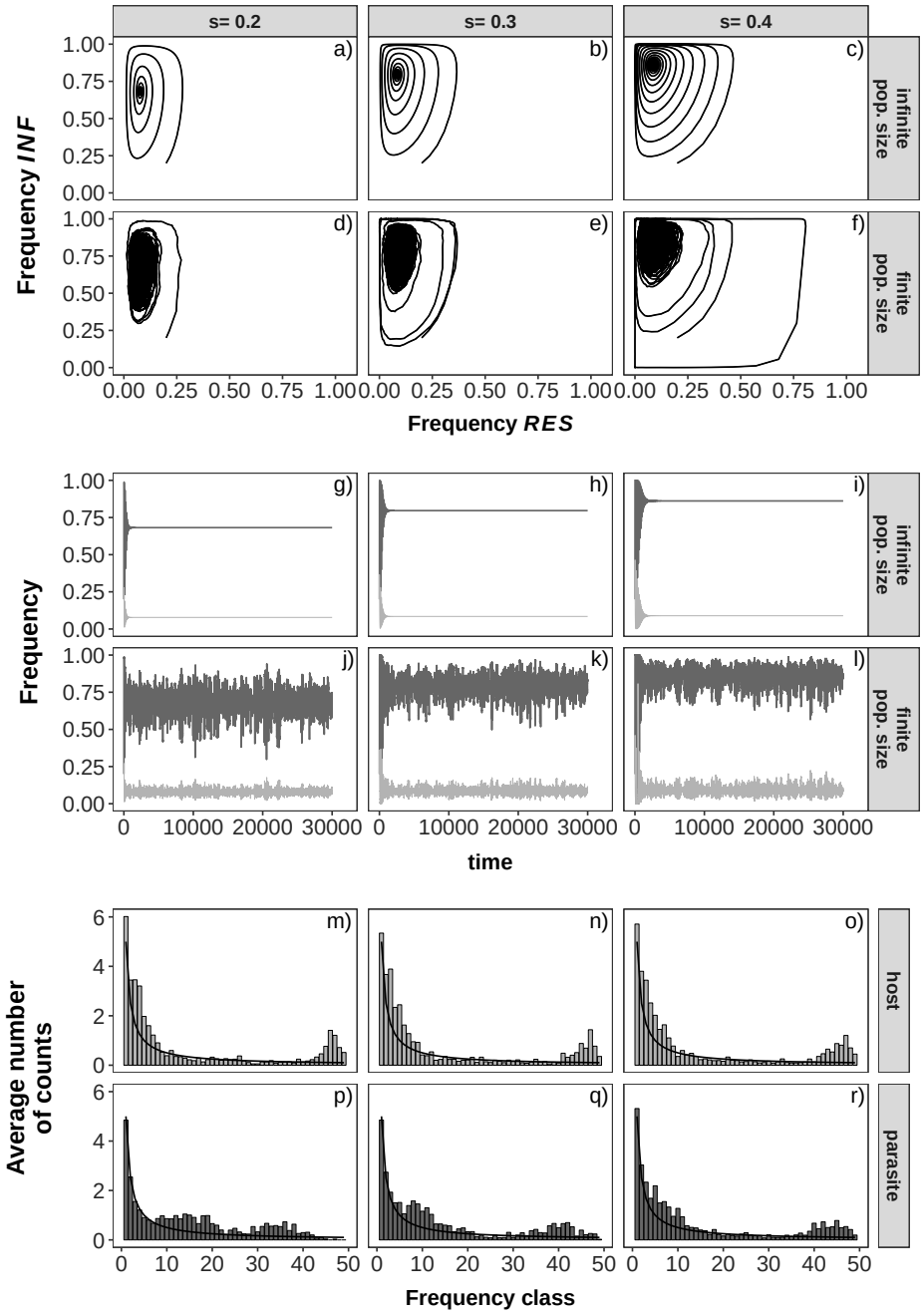
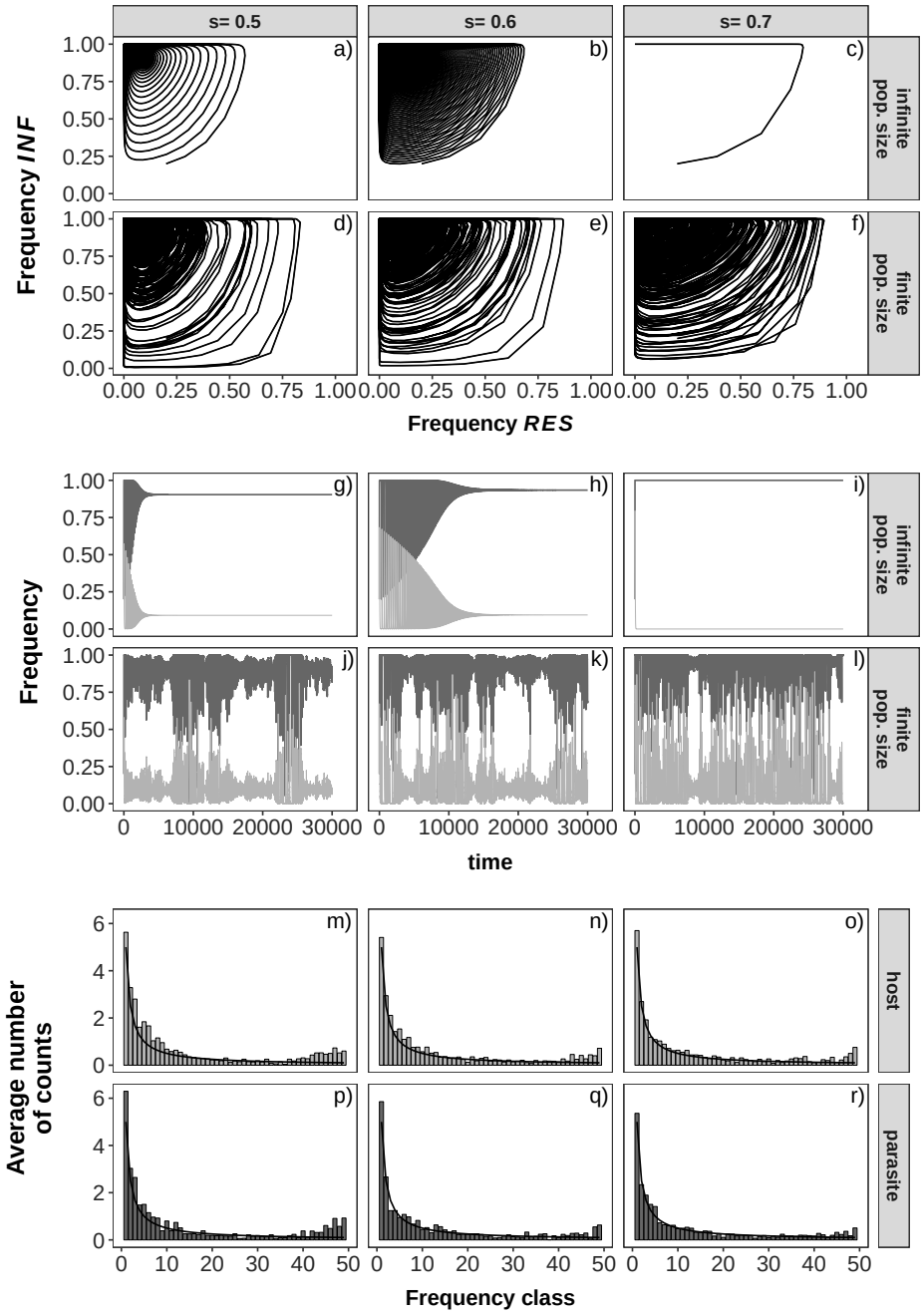


Figure S6: Median of the posterior distribution (y-axis) for the cost of infection s (a-c), cost of resistance c_H (d-f) and cost of infectivity c_P (g-i) compared to the true values (x-axis) for $r = 10$. The results are shown for inference based on host and parasite polymorphism data (left, a+d+g), host polymorphism data only (middle, b+e+h) and parasite polymorphism data only (right, c+f+i). The fixed parameters are chosen as: $N_H = N_P = 10,000$, $\mu_{Rtor} = \mu_{ntoI} = \mu_{rtoR} = \mu_{Iton} = 10^{-5}$, $g_{max} = 30,000$, $\theta_H = 5$, $\theta_P = 5$, $n_H = 50$ and $n_P = 50$. Priors have been chosen as follows: s uniform$(0.1, 0.9)$, c_H uniform$(0.01, 0.35)$, c_P uniform$(0.01, 0.35)$. s, c_H and c_P are inferred simultaneously for all plots.

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



711

bioRxiv preprint doi: https://doi.org/10.1101/625301; this version posted May 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figure S7: Influence of the cost of infection (s) on the coevolutionary dynamics and genomic signatures in **Model A**. The subfigures show the allele frequency trajectory in infinite population size (a-c, g-i), one exemplary allele frequency path in finite population size which takes genetic drift and functional mutations into account (d-f, j-l), the average unfolded host site frequency spectrum of $r = 200$ repetitions (m-o) and the average unfolded parasite site frequency spectrum of $r = 200$ repetitions (p-r).

In subfigures a-f each dot represents the frequency of resistant (*RES*) hosts (x-axis) and infective (*INF*) parasites (y-axis) at the beginning of a single host generation g. The same information is displayed in a slightly different way in subfigures g-l. Here, the frequencies of resistant (*RES*) hosts (light grey) and infective (*INF*) parasites (dark grey) (y-axis) are plotted over time (x-axis). Costs are fixed to $c_H = 0.05$, $c_P = 0.1$. The results in finite population size are plotted for $N_H = N_P = 10,000$, $\mu_{Rtor} = \mu_{ntoI} = \mu_{rtoR} = \mu_{Iton} = 10^{-5}$. The site frequency spectra are shown for $\theta_P = \theta_H = 5$ and $n_H = n_P = 50$.