Cross-Species association statistics for genome-wide studies of host and parasite polymorphism data

Hanna Märkle[a,1], Aurélien Tellier[a,1,*], Sona John[a,1,*]

[a]*Section of Population Genetics, Technical University of Munich, 85354 Freising, Germany*

[*]Corresponding authors. Email addresses: sona.john@tum.de, tellier@wzw.tum.de

[1]All three authors contributed equally.

Abstract

Uncovering the genes governing host-parasite coevolution is of importance for disease management in agriculture and human medicine. The availability of increasing amounts of host and parasite full genome-data in recent times allows to perform cross-species genome-wide association studies based on sampling of genomic data of infected hosts and their associated parasites strains. We aim to understand the statistical power of such approaches. We develop two indices, the cross species association (CSA) and the cross species prevalence (CSP), the latter additionally incorporating genomic data from uninfected hosts. For both indices, we derive genome-wide significance thresholds by computing their expected distribution over unlinked neutral loci, *i.e.* those not involved in determining the outcome of interaction. Using a population genetics and an epidemiological coevolutionary model, we demonstrate that the statistical power of these indices to pinpoint the interacting loci in full genome data varies over time. This is due to the underlying GxG interactions and the coevolutionary dynamics. Under trench-warfare dynamics, CSA and CSP are very accurate in finding out the loci under coevolution, while under arms-race dynamics the power is limited especially under a gene-for-gene interaction. Furthermore, we reveal that the combination of both indices across time samples can be used to estimate the asymmetry of the underlying infection matrix. Our results provide novel insights into the power and biological interpretation of cross-species association studies using samples from natural populations or controlled experiments.

Keywords

population genomics; linkage disequilibrium; single nucleotide polymorphism; host-parasite coevolution

1. Introduction

The increasing availability of host and parasite whole-genome data provides powerful means to detect genes determining the outcome of host-parasite interactions. The simple underlying idea is to perform all possible pairwise comparisons of Single Nucleotide Polymorphisms (SNPs) between samples of hosts and parasites in order to draw correlations with the outcome of infection. Recently, new Genome-Wide Association (GWA) methods to study host-parasite coevolutionary interactions have been proposed and performed (termed co-GWAs, (Ebert 2018, MacPherson et al.

2018, Nuismer et al. 2017, Wang et al. 2018)). However, such analyses rely on performing large scale controlled experiments with numerous host and parasites genotypes (thus can be termed as controlled co-GWAs). A promising less-labour intensive alternative is to perform natural co-GWAs, based on whole-genome data of infected hosts and their associated parasites sampled from natural populations (Ansari et al. 2017, Bartha et al. 2013, Bartoli and Roux 2017). Such data sets inherently contain phenotypic information (the infection outcome) for each sampled host-parasite pair, namely the susceptibility/resistance of host genotypes and infectivity/non-infectivity of parasite genotypes. Accordingly, the causal genetic variants for host susceptibility and parasite infectivity are expected to show statistically significant associations (Host Genotype x Parasite Genotype). A chief hypothesis is that such combination of host and parasite loci should explain a large variance in the infection in contrast to the neutral SNPs in the genome, which by definition have no effect on the interaction outcome. Natural co-GWAs have been applied, to our knowledge, twice to uncover strong associations between human SNPs 1) of the major histocompatibility complex (MHC) and known HIV epitopes (Bartha et al. 2013), and 2) of leukocyte antigen molecules and components of the interferon lambda innate immune system and the hepatitis C virus NS5A protein (Ansari et al. 2017). These studies conclude that the HIV and hepatitis C viruses do adapt to different variants of the MHC or leukocyte/interferon lambda present in the human population (Ansari et al. 2017, Bartha et al. 2013).

In principle, natural co-GWAs can be readily extended to any coevolutionary system where it is possible to call SNPs for a sample of infected hosts and the corresponding infecting parasites (Bartoli and Roux 2017). For example, transcriptome data of infected hosts and the corresponding infecting parasites (Dobon et al. 2016) or whole genome-data from controlled coevolutionary experiments (Frickel et al. 2018) can be readily obtained. In these studies, a current restriction (also a key requirement) to perform natural co-GWAs is to keep track of the co-occurrence of host and parasite genotypes which can be technically challenging as often sequencing of pools of infected hosts and parasites is performed.

One common underlying assumption of host-parasite co-GWAs is that the host genes determining the infection outcome, *i.e.* susceptibility/resistance, are coevolving with the corresponding infectivity genes in the parasite. Coevolution is defined as reciprocal changes in allele frequencies at the coevolving loci which result from selective pressures that two interacting species exert on one another (Janzen 1980). This definition encompasses both, synergistic (symbiosis) and

antagonistic (host-parasite, prey-predator) interactions.

Allele frequency changes at the coevolutionary loci are commonly described by a continuum between two extremes (Woolhouse et al. 2002), namely, the arms-race (Dawkins and Krebs 1979, Stahl and Bishop 2000, Woolhouse et al. 2002) and the trench-warfare (Stahl et al. 1999) dynamics. Under arms-race dynamics, coevolution causes recurrent fixation of alleles at the interacting loci, and accordingly, allelic polymorphism is only transient. In contrast, several alleles are maintained for a long-period of time under trench-warfare dynamics, with frequencies either persistently fluctuating over time or reaching a fixed stable polymorphic equilibrium. Note that in both scenarios, allele frequencies fluctuate over time before reaching fixation or the stable equilibrium.

The speed and type of frequency fluctuations depends on the host and parasite life-history traits (review in Brown and Tellier (2011)), the strength of epidemiological dynamics (Ashby and Boots 2017) and the underlying GxG interactions. Given the assumption that few major genes determine the interaction outcome, these GxG interactions can be captured in a so called infection matrix $\mathcal{A}$. Here, each entry α_{ij} stores the probability that a parasite genotype j can infect a host genotype i (Tab. 1) or equivalently the degree of infection (disease severity). Two well known infection matrices, are the matching-allele (MA) and the gene-for-gene (GFG) model. Both the MA and the GFG models represent some point in a continuum of infection matrices (Agrawal and Lively 2002) and are a subset of more complex matrices (with several alleles or loci, Gandon and Michalakis (2002), Ashby and Boots (2017)). In MA interactions a given parasite genotype can only infect a host when it matches the particular host allele (diagonal coefficients in Tab. 1b). For a 2x2 infection matrix the probabilities to infect the "non-matching" host genotypes can be defined as $1 - c_1$ and $1 - c_2$ (off diagonal coefficients in Tab. 1b). GFG interactions (Tab. 1c) are characterized by a universally susceptible host genotype (here host $i = 1$) and an universally infective parasite (here parasite $j = 2$). Here, the probability that host $i = 2$ is infected by parasite $j = 1$ is denoted by $1 - c$.

In this study we address the following questions. 1) Which statistics can be used in co-GWAs studies to pinpoint the loci under coevolution? 2) What is the power of these statistics to disentangle the effect of neutral loci from that of coevolutionary loci? 3) What is the statistical power of these statistics under various infection matrices and/or fluctuating allele frequencies? 4) Can

Table 1 Infection matrices for coevolutionary models

a) general infection matrix	**b) matching-allele**	**c) gene-for-gene**
$\begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix}$	$\begin{pmatrix} 1 & 1-c_1 \\ 1-c_2 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 1 \\ 1-c & 1 \end{pmatrix}$

The infection matrix $\mathcal{A}$ determines the outcome of the interaction between host genotypes (rows) and parasite genotypes (columns). Each α_{ij} can be interpreted either as the probability for a given individual to be infected or as the degree of infection (disease severity or partial resistance).

these statistics be used to infer the underlying infection matrix based on collected samples without further controlled infection experiments? We first develop two indices the cross-species association (CSA) index (analogous to the measure used in Ansari et al. (2017), Bartoli and Roux (2017)) and the cross species prevalence (CSP) index to measure the association of alleles between the coevolutionary loci in the host and the parasite. Second, we assess the statistical power of these cross-species indices to pinpoint the coevolving loci among the genome-wide neutral SNPs. This is realized by computing the expected distribution of these indices for all possible comparisons between host and parasite neutral loci. Note that we assume sampling from natural populations of hosts and parasites which undergo recombination so that neutral SNPs are unlinked from the loci under coevolution. As a result, we quantify the statistical power of these statistics to detect the loci underlying coevolution over the course of coevolutionary cycles and for different underlying GxG matrices. We demonstrate that performing co-GWAs with our indices across time samples allows to infer the type of dynamics occurring (arms race versus trench warfare) as well as the underlying GxG interaction matrix. We then discuss the applicability of our co-GWAs indices to study coevolution in natural or controlled systems.

2. Methods

2.1. The coevolutionary models

We assume that the outcome of an interaction between a host and a parasite, namely if the host is infected or not, is determined by a single biallelic host and single biallelic parasite locus. These two loci are defined as the coevolving loci. For simplicity, we consider a haploid model for hosts

and parasites. The outcome of the interaction between a particular host and parasite genotype is determined by the infection matrix $\mathcal{A} = (\alpha_{ij})$ with $1 \leq i, j \leq A$, where A is the total number of host (respectively parasite) genotypes (in this study $A = 2$, Table 1).

2.1.1. Model A: population genetics model

First, we use a simple population genetics model (henceforward termed model A) to study the allele frequency changes at the coevolving loci under the assumption of very large (infinite) haploid host and parasite population sizes. We assume that host and parasite generations are discrete and synchronized in terms of reproduction (Tellier and Brown 2007). The frequency of host genotype h_i and parasite genotype p_j in generation $g + 1$ is obtained as:

$$h_{i,g+1} = \frac{h_{i,g} w_{H,i}}{\bar{w}_{H,g}} \text{ , and } p_{j,g+1} = \frac{p_{j,g} w_{P,j}}{\bar{w}_{P,g}}$$

where $w_{H,i}$ ($w_{P,j}$) is the fitness of host genotype i (parasite genotype j). The average fitness of the host (parasite) population, $\bar{w}_{H,g}$ ($\bar{w}_{P,g}$), is obtained as $\sum_{i=1}^{2} w_{H,i} \cdot h_{i.g}$ (respectively, $\sum_{j=1}^{2} w_{P,j} \cdot p_{j,g}$). Every generation g a proportion ϕ_g (*i.e.* the disease prevalence) of the host population interacts with the parasite population in a frequency-dependent manner. Whether a particular interaction between host genotype i and parasite genotype j results in an infection or not depends on the matrix $\mathcal{A}$. An infection reduces the relative fitness of hosts by an amount s (cost of infection). Further, each host genotype i (parasite genotype j) can be associated with some fitness cost c_{H_i} (c_{P_j}), such as a cost of resistance (infectivity). Therefore, the frequencies of the different host and parasite genotypes can be modelled using the following recurrence equations:

$$h_{i,g+1} = \frac{h_{i,g} \cdot (1 - c_{H_i}) \cdot \left(1 - \phi_g \cdot s \cdot \sum_{j=1}^{2} \alpha_{ij} p_{j,g}\right)}{\bar{w}_{H,g}} \tag{1a}$$

$$p_{j,g+1} = \frac{p_{j,g} \cdot (1 - c_{P_j}) \cdot \left(\sum_{i=1}^{2} \alpha_{ij} h_{i,g}\right)}{\bar{w}_{P,g}} \tag{1b}$$

This dynamical system admits an equilibrium point when the conditions $h_{i,g+1} = h_{i,g} = \widehat{h_i}$ and $p_{j,g+1} = p_{j,g} = \widehat{p_j}$ hold for each host genotype i and each parasite genotype j. There are four so called trivial monomorphic equilibrium points at which one host and one parasite allele are fixed,

and one polymorphic equilibrium with frequencies:

$$\hat{h}_1 = \frac{\alpha_{22}(1-c_{P_2}) - \alpha_{21}(1-c_{P_1})}{(\alpha_{11}-\alpha_{21})(1-c_{P_1}) + (\alpha_{22}-\alpha_{12})(1-c_{P_2})} \tag{2a}$$

$$\hat{p}_1 = \frac{c_{H_1} - c_{H_2} + \phi s\,(\alpha_{12}(1-c_{H_1}) - \alpha_{22}(1-c_{H_2}))}{\phi s\,((\alpha_{12}-\alpha_{11})(1-c_{H_1}) + (\alpha_{21}-\alpha_{22})(1-c_{H_2}))} \tag{2b}$$

In line with previous studies, for both the symmetric and asymmetric MA model we assume no costs $c_{H_1} = c_{H_2} = c_{P_1} = c_{P_2} = 0$ (Gandon and Nuismer 2009). For the GFG model we use the infection matrix shown in Tab. 1c) and assume that $0 < c_{H_2}, c_{P_2} < 1$ and $c_{H_1} = c_{P_1} = 0$ (Tellier and Brown 2007). Previous work have shown that Model A only produces arms-race dynamics. Indeed, the trace of the Jacobian evaluated at the equilibrium point is zero and the non-trivial equilibrium point (2a) is an unstable saddle point leading to host and parasite allele frequencies with increasing amplitude and period over time (Tellier and Brown 2007).

2.1.2. Model B: model with epidemiological dynamics and feedback

In model B we consider a continuous time coevolutionary model (Živković et al. 2019) based on a known Susceptible-Infected model (Ashby and Boots 2017, Boots et al. 2014, May and Anderson 1983). This model allows for simultaneous changes in population sizes and allele frequencies. The total number of hosts of type i includes S_i susceptible and $\sum_j I_{ij}$ infected individuals. The change in number of susceptible hosts S_i is given by Eq. 3a and the change in number of infected individuals I_{ij} is given by Eq. 3b.

$$\frac{dS_i}{dt} = S_i\left[b(1-c_{H_i}) - \gamma - \sum_{j=1}^{2}\alpha_{ij}\beta(1-c_{P_j})\sum_{k=1}^{2}I_{kj}\right] + b(1-c_{H_i})(1-s)\sum_{j=1}^{2}I_{ij}, \tag{3a}$$

$$\frac{dI_{ij}}{dt} = -\gamma I_{ij} + S_i\left[\alpha_{ij}\beta(1-c_{P_j})\sum_{k=1}^{2}I_{kj}\right]. \tag{3b}$$

The number of parasites of type j is obtained as $P_j = \sum_i I_{ij}$ and hence, the change in number of parasites of type j is given by $\frac{dP_j}{dt} = \sum_i \frac{dI_{ij}}{dt}$. Hosts reproduce at natural birth rate b and die at natural death rate γ. The total host population size at generation t is $N(t) = \sum_{i,j} I_{ij}(t) + \sum_i S_i(t)$. We assume that there is no vertical disease transmission, and the infections are sustained in the populations through an overlap between generations. Uninfected hosts can get infected by horizontal disease

transmission with rate β. The costs c_{H_i}, c_{P_j} and s are defined as in model A. Previous analyses have shown that depending on the parametrization (chosen infection matrix and parameter values) this model results in a range of different dynamics (arms-race dynamics, trench-warfare dynamics with stable limit cycles and trench-warfare dynamics with a stable attractor) (Ashby and Boots 2017, Živković et al. 2019). We focus here chiefly on the trench warfare outcome, and especially when the dynamics converges to the stable (attractor) polymorphic equilibrium point.

To simulate the dynamics, we discretise model B into small time steps of size δ_t. Hence, one discrete generation t consists of $1/\delta_t$ time steps. The value of δ_t is chosen so that the discretised time dynamics matches with the continuous time expectation. The equilibrium points can be computed for this system (Živković et al. 2019) but as the formulae are complex and not very intuitive we refrain from using them here. The population size changes, allele frequencies changes and corresponding association statistics values are computed over time and at the equilibrium point. The disease prevalence is here an inherent property of the disease dynamics and allele frequencies as defined under the eco-evolutionary feedbacks (Ashby et al. 2019, Boots et al. 2014), and thus varies over time:

$$\phi(t) = \frac{\left(\sum_{j=1}^{2} \beta \left(1 - c_{P_j}\right) \sum_{k=1}^{2} I_{kj}(t)\right)}{N(t)}.$$

2.2. Definition of the association statistics

We assume that n_{T} host individuals have been sampled and genotyped at each biallelic single nucleotide polymorphism (SNP) in the genome, so that two types of hosts are found ($i \in (1, 2)$). The total host sample n_{T} consists of n_{Inf} infected hosts, the infected subsample, and n_{H} non-infected, healthy, hosts, the non-infected subsample. A number of n_{Par} parasite samples is obtained from the n_{Inf} infected hosts (one sample per host) and also genotyped at each biallelic SNP. Accordingly, there are also two parasite types for each biallelic SNP ($j \in (1, 2)$). Note that the sites typically considered here are SNPs as commonly used in GWAs and co-GWAs. Our definition also applies to any type of mutation with two states such as insertion-deletion (of few to many base pairs) or presence/absence polymorphism of larger genomic regions (*e.g.* coding genes).

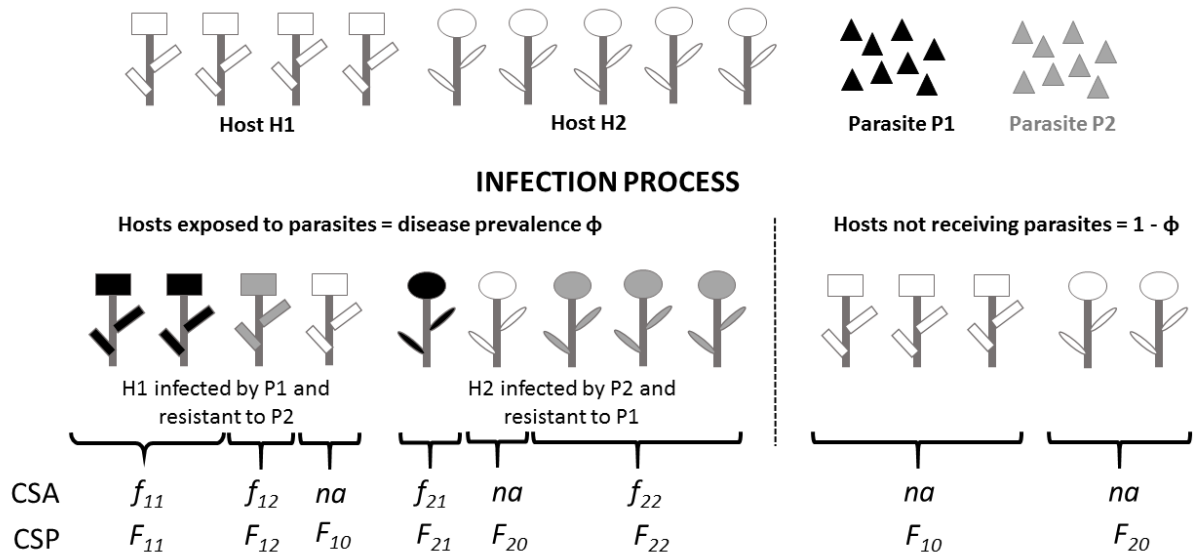

Figure 1 Graphic illustration of the properties of our indices CSA and CSP. The host population consists of two host types H_1 (square) and H_2 (circle) and the parasite population consists of two types P_1 (black) and P_2 (grey). A proportion ϕ of the hosts is exposed to parasites. Hosts which are exposed to the parasite either become infected or they can resist infection. Infected hosts are coloured based on the identity of the infecting parasite genotype (grey or black). f_{ij} is the proportion of hosts with type i which are infected by parasites of type j in the proportion of all infected hosts. F_{ij} is the proportion of hosts of type i being infect by parasites of type j in the whole host population (sum of all hosts). F_{i0} is the proportion of non-infected hosts of type i in the whole host population. F_{i0} is composed of hosts of type i which either did not receive spores $(1 - \phi)$ or which received spores but are resistant to the respective parasite.

2.2.1. The Cross-Species Association index (CSA)

We define the absolute Cross Species Association index (CSA) when sampling n_{Inf} hosts and $n_{\text{Par}} = n_{\text{Inf}}$ parasites as:

$$\text{CSA} = |f_{11}f_{22} - f_{21}f_{12}| \tag{4}$$

Here, f_{ij} is the number of hosts of type i being infected by a parasite of type j divided by the size of the infected subsample (n_{Inf}), so that $\sum_{\forall i,j} f_{ij} = 1$. This statistic is an adaptation of the well-known linkage disequilibrium (LD) measure in population genetics (Charlesworth and Charlesworth 2010, Lewontin and Kojima 1960, p371-373) and related to the statistics performed in Bartha et al. (2013) and Ansari et al. (2017).

Following population genetics theory (Charlesworth and Charlesworth 2010, p371-373), we normalize CSA in two different ways such that the absolute values range from 0 to 1. First, we define CSA′ which is obtained by normalizing each CSA value by the maximum CSA value possible, $\text{CSA}^{\text{max}} = 0.25$. CSA reaches its maximum value when hosts of type 1 are solely infected by parasites of type 1 and hosts of type 2 are solely infected by parasites of type 2 and $f_{11} = f_{22} = 0.5$ (or when hosts of type 1 are solely infected by parasites of type 2 and hosts of type 2 are solely infected by parasites of type 1 and $f_{12} = f_{21} = 0.5$).

$$\text{CSA}' = \frac{\text{CSA}}{\text{CSA}^{\text{max}}} = \frac{|f_{11}f_{22} - f_{21}f_{12}|}{0.25} = 4 \cdot \text{CSA} \tag{5}$$

Our second normalization consists in dividing the CSA value by the square root of the product of the frequencies of the different host and parasite alleles in the infected subsample.

$$\text{CSA}_{\text{r}} = \frac{f_{11}f_{22} - f_{21}f_{12}}{\sqrt{(f_{11}+f_{12})(f_{21}+f_{22})(f_{11}+f_{21})(f_{12}+f_{22})}} \tag{6}$$

We calculate the value of CSA at each generation g (eq. 4) based on our coevolutionary model A (eq. 1a):

$$\text{CSA}_g = \left| \frac{\alpha_{11}h_{1,g}p_{1,g}\alpha_{22}h_{2,g}p_{2,g} - \alpha_{21}h_{2,g}p_{1,g}\alpha_{12}h_{1,g}p_{2,g}}{\Delta^2} \right| \tag{7}$$

where $\Delta = \alpha_{11}h_{1,g}p_{1,g} + \alpha_{22}h_{2,g}p_{2,g} + \alpha_{21}h_{2,g}p_{1,g} + \alpha_{12}h_{1,g}p_{2,g}$ (introduced to make sure in eq. 4 that $\sum_{\forall i,j} f_{ij} = 1$).

For Model B, the CSA at each time step t is obtained as:

$$\text{CSA}(t) = \left| \frac{I_{11}(t) \cdot I_{22}(t) - I_{12}(t) \cdot I_{21}(t)}{(\sum_i \sum_j I_{ij}(t))^2}. \right| \tag{8}$$

Therefore, we can compute CSA$'$ and CSA$_r$ at each generation based on eq. 7 for Model A and based on eq. 8 for Model B.

2.2.2. The Cross-Species Prevalence index (CSP)

We define the Cross Species Prevalence index (CSP) at any generation at which n_{Inf} infected, n_{H} non-infected hosts and $n_{\text{Par}} = n_{\text{Inf}}$ pathogens are sampled.

$$\text{CSP} = \left| \frac{F_{11} + F_{12}}{F_{10}} - \frac{F_{21} + F_{22}}{F_{20}} \right| \tag{9}$$

Here, F_{ij} is the proportion of host type i infected by parasite type j in the total host sample (n_{T}). At the denominator, F_{i0} is the proportion of uninfected hosts of type i in the total sample. By definition, $\frac{n_{\text{Inf}}}{n_T} = F_{11}+F_{12}+F_{21}+F_{22}$, $\frac{n_T - n_{\text{Inf}}}{n_T} = F_{10}+F_{20}$, and $F_{11}+F_{12}+F_{21}+F_{22}+F_{10}+F_{20} = 1$ (see Fig. 1). Note that F_{i0} is composed of individuals 1) which do not encounter any parasite due to the incomplete disease prevalence in the population, and 2) which are exposed to parasites but are resistant.

When eq. 9 is applied to our coevolutionary model A, CSP at each generation g is obtained as:

$$\begin{aligned} \text{CSP}_g = \Bigg| & \frac{\phi_g(\alpha_{11} h_{1,g} p_{1,g} + \alpha_{12} h_{1,g} p_{2,g})}{(1-\phi_g) h_{1,g} + \phi_g((1-\alpha_{11}) h_{1,g} p_{1,g} + (1-\alpha_{12}) h_{1,g} p_{2,g})} \\ & - \frac{\phi_g(\alpha_{21} h_{2,g} p_{1,g} + \alpha_{22} h_{2,g} p_{2,g})}{(1-\phi_g) h_{2,g} + \phi_g((1-\alpha_{21}) h_{2,g} p_{1,g} + (1-\alpha_{22}) h_{2,g} p_{2,g})} \Bigg| \end{aligned} \tag{10}$$

For Model B, the CSP at time t is given by:

$$\text{CSP}(t) = \left| \frac{I_{11}(t) + I_{12}(t)}{S_1(t)} - \frac{I_{21}(t) + I_{22}(t)}{S_2(t)} \right| \tag{11}$$

Irrespective of the model, CSP is only defined as long as there are some uninfected individuals of both host types.

3. Results

3.1. Analytical results for model A

We first present some analytical results by computing CSA and CSP for the population genetics Model A with either a matching-allele (MA) or a gene-for-gene (GFG) interaction to provide some intuition on the behaviour of the presented indices. In the calculations, we only focus on CSA, as it is straightforward to obtain CSA′ and CSA$_r$ by applying the respective normalizations.

3.1.1. Under the Matching Allele infection matrix

For a matching-allele infection matrix and for $c_{H_1} = c_{H_2} = c_{P_1} = c_{P_2} = 0$ the equations for model A (eq. 1a) reduce to:

$$\begin{aligned} h_{1,g+1} &= \frac{h_{1,g}\left(1-\phi_g s\left[p_{1,g}+(1-c_1)p_{2,g}\right]\right)}{\bar{w}_H}, \text{ and } h_{2,g+1} = \frac{h_{2,g}\left(1-\phi_g s\left[p_{2,g}+(1-c_2)p_{1,g}\right]\right)}{\bar{w}_H}. \\ p_{1,g+1} &= \frac{p_{1,g}\left(h_{1,g}+h_{2,g}(1-c_2)\right)}{\bar{w}_P}, \text{ and } p_{2,g+1} = \frac{p_{2,g}(h_{2,g}+h_{1,g}(1-c_1))}{\bar{w}_P}. \end{aligned} \tag{12}$$

By applying eq. 7 and eq. 9 to these MA-equations, we obtain CSA$_{g,\text{MA}}$ and CSP$_{g,\text{MA}}$ at any generation g.

$$\text{CSA}_{g,\text{MA}} = \left|\frac{(c_1+c_2-c_1c_2)h_{1,g}h_{2,g}p_{1,g}p_{2,g}}{\left(1-c_2h_{2,g}p_{1,g}-c_1h_{1,g}p_{2,g}\right)^2}\right|, \tag{13}$$

$$\text{CSP}_{g,\text{MA}} = \left|\frac{\phi_g\left(c_2p_{1,g}-c_1p_{2,g}\right)}{\left(1-\phi_g\left(1-c_1p_{2,g}\right)\right)\left(1-\phi_g\left(1-c_2p_{1,g}\right)\right)}\right|. \tag{14}$$

It is evident that CSP, by contrast to the CSA, does not depend on the frequencies of the different host types but only on the parasite frequencies. Moreover, the CSP cannot be computed if the disease prevalence is at maximum ($\phi_g = 1$) and if neither of the host alleles provides any resistance to any parasite genotype ($c_1 = c_2 = 0$).

Further, the matching-allele model formulated in equation 12 has four monomorphic equilibria and one polymorphic equilibrium. The frequencies of the latter are given by:

$$\widehat{p}_1 = \widehat{h}_2 = \frac{c_1}{c_1+c_2}, \text{ and } \widehat{h}_1 = \widehat{p}_2 = \frac{c_2}{c_2+c_1}. \tag{15}$$

Inserting these equilibrium frequencies into eq. 13 and eq. 14 we can obtain the values of the indices at the polymorphic equilibrium point.

$$\widehat{\text{CSA}}_{\text{MA}} = \frac{c_1^2 c_2^2}{(c_1+c_2)^2\,(c_1+c_2-c_2c_1)}\text{ , and } \widehat{\text{CSP}}_{\text{MA}} = 0. \tag{16}$$

For a matching alleles interaction without any genotype costs, CSP is always zero at the equilibrium point, irrespective of the values of c_1 and c_2. The values of the CSA and CSP display a different behaviour over time and at the equilibrium. Thus, comparing their values over time can yield insights into the asymmetry of the infection matrix.

3.1.2. Under the Gene-For-Gene infection matrix

For a gene-for-gene infection matrix and for $0 < c_{H_2}, c_{P_2} < 1$ and $c_{H_1} = c_{P_1} = 0$, the equations for the coevolutionary model A (eq. 1a) reduce to:

$$\begin{aligned} h_{1,g+1} &= \frac{h_{1,g}(1-s\phi_g)}{\bar{w}_H}\text{ , and } h_{2,g+1} = \frac{h_{2,g}(1-c_{H_2})\,(1-\phi_g s\,[(1-c)p_{1,g}+p_{2,g}])}{\bar{w}_H}, \\ p_{1,g+1} &= \frac{p_{1,g}(h_{1,g}+h_{2,g}(1-c))}{\bar{w}_P}\text{ , and } p_{2,g+1} = \frac{p_{2,g}(1-c_{P_2})}{\bar{w}_P}. \end{aligned} \tag{17}$$

Applying eq. 7 and eq. 9 to the GFG system of equation, yields the following values of $\text{CSA}_{g,\text{GFG}}$ and $\text{CSP}_{g,\text{GFG}}$ at some generation g.

$$\text{CSA}_{g,\text{GFG}} = \frac{c h_{1,g} h_{2,g} p_{1,g} p_{2,g}}{(1-c h_{2,g} p_{1,g})^2}, \tag{18}$$

$$\text{CSP}_{g,\text{GFG}} = \frac{\phi_g c p_{1,g}}{(1-\phi_g)(1-\phi_g(1-c\cdot p_{1,g}))}. \tag{19}$$

As for the MA model, the CSP values, by contrast to the CSA, do not depend on the frequencies of the different host types but only on the the parasite frequencies. The conditions for computing the CSP are more restrictive than under the MA, as CSP is not defined as soon as the disease prevalence is maximum ($\phi_g = 1$) and therefore, all hosts of type 1 are infected.

The polymorphic equilibrium frequencies of this model A with GFG are given by:

$$\begin{aligned} \widehat{p}_1 = \frac{c_{H_2}(1-s\phi)}{c\phi s(1-c_{H_2})} \quad &\text{, and} \quad \widehat{h}_1 = \frac{c-c_{P_2}}{c}, \\ \widehat{p}_2 = 1-\frac{c_{H_2}(1-s\phi)}{c\phi s(1-c_{H_2})} \quad &\text{, and} \quad \widehat{h}_2 = \frac{c_{P_2}}{c}. \end{aligned} \tag{20}$$

Inserting these frequencies into eq. 18 and 19 we can obtain the values of the indices at the polymorphic equilibrium point.

$$\widehat{\text{CSA}}_{\text{GFG}} = \frac{\frac{c_{P_2} c_{H_2} \left(c - c_{P_2}\right)(1 - s\phi)}{c^2 \phi s \left(1 - c_{H_2}\right)} \left(1 - \frac{c_{H_2}(1 - s\phi)}{c\phi s \left(1 - c_{H_2}\right)}\right)}{\left(1 - \frac{c_{P_2} c_{H_2}(1 - s\phi)}{c\phi s \left(1 - c_{H_2}\right)}\right)^2}, \tag{21}$$

$$\widehat{\text{CSP}}_{\text{GFG}} = \frac{c_{H_2}\left(1 - \phi s\right)}{(1 - \phi)\left((1 - \phi)\, s + c_{H_2}\left(1 - s\right)\right)}. \tag{22}$$

To gain a deeper understanding of these results, we conduct numerical simulations for both types of interactions over 500 generations and compare the values of the CSA_r/CSA′/CSP over time to the detection thresholds obtained for neutral loci.

3.2. Detection thresholds for CSA and CSP

In order to evaluate the power of CSA and CSP to pinpoint coevolutionary loci, it is necessary to derive threshold (cut-off) values for these indices based on all possible comparisons of pairs of loci from host and parasite genome data. Any pair of host and parasite SNPs exhibiting a value above the cut-off is classically considered as a strong candidate pair governing the outcome of infection and hence, to be under coevolution. It is common to obtain these cut-off values from the distribution of all empirical values of a given data set (using ad hoc multiple testing correction, see Bartha et al. (2013), Ansari et al. (2017)). By contrast, we assume here that sequences are obtained from a random samples of infected and non-infected hosts from a panmictic, natural or controlled, population with recombination. We thus can derive the expected distribution of CSA (CSA′ and CSA_r, correspondingly) and CSP for pairwise comparisons of many neutral host and parasite SNPs based on population genetic assumptions, namely that allele frequencies are distributed according to a neutral site-frequency spectrum (SFS).

By definition, neutral host and parasite loci are not determining the outcome of infection, and therefore, each neutral host SNP-neutral parasite SNP-pair is characterized by an infection matrix $\mathcal{A}_{neutral} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$. Full genome data usually contain a large number of neutral SNPs which are distributed across the whole genome. Given that the recombination rate is high enough, these neutral SNPs are assumed to evolve independently from one another other and therefore, their allele frequencies in the population and in the sample are mutually independent. In order to obtain

the expected neutral distribution of CSA and CSP we compute first all possible combinations of host and parasite alleles in the sample given a minor allele frequency count and the neutral infection matrix $\mathcal{A}_{\text{neutral}}$. Second, we compute the respective CSA/CSP value for each combination of host and parasite allele frequencies. Third, the expectations of CSA and CSP in a sample from a natural population at mutation-drift equilibrium is obtained by obtaining from the SFS the probability p that the host minor allele frequency count in the sample is v and the probability q that the parasite minor allele frequency count in the sample is w.

Our sample consists of n_{Inf} hosts and one representative parasite strain from each of these infected hosts (thus $n = n_{\text{Inf}} = n_{\text{par}}$). The expected CSA for neutral SNPs ($E(\text{CSA}_{vw})$) is measured for the association between a host SNP with minor allele frequency v and a parasite SNP with minor allele frequency w and is given by:

$$E(CSA_{vw}) = \Omega_{vw} \sum_{k=0}^{l} \frac{\binom{v}{k}\binom{n-v}{w-k}}{\binom{n_{\text{Inf}}}{w}} \left| \frac{k n_{\text{Inf}} - vw}{n_{\text{Inf}}^2} \right| \tag{23}$$

where $l = min(v, w)$, and Ω_{vw} is the normalization for either obtaining CSA′ (with $\Omega_{vw} = 4$, from eq. 5) or CSA$_r$ (with $(\Omega_{vw} = \frac{1}{\sqrt{\frac{v}{n_{\text{Inf}}}\frac{n-v}{n_{\text{Inf}}}\frac{w}{n_{\text{Inf}}}\frac{n-w}{n_{\text{Inf}}}}})$, from eq. 6).
For the CSP, the computation is similar, though more complex, and the expectation is given by:

$$E(CSP_{vw}) = \sum_{z=\rho}^{m-1} \frac{\binom{n_{\text{Inf}}}{z}\binom{n_H}{v-z}}{\binom{n_{\text{T}}}{v} - \sum_{b=0}^{\rho-1}\binom{n_{\text{Inf}}}{b}\binom{n_H}{v-b} - \sum_{b=m}^{v}\binom{n_{\text{Inf}}}{b}\binom{n_H}{v-b}} \left| \frac{a}{\lambda n_H + (v-z)(-1)^{\lambda}} - \frac{n_{\text{Inf}} - a}{(1-\lambda)n_H - (v-z)(-1)^{\lambda}} \right| \tag{24}$$

with $\rho = max(1, v - n_H + 1)$, $m = min(n_{\text{Inf}}, v)$, $a = min(z, n_{\text{Inf}} - z)$, and

$$\lambda = \begin{cases} 0 & \text{for } z \leq n_{\text{Inf}} - z \\ 1 & \text{for } z > n_{\text{Inf}} - z \end{cases}$$

Computational details are given in the Online Supplementary Material. The expectations of CSA and CSP over all pairwise comparisons of neutral host and parasite SNPs, are obtained by weighting each $E(\text{CSA}_{vw})$ and $E(\text{CSP}_{vw})$ value by the probability that a neutral host SNP has minor allele

frequency v (p_v) and a neutral parasite SNP has minor allele frequency w (q_w) in the sample:

$$E(CSA) = \sum_{v=1}^{\lfloor n/2 \rfloor} \sum_{w=1}^{\lfloor n/2 \rfloor} p_v q_w E(CSA_{vw}), \tag{25}$$
$$E(CSP) = \sum_{v=1}^{\lfloor n/2 \rfloor} \sum_{w=1}^{\lfloor n/2 \rfloor} p_v q_w E(CSP_{vw}).$$

These probabilities are obtained from the relative folded site frequency spectrum of the sample (eq. A2). We use the folded SFS, as we assume that there is no outgroup sequence available and thus, the ancestral and derived state are indistinguishable at each SNP. For simplicity, we assume that the allele frequency distribution of neutral SNPs in both, the host and the parasite, follows the SFS under drift-mutation equilibrium for a Wright-Fisher model under constant population size (Durrett (2010), p.50). These probabilities are found in the Appendix (eq. A2). We compute the distributions (and extract the values of the 95 and 99 percentiles) for CSA and CSP for different sample sizes and sampling schemes. In the following we show the results for $n_T = 200$ and $n_{\text{Inf}} = n_{\text{H}} = 100$ (See the Online Supplementary Material for cut-off values under other sampling schemes).

3.3. Numerical simulations: Temporal changes of CSA/CSP and detection thresholds

When simulating an asymmetric MA interaction ($c_1 = 0.9$, $c_2 = 0.7$, model A) over 500 generations (Fig. 2) coevolution results in arms-race dynamics under Model A. We observe that CSA and CSP fluctuate over time due to the coevolutionary cycles and the associated allele frequency changes. Overall, the CSA values decrease over time and are constantly found below the detection threshold after $g = 300$ generations. Therefore, under unstable coevolutionary dynamics with increasing amplitude and period of the coevolutionary cycles (resulting ultimately in fixation of alleles), the associations between hosts and parasites alleles in the infected sample become too weak to be observable (Fig. 2c,d). Under an arms-race, one host allele occurs in very high frequency and the parasite tracks this frequency down over time which generates the coevolutionary cycles. During these cycles there is only a very limited amount of time at which both hosts and parasites alleles are found in intermediate frequencies yielding high values of CSA index. On the other hand, the CSP values under the MA model are consistently high and exhibit enough statistical power to detect the loci under coevolution (Fig. 2d). This demonstrates the importance of obtaining

additional non-infected host samples. Under the same model A with symmetric MA, we find similar outcome as in Fig. 2, albeit the oscillations of CSA and CSP values are perfectly matching the allele frequency cycles and show a regular amplitude pattern (Fig. S8).

Under the GFG model with arms race, CSA values are consistently small with narrow peaks (Fig. 3b,c) which are barely above the detection threshold. The CSP has several peaks above the cut-off value, yet it maybe difficult to detect the coevolutionary locus even when time samples are available if the more stringent 0.99-cut-off level is applied (Fig. 3d). The comparison of MA and GFG arms race dynamics shows that the combination of CSA and CSP values over time gives some indication about the symmetry of the infection matrix. Furthermore, the CSP exhibits the highest power to disentangle loci under coevolution from the neutral background but also to infer the asymmetry of the infection matrix. Generally, model A always results in arms-race dynamics, however the amplitude and the time to fixation are affected by the underlying infection matrix and the coevolutionary costs.

Under Model B trench-warfare dynamics can take place for MA-interactions and allele frequencies can converge to a stable polymorphic equilibrium. Once the allele frequencies are at equilibrium, CSA has a very strong power to distinguish the coevolving loci from the neutral background, while the CSP decreases to zero (as shown above in eq. 16). In Model B, the disease prevalence varies over time as a function of the changes in allele frequencies, so that it is expected that CSP varies over time. However, once the allele frequencies reach a stable equilibrium also the numbers in all host compartments eventually remain constant. When the ratio of infected to non-infected individuals is the same for both host types then CSP drops to zero. The value of CSA at equilibrium depends on the respective equilibrium frequencies and thus, is highest when alleles are in frequency 0.5 (eq. 16). Even if coevolution results in stably sustained cycles, the CSA values remain high as long as the amplitude of the cycles does not become too large and therefore allele do not reach too high or too low frequencies (close to the boundaries). When allele frequency fluctuations do not occur and the system is already at the polymorphic equilibrium, CSA is fixed to a constant and high value (Fig. S9), while the CSP is fixed to zero (under a symmetric MA infection matrix). Note that the epidemiological model B can also generate arms race dynamics with a consequent fixation of one host and one parasite allele for some parameter combinations under a GFG-interaction (Fig. S10). In such cases, the obtained results are similar to those of Fig. 3 with both indices dropping to zero over time. Finally, we note that the two measures of CSA we introduce, CSA' and CSA_r show the same trend and similar power under an arms-race, while

the CSA_r is slightly more precise under trench warfare when allele frequencies reach very high or very low values (close to fixation or loss). Generally, the parasite population is dying out when the disease transmission rate (β) is either too high or too low, irrespective of the underlying infection matrix. For intermediate disease transmission rates under a MA-interaction, we either observe stable limit cycles for large values of s ($s \approx 1$) or convergence to a stable equilibrium for $s < 1$. These qualitative trends are similar when changing the infection matrix from MA to GFG, tough the number of cycles to reach the stable equilibrium varies (Živković et al. 2019). Therefore, the highlighted results from Model B reflect the most commonly observed dynamics in this model.

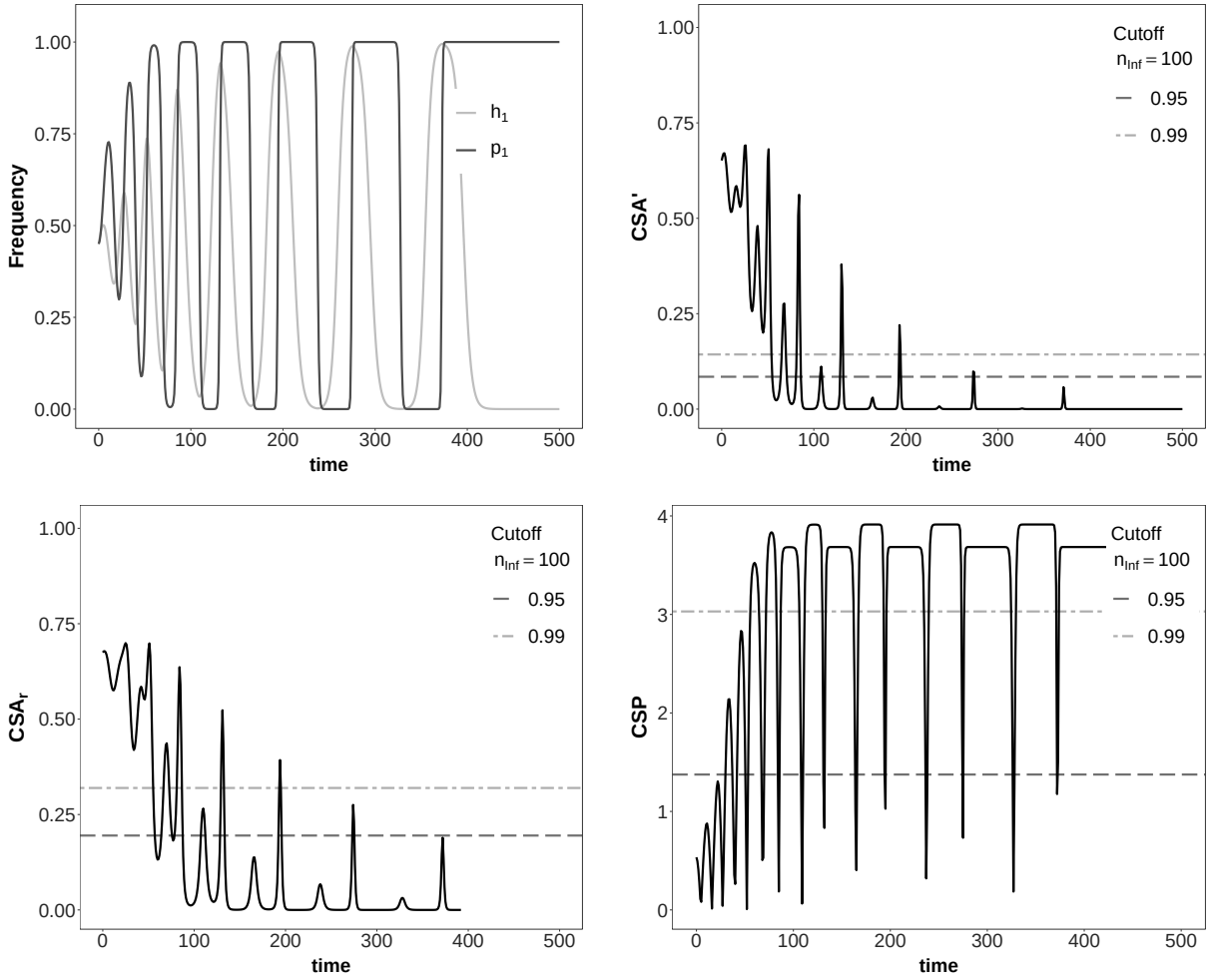

Figure 2 Temporal changes in allele frequencies, CSA', CSA'_r and CSP in an unstable asymmetric MA-model (model A) with one parasite generation per host generation. For each index cut-off values are shown based on the expected neutral distributions for a total host sample size $n_T = 200$ and for $n_{\text{Inf}} = n_{\text{H}} = 100$. The 0.95-cut-off value is shown as a dashed line and the 0.99-cut-off value as a dotted-dashed line. Top left: frequencies of h_1 (light grey) and p_1 (dark grey). Top right: CSA'. Bottom left: CSA_r. Bottom right: CSP. The model parameters are $c_1 = 0.9$, $c_2 = 0.7$, $\phi = 0.8$, $s = 0.35$, initial values $h_{1,g=0} = p_{1,g=0} = 0.45$.

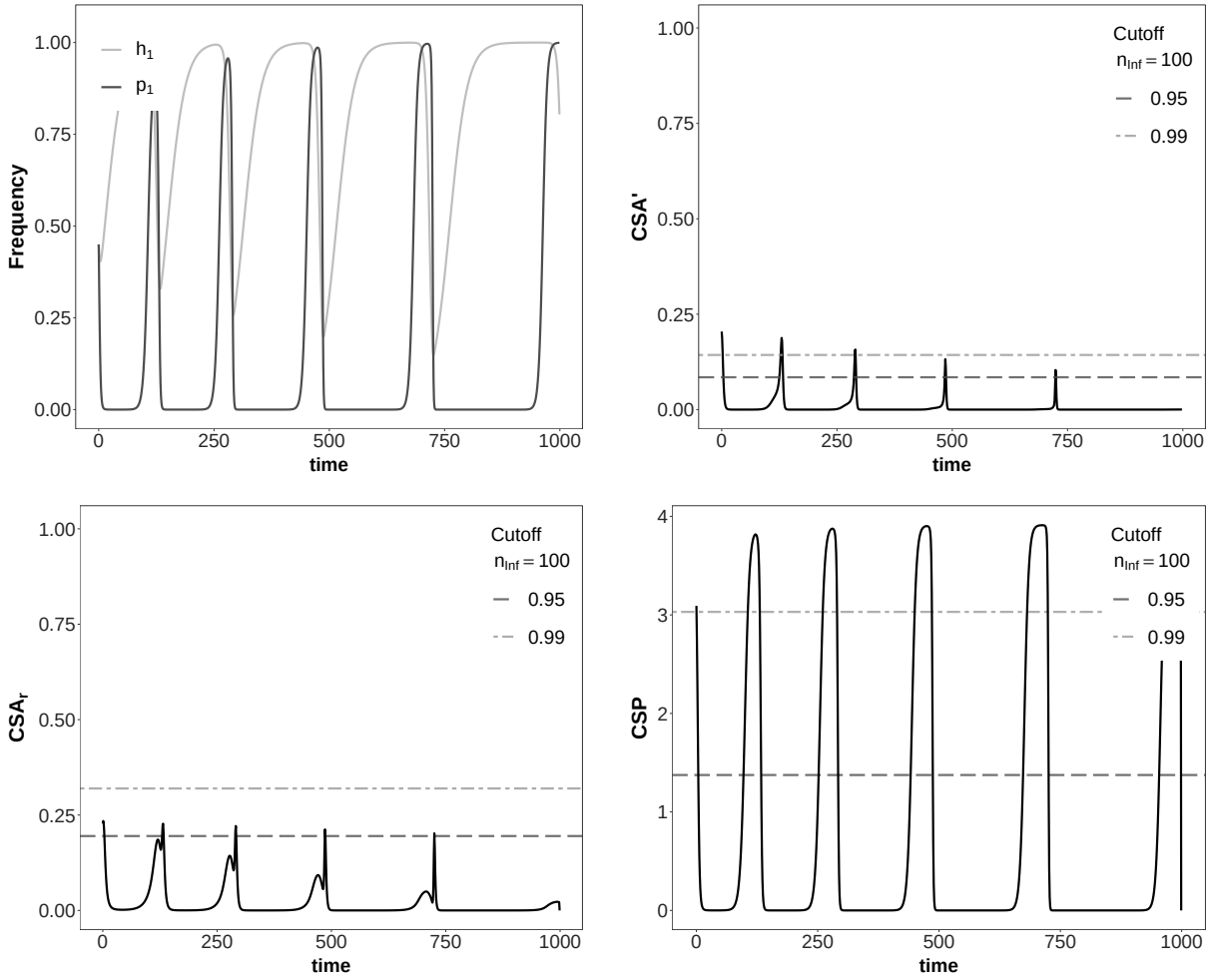

Figure 3 Temporal changes in allele frequencies, $\mathrm{CSA}', \mathrm{CSA}'_r$ and CSP in an unstable GFG-model (model A) with one parasite generation per host generation. For each index cut-off values are shown based on the expected neutral distributions for a total host sample size $n_T = 200$ and for $n_{\mathrm{Inf}} = n_{\mathrm{H}} = 100$. The 0.95-cut-off value is shown as a dashed line and the 0.99-cut-off value as a dotted-dashed line. Top left: frequencies of h_1 (light grey) and p_1 (dark grey). Top right: CSA'. Bottom left: CSA_r. Bottom right: CSP. The model parameters are $c_{H_1} = c_{P_1} = 0$, $c_{H_2} = 0.05$, $c_{P_2} = 0.2$, $\phi = 0.8$, $s = 0.35$, $c = 0.9$, initial values $h_{1,g=0} = p_{1,g=0} = 0.45$.

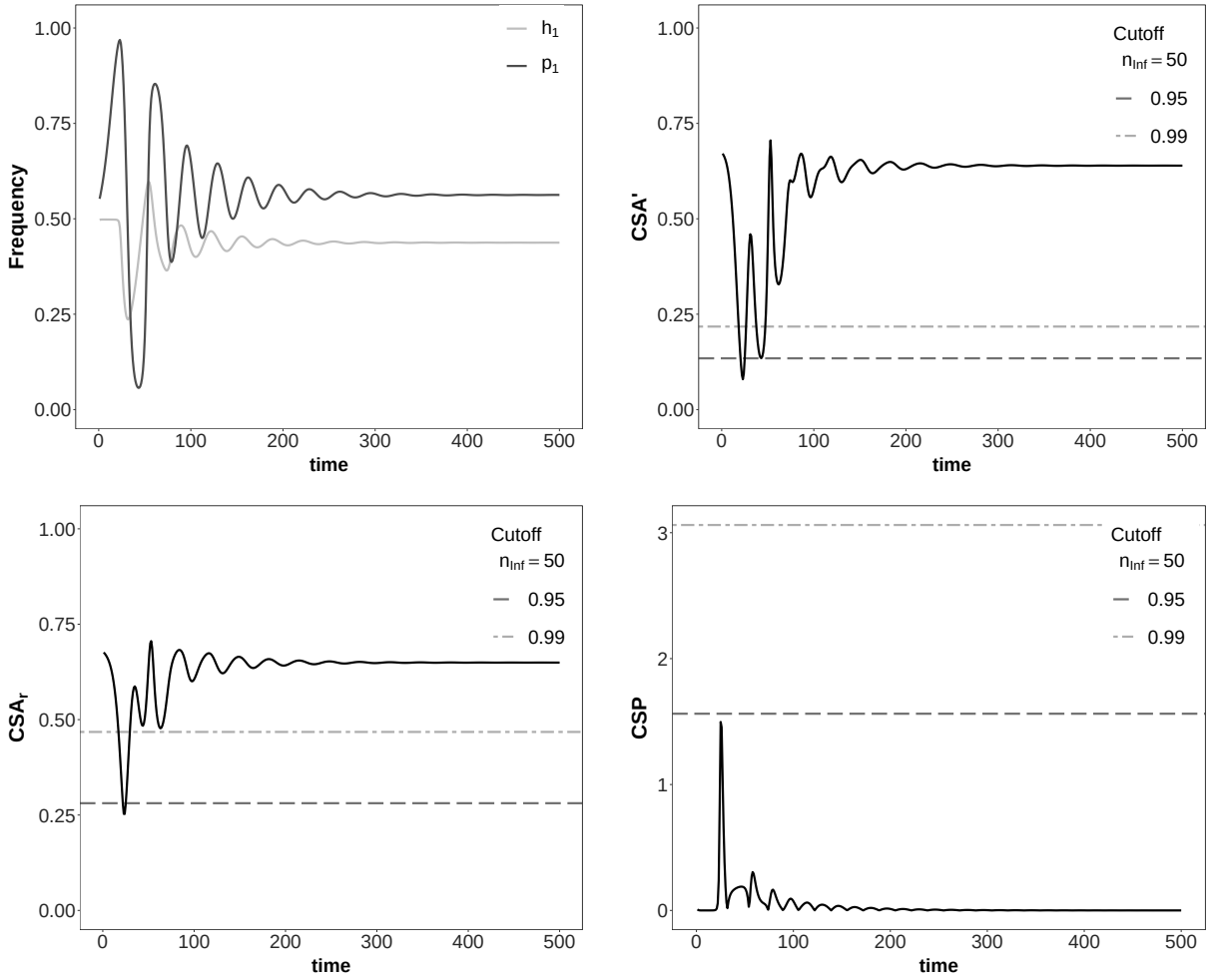

Figure 4 Temporal changes in allele frequencies, CSA′, CSA′$_r$ and CSP in an epidemiological (model B) with an asymmetric MA-infection matrix. For each index cut-off values are shown based on the expected neutral distributions for a total host sample size $n_T = 200$ and for $n_{\text{Inf}} = n_{\text{H}} = 100$. The 0.95-cut-off value is shown as a dashed line and the 0.99-cut-off value as a dotted-dashed line. Top left: frequencies of h_1 (light grey) and p_1 (dark grey). Top right: CSA′. Bottom left: CSA$_r$. Bottom right: CSP. The model parameters are $\beta = 0.00005$, $s = 0.6$, $c_1 = 0.9$, $c_2 = 0.7$, $b = 1$, $\gamma = 0.9$, and $c_{H_1} = c_{P_1} = c_{H_2} = c_{P_2} = 0$. The initial values are $S_{1,t=0} = S_{2,t=0} = 4150$, $I_{11,t=0} = I_{12,t=0} = I_{21,t=0} = I_{22,t=0} = 415$. The time intervals for computations is $\delta_t = 0.001$.

4. Discussion

With the technological advances, it has become feasible to sequence full genomes of several hosts from a given population as well as the parasite strains infecting them. Recent studies by Ansari et al. (2017), Bartha et al. (2013) test for the degree of association between all the host and parasite SNPs in a pairwise manner to obtain a genome wide natural co-GWAs. Here, we show that the power of such studies can be improved by including additional sequence data from non-infected individuals. Further, we derive cut-off values for significant associations based on simple population genetics assumptions. Finally, we demonstrate that the power to identify the loci underlying coevolution and thus, determining the infection outcome and the phenotype, varies in time and depends on the asymmetry of the underlying infection matrix.

We expect *a priori* that the power to detect coevolutionary loci varies in time due to the coevolutionary dynamics and the respective allele frequency changes at the involved loci. Our approach is similar in spirit to studies measuring local adaptation by performing reciprocal transplant or common garden experiments across several host-parasite populations being connected by migration (Gandon and Nuismer 2009, Nuismer et al. 2017). The CSA measure is closely related to the covariance computed in Gandon and Nuismer (2009), Nuismer et al. (2017), which shows also variable statistical power over time to detect coevolution. In contrast to reciprocal transplant or common garden (Gandon and Nuismer 2009, Nuismer et al. 2017) experiments and host-parasite controlled co-GWAs (MacPherson et al. 2018, Wang et al. 2018), the design of our natural co-GWA (cross-species association) study already implicitly contains phenotypic information. The infection experiment has been already ”performed” by nature.

As indicated by analytical and simulation results, the two indices we introduce, CSA and CSP, provide different information regarding 1) the symmetry of the infection matrix, 2) the type of dynamics, and 3) whether the allele frequencies have reached a stable polymorphic equilibrium point. We suggest to obtain samples from several time points rather than a single time point to extract as much information as possible. First, in order to infer the infection matrix, we see that two time samples for a MA or GFG interaction with an arms-race would likely both yield low values of CSA, while the value of CSP would be comparatively high for a MA interaction and low for a GFG interaction. Additional time samples would allow to capture the finer patterns of regular peak behaviour. Second, two or few time samples are enough to allow for the inference of the type

of coevolutionary dynamics, namely arms-race or trench-warfare. A similar idea was proposed in Gandon et al. (2008) for the study of local adaptation, when using phenotypic (infection) data. In our case, increasing amplitudes of allele frequencies (arms race) leads to a decrease of both CSA and CSP values over time, while decreasing amplitudes (trench-warfare dynamics) of coevolutionary cycles generate increasing CSA values. Once these values remain constant across time-samples a stable equilibrium point hast been reached.

One crucial result of our study is the derivation of the neutral expectations for CSA and CSP which allows us to compute detection thresholds for different sample sizes. A major difference to the controlled co-GWAs studies (MacPherson et al. 2018, Wang et al. 2018) and our approach is that the sequenced samples of infected and non-infected hosts are random samples from panmictic reproducing host and parasite populations. This allows us to apply population genetics theory to derive the expected power of our indices. To do so, we assume that host and parasite populations are at drift-mutation equilibrium. Host and parasite population sizes (termed as the demographic history in population genetics) are not necessarily constant over time either due to 1) eco-evolutionary feedback arising from the epidemiological dynamics (as for example in model B), and/or 2) due to abiotic environmental factors such as resource availability and habitat suitability. Irrespective of the source causing the respective population size changes, it is possible to compute the neutral SFS for both demographic scenarios based on previous work (Živković et al. 2019, 2015). This information could be used in eq. 25 to calculate the expectation of CSA and CSP for neutral loci.
Note also that the influence of the available sample sizes on cut-off values can be large, as small sample sizes ($n_T < 25$) decrease substantially the power to detect the coevolving loci (see Supplementary Online Material). We suggest therefore as a strategy to collect at few time points and obtain large sample sizes, rather than smaller sample sizes at many time points.

One simplifying assumption in our approach is the strict one to one relationship between the host and the parasite, *i.e.* one parasite sample obtained per host. However, co-infections are common for many diseases (Alizon et al. 2013, Tollenaere et al. 2016). A solution to deal with such cases is to only use the major parasite strain found on the infected host (Bartha et al. 2013).
The presented results are further based on the assumption that only one major gene per species is involved. However, our indices are applied independently to each pair of host and parasite

loci. Therefore, our approach is also suited to capture several major loci involved, as long as they are freely recombining and as there are no epistatic or pleitropic effects on the infection outcome. In general, the approach presented here is also potentially applicable to detect the major genes which are determining the compatibility between symbionts in mutualistic interactions. It is straightforward to study the power of the presented indices for such types of interaction by adjusting the equations governing the coevolutionary dynamics accordingly.

We conclude by presenting a set of recommendations for applying this method to different host-parasite systems. It is advised to obtain infected and non-infected hosts and parasite random samples at several time points from natural populations or from controlled coevolution experiments. It has been shown that polycyclic parasites, that is parasites with several infection cycles/generations per host generation, track down the host frequencies within a host generation (Brown and Tellier 2011). Therefore, it is expected that the cross-species association method has a high power when applied to parasite with strong life-cycle dependency on their hosts. This effect should be strongly pronounced for parasites which have a much shorter generation time than their hosts (viruses (Ansari et al. 2017, Bartha et al. 2013) or bacteria). For such types of parasites taking serial samples within a single host generations should help to pinpoint coevolutionary loci very accurately. Here, the coevolving loci are expected to show an increasing association with the corresponding host loci over the course of a single host generation. However, the assumption of independence between neutral and coevolving loci can be violated when studying viruses, bacteria or clonal fungi due to the absence of recombination. Our approach to disentangle neutral loci from candidate loci based on threshold values for the indices relies on recombination. In Bartha et al. (2013), the hurdle of linkage was overcome by first performing a phylogeny of the virus samples, and a subsequent identification of clusters of polymorphic SNPs across the phylogeny. Based on these clusters the association study was performed in a second step.

An implicit assumption in Bartha et al. (2013), Ansari et al. (2017), Bartoli and Roux (2017) and our model is that disease transmission is random and panmictic so that potentially every host can get in contact with the disease (no population sub-structuring affecting disease transmission). Thus, it is crucial to assess the extent of population structure before performing a cross-species association study as different populations can be at different stages of the coevolutionary cycle (*e.g.* Sasaki (2000), Gavrilets and Michalakis (2008)). Neglecting population structure in such cases can

result in biased results.

Nevertheless, one could take advantage of the geographic mosaic of coevolution by obtaining sequence data from several populations when sequence data from several time points are not feasible. However, note that the neutral SFS of pooled samples from a spatially structured population does not follow the equations we used here. Hence, this has to be taken into account when obtaining the threshold values for CSA and CSP. A description of the effect of spatial population structure on allele frequency distributions can be for example found in Wakeley and Aliacar (2002) and Staedler et al. (2009).

Host-parasite coevolution is a multifaceted process. We have shown the power of natural co-GWAs to gain insights into this process, especially when time-samples are available, despite potential shortcomings due to model assumptions. A further worthwhile development will be the explicit inclusion of spatial structure of host and parasite populations.

References

Agrawal, A., Lively, C.M., 2002. Infection genetics: gene-for-gene versus matching-alleles models and all points in between. Evolutionary Ecology Research 4, 79–90.

Alizon, S., de Roode, J.C., Michalakis, Y., 2013. Multiple infections and the evolution of virulence. Ecology Letters 16, 556–567. doi:10.1111/ele.12076.

Ansari, M.A., Pedergnana, V., Ip, C.L.C., Magri, A., Von Delft, A., Bonsall, D., Chaturvedi, N., Bartha, I., Smith, D., Nicholson, G., McVean, G., Trebes, A., Piazza, P., Fellay, J., Cooke, G., Foster, G.R., Hudson, E., McLauchlan, J., Simmonds, P., Bowden, R., Klenerman, P., Barnes, E., Spencer, C.C.A., Consortium, S.H., 2017. Genome-to-genome analysis highlights the effect of the human innate and adaptive immune systems on the hepatitis C virus. NATURE GENETICS 49, 666+. doi:10.1038/ng.3835.

Ashby, B., Boots, M., 2017. Multi-mode fluctuating selection in host-parasite coevolution. Ecology Letters 20, 357–365. doi:10.1111/ele.12734.

Ashby, B., Iritani, R., Best, A., White, A., Boots, M., 2019. Understanding the role of eco-evolutionary feedbacks in host-parasite coevolution. Journal of Theoretical Biology 464, 115–125. doi:10.1016/j.jtbi.2018.12.031.

Bartha, I., Carlson, J.M., Brumme, C.J., McLaren, P.J., Brumme, Z.L., John, M., Haas, D.W., Martinez-Picado, J., Dalmau, J., Lopez-Galindez, C., Casado, C., Rauch, A., Guenthard, H.F., Bernasconi, E., Vernazza, P., Klimkait, T., Yerly, S., O'Brien, S.J., Listgarten, J., Pfeifer, N., Lippert, C., Fusi, N., Kutalik, Z., Allen, T.M., Mueller, V., Harrigan, P.R., Heckerman, D., Telenti, A., Fellay, J., to Genome Study, H.G., Study, S.H.C., 2013. A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. eLife 2. doi:10.7554/eLife.01123.

Bartoli, C., Roux, F., 2017. Genome-Wide Association Studies in plant pathosystems: Toward an ecological genomics approach. Frontiers in Plant Science 8. doi:10.3389/fpls.2017.00763.

Boots, M., White, A., Best, A., Bowers, R., 2014. How specificity and epidemiology drive the coevolution of static trait diversity in hosts and parasites. Evolution 68, 1594–1606. doi:10.1111/evo.12393.

Brown, J.K.M., Tellier, A., 2011. Plant-parasite coevolution: Bridging the gap between genetics and ecology, in: VanAlfen, NK and Bruening, G and Leach, JE (Ed.), Annual Review of Phytopathology, VOL 49. volume 49 of *Annual Review of Phytopathology*, pp. 345–367. doi:10.1146/annurev-phyto-072910-09530l.

Charlesworth, B., Charlesworth, D., 2010. Elements of evolutionary genetics. volume 1. Roberts and Company Publishers, Greenwood Village.

Dawkins, R., Krebs, J., 1979. Arms races between and within species. Proceedings of the Royal Society Series B-Biological Sciences 205, 489–511. doi:10.1098/rspb.1979.0081.

Dobon, A., Bunting, D.C.E., Cabrera-Quio, L.E., Uauy, C., Saunders, D.G.O., 2016. The host-pathogen interaction between wheat and yellow rust induces temporally coordinated waves of gene expression. BMC Genomics 17. doi:10.1186/s12864-016-2684-4.

Durrett, R., 2010. Probability models for DNA sequence evolution. volume 2. Springer, New York.

Ebert, D., 2018. Open questions: what are the genes underlying antagonistic coevolution? BMC Biology 16. doi:10.1186/s12915-018-0583-7.

Frickel, J., Feulner, P.G.D., Karakoc, E., Becks, L., 2018. Population size changes and selection drive patterns of parallel evolution in a host–virus system. Nature Communications 9, 1706. URL: http://dx.doi.org/10.1038/s41467-018-03990-7%0Ahttp://www.nature.com/articles/s41467-018-0399 doi:10.1038/s41467-018-03990-7.

Gandon, S., Buckling, A., Decaestecker, E., Day, T., 2008. Host-parasite coevolution and patterns of adaptation across time and space. Journal of Evolutionary Biology 21, 1861–1866. doi:10.1111/j.1420-9101.2008.01598.x.

Gandon, S., Michalakis, Y., 2002. Local adaptation, evolutionary potential and host-parasite coevolution: interactions between migration, mutation, population size and generation time. Journal of Evolutionary Biology 15, 451–462. doi:10.1046/j.1420-9101.2002.00402.x.

Gandon, S., Nuismer, S.L., 2009. Interactions between genetic drift, gene flow, and selection mosaics drive parasite local adaptation. America Naturalist 173, 212–224. doi:10.1086/593706.

Gavrilets, S., Michalakis, Y., 2008. Effects of environmental heterogeneity on victim-exploiter coevolution. Evolution 62, 3100–3116. doi:10.1111/j.1558-5646.2008.00513.x.

Janzen, D., 1980. When is it coevolution. Evolution 34, 611–612. doi:10.2307/2408229.

Lewontin, R., Kojima, K., 1960. The evolutionary dynamics of complex polymorphisms. Evolution 14, 458–472. doi:10.1111/j.1558-5646.1960.tb03113.x.

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. doi:10.1534/genetics.117.300481/-/DC1.1.

May, R., Anderson, R., 1983. Epidemiology and genetics in the coevolution of parasites and hosts. Proceedings of the Royal Society Series B-Biological Sciences 219, 281–313. doi:10.1098/rspb.1983.0075.

Nuismer, S.L., Jenkins, C.E., Dybdahl, M.F., 2017. Identifying coevolving loci using interspecific genetic correlations. Ecology and Evolution 7, 6894–6903. doi:10.1002/ece3.3107.

Sasaki, A., 2000. Host-parasite coevolution in a multilocus gene-for-gene system. Proceedings of the Royal Society B-Biological Sciences 267, 2183–2188. doi:10.1098/rspb.2000.1267.

Staedler, T., Haubold, B., Merino, C., Stephan, W., Pfaffelhuber, P., 2009. The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. Genetics 182, 205–216. doi:10.1534/genetics.108.094904.

Stahl, E., Bishop, J., 2000. Plant-pathogen arms races at the molecular level. Current Opinion in Plant Biology 3, 299–304. doi:10.1016/S1369-5266(00)00083-2.

Stahl, E., Dwyer, G., Mauricio, R., Kreitman, M., Bergelson, J., 1999. Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis. Nature 400, 667–671.

Tellier, A., Brown, J.K.M., 2007. Stability of genetic polymorphism in host-parasite interactions. Proceedings of the Royal Society B-Biological Sciences 274, 809–817. doi:10.1098/rspb.2006.0281.

Tollenaere, C., Susi, H., Laine, A.L., 2016. Evolutionary and epidemiological implications of multiple infection in plants. Trends in Plant Science 21, 80–90. doi:10.1016/j.tplants.2015.10.014.

Wakeley, J., Aliacar, N., 2002. Gene genealogies in a metapopulation (vol 159, pg 893, 2001). Genetics 160, 1263.

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. Proceedings of the National Academy of Sciences of the United States of America , 201710980URL: http://www.ncbi.nlm.nih.gov/pubmed/29848634, doi:10.1073/pnas.1710980115, arXiv:1711.07918.

Woolhouse, M.E.J., Webster, J.P., Domingo, E., Charlesworth, B., Levin, B.R., 2002. Biological and biomedical implications of the co-evolution of pathogens and their hosts. Nature Genetics 32, 569–577. URL: http://www.nature.com/doifinder/10.1038/ng1202-569, doi:10.1038/ng1202-569.

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, doi:10.1101/588202, arXiv:https://www.biorxiv.org/content/early/2019/03/25/588202.full.pdf.

Živković, D., Steinruecken, M., Song, Y.S., Stephan, W., 2015. Transition Densities and Sample Frequency Spectra of Diffusion Processes with Selection and Variable Population Size. Genetics 200, 601+. doi:10.1534/genetics.115.175265.

A1. Appendix: The neutral site frequency spectrum

We assume for simplicity that the allele frequency distribution of neutral SNPs in both, the host and the parasite, follows the site frequency spectrum (SFS) under drift-mutation equilibrium for a Wright-Fisher model under constant population size. Further, we assume that there is no outgroup sequence available, thus the ancestral and derived state are unknown for a given SNP. The expected folded SFS $\eta = \{\eta_1, \ldots, \eta_{\lfloor n/2 \rfloor}\}$ under drift-mutation equilibrium for a sample of size n is given by:

$$\eta_k = \frac{\frac{\theta}{k} + \frac{\theta}{n-k}}{1 + \delta_{k,n-k}} \quad \text{for } 1 \leq k \leq \lfloor n/2 \rfloor \tag{A1}$$

where θ is the population mutation rate, $\lfloor n/2 \rfloor$ denotes the largest integer being smaller or equal to $n/2$ and $\delta_{k,l}$ is Kronecker's delta with

$$\delta_{k,l} = \begin{cases} 0 & \text{for } k \neq l \\ 1 & \text{for } k = l \end{cases}$$

Thus, the probability (p_k) to choose a SNP with minor allele frequency k in a sample of size n, is given by:

$$p_k = \frac{\eta_k}{\sum_{i=1}^{\lfloor n/2 \rfloor} \eta_i} = \frac{\left(\frac{\frac{1}{k} + \frac{1}{n-k}}{1 + \delta_{k,n-k}} \right)}{\sum_{i=1}^{\lfloor n/2 \rfloor} \left(\frac{\frac{1}{k} + \frac{1}{n-k}}{1 + \delta_{k,n-k}} \right)} \tag{A2}$$

These probabilities are independent of the population size and the mutation rate. However note that, changes in population size feed-back on the shape of the SFS and thus, the calculation of these probabilities.

Our computations include singletons, that is alleles with frequency $1/n$ in the sample. However, it is known that the sequencing and detection of singletons can be biased (*e.g.* with NGS technologies or pooling of samples). Therefore, singletons can be also removed from the CSA calculation and the CSP calculations should be adjusted accordingly by constraining the minor allele frequency in the infected subsample to be at least equal to two. Furthermore, if more complex demographic scenarios want to be included via their influence on the neutral host and/or parasite SFS, the probability A2 can be adjusted analytically or from results of coalescent simulations.

S1. Supplementary Online Information

In this supplement we add some details on the computations of the CSA and CSP values for neutral loci as well as additional figures.

S1.1. Cross species association index (CSA)

Remember that we have obtained n_{Inf} host samples and one representative parasite strain from each of these infected hosts. Thus, the host sample size (n_{Inf}) and the parasite sample size (n_{Par}) are the same ($n = n_{\text{Inf}} = n_{\text{par}}$). In order to compute the expected CSA for neutral SNPs we first have to derive an expression for the expected value of CSA ($E(\text{CSA}_{vw})$) measuring the association between a host SNP with minor allele frequency v and a parasite SNP with minor allele frequency w. Therefore, we first compute the number of all such possible combinations. For each combination, the value CSA is $\text{CSA}_{vw,k}$ and the probability of that particular combination is $\binom{v}{k}\binom{n_{\text{Inf}}-v}{w-k}$. The expectation $E(\text{CSA}_{vw})$ is then:

$$
\begin{aligned}
E(CSA_{vw}) &= \Omega_{vw} \sum_{k=0}^{l} \frac{\binom{v}{k}\binom{n_{\text{Inf}}-v}{w-k}}{\binom{n_{\text{Inf}}}{w}} CSA_{vw,k} && \text{(S1)} \\
&= \Omega_{vw} \sum_{k=0}^{l} \frac{\binom{v}{k}\binom{n_{\text{Inf}}-v}{w-k}}{\binom{n_{\text{Inf}}}{w}} \left(\left| \frac{k}{n_{\text{Inf}}} \cdot \frac{n_{\text{Inf}} - v - (w-k)}{n_{\text{Inf}}} - \frac{v-k}{n_{\text{Inf}}} \cdot \frac{w-k}{n_{\text{Inf}}} \right| \right) && \text{(S2)} \\
&= \Omega_{vw} \sum_{k=0}^{l} \frac{\binom{v}{k}\binom{n-v}{w-k}}{\binom{n_{\text{Inf}}}{w}} \left| \frac{k n_{\text{Inf}} - vw}{n_{\text{Inf}}^2} \right| && \text{(S3)}
\end{aligned}
$$

where $l = min(v, w)$.

Here the index k can be interpreted as the number of hosts with the minor allele which are infected by a parasite with the minor allele, or put in a different way, $w-k$ out of the $n_{\text{Inf}} - v$ hosts with the major allele are infected by a parasite with the minor allele. Accordingly, $v-k$ hosts with the minor allele are infected by a parasite which has the major allele, and $n_{\text{Inf}}-v-(w-k)$ hosts with the major allele are infected by parasites with the major allele. We define Ω_{vw} as the normalization for either obtaining CSA′ (with $\Omega_{vw} = 4$, as in eq. 5) or CSA_r (with $(\Omega_{vw} = \frac{1}{\sqrt{\frac{v}{n_{\text{Inf}}}\frac{n-v}{n_{\text{Inf}}}\frac{w}{n_{\text{Inf}}}\frac{n-w}{n_{\text{Inf}}}}})$, from eq. 6).

S1.2. Distribution of CSA for different sample sizes

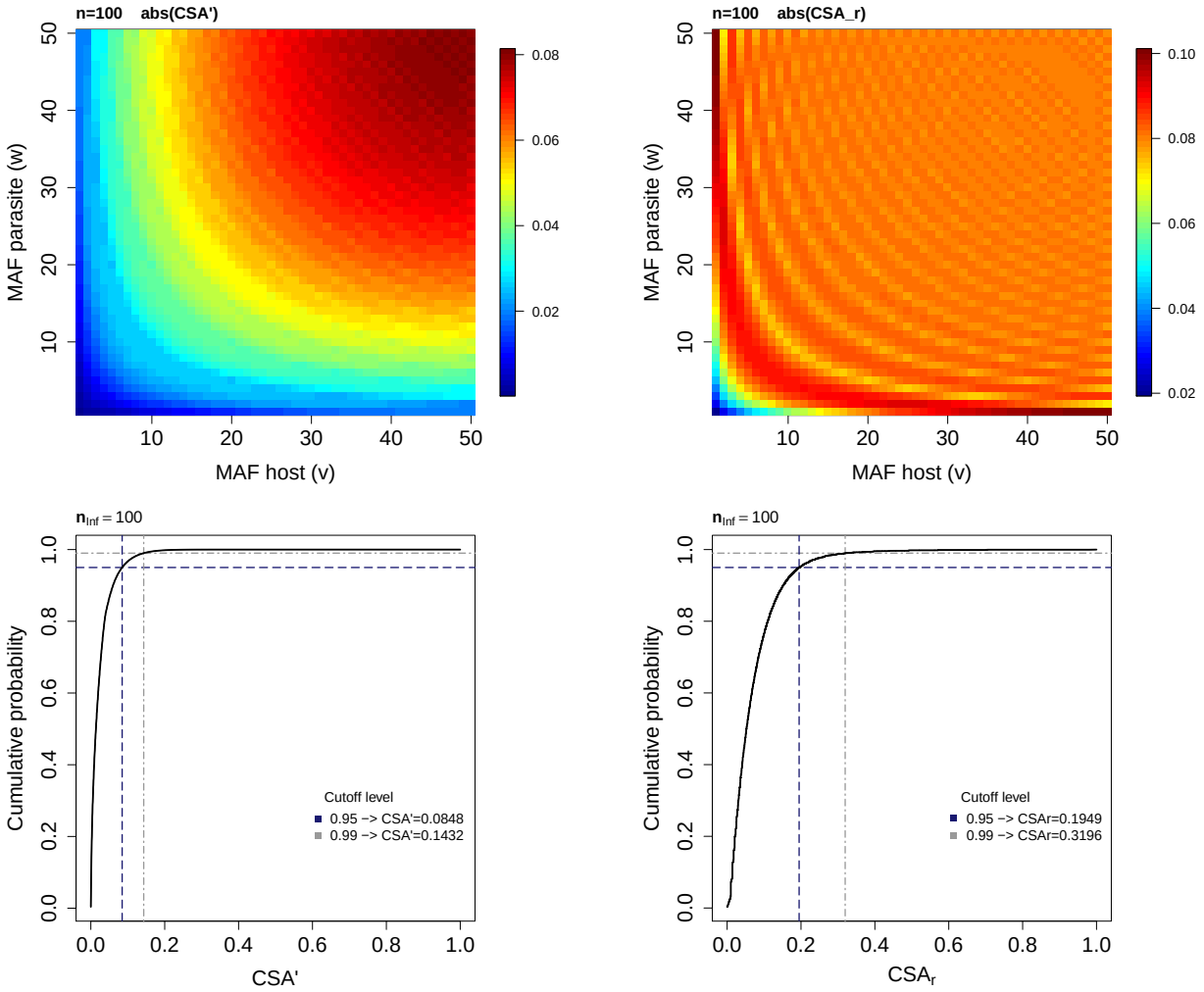

Figure S1 Expected values of CSA′ (top left) and CSA_r (top right) when comparing all neutral host SNPs with minor allele frequency v ($v \in \{1, \ldots, \lfloor n_{\text{Inf}}/2 \rfloor\}$) to all neutral parasite SNPs with minor allele frequency w ($w \in \{1, \ldots, \lfloor n_{\text{Par}}/2 \rfloor\}$ and the resulting expected cumulative distribution function of E(CSA') (bottom left) and E(CSA_r) (bottom right) for a sample size of $n_{\text{Inf}} = n_{\text{Par}} = 100$.

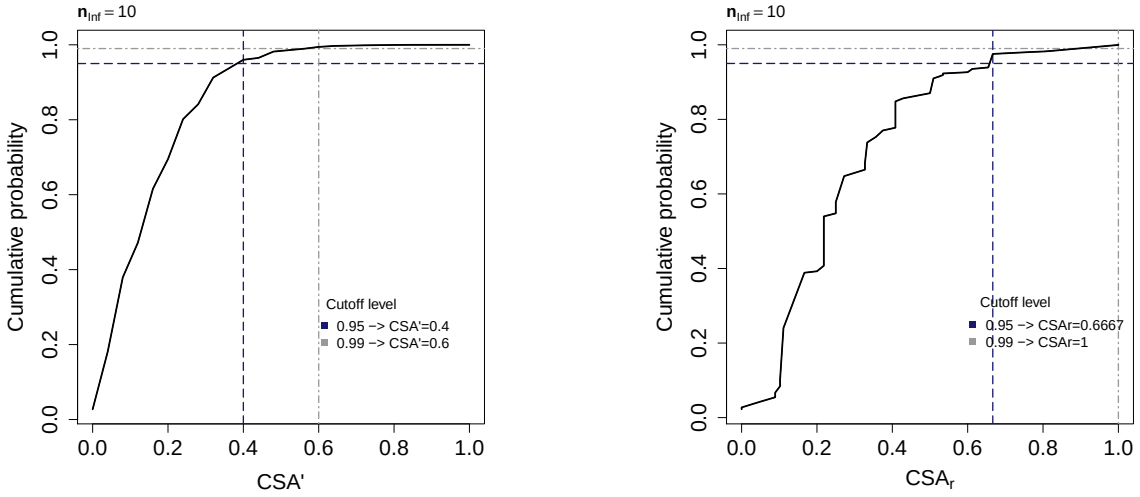

Figure S2 Expected cumulative distribution function of CSA′ (left) and CSA_r (right) when comparing all neutral host SNPs with minor allele frequency v ($v \in \{1, \ldots, \lfloor n/2 \rfloor\}$) to all neutral parasite SNPs with minor allele frequency w ($w \in \{1, \ldots, \lfloor n/2 \rfloor\}$ for a sample size of $n_{\text{Inf}} = n_{\text{Par}} = 10$.

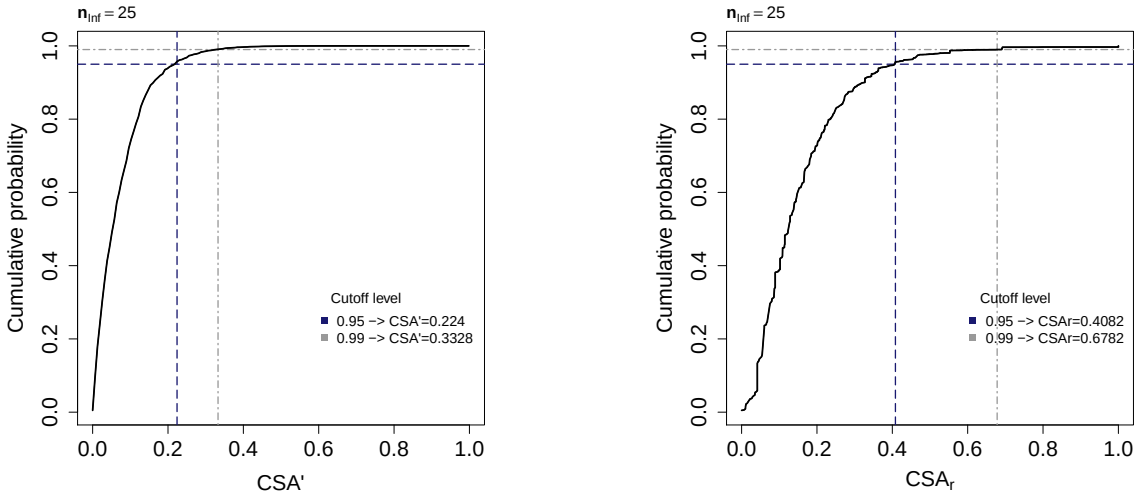

Figure S3 Expected cumulative distribution function of CSA′ (left) and CSA_r (right) when comparing all neutral host SNPs with minor allele frequency v ($v \in \{1, \ldots, \lfloor n/2 \rfloor\}$) to all neutral parasite SNPs with minor allele frequency w ($w \in \{1, \ldots, \lfloor n/2 \rfloor\}$ for a sample size of $n_{\text{Inf}} = n_{\text{Par}} = 25$.

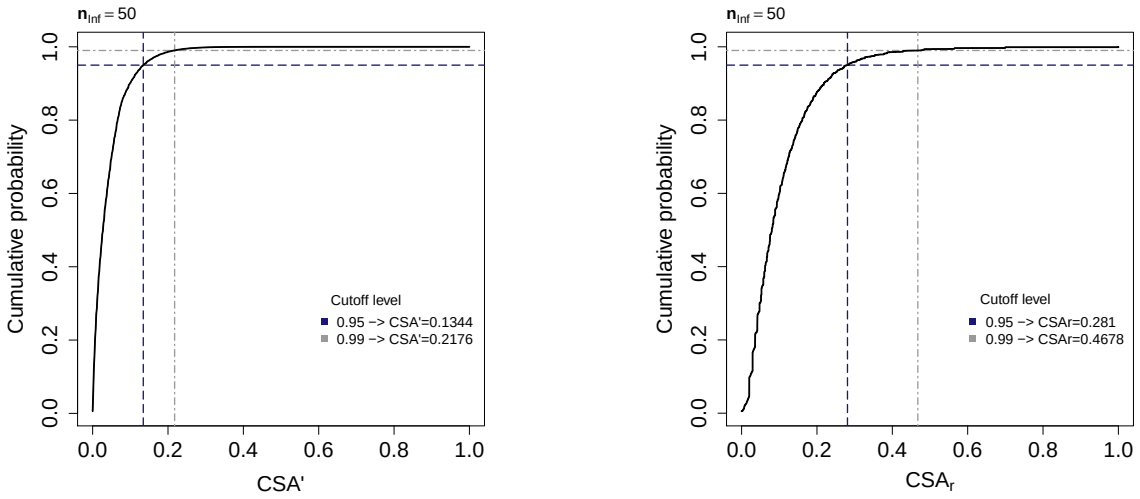

Figure S4 Expected cumulative distribution function of CSA′ (left) and CSA$_r$ (right) when comparing all neutral host SNPs with minor allele frequency v ($v \in \{1, \ldots, \lfloor n/2 \rfloor\}$) to all neutral parasite SNPs with minor allele frequency w ($w \in \{1, \ldots, \lfloor n/2 \rfloor\}$ for a sample size of $n_{\text{Inf}} = n_{\text{Par}} = 50$.

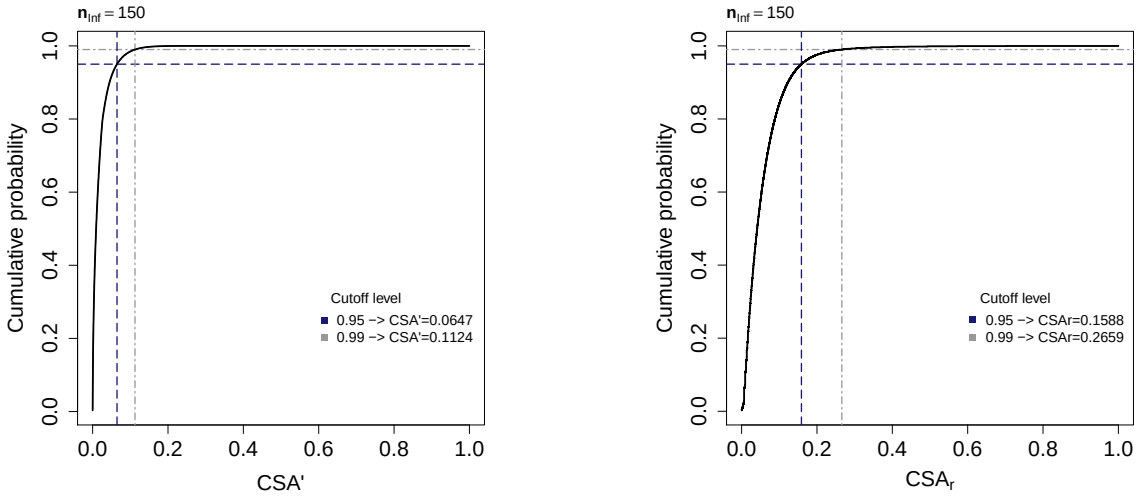

Figure S5 Expected cumulative distribution function of CSA′ (left) and CSA$_r$ (right) when comparing all neutral host SNPs with minor allele frequency v ($v \in \{1, \ldots, \lfloor n/2 \rfloor\}$) to all neutral parasite SNPs with minor allele frequency w ($w \in \{1, \ldots, \lfloor n/2 \rfloor\}$ for a sample size of $n_{\text{Inf}} = n_{\text{Par}} = 150$.

S1.3. Cross species prevalence index (CSP)

We label the host allele with minor frequency in the **infected subsample** as $i = 1$ and the host allele with major frequency in the infected subsample as $i = 2$. Note that the allele with minor allele frequency in the infected subsample is not necessarily the minor allele in the whole sample (see Fig. S6). In cases where both alleles have equal frequencies in the infected subsample, the allele with minor allele frequency in the whole sample will be labelled as 1 and the allele with major

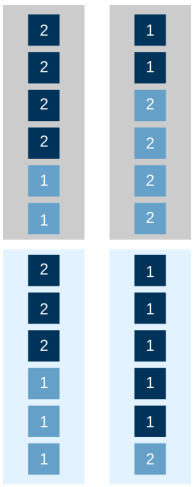

Figure S6 Two possible host configurations when sampling a total number $n_{\mathrm{T}} = 12$ host individuals among which $n_{\mathrm{Inf}} = 6$ individuals are infected (grey box) and $n_{\mathrm{H}} = 6$ individuals are healthy (light blue box) and the minor host allele frequency is $v = 5$. Host individuals which have the minor allele (based on the whole sample) are shown in light blue, host individuals with the major allele (based on the whole sample) are shown in dark blue. Labelling of the alleles for the calculation of CSP is based on the minor allele frequency in the infected subsample. On the left, the minor allele is labelled by 1 as it is also the minor allele in the infected subsample. On the right, the major allele of the total sample is labelled by 1 as it represents the allele with minor allele frequency in the infected subsample.

allele frequency in the whole sample will be labelled as 2. Therefore, F_{11} (F_{12}) is the proportion of hosts with label 1 which are infected by a parasite with the minor (major) allele. F_{21} (F_{22}) is the proportion of hosts with label 2 which are infected by a parasite with the minor (major) allele. Further, F_{10} (respectively F_{20}) is the proportion of non-infected hosts carrying allele 1 (respectively 2). If for a neutral locus there are v minor alleles in the total host sample n_T, these minor alleles can be found with equal probability on each of the n_{T} individuals (irrespective of the infection status) as a neutral SNP does not have an effect on the infection outcome. Similarly, all of the w parasite minor alleles can be randomly assigned to any of the n_{par} parasite individuals which are infecting the n_{Inf} host individuals. Further, note that CSP is only informative when the minor and major allele can be found in both, the infected and the non-infected subsample. Therefore, we exclude SNPs which are singletons in the total host sample (n_{T}). We proceed as follows to obtain the expected CSP for a neutral host SNP with minor allele frequency v and and neutral parasite SNP with minor allele frequency w.

First, we have to find all host combinations (and their probability) for which the minor and major host alleles are found in both the infected and non-infected subsamples. We define z as the number of minor host alleles which are found in the infected subsample for a given combination. Accordingly, the number of major host alleles in the infected subsample is $n_{\mathrm{Inf}} - z$, the number of

minor host alleles in the non-infected subsample is $v - z$ and the number of major host alleles in the non-infected subsample is $n_T - n_{\text{Inf}} - (v - z)$. Based on the resulting composition of the infected subsample the alleles are labelled. The indicator variable λ is used to keep track of whether the minor allele in the total sample is the minor ($\lambda = 0$) or the major ($\lambda = 1$) allele in the infected subsample.

Second, the n_{par} parasites among which w individuals have the minor parasite allele are assigned within the n_{Inf} sample. Hereby, k denotes the number of hosts with label 1 which are infected by a parasite with the minor allele (see CSA). Thus, the expected value of CSP for a SNP with minor allele frequency v in the host and minor allele frequency w in the parasite is given by:

$$E(CSP_{vw}) = \sum_{z=\rho}^{m-1} \frac{\binom{n_{\text{Inf}}}{z}\binom{n_H}{v-z}}{\binom{n_{\text{T}}}{v} - \sum_{b=0}^{\rho-1}\binom{n_{\text{Inf}}}{b}\binom{n_H}{v-b} - \sum_{b=m}^{v}\binom{n_{\text{Inf}}}{b}\binom{n_H}{v-b}} \sum_{k=0}^{min(w,a)} \frac{\binom{a}{k}\binom{n_{\text{Inf}}-a}{w-k}}{\binom{n_{\text{Inf}}}{w}} \left| \frac{\frac{k}{n_{\text{T}}} + \frac{a-k}{n_{\text{T}}}}{\frac{\lambda n_H + (v-z)(-1)^{\lambda}}{n_{\text{T}}}} - \frac{\frac{w-k}{n_{\text{T}}} + \frac{n_{\text{Inf}}-a-(w-k)}{n_{\text{T}}}}{\frac{(1-\lambda)n_H - (v-z)(-1)^{\lambda}}{n_{\text{T}}}} \right| \tag{S4}$$

$$E(CSP_{vw}) = \sum_{z=\rho}^{m-1} \frac{\binom{n_{\text{Inf}}}{z}\binom{n_H}{v-z}}{\binom{n_{\text{T}}}{v} - \sum_{b=0}^{\rho-1}\binom{n_{\text{Inf}}}{b}\binom{n_H}{v-b} - \sum_{b=m}^{v}\binom{n_{\text{Inf}}}{b}\binom{n_H}{v-b}} \sum_{k=0}^{min(w,a)} \frac{\binom{a}{k}\binom{n_{\text{Inf}}-a}{w-k}}{\binom{n_{\text{Inf}}}{w}} \left| \frac{a}{\lambda n_H + (v-z)(-1)^{\lambda}} - \frac{n_{\text{Inf}} - a}{(1-\lambda)n_H - (v-z)(-1)^{\lambda}} \right|$$

$$E(CSP_{vw}) = \sum_{z=\rho}^{m-1} \frac{\binom{n_{\text{Inf}}}{z}\binom{n_H}{v-z}}{\binom{n_{\text{T}}}{v} - \sum_{b=0}^{\rho-1}\binom{n_{\text{Inf}}}{b}\binom{n_H}{v-b} - \sum_{b=m}^{v}\binom{n_{\text{Inf}}}{b}\binom{n_H}{v-b}} \left| \frac{a}{\lambda n_H + (v-z)(-1)^{\lambda}} - \frac{n_{\text{Inf}} - a}{(1-\lambda)n_H - (v-z)(-1)^{\lambda}} \right| \tag{S5}$$

1 $\rho = min(1, v - n_H + 1)$, $m = min(n_{\text{Inf}}, v)$, $a = min(z, n_{\text{Inf}} - z)$

$$\lambda = \begin{cases} 0 & \text{for } z \leq n_{\text{Inf}} - z \\ 1 & \text{for } z > n_{\text{Inf}} - z \end{cases} \tag{S6}$$

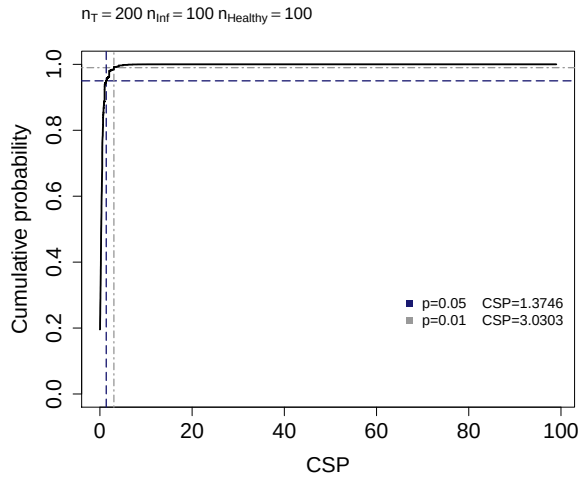

Figure S7 Cumulative distribution function of the expected value of CSP when taking a host sample of total size $n_{\mathrm{T}} = 200$ which includes $n_{\mathrm{Inf}} = 100$ infected hosts and $n_{\mathrm{H}} = 100$ healthy hosts.

The condition that z starts from $\rho = max(1, v - (n_T - n_{\mathrm{Inf}}) + 1)$ is necessary to avoid combinations where 1) no minor allele is found in the infected subsample, and 2) the healthy sample only consists of hosts with the minor allele. The condition that z has values up to $m - 1$ is necessary to avoid two configurations where 1) none of the minor alleles is found in the non-infected subsample, and 2) all individuals in the infected subsample have the minor allele. For a given host allele combination, we perform the labelling step mentioned above by defining $a = min(z, n_{\mathrm{Inf}} - z)$. Then, we assign the n_{Par} parasites, w of them having the minor allele, to the n_{Inf} host. Here, k is the number of hosts with label 1 which are infected by a parasite with the minor allele ($F_{11} \cdot n_T$). Accordingly, $a - k$ is the number of hosts with label 1 which are infected by a parasite with major allele ($F_{12} \cdot n_T$), $w - k$ is the number of hosts with label 2 which are infected by a parasite with minor allele ($F_{21} \cdot n_T$) and $n_{\mathrm{Inf}} - a - (w - k)$ with label 2 which are infected by a parasite with the major allele.

S2. Supplementary figures

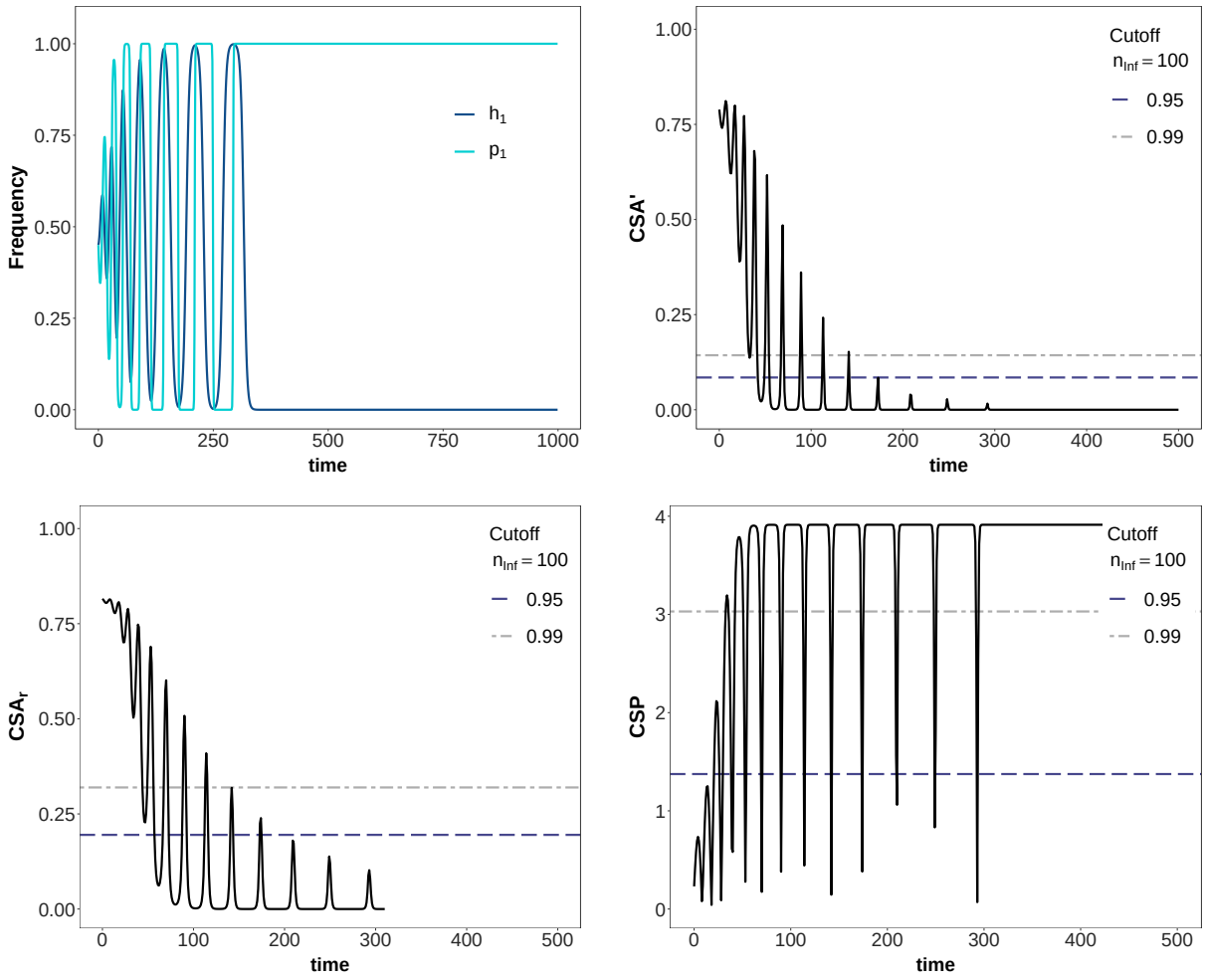

Figure S8 Temporal changes in allele frequencies, CSA$'$, CSA$_r$ and CSP in an unstable MA-model (model A) with one parasite generation per host generation. For each index cut-off values are shown based on the expected neutral distributions for a total host sample size $n_T = 200$ and for $n_{\text{Inf}} = n_{\text{H}} = 100$. The 0.95-cut-off value is shown in blue (dashed line) and the 0.99-cut-off value is shown in grey (dotted-dashed line). Top left: frequencies of h_1 (dark blue) and p_1 (light blue). Top right: CSA$'$. Bottom left: CSA$_r$. Bottom right: CSP. The parameters values of the model are: $c_1 = c_2 = 0.9$, $c_{H_1} = c_{P_1} = c_{H_2} = c_{P_2} = 0$, $\phi = 0.8$, $s = 0.35$, $h_{1,\text{init}} = p_{1,\text{init}} = 0.45$

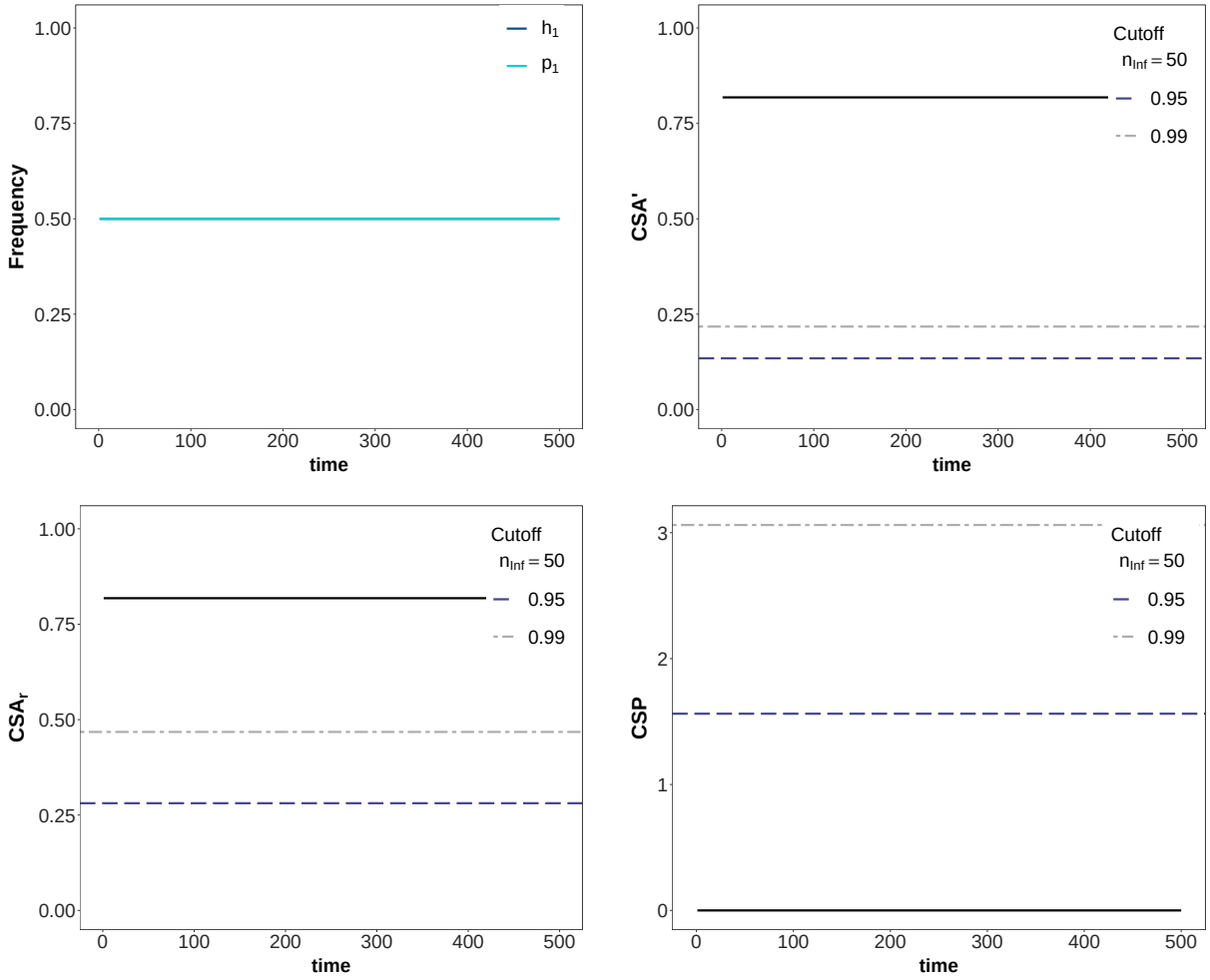

Figure S9 Temporal changes in allele frequencies, CSA′, CSA_r and CSP in an epidemiological model (model B) with a symmetric MA-infection matrix. For each index cut-off values are shown based on the expected neutral distributions for a total host sample size $n_T = 200$ and for $n_{\text{Inf}} = n_{\text{H}} = 100$. The 0.95-cut-off value is shown in blue (dashed line) and the 0.99-cut-off value is shown in grey (dotted-dashed line). Top left: frequencies of h_1 (dark blue) and p_1 (light blue). Top right: CSA′. Bottom left: CSA_r. Bottom right: CSP. The parameters values of the model are: $c_{H_1} = c_{P_1} = c_{H_2} = c_{P_2} = 0$, $\beta = 0.00005$, $s = 0.6$, $c_1 = c_2 = 0.9$, $S_{1,\text{init}} = S_{2,\text{init}} = 4150$, $I_{11} = I_{12} = I_{21} = I_{22} = 415$, $\delta_t = 0.001$, $b = 1$, $\gamma = 0.9$

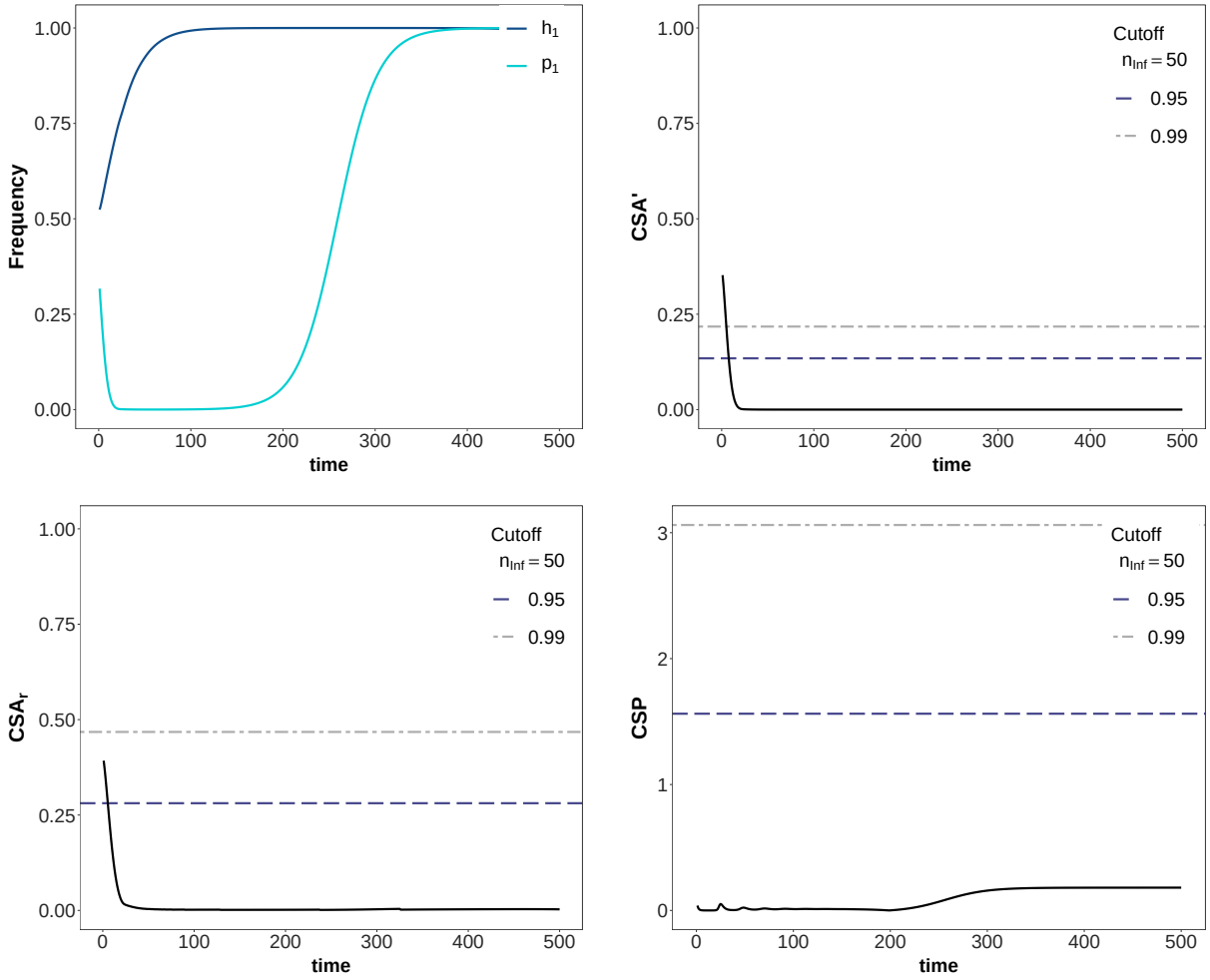

Figure S10 Temporal changes in allele frequencies, CSA′, CSA_r and CSP an epidemiological (model B) with a GFG-infection matrix. For each index cut-off values are shown based on the expected neutral distributions for a total host sample size $n_T = 200$ and for $n_{\text{Inf}} = n_{\text{H}} = 100$. The 0.95-cut-off value is shown in blue (dashed line) and the 0.99-cut-off value is shown in grey (dotted-dashed line). Top left: frequencies of h_1 (dark blue) and p_1 (light blue). Top right: CSA′. Bottom left: CSA_r. Bottom right: CSP. The parameters values of the model are: $c_{H_1} = c_{P_1} = 0$, $c_{H_2} = c_{P_2} = 0.05$, $\beta = 0.00005$, $s = 0.6$,$c = 0.9$, $S_{1,\text{init}} = S_{2,\text{init}} = 4150$, $I_{11} = I_{12} = I_{21} = I_{22} = 415$, $\delta_t = 0.001$, $b = 1$, $\gamma = 0.9$