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

Short title: Host-parasite associations statistics

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

[a]*Professorship for Population Genetics, Department of Life Science Systems, Technical University of Munich, 85354 Freising, Germany*
[b]*now Department of Ecology and Evolution, University of Chicago, 1101 E 57th St, Chicago, IL 60637, USA*

[*]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 coevolutionary interactions is of importance for disease management. The increasing availability of host and parasite full genome-data allows for cross-species genome-wide association studies based on the genomic data of infected hosts and their associated parasite strains sampled from natural populations (*i.e.* natural co-GWAs). Such studies focus on searching for cross-species allelic associations between pairs of host and parasite SNPs. We aim to quantify the power of natural co-GWAs to pinpoint loci under coevolution with respect to the intrinsic complexities of coevolutionary systems, such as the genetic specificity of the interaction and the temporal allele frequency changes resulting from the interaction. Therefore, we develop the cross-species association (CSA) and the cross-species prevalence (CSP) indices, the latter additionally incorporating genomic data from uninfected hosts. To provide an assessment of the statistical power of these indices, we analytically derive their genome-wide False Discovery Rates (FDR) based on the neutral site-frequency spectrum of the host and the parasite population. Using two coevolutionary models, we investigate under which parameter regimes these indices pinpoint the coevolving loci. Under trench warfare dynamics, CSA and CSP are very accurate in pinpointing the loci under coevolution, while under arms race dynamics the power is limited especially for gene-for-gene interactions. Furthermore, we reveal that the combination of both indices across time samples is an indicator for the specificity of the interaction. Our results provide novel insights into the power and biological interpretation of natural cross-species association studies.

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. 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). The simple underlying idea is to compare all Single Nucleotide Polymorphisms (SNPs) between samples of hosts and parasites in a pairwise manner in order to detect significant correlations with the infection outcome (amount of infection, disease severity) or to detect significant allelex frequency correlations between species. However, such analyses rely on running large scale controlled experiments with numerous host and parasite genotypes (and thus can be termed controlled co-GWAs). A promising less-labor intensive alternative is to perform what we will call a natural co-GWAs. Such natural co-GWAs are based on sampling infected hosts and their associated parasites from natural populations (Ansari et al. 2017, Bartha et al. 2013, Bartoli and Roux 2017, Lees et al. 2019) and performing whole-genome sequencing of both interacting partners. The resulting data inherently contains phenotypic information (the infection outcome) about each sampled host-parasite pair, namely the given parasite genotype is infective on the given host genotype and hence, the host is susceptible to that particular parasite. The causal genetic variants for host susceptibility and parasite infectivity are expected to show statistically significant GxG associations (host genotype x parasite genotype) within the sample. A chief hypothesis is that these loci should explain a large proportion of the variance of the infection outcome as compared to loci in the neutral genomic background, which, by definition, have no effect on the infection outcome. To our knowledge, natural co-GWAs have been applied three times. These studies have uncovered strong associations between human SNPs 1) of the major histocompatibility complex (MHC) and known HIV epitopes (Bartha et al. 2013), 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) and 3) of human disease

susceptibility genes and *Streptococcus pneumoniae* genes responsible for invasiveness (Lees et al. 2019). Nevertheless, a key statistical issue, as in all GWAs, is that these natural co-GWAs methods rely on calculating association values for all possible comparisons of host and parasite SNP pairs, thus generating a distribution of millions of association values. As the top candidate loci are likely to be followed up on in molecular studies, there is a need to minimize the False Discovery Rate (FDR), that is the proportion of SNP pairs being identified as underlying the interaction when in fact they do not.

In principle, natural co-GWAs can be easily extended to any coevolutionary system where it is possible to call SNPs for a sample of infected hosts and the associated parasites (Bartoli and Roux 2017). An implicit underlying assumption is that the host and parasite genes revealed to determine the infection outcome are likely to be coevolving. Coevolution is defined as reciprocal changes in allele frequencies which result from selective pressures that two interacting species exert on one another (Janzen 1980). Allele frequency changes at coevolutionary loci are commonly described by a continuum between arms race (Dawkins and Krebs 1979, Stahl and Bishop 2000, Woolhouse et al. 2002) and trench warfare dynamics (Stahl et al. 1999). Arms race dynamics are characterized by recurrent fixation of alleles at the interacting loci and transient allelic polymorphism. Trench warfare dynamics, on the other extreme, result in long-term stable maintenance of several alleles with their frequencies either persistently fluctuating or converging towards a stable polymorphic equilibrium over time. Hence, allele frequency fluctuations are a common feature of all coevolutionary dynamics, irrespective of the ultimate fate of alleles. These allele frequency fluctuations are expected to affect the statistical power of natural co-GWAs over time. Indeed, it has been shown for experimental co-GWAS that the statistical significance of a given SNP pair association does depend on both host and parasite allele frequencies (MacPherson et al. 2018, Nuismer et al. 2017).

The speed and type of coevolutionary dynamics have been demonstrated to depend upon host and parasite life-history traits (reviewed in Brown and Tellier 2011), the strength

of epidemiological dynamics (Ashby and Boots 2017) and the underlying host genotype by parasite genotype (GxG) interactions. Under the assumption that few major genes determine the interaction outcome, the underlying GxG interactions can be captured in a so called infection matrix $\mathcal{A}$. Here, each entry α_{ij} denotes 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) model and the gene-for-gene (GFG) model. 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 $1-\omega_1$ and $1-\omega_2$ (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 $1-\omega$. Both matrices represent some points in a continuum of infection matrices (Agrawal and Lively 2002) and are a subset of more complex matrices with several alleles or loci (Ashby and Boots 2017, Gandon and Michalakis 2002).

Our aim is to evaluate the statistical power and False Discovery Rate (FDR) of natural co-GWAs to pinpoint genes under coevolution given the various kinds of GxG interactions and the varying temporal dynamics of host-parasite coevolution. Therefore, we address the following three questions. First, which statistics (indices) are useful to locate the loci under coevolution? Second, what is the statistical power (FDR) of these indices under various infection matrices and/or fluctuating allele frequencies? Third, can these statistics provide some information about the underlying infection matrix and/or coevolutionary dynamics? To answer these questions, we first develop two indices: the cross-species association (CSA) index (analogous to the measure used in Ansari et al. 2017, Bartha et al. 2013 and suggested in Bartoli and Roux 2017) and the cross-species prevalence (CSP) index. Both indices measure the amount of association of allele frequencies across host and parasite loci. Second, we study six hypothetical numerical examples representing varying amounts

of cross-species association. For these examples, we calculate the introduced indices along with p-values of a corresponding χ^2-test. The p-value allows us to measure the chance that a pair of loci under coevolution would be among the top candidates when correcting for multiple testing. Third, we explicitly calculate the expected distribution of the values of both indices for all possible comparisons between pairs of neutral host and neutral parasite loci (representing the genomic background). Our computations are based on the site-frequency spectrum distribution of neutral allele frequencies for a population with constant size. This neutral distribution of values of our indices allows us to access the False Discovery Rate (FDR), that is the percentage of pairs of neutral alleles which show higher index values than the pair of loci under coevolution. We can show for these same hypothetical examples that our FDR results are in accordance with those of the χ^2-test. Fourth, we present analytical results on the expected values of our indices at the loci under coevolution for two widely used coevolutionary models and for arbitrary infection matrices. Finally, we quantify the statistical power of these statistics to pinpoint coevolving loci over the course of a coevolutionary interaction and for different underlying GxG matrices by terms of simulation. To do so, we calculate both indices over the entire course of a coevolutionary dynamic and access the FDRs based on the neutral distributions. We demonstrate that performing co-GWAs using these indices across time samples not only allows to infer 1) the type of dynamics occurring (arms race versus trench warfare), but also 2) the underlying GxG interaction matrix. We conclude by discussing the applicability of our co-GWAs indices to studies of coevolution in natural or controlled systems.

2. Methods

2.1. Definition of the association indices

We assume that n_{T} host individuals are sampled and genotyped at each biallelic single nucleotide polymorphism (SNP) in the genome, so that at any locus l, 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 bi-allelic SNPs are also a common assumption 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 or transposable elements).

2.1.1. The Cross-Species Association index (CSA)

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

$$\text{CSA} = |f_{11}f_{22} - f_{21}f_{12}|. \tag{1}$$

Here, f_{ij} is the proportion of hosts of type i being infected by a parasite of type j in the infected subsample (n_{Inf}), so that $\sum_{\forall i,j} f_{ij} = 1$. The 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).

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′ by normalizing CSA values by the maximum CSA value possible, $\text{CSA}^{\text{max}} = 0.25$.

$$\text{CSA}' = \frac{\text{CSA}}{\text{CSA}^{\text{max}}} = \frac{|f_{11}f_{22} - f_{21}f_{12}|}{0.25} = 4 \cdot \text{CSA}, \tag{2}$$

CSA is at its maximum 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 $f_{12} = f_{21} = 0.5$).

The 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

$$\text{CSA}_\text{r} = \frac{|f_{11}f_{22} - f_{21}f_{12}|}{\sqrt{(f_{11}+f_{12})(f_{21}+f_{22})(f_{11}+f_{21})(f_{12}+f_{22})}}. \tag{3}$$

2.1.2. The Cross-Species Prevalence index (CSP)

We define the Cross Species Prevalence index (CSP) as

$$\text{CSP} = \left| \frac{F_{11}+F_{12}}{F_{10}} - \frac{F_{21}+F_{22}}{F_{20}} \right|. \tag{4}$$

Here, F_{ij} is the proportion of host type i infected by parasite type j in the total host sample (n_T) and F_{i0} is the proportion of uninfected hosts of type i in the total sample (n_T). By definition, $F_{11}+F_{12}+F_{21}+F_{22} = n_\text{Inf}/n_T$, $F_{10}+F_{20} = (n_T - n_\text{Inf})/n_T$, 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 which either 1) have no interaction with any parasite if the disease prevalence is less than 100%, or 2) are exposed to parasites but resist infection. CSP is defined as long as there are uninfected individuals of both host types present in the population (all $F_{i0} \neq 0$). CSP is thus an indirect association index as it does not explicitly take into account which parasite type infects which host type in the infected subsample. But, it depends on the frequencies of the host types in the different subsamples which are determined by the specificity of the interaction and the parasite allele frequencies.

2.2. Statistics, FDR and distributions of indices' values

When performing a natural co-GWAS among m_H host SNPs and m_P parasite SNPs, there is a potentially very large total number of $m_H * m_P$ comparisons (see Ansari et al. 2017, Bartha et al. 2013). This means that any significance threshold based on neutral SNPs should be stringent enough to have a low number of false candidate loci (a low FDR).

2.2.1. χ^2-test for CSA/CSP

For both CSA and CSP indices, we can compare proportions in a 2x2 contingency table, and thus perform a simple χ^2-test of homogeneity to obtain the p-value for a given pair of SNPs. The corresponding contingency table for CSA is

$$\begin{array}{c} \\ i=1 \\ i=2 \end{array} \begin{array}{c} \begin{array}{cc} j=1 & j=2 \end{array} \\ \begin{pmatrix} n_{\text{Inf}} f_{11} & n_{\text{Inf}} f_{12} \\ n_{\text{Inf}} f_{21} & n_{\text{Inf}} f_{22} \end{pmatrix} \end{array},$$

and the corresponding contingency table for CSP is

$$\begin{array}{c} \\ i=1 \\ i=2 \end{array} \begin{array}{c} \begin{array}{cc} \text{Infected} & \text{Healthy} \end{array} \\ \begin{pmatrix} n_{\text{T}}(F_{11}+F_{12}) & n_{\text{T}} F_{10} \\ n_{\text{T}}(F_{21}+F_{22}) & n_{\text{T}} F_{20} \end{pmatrix} \end{array}.$$

To account for multiple testing, we apply a Bonferroni-correction by dividing the chosen significance threshold (*e.g.* 0.05 or 0.01) by the total number of comparisons ($m_H * m_P$). We provide R-codes to compute the values of CSA and CSP and corresponding χ^2-tests for a list of pairwise comparisons between m_H host and m_P parasite SNPs.

2.2.2. Expected distributions of CSA and CSP for neutral SNPs

To quantify the statistical power (FDR) of our indices without having empirical data at hand, we compute the expected theoretical distribution of CSA and CSP values based on a large number of pairs of neutral host and parasite loci. Full genome data usually contains 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 and thus, both their population and sample allele frequencies are mutually independent. Therefore, we can conveniently draw the sample frequencies for

our neutral loci from an expected neutral site-frequency spectrum for a population with constant size.

By definition, neutral host and parasite loci do not determine the outcome of infection. Therefore, each neutral host SNP-neutral parasite SNP pair is characterized by an infection matrix $\mathcal{A}_{neutral} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$.

We obtain the expected value for CSA as follows:

1. We calculate the expected value of CSA when a host SNP has minor allele frequency v and a parasite SNP has minor allele frequency w in the infected subsample n_{Inf}. Given these minor allele frequencies, we calculate the value of CSA for each possible configuration of alleles and weight it by the probability of the configuration.

$$E(CSA_{vw}) = \Omega_{vw} \sum_{k=0}^{l} \frac{\binom{v}{k}\binom{n_{\text{Inf}}-v}{w-k}}{\binom{n_{\text{Inf}}}{w}} \left| \frac{kn_{\text{Inf}} - vw}{n_{\text{Inf}}^2} \right|, \tag{5}$$

   where $l = min(v, w)$, and Ω_{vw} is the normalization for either obtaining CSA$'$ (with $\Omega_{vw} = 4$, from Eq. 2) 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. 3).
2. Based on the neutral host and parasite site frequency spectra we calculate the probability p_v that a host SNP has minor allele frequency v and the probability p_w that a parasite SNP has minor allele frequency w in a sample of size n_{Inf}.
3. We weight the expected values obtained in step (1) by the probabilities obtained in step (2)

$$E(CSA) = \sum_{v=1}^{\lfloor n/2 \rfloor} \sum_{w=1}^{\lfloor n/2 \rfloor} p_v q_w E(CSA_{vw}). \tag{6}$$

This approach also allows us to generate a probability mass function and a cumulative distribution function for a given sample size, from which the FDR is computed by using the r-package distr (Ruckdeschel et al. 2006).

For the CSP we proceed in a similar way.

1. We calculate the expected value of CSP when a host SNP has minor allele frequency

v in the total host sample n_{T} and a parasite SNP has minor allele frequency w in the infected subsample n_{Inf}. In contrast to the CSA calculations, we first have to access all possible ways of assigning host alleles to the healthy (n_{H}) and infected (n_{Inf}) subsample. We restrict ourselves to configurations where both alleles can be found in both the infected and the healthy subsample.

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

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

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

2. Based on the neutral host and parasite site frequency spectra, we calculate the probability p_v that a host SNP has minor allele frequency v in a sample of size n_{T} and the probability p_w that a parasite SNP has minor allele frequency w in a sample of size n_{Inf}.
3. We weight the expected values obtained in step (1) by the probabilities obtained in step (2)

$$E(CSP) = \sum_{v=1}^{\lfloor n/2 \rfloor} \sum_{w=1}^{\lfloor n/2 \rfloor} p_v q_w E(CSP_{vw}). \tag{7}$$

Without loss of generality, we use the folded SFS of both the host and the parasite. Thus, we assume that there is no outgroup sequence available and 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 (our Eq. S2

is found for example in Durrett 2010, p.50). We compute the distributions for CSA and CSP for different sample sizes and sampling schemes. For any CSA or CSP value at a pair of coevolving loci, we can extract the False Discovery Rate (FDR) from this distribution assuming that there is only one or very few coevolving pairs of loci. The FDR is the proportion of pairs of neutral loci which exhibit a higher index value than a pair of coevolving loci. Based on the design of realistic molecular biology studies, we aim to obtain a maximum of ten candidate pairs of loci. Keeping in mind that both indices are computed for $m_H * m_P$ pairs of loci, the FDR rate therefore must be roughly smaller than $10/(m_H * m_P)$. If not stated otherwise, we show results for $n_T = 200$ and $n_{\text{Inf}} = n_{\text{H}} = 100$ (see the Online Supplementary Material for other sampling schemes Fig. S1 - S5).

2.3. The coevolutionary models

We study the temporal dynamics of CSA and CSP at a pair of coevolving loci using two different models, namely a pure population genetics model and an epidemiological model which incorporates eco-evolutionary feedback. Both models can be flexibly parameterized in terms of the underlying infection matrix and have been shown to exhibit the full range of coevolutionary dynamics.

2.3.1. 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 2007b). The frequency of host genotype h_i and parasite genotype p_j in generation $g + 1$ is

obtained as:

$$h_{i,g+1} = \frac{h_{i,g} \cdot (1 - c_{H_i}) \cdot \left(1 - \phi_g \cdot s \cdot \sum_{j=1}^{2} \alpha_{ij} p_{j,g}\right)}{\bar{w}_{H,g}}, \tag{8a}$$

$$p_{j,g+1} = \frac{p_{j,g} \cdot (1 - c_{P_j}) \cdot \left(\sum_{i=1}^{2} \alpha_{ij} h_{i,g}\right)}{\bar{w}_{P,g}}. \tag{8b}$$

Every generation g, a proportion ϕ_g of the host population (*i.e.* the exposed fraction) interacts with the parasite population in a frequency-dependent manner. The outcome of interactions is determined by the infection matrix $\mathcal{A}$. Therefore, the proportion of infected individuals in the population (*i.e.* the disease prevalence) is $\xi_g = \phi_g \sum_i \sum_j \alpha_{ij} h_{i,g} p_{j,g}$. 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). Unlike the cost of infection (s) which reduces the fitness of only the infected individuals, the cost of resistance (c_{H_i}) and cost of infectivity (c_{P_j}) reduce the fitness of all individuals of the given genotype, irrespective of the infection status. Previous work has shown that this model always generates arms race dynamics (Tellier and Brown 2007b).

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:

$$\hat{h}_1 = \frac{\alpha_{22}(1 - c_{P_2}) - \alpha_{21}(1 - c_{P_1})}{(\alpha_{11} - \alpha_{21})(1 - c_{P_1}) + (\alpha_{22} - \alpha_{12})(1 - c_{P_2})}, \tag{9a}$$

$$\hat{p}_1 = \frac{c_{H_1} - c_{H_2} + \phi s\,(\alpha_{12}(1 - c_{H_1}) - \alpha_{22}(1 - c_{H_2}))}{\phi s\,((\alpha_{12} - \alpha_{11})(1 - c_{H_1}) + (\alpha_{21} - \alpha_{22})(1 - c_{H_2}))}. \tag{9b}$$

Under this model, the value of CSA at generation g (Eq. 1) can be obtained as:

$$\text{CSA}_g = \left| \frac{\alpha_{11} h_{1,g} p_{1,g} \alpha_{22} h_{2,g} p_{2,g} - \alpha_{21} h_{2,g} p_{1,g} \alpha_{12} h_{1,g} p_{2,g}}{\Delta^2} \right|, \tag{10}$$

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. 1 that $\sum_{\forall i,j} f_{ij} = 1$).

When Eq. 4 is applied to our coevolutionary model A, the CSP value at each generation g is obtained as:

$$\text{CSP}_g = \left| \frac{\phi_g(\alpha_{11}h_{1,g}p_{1,g} + \alpha_{12}h_{1,g}p_{2,g})}{(1-\phi_g)h_{1,g} + \phi_g((1-\alpha_{11})h_{1,g}p_{1,g} + (1-\alpha_{12})h_{1,g}p_{2,g})} - \frac{\phi_g(\alpha_{21}h_{2,g}p_{1,g} + \alpha_{22}h_{2,g}p_{2,g})}{(1-\phi_g)h_{2,g} + \phi_g((1-\alpha_{21})h_{2,g}p_{1,g} + (1-\alpha_{22})h_{2,g}p_{2,g})} \right|. \quad (11)$$

Based on these equations we can calculate the value of CSA and CSP at the non-trivial equilibrium point under both a MA and a GFG infection matrix.

2.3.2. 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. 12a and the change in number of infected individuals I_{ij} is given by Eq. 12b.

$$\frac{dS_i}{dt} = S_i \left[b(1-c_{H_i}) - \gamma - \sum_{j=1}^{2} \alpha_{ij}\beta(1-c_{P_j}) \sum_{k=1}^{2} I_{kj} \right] + b(1-c_{H_i})(1-s) \sum_{j=1}^{2} I_{ij}, \quad (12a)$$

$$\frac{dI_{ij}}{dt} = -\gamma I_{ij} + S_i \left[\alpha_{ij}\beta(1-c_{P_j}) \sum_{k=1}^{2} I_{kj} \right]. \quad (12b)$$

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 population through an overlap between generations. Uninfected hosts can get infected by horizontal disease transmission at 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 converge to a stable (attractor) polymorphic equilibrium point.

To simulate the dynamics, we discretize 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 discretized time dynamics match the continuous time dynamics. The changes in population sizes and in allele frequencies, as well as the 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:

$$\xi(t) = \frac{\sum_i \sum_j I_{ij}(t)}{N(t)}.$$

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

$$\text{CSA}(t) = \left| \frac{I_{11}(t) \cdot I_{22}(t) - I_{12}(t) \cdot I_{21}(t)}{(\sum_i \sum_j I_{ij}(t))^2} \right|. \tag{13}$$

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

$$\text{CSP}(t) = \left| \frac{I_{11}(t) + I_{12}(t)}{S_1(t)} - \frac{I_{21}(t) + I_{22}(t)}{S_2(t)} \right|. \tag{14}$$

2.3.3. Numerical simulations

For both models we conduct numerical simulations under a symmetric matching-alleles (Tab. 1a), an asymmetric matching-alleles (Tab. 1b) and a GFG-interaction (Tab. 1c) interaction over 500 generations. In line with previous studies we assume no genotype costs $c_{H_1} = c_{H_2} = c_{P_1} = c_{P_2} = 0$ for both symmetric and asymmetric MA interactions (Gandon and Nuismer 2009). Under a GFG interaction model we assume that $0 < c_{H_2}, c_{P_2} < 1$ and $c_{H_1} = c_{P_1} = 0$ (Tellier and Brown 2007b). We calculate the values of CSA$_r$/CSA$'$/CSP at any generation, based on the equations above (Eqs. 10, 11, 13, 14) and compare them to a FDR of 1%, based on the theoretical SFS distributions for neutral loci computed in section 2.2.2.

We also access the power of using time sample data to pinpoint the coevolutionary locus. As in natural populations it may prove difficult to obtain time samples spanning extended periods of time, we use a simple, realistic and affordable (in terms of sequencing costs) scheme with three time samples taken within an interval of ten generations. In natural systems the state of coevolutionary dynamics is usually unknown, and thus on purpose, we do not optimize the sampling strategy to detect coevolutionary cycles. We present the analysis for two sampling schemes, denoted Z_1 and Z_2, which are chosen to represent the early and later phase of the simulations. For all dynamics, we fix Z_1 to consist of time samples from generations 50, 55 and 60 and Z_2 from generations 250, 255 and 260. All codes used for generating and analyzing the data can be obtained from: URL XY.

3. Results

3.1. Hypothetical numerical examples

In order to explicit the rationale behind our indices, we consider six cases with decreasing levels of association (maximum association for case 1, random association for case 6). Case 1 for CSP in Tab. 3 is chosen such that $F_{i0} \neq 0$ and therefore, CSP is defined. We compute the p-value of a χ^2-test for each case. Both our indices reflect the strength of the association, as the FDR is very low for the example with the strongest association, especially for CSA′ (Table 2, Table 3). Further, we show that our indices and their FDR values show the same trend as the p-value of the corresponding χ^2-test. When correcting for multiple testing by applying a Bonferroni-correction based on a arbitrary (but realistic, see Ansari et al. 2017, Bartha et al. 2013) number of $m_H * m_P = 10^7$ comparisons, the SNP-pairs of cases 1, 2 and 3 in Tab. 2 would be still significant under the χ^2-test. However, the FDR CSA_r for case 2 is around $4e^{-4}$, which would result in $4,000$ false positive SNP pairs under $m_H * m_P = 10^7$. In contrast, the FDR for CSA′ is lower than 10^{-12}, which would correspondingly translate into only a handful (if any) false positive SNP pairs. Nevertheless, our indices are robust and stringent (low FDR) as they will only pinpoint with high accuracy the host-parasite SNP-pair, which represents the highest level of association.

3.2. Analytical results for model A

In the methods and material section we computed the threshold for detecting pairs of coevolving SNPs for a large number of pairwise SNP comparisons. We now quantify the usefulness and power of our statistics to uncover genes under coevolution in a natural setting with coevolutionary dynamics and different underlying infection matrices.

3.2.1. Under a matching-alleles infection matrix

For a matching-allele infection matrix (Table 1b) and for $c_{H_1} = c_{H_2} = c_{P_1} = c_{P_2} = 0$, the frequencies at the non-trivial equilibrium point in model A (Eq. 9) are:

$$\widehat{p}_1 = \frac{\omega_1}{\omega_1 + \omega_2}, \text{ and } \widehat{h}_1 = \frac{\omega_2}{\omega_2 + \omega_1}.$$

Combining these equilibrium frequencies with Eq. 10 and Eq. 11 we can obtain the values of the indices at the polymorphic equilibrium point.

$$\widehat{\text{CSA}}_{\text{MA}} = \frac{\omega_1^2 \omega_2^2}{(\omega_1 + \omega_2)^2 (\omega_1 + \omega_2 - \omega_2 \omega_1)} \text{ , and } \widehat{\text{CSP}}_{\text{MA}} = 0. \tag{15}$$

For a matching alleles (MA) interaction without any genotype costs, CSP is always zero at the equilibrium point (Eq. 15), irrespective of the values of ω_1 and ω_2. Under the MA model, the values of the CSA and CSP display a different behavior over time and the value of CSA at the equilibrium depends on ω_1 and ω_2. This implies that tracking CSA and CSP values over time yields insights into the asymmetry of the infection matrix (Eq. 15, Figs. 2, 4).

3.2.2. Under a gene-for-gene infection matrix

For a gene-for-gene infection matrix (Table 1c) and for $0 < c_{H_2}, c_{P_2} < 1$ and $c_{H_1} = c_{P_1} = 0$, the non-trivial equilibrium frequencies are given by:

$$\widehat{p}_1 = \frac{c_{H_2}(1 - s\phi)}{\omega \phi s (1 - c_{H_2})} \text{ , and } \widehat{h}_1 = \frac{\omega - c_{P_2}}{\omega}.$$

Inserting these frequencies into Eq. 10 and Eq. 11 we can obtain the values of the indices at the polymorphic equilibrium point.

$$\widehat{\text{CSA}}_{\text{GFG}} = \frac{\frac{c_{P_2} c_{H_2}\left(\omega - c_{P_2}\right)(1-s\phi)}{\omega^2 \phi s\left(1-c_{H_2}\right)}\left(1 - \frac{c_{H_2}(1-s\phi)}{\omega\phi s\left(1-c_{H_2}\right)}\right)}{\left(1 - \frac{c_{P_2} c_{H_2}(1-s\phi)}{\omega\phi s\left(1-c_{H_2}\right)}\right)^2}, \tag{16a}$$

$$\widehat{\text{CSP}}_{\text{GFG}} = \frac{c_{H_2}\left(1-\phi s\right)}{\left(1-\phi\right)\left(\left(1-\phi\right)s + c_{H_2}\left(1-s\right)\right)}. \tag{16b}$$

Note that under a gene-for-gene model (GFG), the equilibrium CSP values solely depend on host parameters c_{H_2}, ϕ and s (independent of parasite cost c_{P_2}), whereas the equilibrium CSA values depend on all the parameters in both species (Eq. 16, Fig. 3). The conditions under which CSP is defined are more restrictive under GFG than under MA. Unlike for MA, CSP is not defined when the exposed fraction is maximal ($\phi = 1$) (see Eq. 11).

3.3. Numerical simulations

3.3.1. Arms race dynamics under model A

When simulating an asymmetric MA interaction for one representative parameter combination ($\omega_1 = 0.9$, $\omega_2 = 0.7$) 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 corresponding allele frequency changes. Overall, CSA values decrease over time and result in high FDRs 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 host and parasite alleles in the infected sample become too weak to be observable (Fig. 2b,c). Under an arms race dynamic, one host allele tends to reach a very high frequency and is subsequently tracked down by the parasite which generates the observed coevolutionary cycles. Along these cycles there is only a very limited amount of time at which both host and

parasite alleles are found at intermediate frequencies yielding high values of CSA. On the other hand, the CSP values under the MA model are consistently high with very low FDR, exhibiting thus enough statistical power to detect the loci under coevolution (Fig. 2d,g). This demonstrates the importance of obtaining additional non-infected host samples. Under the same model A with a symmetric MA interaction, we find similar outcomes as in Fig. 2. The oscillations of CSA and CSP values perfectly match the allele frequency cycles and show a regular amplitude pattern (Fig. S8). Under the GFG model with arms race dynamics, CSA values are consistently small with narrow peaks (Fig. 3b,c) which are barely below the 1% FDR threshold (Fig. 3e,f). The CSP has several peaks at relatively low FDR (Fig. 3d,g). Yet, it maybe difficult to detect the coevolutionary loci even when time samples are available if a stringently low FDR is required (Fig. 3g). Comparing the results for a MA and a GFG-interaction under arms race dynamics shows that the combination of CSA and CSP values over time gives some indication about the (a)symmetry of the infection matrix, while the CSP exhibits the highest power to disentangle loci under coevolution from the neutral background. 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.

3.3.2. Trench warfare and arms race dynamics under model B

Under Model B, trench warfare dynamics can take place for MA-interactions and thus, allele frequencies can converge to a stable polymorphic equilibrium. Once the allele frequencies are at equilibrium, CSA accurately distinguishes the coevolving loci from the neutral background (Fig. 4b,c,e,f), while CSP decreases to zero (Fig. 4d). In Model B, the disease prevalence varies over time as a function of the changes in allele frequencies, so that CSP is expected to vary over time. However, once the allele frequencies have converged to a stable equilibrium the numbers in all host compartments eventually remain constant. Whenever the ratio of infected to non-infected individuals is the same for both host types the CSP drops to zero. The value of CSA at equilibrium depends on the respective equilib-

rium frequencies and thus, is highest when alleles are at frequency 0.5 (Eq. 15) resulting in a a very low FDR. Even if coevolution results in stably sustained cycles, the CSA values would remain high as long as the amplitude of the cycles does not become too large and therefore alleles 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) with very low FDR, while the CSP is fixed to zero (under a symmetric MA infection matrix) with a maximum FDR. 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.

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, though 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. 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 assumption, while the CSA_r is slightly more precise under trench warfare dynamics when allele frequencies reach very high or very low values (close to fixation or loss).

3.3.3. Added value of time sampling

Our sampling schemes Z_1 and Z_2 fall into distinct parts of the allele frequency dynamics (Fig. 2-4 and Fig. S8-S10). For the analyzed asymmetric matching-alleles and symmetric matching-alleles interaction under model A the first interval Z_1 is located around a peak with respect to CSA and CSP (Fig. 2 and Fig. S8) whereas it is located in a local mini-

mum for the GFG interaction (Fig. 3). The second time interval Z_2 coincides with very small values for the asymmetric MA-model (Fig. 2), the transition into a peak for the GFG model (Fig. 3) and a peak for the symmetric MA-model (Fig. S8). For the asymmetric MA-interaction under Model B, the first time interval encompasses a transition into a local maximum with respect to CSA with the allele frequencies still showing minor fluctuations but being already close to the equilibrium point (Fig. 4). Correspondingly, the CSP values are already very close to zero. The second time interval falls into the part of the dynamic where the allele frequencies have already converged to the equilibrium and hence CSA and CSP have also converged to constant values. The same pattern can be seen under a corresponding symmetric MA-interaction for both intervals Z_1 and Z_2 (Fig. S9). For the corresponding GFG-interaction both time intervals fall into the part of the dynamic with strong positive selection and therefore, both time intervals are characterized by low CSA and CSP values.

We also compare the χ^2-tests and FDR values across time-samples within time intervals and between time intervals (Tab. 4-6 and Tab. S1-S3). Both, the symmetric and the asymmetric MA-allele interaction in Model B result in significant CSA-χ^2-association tests irrespective of the time-point and interval considered, even if a stringent Bonferroni-correction is applied (Tab. 6, Tab. S2). The corresponding FDRs for CSA$'$ are very low whereas the FDRs for CSA$_r$ would result in large number of false positive when the number of comparisons is large. These results are in stark contrast to all interactions involving arms race dynamics (model A and a GFG-interaction for model B)(Tab. 4, Tab. 5, Tab. S1). For the asymmetric matching-alleles interaction in model A the CSA-χ^2-test for $g = 50$ is significant where it is not for the other two time samples within the same interval Z_1. This reflects the fact that the sample at $g = 50$ exactly falls on the peak for CSA values whereas the other two samples within the same interval already fall slightly outside the peak. On the other hand the CSP-χ^2-tests for the other two time samples within Z_1 are significant as are all three of them in Z_2 (Tab. 4). The corresponding results for the symmetric MA are overall similar, though none of the time samples in the first interval is exactly located on the local maximum of the CSA-peak. The time samples for the GFG-dynamic under model A (Tab. 5,

Fig. 3) and model B (Tab. S3) do not show any significant χ^2-test for both intervals. We also note that in all studied examples the number of false positives, for both CSA and CSP, are fairly high if a large number of loci is to be analyzed irrespectively of the significance of the corresponding χ^2-tests.

4. Discussion

With technological advances, full genome sequencing of several hosts along with the parasite strains infecting them has become affordable. Recent studies (Ansari et al. 2017, Bartha et al. 2013, Lees et al. 2019) tested for the degree of association between different host and parasite SNPs by performing millions of pairwise comparisons to obtain a genome-wide natural co-GWAs. Here we develop two indices, CSA and CSP, measuring the strength of association between host and parasite SNPs. We demonstrate that the statistical power to locate the loci under coevolution depends on the 1) symmetry of the infection matrix, 2) type of dynamics, and 3) state of the allele frequency changes (whether a stable polymorphic equilibrium has been reached or not). We further show that CSA and CSP, provide different information with respect to these properties and can complement each other by having maximum power at different time points.

Based on previous studies it is expected that the power to detect coevolutionary loci varies in time due to the coevolutionary dynamics and the underlying allele frequency changes (Gandon and Nuismer 2009, MacPherson et al. 2018, Nuismer et al. 2017, Nuismer and Week 2019). 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) and Nuismer et al. (2017), which also shows variable statistical power over time to detect coevolution. Therefore, we suggest obtaining samples

from several time points, rather than a single time point, to extract as much information as possible. A similar idea was proposed in Gandon et al. (2008) for the study of local adaptation when using phenotypic (infection) data.

One crucial result of our study is the derivation of an indicative False Discovery Rate (FDR) for the association indices CSA and CSP. To do so, we assume that the analyzed samples originate from panmictic host and parasite populations with constant size. This allows us to apply population genetics theory to obtain the expected power and FDR of our indices for populations at drift-mutation equilibrium. In natural systems host and parasite population sizes (termed as the demographic history in population genetics) can change over time due to 1) eco-evolutionary feedback arising from the epidemiological dynamics (for example in model B), and/or due to 2) abiotic environmental factors such as resource availability and habitat suitability. It is possible to account for both type of demographic changes, by incorporating previous results on how they change the neutral SFS (Živković et al. 2019, 2015) into the calculation of the expected neutral distributions.
Further, it is known from a previous study (Živković et al. 2019) that the changes in neutral SFS under model B due to the coevolutionary cycles are very minor given that the initial population sizes are large enough. As we always consider large initial population sizes in our analysis, we therefore can roughly approximate the SFS by a stationary state SFS.

We assume that coevolution only involves one major bi-allelic locus per species. However, the MHC (major histocompatibility complex) loci in vertebrates (Horton et al. 2004) or NLRs (nucleotide-binding leucine rich repeat immune receptors) in plants (Van de Weyer et al. 2019) represent examples of gene families with complex genomic and functional structures which are likely to be involved into coevolution. We think that our indices are informative when several independent major loci (freely recombining and no epistatic or pleiotropic effects) are involved into the coevolutionary interaction as our indices are applied independently to each pair of host and parasite loci.
In the case of epistasis, it has been shown that positive epistatis increases the specificity

of the parasite whereas negative epistasis widens the range of parasite infectivity (Ashby et al. 2014). Therefore, we expect the power of our indices to pinpoint the coevolving loci to increase (decrease) in the former case (in the latter case).

While we assume bi-allelic SNPs, several haplotypes can form the genetic basis of coevolution. Each haplotype can either consist of several alleles at different sites within a given gene (Seger 1988) or of bi-allelic SNPs across several linked loci (Tellier and Brown 2007a). In such cases the signature of coevolution and association (between host and parasite) would be weak for each single SNP and only visible once the defined haplotypes become the unit of analysis.

In principle, our indices can be extended to more than two-alleles per locus (Seger 1988) or to linked loci (Tellier and Brown 2007a), but theory suggests that, for these genetic structures, the coevolutionary dynamics become more complex (Seger 1988, Tellier and Brown 2007a).

We conclude by presenting a set of recommendations for applying this method to different host-parasite systems. One key requirement of this method is to keep track of the co-occurrence of host and parasite genotypes, which can be technically challenging especially in pool-sequencing approaches (as in Retel et al. 2019). It is advised to obtain random samples from infected and non-infected hosts and parasites at several time points from natural populations or from controlled coevolution experiments. Furthermore, sample sizes do affect the FDR and small sample sizes ($n_T < 25$) decrease the power to detect the coevolving loci substantially (see Supplementary Online Material). Therefore, we suggest to collect large samples at few time points, rather than sampling small samples at several time points. 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 (Tellier and Brown 2007b). Therefore, it is expected that the cross-species association method has a high power when applied to parasites with strong life-cycle dependency on their hosts. This effect should be strongly pronounced for parasites which have a much shorter generation times than their hosts, such as viruses (Ansari

et al. 2017, Bartha et al. 2013) or bacterial pathogens of humans (Lees et al. 2019). For such types of parasites, taking serial samples within a single host generation 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, especially if the infection matrix follows a MA model.

An implicit assumption in Ansari et al. (2017), Bartha et al. (2013), Bartoli and Roux (2017), Lees et al. (2019) and our model is that disease transmission is random and panmictic so that potentially every host can get into close contact with the disease (no population sub-structuring affecting disease transmission). Thus, before performing a cross-species association, it is crucial to assess and correct for 1) population sub-structure and kinship relationships within single populations, and 2) overall spatial structure. Additional bias can be caused by neglecting population structure when different populations are at different stages of the coevolutionary cycle (*e.g.* Gavrilets and Michalakis 2008, Sasaki 2000).

Co-infections are common for many diseases (Alizon et al. 2013, Tollenaere et al. 2016). A solution to deal with such multi-strain infections in a co-GWAS-setting is to base the analysis on the major parasite strain found on the infected host (Bartha et al. 2013).

The independence between neutral and coevolving loci can be violated for viruses, bacteria or clonal fungi due to the potential absence of recombination. This problem can addressed by following the approach outlined in Bartha et al. (2013), by first inferring a phylogeny of the virus samples, followed by a subsequent identification of clusters of polymorphic SNPs across the phylogeny and performing the co-GWAs based on the identified clusters (Ansari et al. 2017, Bartha et al. 2013).

Host-parasite coevolution is a multifaceted process. We have shown that natural co-GWAs have the power to pinpoint genes involved in this process especially when time-samples are available, while we acknowledge potential shortcomings due to model assumptions, and dependency on the underlying infection matrix and coevolutionary allele frequency changes. The approach presented here is also potentially applicable to detect

major genes which are determining the compatibility between symbionts in mutualistic interactions. A further worthwhile development will be the explicit inclusion of spatial structure of host and parasite populations.

Acknowledgments

This work was funded by the German Research Foundation (DFG). HM received support from grant numbers TE809/3-1 and 4-1 (awarded to Aurélien Tellier) within the DFG Priority Programme 1819. SJ received support from grant number TE809/3-2 (awarded to Aurélien Tellier) within the DFG Priority Programme 1819 and grant number TE809/6-2 (awarded to Aurélien Tellier and Wolfgang Stephan) within the DFG Priority Programme 1590.

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., L C Ip, 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., Consortium, S.T.O.P.H.C.V., Hudson, E., McLauchlan, J., Simmonds, P., Bowden, R., Klenerman, P., Barnes, E., Spencer, C.C.A., 2017. Genome-to-genome analysis highlights the effect of the human innate and adaptive immune systems on the Hepatitis C virus. Nature Genetics 49, 666–673. URL: `https://pubmed.ncbi.nlm.nih.gov/28394351`, doi:10.1038/ng.3835. 28394351[pmid].

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., Gupta, S., Buckling, A., 2014. Effects of epistasis on infectivity range during host-parasite coevolution. Evolution 68, 2972–2982. doi:10.1111/evo.12479.

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, 763. URL: `https://www.frontiersin.org/article/10.3389/fpls.2017.00763`, 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. Annual Review of Phytopathology 49, 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.

Durrett, R., 2010. Probability models for DNA sequence evolution. 2nd ed., 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.

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. American 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.

Horton, R., Wilming, L., Rand, V., Lovering, R.C., Bruford, E.A., Khodiyar, V.K., Lush, M.J., Povey, S., Talbot, C.C., Wright, M.W., et al., 2004. Gene map of the extended human MHC. Nature Reviews Genetics 5, 889–899.

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

Lees, J.A., Ferwerda, B., Kremer, P.H.C., Wheeler, N.E., Serón, M.V., Croucher, N.J., Gladstone, R.A., Bootsma, H.J., Rots, N.Y., Wijmega-Monsuur, A.J., Sanders, E.A.M., Trzciński, K., Wyllie, A.L., Zwinderman, A.H., van den Berg, L.H., van Rheenen, W., Veldink, J.H., Harboe, Z.B., Lundbo, L.F., de Groot, L.C.P.G.M., van Schoor, N.M., van der Velde, N., Ängquist, L.H., Sørensen, T.I.A., Nohr, E.A., Mentzer, A.J., Mills, T.C., Knight, J.C., du Plessis, M., Nzenze, S., Weiser, J.N., Parkhill, J., Madhi, S., Benfield, T., von Gottberg, A., van der Ende, A., Brouwer, M.C., Barrett, J.C., Bentley, S.D., van de Beek, D., 2019. Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis. Nature Communications 10, 2176–2176. doi:10.1038/s41467-019-09976-3.

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.

Nuismer, S.L., Week, B., 2019. Approximate Bayesian estimation of coevolutionary arms races. PLOS Computational Biology 15. doi:10.1371/journal.pcbi.1006988.

Retel, C., Kowallik, V., Huang, W., Werner, B., Künzel, S., Becks, L., Feulner, P.G.D., 2019. The feedback between selection and demography shapes genomic diversity during coevolution. Science Advances 5. doi:10.1126/sciadv.aax0530.

Ruckdeschel, P., Kohl, M., Stabla, T., Camphausen, F., 2006. S4 classes for distributions. R News 6, 2–6.

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.

Seger, J., 1988. Dynamics of some simple host-parasite models with more than 2 genotypes in each species. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences 319, 541–555. doi:10.1098/rstb.1988.0064.

Stahl, E., Bishop, J., 2000. Plant-pathogen arms races at the molecular level. Current Opinion in Plant Biology 3. 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., 2007a. Polymorphism in multilocus host-parasite coevolutionary interactions. Genetics 177, 1777–1790. doi:10.1534/genetics.107.074393.

Tellier, A., Brown, J.K.M., 2007b. 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.

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 , 201710980doi:10.1073/pnas.1710980115.

Van de Weyer, A.L., Monteiro, F., Furzer, O.J., Nishimura, M.T., Cevik, V., Witek, K., Jones, J.D.G., Dangl, J.L., Weigel, D., Bemm, F., 2019. A species-wide inventory of NLR genes and alleles in *Arabidopsis thaliana*. Cell 178, 1260–1272.e14. doi:10.1016/j.cell.2019.07.038.

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. BMC Evolutionary Biology 19. URL: `https://doi.org/10.1186/s12862-019-1556-3`, doi:10.1186/s12862-019-1556-3.

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.

Data accessibility statement

A GitHub repository is being created with all the codes used for simulating data, calculating distributions and analyzing results. This will be made publicly available upon acceptance.

Author contributions

All three authors (HM, AT and SJ) designed the study, implemented the simulations, analyzed the results, wrote the manuscript and revised the submitted version of the manuscript.

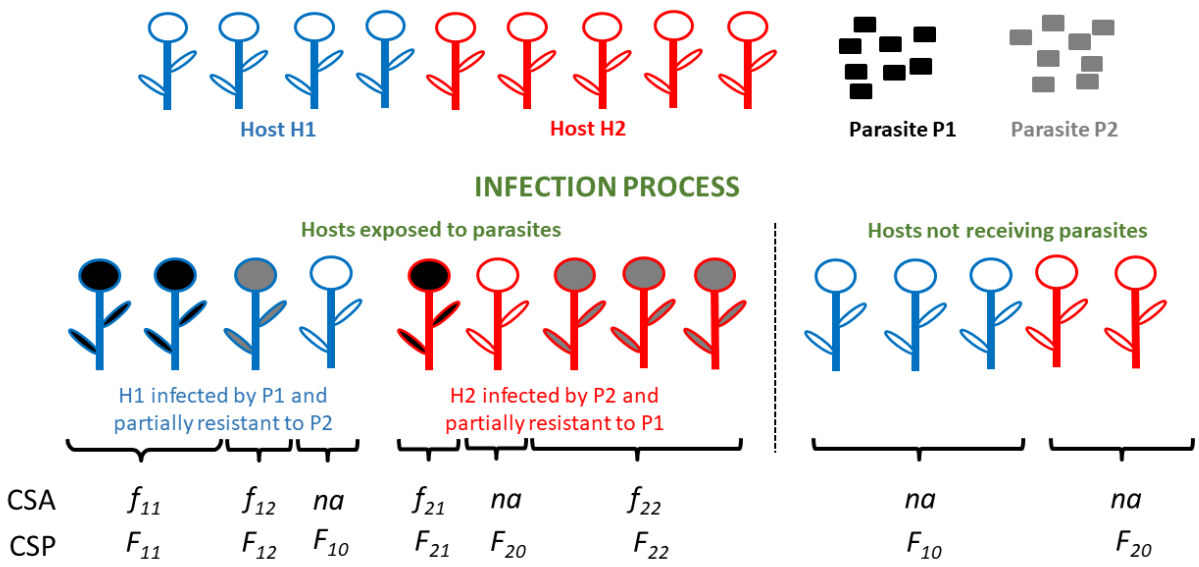

Figure 1 Schematic representation of the CSA and CSP definitions. The host population consists of two host types H_1 (square) and H_2 (circle) and the parasite population consists of two parasite types P_1 (black) and P_2 (grey). A proportion ϕ of the hosts is exposed to parasites. Hosts exposed to the parasite either become infected or they can resist infection. Infected hosts are colored based on the identity of the infecting parasite genotype (grey or black). f_{ij} is the proportion of hosts type i infected by parasites of type j among infected hosts. F_{ij} is the proportion of hosts of type i infected 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 type i hosts which did not receive spores $(1-\phi)$ or received spores and are resistant to the respective parasite.

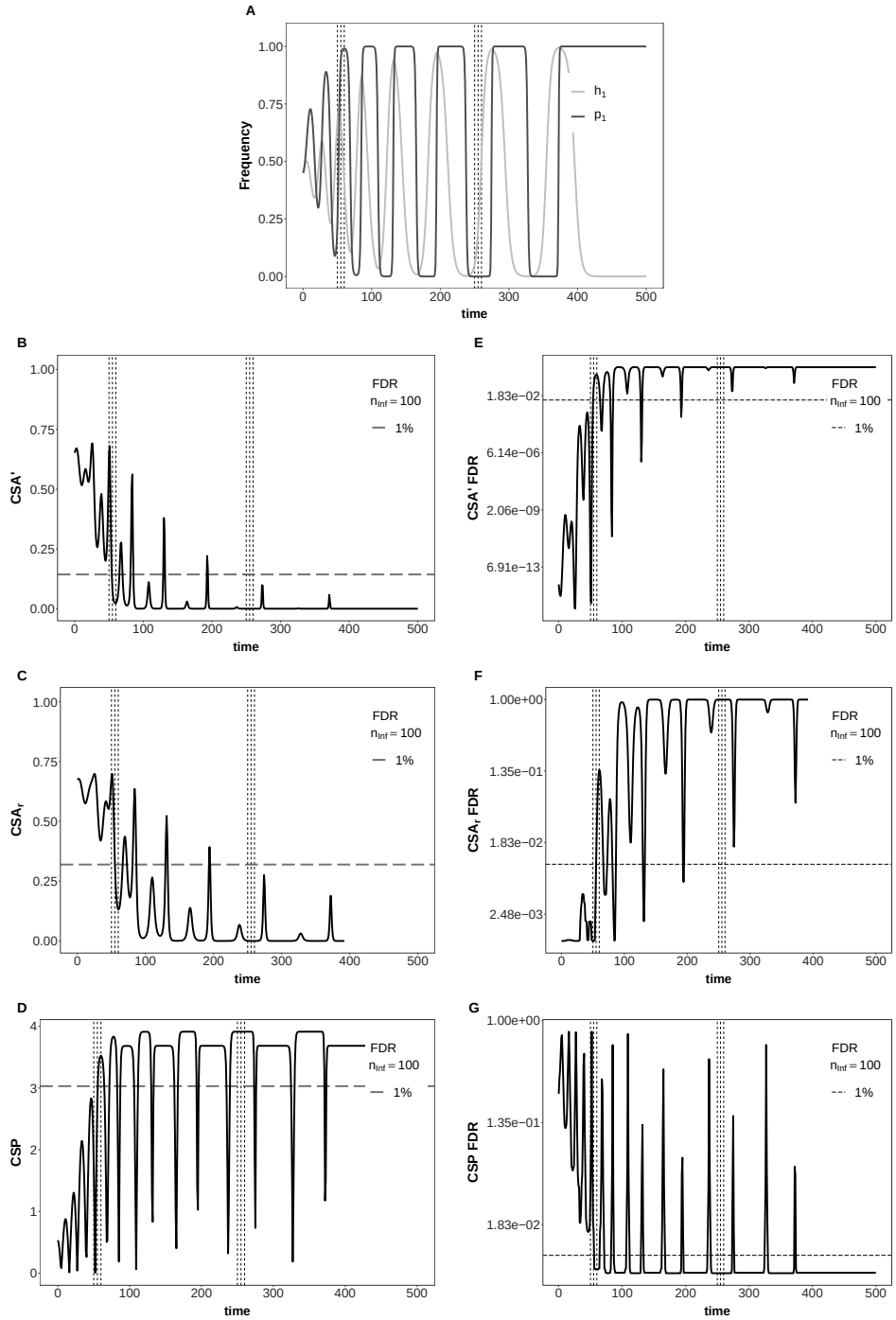

Figure 2 Temporal changes in allele frequencies, $\mathrm{CSA}', \mathrm{CSA}'_r$ and CSP in an unstable asymmetric MA-model (model A). Top: Temporal changes in allele frequencies. Left: Temporal changes of the three association indices. Right: corresponding FDR. Dashed lines correspond to a 1%-FDR, assuming $n_T = 200$ and $n_{\text{Inf}} = n_{\text{H}} = 100$. The model parameters are $\omega_1 = 0.9$, $\omega_2 = 0.7$, $\phi = 0.8$, $s = 0.35$, initial values $h_{1,g=0} = p_{1,g=0} = 0.45$. The y-axes in subfigures E, F and G have been log-transformed for better visualization of low FDR levels.

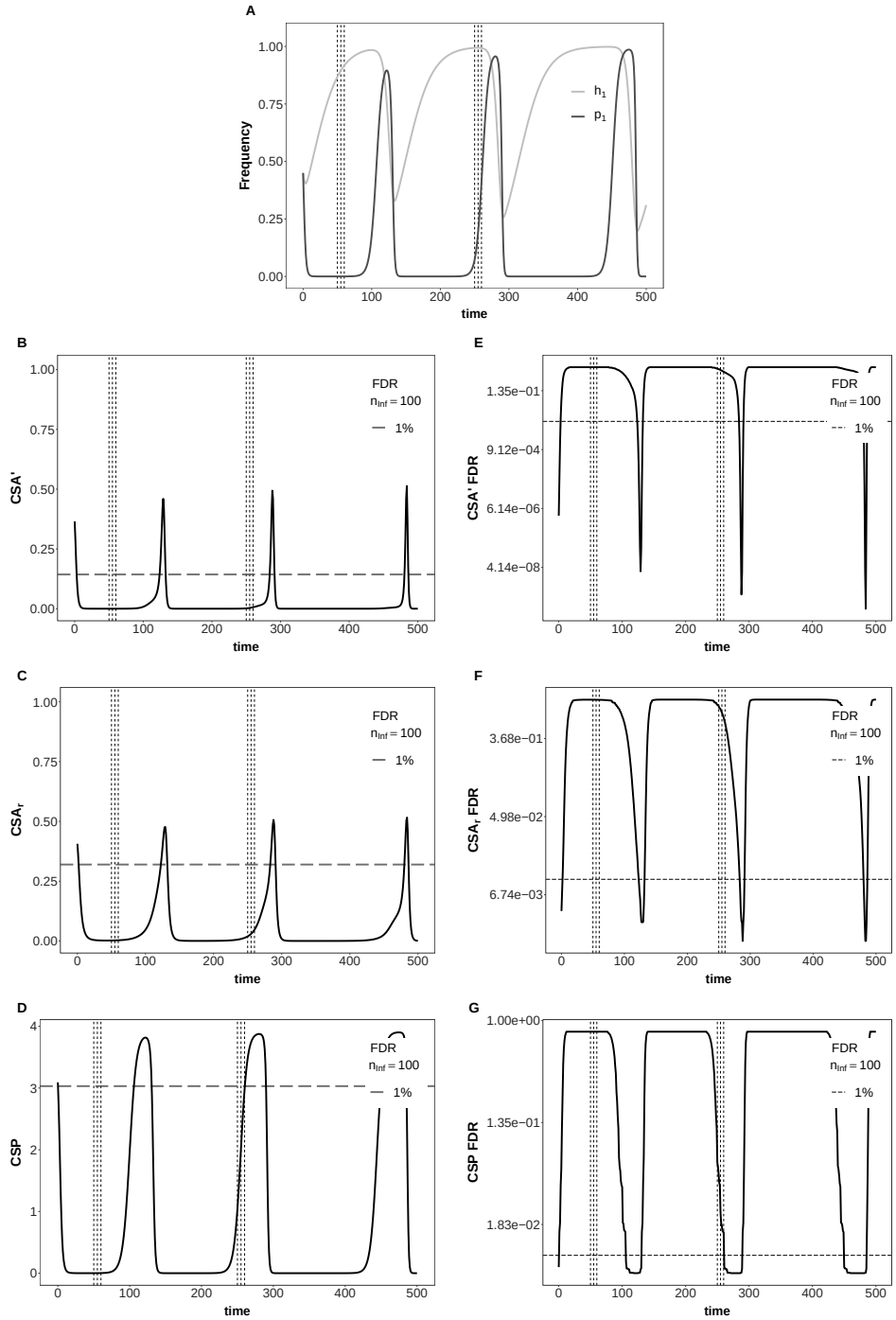

Figure 3 Temporal changes in allele frequencies, $\mathrm{CSA}', \mathrm{CSA}'_r$ and CSP in an unstable GFG-model (model A). Top: Temporal changes in allele frequencies. Left: Temporal changes of the three association indices. Right: corresponding FDR. Dashed lines correspond to a 1%-FDR assuming $n_T = 200$ and $n_{\mathrm{Inf}} = n_{\mathrm{H}} = 100$. The model parameters are $c_{H_1} = c_{P_1} = 0$, $c_{H_2} = 0.05$, $c_{P_2} = 0.2$, $\phi = 0.8$, $s = 0.35$, $\omega = 0.9$, initial values $h_{1,g=0} = p_{1,g=0} = 0.45$. The y-axes in subfigures E, F and G have been log-transformed for better visualization of low FDR levels.

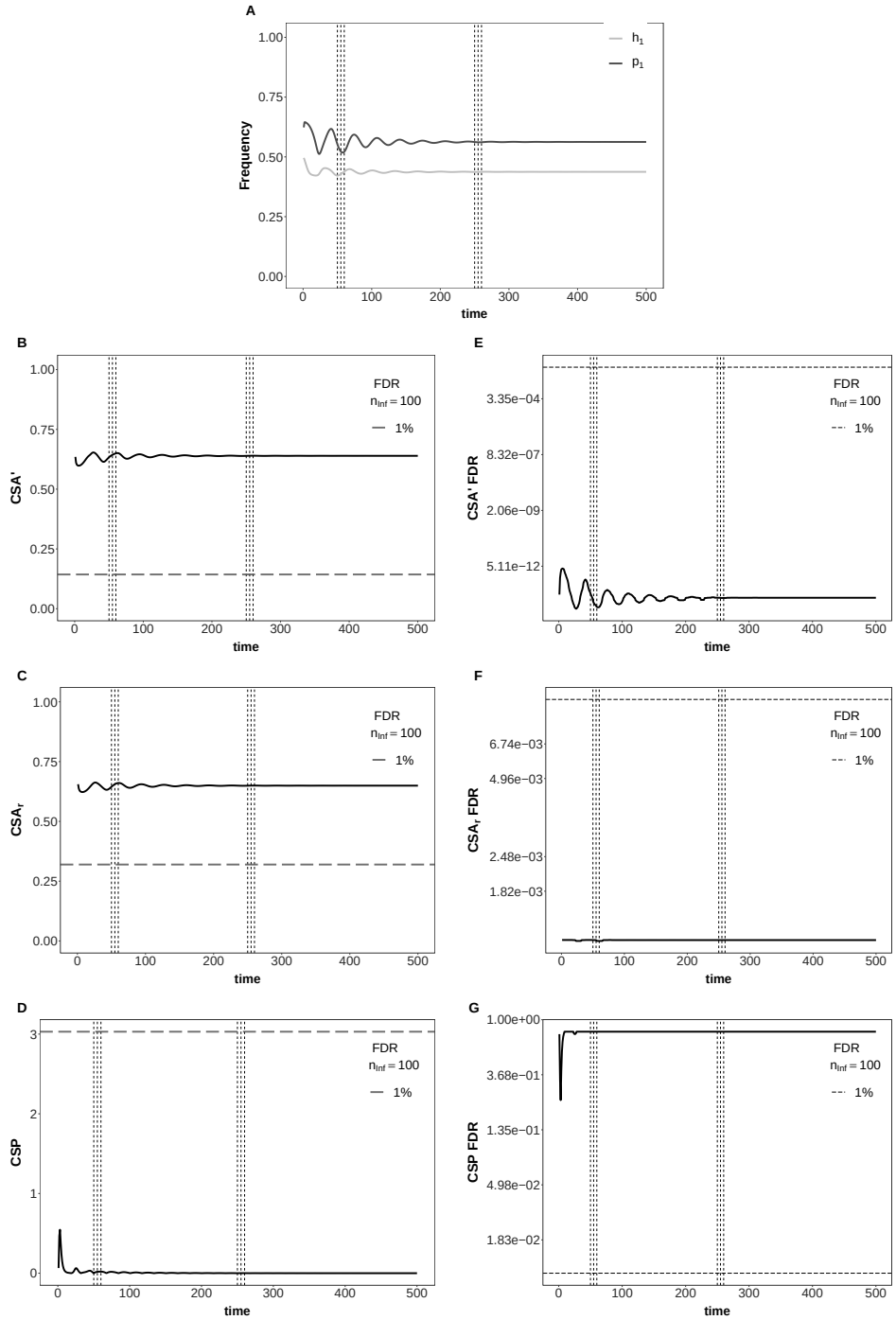

Figure 4 Temporal changes in allele frequencies, CSA', CSA'_r and CSP in an epidemiological (model B) with an asymmetric MA-infection matrix. Top: Temporal changes in allele frequencies. Left: Temporal changes of the three association indices. Right: corresponding FDR. Dashed lines correspond to a 1%-FDR assuming $n_T = 200$ and $n_{\text{Inf}} = n_{\text{H}} = 100$. The model parameters are $\beta = 0.00005$, $s = 0.6$, $\omega_1 = 0.9$, $\omega_2 = 0.7$, $b = 1$, $\gamma = 0.9$, and $c_{H_1} = c_{P_1} = c_{H_2} = c_{P_2} = 0$. The initial values are $S_{1,t=0} = S_{2,t=0} = 41500$, $I_{11,t=0} = I_{12,t=0} = I_{21,t=0} = I_{22,t=0} = 4150$. The time interval for computations is $\delta_t = 0.001$. The y-axes in subfigures E, F and G have been log-transformed for better visualization of low FDR levels.

Table 1 Infection matrices for coevolutionary models. 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 host to be infected or as the degree of infection (disease severity or partial resistance).

a) general infection matrix	**b) matching-allele**	**c) gene-for-gene**
$\begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix}$	$\begin{pmatrix} 1 & 1-\omega_1 \\ 1-\omega_2 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 1 \\ 1-\omega & 1 \end{pmatrix}$

Table 2 Six hypothetical numerical examples to illustrate the strength of association when sampling only infected individuals and the relevant statistics: the p-value of a χ^2-test, CSA value and the corresponding false discovery rate (FDR) based on neutral allele frequencies. The examples are build assuming both host and parasite loci with a minor allele frequency of 0.5. The FDR is calculated for a sample size of $n_{\text{Inf}} = 100$. Asterisks (*) indicate significant p-values after Bonferroni correction for 10^7 comparisons.

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
f_{11}	0.5	0.45	0.40	0.35	0.30	0.25
f_{12}	0	0.05	0.10	0.15	0.20	0.25
f_{21}	0	0.05	0.10	0.15	0.20	0.25
f_{22}	0.5	0.45	0.40	0.35	0.30	0.25
p-value χ^2	$1.52e^{-23*}$	$1.24e^{-15*}$	$1.97e^{-9*}$	$6.33e^{-5}$	$4.55e^{-2}$	1
CSA value	1	0.8	0.6	0.4	0.2	0
FDR CSA_r	$< 10^{-12}$	$3.9e^{-4}$	$1.2e^{-3}$	$4.5e^{-3}$	$4.8e^{-2}$	1
FDR CSA'	$< 10^{-12}$	$< 10^{-12}$	$3.04e^{-12}$	$6.81e^{-7}$	$1.88e^{-3}$	0.996

Table 3 Six hypothetical numerical examples to illustrate the strength of association when sampling infected and non-infected individuals and the relevant statistics: the p-value of a χ^2-test, CSP value and the corresponding false discovery rate (FDR) based on neutral allele frequencies. The examples are build assuming both host and parasite loci with a minor allele frequency of 0.5. The FDR is calculated for sample sizes $n_T = 200$ and $n_{\text{Inf}} = 100$. Asterisks (*) indicate significant p-values after Bonferroni correction for 10^7 comparisons.

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
$F_{11} + F_{12}$	0.49	0.45	0.40	0.35	0.30	0.25
F_{10}	0.01	0.05	0.10	0.15	0.20	0.25
$F_{21} + F_{22}$	0.01	0.05	0.10	0.15	0.20	0.25
F_{20}	0.49	0.45	0.40	0.35	0.30	0.25
p-value χ^2	$5.52e^{-42^*}$	$1.12e^{-29^*}$	$2.15e^{-17^*}$	$1.54e^{-8}$	$4.68e^{-3}$	1
CSP value	48.98	8.889	3.75	1.905	0.833	0
FDR CSP	$< 10^{-12}$	$1.69e^{-4}$	$7.1e^{-3}$	$3.83e^{-2}$	0.12	0.8044

Table 4 Comparison of the χ^2-test results and the FDRs corresponding to the values of CSA', CSA$_r$ and CSP for the six time samples taken under an asymmetric MA for model A (see Fig. 2). The model parameters are $\omega_1 = 0.9$, $\omega_2 = 0.7$, $\phi = 0.8$, $s = 0.35$, with initial values $h_{1,g=0} = p_{1,g=0} = 0.45$ and sample sizes $n_T = 200$ and $n_{\text{Inf}} = 100$. Asterisks (*) indicate significant p-values after Bonferroni correction for 10^7 comparisons.

	g=50	g=55	g=60	g=250	g=255	g=260
χ^2_{CSA}	4.65e+01	1.11e+01	1.74e+00	6.41e-05	1.96e-06	6.71e-07
p-value	8.99e-12*	8.82e-04	1.87e-01	9.94e-01	9.99e-01	9.99e-01
CSA$'$	6.50e-01	7.80e-02	2.37e-02	9.34e-08	4.33e-09	3.53e-09
FDR CSA$'$	6.38e-14	6.04e-02	3.51e-01	9.96e-01	9.96e-01	9.96e-01
CSA$_\text{r}$	6.82e-01	3.33e-01	1.32e-01	8.00e-04	1.40e-04	8.19e-05
FDR CSA$_\text{r}$	1.17e-03	9.51e-03	1.39e-01	9.96e-01	9.96e-01	9.96e-01
χ^2_{CSP}	3.46e+01	4.36e+01	5.83e+01	2.86e+01	7.35e+01	1.06e+02
p-value	3.98e-09	4.04e-11*	2.27e-14*	9.12e-08	1.03e-17*	7.07e-25*
CSP	1.75e+00	2.80e+00	3.52e+00	3.91e+00	3.91e+00	3.91e+00
FDR CSP	3.95e-02	1.58e-02	7.59e-03	7.04e-03	7.04e-03	7.04e-03

Table 5 Comparison of the χ^2-test results and the FDRs corresponding to the values of CSA', CSA$_r$ and CSP for the six time samples taken under a GFG-interaction for model A (see Fig. 3). The model parameters are $c_{H_1} = c_{P_1} = 0$, $c_{H_2} = 0.05$, $c_{P_2} = 0.2$, $\phi = 0.8$, $s = 0.35$, $\omega = 0.9$, with initial values $h_{1,g=0} = p_{1,g=0} = 0.45$, and sample sizes $n_T = 200$ and $n_{\text{Inf}} = 100$.

	g=50	g=55	g=60	g=250	g=255	g=260
χ^2_{CSA}	3.74e-04	5.14e-04	7.94e-04	3.64e-02	8.73e-02	2.19e-01
p-value	9.85e-01	9.82e-01	9.78e-01	8.49e-01	7.68e-01	6.40e-01
CSA$'$	1.47e-05	2.07e-05	3.27e-05	1.50e-03	3.21e-03	6.13e-03
FDR CSA'	9.96e-01	9.96e-01	9.96e-01	8.91e-01	7.76e-01	6.80e-01
CSA$_\text{r}$	1.93e-03	2.27e-03	2.82e-03	1.91e-02	2.95e-02	4.68e-02
FDR CSA$_\text{r}$	9.93e-01	9.92e-01	9.90e-01	8.37e-01	7.14e-01	5.51e-01
χ^2_{CSP}	9.40e-08	2.28e-07	6.99e-07	2.13e-02	1.35e-01	7.58e-01
p-value	1.00e+00	1.00e+00	9.99e-01	8.84e-01	7.14e-01	3.84e-01
CSP	6.41e-04	1.10e-03	2.14e-03	1.04e+00	2.03e+00	2.98e+00
FDR CSP	8.04e-01	8.04e-01	8.04e-01	6.15e-02	2.33e-02	1.58e-02

Table 6 Comparison of the χ^2-test results and the FDRs corresponding to the values of CSA', CSA$_r$ and CSP for the six time samples taken under an asymmetric MA for model B (see Fig. 4). The model parameters are $\beta = 0.00005$, $s = 0.6$, $\omega_1 = 0.9$, $\omega_2 = 0.7$, $b = 1$, $\gamma = 0.9$, and $c_{H_1} = c_{P_1} = c_{H_2} = c_{P_2} = 0$. The initial values are $S_{1,t=0} = S_{2,t=0} = 41500$, $I_{11,t=0} = I_{12,t=0} = I_{21,t=0} = I_{22,t=0} = 4150$, the time interval for computations is $\delta_t = 0.001$, and sample sizes are $n_T = 200$ and $n_{\text{Inf}} = 100$. Asterisks (*) indicate significant p-values after Bonferroni correction for 10^7 comparisons.

	t=50	t=55	t=60	t=250	t=255	t=260
χ^2_{CSA}	4.16e+01	4.30e+01	4.36e+01	4.22e+01	4.22e+01	4.22e+01
p-value	1.15e-10*	5.55e-11*	3.96e-11*	8.40e-11*	8.19e-11*	8.11e-11*
CSA'	6.33e-01	6.43e-01	6.50e-01	6.39e-01	6.40e-01	6.40e-01
FDR CSA'	2.73e-13	1.02e-13	6.47e-14	1.74e-13	1.68e-13	1.68e-13
CSA$_\text{r}$	6.45e-01	6.56e-01	6.61e-01	6.49e-01	6.50e-01	6.50e-01
FDR CSA$_\text{r}$	1.18e-03	1.18e-03	1.17e-03	1.18e-03	1.18e-03	1.18e-03
χ^2_{CSP}	3.05e-06	6.34e-02	8.92e-02	1.30e-07	7.10e-05	8.74e-05
p-value	9.99e-01	8.01e-01	7.65e-01	1.00e+00	9.93e-01	9.93e-01
CSP	1.60e-04	1.78e-02	1.99e-02	2.76e-05	6.44e-04	7.15e-04
FDR CSP	8.04e-01	8.04e-01	8.04e-01	8.04e-01	8.04e-01	8.04e-01