

RESEARCH ARTICLE

Open Access

# Neutral genomic signatures of host-parasite coevolution



Daniel Živković<sup>1\*</sup>, Sona John<sup>1</sup>, Mélissa Verin<sup>1,2</sup>, Wolfgang Stephan<sup>3</sup> and Aurélien Tellier<sup>1\*</sup>

## Abstract

**Background:** Coevolution is a selective process of reciprocal adaptation in hosts and parasites or in mutualistic symbionts. Classic population genetics theory predicts the signatures of selection at the interacting loci of both species, but not the neutral genome-wide polymorphism patterns. To bridge this gap, we build an eco-evolutionary model, where neutral genomic changes over time are driven by a single selected locus in hosts and parasites via a simple biallelic gene-for-gene or matching-allele interaction. This coevolutionary process may lead to cyclic changes in the sizes of the interacting populations.

**Results:** We investigate if and when these changes can be observed in the site frequency spectrum of neutral polymorphisms from host and parasite full genome data. We show that changes of the host population size are too smooth to be observable in its polymorphism pattern over the course of time. Conversely, the parasite population may undergo a series of strong bottlenecks occurring on a slower relative time scale, which may lead to observable changes in a time series sample. We also extend our results to cases with 1) several parasites per host accelerating relative time, and 2) multiple parasite generations per host generation slowing down rescaled time.

**Conclusions:** Our results show that time series sampling of host and parasite populations with full genome data are crucial to understand if and how coevolution occurs. This model provides therefore a framework to interpret and draw inference from genome-wide polymorphism data of interacting species.

**Keywords:** Population genomics, Epidemiological model, SI model, Population dynamics

## Introduction

Host-parasite antagonistic interactions are a role model for observing and studying rapid evolutionary change as well as feedbacks between ecological and evolutionary forces and time scales. Coevolution, defined here as the reciprocal adaptation of hosts and their parasites, typically generates significant phenotypic and genetic diversity for host resistance and for parasite infectivity and virulence. Such changes in the genetic composition of the interacting species at the key underpinning loci, drive, and are driven by, short-term epidemiological (ecological) dynamics. To develop infectious disease epidemiology as a predictive science, there is thus a need to understand the synergy of fast evolution and within and between populations disease dynamics [1], the so-called eco-evolutionary feedbacks [2].

Coevolution as determined by changes in allele frequencies over time at the interacting genes, is observable as coevolutionary cycles driven by negative indirect frequency-dependent selection [3, 4]. Theory predicts that a continuum of dynamics of allele frequency cycles occurs, characterized by their stability, period and amplitude, and ranging between two extremes: the arms race and the trench warfare dynamics. The arms race is defined as the recurrent fixation of alleles at these major loci in host and parasite populations [5, 6], while trench warfare maintains cycling over a long period of time [7] (also called the Red Queen dynamics [6], or Fluctuation Selection Dynamics [2]). The transition between these types of dynamics depends on the occurrence and strength of negative direct frequency-dependent selection [4], which stabilizes cycles and is generated by several host and parasite life history traits (reviewed in [8]).

Theory also predicts that the coevolutionary dynamics, either by arms race or trench warfare, can be observed in the patterns of polymorphism at these loci, namely

\*Correspondence: [daniel.zivkovic@gmx.at](mailto:daniel.zivkovic@gmx.at); [tellier@wzw.tum.de](mailto:tellier@wzw.tum.de)

<sup>1</sup>Section of Population Genetics, Technical University of Munich, Freising, Germany

Full list of author information is available at the end of the article



in the frequencies of Single Nucleotide Polymorphisms (SNPs). The arms race is expected to generate recurrent selective sweeps, while trench warfare generates balancing selection (but see [9] for more complex but realistic predictions). These predictions form the basis of scans for genes under coevolution in host or parasite genomes relying on the prevalent perception that natural selection acts only at few loci, while neutral forces, such as demographic histories, affect the whole genome. Detecting genes under coevolution entails therefore to disentangle the signatures of arms race or trench warfare from the polymorphism patterns observed in genome-wide data.

Besides the allelic coevolutionary cycles at the host and parasite interacting loci, size fluctuations of host and parasite populations are also predicted to occur and are indeed observed and quantified in controlled experiments [10]. These changes in population size over time are induced by reciprocal selection among the antagonists and are an inherent property of host-parasite coevolution under epidemiological dynamics (such as the Susceptible-Infected or Susceptible-Infected-Recovered models, [11]) but also prey-predator (Lotka-Volterra) dynamics. In a more complex coevolutionary system with several host and parasite genotypes being present at major genes of interaction, cycles of coevolution do occur, thereby generating a fluctuation of the numbers of hosts and parasites over time [11] as an epidemiological feedback [2]. Several episodes of coevolution proceed with increasing and decreasing disease prevalence depending on the cycling of resistance and infectivity alleles. The epidemiological feedback generates negative direct frequency-dependent selection, thus stabilizing the frequencies of alleles and maintaining long-term diversity at the interacting loci [12]. Coevolutionary models based on Lotka-Volterra dynamics have similar characteristics [13–16]. Population size changes due to coevolution should affect the whole genome polymorphism of both antagonistic species, an effect which we term as the co-demographic history. When studying host and parasite polymorphism data, two sources of demographic variation generating genetic drift can therefore be defined: 1) the population or species demographic history (e.g. colonisation of new habitats or recolonisation), and 2) the co-demographic history due to coevolutionary and epidemiological dynamics. Both types of demographic events affect the ability to detect genes under coevolution using scans for arms race or trench warfare signatures. Moreover, there is currently no theoretical prediction regarding the signature of co-demographic history on genome-wide polymorphism in hosts and parasites.

Our aim in this study is to propose the first model of neutral polymorphism generated by the co-demographic history of host and parasite populations. First, we establish an epidemiological model describing changes in the

numbers of healthy and infected hosts over time focusing on biallelic gene-for-gene and matching-allele infections and initially assigning one parasite per host. Second, we utilize an analytical result [17] for the neutral site frequency spectrum (SFS) under arbitrary deterministic population size changes and apply it to the host and parasite populations. We show that these population size changes can be quite drastic in the parasite and occur on a time scale slow enough to leave a corresponding signature in the SFS over time. Conversely, changes in the host size are barely detected in the polymorphism data. Finally, since such recurrent bottlenecks in parasites cause a reduced amount of polymorphism, we further discuss the impacts of multiple parasites per host and multiple parasite generations per host generation.

## Results

### Key characteristics of our model

Our model presents three key features to keep in mind. The first key aspect of our eco-evolutionary framework is that changes in the population size are a direct consequence of the dynamics of the model (Eqs. (1) and (2), Additional file 1: S11-S13, driven by a single locus underpinning coevolution, and not assumed to follow an arbitrarily chosen function of time as in the majority of the population genetics literature. We assume that there is no mutation between alleles at the coevolving loci. Note also that we assume implicitly that host and parasite do undergo recombination in their genome, so that allele frequencies at neutral loci are not linked to those at the coevolving loci.

The second crucial point is the definition of the host and parasite time scales of evolution as determined by the generation times and the population mutation rates of the antagonistic species [18]. If viral, bacterial or fungal parasites often have higher mutation rates than their hosts, their effective population size may not always be larger than that of the hosts and at the onset of an epidemics. The reference population size  $N_{ref}$  at the onset of an epidemics is important because 1) it sets up the initial available diversity, and 2) it defines the time scale for genetic drift in host and parasite and the timing of new neutral mutations occurring with rate  $\theta$ . In our population genetics setting, time is scaled in units of  $N_{ref}$  generations, whereas the host-parasite model specified in Eqs. (1) and (2) runs on arbitrary continuous time reflecting calendar time (in weeks, months or years). If calendar time is equivalent for both species, the scaled time based on  $N_{ref}$  defines the changes occurring in the observed polymorphism over time. We exemplify the difference in time scale and its influence on polymorphism data by a simplified bottleneck model. Two populations with different initial population sizes  $N_{ref}$  experience a size change of the same magnitude and for the same number of generations

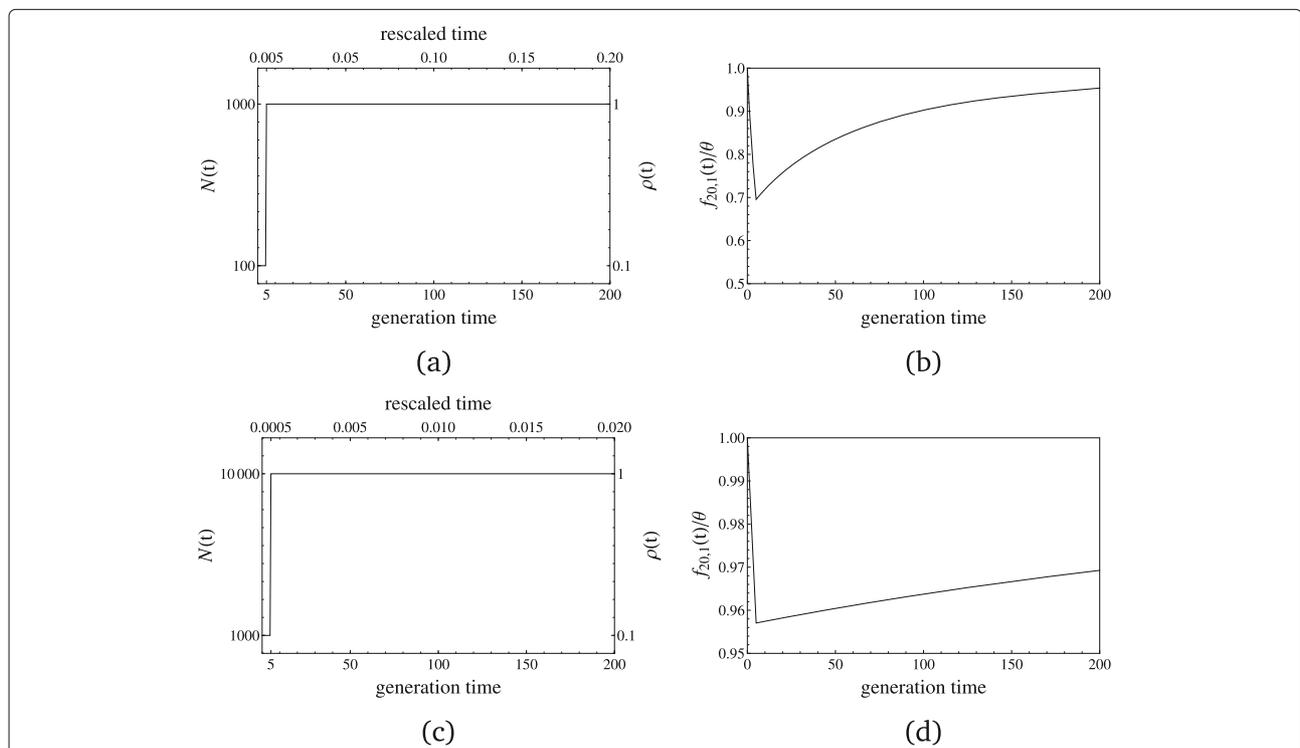
on calendar time scale (bottom x-axes) but for different rescaled time with respect to  $N_{\text{ref}}$  (top x-axes of Fig. 1a and c). Consequently, a size change of the same magnitude but based on two different initial population sizes  $N_{\text{ref}}$  affects the SFS similarly regarding the course of time but with very different strength and detectability (y-axes of Fig. 1b and d). Note that we use the absolute number of singletons scaled by  $\theta$  in Fig. 1 for illustrative purposes, since the difference among the two different population sizes is most pronounced. The difference can only be observed for low-frequency derived variants and eventually vanishes for intermediate to high-frequency derived allelic classes.

The third important feature of our model is the underlying assumption regarding genetic drift. The theoretical result for our SFS computations (Equation 33 in [17], and Additional file 1: SI4) is based on a continuous time approximation of the Wright-Fisher model assuming descendants picking their parents at random from non-overlapping discrete generations. However, in our model, there is an overlap of generations in the host and parasite populations to allow the transmission of disease. A higher overlap of generations occurs for lower values of the

death rate  $d$ . In order to apply the theoretical results, we therefore focus specifically on scenarios where the death rate  $d$  is close to the birth rate  $b = 1$ , so that most of the population is replaced within a 'generation' as defined by  $1/d$ . In addition, we build a simulation method that incorporates the overlap of generations in hosts and pathogens during the drawing of the next generation's allele frequencies.

**Effect of the various parameters on the dynamical system**

We are only interested in situations where at least one host and a parasite genotype survive and both populations coexist. Therefore, we derive first the criteria for the disease to spread in the population via the reproduction ratios (Additional file 1: SI2). Then, we scan the parameter space of our epidemiological model to determine the behavior of the coevolutionary dynamics and the speed of cycling. We thus eliminate situations where the disease spreads but the host and parasite populations finally collapse and get extinct or rise jointly in size. The behavior of the allele frequency cycling is determined by the state of the fixed point (computed in Additional file 1: SI3 for  $A = 2$ ). The cycling can be stable (regular cycles as fluctuating

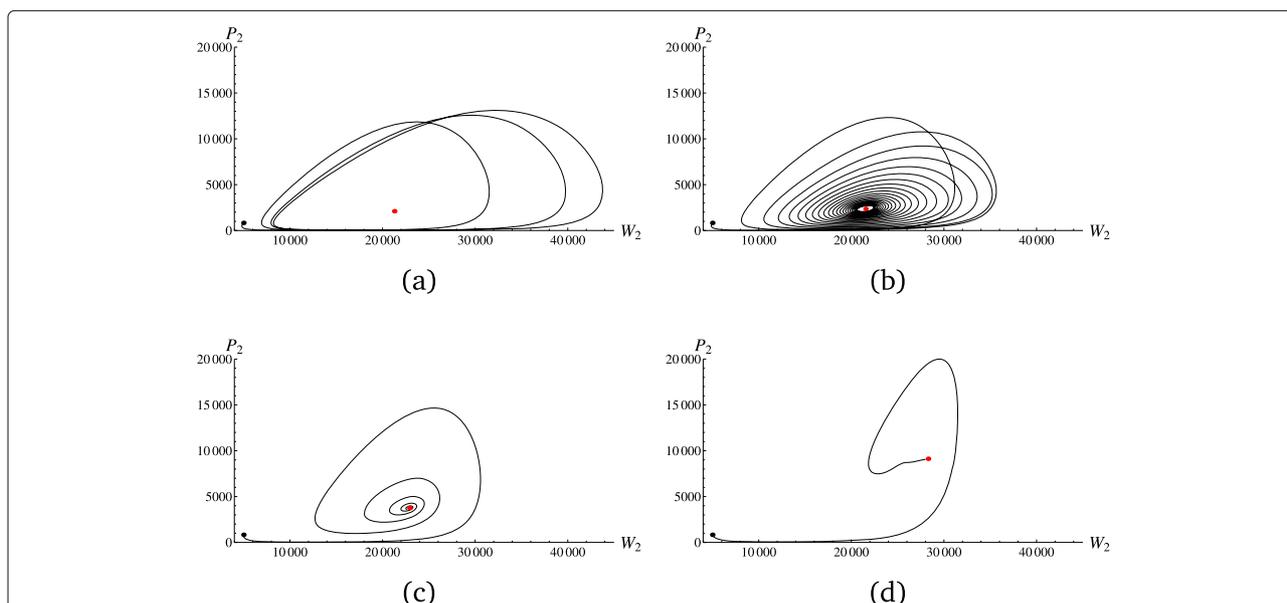


**Fig. 1** A simple bottleneck model with an initial population size of 1000 (a) and 10000 (c) is considered, where the population size drops instantaneously to one tenth of its original size at time zero for 5 generations before recovering again to its original size. Original time in generations is opposed to time scaled in units of the initial population size on the x-axes and the population size changes are given in absolute and relative numbers on the y-axes. The absolute number of singletons scaled by  $\theta$ ,  $f_{20,1}(t)/\theta$ , are illustrated for the respective cases in (b) and (d). The bottleneck is more apparent in (b) than in (d), where the changes of  $f_{20,1}(t)/\theta$  stay within a five percent margin. This is due to slower rescaled time in the first scenario (a), where the population size drop affects polymorphisms prolongedly

selection) or damp off to the fixed polymorphic state. The stability behavior of the hosts and the parasites, whose frequencies are not explicitly given by Eqs. (1) and (2), cannot be explicitly determined by means of a Jacobian matrix and only be determined by numerically solving these equations.

For the stability analysis, the initial conditions were chosen so that  $N^W$  is close to 10,000 and the infected individuals make up 20% of the healthy ones ( $H_1 = H_2 = 4150, I_{11} = I_{12} = I_{21} = I_{22} = 415$ ). The birth rates  $b$  were fixed to one and  $c_H = c_P = 0.05$  for the MA model and  $c_{H_1} = c_{P_2} = 0.05, c_{H_2} = c_{P_1} = 0$  for the GFG model (see [19, 20] for comparable costs). The remaining parameters are given in the color-coded figures of Additional file 1: SI5. We summarize the results of the stability analysis as follows. An increasing difference between the birth and the death rates results in 1) wider parameter ranges of the mortality  $\delta$  and the disease transmission rate  $\beta$  for which cycles may occur, and 2) also increases the number of cycles over a given time interval. While the number of cycles generally increases with increasing values of  $\delta, \beta$  affects the number of cycles most distinctly for  $s = 1$ , for which smaller values of  $\beta$  lead to a reduced number of cycles. The speed of cycling for MA and GFG models is equivalent, if costs are set to zero. Thus, it may be unrealistic to aim to infer the model itself (MA or GFG) based on polymorphism data. In the following, we focus

on the GFG model and present according results for the MA model in the supplement. As illustrated in Fig. 2, besides the difference between  $b(1 - c_H)$  and  $d$ , the selection coefficient  $s$  is the crucial parameter that determines the number of cycles per unit of time. A limit cycle is observed for  $s = 1$ , a case defined as castrating parasite. The cycling appears faster and with quicker damping off with smaller values of  $s$  [2]. Note that applying the parameter values from Fig. 2 to the MA model ( $d = 0.9$  and using the MA specific costs of  $c_H = c_P = 0.05$ ) results in a loss of all parasite alleles for any value of  $s$ , while for a death rate of  $d = 0.6$  a limit cycle occurs for  $s = 1$ , cycling towards the fixed points for  $s = 0.9$  and  $s = 0.6$  but not for  $s = 0.3$ . When  $d = 0.3$ , a limit cycle does occur for  $s = 1$  and a cycling towards the fixed point only for  $s = 0.9$ . The occurrence and speed of cycling is therefore not only determined by the coevolutionary ( $s, c_H, c_P, \alpha$ ) and epidemiological parameters ( $\beta, \delta$ ), but also by the ecological characteristics of the species and the environment controlling the birth  $b$  and death  $d$  rates. An example of a MA model with a death rate  $d = 0.9$  and slow cycles that will be studied alongside the GFG example of Fig. 2 is given in Additional file 1: SI6. Note that changing the birth rate to values other than one and employing various initial allele frequencies of healthy and infected individuals give results that correspond to the ones presented here.



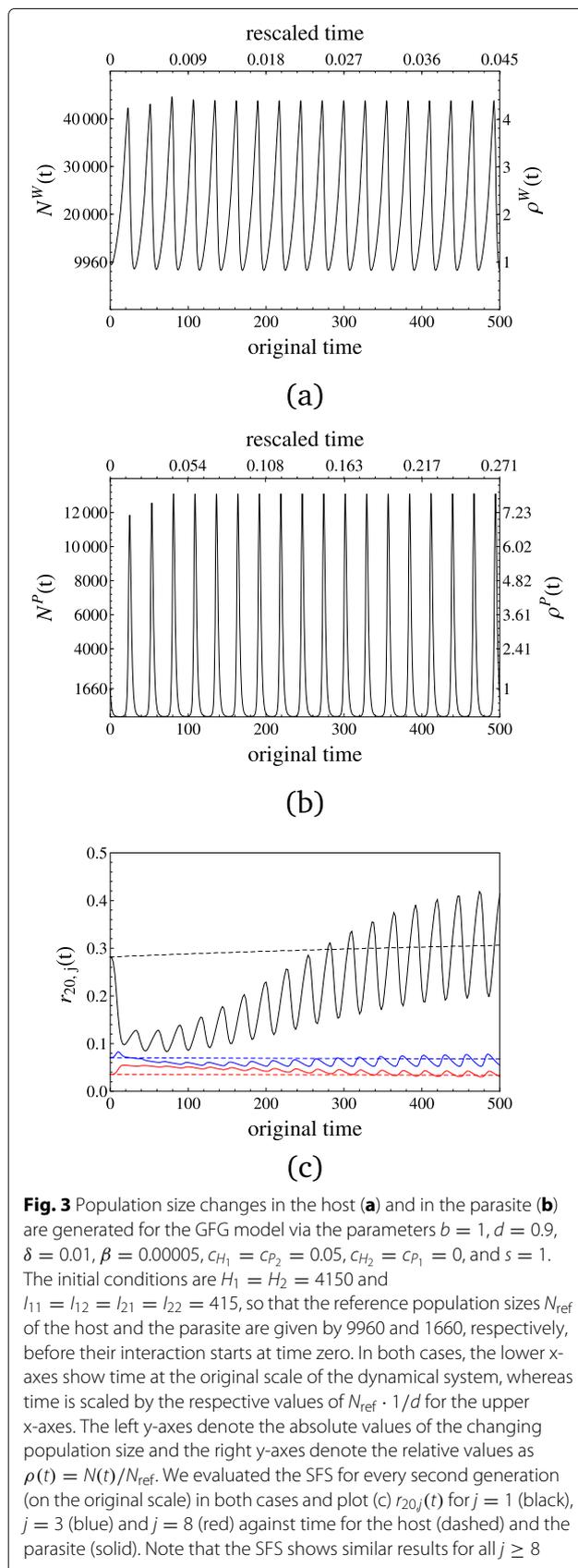
**Fig. 2** Parasite and host alleles of genotype two are plotted against each other for the GFG model over time by numerically solving (1) and (2) for the following parameter values (being equivalent for both genotypes):  $b = 1, d = 0.9, \delta = 0.01, \beta = 0.00005, c_{H_1} = c_{P_2} = 0.05$  and  $c_{H_2} = c_{P_1} = 0$ . The initial conditions are  $H_1 = H_2 = 4150$  and  $I_{11} = I_{12} = I_{21} = I_{22} = 415$ . The selection coefficients  $s$  are given by **a** 1, **b** 0.9, **c** 0.6 and **d** 0.3. The parametric plots are shown for **a** 100, **b** 500, **c** 120 and **d** 96 time steps, which are the minimum amounts of time to complete one orbit of the final limit circle (**a**), come close to the fixed point (**b**), (**c**), or to reach the fixed point (**d**). The initial and fixed points are colored in black and red, respectively

### Analyzing the polymorphism patterns of hosts and parasites

#### One parasite per host and equal generation times

We provide detailed results for a few examples of GFG and MA models to highlight the key features regarding changes in the host and parasite SFS. Mutation rate and genetic drift occur on different time scales for hosts and parasites. The population size changes in the parasite occur on a relatively slower time scale compared to the host (Fig. 3). Moreover, the amplitude of parasite population size fluctuations is much more pronounced than in the host, the number of infected individuals, i.e. parasites, fluctuating between about twenty and 13,000 individuals. The relatively weak and fast fluctuations of the host cannot be observed in the SFS except for a slight increase in the number of singletons over time. In contrast, the strong and relatively slow changes in the parasite population size are clearly reflected by the entire SFS. The number of singletons decreases first due to the initial decline in the population size before tending to increase over time. For intermediate to high-frequency derived allelic classes, an initial increase in allele frequencies is followed by a decrease over time. A similar result for the MA model is given in Additional file 1: SI7. We also observed that signatures of coevolutionary cycles in the host and parasite SFS depend mostly on the speed of their fluctuations in terms of population size scaled generation times, whereas the magnitude of the population size changes has a small influence on the SFS (rather apparent in Fig. 2 than in Fig. 3). We also evaluated one of the slowest cycling examples for a death rate of  $d = 0.3$  shown in the first panel of Figure SI5.2.2 in Additional file 1: SI5 to illustrate that despite the difference in time scale in hosts and parasites, cycles are also barely detectable (even in the singleton class) in the parasite SFS (Additional file 1: SI8).

To detect changes in the SFS in host and parasite, it is also important to assess if enough genetic diversity can be observed over time. We find that the absolute number of polymorphisms strongly decreases over the considered time interval. For example in the GFG model (Fig. 3) the number of segregating sites in the parasite decreases to about five percent and for the MA model (Additional file 1: SI7) even down to about one percent of their initial values. We also contrast two scenarios: 1) an initial total host population size of 100,000 and 2% of initial disease prevalence, and 2) an initial population size of 10,000 with 50% initial prevalence (for GFG and MA models in Additional file 1: SI9). We observe that the initial prevalence defining the parasite population size plays a crucial role. Scenario two shows stronger fluctuations in the relative SFS (exemplified by the number of parasite singletons in the Figures of Additional file 1: SI9) and a more drastic decrease in the absolute number



**Fig. 3** Population size changes in the host **(a)** and in the parasite **(b)** are generated for the GFG model via the parameters  $b = 1, d = 0.9, \delta = 0.01, \beta = 0.00005, c_{H1} = c_{P2} = 0.05, c_{H2} = c_{P1} = 0,$  and  $s = 1$ . The initial conditions are  $H_1 = H_2 = 4150$  and  $I_{11} = I_{12} = I_{21} = I_{22} = 415$ , so that the reference population sizes  $N_{ref}$  of the host and the parasite are given by 9960 and 1660, respectively, before their interaction starts at time zero. In both cases, the lower x-axes show time at the original scale of the dynamical system, whereas time is scaled by the respective values of  $N_{ref} \cdot 1/d$  for the upper x-axes. The left y-axes denote the absolute values of the changing population size and the right y-axes denote the relative values as  $\rho(t) = N(t)/N_{ref}$ . We evaluated the SFS for every second generation (on the original scale) in both cases and plot (c)  $r_{20,j}(t)$  for  $j = 1$  (black),  $j = 3$  (blue) and  $j = 8$  (red) against time for the host (dashed) and the parasite (solid). Note that the SFS shows similar results for all  $j \geq 8$

of segregating sites over time than under scenario one. This shows that larger fluctuations in the relative SFS over time, which are in principle detectable in time sample polymorphism data, go along with a stronger decrease in the total amount of observed polymorphisms (due to more drastic population bottlenecks). As a guideline, we provide an estimate of the possibility to detect changes in the SFS based on a sufficient number of segregating sites (the numerical minima and maxima in Figure SI9.1 of Additional file 1: SI9 yield Table 1). Cycles are more likely observed in parasites with large genome mutation rates such as fungi or protozoans in contrast to bacteria (Table 1).

When comparing our computationally advantageous approach based on the Wright-Fisher model with our stochastic simulations, we find that polymorphism signatures as exemplarily measured by  $\Pi_{20}(t)$  agree in general for host and parasite over time (Fig. 4). The parasite sample shows less polymorphism in the simulations with overlap while for the host this difference is negligible. The net effect of generation overlap under strong population bottlenecks with more dominant decline than expansion phases is an even stronger decrease of the effective population size and thus the amount of polymorphism, as seen in the parasite. This is due to less newborns contributing fewer novel mutations in the model with overlap and the different sampling schemes of newborn and overlapping individuals. Whenever population size decreases, new offspring individuals are present in smaller proportions than overlapping ones, whereas the reverse occurs when population size increases. The fraction of new offspring follows a sampling with replacement as for the Wright-Fisher model, while the overlap fraction is drawn without replacement. The difference of these two sampling schemes becomes apparent during phases of small population sizes. During decline (expansion) phases, the increased fraction of overlapping (newborn) individuals leads to stronger (lesser) deviations from the Wright-Fisher expectations. Consequently, as the parasite population is experiencing a drastic population decrease over time and several cycles, the amount of diversity differs between both approaches in contrast to the host population experiencing a slight increase in size over time.

**Multiple parasites per host and polycyclic diseases**

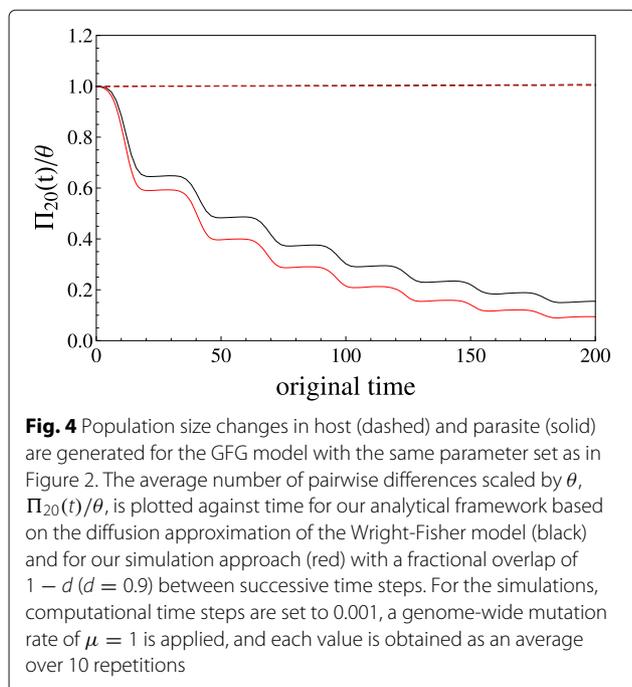
We extend our predictions for two classic deviations from our model. First, multiple or even a large number of parasites, as denoted by  $F$ , often infect a single host. We assume here the simple case of several strains of the same species and of the same allele at the interacting loci that simultaneously infect the host. The parasite strains differ at the neutral loci along the genomes but do not interact with each other for the infection, multiplication and transmission processes. Considering this effect on the SFS leads to an increased scaled mutation rate  $F\theta$  thereby increasing the number of segregating sites that can be detected by full-genome sequencing. Concurrently, relative time is sped up by  $F$  due to the larger initial (reference) population size of the parasite. Therefore, two opposite effects are expected when increasing the number of parasites per host, an increase in the amount of polymorphism comes at the cost of a reduced amount of detectable cycles. Second, host and parasite generation times may differ from one another with parasites often exhibiting smaller generation times especially for virus, bacteria or fungi. We assume here the simple case of a pathogen strain of a polycyclic disease, which undergo several consecutive independent generations within one host individual before disease transmission. The parasite generations act additively to define the amount of damage to the host summarized by the parameters  $\delta$  and  $s$ . We define  $E$  parasite generations per host generation so that the relative time for the parasite is slowed down by  $E$ . This rescaling is expected to enhance the detectability of coevolutionary cycles using the parasite SFS. We investigate the joint impact of multiple parasites per host and polycyclic diseases in Additional file 1: SI10.

We compute the SFS over time for the GFG above (Fig. 3) with nine different combinations of values for  $E$  and  $F$  (see Additional file 1: SI10). To evaluate fluctuations in the polymorphism pattern over time, we compute the relative rate of change of  $\Pi_{20}(t)$  at various equidistantly distributed sampling points over time. These points are chosen to be fixed in time and independent of parasite generation time, so that the various cases are equivalently clocked on the original time scale of the dynamical system. As illustrated in Additional file 1: SI10 more sampling

**Table 1** [minimum, maximum] of the number of segregating sites in full genome sequences of parasites following the GFG model and depending on the initial prevalence and population size

Initial prevalence, $N_{ref}$ parasites	Fungi/Trypanosoma/Nematodes, $\mu = 1$	Viruses, $\mu = 0.1$	Bacteria, $\mu = 0.001$
2%, 2000	[8232, 14191]	[823, 1419]	[8; 14]
20%, 1660	[560; 11779]	[56; 1178]	[1; 12]
50%, 3332	[48; 23642]	[5; 2364]	[0; 24]

The minima are determined by the results in Additional file SI9 and the maxima are the initial numbers of segregating sites. The per genome mutation rate,  $\mu$ , is an approximation based on genome length and per site mutation rate from different typical estimates [21]



points are needed to capture the cycling for multiple parasites per host ( $F$ ), whereas less samples are needed for polycyclic diseases. We therefore show here that the number of time samples necessary to recover the number of cycles can be determined knowing the biology of host and parasite.

**Discussion**

We develop here a model to analyze the evolution of neutral genome-wide polymorphism of coevolving host and parasite populations. Variations in polymorphism reflect the co-demographic history driven by eco-evolutionary feedbacks between 1) the frequency changes in host resistance and parasite infectivity over time due to frequency-dependent selection, and 2) the ecological changes in host and parasite population sizes. While antagonistic or synergistic coevolution is a process driven by natural selection, our approach is the first to provide a description of the consequences of coevolutionary dynamics on neutral genome-wide polymorphism.

We demonstrate that using time series sampling data, i.e. population samples of hosts and parasites at different time points, it is possible to track the existence and speed of eco-evolutionary cycles using polymorphism data. Our main result states that cycles of coevolution are detectable in the parasite and barely in the host population. This is due to a fundamental difference in the time scale of neutral evolution between interacting species crucially depending on the initial population size

at the onset of epidemics (i.e. the start of the coevolutionary history). Furthermore, the parasite population size fluctuations are more pronounced than that of the host. This is a universal characteristics of epidemiological (and Lotka-Volterra) models [14] matching well the observed experimental patterns [10]. If the parasite evolution time scale is adequately slow, it is thus easier to observe strong fluctuations in population size in the SFS than weaker ones. Note that pathogens with infection rates strongly determined by environmental conditions (such as plant pathogens) do show weaker eco-evolutionary feedbacks [22] and population size fluctuations than in our model. The time scale of neutral evolution is also determined by the parasite generation time and number of parasites per infected host. We study here one coevolutionary run starting by the introduction of a parasite population into a larger host population and generating a dynamics over several hundreds of generations. If new infectivity or resistance alleles appear by mutation, the epidemics and the cycling behavior are affected, and our model should be reset to evaluate a new run. The time scale that we investigate is therefore intermediate between the classic expectations of long coevolution and its signatures at interacting loci [6, 9] and the short-term epidemiological dynamics (with susceptible hosts and one parasite type) [23].

Polymorphism data can be used to detect coevolutionary cycles, if such cycles run sufficiently long at adequately low speed. We indeed predict that long term occurrence of cycles should be searched for in parasites that strongly decrease the host fitness due to high disease severity  $s$  (parasite effect on fecundity). For low to moderate disease transmission and smaller values of  $s$  the internal polymorphic equilibrium point is a stable attractor, meaning that cycles damp off quickly towards a polymorphic equilibrium at which population size and allele frequencies are fixed (Fig. 2). For high values of the disease transmission rate  $\beta$  and parasite virulence  $\delta$  (effect of parasite on mortality), the internal polymorphic equilibrium point is an unstable point [2], and a monomorphic equilibrium is reached with a fixed population size [12, 14]. Cycles should be slow enough to be observable in polymorphism data, and our results challenge the classic assumption that coevolutionary cycles are too fast for being observed in the SFS. Interestingly, the speed of cycling depends mainly on two ecological parameters, i.e. the birth rate  $b$  and the death rate  $d$  of the host, and to a lesser extent on the coevolutionary parameters ( $s$ , costs of resistance and virulence).

Eco-evolutionary cycles occur and are observable in polymorphism data, when the effect of the parasite on host fecundity  $s$  is sufficiently strong. We predict that our results are applicable to many host-parasite systems with castrating parasites, whose transmission is associated with

host death (algae-rotifer [24], bacteria-phage [25] and *Daphnia magna*-bacterium [26]). For plant pathogens, cycles may be less observable because the disease severity can range from low to very high (or even castrating, [27]), but often depends on abiotic factors [28]. Most epidemiological studies have focused on the evolution of virulence (effect of parasite on host mortality) and disease transmission within the short duration of an epidemic. However, to use polymorphism data for the study of eco-evolutionary dynamics, parasite virulence is not an essential parameter to be measured or estimated, as it is more useful to quantify the difference between the host's birth and natural death rates. Another practical reason for favoring hosts with high death rates is that our SFS computations are based on the Wright-Fisher model assuming non-overlapping generations [17]. Note that in epidemiological models [2, 12, 14, 22] overlapping generations in the host are a necessary assumption, since a disease can only be transmitted among living hosts. We wrote a simulation code that explicitly accounts for less newborns contributing fewer mutations while generations overlap (and compared to the Wright-Fisher model) to evaluate the SFS. The simulations show that our Wright-Fisher approximation is robust with respect to overlapping generations for hosts with high death rates.

We only consider neutral sites because arbitrary demographic changes can be used in the analytic solution for the SFS as needed to cope with the complex demographics arising from our dynamical system. The frequency spectrum for sites under selection can only be computed for piecewise changes in the population size [29]. This approach is not readily applicable for such complex demographics because their discretization would be computationally cumbersome.

Time sampling is crucial for capturing cycles. These can be observed in the polymorphism data for several hundreds of parasite generations, if the genome mutation rate and the effective population size are sufficiently large (see Table 1) and if SNPs are sampled at appropriate time points (see Additional file 1: SI10). Combining these results with those of the stability analysis (see Additional file 1: SI5), the number of time series samples needed to capture all cycles decreases with smaller disease severity  $s$  as the dynamics tends to stabilize and exhibits shorter cycling periods. In the case of castrating parasites, sampling every 20 host generations should cover most cycles. The less aggressive the parasite, the more time samples are needed. When the generation overlap in the host is higher,  $d = 0.3$ , cycles can be barely detected in host and parasite. Using such time sampling data, one can first estimate changes in host and parasite population sizes based on variation of allele frequencies between time points [30, 31]. As the second step, we can detect loci driven

by selection [32, 33] in the genomes of the interacting species.

To test our predictions, time samples can be readily obtained in experimental coevolution set-ups [10, 24, 25], whereas this may be more complex for natural populations. Nevertheless, samples from the past can be obtained for crustaceans (*Daphnia*, [26]) from dormant stages deposited in sediments, and for plant species from seeds in the soil (possibly using ancient DNA recovery techniques).

Caution should be used when analyzing such data in the light of our results as these are based on a simple coevolutionary scenario with a single pair of loci driving the eco-evolutionary dynamics. In many hosts with available data, the genetic architecture of resistance can, however, be explained by epistasis between several loci [34–36] or few genes with major effect in plants [5, 7, 37]. Finally note that we assume implicitly that recombination occurs in host and parasite genomes so that neutral loci evolve independently from the coevolutionary loci. This condition may not be fulfilled in some host or parasite species and the SFS can be biased due to linkage disequilibrium.

## Conclusions

We demonstrated that eco-evolutionary cycles occur and are observable in polymorphism data. We predict that parasites undergoing several generations per host generation but producing small amount of pathogen propagules per host should be the species exhibiting most clearly the signature of co-demographic dynamics in polymorphism data. Our results pave the way to use time sample genomic data of hosts and parasites from wild or experimental populations, to analyze, infer and take into account the co-demographic history of the antagonistic species and scan for genes under coevolution with greater accuracy. More generally, our results of tracking the changes in SFS over time can be extended to other biological systems exhibiting cycling population sizes at ecological time scales, such as simple prey-predator or complex trophic (species) network interactions.

## Methods

### Modeling a single coevolving locus

We have a haploid one locus model with  $A$  alleles in the host and in the parasite. The infection matrix is given by  $\Lambda = (\alpha_{ij})$  with  $1 \leq i, j \leq A$ . The entries  $\alpha_{ij}$  give the probability that once encountered hosts of genotype  $i$  are infected by parasites of genotype  $j$ . Examples of simple infection matrices for two alleles are given in Table 2.

In analogy to [38] the changes of host and parasite allele frequencies over time are determined by the following coupled differential equations:

**Table 2** Infection matrices for four coevolution models

matching-allele	inverse matching-allele	gene-for-gene	inverse gene-for-gene
$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$

The infection matrices determine the outcome of the interaction between host genotypes (rows) and parasite genotypes (columns). To keep the illustration simple, the rates  $\alpha_{ij}$  are either chosen as one for infection or as zero for resistance. Matching-allele and gene-for-gene models are shown as well as their inverse versions

$$\frac{dH_i}{dt} = H_i \left[ b_i(1 - c_{H_i}) - d_i - \sum_{j=1}^A \alpha_{ij} \beta_{ij} (1 - c_{P_j}) \sum_{k=1}^A I_{kj} \right] + b_i(1 - c_{H_i}) \sum_{j=1}^A (1 - s_{ij}) I_{ij}, \quad (1)$$

$$\frac{dI_{ij}}{dt} = I_{ij}(-d_i - \delta_{ij}) + H_i \left[ \alpha_{ij} \beta_{ij} (1 - c_{P_j}) \sum_{k=1}^A I_{kj} \right]. \quad (2)$$

In Eqs. (1) and (2),  $H_i$  is the number of healthy (*i.e.* non-infected) individuals of genotype  $i$ , and  $I_{ij}$  denotes the number of host genotype  $i$  infected by parasite genotype  $j$ .  $b_i$  and  $d_i$  are the birth and natural death rates (*i.e.* independent of the disease) of host genotype  $i$ , respectively, and  $\delta_{ij}$  is the disease induced death rate caused by pathogen  $j$  on host genotype  $i$  (*i.e.* the effect of pathogen on host mortality [38]).  $\beta_{ij}$  is the disease transmission rate between a parasite of genotype  $j$  and a host of genotype  $i$ .  $c_{H_i}$  and  $c_{P_j}$  are the costs for the hosts and the parasites of carrying genotype  $i$  and  $j$ , respectively.  $s_{ij}$  with  $0 \leq s_{ij} \leq 1$  is the decrease of reproductive fitness of host genotype  $i$  due to an infection of parasite  $j$ , *i.e.* the effect of pathogen on host fecundity. Due to the large number of parameters in our epidemiological model, we investigate a simplified version with two alleles ( $A = 2$ ) in the host and in the parasite except for the evaluation of the host effective population size over time (Additional file 1: SI1) and the reproduction ratios (Additional file 1: SI2), which can be computed for arbitrary  $A$ . Due to the purely deterministic setting, an allele is considered as lost as soon as its count takes a value below one and cannot be introduced by mutation according to our model. We only allow rates and costs to differ among the genotypes in Additional file 1: SI1-2 and for the calculation of the fixed points (Additional file 1: SI3) of the dynamical system (1) and (2). Otherwise, we assume  $b_i = b$ ,  $d_i = d$ ,  $\delta_{ij} = \delta$ ,  $\beta_{ij} = \beta$  and  $s_{ij} = s$ . While we evaluate the fixed points (Additional file 1: SI3) for all four coevolutionary models, only the matching-allele (MA) and the gene-for-gene (GFG) model are investigated in further detail, since the inverse matching-allele (iMA) model is symmetric (and therefore equivalent) to the MA model for  $A = 2$  and the inverse gene-for-gene (iGFG) model is restricted in its behavior compared to the GFG model.

For the MA model symmetric costs are chosen and (except for Additional file 1: SI1-3) assumed to be independent of the host ( $c_{H_i} = c_H$ ) and the parasite ( $c_{P_j} = c_P$ ) genotype. For the GFG model costs are chosen asymmetrically as follows (except for the choice of arbitrary costs in Additional file 1: SI1-3). Since  $H_1$  defends itself successfully against  $P_1$ , and  $P_2$  can infect both host genotypes, only  $H_1$  and  $P_2$  are assumed to carry realistically small costs  $c_{H_1}$  and  $c_{P_2}$  ( $c_{H_2} = c_{P_1} = 0$ , see [4]).

The total number of hosts of genotype  $i$  is given by  $W_i = H_i + \sum_{j=1}^A I_{ij}$ . The number  $P_j$  of parasites of genotype  $j$  is only implicitly given in Eqs. (1) and (2) by the number of infected individuals; assuming one parasite per infected genotype we have  $P_j = \sum_{i=1}^A I_{ij}$ . The change in the effective population size of the host over time  $N^W = \sum_{i=1}^A W_i$  is obtained by numerically solving (1) and (2). The respective differential equation and a condition for obtaining a constant population size are given in Additional file 1: SI1. The effective population size of the parasite is obtained as  $N^P = \sum_{j=1}^A P_j$  (as we assume first one parasite per host).

### Assessing the effect of the coevolving locus on neutral polymorphism patterns

To evaluate the impact of host-parasite interactions on neutrally evolving and genome-wide distributed SNPs over time for interesting cases, we utilize the SFS (Additional file 1: SI4), which is the distribution  $f_{n,i}(t)$  of the number of times  $i$  a mutant allele is observed across sites in a sample of  $n$  DNA sequences at time  $t$ . While  $f_{n,i}(t)$  takes allele counts in absolute numbers, its relative version  $r_{n,i}(t) = f_{n,i}(t) / \sum_{k=1}^{n-1} f_{n,k}(t)$  is normalized by the total number of segregating sites. For this purpose, the deterministic trajectories of time-varying host and parasite population sizes are employed into the analytical result [17] for the neutral SFS. This application requires an appropriate scaling: we define a relative population size function  $\rho(t)$  as the ratio of the population size  $N(t)$  at time  $t$  scaled by the reference population size  $N_{ref}$  at the time of the infection, which initiates the coevolutionary history, *i.e.*  $\rho(t) = N(t)/N_{ref}$ , and denote the population mutation rate as  $\theta = 2N_{ref}\mu$ . We compute the changes in frequencies of neutral alleles generated by the co-demographic scenario in terms of the SFS (also in relative numbers) and the average number of pairwise differences  $\Pi_n(t) = 1/\binom{n}{2} \sum_{k=1}^{n-1} k(n-k)f_{n,k}(t)$  for  $n = 20$  hosts and parasites (Additional file 1: SI4). Our forward

approach is adequately suited for the analysis of time-series data in contrast to the corresponding coalescent result for the neutral allelic spectrum [39, 40], where only a single time point can be immediately evaluated for a given demographic history. Note also that we assume implicitly that host and parasite do undergo recombination in their genome, so that neutral loci are not in linkage disequilibrium with the coevolving loci.

### Modeling overlapping generations

A simulation protocol (code in C available from the authors upon request) was designed that accounts for the effect of overlapping generations in the stochastic sampling of the host and parasite SFS. The differential Eqs. (1) and (2) are discretized by choosing a sufficiently small value for  $\Delta t$  (its infinitesimal version being  $dt$ ) ensuring that the coevolutionary dynamics match the numerical evaluations from *Mathematica*. At every discrete generation  $\tau$ , the current population size  $N_\tau$  consists of  $N_o(\tau) = (1 - d) N(\tau - 1)$  individuals that did not die and therefore overlap and  $N_b = N(\tau) - N_o(\tau)$  newborns (*i.e.* newly infected hosts in case of the parasite). As before, we assume that the SFS of both populations is in equilibrium at start of the infection, *i.e.* the population SFS at time zero is given by  $f_x(0) = \theta/x$  with  $\theta = 2N_{\text{ref}}\mu$  and  $\mu$  being the per generation mutation rate of an entire genome. The SFS of each population is recursively evaluated as follows. For a fixed generation  $\tau_0$ , the  $N(\tau_0)$  alleles are sampled from the pool of size  $N(\tau_0 - 1)$  and according to the allele frequencies at generation  $\tau_0 - 1$ . The newborns and the overlap fraction are, respectively, obtained via sampling with and without replacement. New mutants arising only in newborns and as a single copy at a previously monomorphic site are obtained per generation as a Poisson random variable with mean  $2N_b \mu$  and added to the singleton class of the SFS. Once the population SFS  $f_x(t)$  is computed for a given time interval, its sample version  $f_{n,i}(t)$  is readily obtained via binomial sampling. Note that the number of novel mutations and therefore the amount of polymorphism over time is reduced by definition in this model with overlap compared to the Wright-Fisher model, where all individuals are newborns when descendants replace their parental generation.

### Supplementary information

**Supplementary information** accompanies this paper at <https://doi.org/s12862-019-1556-3>.

**Additional file 1:** In the Supplementary Information, we provide additional methodological details as well as analytical and computational results.

### Abbreviations

SNP: Single nucleotide polymorphism; SFS: Site frequency spectrum;

### Authors' contributions

WS and AT initiated the study; DZ and AT designed the study; DZ performed the analytical work; DZ, MV and SJ conducted the simulations; DZ and AT wrote the manuscript with input from all authors. All authors agreed on the interpretation of data and approved the final version of the manuscript.

### Funding

DZ and SJ were supported by the Deutsche Forschungsgemeinschaft (DFG) grant STE325/14 from the Priority Program SPP1590 "Probabilistic Structures in Evolution" to WS. MV was supported by DFG grant TE809/1 to AT, and SJ by grant TE809/3 from the SPP1819 "Rapid Evolutionary Adaptation" to AT. The DFG played no role in the design of the study, analyses, interpretation of data and writing of the manuscript.

### Availability of data and materials

Mathematica files and our C code for simulations are available upon request from the corresponding authors.

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing Interests

The authors declare no conflict of interest.

### Author details

<sup>1</sup>Section of Population Genetics, Technical University of Munich, Freising, Germany. <sup>2</sup>Department of Mathematics and Statistics, Queen's University, Kingston, Ontario, Canada. <sup>3</sup>Leibniz Institute for Evolution and Biodiversity Science, Berlin, Germany.

Received: 10 April 2019 Accepted: 9 December 2019

Published online: 19 December 2019

### References

- Gandon S, Day T, Metcalf CJE, Grenfell BT. Forecasting epidemiological and evolutionary dynamics of infectious diseases. *Trends Ecol Evol.* 2016;31:776–88.
- Ashby B, Boots M. Multi-mode fluctuating selection in host–parasite coevolution. *Ecol Lett.* 2017;20:357–65.
- Clarke B, O'Donald P. Frequency-dependent selection. *Heredity.* 1964;19:201–6.
- Tellier A, Brown JK. Stability of genetic polymorphism in host–parasite interactions. *Proc R Soc Lond B Biol Sci.* 2007;274:809–17.
- Bergelson J, Kreitman M, Stahl EA, Tian D. Evolutionary dynamics of plant R-genes. *Science.* 2001;292:2281–5.
- Woolhouse ME, Webster JP, Domingo E, Charlesworth B, Levin BR. Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nat Genet.* 2002;32:569–77.
- Stahl EA, Dwyer G, Mauricio R, Kreitman M, Bergelson J. Dynamics of disease resistance polymorphism at the Rpm1 locus of *Arabidopsis*. *Nature.* 1999;400:667–71.
- Brown JK, Tellier A. Plant-parasite coevolution: bridging the gap between genetics and ecology. *Annu Rev Phytopathol.* 2011;49:345–67.
- Tellier A, Moreno-Gómez S, Stephan W. Speed of adaptation and genomic footprints of host–parasite coevolution under arms race and trench warfare dynamics. *Evolution.* 2014;68:2211–24.
- Frickel J, Feulner PG, Karakoc E, Becks L. Population size changes and selection drive patterns of parallel evolution in a host–virus system. *Nat Commun.* 2018;9(1):1706.
- May RM, Anderson R. Epidemiology and genetics in the coevolution of parasites and hosts. *Proc R Soc Lond B Biol Sci.* 1983;219:281–313.
- Tellier A, Brown JKM. The Influence of Perenniality and Seed Banks on Polymorphism in Plant-Parasite Interactions. *Am Nat.* 2009;174:769–79.
- Frank SA. Coevolutionary genetics of hosts and parasites with quantitative inheritance. *Evol Ecol.* 1994;8:74–94.
- Gokhale CS, Papkou A, Traulsen A, Schulenburg H. Lotka–Volterra dynamics kills the Red Queen: population size fluctuations and associated

- stochasticity dramatically change host-parasite coevolution. *BMC Evol Biol.* 2013;13:254.
15. Song Y, Gokhale CS, Papkou A, Schulenburg H, Traulsen A. Host-parasite coevolution in populations of constant and variable size. *BMC Evol Biol.* 2015;15:212.
  16. Rabajante JF, Tubay JM, Ito H, Uehara T, Kakishima S, Morita S, Yoshimura J, Ebert D. Host-parasite Red Queen dynamics with phase-locked rare genotypes. *Sci Adv.* 2016;2:1501548.
  17. Živković D, Stephan W. Analytical results on the neutral non-equilibrium allele frequency spectrum based on diffusion theory. *Theor Popul Biol.* 2011;79:184–91.
  18. Gandon S, Michalakis Y. Local adaptation, evolutionary potential and host–parasite coevolution: interactions between migration, mutation, population size and generation time. *J Evol Biol.* 2002;15:451–62.
  19. Thrall PH, Burdon JJ. Evolution of virulence in a plant host–pathogen metapopulation. *Science.* 2003;299:1735–7.
  20. Tian D, Traw MB, Chen JQ, Kreitman M, Bergelson J. Fitness costs of R-gene-mediated resistance in *Arabidopsis thaliana*. *Nature.* 2003;423:74–7.
  21. Lynch M, Ackerman MS, Gout J-F, Long H, Sung W, Thomas WK, Foster PL. Genetic drift, selection and the evolution of the mutation rate. *Nature Rev Genet.* 2016;17:704–14.
  22. Ashby B, Iritani R, Best A, White A, Boots M. Understanding the role of eco-evolutionary feedbacks in host–parasite coevolution. *J Theor Biol.* 2019;464:115–25.
  23. Stadler T, Kouyos R, von Wyl V, Yerly S, Böni J, et al. Estimating the basic reproductive number from viral sequence data. *Mol Biol Evol.* 2012;29:347–357.
  24. Becks L, Ellner SP, Jones LE, Hairston NG. The functional genomics of an eco-evolutionary feedback loop: linking gene expression, trait evolution, and community dynamics. *Ecol Lett.* 2012;15:492–501.
  25. Hall AR, Scanlan PD, Morgan AD, Buckling A. Host–parasite coevolutionary arms races give way to fluctuating selection. *Ecol Lett.* 2011;14:635–42.
  26. Decaestecker E, Gaba S, Raeymaekers JAM, Stoks R, Van Kerckhoven L, Ebert D, De Meester L. Host–parasite ‘Red Queen’ dynamics archived in pond sediment. *Nature.* 2007;450:870–3.
  27. Clay K. Parasitic castration of plants by fungi. *Trends Ecol Evol.* 1991;6:162–6.
  28. Agrios GN. *Plant Pathology*, 5th. New York: Elsevier Academic Press; 2005.
  29. Živković D, Steinrücken M, Song YS, Stephan W. Transition densities and sample frequency spectra of diffusion processes with selection and variable population size. *Genetics.* 2015;200:601–17.
  30. Waples RS. A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics.* 1989;121(2):379–91.
  31. Jorde PE, Ryman N. Unbiased estimator for genetic drift and effective population size. *Genetics.* 2007;177(2):927–35.
  32. Malaspina A-S, Malaspina O, Evans SN, Slatkin M. Estimating allele age and selection coefficient from time-serial data. *Genetics.* 2012;192(2):599–607.
  33. Foll M, Poh Y-P, Renzette N, Ferrer-Admetlla A, Bank C, Shim H, Malaspina A-S, Ewing G, Liu P, Wegmann D, Caffrey DR, Zeldovich KB, Bolon DN, Wang JP, Kowalik TF, Schiffer CA, Finberg RW, Jensen JD. Influenza virus drug resistance: A time-sampled population genetics perspective. *PLoS Genet.* 2014;10:1004185.
  34. Papkou A, Guzella T, Yang W, Koepper S, Pees B, Schalkowski R, Barg M-C, Rosenstiel PC, Teotónio H, Schulenburg H. The genomic basis of Red Queen dynamics during rapid reciprocal host–pathogen coevolution. *Proc Natl Acad Sci.* 2019;116(3):923–8.
  35. Hall MD, Routtu J, Ebert D. Dissecting the genetic architecture of a stepwise infection process. *Mol Ecol.* 2019;28(17):3942–57.
  36. Bento G, Routtu J, Fields PD, Bourgeois Y, Du Pasquier L, Ebert D. The genetic basis of resistance and matching-allele interactions of a host–parasite system: The *Daphnia magna*–*Pasteuria ramosa* model. *PLoS Genet.* 2017;13:1006596.
  37. Stam R, Silva-Arias GA, Tellier A. Subsets of nlr genes show differential signatures of adaptation during colonization of new habitats. *New Phytol.* 2019;224(1):367–79.
  38. Boots M, White A, Best A, Bowers R. How specificity and epidemiology drive the coevolution of static trait diversity in hosts and parasites. *Evolution.* 2014;68:1594–606.
  39. Griffiths RC, Tavaré S. The age of a mutation in a general coalescent tree. *Stoch Model.* 1998;14:273–95.
  40. Živković D, Wiehe T. Second-order moments of segregating sites under variable population size. *Genetics.* 2008;180:341–57.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://www.biomedcentral.com/submissions)

