Technische Universität München
Wissenschaftszentrum Weihenstephan für Ernährung,
Landnutzung und Umwelt
Professur für Populationsgenetik

Understanding and inferring coevolution from host and parasite genomic data

Hanna Julia Märkle

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat)

genehmigten Dissertation.

Vorsitzender:
Prof. Dr. Hanno Schaefer

Prüfende der Dissertation:
1. Prof. Dr. Aurélien Tellier
2. Prof. Mike Boots, Ph.D.
   University of California, Berkeley
3. Prof. Bruce McDonald, Ph.D.
   ETH Zürich

Die Dissertation wurde am 04.07.2019 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 27.09.2019 angenommen.

Table of contents

List of figures

List of tables

Acknowledgments

I'm very grateful to all the people who have supported me in many different ways during my PhD. Without you I would presumably not be where I am now, namely at the stage of writing the acknowledgments section of my thesis. First, I would like to thank Aurélien. Your lectures about host-parasite coevolution and population genetics made me curious about these two very fascinating topics. Thank you very much for all the trust, for all the opportunities you offered to me to mature as a scientist, for sharing your always encouraging passion about science and for your positive motivation when the project did not work the way I wanted it to work. I further would like to thank Hanno for being the chair person of my committee and Prof. Dr. Mike Boots and Prof. Dr. Bruce McDonald for reviewing my thesis.

Cas, having endless discussions with you about host-parasite coevolution which finally ended into writing a review together was a great pleasure and gave me a lot of motivation. Melissa, I will never forget our extensive discussions about the sense and nonsense of some definitions, sharing conference experiences of different kinds and all the cocktail evenings with and without bubbles we had together.

Diala and Thibaut, you have been wonderful office mates. Thibaut, thanks for the many times you made me smile by joining conversations at random points with random stuff, for cheering me up in difficult moments and of course for the long lasting discussion how the word 'Bounty' is pronounced correctly. Diala, you always helped me to give my thoughts the proper structure, when I wanted too much at the same time and I enjoyed a lot the discussions be it on philosophical or scientific questions while walking home after a long day in the office.

Sona, thanks for the hours we spent together wrapping our heads around combinatorics and coding. Although, our ways of thinking are so different we have achieved a lot of output together.

Silke, what would I have done without all your advice and help while dealing with administrative questions and problems and the short break chats we had. I also would like to thank Daniela for the many enjoyable lunch breaks and small talks. Remco, thanks

for always being ready to listen to my questions about molecular details of host-parasite coevolution, for listening to my thoughts on R-gene evolution, for guiding me through my first real conference which was a bit chaotic and for all the enjoyable bar evenings. Saurabh, thank you so much for introducing me to awk and bioinformatics, for patiently answering all my bioinformatics questions at the beginning of my PhD-thesis and for all the chats we had between office doors. Drazen, thanks for keeping my computer and the cluster always running.

Of course I would also like to thank all current and former lab members. Sharing lunch breaks and bar evenings with you made being in the lab a true pleasure. Further, I wish to thank Pavlos and his group for the great research stay I had in Heraklion. I enjoyed the time in your lab a lot. Your guided advice while working with the ms-code greatly improved my understanding of the coalescent and allowed me to grasp the essentials of C-programming.

Last but no least I would like to thank all my great friends and my family. Thank you for always encouraging me to follow my passion for science, for listening patiently to my long-lasting remarks about scientific problems and for always supporting me whenever I needed your support.

Author contributions

Chapter 1 Hanna Märkle wrote the chapter

Chapter 2 Cas Retel and Hanna Märkle wrote the first draft of the manuscript and designed the figures. All authors wrote and revised the final manuscript. The chapter has been published as:
Retel C.; Märkle H.; Becks L.; and Feulner PGD. 2019. Ecological and Evolutionary Processes Shaping Viral Genetic Diversity. Viruses. 2019; 11(3):220. doi: 10.3390/v11030220

Chapter 3 Hanna Märkle wrote the chapter

Chapter 4 Hanna Märkle, Sona John and Aurélien Tellier jointly designed the study, performed the analytical computations and simulations, analysed the results and wrote a first version of the manuscript. Hanna Märkle wrote the presented version of the manuscript.

Chapter 5 Hanna Märkle and Aurélien Tellier designed the study and wrote the manuscript. Hanna Märkle performed the simulations and analysed the results. This chapter has been submitted to PLoS Computational Biology and is currently under review.

Chapter 6 Hanna Märkle, Sona John and Aurélien Tellier designed the study. Hanna Märkle and Sona John implemented the code. Hanna Märkle wrote the presented version of the manuscript

Chapter 7 Hanna Märkle wrote the chapter.

Abstract

Host-parasite coevolution is a biologically relevant process in all types of environments. It describes the process by which hosts and parasites exert reciprocal selective pressure on one another resulting in evolutionary changes (allele frequencies, genotype frequencies, phenotype frequencies) in both species. The resulting coevolutionary dynamics can be placed in a continuum ranging from successive fixation of novel beneficial types (arms-race dynamics) to stable maintenance of several types in both species (trench-warfare dynamics). Understanding host-parasite coevolution has been an active research field for decades due to its relevance for food production and medicine and its suggested role in facilitating the evolution of sex. Much research effort has been put into detecting the underlying genomic basis, understanding the maintenance of genetic and allelic polymorphism at the involved loci and into improving our understanding of the temporal dynamics of host-parasite coevolution.

The increasing availability of host and parasite full genome data offers promising ways to further enhance our knowledge on all of these aspects. However, this requires an understanding of 1) how coevolutionary dynamics interact with and shape the polymorphism patterns at the coevolving loci, 2) how these polymorphism patterns vary over time, 3) how much information about the underlying coevolutionary process is contained in the polymorphism data of the coevolving loci and 4) how this information can be extracted efficiently and in meaningful ways.

The aim of this thesis is to tackle several aspects of these questions. First, we outline the relevant processes shaping genomic diversity at the coevolving genes and at a genome-wide level. In the next step, we test for the suitability of cross-species genome-wide association studies to detect the coevolving loci. Such studies are based on collecting genomic data of infected hosts and their associated parasites strains. Here, we especially test for the effect of two forms of host genotype x parasite genotype (GxG) interactions (gene-for-gene and matching alleles) and two types of coevolutionary dynamics (arms-race and trench-warfare) on the detection of the coevolving loci. We define two indices, the cross-species association index and the cross-species prevalence index and derive their expected distri-

bution for neutral loci. We show that the underlying GxG interactions and the temporal dynamics determine the power for discriminating between coevolving loci from the neutral background.
In the next chapter, we first analyse how the coevolutionary dynamics in combination with genetic drift relate to genetic signatures at the coevolving loci. Therefore, we couple a gene-for-gene coevolution model with coalescent simulations. We show that the polymorphic equilibrium frequencies determine the strength of balancing selection signatures under trench-warfare dynamics. These equilibrium frequencies are affected by several evolutionary costs such as cost of resistance, infection and infectivity. Therefore, we test as a proof-of-principle whether information about these costs can be recovered by jointly analyzing host and parasite polymorphism data at the coevolving loci from repeated experiments. Using an Approximate Bayesian Computation approach, we can show that it is possible to infer these costs if either a) some costs are known or b) the host and parasite population sizes are known. Most strikingly, the host polymorphism data are informative about costs applying to the parasite and vice-versa, underscoring the reciprocal nature of the interaction.
Finally, we outline ideas for a simulator which can be used to simulate simultaneously the host and parasite population sizes and allele frequency changes arising from coevolution, the resulting haplotypes at the coevolving loci and the genome-wide neutral site frequency spectrum. This simulator will likely provide valuable future insights on how eco-evolutionary feedbacks arising from host-parasite coevolution shape the genomic signatures at the coevolving loci. We will reveal also how the coevolving loci can be distinguished from neutral loci when coevolution causes simultaneous changes in population sizes and how different genomic information can be optimally combined in order to infer information about the short-term and long-term coevolutionary dynamics from them.
We conclude by discussing the value and limits of genomic data for understanding host-parasite coevolution, and the means by which the approaches outlined in this thesis can be further combined and used in combination with additional sources of information.

Zusammenfassung

Die Koevolution zwischen Wirt und Parasit ist ein allgegenwärtiger Prozess in sämtlichen Lebensräumen. Unter Wirt-Parasit-Koevolution versteht man einen Prozess, in dem ein Wirt und ein Parasit einen wechselseitigen Selektionsdruck aufeinander ausüben, sowie die sich daraus ergebenden evolutionären Veränderungen (relative Allelhäufigkeiten, Genotyphäufigkeiten und Phänotyphäufigkeiten). Koevolutionäre Dynamiken können in einen Gradienten eingeordnet werden, der von der wiederholten Fixierung von neuen vorteilhaften Typen (Arms-Race dynamics, Wettrüsten) bis hin zum fortlaufenden Erhalt von mehreren Typen in beiden Partnern (Trench-Warfare dynamics, Grabenkrieg) reicht.
Das Verständnis von Wirt-Parasit-Koevolution ist, aufgrund ihrer Relevanz für die landwirtschaftliche Lebensmittelproduktion, die Medizin und ihrer vermuteten Rolle in der Evolution sexueller Reproduktionsysteme, ein Forschungsfeld mit langer Tradition. Dabei wurde vor allem viel daran geforscht, wie die Gene, die der Interaktion zugrunde liegen, identifiziert werden können. Ein weiteres Augenmerk lag auf dem Verständnis des Erhalts von genetischer Diversität und der Diversität von Allelen an den involvierten Genorten, sowie den temporären Dynamiken von Wirt-Parasit-Koevolution.
Die zunehmende Verfügbarkeit von kompletten Wirt- und Parasit-Genomen bietet vielversprechende Möglichkeiten, um das Wissen hinsichtlich all dieser Aspekte zu erweitern. Jedoch wird dazu ein detailiertes Verständnis über folgende Fragestellungen benötigt:
1. Wie interagieren die koevolutionären Dynamiken mit der genetischen Variation an den koevolvierenden Genorten und wie prägen sie diese? 2. Wie verändert sich diese genetische Variation über die Zeit? 3. Wie viel Information wird über die zugrunde liegenden koevolutionären Prozesse in der genetischen Variation gespeichert? 4. Wie kann diese Information effizient und sinnvoll extrahiert werden?
Das Ziel dieser Arbeit ist es, diese Fragen näher zu untersuchen. Zunächst werden dazu relevante Prozesse beschrieben, die die genetische Diversität auf der Ebene der koevolvierenden Gene und der Ebene des Genoms formen. Im nächsten Schritt wird die Eignung von so genannten 'Cross-Species-Genome-Wide-Association'-Studien zur Entdeckung von koevolvierenden Genorten betrachtet. Solche Studien basieren darauf, dass gleich-

zeitig Genomdaten von infizierten Wirten und den Parasitenstämmen, die den jeweiligen Wirt infizieren, gesammelt werden. In dieser Arbeit wird vor allem untersucht, wie sich zwei unterschiedliche Formen von Wirt-Genotyp x Parasit-Genotyp-Interaktionen (GxG), genauer gesagt Gene-for-Gene und Matching-Alleles, und unterschiedliche koevolutionäre Dynamiken (Arms-Race und Trench-Warfare) auf das Auffinden der koevolvierenden Gene auswirken. Dazu werden zwei Indices definiert, der Cross-Species-Association-Index und der Cross-Species-Prevalence-Index und deren erwartete Verteilung für neutrale Genorte hergeleitet. Es kann gezeigt werden, dass die Unterscheidbarkeit zwischen dem koevolvierenden Genort und dem neutralen Hintergrund im Wesentlichen durch die jeweilige GxG-Interaktion und die jeweilige Form der koevolutionären Dynamiken bestimmt wird. Im nächsten Kapitel wird analysiert, wie sich die Kombination von koevolutionären Dynamiken und genetischer Drift auf die genetischen Signaturen an den koevolvierenden Genorten auswirkt. Dazu wird ein Gene-for-Gene-Koevolutionsmodell mit Coalescent-Simulationen verknüpft. Dabei zeigt sich, dass die Ausprägung der Signaturen balancierender Selektion unter Trench-Warfare-Dynamiken durch die relativen Häufigkeiten der einzelnen Allele im polymorphen Gleichgewicht bestimmt wird. Diese relativen Allelhäufigkeiten werden durch mehrere evolutionäre Kosten, wie die Kosten für Resistenz, Infektivität oder Infektion, bestimmt. Deshalb wird getestet, ob Wirt- und Parasit-Genomdaten der koevolvierenden Genorte, die in wiederholten Experimenten erhoben wurden, Aufschlüsse über diese Kosten geben können. Durch die Verwendung einer Approximate-Bayesian-Computation-Methode kann gezeigt werden, dass es möglich ist, Rückschlüsse auf diese Kosten zu ziehen. Dafür müssen entweder einige dieser Kosten oder die Wirt- und Parasitpopulationsgrößen bekannt sein. Es zeigt sich, dass anhand der Daten des Wirtes Rückschlüsse über die Kosten, die den Parasiten beeinflussen, gezogen werden können und umgekehrt. Dies spiegelt die Natur der wechselseitige Interaktion wider.
Schließlich wird die Idee für einen Simulator skizziert, mit dem folgende Merkmale gleichzeitig simuliert werden können: die Veränderungen der Populationsgrößen des Wirtes und des Parasiten, die Veränderungen der relativen Allelhäufigkeiten im Wirt und Parasit, die daraus resultierenden Haplotypen an den koevolvierenden Genorten und das genomweite Site-Frequency-Spectrum. Mit diesem Simulator können in Zukunft wahrscheinlich nützliche Erkenntnisse darüber gewonnen werden, wie sich eco-evolutionäre Feedbacks, die durch Wirt-Parasit-Koevolution entstehen, auf die Genomdaten der koevolvierenden Genorte auswirken. Zudem kann untersucht werden, inwiefern koevolvierende von neutralen Genorte

unterschieden werden können, wenn sich durch Koevolution gleichzeitig die relativen Allelhäufigkeiten und Populationsgrößen verändern. Außerdem kann der Simulator genutzt werden, um aufzuzeigen, wie unterschiedliche genomische Informationen optimal kombiniert werden können, um daraus Rückschlüsse über kurzfristige und langfristige koevolutionäre Dynamiken zu ziehen.
Zum Schluss wird der Nutzen von Genomdaten zum besseren Verständnis von Wirt-Parasit-Koevolution diskutiert, welche Beschränkungen dabei bestehen und wie die in dieser Arbeit vorgestellten Methoden miteinander und mit weiteren Informationsquellen verknüpft werden können.

1
Introduction

Host-parasite coevolution is a highly dynamic and interactive process which not only affects the evolutionary history of the host and the parasite but also has the potential to affect the structure of communities and biodiversity at all levels (Bohannan and Lenski 2000; Koskella and Brockhurst 2014). Coevolutionary interactions are characterised by reciprocal evolutionary changes in response to the selective pressures two species exert on one another (Janzen 1980; Woolhouse *et al.* 2002). Put in a different way coevolutionary interactions are biotic interactions with evolutionary consequences in both interacting partners. Thus, host-parasite coevolution has attracted the considerable attention of ecologists and evolutionary biologists. Due to the potentially devastating effects of disease on food production and human health, host-parasite coevolution is also of central relevance for agriculture and medicine.

A large body of literature has studied and sought to understand the short and long-term consequences of host-parasite coevolutionary interactions such as population size changes, trait evolution (Boots and Haraguchi 1999; Boots *et al.* 2014; Thrall and Burdon 2003) and maintenance of polymorphism (Frank 1992; Leonard 1994; Sasaki 2000). In the short run and thus, on ecologically relevant time scales, severe epidemics can drastically alter the population sizes of the host (Frick *et al.* 2010; Ebert *et al.* 2000; Berger *et al.* 1998). Conversely, levels of host resistance variation have been shown to affect disease levels in local populations (Thrall and Burdon 2000; Laine 2004). In the long run the selective nature of host-parasite coevolution affects the genetic diversity at the loci involved into the coevolutionary interaction. Adaptive evolutionary changes can be very rapid in host-parasite coevolution (Brown 2015; Paterson *et al.* 2010; Obbard *et al.* 2006). Thus, ecological changes (population size changes) and evolutionary changes can feedback on each other in host-parasite coevolution (Frickel *et al.* 2016), resulting in so called eco-evolutionary feedback loops (Kokko and López-Sepulcre 2007; Schoener 2011; Bailey *et al.* 2009; Post and Palkovacs 2009).

The pivotal role of host-parasite interactions is reflected by the magnitude of evolved host defense mechanisms and the equally large diversity of parasite strategies to circumvent or exploit these defense mechanisms. Plants for example have evolved a two-layered immune

system (Jones and Dangl 2006; Chisholm *et al.* 2006; Dodds and Rathjen 2010). On a first level pattern recognition receptors (PRRs) can detect molecular patterns which are characteristic for a particular class of pathogens/microbes, such as flagellin for bacteria. The recognition of these molecular patterns, called PAMPs/MAMPs (pathogen/microbe associated molecular patterns) by PRRs results in PAMP/MAMP-triggered immunity (PTI). This layer of the plant immune system thus shares common features with the innate immune system in vertebrates. However, pathogens have evolved means to circumvent PTI by releasing so called effectors which interfere with the signaling cascade in PTI. The second layer of the plant immune system consist of effector triggered immunity. Here, plant resistance genes (R-genes) either directly or indirectly recognize specific pathogen effectors in distinct ways (Kourelis and van der Hoorn 2018) which results in effector-triggered immunity (ETI).

Genomic and molecular analysis have shown that many plant species have evolved a rich repertoire of resistance genes (R-genes) and similarly parasites have evolved a large arsenal of effector genes. Most of these genomic analyses have been performed by applying population genetic/genomic techniques to genomic data from one of the coevolving partners from a single time point. This techniques have been originally developed for single species in order to a) to detect loci which are associated with particular phenotypes or b) to understand the relative strength of the different evolutionary forces. Based on these methods several previously unidentified new resistance and effectors gene could be uncovered. Genomic analysis have further demonstrated the long-term maintenance of polymorphism at R-genes in different plant species (Caicedo and Schaal 2004; Hoerger *et al.* 2012; Rose *et al.* 2007; Karasov *et al.* 2014; Stahl *et al.* 1999; Koenig *et al.* 2019) and signatures of positive selection at several R-genes and effector genes (Obbard *et al.* 2011; Schweizer *et al.* 2018).

Recently new methods have been developed which take simultaneously information from the host and the parasite into account (Wang *et al.* 2018; MacPherson *et al.* 2018; Nuismer *et al.* 2017; Bartha *et al.* 2013) (see also chapter 2 for more information) in order to identify genes under coevolution. Thus, these methods account explicitly for the reciprocal nature of coevolutionary interactions. MacPherson *et al.* (2018) have further shown that ignoring genetic structure of the coevolving partner can yield biased effect sizes in genome-wide association studies.

The method by Bartha *et al.* (2013) relies on obtaining genomic data from infected hosts and the parasite strains which infect them. Based on this data they perform a genome-to-

genome association study. So far, nobody has tested how the ability to detect loci under coevolution by means of such an approach depends on temporal dynamics of host-parasite coevolution and the involved host genotype x parasite genotype interactions. Therefore, this question will be examined in further detail in chapter 4 of this thesis.
However for disease management, it is not only important to identify loci under coevolution, but also to understand the processes which have given rise to the observed polymorphism patterns (Gandon *et al.* 2016). There are expectations from theory how particular host-parasite coevolutionary dynamics link to the polymorphism patterns, so called signatures, at the coevolving loci. However, these dynamics can be due to different processes and can also interact with other evolutionary forces such as mutation and genetic drift. Thus the aim of chapter 2 is to outline the ecological and evolutionary processes which are potentially relevant. Although within this chapter, these processes are specifically described and illustrated using examples from host-virus coevolution they do not exclusively apply to host-virus coevolution. Most of them are also relevant for any other host-parasite system although their relative effect sizes will vary depending on the biology of the species. In chapter 5 we will investigate how coevolutionary dynamics in combination with genetic drift link to genomic signatures at the coevolving genes. A more detailed information about the underlying theory and models can be found in chapter 3. In chapter 5 we will further show as a proof-of-principle that information about the past coevolutionary history, namely several fitness costs, can be recovered from genomic data. Therefore, we use Approximate Bayesian Computation (ABC) which is a commonly used inference method in population genetics and which will be introduced in more detail in chapter 3. In chapter 6 an idea for a simulator is outlined which can simultaneously keep track of host-parasite coevolutionary dynamics in the short run and the long run and how these dynamics affect polymorphism patterns at a genome-wide level and at the level of the coevolving genes in both species. The core novelty of this simulator is to keep track of host population size changes, parasite population sizes changes, haplotypes at the coevolving loci and the genome-wide site frequency spectrum over time. Thus, this simulator will allow for investigating 1) the temporal changes in genomic diversity in the host and the parasite in more detail and for investigating 2) how the ability to detect coevolving loci changes over time. So far it is unknown how host-parasite coevolutionary dynamics resulting in simultaneous changes of populations sizes and allele frequencies shape the signatures at the coevolving genes. Further, this simulator can be also used to address the question how time-sampled data can be used efficiently to gain a better understanding

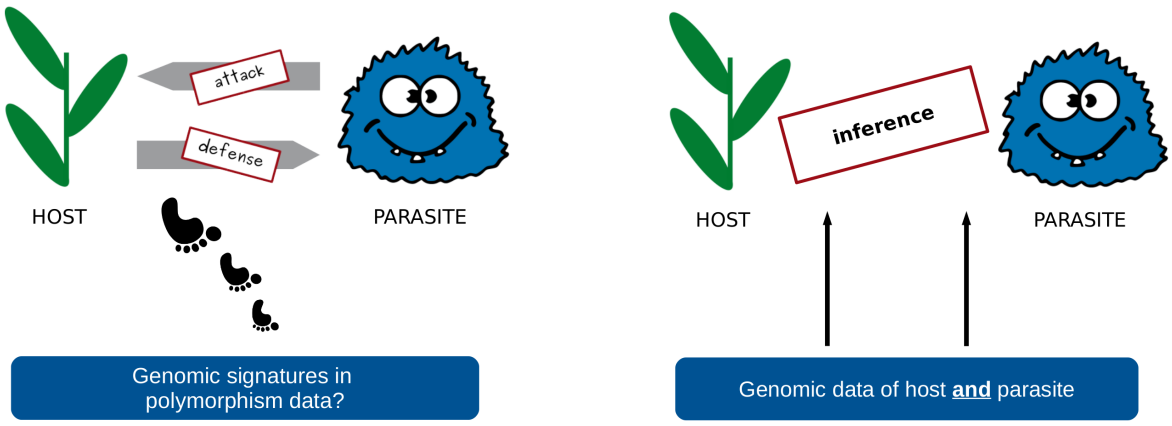


Fig. 1.1: Schematic of the two main topics of this thesis. First, we aim to understand how host-parasite coevolution interacts with and shapes host and parasite polymorphism data. Second, given that we have sampled host and parasite polymorphism data, which information about the coevolutionary interaction can we recover from these data and by means of which methods?

of past coevolutionary dynamics and to devise optimal sampling schemes.

Therefore, the two main topics of these thesis can be summarized as follows (see Fig. 1.1): 1) How do coevolutionary interactions shape the polymorphism data of the interacting partners? 2) Which information about the coevolutionary interaction can be recovered by applying inference methods to full genome data of the species over time?

2
Ecological and evolutionary processes shaping viral genetic diversity

This chapter has been published as:
Retel, C.*, Märkle, H.*, Becks, L.‡, and Feulner PGD.‡2019. Ecological and Evolutionary Processes Shaping Viral Genetic Diversity.
Viruses, 11(3): 220. doi: 10.3390/v11030220
* authors contributed equally
‡ authors contributed equally

2.1 Abstract

The contemporary genomic diversity of viruses is a result of the continuous and dynamic interaction of past ecological and evolutionary processes. Thus, genome sequences of viruses can be a valuable source of information about these processes. In this review, we first describe the relevant processes shaping viral genomic variation, with a focus on the role of host–virus coevolution and its potential to give rise to eco-evolutionary feedback loops. We further give a brief overview of available methodology designed to extract information about these processes from genomic data. Short generation times and small genomes make viruses ideal model systems to study the joint effect of complex coevolutionary and eco-evolutionary interactions on genetic evolution. This complexity, together with the diverse array of lifetime and reproductive strategies in viruses ask for extensions of existing inference methods, for example by integrating multiple information sources. Such integration can broaden the applicability of genetic inference methods and thus further improve our understanding of the role viruses play in biological communities.

Keywords: genetic diversity; viral population genetics; host-virus coevolution; eco-evolutionary feedback

2.2 Introduction

Viruses are ubiquitous and diverse (Koonin *et al.* 2006; Suttle 2005), adapt rapidly (Martiny *et al.* 2014), and often engage in intimate relationships with their host (Clokie *et al.* 2011). Delimitation and discovery of new viruses has been strongly eased by the advent of new sequencing technologies, and genetic barcoding is nowadays a standard technique in virology (Radford *et al.* 2012; Shi *et al.* 2016; Simmonds 2018). The contemporary genome sequence and genetic variation within a given virus strain is shaped by the ongoing interaction of dynamic ecological and evolutionary forces. Thus, besides species delimitation, sequence data can provide valuable information about the evolutionary history of viruses and have been used to track transmission patterns (Gire *et al.* 2014), to date the emergence of new viruses (Faria *et al.* 2014; Saxenhofer *et al.* 2017) and host shifts (Longdon *et al.* 2014), and to identify genes that are under selection (Foll *et al.* 2014). However, extracting such information requires an understanding of all other factors which can (interactively) affect genetic change and create, maintain, and/or deplete genetic variation. Our aim is to review the relevant evolutionary and ecological processes that affect viral genetic diversity and outline how different processes interact and temporally vary (or not). In this review we focus on antagonistic coevolutionary and eco-evolutionary feedback interactions between host and virus species, without restricting ourselves to a particular group of viruses. The influence of abiotic factors such as temperature, CO_2 concentration, and UV radiation on viruses and the interaction with the host has been reviewed elsewhere (Danovaro *et al.* 2011; Horas *et al.* 2018). We will furthermore give a brief overview of existing methodology designed to infer aspects of evolutionary history based on genetic data and describe some recently developed tools for the simultaneous analysis of host and virus genetic data.

2.3 Viral population genetics

Variation at the sequence level of viruses is created by mutations which range from changes at single bases (single-nucleotide polymorphism, SNP) up to rearrangements of the genome

architecture. Variant forms of a genetic sequence are called **alleles**, and the position at which they occur is referred to as **locus**. Viral mutation rate estimates range from 10^{-8} to 10^{-6} changes per base pair per cellular infection (generation) for DNA viruses and from 10^{-6} to 10^{-4} for RNA viruses (Drake *et al.* 1998; Duffy *et al.* 2008; Sanjuán *et al.* 2010). These rates are high compared to microbes such as *E. coli* and *S. saccharomyces* (both $< 10^{-9}$ (Jee *et al.* 2016; Lang and Murray 2008)). The total mutation supply in a population per generation not only depends on the mutation rate per sequence per generation (μ) but also on the effective population size N_e (see below) of the focal population. The **population mutation rate** $\theta = 4N_e\mu$ captures this interplay and represents the expected number of accumulated differences between a pair of randomly chosen sequences in a population (Hein *et al.* 2004). The ultimate fate of a mutation, i.e., fixation, loss, or maintenance at intermediate frequency—and by extension the total amount of genomic variation in a population, is determined by the interaction between genetic drift, selection, recombination, and migration. In this review, we pay less attention to viral recombination (Meier-Kolthoff *et al.* 2018; Pérez-Losada *et al.* 2015) and the concepts of spatial structure and migration (Berngruber *et al.* 2015; Declerck *et al.* 2013) but focus on drift and selection because they are especially relevant for microbial viruses.

Genetic drift describes the process of stochastic changes in allele frequencies due to random sampling of offspring from the parental generation. Generally, the strength of genetic drift depends mainly on the effective population size, with smaller populations experiencing stronger drift. The **effective population size** (N_e) corresponds to the size of an idealized population (satisfying the assumptions of the so-called Wright–Fisher model of population genetics: constant population size, non-overlapping generations, diploid individuals, equal sex ratio, no selection, no recombination, small variance in offspring numbers, and random mating among individuals) which experiences the same amount of stochastic genetic change as the population analyzed (Charlesworth 2009). The ratio of N_e to census population size N is affected by factors such as the mode of reproduction and temporal variation in population size (Ellegren and Galtier 2016). Viruses possess several characteristics that reduce the N_e-to-N ratio. Population sizes of viruses infecting several globally important phytoplankton species can fluctuate by orders of magnitude within a season (Brussaard *et al.* 2005; Castberg *et al.* 2001; Johannessen *et al.* 2017; Yoshida *et al.* 2008). Viruses typically also have skewed offspring distributions, with a lot of virions never successfully reproducing and a few contributing disproportionately large amounts of genetic material to the next generation (Sanjuán 2018). For example, the RNA virus

vesicular stomatitis virus and the dsDNA virus chlorovirus PBCV-1 can produce burst sizes ranging from 50 to 8000 and 100 to 350 particles per replication event, respectively (Zhu *et al.* 2009; Van Etten *et al.* 1983). Both fluctuating population size and skewed offspring distributions increase the relative importance of drift. Hence, viruses experience stronger drift than other organisms with similar census population sizes.

Besides genetic drift, the type and strength of selection influences the probability and rate by which alleles increase or decrease in frequency in a population. The term **fitness** captures the number of offspring any individual possessing a particular genotype is expected to contribute to the next generation. **Positive selection** describes selection on constantly beneficial alleles (Weigand and Leese 2018), which are expected to increase in frequency across generations until they reach fixation, meaning that every individual in the population possesses the allele and variation at the locus is lost. Opposed to positive selection, **purifying selection** captures the process of selection against deleterious mutations. **Balancing selection** summarizes any form of selection which maintains variation in the population (i.e., more than one allele at a locus) (Charlesworth 2006).

Alleles under positive selection can decrease in frequency due to genetic drift. Therefore, there is always a chance that they are lost from a population, especially when their frequency is low (Fig. 2.1). In a Wright–Fisher type population, the probability of fixation of a beneficial mutation present in a single individual, provided that it has a weak selective advantage s and population size is large, is approximately $2sN_e/N$ (Kimura 1962; Otto and Whitlock 1997). Skewed offspring distributions as seen in many viruses increase the probability that beneficial mutations reach fixation (Der *et al.* 2011; Irwin *et al.* 2016) and decrease the expected time this takes (Eldon and Wakeley 2009). For these reasons, we expect frequency changes of alleles under selection in virus populations to be comparatively rapid. Alleles that have no or small fitness effects which are physically associated with (linked to) a positively selected mutation are expected to change simultaneously in frequency with the mutation under selection, a process referred to as **genetic hitchhiking** (Barton 1998; Kaplan *et al.* 1989; Maynard Smith and Haigh 1974). Being "dragged along" with alleles under positive selection increases the variance in temporal frequency changes of genetic hitchhikers compared to those of neutrally evolving loci (Kosheleva and Desai 2013; Schiffels *et al.* 2011). Genetic diversity and frequency changes at neutral sites can be further affected by purifying selection against deleterious mutations, a process termed **background selection** (Charlesworth *et al.* 1993). Background selection decreases genetic variation at linked sites (Ewing and Jensen 2016; Good *et al.* 2014) and

has the potential to slow down or even impede the expected frequency increase of linked adaptive alleles (Charlesworth 2013). Associations between physically linked sites can be broken up by **recombination**. In the absence of recombination, two beneficial mutations that arise independently in different lineages will never be combined in a single genotype (Felsenstein 1974; Muller 1964). Rather, there will be competition among the offspring of these two lineages: a process termed **clonal interference** (Gerrish and Lenski 1998; Park and Krug 2007). Clonal interference can result in hampered frequency increases and the extinction of one of the two lineages (Lang *et al.* 2011; Schiffels *et al.* 2011).
The combination of high mutation rate and large population size leads to a high supply of de novo mutations in virus populations. This increases the likelihood that multiple mutations with varying effects on fitness segregate simultaneously in virus populations (Desai *et al.* 2007; Neher 2013), and both clonal interference and genetic hitchhiking are then likely to occur (Cvijović *et al.* 2018). However, as viral genomes are in most cases rather small (but see recently discovered giant viruses, e.g., (Colson *et al.* 2017)) and densely packed with protein-coding regions, most mutations are likely to be highly deleterious. This suggests a prominent role for background selection in most viruses (Irwin *et al.* 2016) and decreases the potential for multiple mutations to segregate simultaneously (but see (Renzette *et al.* 2015), where genetic variation was found at hundreds of sites in human cytomegalovirus populations within host individuals). There is currently not enough empirical data available to make general statements about the occurrence and relative strength of interference between deleterious and beneficial mutations in viruses (Renzette *et al.* 2016). In summary, the observed variation at the genomic level results from a complex interplay between mutation supply, drift, and selection, and their individual contributions depend on the biology of the particular virus.

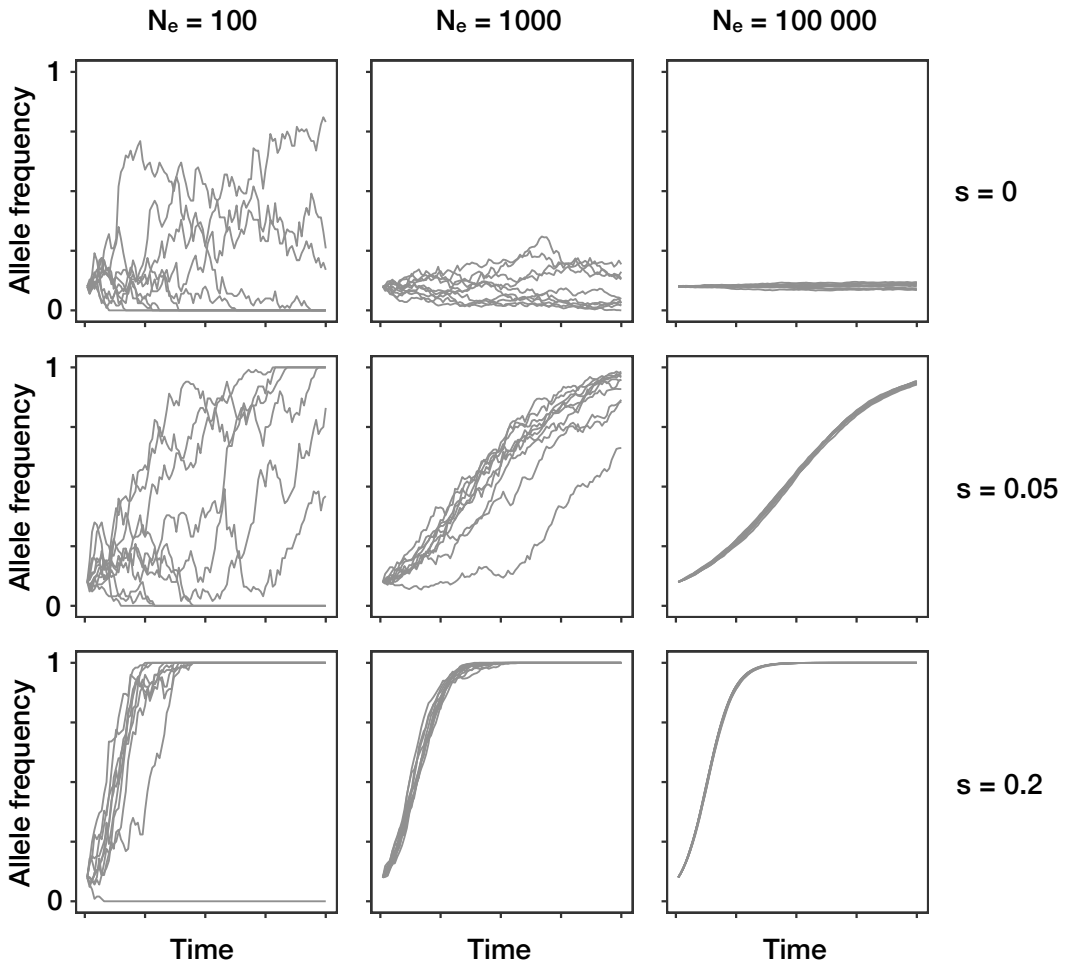


Fig. 2.1: The combined effects of drift and selection on genetic change over time (x-axis). Shown are simulated allele frequencies (y-axis) of a focal allele for different combinations of effective population size N_e (columns) and selection coefficient s (rows). A positive selection coefficient ($s > 0$) indicates a selective advantage of the focal allele compared to the other allele, if $s = 0$ both alleles are neutral and thus, allele frequency changes are only due to genetic drift. Each panel shows the results of ten independent replicates with an initial frequency of 0.1 for the focal allele. Note that when effective population size is small, even positively selected alleles sometimes go extinct due to drift (left column, middle and bottom row). Absolute frequencies k of the allele in generation $t + 1$ were obtained by randomly drawing from a binomial distribution with $P(X = k) = \binom{N_e}{k} p_{t+1}^k (1 - p_{t+1})^{n-k}$ and $p_{t+1} = \frac{p_t(1+s)}{\bar{w}}$, where w denotes the average fitness of the population and p_{t+1} denotes the expected frequency of the focal allele without drift in the next generation $t + 1$.

2.4 Host-virus coevolution

Because viruses depend on their hosts for replication, their genome evolution is also strongly influenced by their host (Bozick and Real 2015; Simmonds *et al.* 2018). Similarly, hosts are under constant pressure to reduce the detrimental fitness effects of viruses (Bernatchez and Landry 2003; Rodriguez-Valera *et al.* 2009). This reciprocal evolutionary interplay is called coevolution when adaptation of one species changes selection on the interacting partner and vice versa (Burmeister *et al.* 2016; Janzen 1980; Nuismer *et al.* 2010). The dynamics and genetic consequences of host–virus coevolution in particular and of antagonistic species interactions in general have been enigmatic research topics dedicated to understanding the maintenance of genetic diversity (Clarke 1979), the evolution of sex (Hamilton *et al.* 1990), patterns of local adaptation (Thompson 2005), and the speed of evolution (Paterson *et al.* 2010).

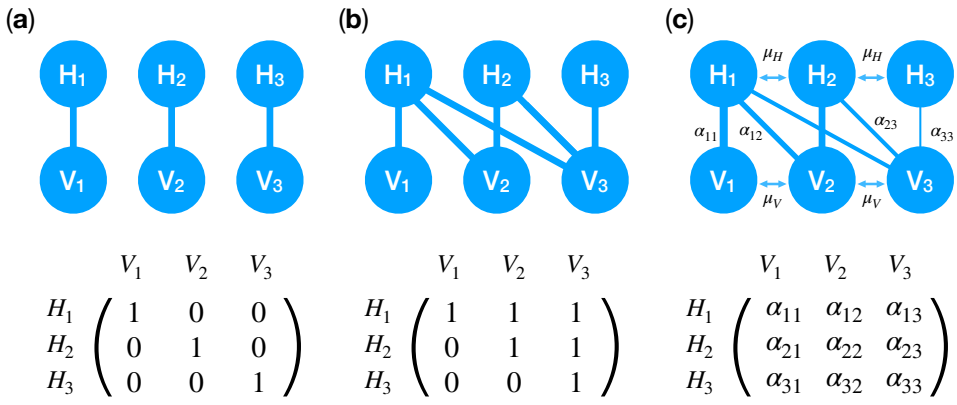
An important determinant of host–virus coevolutionary dynamics is the number of genotypes per population and how these interact with antagonist genotypes, captured in an **infection matrix** (Agrawal and Lively 2002). The two opposite ends of the continuum of possible infection matrices are, on the one hand, **matching alleles**, where every virus genotype can only infect one host genotype (Ebert 1994; Spanakis and Horne 1987) (Fig. 2.2a), and, on the other, **gene-for-gene** interactions (Fig. 2.2b), where virus genotypes infect a broad range of host genotypes (Flor 1956). There is a range of possible infection matrices in between these two extremes (Engelstädter 2015; Forde *et al.* 2008) (Fig. 2.2c). Viruses infecting bacterial hosts completely span this range (Dennehy 2012; Koskella and Meaden 2013). Importantly, adaptation of one or both interacting partners can lead to changes in the underlying infection matrix and in the resulting coevolutionary dynamics (Frickel *et al.* 2016; Poullain *et al.* 2008; Scanlan *et al.* 2011).

The coevolutionary dynamics of genes underlying the molecular interaction between host and virus are separated into **arms race** or **fluctuating selection dynamics** (also called trench-warfare or Red Queen dynamics) (Woolhouse *et al.* 2002) although these two categories rather describe the two end points of a continuum (Agrawal and Lively 2002). In arms race dynamics, coevolution is driven by the reciprocal consecutive increases in frequency of novel genotypes which provide an evolutionary advantage (e.g., the ability to target a novel outer cell membrane protein (Meyer *et al.* 2012)), ultimately resulting in fixation. On the genomic level, these frequency increases result in a reduction of genetic diversity at linked loci (Barton 1998; Maynard Smith and Haigh 1974), also referred to

as a selective sweep. Arms race dynamics are characterized by directional changes in phenotype distributions, such as a monotonic increase in viral infectivity (Gómez *et al.* 2015; Marston *et al.* 2012; Schiffels *et al.* 2011). Known examples of arms races in (semi-)natural species interactions are *Flavobacterium* phage coevolution in fish farms (Laanto *et al.* 2017) and *Drosophila* resistance to sigma virus (Wilfert and Jiggins 2012).

Fluctuating selection dynamics, a form of balancing selection, occur when the fitness of multiple functional genotypes in both species negatively depends on their frequency in the population. This results in fluctuations of their relative frequencies and thus maintenance of several genotypes in both interacting species. The underlying mechanism, here outlined for two host and two virus genotypes, is as follows: selection in the virus favors efficient exploitation of the most common host genotype A. This confers a selective advantage to a rarer host genotype B, which is thus expected to increase in frequency over the course of generations. Once host genotype B becomes the most common one, selection in the virus no longer favors exploitation of host genotype A, granting another virus capable of infecting host genotype B the selective advantage. Such frequency-dependent selection patterns can lead to perpetual oscillations of functional allele frequencies in both coevolving populations, with allele frequencies in the virus population following those in the host (Rosenzweig and MacArthur 1963). Fluctuating selection dynamics are characterized by oscillating phenotype distributions (Betts *et al.* 2014; Papkou *et al.* 2018) and are expected to result in higher levels of genomic variation at functional and associated loci than expected under neutrality. Negative frequency-dependent host–parasite interactions have been found in a taxonomically wide range of antagonistically coevolving systems, such as flax and its fungal pathogen flax rust (Thrall *et al.* 2012), *Daphnia magna* and its bacterial endoparasite *Pasteuria ramosa* (Decaestecker *et al.* 2007), and *Pseudomonas* with naturally associated lytic viruses (Gómez and Buckling 2011).

In summary, coevolutionary dynamics at the functional loci can be classified based on their effect on the number and frequencies of functional genotypes in both antagonistic species over time. The occurring type of dynamics depends on various factors, such as the number of functional genes, the underlying infection matrix of different genotypes, and the ecology of the interacting species. The effect on the genomic diversity at the interacting functional loci is determined by the interaction of the particular dynamics with de novo mutations, standing genetic variation, recombination, and the amount of genetic drift.



$$
\begin{array}{c} \\ H_1 \\ H_2 \\ H_3 \end{array}
\begin{array}{c} \begin{array}{ccc} V_1 & V_2 & V_3 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array}
\qquad
\begin{array}{c} \\ H_1 \\ H_2 \\ H_3 \end{array}
\begin{array}{c} \begin{array}{ccc} V_1 & V_2 & V_3 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \end{array}
\qquad
\begin{array}{c} \\ H_1 \\ H_2 \\ H_3 \end{array}
\begin{array}{c} \begin{array}{ccc} V_1 & V_2 & V_3 \end{array} \\ \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{pmatrix} \end{array}
$$

Fig. 2.2: Overview of possible types of coevolutionary interactions between host genotypes (H_i) and virus genotypes (V_j). The top row shows a graphical representation of potential interactions between different host and parasite genotypes. Lines indicate that a virus with genotype j can infect a host with genotype i. The corresponding infection matrices are shown in the bottom line. Entries in the infection matrix which are equal to 0 (1) indicate that the host genotype in row i is fully resistant (susceptible) to the virus genotype in column j. (a) In a matching-allele system each virus genotype can successfully infect only one host genotype. (b) In gene-for-gene systems there is one universally infective virus genotype (here V_3) which is able to infect all host genotypes. Most coevolutionary interactions fall onto a continuum between these two extremes and can be captured in a correspondingly parameterized infection matrix as illustrated in (c). α_{ij} reflects the rate of success for virus genotype j to infect host genotype i. Every α_{ij} can take values between 0 and 1. Genotype-altering mutations happen at rate μ_H in the host and μ_P in the parasite.

2.5 Eco-evolutionary feedbacks in viruses

Host–virus systems are likely to be subject to feedbacks between ecological and evolutionary change. Population densities affect encounter rates between antagonistic individuals. Therefore, the strength of antagonistic selection varies in concert with population size. Population sizes, in turn, mediate the strength of drift and supply of de novo mutations (section 2). For these reasons, abundance (ecological) and allele frequency (evolutionary) dynamics often reciprocally influence each other (Kokko and López-Sepulcre 2007) in host–virus interactions (Papkou *et al.* 2016) and interactively determine coevolutionary genetic change (Fig. 2.3) (Becks *et al.* 2012; Frickel *et al.* 2018).

Reciprocal effects between ecological and evolutionary change are especially important to consider when evolutionary and ecological changes occur on similar timescales, i.e., when evolution is rapid (Hairston *et al.* 2005; Messer *et al.* 2016). How often contemporary evolution has a considerable influence on community dynamics is an outstanding question in ecology and evolution (Koch *et al.* 2014; Thompson 2009). Rapid evolution of resistance

has been experimentally shown to change the effects chlorovirus has on the population dynamics of its host (Frickel *et al.* 2016) and to facilitate coexistence with a third species (Frickel *et al.* 2017). Furthermore, rapid host resistance evolution has been demonstrated to alter the effects myovirus has on marine microbial food web structure (Lennon and Martiny 2008). With their short generation times and large population sizes, both viruses and microbes are likely to display rapid adaptive responses, and hence to be involved in eco-evolutionary feedbacks.

Analytical predictions on allele frequency dynamics and equilibrium states change when population sizes are allowed to vary and determine the strength of antagonistic selection (Govaert *et al.* 2019; Luo and Koelle 2013). Mechanistic predictions on reciprocal allele frequency changes can then be done by combining information on phenotypic traits and abundances of both populations (Velzen and Gaedke 2017; van Velzen and Gaedke 2018). Balancing selection—thus, maintenance of higher-than-expected levels of genetic diversity—becomes more likely and can occur even when the infection matrix conforms to a 'true' gene-for-gene system (see Fig. 2.2b), where virus genotypes are equally successful on all host genotypes (Best *et al.* 2017). However, population bottlenecks (drastic reductions in size) increase the probability that stochastic fixation events occur (Gokhale *et al.* 2013). Such events remove functional genotypes and subsequently diminish genomic variation. Even if no genotypes are stochastically lost, the traditionally predicted simple harmonic oscillations of allele frequencies are either replaced by more complex combinations of sinusoidal functions (Song *et al.* 2015), or allele frequencies stabilize but species abundances fluctuate (De Andreazzi *et al.* 2018; MacPherson and Otto 2018) in models including eco-evolutionary feedback effects.

In summary, strong reciprocal fitness effects that cause fluctuations in population size and the potential to adapt rapidly make microbial host–virus interactions likely subject to eco-evolutionary feedback dynamics. To which extent the integration of such feedback is necessary to correctly interpret the genetic signature of coevolution is an unresolved question in evolutionary genetics.

Fig. 2.3: Potential interactions between ecological processes and evolutionary processes in host–virus coevolution and how they interact with and shape genomic variation. In the evolution section, boxes in the top row correspond to genomic variation and those in the second row (bottom) to evolutionary forces governing them. Eco-evolutionary feedback loops take place when coevolutionary selection alters population sizes and/or other ecological processes (third row and below), which in turn alters how the different genomic forces affect genomic variation. The dots at the bottom indicate that the above-mentioned ecological processes by no means constitute an exhaustive list. Features of the host are always presented in green and features of the virus in orange. Selection imposed by abiotic variables is not included in this figure.

2.6 Genomic inference methods

After having outlined the various processes which can affect and interact with viral genomic diversity, we will now give an introductory overview of available inference methods that can be used to extract information about these processes from genomic data. We will start with methods which are traditionally used to analyze genomic data of a single species. Then, we present some recently developed methods which take into account the reciprocal nature of host–virus coevolutionary interactions.

Outlier scans can be used to search for loci that are putatively under selection (Fig. 2.4a). Genomes from a population sample are scanned for loci which show either elevated or decreased levels of genetic diversity and/or linkage disequilibrium compared to the genome-wide average. Deviations from the average are interpreted as evidence for selection having acted (Aguileta *et al.* 2009; Nielsen 2005; Vitti *et al.* 2013; Weigand and Leese 2018). Depending on the type of available data, these scans are often based on summaries of the sequence data, such as the site frequency spectrum (from which statistics such as Tajima's D can be calculated), haplotype distribution, or when multiple populations are compared, e.g., by differentiation measures such as F_{ST}.

The first step in outlier scanning is the establishment of the demographic history of the population based on diversity patterns of putatively neutrally evolving loci. This step is crucial, as various demographic scenarios can produce genomic signatures which are very similar to those of positive selection (recent population expansion after a bottleneck) or of balancing selection (population decline) (Bank *et al.* 2014; Crisci *et al.* 2012; Hoban *et al.* 2016). It is further important to note that not accounting for background selection can result in biased demographic inference, most pronounced when it is at intermediate levels (Bank *et al.* 2014; Ewing and Jensen 2016). The second step involves comparing the diversity per locus to the expected neutral distribution given the established population demography. Loci under positive selection are expected to show lower levels of diversity and higher linkage disequilibrium with neighboring regions. Loci under balancing selection will on the other hand show elevated levels of nucleotide diversity, detection of which can be hard depending on the timescale at which selection has acted (Fijarczyk and Babik 2015).

The power of jointly inferring demography and selection can be increased by sampling genome data or collecting allele frequencies at several time points (Fig. 2.4b; see, e.g., Foll *et al.* (2014)). An overview on existing methods to analyze such time-sampled data,

including the advantages and potential biases, is given in (Bank *et al.* 2014).
If phenotype data of sequenced individuals are available, it is possible to perform a genome-wide association study (GWAS; Fig. 2.4c), which searches for alleles that are statistically associated with an observed phenotype (e.g., Gutierrez *et al.* (2018)). It should be noted that GWAS studies can yield different effect sizes for a single locus if the genetic structure of the coevolving partner is not taken into account (MacPherson *et al.* 2018). The authors of this publication further demonstrated the value of integrating genomic information of both co-evolving partners simultaneously into a "two-species co-GWAS" (MacPherson *et al.* 2018).
Methods jointly integrating genome data from both coevolving partners are very likely to increase our understanding of the genetic basis underlying coevolution, as the genomes of both partners contain pieces of information on their joint coevolutionary history (Fig. 2.4d). Wang *et al.* (2018) proposed such a method, called Analysis with a Two-Organism Mixed Model or ATOMM, which aims to associate the outcome of reciprocal infection experiments (e.g., the level of quantitative resistance) with genetic variants in the host and parasite genome, simultaneously (Wang *et al.* 2018). Their method also accounts for a latent population structure and allows for different types of genetic variants including insertion/deletion polymorphisms. Nuismer, Jenkins, and Dybdahl proposed a framework to identify coevolutionary loci by measuring the spatial covariation of marker frequencies in the host and parasite across several populations (Nuismer *et al.* 2017). They showed that the performance of their method mainly depends on the strength of local adaptation between host and parasite, the number of populations being sampled, and the genomic architecture of the trait.
In summary, we have presented an overview of genetic inference methods. Detecting the genomic basis of coevolution can be achieved by incorporating different sources of information (genotypic information, phenotypic information, information from both interacting partners, or from multiple time points) and combining different sets of methods which are most appropriate for the given system. The overview above also underlines that currently much work is being done on the development and extension of methods that are specifically tailored for the analysis of coevolving systems.

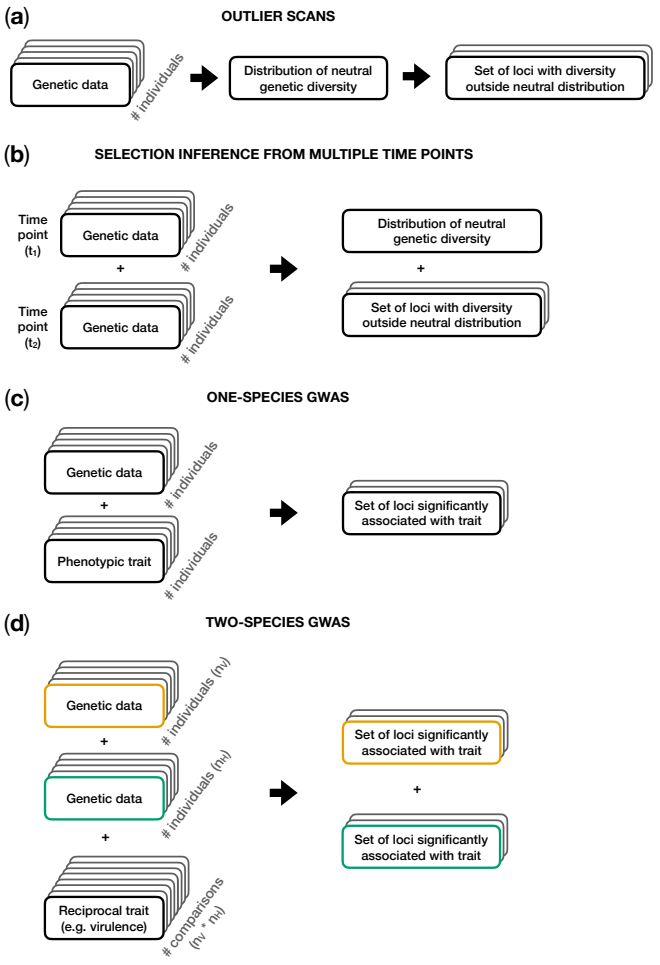


Fig. 2.4: Analysis steps involved in the genetic inference methods outlined in section 2.6. (a) In outlier scans, genetic data are used to obtain an estimate of the demographic history and the distribution of neutral diversity given this demography. Loci which are at the extremes or even outside of this neutral distribution are subsequently identified as putatively under selection. (b) Genetic data from multiple time points allow the calculation of changes in allele frequencies over time. This increases the power to jointly estimate the demography and identify loci under selection. (c) Genome-wide association studies (GWAS) are performed with phenotypic and genetic information from a sample of individuals within a population to detect associations between genetic variants and a certain phenotype (e.g., quantitative virulence). (d) Two-species GWAS integrates genomic information from a sample of n_H host individuals and n_V virus individuals and phenotypic outcome of all $n_H * n_V$ pairwise interactions. Data from the virus are shown in yellow. Data from the host are shown in green.

2.7 Discussion

In this review, we have outlined the evolutionary processes shaping the genomic diversity of viruses and highlighted how they can (individually and interactively) affect genomic diversity (see Fig. 2.3 for an overview). Many analysis tools have been developed by population geneticists to infer the presence and strength of evolutionary forces from genome data of a single species (Fig. 2.4). When applying them, one must be aware of their respective underlying assumptions and to which extent they fit the biology of the virus being studied. Recently developed methods have started to extend beyond a single species analysis framework, enabling the integration of genetic and phenotypic data of coevolving virus and host populations, thus, explicitly taking the reciprocal nature of antagonistic coevolution into account. Such methods provide promising opportunities to identify previously undetected targets of selection, such as resistance genes, virulence factors, or regulatory regions, which will deepen our insights into the molecular basis of host–virus interactions and coevolution.

When viruses are involved in complex eco-evolutionary feedback loops, species abundances, phenotypic trait distributions, and allele frequencies all change continuously and simultaneously (Frickel *et al.* 2016). In such cases, sampling genomic data at several points in time while simultaneously keeping track of phenotype data and population size data enables us to establish links between genetic, phenotypic, and population size changes. Time-sampled genomic data specifically allow for a more precise quantification of the strength of selection (Renzette *et al.* 2013) and offer more powerful means to disentangle the effects of various ecological and evolutionary forces on genome-wide diversity dynamics (Bank *et al.* 2014; Foll *et al.* 2014; Pennings *et al.* 2014). With their diverse array of life-history traits, life-time strategies, often comparatively small genomes, and short generation times, viruses offer a great opportunity to study the dynamics of complex biological systems in real time. Such a time-resolved multifarious view on ecological and evolutionary dynamics will also increase our mechanistic understanding of the role viruses play in natural ecosystems.

The possibility to sample and analyze data from repeated experiments provides further insights into the diversity of the possible paths antagonistic coevolution can take (Desai 2013; Frickel *et al.* 2018) and how the interaction between different kinds of mutations (beneficial, neutral, and deleterious) will shape the resulting eco-evolutionary dynamics (Hinkley *et al.* 2011; Jiang *et al.* 2016; Kryazhimskiy *et al.* 2011). Analyzing replicated

viral genomic data from, e.g., microcosm experiments or different populations with similar environmental properties will allow us to identify conditions under which viral evolution is predictable and will thus aid in understanding and predicting, e.g., disease outbreaks (Russell *et al.* 2012). Several challenges remain to be addressed. First, we are not aware of any genomic inference methods that simultaneously take coevolutionary change and ecological population size changes into account. Second, it is important to increase the discussion on optimal sampling schemes (in terms of replication, temporal sampling density, and specific sampling times) to capture as much relevant information as possible in a time- and cost-effective way. Third, there are limits as to how much genomic data can tell us about such highly complex systems, and these limits should be investigated carefully. For all of this, advancement will crucially depend on ongoing exchange between empiricists and theoreticians from various fields, such as virology, ecology, evolutionary biology, and population genetics.

In summary, we have shown in our review how genomic data of viruses—besides helping to delimitate species—offer a powerful source of information to elucidate past ecological and evolutionary processes, to study the genomic basis of adaptation, to improve our understanding of evolution under species interactions, and to shed light on reciprocal interactions between ecological and evolutionary change. These are exciting times in which more and finer-scaled genetic data is increasingly available, and substantial progress is being made in the development of methods linking such data to theory. Both are vital to increase our understanding on how viruses interact with their hosts, how this shapes genomic diversity of both interacting partners, and how this feeds back into ecological processes.

Author Contributions: CR and HM wrote the first draft of the manuscript. All authors wrote and revised the final manuscript.

Funding: HM received support from grants from the German Research Foundation (DFG), grant numbers TE 809/3-1 and 4-1 (awarded to Aurélien Tellier). LB received support from the German Research Foundation (DFG) grant number BE 4135/3-1, and PGDF from the Swiss National Science Foundation (SNSF), grant number 310030E-160812/1.

Acknowledgments: The authors thank Aurélien Tellier for helpful comments and discussions on early versions of the manuscript, and two anonymous reviewers for constructive feedback.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study, in the decision to publish this review, or in the writing of the manuscript.

3
Extended introduction

After having outlined a) some general motivations to study host parasite coevolution, b) the research questions which will be addressed in this thesis and c) the processes which are potentially relevant in host-parasite coevolution, this chapter is meant to introduce some theoretical background information and modelling approaches in more detail. First, we introduce two types of models which are frequently used to model host-parasite interactions and coevolution. The first type of models, population genetics models, can be used to understand the maintenance of allelic polymorphism at the coevolving genes (Frank 1992; Tellier and Brown 2007b,a; Leonard 1994; Sasaki 2000). The second type of models, SI-models (Anderson and May 1979), has been used to understand the temporal dynamics of epidemics and the evolution of host and parasite traits. Third, we give some more detailed information about population genetics. Here, genetic drift is described in more detail as drift becomes relevant in chapter 5 when linking host-parasite coevolutionary dynamics to genomic signatures. Further, we give some brief introduction into the standard coalescent (Kingman 1982a,b) which is a probabilistic model to describe the ancestral relationship of a sample of individuals. It is a powerful framework to simulate genomic data for a sample of individuals under a particular demographic scenario. Further, the expectations of some frequently used summary statistics such as the site frequency or Tajima's D can be derived from it. And finally, a short overview is given about Approximate Bayesian Computation, which is a widely used inference method in population genetics and on which the inference in chapter 5 is based.

3.1 Modelling host-parasite coevolution

3.1.1 Population genetics model

Coevolutionary changes, namely changes in allele frequencies or phenotypes, can be modeled by a population genetics model. Population genetic models are based on the assumption that selection is the main evolutionary force within a population. Therefore, one core assumption of such models is that the populations sizes are infinite, or extremely large,

so that the effect of genetic drift becomes negligible. In the next paragraphs we will first introduce the general form of a population genetics model for haploid populations (each individual has a single copy of the gene) with discrete generations. Further, we will outline the general model for diploid populations with discrete generations and for a haploid populations with overlapping generations. Based on the model for haploid populations with discrete generations, we will introduce the general form of coevolutionary models when hosts and parasites are haploid. Further, we will describe the features and costs which are frequently incorporated in such models and how they relate to the biology of the two coevolving species. By doing so, we will introduce three models which will be used in this thesis and one well known model of the host-parasite coevolutionary literature, namely the Leonard model (Leonard 1994). A non-exhaustive list of coevolutionary models from the literature and their underlying assumptions can be found in Tab. 3.1.

General form of population genetics models

A population genetics model in its most simple form tracks the allele frequencies at a single locus with two alleles, A and a, in a haploid population with discrete generations. Central to any population genetics models is the calculation of the fitness for each allele/genotype. The term fitness expresses the exptected life-time reproductive success of individuals with a particular genotype. Based on the fitness and the frequency of a particular allele in the current generation g, the allele frequencies in the next generation $g+1$ can be obtained as (Otto and Day 2007, p.64):

$$p_{g+1} = \frac{p_g w_A}{p_g w_A + q_g w_a} = \frac{p_g w_A}{p_g w_A + (1-p_g) w_a} = \frac{p_g w_A}{\bar{w}} \tag{3.1}$$

Here, p_g (q_g) denotes the allele frequency of allele A (a) in generation g. The frequency of allele a can be alternatively expressed as $1-p_g$, since the frequencies of both alleles must sum up to one. w_A (w_a) is the relative fitness of allele A (a) and $\bar{w}$ is the average fitness of the whole population. The average population fitness, $\bar{w}$, can be obtained by weighting the fitness of each allele, w_A and w_a respectively, by its current frequency in the population.

It is straightforward to extend this model to diploid populations (each individual has two copies of each gene). In order to calculate the new frequency of allele A, one first has to compute its current frequency based on the frequencies of the A-homozygotes (individuals

having two copies of the A-allele) and the frequency of heterozygotes (individuals with one copy of the A and one copy of the a allele) and the respective fitness of the A-homozygotes and the heterozygotes. Accordingly, the recurrence equation for a model of natural selection in a diploid population is given by (see Otto and Day (2007), p.71):

$$p_{g+1} = \frac{p_g \left(p_g w_{AA} + (1-p_g) w_{Aa}\right)}{p_g \left(p_g w_{AA} + (1-p_g) w_{Aa}\right) + (1-p_g)\left(p_g w_{Aa} + (1-p_g) w_{aa}\right)}, \tag{3.2}$$

with w_{AA} and w_{aa} being the fitness of the homozygotes and w_{Aa} being the fitness of the heterozygote.
Similarly, the model in equation 3.1 for haploid populations with continuous (overlapping generations) can be written as (Otto and Day 2007, p.67):

$$\frac{dp}{dt} = s\, p(t)\left(1 - p(t)\right), \tag{3.3}$$

where s is the selection coefficient, namely the difference in growth rates of alleles A and a (Otto and Day 2007, p.69). Whether to use a discrete-time or a continuous-time model depends on the reproductive system and the age structure in the population of concern. Discrete-time models are appropriate for species where all individuals reproduce at the same time and die afterwards. Therefore, discrete-time models are for example well suited for annual plants, soil-borne pathogens or pathogens which sporulate at the end of the host season, survive on an intermediate host and than spread at the beginning of the next season again. If on the other hand, hosts and/or parasites are continuously dying and reproducing, continuous-time models are a more appropriate choice. Such populations are characterized by so called overlapping generations, meaning that parents and their offspring live at the same time. Coevolutionary systems which are characterized by overlapping generations are for example algae or bacteria coevolving with phages.

General model form coevolutionary population genetics model for major genes in haploid populations

In this thesis we will for the most part focus on coevolutionary interactions between a haploid host and a haploid parasite with discrete generations. Further, we assume that the coevolutionary interaction is driven by a bi-allelic major locus in the host and a bi-allelic major locus in the parasite. By writing equation 3.1 once for the host and once for

the parasite, the coevolutionary system can be described by the following two recurrence equations:

$$h_{i,g+1} = \frac{h_{i,g} w_{H_i,g}}{\bar{w}_{H,g}} \tag{3.4}$$

$$p_{j,g+1} = \frac{p_{j,g} w_{P_j,g}}{\bar{w}_{P,g}} \tag{3.5}$$

Here, $h_{i,g}$ denotes the frequency of the i-th host allele in generation g and $p_{j,g}$ is the frequency of the j-th parasite allele in generation g. Similarly, $w_{H_i,g}$ ($w_{P_j,g}$) is the fitness of the i-th host (j-th parasite) allele in generation g and $\bar{w}_{H,g}$ ($\bar{w}_{P,g}$) is the average fitness of the host (parasite population). Inherent to the coevolutionary interaction is the idea that the fitness of the various host (parasite) types depends on the outcome of the interaction with the various parasite (host) types and the composition of the parasite (host) population in terms of allele frequencies.

The infection matrix α

The infection matrix A stores the outcome of all host genotype x parasite genotype interactions. It is a $m \cdot n$ matrix with m being the number of host types and n being the number of parasite types. Each entry α_{ij} in this matrix denotes the probability that a parasite of type j can infect a host of type i. For a simple system with two host and parasite types this matrix writes as:

$$\begin{array}{c} \\ H_1 \\ H_2 \end{array} \begin{array}{c} \begin{array}{cc} P_1 & P_2 \end{array} \\ \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} \end{array}. \tag{3.6}$$

Two well studied types of infection matrices are the gene-for-gene and the matching-allele matrix. The gene-for-gene matrix is based on the work by Flor (1956, 1971) on the *Linum marginale–Melampsora lini* system. Gene-for-gene interactions are characterized by a universally infective parasite (a parasite which can infect all hosts) and a generally susceptible host (a host which can be infected by any parasite). Therefore, this infection matrix captures well the molecular interactions underlying effector-triggered immunity (Dodds and Rathjen 2010). One well studied other type is matching-alleles interactions, where each parasite type can only infect a matching host type (Luijckx *et al.* 2013). This

type of infection matrix fits best to systems with self-non-self recognition such as the MHC-complex in vertebrates. Generally, gene-for-gene and matching-alleles infection matrices represent some points in a continuum of infection matrices with varying specificity (Agrawal and Lively 2002; Engelstaedter 2015).

Fitness costs/gains in coevolutionary models

Infection decreases the fitness of a host by an amount s, the cost of infection. For parasites either the fitness of a parasite which can not infect a host, decreases by an amount c compared to parasite which infect successfully (Tellier and Brown 2007b) or the fitness of parasites which infect hosts successfully increases by some amount ζ compared to parasites which fail to infect hosts (Leonard 1994, see infobox below). Further each host (parasite) genotype can be associated with some fitness cost $c_{H_i}(c_{P_j})$. For the host this can be some cost of resistance (Tian *et al.* 2003; Bergelson and Purrington 1996; Karasov *et al.* 2014). Similarly for the parasite this could reflect a cost of infectivity/virulence (Bahri *et al.* 2009; Thrall and Burdon 2003; Montarry *et al.* 2010). Taking the infection matrix and these potential costs into account, we can rewrite equation 3.4 as:

$$h_{i,g+1} = \frac{h_{i,g}(1-c_{H_i})\left(1-\phi_g s\sum_{j=1}^{2}\alpha_{ij}p_{j,g}\right)}{\bar{w}_{H,g}} \tag{3.7}$$

$$p_{j,g+1} = \frac{p_{j,g}(1-c_{P_j})\left(1-c\sum_{i=1}^{2}(1-\alpha_{ij})\,h_{i,g}\right)}{\bar{w}_{P,g}} \tag{3.8}$$

Inherent to these equations are two further assumption. First, hosts and parasites interact in a frequency-dependent manner and thus, disease transmission is frequency-dependent. Second, a proportion ϕ_g of the host population interacts with the parasite population in generation g.

Type of disease transmission

In our model we assume that the disease transmission is frequency-dependent. Thus, the probability whether a host of type i encounters a parasite of type j is only dependent on their respective frequencies h_i and p_j. Such a disease transmission mechanism could for example apply to sexually transmitted diseases (Antonovics 2017). Frequency-dependent

disease transmission is the only disease transmission type which can incorporated in classic population genetics model. However, an other type of disease transmission, namely density-dependent transmission, can be relevant when populations are of finite size. Here, the frequency of encounters between hosts of type i and parasites of type j depends on their respective densities (numbers).

Disease prevalence

Further, we assume that every generation a proportion ϕ_g ($0 \leq \phi_g \leq 1$) of the host population interacts with the parasite population. A disease prevalence of $\phi_g < 1$ could be for example due to parasites with limited dispersal.

The simple coevolution model for a gene-for-gene system

In a gene-for-gene interaction there are two host types, namely, susceptible hosts (*res*) and resistant hosts (*RES*), and two parasite types, namely, infective parasites (*INF*) and non-infective parasites (*ninf*)-parasites. Note that in the case of diploid hosts it is usually assumed that resistance is dominant to susceptibility and non-infectivity is dominant to infectivity. Therefore from a molecular perspective, it would be natural to abbreviate the infective parasite by small letters and the non-infective parasite by capital letters. However, in order to be consistent with other papers from our group (Tellier *et al.* 2014; Verin and Tellier 2018) we use the former notation. Further, note that in the plant pathology literature the infectivity allele is frequently called the virulence allele and the non-infective allele is called the avirulence allele.
The frequency of resistant (susceptible) hosts is denoted by R_g (r_g) and the frequency of non-infective (infective) parasites is denoted by A_g (a_g). Based on experiments it has been shown that resistance can be costly (Karasov *et al.* 2014; Bergelson and Purrington 1996). Therefore, the model for a gene-for-gene interaction incorporates a fitness cost for the resistant host genotype (in the following paragraphs c_H for short). Further, we assume a cost of infectivity for infective (virulent sensu plant pathology literature) parasites (Thrall and Burdon 2003; Bahri *et al.* 2009; Montarry *et al.* 2010) and that the disease prevalence

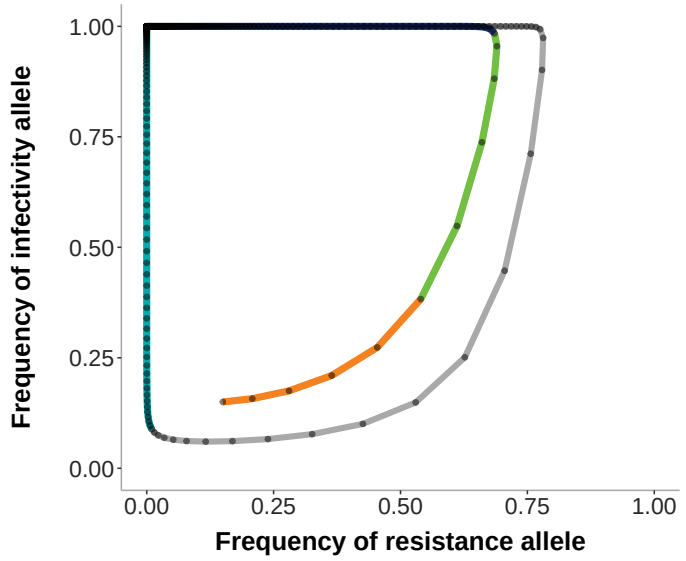


Fig. 3.1: Coevolutionary dynamics in a monocyclic gene-for-gene-model. Each dot represents the frequency of resistance (x-axis) and frequency of infectivity (y-axis) in a single generation g. The dynamics were simulated based on eq. 4.6 for $c_H = 0.05$, $c_P = 0.1$, $s = 0.4$, $c = 1$, $R_0 = a_0 = 0.15$

is $\phi = 1$. Therefore, the model writes as (Tellier and Brown 2007b):

$$R_{g+1} = \frac{R_g(1-c_H)\,(1-sa_g)}{\bar{w}_H} \tag{3.9}$$

$$r_{g+1} = \frac{r_g\,(1-s)}{\bar{w}_H} \tag{3.10}$$

$$A_{g+1} = \frac{A_g((1-c)R_g + r_g)}{\bar{w}_P} \tag{3.11}$$

$$a_{g+1} = \frac{a_g\,(1-c_P)\,[R_g + r_g]}{\bar{w}_P} \tag{3.12}$$

The dynamics of this model (and other bi-allelic coevolution models) can be plotted in a 2-D plane where the frequency of one of the two host alleles is plotted on the x-axis and the frequency of one of the two parasite alleles is plotted on the y-axis. Each point in the resulting plot represents the frequencies of these alleles in a single generation g. By connecting the points of consecutive generations, we obtain the so-called allele frequency trajectory. Fig. 3.1 shows one exemplary trajectory of the model.

The dynamics can be understood as follows. Initially, the proportion of infective parasite and resistant hosts is small. Therefore, resistant hosts have a selective advantage compared to susceptible hosts as the cost of resistance is smaller compared to the cost of infection in

susceptible hosts. Therefore, the frequency of resistant hosts increases (orange part of the trajectory). Due to the increase of resistance in the host population, infectivity becomes advantageous compared to non-infectivity. Infective parasites can infect both host types and this outweighs the fitness loss due to the cost of infectivity. Accordingly, the frequency of the infectivity allele increases almost to fixation (green part of the trajectory). However, due to the very high frequency of infective parasites, the fitness of resistant compared to susceptible hosts decreases. Resistant hosts have to pay now both, the cost of resistance and the cost of infection, in almost all interactions. As the cost of resistance is small in the example shown, the frequency of resistant hosts only marginally decreases every generation (blue part of the trajectory) until almost the whole host population becomes susceptible again. This in turn selects against infective parasites. Therefore, the frequency of non-infective parasites increases again (lightblue part of the trajectory) up to the point the next cycle starts.
In the shown example, the allele frequencies spiral outwards towards the corners of the plane. Each of the four corners of the plane represents a monomorphic equilibrium point. At a monomorphic equilibrium both the host and the parasite population are fixed for either of the two alleles. However, the model also has a fifth equilibrium point where both alleles are maintained. However, as soon as the system is slightly displaced from this polymorphic equilibrium the allele frequencies 'move' away from it and finally get fixed at one of the monomorphic equilibrium. Accordingly, the polymorphic equilibrium point is unstable (Tellier and Brown 2007b) in this model.

Monocylic disease vs polycylic disease

For now we have only considered a so-called monocyclic disease in our model. Monocyclic diseases are characterized by a single infection cycle per host season/generation. However, many diseases are characterized by more than one infection cycle per season. Such diseases are termed polycyclic. The above presented model can be extended to a polycyclic disease with T parasite generations per host generation g. When there are $T = 2$ parasite generations per host generation the coevolutionary interaction can be described by the following three recurrence equations (still assuming discrete host and parasite generations)

(Tellier and Brown 2007b).

$$a_{g,2} = \frac{a_{g,1} \cdot (1 - c_P)}{a_{g,1} \cdot (1 - c_P) + A_{g,1} \cdot r_g} \tag{3.13}$$

$$a_{g+1,1} = \frac{(1 - c_P) \cdot [R_g (A_{g,1} a_{g,2} + a_{g,1}) + r_g a_{g,1}]}{(1 - c_P) \cdot [R_g (A_{g,1} a_{g,2} + a_{g,1}) + r_g a_{g,1}] + r_g A_{g,1}} \tag{3.14}$$

$$R_{g+1} = \frac{R_g \cdot (1 - c_H) [A_{g,1} A_{g,2} + A_{g,1} a_{g,2} (1 - s_2) + a_{g,1} (1 - s_1)]}{R_g \cdot (1 - c_H) [A_{g,1} A_{g,2} + A_{g,1} a_{g,2} (1 - s_2) + a_{g,1} (1 - s_1) + r_g (1 - s_1)]} \tag{3.15}$$

The presented model has the assumption that parasites from the second generation infect the same host as their parent. Accordingly, resistant and susceptible hosts infected with infective parasites in the first parasite generation, stay infected with infective parasites in the second parasite generation. Further, susceptible host which are infected by non-infective parasites in the first generation stay infected by non-infective parasites. When hosts are infected for two consecutive parasite generation they bear a cost of infection, s_1. Resistant host which are attacked by non-infective parasite in the first generation can be attacked by any parasite type in the second parasite generation. Here again non-infective parasites fail to infect, whereas infective parasite can infect successfully. In the latter case resistant hosts loose an amount s_2 of their fitness.

Auto vs alloinfections

This model incorporates an additional assumption, which is of relevance for polycyclic diseases, namely whether a parasite was produced on the host it is currently infecting or on a different host. In the previous model parasites infected the same host as their parents. Such infections are called auto-infections (Mundt 2009). Auto-infections can for example occur when parasites are splashed by water from one plant part to another part. Opposed to auto-infections are allo-infections. Here, parasites infecting a particular host have been produced on a distinct host (Mundt 2009). Taking the possibility of allo-infections at rate ψ into account, the previous model can be rewritten as (Tellier and Brown 2007b):

$$a_{g,2} = \frac{a_{g,1} \cdot (1 - c_P)}{a_{g,1} \cdot (1 - c_P) + A_{g,1} \cdot r_g} \tag{3.16}$$

$$a_{g+1,1} = \frac{(1 - c_P) \cdot [R_g A_{g,1} a_{g,2} + r_g A_{g,1} a_{g,2}(1 - \psi) + a_{g,1}(\psi + a_{g,2}(1 - \psi))]}{\begin{pmatrix} r_g \cdot (\psi A_{g,1} + A_{g,2}(1 - \psi)) \\ + (1 - c_P) \cdot [R_g A_{g,1} a_{g,2} + r_g A_{g,1} a_{g,2}(1 - \psi) + a_{g,1}(\psi + a_{g,2}(1 - \psi))] \end{pmatrix}} \tag{3.17}$$

$$R_{g+1} = \frac{\begin{pmatrix} R_g \cdot (1 - c_H)[A_{g,1} A_{g,2} + (1 - s_2)(A_{g,1} a_{g,2} \\ + a_{g,1} A_{g,2}(1 - \psi)) + (1 - s_1)(a_{g,1}\psi + a_{g,1} a_{g,2}(1 - \psi))] \end{pmatrix}}{\begin{pmatrix} R_g \cdot (1 - c_H)[A_{g,1} A_{g,2} + (1 - s_2)(A_{g,1} a_{g,2} + a_{g,1} A_{g,2}(1 - \psi)) \\ + (1 - s_1)(a_{g,1}\psi + a_{g,1} a_{g,2}(1 - \psi))] + r_g(1 - s_1) \end{pmatrix}} \tag{3.18}$$

Genomic architecture of resistance

For now we have only considered the coevolution between a single major host locus and a single major parasite locus. However for many hosts and parasites, the interaction is presumably driven by several loci with major effects. One way to extend the above presented model towards L major interacting loci is to represent each host (parasite) genotype by a string of length L. Each entry of this string depicts the allelic state at one of the L major loci. For haploid species with biallelic loci the string can represented as a binary number and accordingly there is a total number of 2^L different host and different parasite genotypes. Therefore, the dimension of the infection matrix becomes $2^L * 2^L$, where each entry gives the probability that a host genotype i can be infected by a parasite genotype j. In a gene-for-gene like interaction each entry in the host string which is equal to 1 (0) can be interpreted as a resistance (susceptibility) allele. Similarly, each entry in the parasite string which is equal to 1 (0) denotes a infectivity (non-infective) allele. Accordingly, a host genotype i is (partially) resistant to a parasite genotype j if the host genotype for at least one locus l has the resistance gene and the parasite genotype has the corresponding non-infectivity allele. The gene-for-gene matrix for an interaction

with $L = 3$ host and parasite loci involved looks for example as follows:

$$\begin{array}{c} \\ 000 \\ 100 \\ 010 \\ 001 \\ 110 \\ 101 \\ 011 \\ 111 \end{array} \begin{array}{c} \begin{array}{cccccccc} 000 & 100 & 010 & 001 & 110 & 101 & 011 & 111 \end{array} \\ \left(\begin{array}{cccccccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array} \right) \end{array}. \tag{3.19}$$

In such multi-loci models the cost of resistance is usually integrated as a function of the number of resistance allele a particular host genotype i is carrying and the cost of infectivity becomes a function of the number of infectivity alleles a particular parasite genotype j is carrying. Such multi-locus models have been for example analysed by Tellier and Brown (2007a) and Sasaki (2000).

Box 1.1: Leonard model

The Leonard model is based on the following assumptions:

- Diploid hosts and haploid parasites
- gene-for-gene interaction
- coevolution between a bi-allelic host (resistance allele and susceptible allele) and bi-allelic parasite locus (infectivity allele and non-infectivity allele)
- resistance fully dominant to susceptibility
- non-infectivity fully dominant to infectivity
- discrete host and parasite generations
- delayed feedback between host and parasite. Fitness of a given parasite allele depends on the frequencies of the host genotypes in generation g, fitness of the host alleles depends on the frequencies of the parasite alleles at the beginning of generation $g+1$.

The model involves the following costs:

Notation Leonard (1994)	Notation used	Description
k	c_P	cost of infectivity (virulence)
t	c	effectiveness of resistance
a	ζ	advantage of virulent parasite on resistant host
c	c_H	cost of resistance
s	s	cost of infection

The notations for the frequencies have been adapted to the style of the thesis. Coevolution follows the following recurrence equations:

$$\begin{aligned} {}_{a_{g+1}} &= \frac{a_g\left[1-c_P+(1-r_g^2)\cdot g_P\right]}{1-(1-r_g^2)\cdot c+a\cdot\left[(1-r_g^2)\cdot(\zeta+t)-c_P\right]} \\ R_{g+1} &= \frac{R_g\left[1-c_H-s\cdot(1-c)+a_{g+1}\cdot s\cdot(c_P-\zeta-c)\right]}{1-s+a_{g+1}\cdot c_P\cdot s+(1-r_g^2)\cdot\left[c\cdot s-c_H-a_{g+1}\cdot s\cdot(\zeta+c)\right]} \end{aligned} \tag{3.20}$$

The polymorphic equilibrium frequencies are:

$$\begin{aligned} (1-\hat{r}^2) &= \frac{c_P}{\zeta+c} \\ \hat{a} &= \frac{cs-c_H}{cs-\zeta s} \end{aligned} \tag{3.21}$$

Tab. 3.1: Non-exhaustive list of population genetics model and their underlying assumptions which have been used to model host-parasite coevolution.

Model	# of loci	ploidy host	ploidy parasite	generations	disease type	infection matrix	delay
Leonard (1994)	single biallelic/single biallelic	diploid	haploid	discrete	monocyclic	GFG	yes
Tellier and Brown (2007b)	single biallelic/single biallelic	haploid	haploid	discrete	monocyclic	GFG	no
	single biallelic/single biallelic	haploid	haploid	discrete	polycyclic	GFG	no
Tellier and Brown (2007a)	two (biallelic)/two (biallelic)	haploid	haploid	discrete	monocyclic	GFG	no
Tellier and Brown (2007a)	two (biallelic)/two (biallelic)	haploid	haploid	discrete	polycyclic	GFG	no
Sasaki (2000)	L biallelic/L biallelic	haploid	haploid	discrete	monocyclic	GFG	no
Sasaki	single biallelic/single biallelic	haploid	haploid	discrete	monocyclic	GFG	no
Engelstaedter (2015)	single biallelic/single biallelic	diploid	haploid	discrete	monocyclic	general	no
Fenton *et al.* (2009)	single biallelic/single biallelic	haploid	haploid	discrete	monocyclic	general	no
Fenton *et al.* (2009)	two biallelic/two biallelic	haploid	haploid	discrete	monocyclic	IGFG	no
Seger (1988)	single biallelic/single biallelic	haploid	haploid	discrete	monocyclic	MA	no
Seger (1988)	single n-alleles/single n-alleles	haploid	haploid	discrete	monocyclic	MA	no
Seger (1988)	two loci biallelic/one locus four alleles	haploid	haploid	discrete	monocyclic	MA	no
Agrawal and Lively (2002)	two loci biallelic/two loci biallelic	haploid	haploid	discrete	monocyclic	general	no
Segarra (2005)	single biallelic/single biallelic	haploid	haploid	discrete	monocyclic	GFG	no
Segarra (2005)	two loci biallelic/two loci biallelic	haploid	haploid	discrete	monocyclic	GFG	no
Segarra (2005)	two loci biallelic/two loci biallelic	haploid	haploid	discrete	monocyclic	GFG	no
Jayakar (1970)	single biallelic/single biallelic	haploid	haploid	discrete	monocyclic	GFG	no
Jayakar (1970)	single biallelic/single biallelic	diploid	haploid	discrete	monocyclic	GFG	no
Jayakar (1970)	single biallelic/single biallelic	diploid	diploid	discrete	monocyclic	GFG	no

3.1.2 SI-models

SI (Susceptible-infected)-models (Anderson and May 1979) are widely used in epidemiological modeling for understanding the 1) temporal dynamics of an epidemic, 2) the conditions for an epidemic to start and to persist, 3) the effect of different intervention schemes on an epidemic and 4) to understand the evolution of host and parasite life-history traits such as virulence and resistance. A SI model keeps track of the number of susceptible (S) host individuals, the healthy compartment, and the number of infected (I) host individuals, the infected compartment, over time. The number changes in these compartments are usually modeled by a set of coupled differential equations. The use of differential equations implies implicitly that host and parasite generations are overlapping. Usually, these models include at least the following four parameters (see Fig. 3.2): b, the natural host birth rate, d, the natural host death rate, β, the disease transmission rate and γ, the parasite induced death rate or virulence sensu animal literature. The number of parasites can be obtained indirectly via the number of infected hosts.
The density of susceptible hosts increases by birth from susceptible individuals. It further increases by birth of new individuals from the infected class given that the disease is only transmitted horizontally (only among individuals but not from parent to offspring) and the parasite is not sterilizing the host. The number of susceptible individuals increases due to disease transmission at rate β when susceptible hosts interact with infected hosts. In SI-models, disease transmission can be either included in a density-dependent or a frequency-dependent manner. A discussion about disease transmission mechanisms and how they could apply to different types of diseases can be found in Antonovics (2017).
Both types of hosts (susceptible and infected ones) die at rate d. Further, the number of infected hosts decreases by parasite induced death and increases due to new infections. One central quantity in epidemiological models is the basic reproduction number R_0. It captures the number of secondary infection which are produced by a single infected individual which is placed into a fully susceptible host population (May and Anderson 1983). This number therefore determines whether a disease spreads (that is $R_0 > 1$) in a population or not (that is $R_0 \leq 1$). SI models can be readily extended to SIR models, where hosts can recover from the disease at rate ν. Further, such models can be extended to include more than one host type or more than one parasite type by increasing the number of compartments accordingly.

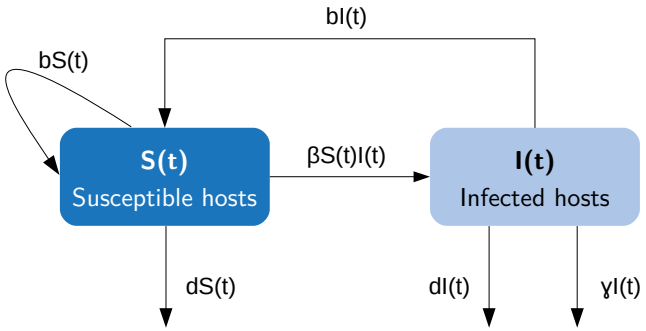


Fig. 3.2: Basic susceptible-infected model with density-dependent disease transmission. The number of susceptible hosts $S(t)$ decreases by new infections and natural death d and increases by natural birth from susceptible and infected hosts. The number of infected hosts increases by new infections and decrease by natural death and disease induced death. The parameters are: $\beta =$ disease transmission rate, $d =$ natural death rate, $b =$ natural birth rate and $\gamma =$ disease induced death rate.

3.2 Population genetics

The field of population genetics deals with the genetic composition and the evolutionary changes of genetic variants in a population and how they are shaped by the interplay of the evolutionary forces. Remember from chapter 2 that the evolutionary forces shaping genomic data are mutation, selection, recombination, genetic drift and migration. Selection acts on the phenotype and results in changes of allele frequencies at the genes underlying the phenotype. Here, an allele is defined as a variant form of a locus (can be a gene, a SNP, ...). These variant forms arise by mutations which are the result of erroneous DNA/RNA replication and range from changes at single base positions (Single Nucleotide Polymorphism), the insertion or deletion of several base pairs up to chromosomal rearrangements. The allelic state before the mutation occurred is called the ancestral state, the allelic state created by a mutation is called the derived state. Mutations create new diversity on which other forces such as selection, demography and recombination can act on. Thus, the ultimate fate (loss, fixation, maintenance) of a mutation is determined by the strength of the other evolutionary forces. A commonly accepted view is that the majority of *de novo* mutations has a neutral or slightly deleterious effect (but see Kern and Hahn (2018) for starting a recent controversial discussion (Jensen *et al.* 2019)). Here, the word neutral implies that the particular mutation does not alter the fitness of its carriers

compared to individuals with the ancestral state. The frequencies of neutral mutations are mainly affected by demography but can also change due to linkage to selected sites (Hill-Robertson-effect, Hill and Robertson (1966)).
In general linkage between sites can be broken up by recombination. Recombination can combine genetic information from both parents in the sequence of an offspring. Once a recombination event happens, the sequence of the resulting offspring contains sequence information from both parents on a single chromosome. It inherits all the sequence information to the right of the recombination event from one of the two parents and all sequence information to the left of the recombination event from the other parent. The recombination rate therefore also determines whether the allele frequencies, and thus the evolutionary trajectories, at two distinct loci in a genome are independent of each other or not. If two loci are independent of each other the chance for a single individual to have a particular allele A on the first locus and a particular allele B on the second locus is just equal to the product of the frequencies of these two allele in the population ($p_A \cdot p_B$) (Charlesworth and Charlesworth 2010). Conversely, if two loci are fully linked a particular allele at the first locus always comes together with a particular allele at the other locus. In this case the two loci are in strong linkage disequilibrium. The amount of linkage disequilibrium between two loci can be measured as:

$$D = g_{AB}g_{ab} - g_{aB}g_{Ab}, \tag{3.1}$$

where g_{AB} is the frequency of individuals in the population which have allele A at the first locus and allele B at the second locus and so forth (see e.g. Lewontin and Kojima (1960)). If $|D| = 1$ the two loci are in complete linkage and if $|D| = 0$ the two loci are completely independent of each other.

3.2.1 Genetic drift

Genetic drift describes the random changes in allele frequencies arising from stochastic effects in offspring reproduction. The standard model for capturing the effect of genetic drift is the so-called Wright-Fisher model which was introduced by Fisher and Wright in the 1930s. This model is based on the assumptions that the population is haploid with a constant size of $2N$ haploid indviduals (which corresponds to a diploid population with size N), non-overlapping and discrete generations, equal offspring distribution for each individual, random mating and panmixia and absence of selection and recombination.

The model tracks the temporal changes in allele frequencies of neutral alleles (therefore, selection is absent) at a given locus and is used to understand the time to fixation and the probability of fixation for an allele with some initial frequency p. For a bi-allelic locus there are two alleles, A and a, respectively. Allele A is in frequency p_t in generation t and allele a is in some frequency $q_t = 1 - p_t$. It is intuitive that in an extremely large population where each individual gets one child on average (arising from assumption that the population size stays constant and that all the individuals have the same offspring distribution), the frequencies of both alleles should remain constant over time. However, the number of offspring per individual v_i is a random variable which follows a Poisson distribution with mean 1 for a Wright-Fisher population. Thus, some individuals by chance contribute more than one offspring to the next generation whereas some individuals do not contribute any offspring. Individuals from generation t produce offspring, which will compose the population of generation $t + 1$ and die immediately after reproduction. As all alleles have equal fitness, each individual in generation t has the same chance to be the parent of a particular individual in generation $t + 1$. Therefore, the number of individuals in generation $t + 1$ with allele A can be obtained by sampling from a binomial distribution with mean p_t.

$$P(N_{A,t+1} = k) = \binom{2N}{k} p_t^k q_t^{2N-k} \tag{3.2}$$

The probability that an allele which is in initial frequency $p_0 = \frac{1}{2N}$ in the population is fixed is $\frac{1}{2N}$ (see e.g Kimura (1962)). Thus, the amount of genetic drift is larger in small compared to large populations. The expected time $\bar{t}(p_0)$ to fixation (conditional on fixation) for an allele with intial frequency p_0 is equal to (Kimura and Ohta 1969):

$$\bar{t}(p_0) = -\frac{1}{p_0} \left(4N_e \left(1 - p_0\right) log \left(1 - p_0\right)\right) \tag{3.3}$$

3.2.2 The standard coalescent

In the previous section we have described the effect of genetic drift forward in time. However, in very few cases one can possibly track a population forward in time. It is rather common that we have sampled data from a present day population and based on this sample we aim to infer properties of the past evolutionary history in a retrospective way. The coalescent is a probabilistic model to describe the genealogical relationship for

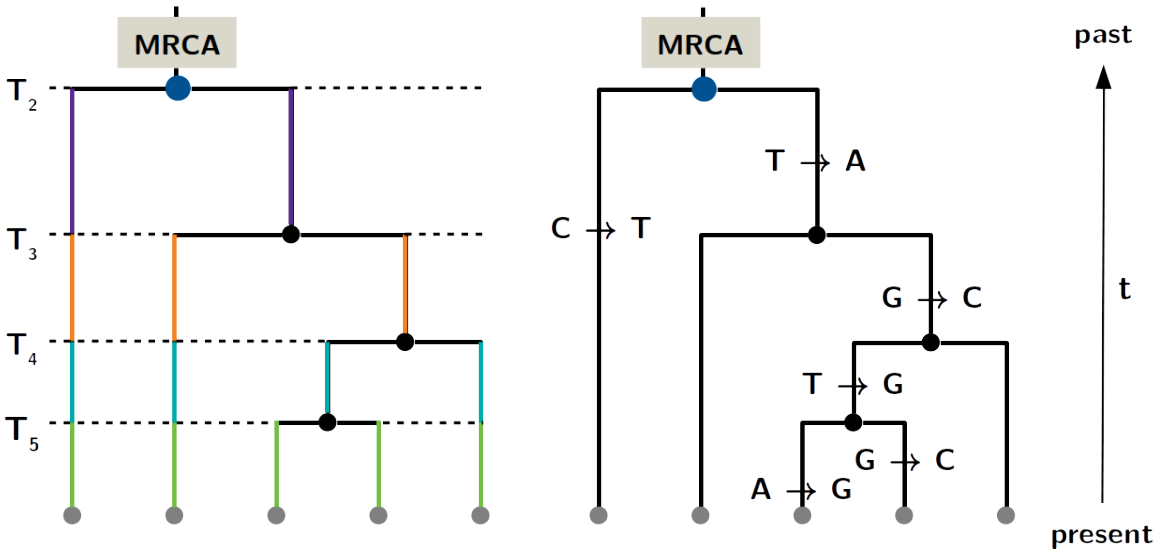


Fig. 3.3: Coalescent tree for a sample of size $n = 5$. Coalescent events are indicated by a black circle and reduce the number of ancestral lineages by one. The most recent common ancestor of the whole sample is shown in blue. The tree on the left contains all T_i ($i \in \{2, \ldots, n\}$) coalescent times which are also reflected by the coloured edges. The same tree is shown on the right this time with the mutation process on top. Time in the coalescent is measured from the present to the past.

a sample of size n up to the most recent common ancestor (MRCA) of all individuals in the sample. The standard n- coalescent (Kingman 1982a,b) can be used as a continuous-time approximation to describe the ancestral process of a Wright-Fisher population when the population size is large ($2N \to \infty$) and the sample size is small compared to the population size ($n << 2N$). The process starts with n samples (see grey circles in Fig. 3.3). Every time two lineages find their common ancestor t generations back in time, their ancestral lineages coalesce (black circles in Fig. 3.3) and the number of lineages ancestral to the sample reduces by one. This process repeats until the last two ancestral lineages find their most recent common ancestor (blue circle in Fig. 3.3). Therefore, the process can be illustrated by a bifurcating rooted tree, the coalescent tree. Each of the n leaves of this tree represents one of the samples and each node represents a coalescent event. The statistical properties of this process can be derived as follows (Hein *et al.* 2004; Wakeley 2008; Donnelly and Tavare 1995, and literature cited herein). For a sample of size $n = 2$ in a population of size $2N$, the probability that these two samples find their most recent common ancestor in the previous generation is equal to $\frac{1}{2N}$. The reasoning behind is as follows: The first individual chooses any of the parents. The probability that the second individual has the exact same parent is equal to $\frac{1}{2N}$ due to the assumption of

constant population size and equal fitness of all individuals in the parental population. Thus, the probability of a coalescent event in the previous generation $t = 1$ is equal to $\frac{1}{2N}$. Correspondingly, the probability of no coalescent event happening is $1-\frac{1}{2N}$. Therefore, the probability that $n = 2$ samples find their MRCA T_2 generations ago, follows a geometric distribution given by:

$$Pr(T_2 = t) = \left(1 - \frac{1}{2N}\right)^{t-1} \cdot \frac{1}{2N} \tag{3.4}$$

Thus, the expected time to the MRCA of a sample of size $n = 2$ is equal to $E[T_2] = 2N$ and the corresponding variance is equal to $\frac{1-\frac{1}{2N}}{(\frac{1}{2N})^2} = 2N(2N-1)$.
For a sample of size n the probability that no coalescent event happens in the previous generation and, accordingly the probability that n samples have n ancestors in the previous generation is given by (Hein *et al.* 2004 p.22, Wakeley 2008 p.68):

$$\begin{aligned} G_{k,k} &= 1 \cdot \frac{2N-1}{2N} \cdot \frac{2N-2}{2N} \cdot \ldots \cdot \frac{2N-k+2}{2N} \cdot \frac{2N-k+1}{2N} \\ &= \prod_{i=1}^{k-1} \left(1 - \frac{i}{2N}\right) \\ &= \left(1 - \sum_{i=1}^{k-1} \frac{i}{2N}\right) + \mathcal{O}\left(\left(\frac{1}{2N}\right)^2\right) \\ &\approx 1 - \frac{\binom{k}{2}}{2N} \qquad (N \to \infty) \end{aligned} \tag{3.5}$$

The first individual picks a parent with probability $p_1 = 1$. The probability that the second out of the n individuals chooses a different parent is equal to $p_2 = \frac{2N-1}{2N}$, the probability that the third individual chooses a parent which has not been previously chosen is $p_3 = \frac{2N-2}{2N}$ and so forth. Therefore, the probability that two samples in a sample of size k find their MRCA in the previous generation is approximately

$$G_{n,n-1} = \binom{n}{2} \frac{1}{2N} \tag{3.6}$$

if $n << 2N$ and $2N$ is large (the probability that more than two individuals find the same parent become negligible under this assumption). Thus, the probability that no pair out

of the n lineages has found their MRCA up to t generations in the past is:

$$\begin{aligned} Pr(T_n > t) &\approx \left(1 - \frac{\binom{n}{2}}{2N}\right)^t \\ &\approx e^{\frac{\binom{n}{2}}{2N}t} \\ &\approx e^{\binom{n}{2}t_c} \quad \text{with } t_c = \frac{t}{2N} \end{aligned} \tag{3.7}$$

Therefore, the time T_n to a coalescent event in a sample of size n can be approximated by an exponential distribution with rate $\binom{n}{2}$ when time is scaled in units of $2N$, the coalescent time scale. Accordingly, the expected time to a coalescent event in a sample of size $k = n$ is $E[T_n] = \frac{2}{n(n-1)}$. When a coalescent event happens, two ancestral lineages fuse into a single ancestral lineage and therefore, the number of ancestral lineages reduces to $n - 1$. The waiting time T_{n-1} to the next coalescent event among the $k = n - 1$ remaining ancestral lineages is again exponentially distributed, but this time with rate $\lambda_{n-1} = \binom{n-1}{2}$.
As the coalescent times are independent of each other, the distribution of the time to the MRCA (T_{MRCA}) for a sample of size n can be obtained as the convolution of the T_k ($k \in \{2, \ldots, n\}$) waiting times which are exponential random variabls with rates $\lambda_n = \binom{n}{2}, \lambda_{n-1} = \binom{n-1}{2}, \ldots, \lambda_3 = \binom{3}{2}, \lambda_2 = \binom{2}{2}$. Therefore, the expected time to the MRCA of the whole sample is (Hein *et al.* 2004 p.26, Wakeley 2008 p.76):

$$E(MRCA) = \sum_{k=2}^{n} E(T_k) = 2\sum_{k=2}^{n} \frac{1}{k(k-1)} \tag{3.8}$$

To obtain the expected total length $E(L_n)$ of all branches in the coalescent tree one has to proceed as follows. For an expected amount of time $E(T_k)$ there are k lineages present. Therefore, the expected length of all branches in this time interval is $kE[T_k]$ (see Fig. 3.3). Accordingly, one can obtain the expected total length of the coalescent tree as (Hein *et al.* 2004 p.27, Wakeley 2008 p.76):

$$E(L_n) = \sum_{k=2}^{n} k \cdot E(T_k) = \sum_{k=2}^{n} k\frac{2}{k(k-1)} = 2\sum_{k=1}^{n-1} \frac{1}{k} \tag{3.9}$$

The branches in the coalescent tree can be categorized based on the number of descendants

they have. Branches which have only one single descendant, namely one of the samples, are called external branches. Branches with several descendants are called internal branches. A coalescent tree for a sample of size n has a total number of $2n-2$ branches, from which n branches are external branches and $n-2$ branches are internal branches.

3.2.3 The standard coalescent with mutations

Due to the assumption of selective neutrality, the mutational process can be considered independently of the genealogical process, as neutral mutations do not affect the chance that an individual is chosen as a parent. Therefore, sequences under the coalescent with mutation can be obtained by first simulating the genealogy and afterwards placing the mutations on the genealogy (see Fig. 3.3). The mutational process itself can follow different mutation models. Which mutation model is most appropriate for a given species depends on the genome size and the biology of the species of concern.
In the **infinite alleles model** (Kimura and Crow 1964) each mutation creates a new allele in the population. Under the **infinite sites model** each mutations falls on a previously unmutated site (Kimura 1969). Therefore, there are at most two different states at a single position when comparing samples from several individuals of the population. Further, no recurrent mutations are happening at a single position in the genome. Accordingly under the infinite sites model, identity by state (for example when two individuals have both an adenine at some base position) also implies identity by descent (they have a common ancestor from which they inherited the mutation). The infinite sites mutation model is appropriate when the mutation rate/base/generation is very low and the sequence is very large. If information from an outgroup is available the ancestral state (the state before the mutation happened) is usually labelled as 0 and the derived state (the state created by the mutation) as 1.
In contrast to the infinite sites model, mutations can recurrently happen at the same position in the genome in a **finite sites model** (Jukes and Cantor 1969). Thus, more than two different states can be found at any position in the genome. Further under a finite sites model, identity by state does no longer imply identity by descent due to the possibility of back mutations. The finite sites model adds more biological reality for species which are characterized by high mutation rates and small genome sizes. Jukes and Cantor (1969) introduced the first, very basic, finite sites model. Their model relies on the assumption that all bases have equal frequencies along the whole sequence. Further, mutations happen

to any other base with equal probability. More refined models additionally take the non-equal distribution of bases along the sequence and transition to transversion biases into account (Tamura 1992; Tamura and Nei 1993; Felsenstein 1981; Kishino and Hasegawa 1989).

One core parameter in population genetics is the so called population mutation rate θ. The population mutation rate $\theta = 4 \cdot N_e\mu$ can be interpreted as the expected number of mutations which have accumulated in a pair of individuals since their most recent common ancestor (Hein *et al.* 2004, p.40). The idea behind this interpretation is as follows the expected time to coalescent for a sample of size $n = 2$ is $\binom{2}{2} = 1$ on the coalescent time scale. μ is the mutation rate per generation per genome. To obtain the real time, the time in units of the coalescent time scale has to be rescaled by $2N$. Accordingly, the total length of the branches separating the two individuals from each other is $2E[T_2] \cdot 2N = 4N$ and the expected number of mutations which have accumulated since the most recent common ancestor is $\theta = 4N\mu$.

The number of mutations per unit coalescent time is Poisson distributed with mean $\theta/2$. Therefore, the time to a mutation event is exponentially distributed with mean $\theta/2$. Accordingly the expected number of segregating sites for a sample of size n can be obtained by weighting the expected number of mutations per unit time interval on the coalescent time scale by the expected length of the coalescent tree.

$$E(S_n) = \frac{\theta}{2}E[L_n] = \frac{\theta}{2} \cdot 2\sum_{k=1}^{n-1}\frac{1}{k} = \theta\sum_{k=1}^{n-1}\frac{1}{k} \tag{3.10}$$

In general there are two different possibilities to simulate a neutral coalescent tree with mutations following an infinite sites model.

Recipe 1 (see Hein *et al.* (2004), p.42-43)

1. Set n=k
2. While $k > 1$
   - Draw a waiting time t_w from an exponential distribution with rate $\lambda = (\binom{k}{2} + \frac{k\theta}{2})$. The underlying rationale is that the coalescent process and the mutation process are two independent Poisson processes. Accordingly, the waiting time to the first event (either a coalescent event or a mutation event) is exponentially

distributed with rate λ which is equal to the sum of the rates of both Poisson processes.

- With probability $\frac{\frac{\theta}{2}}{\frac{\theta}{2}+\binom{k}{2}}$ the event is a mutation event and with probability $\frac{\binom{k}{2}}{\frac{\theta}{2}+\binom{k}{2}}$ the event is a coalescent event
- If the event is a coalescent randomly choose two of the k lineages to coalesce at time t_w and set $k = k - 1$.
- It the event is mutation event choose any of the k lineages with equal probability and place a mutation on this lineage at time t_w

Recipe 2 (see Hein *et al.* (2004), p.41-42) (Hudson 2002))

1. Set n=k
2. While $k > 1$
   - Draw the time to the next coalescent event t_k from an exponential distribution with mean $\binom{k}{2}$
   - Randomly choose two of the k individuals to coalesce at time T_k
   - Decrease k by one
3. Draw the number of mutations on the whole genealogy from a Poisson distribution with rate $\lambda = \frac{\theta}{2} \sum_{k=2}^{n} t_k$
4. Distribute the K mutations uniformly on the genealogy

Therefore, the coalescent tree not only provides means to obtain the statistical properties of the genealogy of a sample of size n but also provides means to simulate sequences for a sample of size n.

Based on the example shown in Fig. 3.3 the resulting sequences could look as follows (each segregating site shown in red):

Sequence 1	A	T	A	T	G	C	T	A	A	C	T	G	T	A	T	G	T
Sequence 2	A	T	A	T	G	C	T	A	A	C	T	G	A	A	C	G	T
Sequence 3	A	T	A	T	C	C	T	G	A	C	T	G	A	A	C	G	G
Sequence 4	A	T	A	T	C	C	T	A	A	C	T	C	A	A	C	G	G
Sequence 5	A	T	A	T	C	C	T	A	A	C	T	G	A	A	C	G	T

The site frequency spectrum (SFS

For many applications it is valuable to condense the information contained in the sequence data into summary statistics. The **site frequency spectrum** (also allele frequency spectrum) is a summary statistic of the sequences of n-sampled individuals. It condenses the information about the frequencies of all segregating sites in the sample. The site frequency spectrum comes in two flavours depending on whether the ancestral state at each particular position in the sequence is known or not. If an outgroup sequence is available and thus, the (likely) ancestral state of the allele can be determined, the unfolded site frequency spectrum is usually used. It summarizes the number of segregating sites in the sample with derived allele frequency i. Thus, the site frequency spectrum is a vector $\boldsymbol{\xi} = (\xi_1, \xi_2, ..., \xi_{n-2}, \xi_{n-1})$ of size $n-1$ and the sum of all vector entries is equal to the number of segregating sites ($|\boldsymbol{\xi}| = S$).
Whenever the ancestral states are unknown, the folded site frequency spectrum has to be used. The folded site frequency spectrum is a vector $\boldsymbol{\eta}$ of length $\lfloor n/2 \rfloor$ where each entry η_i gives the number of segregating sites in the sample with minor allele frequency i. The folded site frequency spectrum links to the unfolded site frequency spectrum as follows:

$$\eta_i = \frac{\xi_i + \xi_{n-i}}{1 + \delta_{i,n-i}} \quad 1 \leq i \leq \lfloor n/2 \rfloor$$
$$\delta_{i,n-i} = \begin{cases} 0 & i \neq j \\ 1 & i = j \\ , \end{cases} \tag{3.11}$$

which corresponds to collapsing ξ_1 and ξ_{n-1} into a single class η_1 and so forth. The expected unfolded site frequency spectrum for a Wright-Fisher population at equilibrium is given by:

$$\xi_i = \frac{\theta}{i} \tag{3.12}$$

Therefore, given that the ancestral state is known, the sequences from the samples in Fig.3.3 can be converted to the unfolded site frequency spectrum as shown in Fig. 3.4.

G	A	G	T	C	T							
G	A	G	T	T	T		0	0	0	0	1	0
G	A	G	A	C	T		0	0	0	1	0	0
C	G	G	A	C	G		1	1	0	1	0	1
C	A	C	A	C	G		1	0	1	1	0	1
C	A	G	A	C	T		1	0	0	1	0	0

$\xi = (\xi_1, \xi_2, \xi_3, \xi_4) = (3, 1, 1, 1)$

Fig. 3.4: Converting the sequences from the example shown in Fig. 3.3 to an unfolded site frequency spectrum when the ancestral state at each site is known.

However, if the ancestral state is unknown, only the folded site frequency spectrum can be calculated (see Fig.3.5) for the sequences of example Fig. 3.3.

G	A	G	T	T	T		1	0	0	1	1	0
G	A	G	A	C	T		1	0	0	0	0	0
C	G	G	A	C	G		0	1	0	0	0	1
C	A	C	A	C	G		0	0	1	0	0	1
C	A	G	A	C	T		0	0	0	0	0	0

$\eta = (\eta_1, \eta_2) = (4, 2)$

Fig. 3.5: Converting the sequences from the example shown in Fig. 3.3 to a folded site frequency spectrum. Note that it is assumed that the ancestral state is unknown. Therefore, the minor allele at each sites was labeled with 1 and the major allele with 0.

The site frequency spectrum is intimately linked to the ancestral process of the sample. When mutations follow an infinite sites model, we know immediately that a mutation which has given rise to a singleton (a segregating site in frequency one) must have happened on an external branch. Similarly, each segregating site which is in frequency two, a so-called doubleton, must be due to a mutation on a branch which is the ancestor of exactly two of the nowadays samples. The site frequency spectrum can be roughly

split into three categories, namely the low frequency classes, which include singletons and doubletons, the intermediate frequency classes and the high frequency classes. The site frequency spectrum can be further summarized by other summary statistics.
Additionally, as the expected number of segregating sites in each frequency class is known under neutrality, deviations from the expected neutral site frequency spectrum are indicative about that one of the assumption which have been used to derive the coalescent must have been violated in the population of concern.

3.2.4 Non constant population size

There are several ways how the population size can vary over time. Two possible simple scenarios are 1) a bottleneck backward in time (which corresponds to a population expansion forward in time) and 2) a population expansion backward in time (which corresponds to a population bottleneck forward in time). In this section we will only verbally describe how these different scenarios will affect the distribution of branch lengths in the coalescent tree and the expected time to the most recent common ancestor. When the population size is decreasing backward in time, the number of individuals which can be chosen as parent decreases. Accordingly, the probability of a coalescent event to happen increases. Therefore, a population bottleneck backward in time decreases the size of the internal branches compared to the external branches and the resulting genealogies are more star-like. This results in an excess of low frequency variants and a decrease in intermediate frequency classes compared to the neutral site frequency spectrum. It is further expected that the variance in the number of segregating sites and expected coalescent times between several independent loci in the genome is low (Wakeley 2008, p.120).
On the other hand if the population is expanding backward in time and thus, there is a population bottleneck forward in time, it is expected that the internal branches are longer compared to external branches (Wakeley 2008, p.120).

3.2.5 Population structure

Population structure can be due to different processes. Either individuals are splitted up into several populations which are connected by migration with varying rate or there can be also assortative mating within a single population and accordingly the chance of picking a particular parent backward in time is not the same for all individuals. If a population is structured, the ancestral process has to be modelled by a so-called structured

coalescent (Wakeley 2008, p.131). In a structured coalescent the samples are taken from l-demes in the population. Therefore, an additional event, namely migration, has to be accounted for besides coalescent and mutation. Backward in time samples can migrate from population j to population i which corresponds to a migration event from population i to population j forward in time.
In general, population structure tends to increase the length of internal branches compared to external branches and accordingly results in an excess of intermediate frequency variants. The prolongation of the internal compared to external branches arises from the fact that in order to find the most recent common ancestor of the whole sample the last two remaining lineages have to first migrate to the same deme and coalesce there. If migration rates are low this event can reach long way back into the past. If populations are completely isolated for some amount of time the variance in coalescent time among independent sites tends to be small. However, if migration rates are low there can be considerable variation in coalescent times and number of segregating sites among independent sites (Wakeley 2008, p.120).

3.2.6 Selection

Selection also changes the shape of the coalescent tree compared to the neutral expectation. Under positive selection, the frequency of a particular allele is increasing forward in time. Further, individuals with the beneficial allele have a higher chance to be selected as a parents. However, initially the frequency of a benefical allele is assumed to be low in a population. Therefore, the genealogical process resembles the one of a a population decline backward in time. Accordingly, the genealogy tends to be star like and the site frequency spectrum at the selected locus is characterized by an excess of low frequency and high frequency variants. However, a selective sweep only affects the locus under selection and not all sides in the genome. Therefore, one way to distinguish a population expansion from a selective sweep is to check if all independent sites in the genome show an excess of low frequency variants or if the excess of low frequency variants is restricted to a single locus.
In contrast to positive selection, balancing selection affects the genealogical relationship among the samples in a way similar to population structure (Hein *et al.* 2004, p.118-119). Here the individuals carrying the different alleles can be thought of individuals belonging to several demes. If no recurrent mutations events between alleles take place the coales-

cent process resembles the one of two populations which became completely isolated from each other after the split from the ancestral population. Accordingly, first all lineages having a particular allele coalesce within their deme and then find their most recent common ancestor only at the time when a mutation created the derived allele. If recurrent mutations are taking place the ancestral process is similar to the one of two demes being connected by migration. However, 'migration' events for two alleles being maintained by balancing selection are recurrent mutation between alleles. In either type of scenario balancing selection results in an relative expansion of the internal branches compared to the external branches. Accordingly, the site frequency spectrum is expected to show an excess of intermediate frequency variants.

3.2.7 Recombination

Recombination reshuffles information from the parents. Seen forward in time a recombination event combines the information of both parents into a single sequence. All parts of the sequence left to the recombination event is obtained from the parent on the left and everything to the right hand of the recombination parents is obtained from the other parent. Accordingly backward in time, a recombination event corresponds to the split of an ancestral lineage into two ancestral lineages (Hudson 1983). Given that there is recombination at the locus of concern, the ancestral process of its sequence can be no longer describe by a coalescent tree. Rather, the ancestral process has to be described by a graph, the so called ancestral recombination graph, into which the coalescent tree of each site in the sequence is embedded (Hein *et al.* 2004, p.139-142).

3.2.8 Summary statistics based on the site frequency spectrum to detect deviations from neutrality

There are several summary statistics which further summarize the site frequency spectrum. Many of them give an estimate of the population mutation rate under the assumption of constant population size, absence of recombination and selection and random mating. As all of them are based on different parts of the site frequency spectrum and therefore, are differently affected by deviations from neutrality they can be used to detect signatures of selection or population structure. One well known summary statistic of the site frequency spectrum is the average number θ_π of of pairwise nucleotide differences π_{ij} for a sample

of size n (Nei and Tajima 1981).

$$\theta_\pi = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=2}^{n} \pi_{ij} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i(n-i)\xi_i \tag{3.13}$$

θ_π is increasing with an excess of intermediate frequency variants. One famous summary statistic is Tajima's D which combines the Watterson estimator (Watterson 1975) of the population mutation rate and θ_π. The Watterson estimator is obtained as $\theta_W = S/\sum_{i=1}^{n-1} \frac{1}{i}$, where S is the number of segregating sites in the sample. Under neutrality it is expected that $\theta_\pi - \theta_W = 0$. Therefore, Tajima (1989) proposed the following summary statistic to the test for deviations from the neutral model:

$$T_D = \frac{\theta_\pi - \frac{S}{a_n}}{\sqrt{\hat{V}\left[\theta_\pi - \frac{S}{a_n}\right]}} \quad \text{with} \quad a_n = \sum_{i=1}^{n-1} \frac{1}{i} \tag{3.14}$$

Accordingly, an excess of intermediate frequency variants results in positive values of Tajima's D and a deficit of intermediate frequency variants results in negative values of Tajima's D. Therefore, positive values of Tajima's are indicative of balancing selection, population expansions backward in time (population decrease forward in time) and structured populations. Negative values of Tajima's D on the other hand are indicative of selective sweeps and populations decrease backward in time (population expansions forward in time). Fu and Li (1993) derived two further summary statistics which are based on a similar idea. Fu and Li's D Fu and Li (1993) is based on comparing the Watterson estimator to the number of singletons, whereas Fu and Li's F Fu and Li (1993) is based on comparing θ_π to the number of singletons.

3.3 Approximate Bayesian Computation

Extracting information about evolutionary relevant processes from genomic data requires the use of statistical inference methods. For many cases the main interest is to infer properties of the underlying process (the Model M) and the parameters associated with the process ($\mathbf{\Theta} = \{\Theta_1, \Theta_2, \ldots\}$) based on an observed data set D. To give a short example, let us assume we aim to infer the rate of an exponential population expansion from sequences of n-individuals. In this case the sequences of the n-individuals would be the observed

data set D, the model M is an exponential population expansion and the parameter to be inferred $\Theta = \{\Theta_1\}$ is the rate of the expansion. One widely used method to tackle such problems, especially when the underlying models are complex, is Approximate Bayesian Computation. Baye's theorem states that:

$$p(\Theta|D) = \frac{p(D|\Theta)\pi(\Theta)}{p(D)} \tag{3.1}$$

$p(\Theta|D)$ is called the posterior distribution and gives the probability distribution of the parameters of interest given the data. $p(D|\Theta)$ is called the likelihood and gives the probability of observing the data conditional on the parameters, $\pi(\Theta)$ is the prior distribution of the parameters and $p(D)$ is called the marginal density of the data. Generally, the Bayesian approach can be summarized as updating our prior knowledge ($\pi(\Theta)$) on the parameters of interest once we have seen the data. This updated knowledge is captured in the posterior distribution. Yet, for biologically complex models it can be hard or even impossible to calculate the likelihood or/and the marginal density. In Approximate Bayesian computation methods the computation of the likelihood is circumvented by performing simulations. An overview about the historical developments of Approximate Bayesian Computation in population genetics is e.g given in Beaumont (2010) and Sunnaker *et al.* (2013). Generally, Approximate Bayesian Computation methods can be roughly divided into three different approaches (Csillery *et al.* 2010; Sunnaker *et al.* 2013).

- Rejection algorithms (Pritchard *et al.* 1999; Beaumont *et al.* 2002; Wegmann *et al.* 2010)
- Markov chain monte carlo (MCMC) without likelihood methods (Marjoram *et al.* 2003; Wegmann *et al.* 2009, 2010)
- Sequential Monte Carlo without likelihoods (Sisson *et al.* 2007; Beaumont *et al.* 2009; Wegmann *et al.* 2010)

Irrespective of the chosen approach, the observed data D are usually summarized by a set of summary statistics $\boldsymbol{s}_{\text{obs}} = \{s_{1,\text{obs}}, s_{2,\text{obs}}, \ldots\}$ which are sufficient about the data. Sufficient summary statistics capture all the relevant information about the data (Wegmann *et al.* 2010, and corresponding manual of ABCToolbox). Converting the data set into a set of summary statistics reduces the dimensionality of the data and hence, is supposed to lower the effect of 'curse of dimensionality' problems. In the example from before the

sequences could be summarized by the number of segregating sites or the site frequency spectrum. Common to all three approaches is to simulate a large amount of data sets for the given model and for different values of the parameters of interest. For each simulated data set i the same set of summary statistics $\boldsymbol{s_i}$ as for the observed data is calculated and compared to the statistics of the observed data set. However, the three approaches differ in the way how the posterior distribution is obtained.

Rejection algorithms are based on simulating a large number N of datasets under a particular model M and by randomly drawing parameter values from the prior distribution $\pi(\Theta)$ for each simulation. After all simulations have been performed, each simulated data set D' is compared to the observed data set D. A simulation i is accepted when the summary statistics of the simulated data set $\boldsymbol{s_i}$ are sufficiently close to the summary statistics of the observed data set (s_{obs}), that is $\|s_i - s_{\text{obs}}\| < \epsilon$. Here, $\|\cdot\|$ is a metric to calculate the distance between the simulated and the observed data set. The tolerance ϵ can be either fixed or chosen in such a way that a certain proportion of all simulated data sets is accepted. Based on all retained simulations the posterior distribution of the parameters of interest is either obtained by immediately fitting a kernel to the parameter values of the retained simulations or by first performing some post-sampling adjustment to account for the distance between the summary statistics of the observed data set and the retained simulations and then applying a kernel density estimation to the adjusted parameter values. The post-sampling adjustment can be for example achieved by performing a local linear regression (Beaumont *et al.* 2002) or by applying a general linear model (Leuenberger and Wegmann 2010). In chapter 5 we will use the rejection algorithm with a post-sampling adjustment based on a general linear model as implemented in ABCtoolbox (Wegmann *et al.* 2010).

MCMC without likelihood methods (Marjoram *et al.* 2003) are based on drawing an initial value $\Theta^{(0)}$ from the prior distribution. At each iteration t of the Markov chain a new set of parameters Θ' is proposed based on some transition kernel $K(\Theta'|\Theta^{(t-1)})$ and the parameters of the previous iteration. A data set D' is simulated with the proposed values Θ'. If the summary statistics of the simulated data set are within the tolerance distance to the observed data set $\|s' - s_{\text{obs}}\| < \epsilon$, then set $\Theta^{(t)} = \Theta'$ to be the new state of the chain with probability $\frac{\pi(\Theta')K(\Theta^{(t-1)}|\Theta')}{\pi(\Theta^{(t-1)})K(\Theta'|\Theta^{(t-1)})}$ otherwise set $\Theta^{(t)} = \Theta^{(t-1)}$. With an an increasing number of iterations the Markov chain approximately converges to the posterior distribution (Marjoram *et al.* 2003; Beaumont 2010; Wegmann *et al.* 2009).

Sequential Monte Carlo methods (Sisson *et al.* 2007; Beaumont *et al.* 2009) com-

bine properties of both other approaches. First, parameter values are drawn from the prior distribution until N simulations have been accepted based on an initial tolerance level. Based on a density kernel of these accepted parameter values new parameter values are proposed until again N simulations have been accepted, this time based on a lower tolerance level.

4

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

4.1 Abstract

Uncovering the genes governing the outcome of host-parasite interactions, and thus expected to be under coevolution, is of importance for disease management in agriculture and human medicine. Increasing amounts of host and parasite full genome-data offer new perspectives to gain such information. One promising approach is to perform cross-species genome-wide association studies based on genomic data of infected hosts and their associated parasites strains. We aim to understand the power of such an approach for different types of coevolutionary dynamics and host-parasite genotype-by-genotype (GxG) interactions over the course of the coevolutionary history. Therefore, we use two indices, the cross species association (CSA) and the cross species prevalence (CSP), the latter additionally incorporating genomic data from uninfected hosts. For both indices, we derive genome-wide significance thresholds by computing their expected distribution over all neutral loci, i.e. those not involved in determining the outcome of interaction. Using two types of coevolutionary models, namely a population genetics and an epidemiological model, we demonstrate that the power of these indices to detect the interacting loci depends on 1) the type of GxG interactions, 2) type of coevolutionary dynamics, and 3) varies over time. When coevolution follows trench-warfare dynamics, CSA and CSP are very accurate in pinpointing the loci under coevolution. However, under arms-race dynamics, the association indices have limited power especially when the GxG interactions are asymmetric, such as in a gene-for-gene interaction. Furthermore, we reveal that the combination of both indices across time samples is a good indicator of the underlying infection matrix. Thus, our results provide novel insights into the power and biological interpretation of cross-species association studies using samples from natural populations or controlled experiments.

Keywords

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

4.2 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 understand host-parasite coevolutionary interactions have been proposed and performed (Ebert 2018; Wang *et al.* 2018; MacPherson *et al.* 2018; Nuismer *et al.* 2017). However, such analyses rely on performing controlled experiments for large host and parasites genotype sample sizes. A promising less-labour intensive alternative is to perform cross-species association studies based on whole-genome data of infected hosts and their associated parasites (Bartha *et al.* 2013; Bartoli and Roux 2017). Such data sets inherently contain phenotypic information for each sampled host-parasite pair, namely the particular host is susceptible to the particular parasite and the particular parasite is infective on the particular host. Accordingly, the causal genetic variants for host susceptibility and parasite infectivity are expected to show statistically significant associations (Host Genotype x Parasite Genotype) and thus, to be distinguishable from neutral variants without any effect on the interaction outcome. This approach has been applied to uncover strong associations between SNPs in the human major histocompatibility complex (MHC) and amino acids in known HIV epitopes (Bartha *et al.* 2013).

In principle, such an approach can be readily extended to any coevolutionary system where it is possible to call SNPs for a sample of infected hosts and the corresponding infecting parasites (Bartoli and Roux 2017). For example, it could be applied to transcriptome data of infected hosts and the corresponding infecting parasites (Dobon *et al.* 2016) or to whole genome-data from controlled coevolutionary experiments.

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

synergistic (symbiosis) and antagonistic (host-parasite, prey-predator) interactions.
Allele frequency changes at the coevolutionary loci are commonly described by a continuum between two extremes (Woolhouse *et al.* 2002), namely, arms-race (Stahl and Bishop 2000; Woolhouse *et al.* 2002; Dawkins and Krebs 1979) and trench-warfare (Stahl *et al.* 1999) dynamics. In arms-race dynamics, covevolution causes recurrent fixation of alleles at the interacting loci, and accordingly, allelic polymorphism is only transient. In contrast, several alleles are maintained for a long-period of time in trench-warfare dynamics, with their individual frequencies either persistently fluctuating over time or converging towards a stable polymorphic equilibrium. Note that in both scenarios, allele frequencies fluctuate over time before reaching fixation or the stable equilibrium.
The speed and type of frequency fluctuations depends on the underlying GxG interactions. Given the assumption that few major genes determine the interaction outcome, these GxG interactions can be captured in a so called infection matrix $\mathcal{A}$. Here, each entry α_{ij} stores the probability that a parasite genotype j can infect a host genotype i (Tab. 4.1). Two well known infection matrices, are the matching-allele (MA) and the gene-for-gene (GFG) model. Both the MA and the GFG model represent some point in a continuum of infection matrices (Agrawal and Lively 2002) and are a subset of more complex matrices (with several alleles or loci, Gandon and Michalakis (2002), Ashby and Boots (2017)). In MA interactions a given parasite genotype can only infect a host when it matches the particular host allele (diagonal coefficients in Tab. 4.1b). For a 2x2 infection matrix the probabilities to infect the "non-matching" host genotypes can be defined as $1 - c_1$ and $1 - c_2$ (off diagonal coefficients in Tab. 4.1b). GFG interactions (Tab. 4.1c) are characterized by a universally susceptible host genotype (here host $i = 1$) and an universally infective parasite (here parasite $j = 2$). Here, the probability that host $i = 2$ is infected by parasite $j = 1$ is denoted by $1 - c$.

This continuum of coevolutionary dynamics in combination with the varying extent of specificity in GxG host-parasite interactions gives rise to several important questions in the context of cross-species association studies: 1) Which statistics can be used in cross-species association studies? 2) What is the power of these statistics to distinguish neutral from coevolutionary loci? 3) How is their power affected by allele frequencies fluctuations and the underlying infection matrix? 4) Are combinations of these statistics indicative of the underlying infection matrix? To answer these questions, we first use and define two indices namely, the cross-species association (CSA) index (based on Bartoli and Roux (2017)) and the cross species prevalence (CSP) index to measure the association of al-

Tab. 4.1: Infection matrices for coevolutionary models

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

The infection matrix $\mathcal{A}$ determines the outcome of the interaction between host genotypes (rows) and parasite genotypes (columns). A rate $\alpha_{ij} = 1$ indicates infection while $\alpha_{ij} = 1 - c_x$ indicates the degree of infection (or degree of resistance).

leles between the coevolutionary locus in the host and the parasite. Second, we predict the power of these cross-species indices to distinguish coevolving loci from neutral loci by computing the expected distribution for associations between neutral host SNPs to a neutral parasite SNPs. As a result, we quantify the statistical power of these statistics to detect the loci underlying coevolution over the course of coevolutionary cycles, and show that applying our indices to host and parasite samples from different time points gives an indication about the underlying GxG interaction matrix at the coevolutionary loci. We then discuss the applicability and usefulness of our GWA indices when applying them to samples from natural and controlled coevolutionary experiments.

4.3 Methods

4.3.1 The coevolutionary models

We assume that the outcome of an interaction between a host and a parasite, namely if the host is infected or not, is determined by a single biallelic host and single biallelic parasite locus. These two loci are defined as the coevolving loci. For simplicity, we consider a haploid model for hosts and parasites. The outcome of the interaction between a particular host and parasite genotype is determined by the infection matrix $\mathcal{A} = (\alpha_{ij})$ with $1 \leq i, j \leq A$ (Table 4.1).

Model 4A: population genetics model

First, we use a simple population genetics model (henceforward termed model 4A) to study the allele frequency changes at the coevolving loci under the assumption of very large (infinite) host and parasite population sizes. We assume that host and parasite generations are discrete and synchronized in terms of reproduction. The frequency of host genotype h_i and parasite genotype p_j in generation $g+1$ is obtained as:

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

where $w_{H,i}$ ($w_{P,j}$) is the fitness of host genotype i (parasite genotype j). The average fitness of the host (parasite) population, $\bar{w}_{H,g}$ ($\bar{w}_{P,g}$), is obtained as $\sum_{i=1}^{2} w_{H,i} \cdot h_{i.g}$ ($\sum_{j=1}^{2} w_{P,j} \cdot p_{j,g}$). Every generation g a proportion ϕ_g, *i.e.* the disease prevalence, of the host population interacts with the parasite population in a frequency-dependent manner. Note that ϕ_g can be constant over time or depend on the epidemiological dynamics of the system (see Model 4B), but in principle its value can be measured experimentally at each generation g. Whether a particular interaction between host genotype i and parasite genotype j results in an infection depends on the matrix $\mathcal{A}$. An infection reduces the relative fitness of hosts by an amount s (cost of infection). Further, each host genotype i (parasite genotype j) can be associated with some fitness cost c_{H_i} (c_{P_j}), such as a cost of resistance (infectivity). Therefore, the frequencies of the different host and parasite genotypes can be modelled using the following recurrence equations:

$$h_{i,g+1} = \frac{h_{i,g} \cdot (1 - c_{H_i}) \cdot \left(1 - \phi_g \cdot s \cdot \sum_{j=1}^{2} \alpha_{ij} p_{j,g}\right)}{\bar{w}_{H,g}} \tag{4.1a}$$

$$p_{j,g+1} = \frac{p_{j,g} \cdot (1 - c_{P_j}) \cdot \left(\sum_{i=1}^{2} \alpha_{ij} h_{i,g}\right)}{\bar{w}_{P,g}} \tag{4.1b}$$

This dynamical system is at 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:

$$\hat{h}_1 = \frac{\alpha_{22}(1-c_{P_2}) - \alpha_{21}(1-c_{P_1})}{(\alpha_{11}-\alpha_{21})(1-c_{P_1}) + (\alpha_{22}-\alpha_{12})(1-c_{P_2})} \tag{4.2a}$$

$$\hat{p}_1 = \frac{c_{H_1} - c_{H_2} + \phi s\left(\alpha_{12}(1-c_{H_1}) - \alpha_{22}(1-c_{H_2})\right)}{\phi s\left((\alpha_{12}-\alpha_{11})(1-c_{H_1}) + (\alpha_{21}-\alpha_{22})(1-c_{H_2})\right)} \tag{4.2b}$$

In line with previous studies, for both the symmetric and asymmetric MA model we assume no costs $c_{H_1} = c_{H_2} = c_{P_1} = c_{P_2} = 0$ (Gandon and Nuismer 2009). For the GFG model we use the infection matrix shown in Tab. 4.1c) and assume that $0 < c_{H_2}, c_{P_2} < 1$ and $c_{H_1} = c_{P_1} = 0$ (Tellier and Brown 2007b).

Model 4B: model with epidemiological dynamics and feedback

In model 4B we consider a continuous time coevolutionary model (Živković *et al.* 2019) based on a known Susceptible-Infected model (May and Anderson 1983; Boots *et al.* 2014; Ashby and Boots 2017). This model allows for modelling simultaneously the changes in allele frequencies and changes in population sizes arising from epidemiological feedback. 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 interior equilibrium point) (Živković *et al.* 2019; Ashby and Boots 2017). We focus here chiefly on the trench warfare outcome. 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. 4.3a and the change in number of infected individuals I_{ij} is given by Eq. 4.3b.

$$\frac{dS_i}{dt} = S_i\left[b(1-c_{H_i}) - d - \sum_{j=1}^{2}\alpha_{ij}\beta(1-c_{P_j})\sum_{k=1}^{2}I_{kj}\right] + b(1-c_{H_i})(1-s)\sum_{j=1}^{2}I_{ij}, \tag{4.3a}$$

$$\frac{dI_{ij}}{dt} = -dI_{ij} + S_i\left[\alpha_{ij}\beta(1-c_{P_j})\sum_{k=1}^{2}I_{kj}\right]. \tag{4.3b}$$

The number of parasites of type j is obtained as $P_j = \sum_i I_{ij}$ and hence, the change in number of parasites of type j is given by $\frac{dP_j}{dt} = \sum_i \frac{dI_{ij}}{dt}$. Hosts reproduce at natural birth rate b and die at natural death rate d. The total host population size at time step t is $N = \sum_{i,j} I_{ij} + \sum_i S_i$. We assume that there is no vertical transmission of disease, and the infections are sustained in the populations through an overlap between generations. The newborn hosts are always susceptible, and they can get infected by the horizontal transfer of infections with rate β. The costs c_{H_i}, c_{P_j} and s are defined as in model 4A.
To simulate the dynamics, we discretize model 4B into small time steps of size δ_t. Hence, one discrete time step t consists of $\frac{1}{\delta_t}$ time steps. The value of δ_t is chosen so that the continuous and discretized time dynamics match. The population size changes, allele frequencies changes and corresponding association statistics values are computed over time and at the equilibrium point. The equilibrium points can be computed for this system (Živković *et al.* 2019) but as the formulae are complex and not very intuitive we refrain from using them here. The disease prevalence is here a inherent property of the disease dynamics and allele frequencies as defined under the eco-evo feedbacks (Boots *et al.* 2014; Ashby *et al.* 2019), and thus varies over time:

$$\phi_t = \frac{\left(\sum_{j=1}^{2} \beta \left(1 - c_{P_j}\right) \sum_{k=1}^{2} I_{kj}\right)}{N}. \tag{4.4}$$

4.3.2 Definition of the association statistics

We assume that n_{T} host individuals have been sampled and genotyped at each biallelic single nucleotide polymorphism (SNP) in the genome, so that two types of hosts are found ($i \in (1, 2)$). The total host sample n_{T} consists of n_{Inf} infected hosts, the infected subsample, and n_{H} non-infected, healthy, hosts, the non-infected subsample. A number of n_{Par} parasite samples is obtained from the n_{Inf} infected hosts (one sample per host) and also genotyped at each biallelic SNP. Accordingly, there are also have two parasite types for each biallelic SNP ($j \in (1, 2)$).

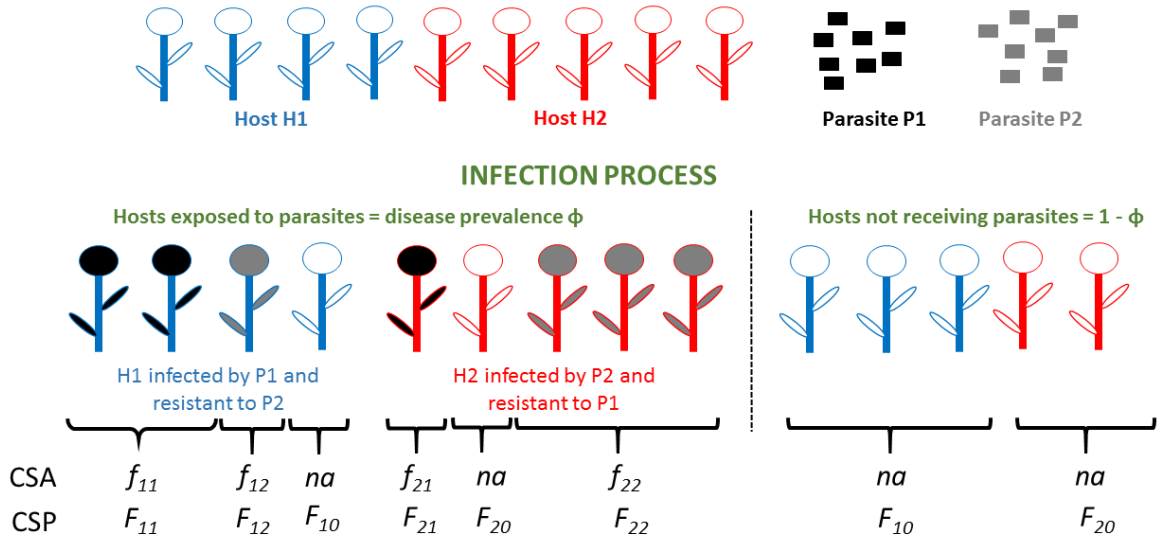


Fig. 4.1: Graphic illustration of the properties of our indices CSA and CSP. The host population consists of two host types H_1 (blue) and H_2 (red). The parasite population consists of two types P_1 (black) and P_2 (grey). A proportion ϕ of the hosts is exposed to parasites. Hosts which are exposed to the parasite either become infected or they can resist infection. Infected hosts are colored based on the identity of the infecting parasite genotype (grey or black). f_{ij} is the proportion of hosts with type i which are infected by parasites of type j in the proportion of all infected hosts. F_{ij} is the proportion of hosts of type i being infect by parasites of type j in the whole host population (sum of all hosts). F_{i0} is the proportion of non-infected hosts of type i in the whole host population. F_{i0} is composed of hosts of type i which either did not receive spores $(1 - \phi)$ or which received spores but are resistant to the respective parasite.

The Cross-Species Association index (CSA)

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

$$\text{CSA} = |f_{11}f_{22} - f_{21}f_{12}| \tag{4.5}$$

Here, f_{ij} is the number of hosts of type i being infected by a parasite of type j divided by the size of the infected subsample (n_{Inf}), so that $\sum_{\forall i,j} f_{ij} = 1$. This statistic is an adaptation of the well-known linkage disequilibrium (LD) measure in population genetics (Lewontin and Kojima 1960; Charlesworth and Charlesworth 2010, p.371-373) and related to the statistics performed in Bartha *et al.* (2013).
Following population genetics theory (Charlesworth and Charlesworth 2010, p.371-373), we normalize CSA in two different ways such that the absolute values range from 0 to 1. First, we define CSA$'$ which is obtained by normalizing each CSA value by the maximum CSA value possible, $\text{CSA}^{\max} = 0.25$. CSA reaches its maximum value when hosts of type 1 are solely infected by parasites of type 1 and hosts of type 2 are solely infected by parasites of type 2 and $f_{11} = f_{22} = 0.5$ (or when hosts of type 1 are solely infected by parasites of type 2 and hosts of type 2 are solely infected by parasites of type 1 and $f_{12} = f_{21} = 0.5$).

$$\text{CSA}' = \frac{\text{CSA}}{\text{CSA}^{\max}} = \frac{|f_{11}f_{22} - f_{21}f_{12}|}{0.25} = 4 \cdot \text{CSA} \tag{4.6}$$

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

$$\text{CSA}_{\text{r}} = \frac{f_{11}f_{22} - f_{21}f_{12}}{\sqrt{(f_{11} + f_{12})(f_{21} + f_{22})(f_{11} + f_{21})(f_{12} + f_{22})}} \tag{4.7}$$

We calculate the value of CSA at each generation g (eq. 4.5) based on our coevolutionary model 4A (eq. 4.1a):

$$\text{CSA}_g = \left| \frac{\alpha_{11}h_{1,g}p_{1,g}\alpha_{22}h_{2,g}p_{2,g} - \alpha_{21}h_{2,g}p_{1,g}\alpha_{12}h_{1,g}p_{2,g}}{\Delta^2} \right| \tag{4.8}$$

where $\Delta = \alpha_{11}h_{1,g}p_{1,g} + \alpha_{22}h_{2,g}p_{2,g} + \alpha_{21}h_{2,g}p_{1,g} + \alpha_{12}h_{1,g}p_{2,g}$ (introduced to make sure in eq. 4.5 that $\sum_{\forall i,j} f_{ij} = 1$). Note that the disease prevalence (ϕ) does drop out because the frequencies of h_1, h_2, p_1, p_2 are the same in the proportion of the population unexposed to parasites $(1-\phi)$ as the one exposed to parasites (ϕ). Therefore, $h_1 + h_2 = 1$ and $p_1 + p_2 = 1$ also holds for the hosts exposed to parasites (see Fig. 4.1).
For Model 4B, the CSA at each time step t is obtained as:

$$\mathrm{CSA}_t = \left| \frac{I_{11} \cdot I_{22} - I_{12} \cdot I_{21}}{(\sum_i \sum_j I_{ij})^2}. \right| \tag{4.9}$$

Therefore, we can compute CSA' and CSA_r at each generation based on eq. 4.8 for Model 4A and based on eq. 4.9 for Model 4B.

The Cross-Species Prevalence index (CSP)

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

$$\mathrm{CSP} = \left| \frac{F_{11} + F_{12}}{F_{10}} - \frac{F_{21} + F_{22}}{F_{20}} \right| \tag{4.10}$$

Here, F_{ij} is the proportion of host type i infected by parasite type j in the total host sample (n_{T}). At the denominator, F_{i0} is the proportion of uninfected hosts of type i in the total sample. By definition, $\frac{n_{\mathrm{Inf}}}{n_T} = F_{11} + F_{12} + F_{21} + F_{22}$, $\frac{n_T - n_{\mathrm{Inf}}}{n_T} = F_{10} + F_{20}$, and $F_{11}+F_{12}+F_{21}+F_{22}+F_{10}+F_{20} = 1$ (see Fig. 4.1). Note that F_{i0} is composed of individuals which 1) did not encounter any parasite due to the incomplete disease prevalence in the population, and 2) which got exposed to parasites but were resistant.
When eq. 4.10 is applied to our coevolutionary model 4A, CSP at each generation g is obtained as:

$$\mathrm{CSP}_g = \left| \frac{\phi_g(\alpha_{11}h_{1,g}p_{1,g} + \alpha_{12}h_{1,g}p_{2,g})}{(1-\phi_g)h_{1,g} + \phi_g((1-\alpha_{11})h_{1,g}p_{1,g} + (1-\alpha_{12})h_{1,g}p_{2,g})} - \frac{\phi_g(\alpha_{21}h_{2,g}p_{1,g} + \alpha_{22}h_{2,g}p_{2,g})}{(1-\phi_g)h_{2,g} + \phi_g((1-\alpha_{21})h_{2,g}p_{1,g} + (1-\alpha_{22})h_{2,g}p_{2,g})} \right| \tag{4.11}$$

For Model 4B, the CSP at time t is given by:

$$\mathrm{CSP}_t = \left| \frac{I_{11} + I_{12}}{S_1} - \frac{I_{21} + I_{22}}{S_2} \right| \tag{4.12}$$

Irrespective of the model, CSP is only defined as long as some hosts are not infected.

4.3.3 Detection thresholds for CSA and CSP

In order to evaluate the power of CSA and CSP to pinpoint coevolutionary loci, it is necessary to derive threshold (cut-off) values for these indices based on all possible comparisons of pairs of loci from host and parasite genome data. Any pair of host and parasite SNPs exhibiting a value above the cut-off would be considered as a strong candidate pair for governing the outcome of infection and hence, to be under coevolution. It is common to obtain these cut-off values from the distribution of all empirical values of a given data set (using ad hoc multiple testing correction). By contrast, as here we have obtained sequences from a random samples of infected and non-infected hosts from the population (natural or controlled), we can derive the expected distribution of CSA (CSA' and CSA_r, correspondingly) and CSP for pairwise comparisons of neutral host and parasite SNPs based on classic population genetic assumptions. We present the ideas underlying these cut-off calculations in a nutshell, and refer the indefatigable reader to the appendix for the detailed explanations and calculations.

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

Full genome data usually contain a large number of neutral SNPs which are distributed across the whole genome. Given that the recombination rate is high enough, these neutral SNPs evolve independently of each other and therefore, also their allele frequencies in the population and in the sample are mutually independent. In order to obtain the expected neutral distribution of CSA and CSP we have to answer the following three questions for each neutral host SNP – neutral parasite SNP comparison in the sample: 1) what is the probability p that the host minor allele frequency count in the sample is v and what is the probability q that the parasite minor allele frequency count in the sample is w, 2) given the minor allele frequency counts and the neutral infection matrix $\mathcal{A}_{\mathrm{neutral}}$ what

are possible combinations of host and parasite alleles in the sample, and 3) what is the respective CSA/CSP value for each combination of SNPs. The minor allele frequencies of all neutral SNPs in the host and parasite samples can be summarized by the so-called folded site-frequency spectrum (SFS). The expected site frequency spectra are known under the assumption of mutation-drift balance and constant population size. Namely, the expected number of SNPs with minor allele frequency k is proportional to the host or parasite population mutation rate θ. We obtain the relative SFS, *i.e.* the proportion of SNPs with minor allele frequency k in the sample, for hosts and parasites by dividing the absolute site frequency spectrum by the respective θ. Therefore, the effective population sizes and mutation rates do not factor in the calculations of the exact distributions of CSA/CSP. We obtain the distribution for different sample sizes. In the results we will present the cut-offs corresponding to the 95 and 99 high percentiles for $n_T = 200$ and $n_{\text{Inf}} = n_{\text{H}} = 100$ (but see the appendix for the cutoff values for other sampling schemes).

4.4 Results

4.4.1 Analytical results for model 4A

We first present some analytical results by computing CSA and CSP for the population genetics models with a matching-alleles and a gene-for-gene interaction to provide some intuition on the behaviour of the presented indices. In the calculations, we only focus on CSA, as it is straightforward to obtain CSA$'$ and CSA$_r$ by applying the respective normalizations.

Under the Matching Allele infection matrix

For a matching-allele infection matrix and for $c_{H_1} = c_{H_2} = c_{P_1} = c_{P_2} = 0$ the equations for model 4A (eq. 4.1a) reduce to:

$$\begin{aligned} h_{1,g+1} &= \frac{h_{1,g}\left(1-\phi_g s\left[p_{1,g}+(1-c_1)p_{2,g}\right]\right)}{\bar{w}_H}, \text{ and } h_{2,g+1} = \frac{h_{2,g}\left(1-\phi_g s\left[p_{2,g}+(1-c_2)p_{1,g}\right]\right)}{\bar{w}_H}. \\ p_{1,g+1} &= \frac{p_{1,g}\left(h_{1,g}+h_{2,g}(1-c_2)\right)}{\bar{w}_P}, \text{ and } p_{2,g+1} = \frac{p_{2,g}(h_{2,g}+h_{1,g}(1-c_1))}{\bar{w}_P}. \end{aligned} \tag{4.1}$$

By applying eq. 4.8 and eq. 4.10 to these MA-equations, we obtain $\mathrm{CSA}_{g,\mathrm{MA}}$ and $\mathrm{CSP}_{g,\mathrm{MA}}$ at any generation g.

$$\mathrm{CSA}_{g,\mathrm{MA}} = \left| \frac{(c_1 + c_2 - c_1c_2)h_{1,g}h_{2,g}p_{1,g}p_{2,g}}{\left(1 - c_2h_{2,g}p_{1,g} - c_1h_{1,g}p_{2,g}\right)^2} \right|, \tag{4.2}$$

$$\mathrm{CSP}_{g,\mathrm{MA}} = \left| \frac{\phi_g \left(c_2p_{1,g} - c_1p_{2,g}\right)}{\left(1 - \phi_g \left(1 - c_1p_{2,g}\right)\right)\left(1 - \phi_g \left(1 - c_2p_{1,g}\right)\right)} \right|. \tag{4.3}$$

It is noticeable that CSP, by contrast to the CSA, does not depend on the frequencies of the different host types but only on the parasite frequencies. Moreover, the CSP cannot be computed if the disease prevalence is at maximum ($\phi = 1$) and if neither of the host alleles provides any resistance to any parasite genotype ($c_1 = c_2 = 0$)
Further, the matching-allele model formulated in equation 4.1 has four monomorphic equilibria and one polymorphic equilibrium with frequencies given by:

$$\widehat{p}_1 = \widehat{h}_2 = \frac{c_1}{c_1 + c_2} \text{ , and } \widehat{h}_1 = \widehat{p}_2 = \frac{c_2}{c_2 + c_1}. \tag{4.4}$$

Inserting these equilibrium frequencies into eq. 4.2 and eq. 4.3 we can obtain the values of the indices at the polymorphic equilibrium point.

$$\widehat{\mathrm{CSA}}_{\mathrm{MA}} = \frac{c_1^2c_2^2}{\left(c_1 + c_2\right)^2\left(c_1 + c_2 - c_2c_1\right)} \text{ , and } \widehat{\mathrm{CSP}}_{\mathrm{MA}} = 0. \tag{4.5}$$

For a matching alleles interaction without any genotype costs, CSP is always zero at the equilibrium point, irrespective of the values of c_1 and c_2. The values of the CSA and CSP display a different behavior over time and at the equilibrium. Comparing their values over time and at the equilibrium can give a good indication about the asymmetry of the infection matrix.

Under the Gene-For-Gene infection matrix

For a gene-for-gene infection matrix and for $0 < c_{H_2}, c_{P_2} < 1$ and $c_{H_1} = c_{P_1} = 0$, the equations for the coevolutionary model 4A (eq. 4.1a) reduce to:

$$\begin{aligned} h_{1,g+1} &= \frac{h_{1,g}(1 - s\phi_g)}{\bar{w}_H} \text{ , and } h_{2,g+1} = \frac{h_{2,g}(1 - c_{H_2})\left(1 - \phi_g s\left[(1-c)p_{1,g} + p_{2,g}\right]\right)}{\bar{w}_H}, \\ p_{1,g+1} &= \frac{p_{1,g}(h_{1,g} + h_{2,g}(1-c))}{\bar{w}_P} \text{ , and } p_{2,g+1} = \frac{p_{2,g}(1 - c_{P_2})}{\bar{w}_P}. \end{aligned} \tag{4.6}$$

Applying eq. 4.8 and eq. 4.10 to the GFG system of equation, yields the following values of $\mathrm{CSA}_{g,\mathrm{GFG}}$ and $\mathrm{CSP}_{g,\mathrm{GFG}}$ at some generation g.

$$\mathrm{CSA}_{g,\mathrm{GFG}} = \frac{ch_{1,g}h_{2,g}p_{1,g}p_{2,g}}{(1-ch_{2,g}p_{1,g})^2}, \tag{4.7}$$

$$\mathrm{CSP}_{g,\mathrm{GFG}} = \frac{\phi_g c p_{1,g}}{(1-\phi_g)(1-\phi_g(1-c\cdot p_{1,g}))}. \tag{4.8}$$

As for the MA model, the CSP values, by contrast to the CSA, do not depend on the frequencies of the different host types but only on the the parasite frequencies. The conditions for computing the CSP are more restrictive than under the MA, as CSP is not defined as soon as the disease prevalence is maximum ($\phi = 1$) and therefore, all hosts of type 1 are infected.
The polymorphic equilibrium frequencies of this model 4A with GFG are given by:

$$\begin{aligned} \widehat{p}_1 &= \frac{c_{H_2}(1-s\phi)}{c\phi s(1-c_{H_2})} \quad \text{, and} \quad \widehat{h}_1 = \frac{c-c_{P_2}}{c}, \\ \widehat{p}_2 &= 1-\frac{c_{H_2}(1-s\phi)}{c\phi s(1-c_{H_2})} \quad \text{, and} \quad \widehat{h}_2 = \frac{c_{P_2}}{c}. \end{aligned} \tag{4.9}$$

Inserting these frequencies into eq. 4.7 and 4.8 we can obtain the values of the indices at the polymorphic equilibrium point.

$$\widehat{\mathrm{CSA}}_{\mathrm{GFG}} = \frac{\frac{c_{P_2}c_{H_2}\left(c-c_{P_2}\right)(1-s\phi)}{c^2\phi s\left(1-c_{H_2}\right)}\left(1-\frac{c_{H_2}(1-s\phi)}{c\phi s\left(1-c_{H_2}\right)}\right)}{\left(1-\frac{c_{P_2}c_{H_2}(1-s\phi)}{c\phi s\left(1-c_{H_2}\right)}\right)^2}, \tag{4.10}$$

$$\widehat{\mathrm{CSP}}_{\mathrm{GFG}} = \frac{c_{H_2}\left(1-\phi s\right)}{\left(1-\phi\right)\left(\left(1-\phi\right)s+c_{H_2}\left(1-s\right)\right)}. \tag{4.11}$$

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

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

When simulating an asymmetric MA interaction ($c_1 = 0.9$, $c_2 = 0.7$, model 4A) over 500 generations (Fig. 4.2) coevolution results in arms-race dynamics. We observe that CSA

and CSP fluctuate over time due to the coevolutionary cycles and the associated allele frequency changes. Overall, the CSA values decrease over time and are constantly found below the detection threshold after $g = 300$ generations. Therefore, under unstable coevolutionary dynamics with increasing amplitude and period of the coevolutionary cycles (resulting ultimately in fixation of alleles), the associations between hosts and parasites alleles in the infected sample become too weak to be observable (Fig. 4.2c,d). Under an arms-race, one host allele occurs in very high frequency and the parasite tracks this frequency down over time which generates the coevolutionary cycles. During these cycles there is only a very limited amount of time at which both hosts and parasites alleles are found in intermediate frequencies yielding high values of CSA index. On the other hand, the CSP values under the MA model are consistently high and exhibit enough statistical power to detect the loci under coevolution (Fig. 4.2d). This demonstrates the importance of obtaining additional non-infected host samples. Under the same model 4A with symmetric MA, we find similar outcome as in Fig. 4.2, albeit the oscillations of CSA and CSP values are perfectly matching the allele frequency cycles and show a regular amplitude pattern (Fig. 4.B.1).

Under the GFG model with arms race, CSA values are consistently small with narrow peaks (Fig. 4.3b,c) which are barely above the detection threshold. The CSP has several peaks above the cut-off value, yet it maybe difficult to detect the coevolutionary locus even when time samples are available if the more stringent 0.99-cutoff level is applied (Fig. 4.3d). The comparison of MA and GFG arms race dynamics shows that the combination of CSA and CSP values over time gives some indication about the symmetry of the infection matrix. Furthermore, the CSP exhibits the highest power to disentangle loci under coevolution from the neutral background but also to infer the asymmetry of the infection matrix.

Under Model 4B trench-warfare dynamics can take place for MA-interactions and allele frequencies can converge to a stable polymorphic equilibrium. Once the allele frequencies are at equilibrium, CSA has a very strong power to distinguish the coevolving loci from the neutral background, while the CSP decreases to zero (as shown above in eq. 4.5). In Model 4B, the disease prevalence varies over time as a function of the changes in allele frequencies, so that it is expected that CSP varies over time. However, once the allele frequencies reach a stable equilibrium also the numbers in all host compartments eventually remain constant. When the ratio of infected to non-infected individuals is the same for both host types then CSP drops to zero. The value of CSA at equilibrium depends on the

respective equilibrium frequencies and thus, is highest when alleles are in frequency 0.5 (eq. 4.5). Even if coevolution results in stably sustained cycles, the CSA values remain high as long as the amplitude of the cycles does not become too large and therefore 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. 4.B.2), while the CSP is fixed to zero (under a symmetric MA infection matrix). Note that the epidemiological model 4B 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. 4.B.3). In such cases, the obtained results are similar to those of Fig. 4.3 with both indices dropping to zero over time. Finally, we note that the two measures of CSA we introduce, CSA' and CSA_r show the same trend and similar power under an arms-race, while the CSA_r is slightly more precise under trench-warfare when allele frequencies reach very high or very low values (close to fixation or loss).

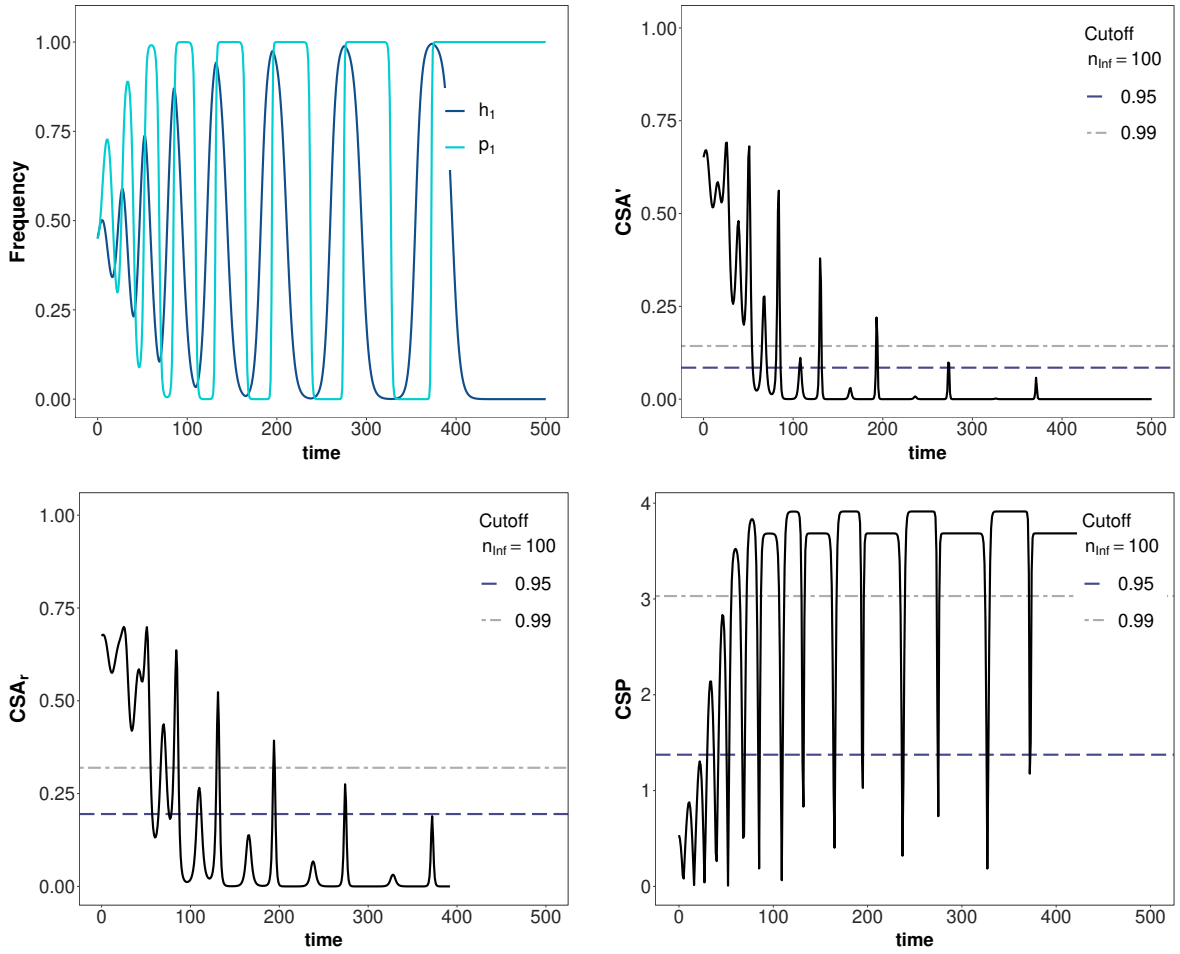


Fig. 4.2: Temporal changes in allele frequencies, $\mathrm{CSA}', \mathrm{CSA}'_r$ and CSP in an unstable asymmetric MA-model (model 4A) with one parasite generation per host generation. For each index cutoff values are shown based on the expected neutral distributions for a total host sample size $n_T = 200$ and for $n_{\mathrm{Inf}} = n_{\mathrm{H}} = 100$. The 0.95-cutoff value is shown in blue (dashed line) and the 0.99-cutoff value is shown in grey (dotted-dashed line). Top left: frequencies of h_1 (darkblue) and p_1 (lightblue). Top right: CSA'. Bottom left: CSA_r. Bottom right: CSP. The model parameters are $c_1 = 0.9$, $c_2 = 0.7$, $\phi = 0.8$, $s = 0.35$, initial values $h_{1,g=0} = p_{1,g=0} = 0.45$.

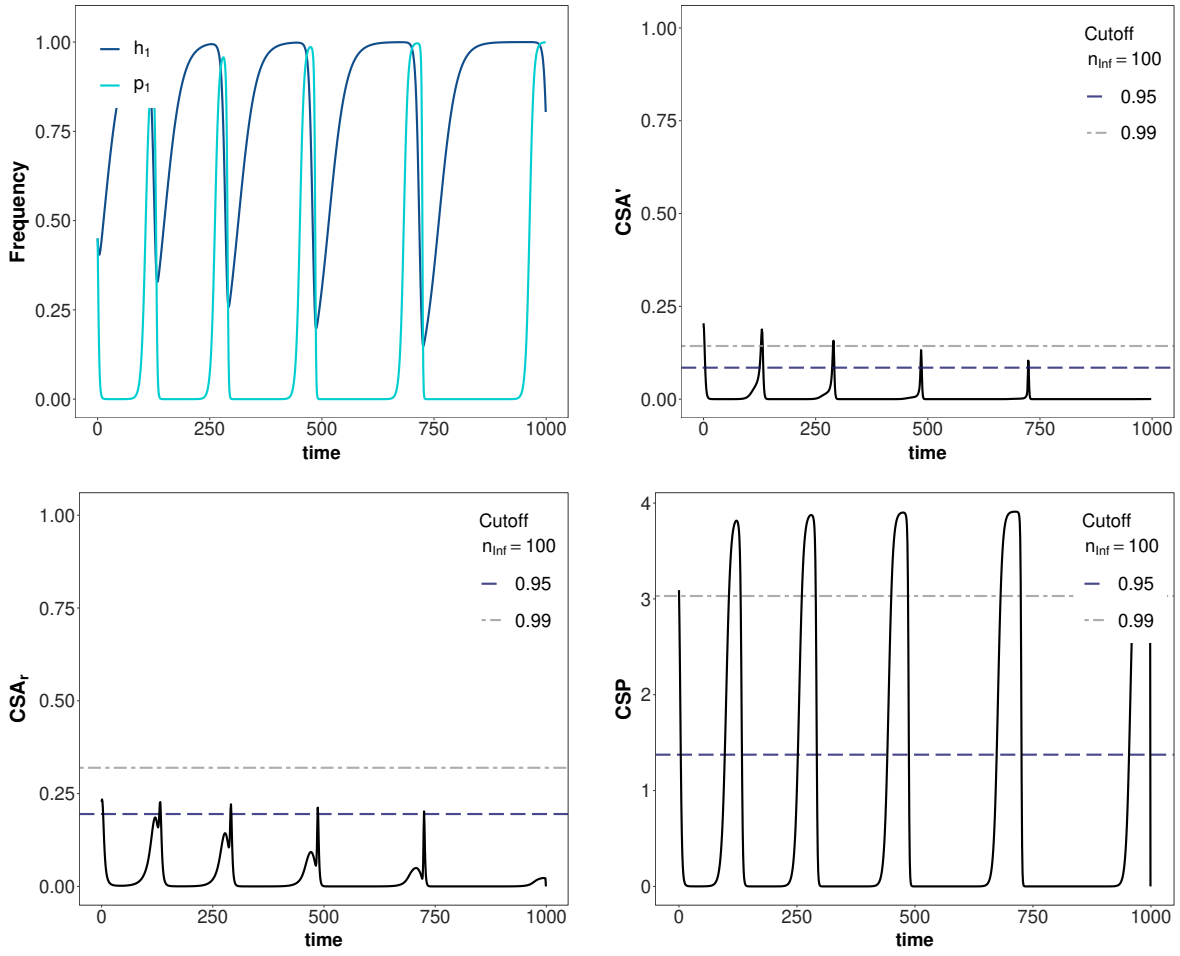


Fig. 4.3: Temporal changes in allele frequencies, $\mathrm{CSA}', \mathrm{CSA}'_r$ and CSP in an unstable GFG-model (model 4A) with one parasite generation per host generation. For each index cutoff values are shown based on the expected neutral distributions for a total host sample size $n_T = 200$ and for $n_{\mathrm{Inf}} = n_{\mathrm{H}} = 100$. The 0.95-cutoff value is shown in blue (dashed line) and the 0.99-cutoff value is shown in grey (dotted-dashed line). Top left: frequencies of h_1 (darkblue) and p_1 (lightblue). Top right: CSA'. Bottom left: CSA_r. Bottom right: CSP. The model parameters are $c_{H_1} = c_{P_1} = 0$, $c_{H_2} = 0.05$, $c_{P_2} = 0.2$, $\phi = 0.8$, $s = 0.35$, $c = 0.9$, initial values $h_{1,g=0} = p_{1,g=0} = 0.45$.

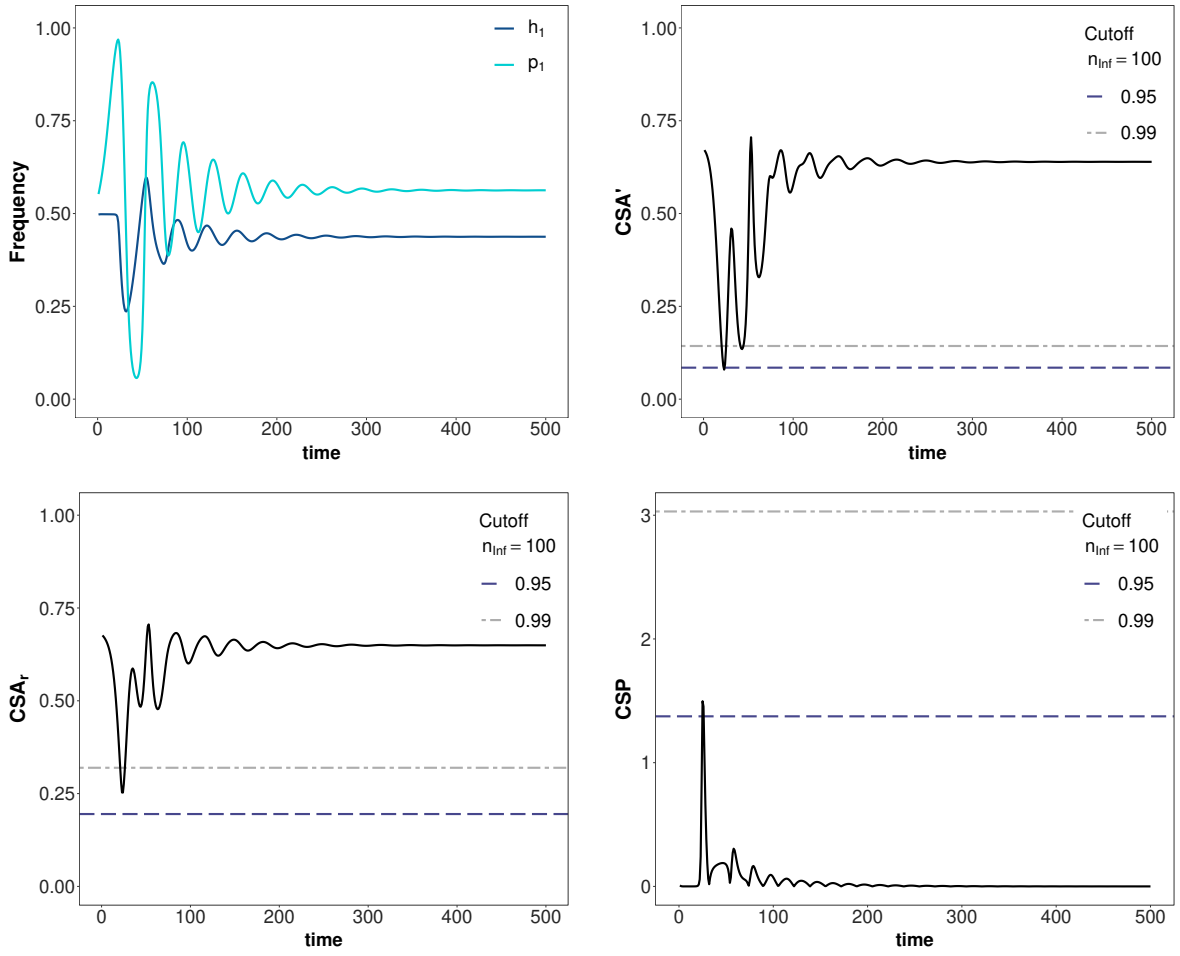


Fig. 4.4: Temporal changes in allele frequencies, $\mathrm{CSA}', \mathrm{CSA}'_r$ and CSP in an epidemiological (model 4B) with an asymmetric MA-infection matrix. For each index cutoff values are shown based on the expected neutral distributions for a total host sample size $n_T = 200$ and for $n_{\mathrm{Inf}} = n_{\mathrm{H}} = 100$. The 0.95-cutoff value is shown in blue (dashed line) and the 0.99-cutoff value is shown in grey (dotted-dashed line). Top left: frequencies of h_1 (darkblue) and p_1 (lightblue). Top right: CSA'. Bottom left: CSA_r. Bottom right: CSP. The model parameters are $\beta = 0.00005$, $s = 0.6$, $c_1 = 0.9$, $c_2 = 0.7$, $b = 1$, $d = 0.9$, and $c_{H_1} = c_{P_1} = c_{H_2} = c_{P_2} = 0$. The initial values are $S_{1,t=0} = S_{2,t=0} = 4150$, $I_{11,t=0} = I_{12,t=0} = I_{21,t=0} = I_{22,t=0} = 415$. The time intervals for computations is $\delta_t = 0.001$.

4.5 Discussion

With the technological advances, it has become feasible to sequence full genomes of several hosts from a given population as well as the parasite strains infecting them. In a previous study, all host and parasite SNPs in the samples have been compared in a pairwise manner for their degree of association (Bartha *et al.* 2013), generating a genome wide cross species association study (sensu Bartoli and Roux (2017)). We showed here, that the power of such studies can be improved if additional sequence data from non-infected hosts are available. Further, we derived cut-off values for significant associations based on simple population genetics assumptions. Finally, we demonstrated that the power to identify the loci underlying coevolution and thus, determining the infection outcome and the phenotype, varies in time and depends on the asymmetry of the underlying infection matrix.

We expected *a priori* that the power to detect coevolutionary loci varies in time due to the coevolutionary dynamics and the respective allele frequency changes at the involved loci. Our approach and results are 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. A large body of literature has shown that the power to detect coevolutionary loci based on local adaptation between the host and the parasite varies over time due to the reciprocal changes in allele frequencies (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 shows also variable statistical power over time to detect coevolution. In contrast to reciprocal transplant or common garden (Gandon and Nuismer 2009; Nuismer *et al.* 2017) and host-parasite co-GWAs (MacPherson *et al.* 2018; Wang *et al.* 2018) experiments, the study design in cross-species association studies already implicitly contains phenotypic information on the host and the parasite, as the infection experiment has already been "naturally" performed in the population. A further difference to the co-GWAs studies (MacPherson *et al.* 2018; Wang *et al.* 2018) is that the sequenced samples of infected and non-infected hosts are random samples from panmictic reproducing host and parasite populations. This allows us to apply population genetics theory to derive the expected power of our indices.

Our results indicate that the CSA and CSP are differently affected by the reciprocal changes in allele frequencies. The non-infected hosts are composed of individuals which

either did not receive spores or which received spores but are resistant. The comparison of the genotype frequencies among uninfected and infected host, thus provides additional information about the distribution of allele frequencies in the non-infected and the infected population subsample. An implicit assumption in Bartha *et al.* (2013), Bartoli and Roux (2017) and our model is that disease transmission is random and panmictic so that potentially every host can get in contact with the disease (no population sub-structuring affecting disease transmission). Thus, it is crucial to assess the extent of population structure before performing a cross-species association study as different populations can be at different stages of the coevolutionary cycle (Gavrilets and Michalakis 2008).
As indicated by analytical and simulation results, the two indices we introduce, CSA and CSP, provide different information regarding 1) the symmetry of the infection matrix, 2) the type of dynamics, and 3) whether the allele frequencies have reached a stable polymorphic equilibrium point. Rather than focusing on the limitations of these indices and cross-species association studies, we focus here on how this information can be potentially used to draw inference about the three above mentioned coevolutionary characteristics. Our results suggest that obtaining samples from several time points is likely to give more accurate inference results than samples from a single time point. First, in order to infer the infection matrix, we can see that two time samples for a MA or GFG interaction with an arms-race would likely both yield low values of CSA, while the value of CSP value would be comparatively high for a MA interaction and low for a GFG interaction. Additionally, time samples from several time-points which are dense enough to capture the finer patterns of regular peak behaviour can be informative about the asymmetry in a MA-alleles infection matrices. Second, two or few time samples would be enough to allow inference of the type of coevolutionary dynamics, namely arms-race vs. trench-warfare. A similar idea was proposed in Gandon *et al.* (2008) for the study of local adaptation, when using phenotypic (infection) data. In our case, we see that increasing amplitudes of allele frequencies (arms-race) leads to a decrease of both CSA and CSP values over time, while decreasing amplitudes (trench-warfare dynamics) of coevolutionary cycles generate increasing CSA and CSP values. Under perfectly stable allele frequency dynamics with a given amplitude and period, the CSA and CSP would oscillate regularly as well, with their mean value depending on the amplitude of the coevolutionary cycles. If the cycling amplitudes are large and close to the boundaries, the CSA and CSP values will be small. Note that our population genetics model is based on the assumption of infinite population size. However, realistic models with finite population size predict that fixation of

alleles (arms-race dynamics) is likely to occur due to the effect of genetic drift when their frequencies are close to the boundary (Tellier *et al.* 2014; Gokhale *et al.* 2013). So the power of performing CSA/CSP measures should be higher in populations with large size (approx. higher than 5,000 hosts and parasites, (Tellier *et al.* 2014)) Third, it follows once a stable polymorphic equilibrium is reached, few time points will reveal constant or close to constant values of the CSA and CSP. Therefore, the CSA and CSP indices could be also introduced as summary statistic into the inference approach presented in chapter 5 in order to estimate the values of the coevolutionary parameters.

One crucial result of our study is the derivation of the neutral expectations for CSA and CSP which allows us to compute detection thresholds for different sample sizes. To do so, we assume that host and parasite populations are at drift-mutation equilibrium, so the expected distribution of neutral allele frequencies is given by a neutral SFS under the assumption of constant population size. However, host and parasite population sizes both change over time to 1) eco-evolutionary feedback arising from the epidemiological dynamics (as for example in model 4B), and/or 2) due to abiotic environmental factors such as resource availability and habitat suitability. Irrespective of the source causing the respective population size changes, it is possible to compute the neutral SFS for both scenarios based on previous work (Živković *et al.* 2015, 2019). Further, the influence of the available sample sizes on cut-off values can be large. Small sample sizes ($n_T < 25$) decrease substantially the power to detect the coevolving loci.

One simplifying assumption in our approach is the strict one to one relationship between the host and the parasite. This also implies that one parasite sample is obtained per host. However, there is evidence that for many diseases co-infections are common (Alizon *et al.* 2013; Tollenaere *et al.* 2016). In such cases, a solution would be to use only the major strain of parasite found per host individual for sequencing as in Bartha *et al.* (2013). Finally, we also assume a simple single biallelic locus model, haploid host and haploid parasite coevolutionary model. One can apply the tests and indices independently to each locus in the host genome and each parasite locus. As long as few freely recombining genes with major effects determine the outcome of the interaction the approach should be still suited to capture the loci under coevolution assuming no pleiotropic or epistatic effects between these genes. The biallelic locus should be understood here as the simple case of two alleles at a single SNP but also applies to presence/absence polymorphism of a gene (or gene domain). In general, the approach presented here is also potentially applicable to detect the major genes which are determining the compatibility between

symbionts in mutualistic interactions. It is straightforward to study the power of the presented indices for such types of interaction by adjusting the equations governing the coevolutionary dynamics accordingly.

We conclude by presenting a set of recommendations for applying this method to different host-parasite systems. It is advised to obtain infected and non-infected hosts and parasite random samples at several time points from natural populations or from controlled experiments. The cross-species association method has a high power when applied to parasite with strong life-cycle dependency on their hosts, in contrast to generalists, non-biotrophic or parasites with complex life-cycles with stages on different hosts. Further, polycyclic parasites, that is parasites with several infection cycles/generations per host generation, track down the host frequencies within a host generation (Brown and Tellier 2011). This effect should be strongly pronounced for parasites with much shorter generation time than their hosts (viruses (Bartha *et al.* 2013) or bacteria). For such types of parasites taking serial samples within a single host generations should help to pinpoint coevolutionary loci very well. These loci are expected to show an increasing association with the corresponding host loci over the course of a single host generation.

A difficulty which can arise in viruses or bacteria is the absence of recombination within the genome or that recombination rates are very low. Thus, the assumption of independence between the neutral and coevolving loci is violated. However, our computations for obtaining threshold values for the cross species association indices rely on recombination in order to disentangle candidate loci from neutral background. In Bartha *et al.* (2013), this hurdle was overcome by first performing a phylogeny of the virus samples, and then clusters of polymorphic SNPs could be identified across the phylogeny. Based on these clusters the association study was performed in a second step.

For parasites of annual plants or of invertebrate hosts with short life spans, it is expected that temporal samples between years can yield enough power to detect the coevolutionary loci by comparing indices values between time samples. In the case of species with dormant stages (seed or egg banks), an adjustment of the neutral SFS computation may be necessary per time sample.

An alternative solution to sequence data from several time points could be the use of sequence data from several host and parasite populations which are connected by low migration rates. If data from several populations are to be used, we suggest to compute the CSA/CSP indices per population as well as based on pooling all samples together. However, note that the neutral SFS of pooled samples from a spatially structured population

does not follow the equations we used here. A description of the effect of spatial population structure on allele frequency distributions can be for example found in Wakeley and Aliacar (2002) and Staedler *et al.* (2009).

Supplementary information

4.A Cross species summary statistics for neutral loci

Our aim is to derive the expectations and confidence intervals for both statistics, CSA and CSP, when measuring associations between a random neutral host and parasite locus. Neutral loci are defined as SNPs without any influence on the infection outcome. Thus, they segregate independently of the locus under coevolution. Put in a different way, interactions between neutral SNPs can be seen as interactions where all entries α_{ij} in the infection matrix $\mathcal{A}$ are $\alpha_{ij} = 1$. We assume that the allele frequency distribution of neutral SNPs in both, the host and the parasite, follows the site frequency spectrum (SFS) under drift-mutation equilibrium for a Wright-Fisher model. Further, we assume that there is no outgroup sequence available, thus the ancestral and derived state are unknown for a given SNP. The expected folded SFS $\eta = \{\eta_1, \ldots, \eta_{\lfloor n/2 \rfloor}\}$ under drift-mutation equilibrium for a sample of size n is given by:

$$\eta_k = \frac{\frac{\theta}{k} + \frac{\theta}{n-k}}{1 + \delta_{k,n-k}} \quad \text{for } 1 \leq k \leq \lfloor n/2 \rfloor \tag{4.1}$$

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

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

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

$$p_k = \frac{\eta_k}{\sum_{i=1}^{\lfloor n/2 \rfloor} \eta_i} = \frac{\left(\frac{\frac{1}{k} + \frac{1}{n-k}}{1 + \delta_{k,n-k}} \right)}{\sum_{i=1}^{\lfloor n/2 \rfloor} \left(\frac{\frac{1}{k} + \frac{1}{n-k}}{1 + \delta_{k,n-k}} \right)} \tag{4.2}$$

Our computations include singletons, that is alleles with frequency $1/n$ in the sample. However, it is known that the sequencing and detection of singletons can be biased (*e.g.* with NGS technologies or pooling of samples). Therefore, singletons can be also removed from the CSA calculation and the CSP calculations should be adjusted accordingly by constraining the minor allele frequency in the infected subsample to be at least equal to two.

4.A.1 Cross species association index (CSA)

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

$$\begin{aligned}
E(CSA_{vw}) &= \Omega_{vw} \sum_{k=0}^{l} \frac{\binom{v}{k}\binom{n_{\text{Inf}}-v}{w-k}}{\binom{n_{\text{Inf}}}{w}} CSA_{vw,k} && (4.3)\\
&= \Omega_{vw} \sum_{k=0}^{l} \frac{\binom{v}{k}\binom{n_{\text{Inf}}-v}{w-k}}{\binom{n_{\text{Inf}}}{w}} \left(\left| \frac{k}{n_{\text{Inf}}} \cdot \frac{n_{\text{Inf}} - v - (w-k)}{n_{\text{Inf}}} - \frac{v-k}{n_{\text{Inf}}} \cdot \frac{w-k}{n_{\text{Inf}}} \right| \right) && (4.4)\\
&= \Omega_{vw} \sum_{k=0}^{l} \frac{\binom{v}{k}\binom{n-v}{w-k}}{\binom{n_{\text{Inf}}}{w}} \left| \frac{k n_{\text{Inf}} - vw}{n_{\text{Inf}}^2} \right| && (4.5)
\end{aligned}$$

where $l = min(v, w)$.
Here the index k can be interpreted as the number of hosts with the minor allele which are infected by a parasite with the minor allele, or put in a different way, $w - k$ out of the $n_{\text{Inf}} - v$ hosts with the major allele are infected by a parasite with the minor allele. Accordingly, $v - k$ hosts with the minor allele are infected by a parasite which has the major allele, and $n_{\text{Inf}} - v - (w - k)$ hosts with the major allele are infected by parasites with the major allele. We define Ω_{vw} as the normalization for either obtaining CSA$'$ (with $\Omega_{vw} = 4$) or CSA$_r$ (with $\Omega_{vw} = \frac{1}{\sqrt{\frac{v}{n_{\text{Inf}}}\frac{n-v}{n_{\text{Inf}}}\frac{w}{n_{\text{Inf}}}\frac{n-w}{n_{\text{Inf}}}}}$).
Second, to calculate the expectation of CSA over all pairwise comparisons of neutral host and parasite SNPs, we have to weight each $E(\text{CSA}_{vw})$ value by the probability that a neutral host SNP has minor allele frequency v (p_v) and a neutral parasite SNP has minor allele frequency w (q_w) in the sample. These probabilities can be obtained by using eq. 4.2. Therefore, the expected CSA value for a randomly chosen pair of neutral host and parasite SNPs is given by:

$$E(CSA) = \sum_{v=1}^{\lfloor n/2 \rfloor} \sum_{w=1}^{\lfloor n/2 \rfloor} p_v q_w E(CSA_{vw}) \tag{4.6}$$

4.A.2 Distribution of CSA for different sample sizes

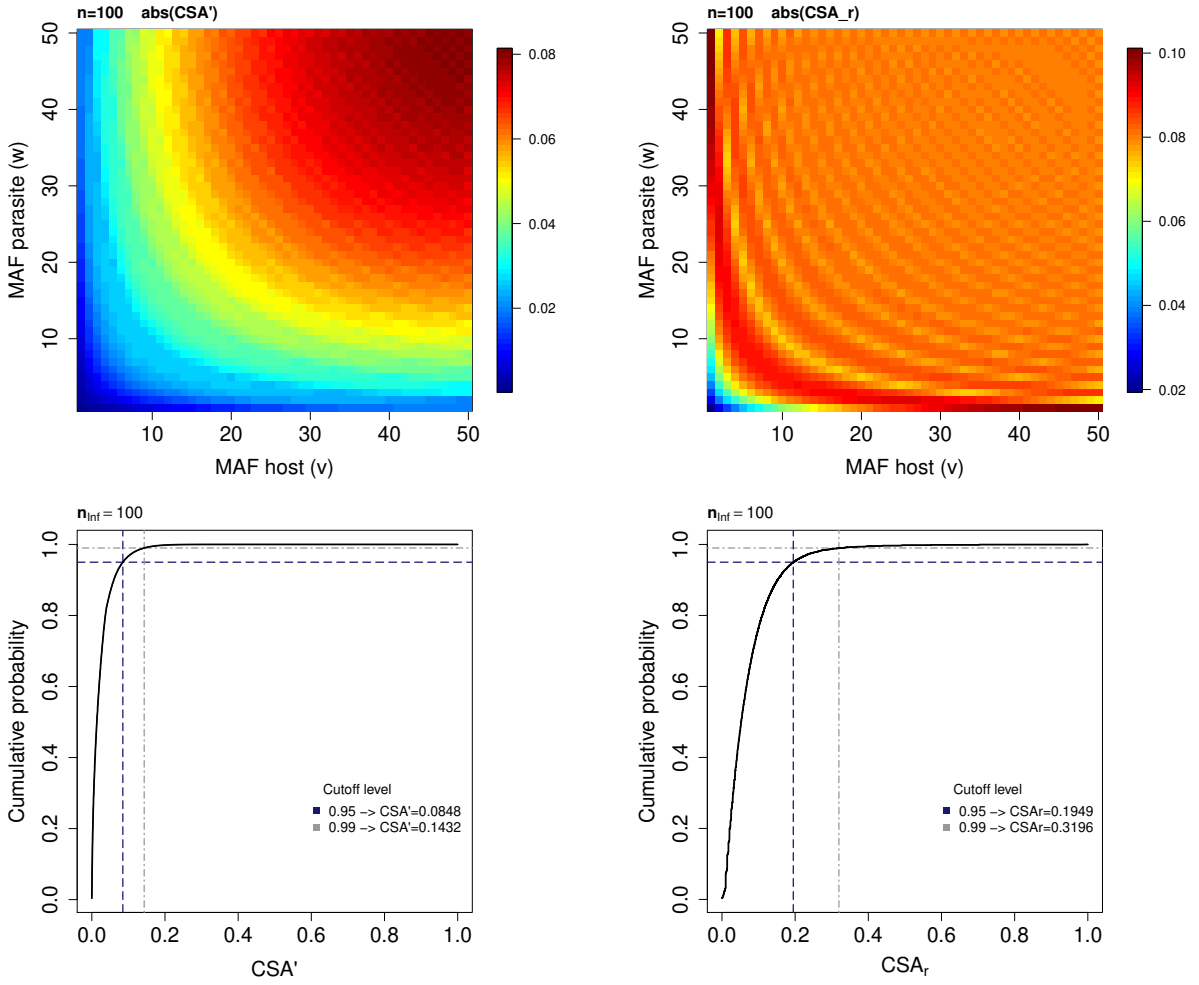


Fig. 4.A.1: Expected values of CSA′ (top left) and CSA_r (top right) when comparing all neutral host SNPs with minor allele frequency v ($v \in \{1, \ldots, \lfloor n_{\text{Inf}}/2 \rfloor\}$) to all neutral parasite SNPs with minor allele frequency w ($w \in \{1, \ldots, \lfloor n_{\text{Par}}/2 \rfloor\}$ and the resulting expected cumulative distribution function of E(CSA') (bottom left) and E(CSA_r) (bottom right) for a sample size of $n_{\text{Inf}} = n_{\text{Par}} = 100$.

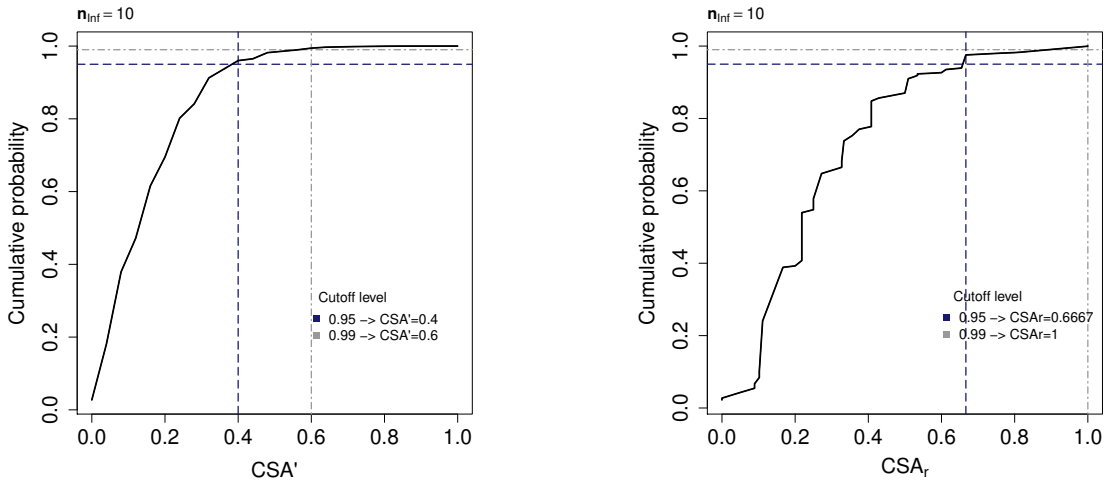


Fig. 4.A.2: Expected cumulative distribution function of CSA' (left) and CSA_r (right) when comparing all neutral host SNPs with minor allele frequency v ($v \in \{1, \ldots, \lfloor n/2 \rfloor\}$) to all neutral parasite SNPs with minor allele frequency w ($w \in \{1, \ldots, \lfloor n/2 \rfloor\}$ for a sample size of $n_{\mathrm{Inf}} = n_{\mathrm{Par}} = 10$.

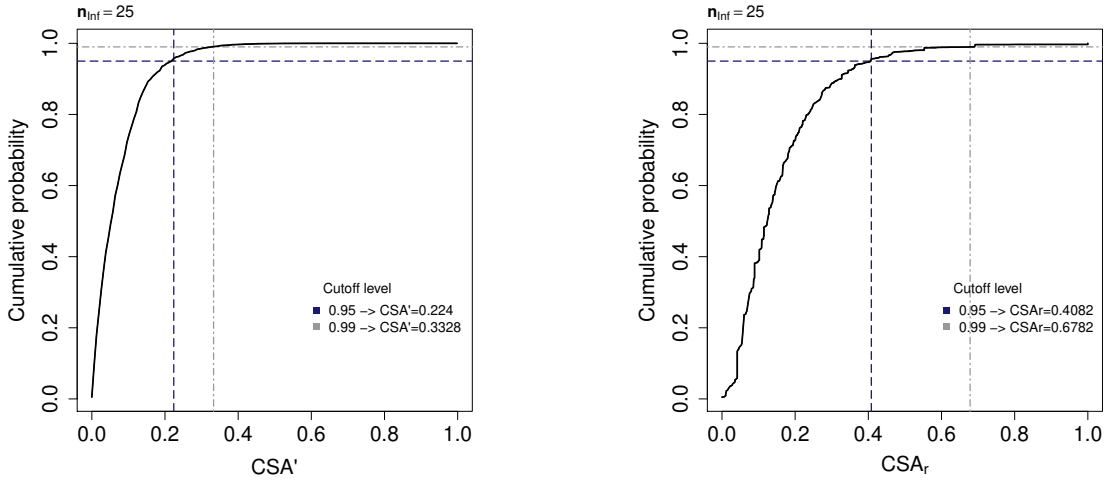


Fig. 4.A.3: Expected cumulative distribution function of CSA' (left) and CSA_r (right) when comparing all neutral host SNPs with minor allele frequency v ($v \in \{1, \ldots, \lfloor n/2 \rfloor\}$) to all neutral parasite SNPs with minor allele frequency w ($w \in \{1, \ldots, \lfloor n/2 \rfloor\}$ for a sample size of $n_{\mathrm{Inf}} = n_{\mathrm{Par}} = 25$.

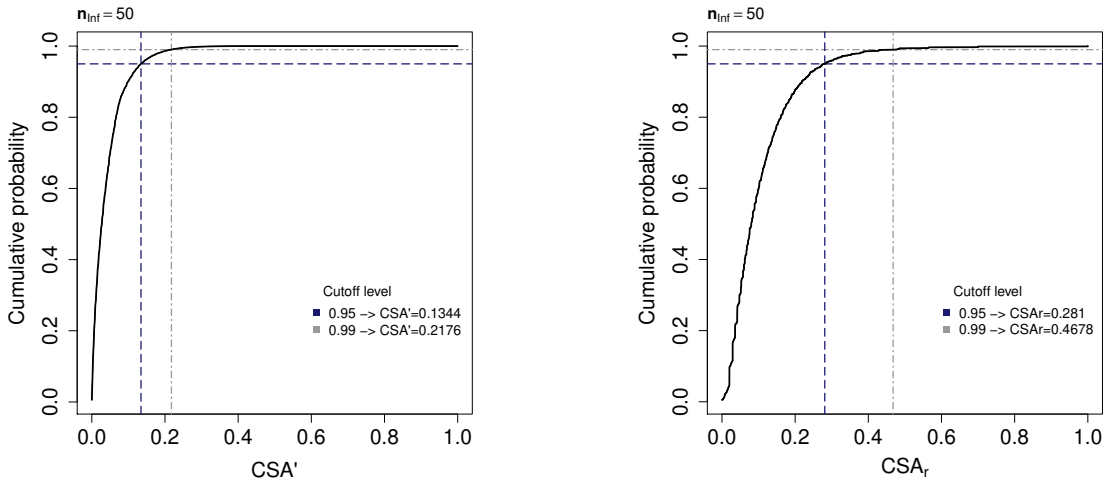


Fig. 4.A.4: Expected cumulative distribution function of CSA′ (left) and CSA_r (right) when comparing all neutral host SNPs with minor allele frequency v ($v \in \{1, \ldots, \lfloor n/2 \rfloor\}$) to all neutral parasite SNPs with minor allele frequency w ($w \in \{1, \ldots, \lfloor n/2 \rfloor\}$ for a sample size of $n_{\mathrm{Inf}} = n_{\mathrm{Par}} = 50$.

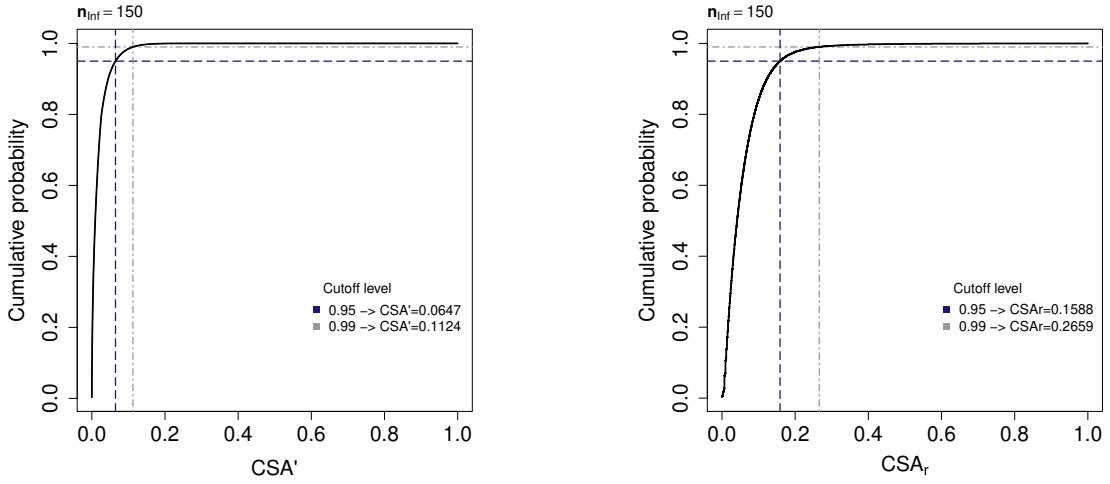


Fig. 4.A.5: Expected cumulative distribution function of CSA′ (left) and CSA_r (right) when comparing all neutral host SNPs with minor allele frequency v ($v \in \{1, \ldots, \lfloor n/2 \rfloor\}$) to all neutral parasite SNPs with minor allele frequency w ($w \in \{1, \ldots, \lfloor n/2 \rfloor\}$ for a sample size of $n_{\mathrm{Inf}} = n_{\mathrm{Par}} = 150$.

4.A.3 Cross species prevalence index (CSP)

We label the host allele with minor frequency in the **infected subsample** as $i = 1$ and the host allele with major frequency in the infected subsample as $i = 2$. Note that the allele with minor allele frequency in the infected subsample is not necessarily the minor allele

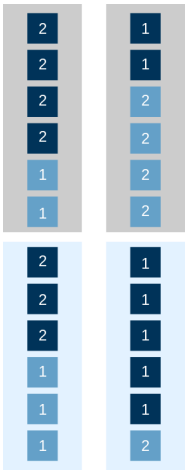


Fig. 4.A.6: Two possible host configurations when sampling a total number $n_{\mathrm{T}} = 12$ host individuals among which $n_{\mathrm{Inf}} = 6$ individuals are infected (grey box) and $n_{\mathrm{H}} = 6$ individuals are healthy (lightblue box) and the minor host allele frequency is $v = 5$. Host individuals which have the minor allele (based on the whole sample) are shown in lightblue, host individuals with the major allele (based on the whole sample) are shown in darkblue. Labelling of the alleles for the calculation of CSP is based on the minor allele frequency in the infected subsample. On the left, the minor allele is labelled by 1 as it is also the minor allele in the infected subsample. On the right, the major allele of the total sample is labelled by 1 as it represents the allele with minor allele frequency in the infected subsample.

in the whole sample (see Fig. 4.A.6). In cases where both alleles have equal frequencies in the infected subsample, the allele with minor allele frequency in the whole sample will be labelled as 1 and the allele with major allele frequency in the whole sample will be labelled as 2. Therefore, F_{11} (F_{12}) is the proportion of hosts with label 1 which are infected by a parasite with the minor (major) allele. F_{21} (F_{22}) is the proportion of hosts with label 2 which are infected by a parasite with the minor (major) allele. Further, F_{10} (respectively F_{20}) is the proportion of non-infected hosts carrying allele 1 (respectively 2). If for a neutral locus there are v minor alleles in the total host sample n_T, these minor alleles can be found with equal probability on each of the n_{T} individuals (irrespective of the infection status) as a neutral SNP does not have an effect on the infection outcome. Similarly, all of the w parasite minor alleles can be randomly assigned to any of the n_{par} parasite individuals which are infecting the n_{Inf} host individuals. Further, note that CSP is only informative when the minor and major allele can be found in both, the infected and the non-infected subsample. Therefore, we exclude SNPs which are singletons in the total host sample (n_{T}). We proceed as follows to obtain the expected CSP for a neutral host SNP with minor allele frequency v and and neutral parasite SNP with minor allele frequency w.

First, we have to find all host combinations (and their probability) for which the minor and major host alleles are found in both the infected and non-infected subsamples. We define z as the number of minor host alleles which are found in the infected subsample for a given combination. Accordingly, the number of major host alleles in the infected subsample is $n_{\text{Inf}}-z$, the number of minor host alleles in the non-infected subsample is $v-z$ and the number of major host alleles in the non-infected subsample is $n_T - n_{\text{Inf}} - (v - z)$. Based on the resulting composition of the infected subsample the alleles are labeled. The indicator variable λ is used to keep track of whether the minor allele in the total sample is the minor ($\lambda = 0$) or the major ($\lambda = 1$) allele in the infected subsample.
Second, the n_{par} parasites among which w individuals have the minor parasite allele are assigned within the n_{Inf} sample. Hereby, k denotes the number of hosts with label 1 which are infected by a parasite with the minor allele (see CSA). Thus, the expected value of CSP for a SNP with minor allele frequency v in the host and minor allele frequency w in the parasite is given by:

$$E(CSP_{vw}) = \sum_{z=\rho}^{m-1} \frac{\binom{n_{\mathrm{Inf}}}{z}\binom{n_H}{v-z}}{\binom{n_{\mathrm{T}}}{v} - \sum_{b=0}^{\rho-1}\binom{n_{\mathrm{Inf}}}{b}\binom{n_H}{v-b} - \sum_{b=m}^{v}\binom{n_{\mathrm{Inf}}}{b}\binom{n_H}{v-b}} \sum_{k=0}^{min(w,a)} \frac{\binom{a}{k}\binom{n_{\mathrm{Inf}}-a}{w-k}}{\binom{n_{\mathrm{Inf}}}{w}} \left| \frac{\frac{k}{n_{\mathrm{T}}} + \frac{a-k}{n_{\mathrm{T}}}}{\frac{\lambda n_H + (v-z)(-1)^{\lambda}}{n_{\mathrm{T}}}} - \frac{\frac{w-k}{n_{\mathrm{T}}} + \frac{n_{\mathrm{Inf}}-a-(w-k)}{n_{\mathrm{T}}}}{\frac{(1-\lambda)n_H - (v-z)(-1)^{\lambda}}{n_{\mathrm{T}}}} \right| \tag{4.7}$$

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

$$E(CSP_{vw}) = \sum_{z=\rho}^{m-1} \frac{\binom{n_{\mathrm{Inf}}}{z}\binom{n_H}{v-z}}{\binom{n_{\mathrm{T}}}{v} - \sum_{b=0}^{\rho-1}\binom{n_{\mathrm{Inf}}}{b}\binom{n_H}{v-b} - \sum_{b=m}^{v}\binom{n_{\mathrm{Inf}}}{b}\binom{n_H}{v-b}} \left| \frac{a}{\lambda n_H + (v-z)(-1)^{\lambda}} - \frac{n_{\mathrm{Inf}} - a}{(1-\lambda)n_H - (v-z)(-1)^{\lambda}} \right| \tag{4.8}$$

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

$$\lambda = \begin{cases} 0 & \text{for } z \leq n_{\mathrm{Inf}} - z \\ 1 & \text{for } z > n_{\mathrm{Inf}} - z \end{cases} \tag{4.9}$$

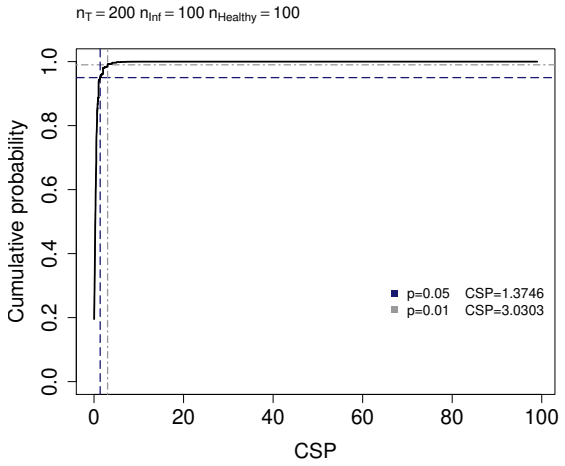


Fig. 4.A.7: Cumulative distribution function of the expected value of CSP when taking a host sample of total size $n_{\mathrm{T}} = 200$ which includes $n_{\mathrm{Inf}} = 100$ infected hosts and $n_{\mathrm{H}} = 100$ healthy hosts.

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

Tab. 4.A.1: Parameter values used to simulate the data for evaluating the potential of CSA and CSP to distinguish between neutral and coevolving loci

	Infection matrix	**unstable popgen**	**epidemiological**
MA	$\begin{pmatrix} 1 & 1-c \\ 1-c & 1 \end{pmatrix}$	$c = 0.9$, $\phi = 0.8$, $c_{H_1} = 0$, $c_{H_2} = 0$, $c_{P_1} = 0$, $c_{P_2} = 0$, $s = 0.35$, $p_1 = 0.45$, $h_1 = 0.45$	$c = 0.9$, $c_{H_1} = 0$, $c_{H_2} = 0$, $c_{P_1} = 0$, $c_{P_2} = 0$, $s = 1\&0.6$, $S_1 = S_2 = 4150$, $I_{11} = I_{12} = I_{21} = I_{22} = 415$, $\beta = 0.00005$, $b = 1$, $d = 0.9$, $\delta_t = 0.001$
MA asymmetric	$\begin{pmatrix} 1 & 1-c_1 \\ 1-c_2 & 1 \end{pmatrix}$	$c_1 = 0.9$, $c_2 = 0.7$, $\phi = 0.8$, $c_{H_1} = 0$, $c_{H_2} = 0$, $c_{P_1} = 0$, $c_{P_2} = 0$, $s = 0.35$, $p_1 = 0.45$, $h_1 = 0.45$	$c_1 = 0.9$, $c_2 = 0.7$, $c_{H_1} = 0$, $c_{H_2} = 0$, $c_{P_1} = 0$, $c_{P_2} = 0$, $s = 1\&0.6$, $S_1 = S_2 = 4150$, $I_{11} = I_{12} = I_{21} = I_{22} = 415$, $\beta = 0.00005$, $b = 1$, $d = 0.9$, $\delta_t = 0.001$
GFG	$\begin{pmatrix} 1 & 1 \\ 1-c & 1 \end{pmatrix}$	$c = 0.9$, $\phi = 0.8$, $c_{H_1} = 0$, $c_{H_2} = 0.05$, $c_{P_1} = 0$, $c_{P_2} = 0.2$, $s = 0.35$, $p_1 = 0.45$, $h_1 = 0.45$	$c = 0.9$, $c_{H_1} = 0$, $c_{H_2} = 0.05$, $c_{P_1} = 0$, $c_{P_2} = 0.05$, $s = 1\&0.6$, $S_1 = S_2 = 4150$, $I_{11} = I_{12} = I_{21} = I_{22} = 415$, $\beta = 0.00005$, $b = 1$, $d = 0.9$, $\delta_t = 0.001$

4.B Supplementary figures

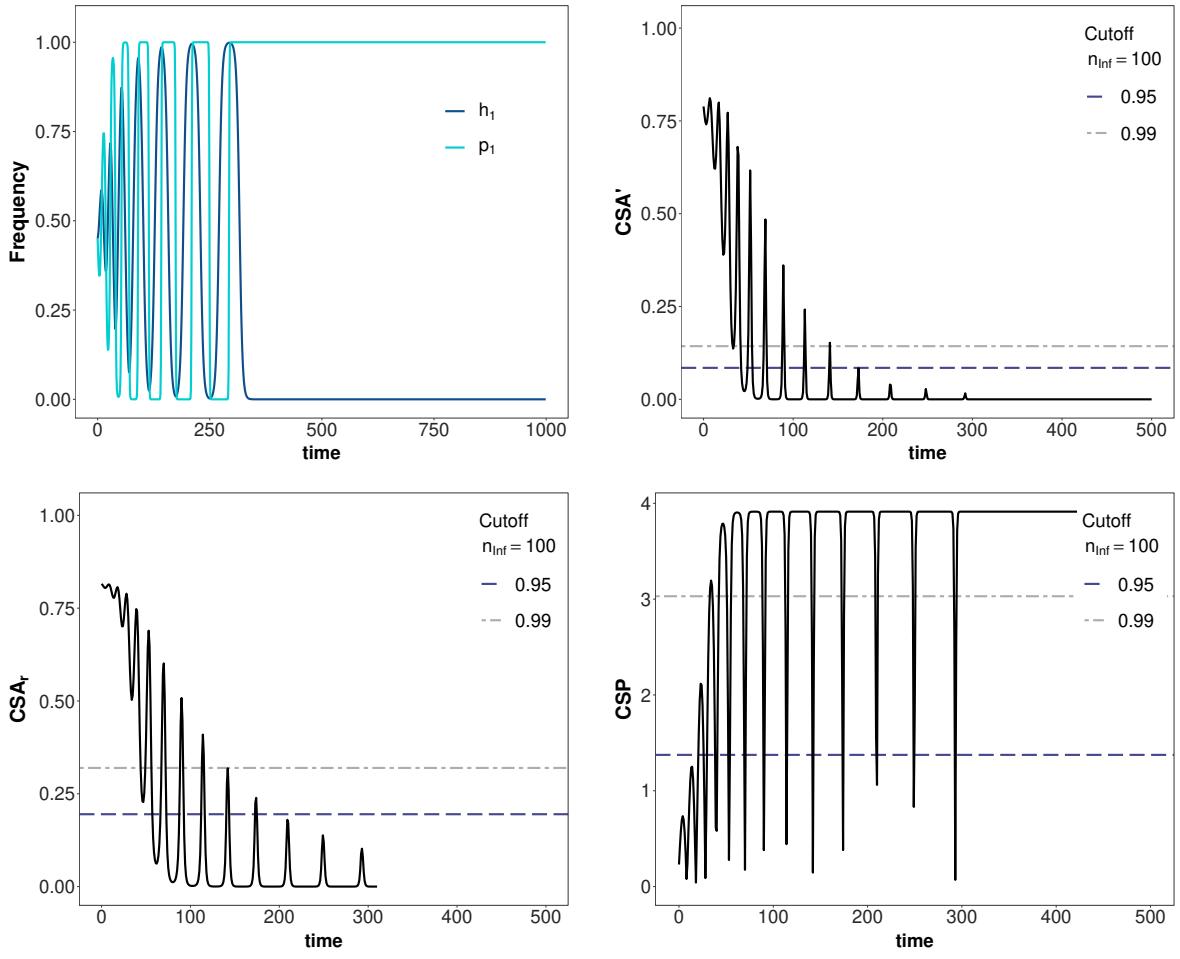


Fig. 4.B.1: Temporal changes in allele frequencies, CSA′, CSA_r and CSP in an unstable MA-model (model 4A) with one parasite generation per host generation. For each index cutoff values are shown based on the expected neutral distributions for a total host sample size $n_T = 200$ and for $n_{\mathrm{Inf}} = n_{\mathrm{H}} = 100$. The 0.95-cutoff value is shown in blue (dashed line) and the 0.99-cutoff value is shown in grey (dotted-dashed line). Top left: frequencies of h_1 (darkblue) and p_1 (lightblue). Top right: CSA′. Bottom left: CSA_r. Bottom right: CSP. The parameters values of the model are: $c_1 = c_2 = 0.9$, $c_{H_1} = c_{P_1} = c_{H_2} = c_{P_2} = 0$, $\phi = 0.8$, $s = 0.35$, $h_{1,\mathrm{init}} = p_{1,\mathrm{init}} = 0.45$

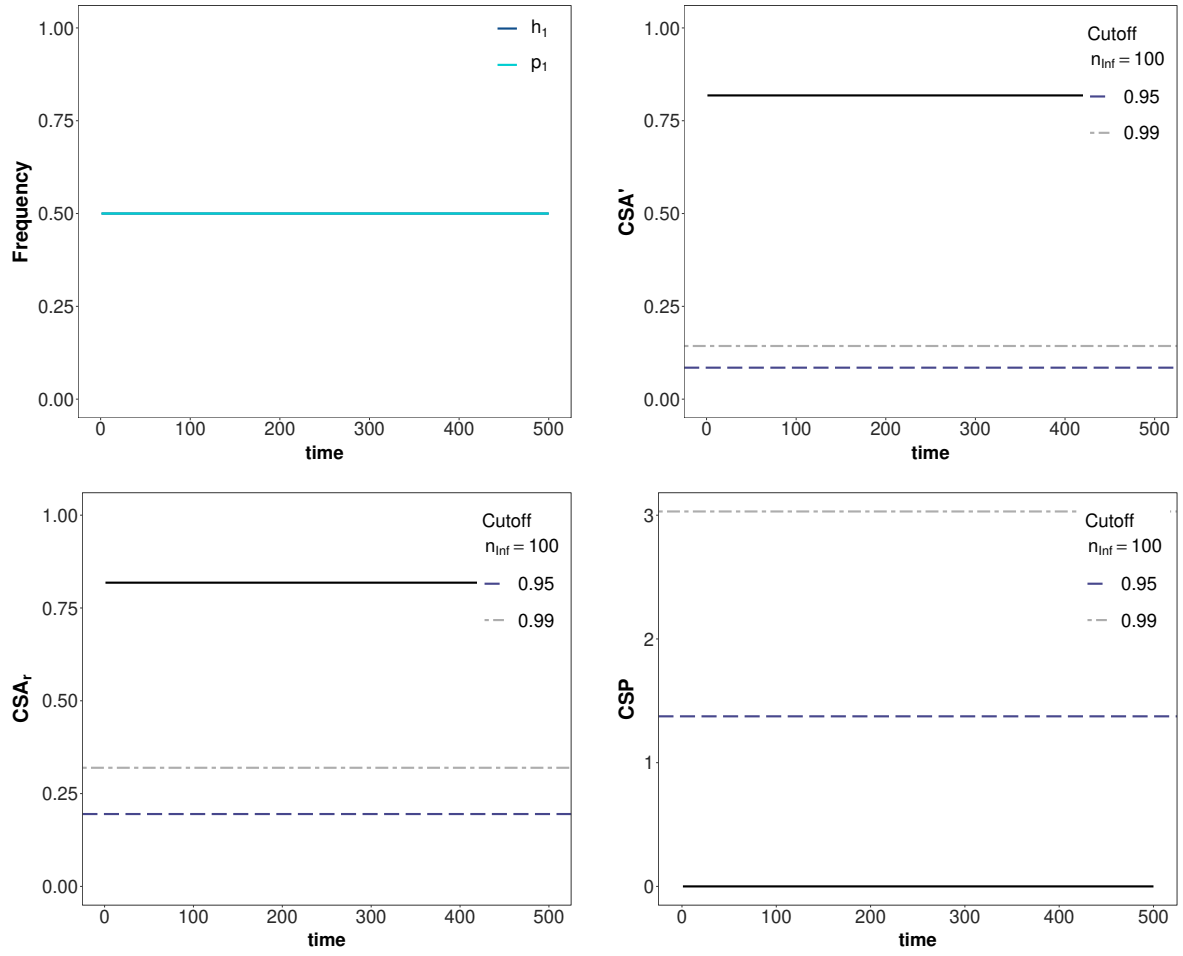


Fig. 4.B.2: Temporal changes in allele frequencies, CSA′, CSA_r and CSP in an epidemiological model (model 4B) with a symmetric MA-infection matrix. For each index cutoff values are shown based on the expected neutral distributions for a total host sample size $n_T = 200$ and for $n_{\mathrm{Inf}} = n_{\mathrm{H}} = 100$. The 0.95-cutoff value is shown in blue (dashed line) and the 0.99-cutoff value is shown in grey (dotted-dashed line). Top left: frequencies of h_1 (darkblue) and p_1 (lightblue). Top right: CSA′. Bottom left: CSA_r. Bottom right: CSP. The parameters values of the model are: $c_{H_1} = c_{P_1} = c_{H_2} = c_{P_2} = 0$, $\beta = 0.00005$, $s = 0.6$, $c_1 = c_2 = 0.9$, $S_{1,\mathrm{init}} = S_{2,\mathrm{init}} = 4150$, $I_{11} = I_{12} = I_{21} = I_{22} = 415$, $\delta_t = 0.001$, $b = 1$, $d = 0.9$

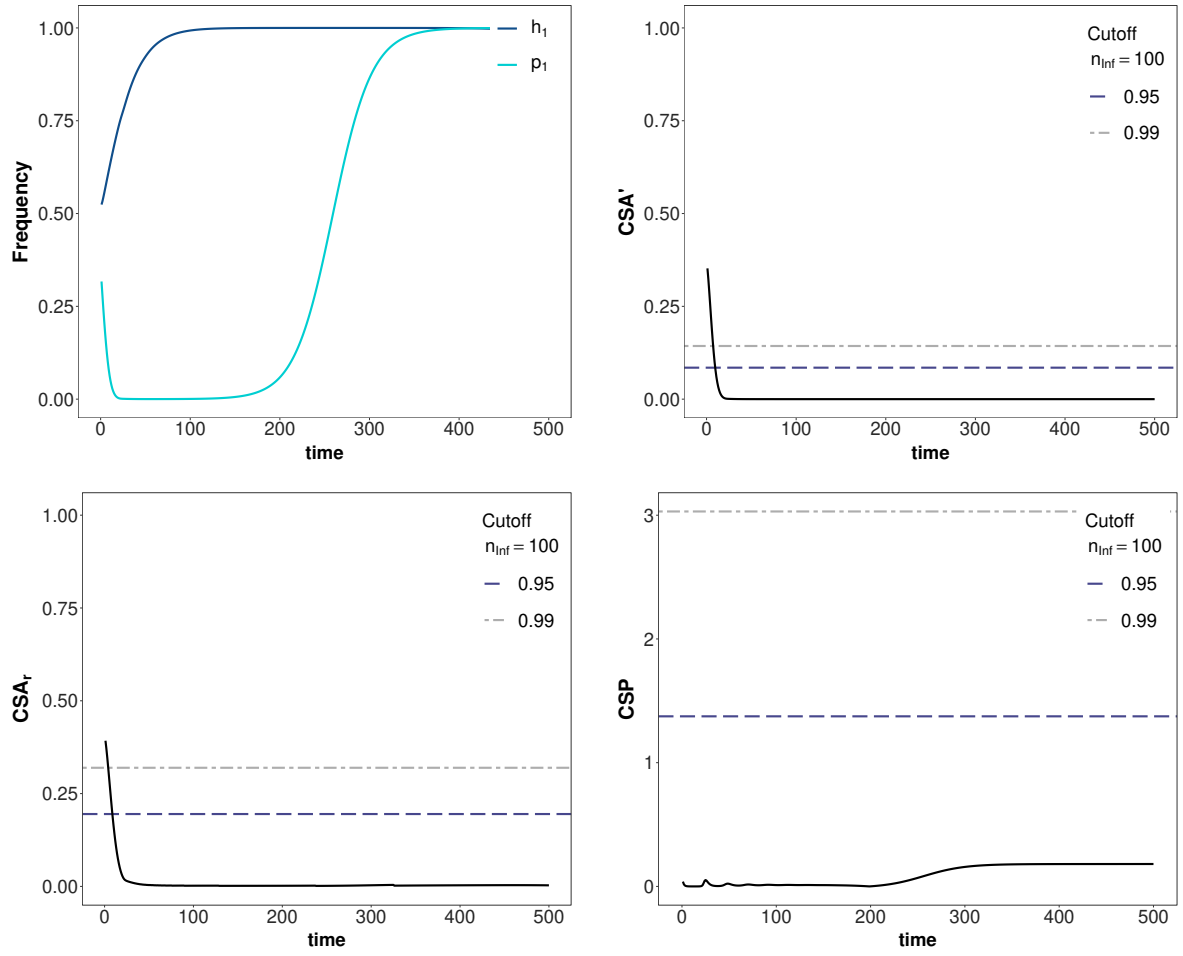


Fig. 4.B.3: Temporal changes in allele frequencies, CSA′, CSA_r and CSP an epidemiological (model 4B) with a GFG-infection matrix. For each index cutoff values are shown based on the expected neutral distributions for a total host sample size $n_T = 200$ and for $n_{\text{Inf}} = n_{\text{H}} = 100$. The 0.95-cutoff value is shown in blue (dashed line) and the 0.99-cutoff value is shown in grey (dotted-dashed line). Top left: frequencies of h_1 (darkblue) and p_1 (lightblue). Top right: CSA′. Bottom left: CSA_r. Bottom right: CSP. The parameters values of the model are: $c_{H_1} = c_{P_1} = 0$, $c_{H_2} = c_{P_2} = 0.05$, $\beta = 0.00005$, $s = 0.6$,$c = 0.9$, $S_{1,\text{init}} = S_{2,\text{init}} = 4150$, $I_{11} = I_{12} = I_{21} = I_{22} = 415$, $\delta_t = 0.001$, $b = 1$, $d = 0.9$

5

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

This chapter is currently under review in PLoS Computational Biology. The formatting has been adjusted to the layout of this thesis.

5.1 Abstract

There is a long-standing interest in understanding host-parasite coevolutionary dynamics and associated fitness effects. Increasing amounts of genomic data offer a promising source to identify candidate loci and to infer core parameters of the past coevolutionary history. However, designing methods to extract such information requires to understand 1) how coevolutionary dynamics and genetic drift jointly shape genetic diversity at the coevolving loci, and 2) the extent to which genomic signatures at the coevolving loci are informative about the parameters of interest. By coupling a gene-for-gene model with coalescent simulations, we show that under trench-warfare dynamics the allele frequencies at the internal equilibrium point determine the strength of the resulting balancing selection signatures. As these equilibrium frequencies are differentially affected by the costs of resistance, infectivity and infection, we suggest that the signatures at the coevolving host and parasite loci potentially contain enough information about these costs. As a proof-of-principle, we apply an Approximate Bayesian Computation approach to infer these costs by jointly integrating host and parasite polymorphism data at the coevolving loci from repeated experiments. First, we demonstrate that the cost of infection and host and parasite population sizes can be inferred when the costs of resistance and infectivity are known. Second, joint inference of all three costs works reasonably well when population sizes are known. Third, polymorphism data of the parasite are informative about costs applying to the host and vice-versa. We discuss the implications of our results for genomic-based inference of host-parasite coevolution.

Author summary

It is of importance for agriculture and medicine to understand host-parasite antagonistic coevolutionary dynamics and the deleterious associated fitness effects, as well as to reveal the genes underpinning these interactions. The increasing amount of genomic data for hosts and parasites offer a promising source to identify such candidate loci, but also to use statistical inference methods to reconstruct the past coevolutionary history. In our study we attempt to draw inference on the past coevolutionary history at key host and parasites loci using sequence data from several individuals and across several experimental replicates. We demonstrate that using a Bayesian statistical method, it is possible to estimate the parameters driving the interaction of hosts and parasites at these loci for thousands of generations. The main parameter that can be estimated is the fitness loss by hosts upon infection. Our method and results can be applied to experimental coevolution data with sequences at the key candidate loci providing enough repetitions and large enough population sizes. As a proof of principle, our results open the door to reconstruct past coevolutionary dynamics using sequence data of interacting species.

5.2 Introduction

Host-parasite coevolution is an ubiquitous process and has been demonstrated in terrestrial (Thrall *et al.* 2012), limnological (Decaestecker *et al.* 2007) and marine environments (Martiny *et al.* 2014). It describes the process of parasites and hosts exerting reciprocal selective pressures on one another. Therefore, coevolutionary dynamics are expected to substantially interact with and shape neutral genomic diversity linked to the coevolving sites. The latter can be single or multiple SNPs in coding or non-coding parts of genes (Rose *et al.* 2007; Hoerger *et al.* 2012), insertions/deletions (Stahl *et al.* 1999) or distributed across a gene network (Shin and MacCarthy 2015). Accordingly, the polymorphism patterns at the coevolving loci are expected to be distinct from loci not involved into the coevolutionary interaction. Therefore, host and parasite genomic data are valuable source to identify loci under coevolution and to infer their past coevolutionary history.
On the one hand, signatures of positive selection which are characterized by lower genetic diversity compared to the genome-wide average and increased levels in linkage disequilibrium (Maynard Smith and Haigh 1974) are expected to arise under so called arms-race

dynamics (Woolhouse *et al.* 2002; Holub 2001). In arms race dynamics, frequencies of new beneficial alleles (such as new resistance or infectivity alleles) arising by *de novo* mutations increase towards fixation in both interacting partners. Accordingly, alleles are recurrently replaced and thus, are short lived and polymorphism is only transient (Woolhouse *et al.* 2002; Holub 2001). On the other hand, signatures of balancing selection being characterized by higher than average diversity (Charlesworth 2006) are expected to be the result of so called trench-warfare dynamics (also referred to as Red Queen dynamics) (Stahl *et al.* 1999; Woolhouse *et al.* 2002). In this type of dynamics, several alleles are stably maintained over large time periods in both coevolving species. Hereby, allele frequencies either converge towards a stable equilbrium or they fluctuate persistently over time. Based on these classic expectations, genomic studies have unravelled positive and balancing selection signatures at various resistance genes (Stahl *et al.* 1999; Bakker *et al.* 2006; Karasov *et al.* 2014; Rose *et al.* 2007; Hoerger *et al.* 2012; Caicedo and Schaal 2004; Obbard *et al.* 2011) and effector genes (Schweizer *et al.* 2018).

An additional difficulty for coevolutionary analyses, is that there is a continuum between arms-race and trench-warfare dynamics and the dynamics are in fact strongly affected by the type and strength of various forms of selection (negative indirect frequency-dependent selection, negative direct frequency-dependent selection, overdominant selection) and their interplay with genetic drift (Tellier *et al.* 2014; Gokhale *et al.* 2013) and mutation (Ejsmond and Radwan 2015; Salathe *et al.* 2005). In other words, the expectations above are probably too simple to be accurately applicable as the effect of genetic drift and mutation affecting the outcome of coevolutionary dynamics is ignored (Tellier *et al.* 2014; Gokhale *et al.* 2013; Ejsmond and Radwan 2015). The force underlying coevolution is frequency-dependent selection, which strength depends on the frequency of particular alleles. Negative indirect frequency-dependent selection (niFDS) takes place when the fitness of a particular host allele increases with decreasing frequencies of a particular allele in the parasite and vice-versa (Seger 1988; Tellier and Brown 2007b). When the fitness of a host or parasite allele decreases with its own frequency negative direct frequency-dependent selection (ndFDS) is acting. ndFDS can be promoted by factors such as asynchrony between host and parasite life-cycles (overlapping parasite generations, several parasite generations per host generation) or epidemiological feedback due to density dependent disease transmission (Brown and Tellier 2011). Overdominant selection or some form of ndFDS are a necessary but not always sufficient condition for trench-warfare dynamics to take place in single locus host-parasite coevolutionary interactions (Tellier and Brown

2007b; Ejsmond and Radwan 2015). Even with some form of ndFDS acting, arms-race dynamics can take place if either the strength of ndFDS compared to niFDS is weak or genetic drift is causing random loss of alleles.

The exact nature of the dynamics, such as the equilibrium frequencies of alleles and the period and amplitude of coevolutionary cycles, is further affected by the way host and parasite genotypes interact at the molecular level and the fitness costs associated with the coevolutionary interaction. The interaction at the molecular level is captured by the infection matrix which stores the specificity and level of infection in all possible pairwise interactions between host and parasite genotypes (Kwiatkowski *et al.* 2012). One well studied type of interaction is the gene-for-gene (GFG) interaction which presents one endpoint of a continuum of infection matrices (Agrawal and Lively 2002; Engelstaedter 2015). GFG-interactions are characterized by one universally infective parasite genotype and one universally susceptible host type. Such interactions have been found for example in the Flax-*Melampsora lini* system (Flor 1971).

A fitness cost which has been shown to crucially affect the coevolutionary dynamics is the loss in fitness due to infection (Tellier and Brown 2007b; Tellier *et al.* 2014). In addition, costs of resistance such as reduced competitive ability or fertility in absence of the parasite (Kraaijeveld and Godfray 1997; Bergelson and Purrington 1996; Lenski 1988) and costs of infectivity such as reduced spore production of infective pathogens (Thrall and Burdon 2003) can further alter the dynamics. These costs also determine the equilibrium frequencies of the coevolutionary system (Leonard 1994; Frank 1992) at which one or several alleles are maintained or around which allele frequencies cycle. An important result from previous theoretical investigations (Leonard 1994; Frank 1992) is that the equilibrium frequencies in the parasite population depend on the parasite fitness costs (cost of infectivity) and vice-versa (cost of resistance and cost of infection).

Given this continuum of coevolutionary dynamics, it is necessary to gain a deeper and refined understanding on how the interaction between allele frequency dynamics at the coevolving loci, genetic drift and mutation shapes the resulting genomic signatures at the coevolutionary loci and linked neutral sites. This is an important step for the development and application of methods designed to draw inference on the coevolutionary history.

Therefore, our first aim is to explicitly investigate the links between coevolutionary dynamics and the resulting signatures by using a model spanning a range of coevolutionary dynamics inbetween arms-race and trench-warfare with varying equilibrium frequencies. We can show that the resulting coevolutionary signatures under trench-warfare dynam-

ics are strongly affected by the equilibrium frequencies of alleles. As these equilibrium frequencies are governed by coevolutionary costs, our second aim is to infer information about these costs from polymorphism data of the coevolving loci. Therefore, we make use of Approximate Bayesian Computation (Csillery *et al.* 2010; Beaumont *et al.* 2002; Sunnaker *et al.* 2013). Approximate Bayesian computation (ABC) is an inference method which can be used in situations where likelihood calculations are intractable, as it is the case for the coevolutionary models (Nuismer and Week 2019). The core of ABC is to perform a huge amount of simulations under a model which is expected to have given rise to the observed data. Each single simulation is run with different values for the parameters to be inferred and the resulting summary statistics are compared to the same set of statistics calculated from the observed data. We assume that our observed data set consists of the average summary statistics from either $r = 200$ or $r = 10$ independent replicates of a coevolutionary experiments where for each replicate $n = 50$ host and parasite sequences from the coevolutionary loci have been obtained.

5.3 Materials and methods

5.3.1 Simulation of polymorphism data

We model coevolution between a single haploid host and a single haploid parasite species. The coevolutionary interaction in both species is driven by a single bi-allelic functional site (SNP, indel, ...). This functional site is located in the coevolutionary locus which encompasses several other neutral sites. Hosts are either resistant (*RES*) or susceptible (*res*) and parasites are either non-infective (*ninf*) or infective (*INF*). Thus, the model follows a gene-for-gene interaction with the following infection matrix:

$$\begin{array}{c} \\ RES \\ res \end{array} \begin{array}{c} \begin{array}{cc} ninf & INF \end{array} \\ \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}. \end{array} \tag{5.1}$$

A 1-entry in the infection matrix indicates that the parasite is able to infect the host and a 0-entry indicates that the host is fully resistant towards the parasite. To obtain polymorphism data at these major coevolutionary loci we combine a forward-in-time coevolu-

tionary model (Tellier and Brown 2007b) including genetic drift and recurrent mutations with backward-in-time coalescent simulations (Tellier *et al.* 2014).

Forward in time coevolution model

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

1. Using a discrete-time gene-for-gene coevolution model (shown below) we compute the expected allele frequencies in the next generation (in infinite population size)

2. We incorporate genetic drift by performing a binomial sampling based on the frequency of the *RES*-allele (*INF*-allele) after selection and the finite and fixed haploid host population size N_H (parasite population size N_P).

3. We allow for recurrent allele mutations to take place and change genotypes from *RES* to *res* at rate μ_{Rtor} or *res* to *RES* at rate μ_{rtoR} in the host and from *ninf* to *INF* at rate μ_{ntoI} and from *INF* to *ninf* at rate μ_{Iton} in the parasite. Henceforward, we call such mutations as functional mutations. We set all functional mutation rates to $\mu_{Rtor} = \mu_{ntoI} = \mu_{rtoR} = \mu_{Iton} = 10^{-5}$.

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

We denote the frequency of resistant (susceptibe) by R (r) and the frequency of infective parasite (non-infective parasites) by a (A). The coevolution model (henceforward termed **model 5A**) is based on the polycyclic auto-infection model in (Tellier and Brown 2007b). This population genetics model (*sensu* Ashby *et al.* (2019)) assumes host and parasite populations to be constant regardless of the disease prevalence and non-overlapping host and parasite generations, and as such is probably more suited to describe plant-parasite or invertebrate-parasite systems. Polycyclic diseases are characterized by more than one infection cycle per season. For simplicity, the model is based on $T = 2$ infection cycles per discrete host generation g each caused by a single discrete parasite generation t ($t \in \{1, 2\}$). An auto-infection refers to an infection where a parasite re-infects the host individual on which it was produced. Therefore, resistant (R_g) and susceptible hosts (r_g) which are infected by infective parasites ($a_{g,1}$) in the first infection cycle ($t = 1$) stay

infected by infective parasites in the second infection cycle ($t = 2$). This causes a fitness reduction s_1 (cost of infection). The same applies to susceptible host (r_g) infected by non-infective parasites ($A_{g,1}$) in the first infection cycle ($t = 1$). Resistant hosts which are attacked by non-infective parasites in the first infection cycle ($t = 1$) resist infection. In the second infection cycle ($t = 2$), this fraction of resistant hosts ($R_g \cdot A_{g,1}$) either receives a non-infective parasite ($A_{g,2}$) resulting in no fitness loss or an infective parasite ($a_{g,2}$) resulting in a reduced cost of infection s_2. Host resistance comes at cost c_H (cost of resistance) and infectivity in the parasite comes at cost c_P (cost of infectivity).
The allele frequencies of resistant hosts (R_g), susceptible hosts (r_g), non-infective parasites ($A_{g,t}$) and infective parasites ($a_{g,t}$) are given by the following recursive equations (for the corresponding fitness matrices see 5.B.1):

$$a_{g,2} = \frac{a_{g,1} \cdot (1 - c_P)}{a_{g,1} \cdot (1 - c_P) + A_{g,1} \cdot r_g} \tag{5.2a}$$

$$a_{g+1,1} = \frac{(1 - c_P) \cdot [R_g \left(A_{g,1}a_{g,2} + a_{g,1}\right) + r_g a_{g,1}]}{(1 - c_P) \cdot [R_g \left(A_{g,1}a_{g,2} + a_{g,1}\right) + r_g a_{g,1}] + r_g A_{g,1}} \tag{5.2b}$$

$$R_{g+1} = \frac{R_g \cdot (1 - c_H) \left[A_{g,1}A_{g,2} + A_{g,1}a_{g,2}(1 - s_2) + a_{g,1}(1 - s_1)\right]}{R_g \cdot (1 - c_H) \left[A_{g,1}A_{g,2} + A_{g,1}a_{g,2}(1 - s_2) + a_{g,1}(1 - s_1) + r_g(1 - s_1)\right]} \tag{5.2c}$$

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

$$\begin{aligned} \hat{a} &\approx \frac{s_2 + s_1 - \sqrt{(s_2 + s_1)^2 - 4s_2(s_1 - c_H)}}{2s_2(1 - c_H)} \\ \hat{R} &\approx \frac{c_P}{2 - c_P - \hat{a}} \\ &\approx \frac{2c_P \cdot s_2 \cdot (1 - c_H)}{s_2(3 - 4c_H - 2c_P(1 - c_H)) - s_1 + \sqrt{(s_2 + s_1)^2 - 4s_2(s_1 - c_H)}} \end{aligned} \tag{5.3}$$

For investigating the link between coevolutionary dynamics in infinite population size and genomic signatures in finite population size, we further use these recurrence equation to simulate allele frequency trajectories in infinite population size for $g_{max} = 30,000$ generations (without genetic drift and recurrent mutations) for all pairwise combinations of $s = \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$, $c_P = \{0.1, 0.3\}$ and $c_H = \{0.05, 0.1\}$.
Additionally, we use two extensions, 5**B** and 5**C** respectively, of the basic model to check

for the generality of our results. In model 5**B**, we extend model 5**A** to more than two parasite ($T > 2$) generations per host generation g (see 5.B). Model 5**C** extends model 5**A** (keeping $T = 2$) by allowing for allo-infections to take place at rate $(1 - \psi)$ in the second parasite generation ($t = 2$) within host generation g (see 5.B).

Backward in time coalescent

To obtain polymorphism data at the coevolutionary loci we combine the obtained frequency paths which include genetic drift and recurrrent mutations with coalescent simulations separately for the host and the parasite. Therefore, we first rescale time for the frequencies paths appropriately (for more information see 5.B). Based on these time-rescaled frequency paths we launch a modified version of msms (Ewing and Hermisson 2010; Tellier *et al.* 2014) once for each species. We set the sample size to $n_H = 50$ for the host ($n_P = 50$ for the parasite). For both species we assume realistically a non-recombining locus of length 2500 bp and a per site neutral mutation rate of 10^{-7}. Accordingly, the neutral population mutation rate is $\theta_H = 2 \cdot N_H \cdot 2500 \cdot 10^{-7}$ for the host ($\theta_P = 2 \cdot N_P \cdot 2500 \cdot 10^{-7}$ for the parasite). Note that the same trends for the summary statistics are obtained when assuming a larger locus length or higher per site mutation rate. Based on the respective msms-output we calculate eight summary statistics for each species which are based on the site frequency spectrum (SFS) of the respective coevolving locus (5.D.2). We only use summary statistics based on the unfolded site frequency spectrum (SFS), as it can be hard to obtain unbiased haplotype summary statistics depending on the sequence method. In addition to these 16 summary statistics we calculate an additional summary statistic (**P**airwise **M**anhattan **D**istance) which is combining information from host and parasite polymorphism data (see 5.C).

5.3.2 Generating pseudo-observed data sets (PODs)

To understand the link between coevolutionary dynamics in finite population size and genomic signatures at the coevolving loci we simulate pseudo-observed data sets (PODs) for various costs of infection (s). This parameter strongly influences the coevolutionary dynamics, namely the allele frequencies at the non-trivial equilibrium point and the stability of the internal equilibrium point (Fig. 5.1, Tab. 5.B.2, Tellier and Brown (2007b)). Therefore by changing s, we can investigate a continuum between arms-race and trench-warfare dynamics with varying internal equilibrium frequencies. We vary s from 0.2 to

0.8 in steps of 0.1. We simulate $r = 200$ repetitions for each value of s using the above mentioned forward-backward approach, and afterwards average the summary statistics across the $r = 200$ repetitions/per parameter combination/per model. For the so called standard case we fix the parameters as follows: $c_H = 0.05$, $c_P = 0.1$, $N_H = N_P = 10,000$, $n_H = n_P = 50$. We extend this standard case in two ways. First, we simulate data for various combinations of host population size $N_H = (5,000; 10,000; 15,000)$ and parasite population size $N_P = (5,000; 10,000; 15,000)$. Second, we assess the signatures for combinations of $c_H = (0.05, 0.1)$ and $c_P = (0.1, 0.3)$ while fixing the population sizes to their standard values.

5.3.3 Performing ABC on the pseudo-observed data sets for model 5A

As a proof of principle we aim to infer different parameters determining the coevolutionary dynamics in model 5**A**. In scenario 1, we aim to infer simultaneously the cost of infection (s), the host population size (N_H) and the parasite population size (N_P) assuming that we know the true cost of resistance c_H and the true cost of infectivity c_P. This scenario mimics systems where experimental measures of the costs of resistance or infectivity have been performed (Tian *et al.* 2003; Thrall and Burdon 2003) and thus, these parameters can be fixed in our method. In scenario 2, our goal is to infer simultaneously the cost of infection (s), the cost of infectivity (c_P) and the cost of resistance (c_H) assuming we know the true host (N_H) and parasite population sizes (N_P). Scenario 2 is motivated by the assumption that an independent estimate of the effective population size can be obtained by using full-genome data of loci unlinked to the coevolutionary locus. We also test for the effect of the number of repetitions on the inference results. Thus, we base our inference on the average summary statistics of $r = 200$ and $r = 10$ repetitions. Besides the effect of the number of repetitions, we assess how the type of polymorphism data available affects the accuracy of inference. Therefore, we perform inference based on a) polymorphism data of both, the host and the parasite, b) polymorphism data of the host and c) polymorphism data of the parasite. For the sampling step of the ABC we use the ABCsampler from ABCtoolbox (Wegmann *et al.* 2010) and perform 100,000 simulations using the standard sampler. The chosen priors, complex parameters and fixed parameters can be found in Tab. 5.D.3. For the estimation step we retain the 1% best simulations and apply the post-sampling adjustment (general linear model) as implemented in ABCestimator (Wegmann *et al.* 2010) (note that similar results are obtained when using the 0.2% best simulations).

All codes and pipelines used are available upon request and will be placed on a publicly available repository.

5.4 Results

5.4.1 Link between coevolutionary dynamics and sequence data

The internal equilibrium frequency of the *RES*-allele mainly increases with increasing cost of infectivity (c_P) (Fig. 5.1 a+b vs. Fig. 5.1 c+d), increases very slightly with increasing cost of infection (s) and remains almost unaffected by changing costs of resistance (c_H) (Fig. 5.1 a+c vs. Fig. 5.1 b+d). The opposite is true for the parasite. Here, the equilibrium frequency of the infective (*INF*)-parasite rises mainly with increasing cost of infection (s) (Fig. 5.1). Higher costs of resistance (c_H) decrease the equilibrium frequency of *INF*-parasites (Fig. 5.1 a+c vs. Fig. 5.1 b+d) for a given value of s. In contrast to the host, the equilibrium frequencies in the parasite are almost unaffected by changes in the cost of infectivity (c_P). For high costs of infection s the dynamics are always switching to arms-race dynamics, irrespectively of the underlying costs of resistance and infectivity (Fig. 5.1, 5.A.7). A large body of theoretical studies has dealt with the dynamics and equilibrium properties of coevolutionary models (i.e. Leonard (1994), Sasaki (2000) and Tellier and Brown (2007b)). However, the result for the used model are shown here, to make it easier to grasp the link between coevolutionary dynamics in infinite populations and genomic signatures.

The changes in equilibrium frequencies with changing cost of infection (s), cost of resistance (c_H) and changing cost of infectivity (c_P) are reflected by the resulting genomic signatures at the coevolving loci (Fig. 5.2). We summarize the genomic signatures of coevolution chiefly as the value of Tajima's D because its behaviour is well known under selective sweeps and balancing selection (Fig. 5.2, 5.3). However, this statistic is also affected by demography, negative selection and linked selection which we do not take into account here (see discussion for the effect of demography). Generally, the strongest signatures of balancing selection can be observed when the equilibrium frequencies of *INF*-parasites or *RES*-hosts are close to 0.5 (see Fig. 5.1, Tab. 5.B.2, Fig. 5.2). The strength of the signatures declines the further the equilibrium frequencies move away from 0.5.

The genomic signature in the parasite changes strongly with changing cost of infection

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

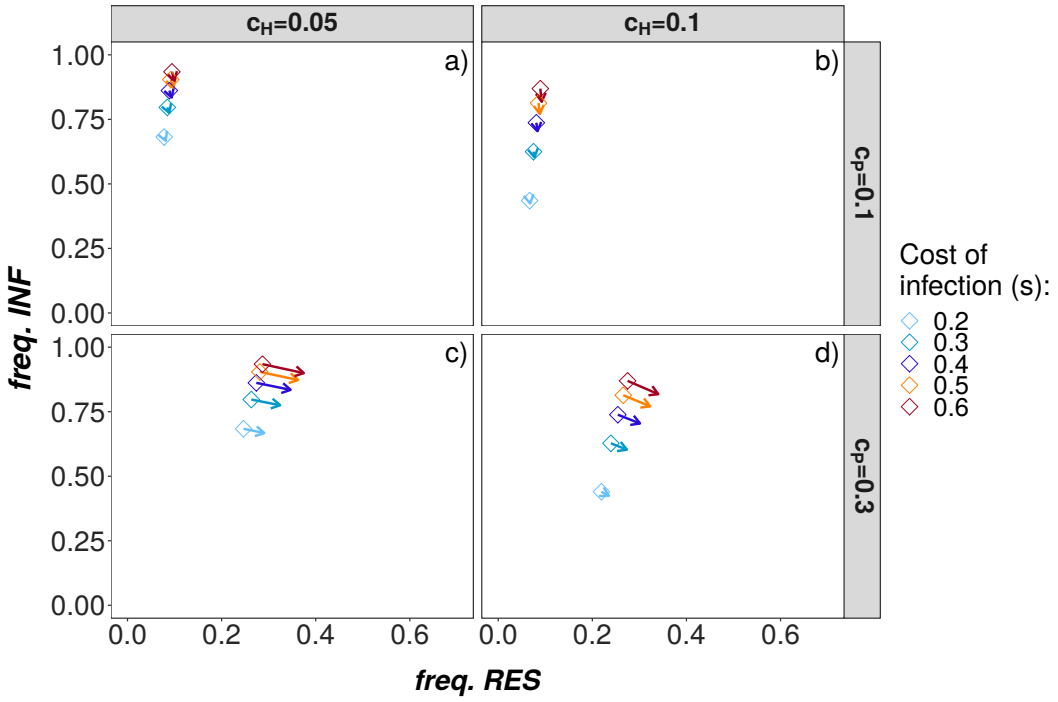


Fig. 5.1: Deterministic equilibrium frequencies for model 5A (pure autoinfection model with $T = 2$ parasite generations) for different combinations of cost of resistance $c_H = (0.05, 0.1)$ (columns), cost of infectivity $c_P = (0.1, 0.3)$ (rows) and cost of infection $s = (0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8)$ (color of the squares). Only combinations with trench-warfare dynamics are shown. Centres of the squares represent the equilbrium frequencies obtained by simulating numerically the recursion equations Eq. 5.2 for 30,000 generations starting with an initial frequency of $R_0 = 0.2$ resistant hosts and $a_0 = 0.2$ infective parasites. Heads of the arrows represent the equilibrium frequencies based on Eq. 5.3 which slightly differ from the numerical computations due to analytical approximations.

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

The combination of host and parasite signatures holds the highest information content about the fitness parameters guiding the coevolutionary dynamics. This is due to the fact that the equilibrium frequencies in the host and parasite are differentially affected by

these fitness parameters. The genomic signature in the host is mainly indicative about the cost of infectivity (c_P), a cost which is affecting the parasite fitness, whereas the signature in the parasite is mainly informative about the costs of resistance (c_H) and infection (s), parameters with a direct fitness effect in the host (Fig. 5.2). This results from the action of niFDS. Increasing costs of resistance (c_H) disfavor resistant hosts. Thus, the frequency of infective parasites is decreasing which results in lower equilibrium frequencies of *INF*-parasites. The opposite is true for increasing costs of infectivity (c_P). This cost reduces the fitness of *INF*-parasites which in turn favors *RES*-hosts compared to *res*-hosts.
The qualitative changes of the genomic signatures for changing costs of infection in the standard case remain similar even when population sizes differ in both interacting partners (Fig. 5.3). However, the strength of genomic signatures is affected by the population sizes. The strongest signature of balancing selection in the parasite is found when the parasite population size is small compared to the host population size (Fig. 5.3 c). Here, the large host population size reduces the amount of genetic drift in the host. Thus, there are less allele frequency fluctuations in the host around the internal equilibrium point. This in turn, also reduces allele frequency fluctuations in the parasite.
Overall, there is a strong link between the equilibrium frequencies under trench-warfare dynamics and the resulting genomic signatures. We obtain similar results when we slightly modify the assumptions about the coevolutionary interaction by either a) extending the model to more than two parasite generations per host generation (**model 5B**, Fig. 5.A.1, Fig. 5.A.3 c+d) or b) allowing for allo-infections at rate $1 - \psi$ in the second parasite generation within host generation g (**model 5C**, Fig. 5.A.2, Fig. 5.A.3 a+b). Increasing the number of parasite generations extends the parameter space in which trench-warfare dynamics occur. Here, the strongest signatures of balancing selection are found for intermediate costs of infection (Fig. 5.A.3). Allowing for allo-infections, decreases the parameter space in which trench-warfare dynamics take place and thus, also the range in which balancing selection signatures can be observed in both interacting partners (Fig. 5.A.3 a+b).
The strong link between equilibrium frequencies and resulting genomic signatures can be explained in terms of a structured coalescent tree. The coalescent tree in both coevolving species consists of two demes (*RES* and *res* for the host and *INF* and *ninf* for the parasite). As we assume no recombination within the coevolutionary locus new mutations are usually linked to the functional allele in which they arose. When a functional mutation is taking place the affected lineage is migrating from one deme to the other.

When the frequencies of both allles (*ninf* and *INF* in the parasite or *RES* and *res* in the host) are fairly similar they have equal contributions to the sample. Thus, the underlying coalescent tree is well balanced. Accordingly, we observe an excess of intermediate frequency variants in the SFS. As the equilibrium frequencies move away from 0.5, the average sample configuration changes and the coalescent tree becomes less balanced (see Fig. 5.A.7 a-f and VII-XII). Therefore, the number of SNPs at intermediate frequencies drops and Tajima's D decreases (Fig. 5.2). Note that two trench-warfare models with different parameters, but same equilibrium frequencies would exhibit the same genomic signatures if reaching the stable polymorphic point. If there is cycling, rather than a fixed-point equilibrium, then the dynamics and thus the polymorphism data do also depend on the amplitude and period of the fluctuations, and likely generate different signatures between the two models.

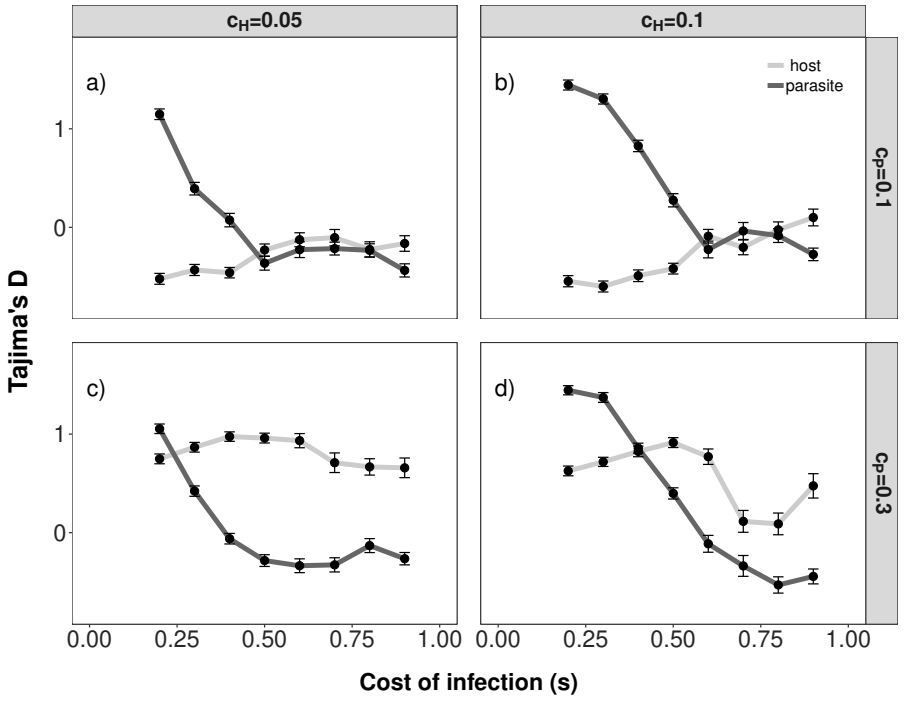


Fig. 5.2: Tajima's D (y-axis) for model 5A for various cost of infection s (x-axis). The results are shown for different combinations of c_P ($c_P = 0.1$ top, $c_P = 0.3$ bottom) and c_H ($c_H = 0.05$ left, $c_H = 0.1$ right). The mean and standard error of Tajima's D of the parasite population (dark grey) and of the host population (light grey) are plotted for $r = 200$ repetitions. The other parameters are fixed to: $N_H = N_P = 10,000$, $n_H = n_P = 50$, $\theta_H = \theta_P = 5$, $\mu_{Rtor} = \mu_{rtoR} = \mu_{ntoI} = \mu_{Iton} = 10^{-5}$.

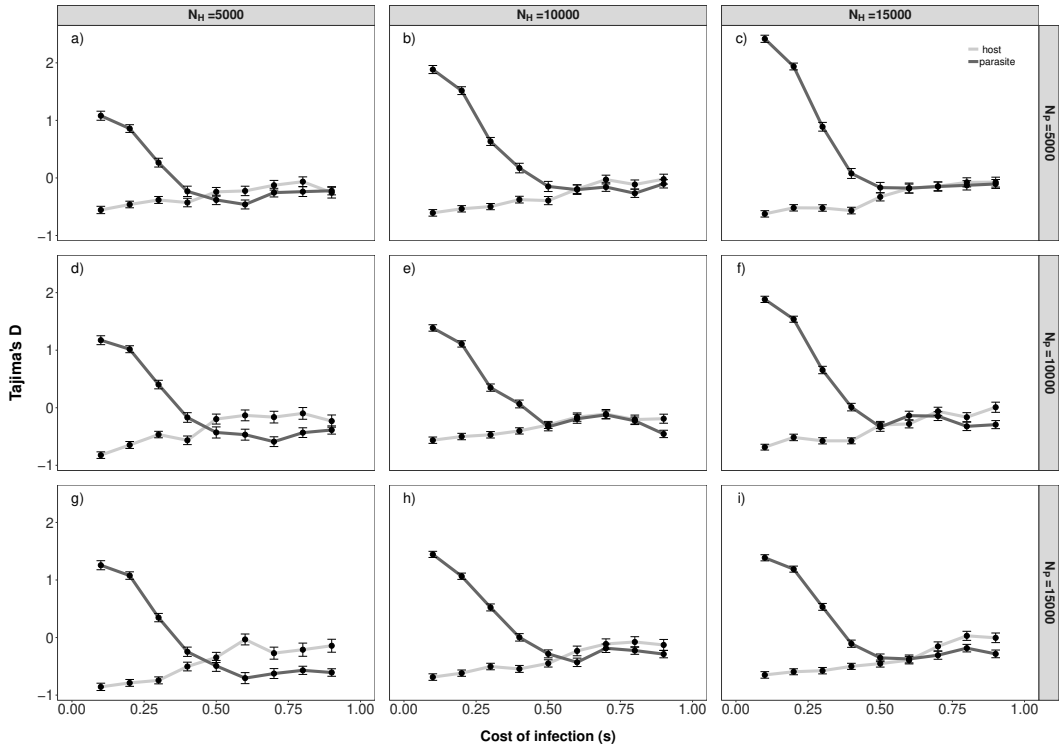


Fig. 5.3: Tajima's D (y-axis) for **model 5A** for various cost of infection s (x-axis) and different combinations of N_P ($N_P = 5,000$ top, $N_P = 10,000$ middle, $N_P = 15,000$ bottom) and N_H ($N_H = 5,000$ left, $N_H = 10,000$ middle, $N_H = 15,000$ right). The mean and standard error of Tajima's D of the parasite population (dark grey) and of the host population (light grey) are plotted for $r = 200$ repetitions. Note that subfigure e corresponds to Fig. 5.2 a. The other parameters are fixed to: $c_H = 0.05$, $c_P = 0.1$, $\theta_H = N_H/2000$, $\theta_P = N_P/2000$, $n_H = n_P = 50$, $\mu_{Rtor} = \mu_{rtoR} = \mu_{ntoI} = \mu_{Iton} = 10^{-5}$.

5.4.2 Inference of coevolutionary dynamics from polymorphism data

Our results indicate that it is possible to infer the cost of infection using polymorphism data from the host and parasite (Fig. 5.4, Fig. 5.5). The accuracy of inference mainly depends on four factors being 1) the true value of the cost of infection, 2) the type of polymorphism data being used (host and parasite together, only host or only parasite), 3) the number of available repetitions and 4) the type of known parameters.
Inferences of the cost of infection and of the population sizes are the most accurate if the number of repetitions is high ($r = 200$), and host and parasite polymorphism data are both available (Fig. 5.4, Fig. 5.A.4, Fig. 5.A.5). Generally, inference for Scenario 1 works best if host and parasite data are used together, irrespectively of the number of repetitions available (compare Fig. 5.4 a to Fig. 5.4 b+c; Fig. 5.4 d to Fig. 5.4 e+f).

Using only parasite polymorphism data is also quite accurate for small to intermediate values of the cost of infection ($s < 0.6$) (Fig. 5.4 c+f) where trench-warfare dynamics take place and SFS of the parasite changes pronouncedly with s (Fig. 5.A.7). In contrast, using only host polymorphism data shows markedly less accuracy in the same parameter range (Fig. 5.4 b+e), especially if the number of available repetitions is low. For low costs of infection the respective equilibrium frequencies of the *RES*-genotype are close to zero (< 0.1) and increase only very slightly when s increases (Tab. 5.B.2). Thus, the host sample mostly consists of polymorphism data from *res*-hosts. Accordingly, the coalescent tree consists of a very large subtree containing the *res*-samples and a very small subtree containing the *RES*-samples and the overall tree looks almost neutral (Fig. 5.2). The estimation accuracy of the cost of infection using only host information diminishes in the transition between trench-warfare and arms-race dynamics (around $s \approx 0.6$), especially if the number of repetitions is low ($r = 10$). In this range, fixation of alleles in both species can either happen due to genetic drift or due to the inherent dynamics of the coevolutionary interaction. This effect decreases the accuracy of parameter estimation even if host and parasite polymorphism data are available (Fig. 5.4 a+d, Fig. 5.2).
The results indicate that the availability of more repetitions increases the accuracy of inference (compare Fig. 5.4 a-c to Fig. 5.4 d-f). There are three sources of stochasticity affecting the polymorphism data at the coevolutionary loci: 1) The effect of genetic drift on the allele frequency trajectory under coevolution, 2) the stochasticity in the coalescent process for a given allele frequency trajectory and 3) the stochasticity in the neutral mutation process on top of the coalescent process. As the first type of stochasticity affects the 'population' sizes of the functional alleles in the host (in the parasite) over time it also has a subsequent effect on the other two sources of stochasticity. Using data from several repetitions allows to better handle and to average out the effect of genetic drift on the variability of the allele frequency path and its subsequent effect on the observed summary statistics. This is especially helpful in the range of parameter values where the dynamics switch from arms-race to trench-warfare.
Like in scenario 1 inference for scenario 2 works best if data from both the host and the parasite are available for a large amount of repetitions $r = 200$. However, the accuracy of inference for the cost of infection s is generally not as accurate as in scenario 1. Simultaneous inference of all three parameters in scenario 2 is most accurate for intermediate costs of infection and if both, host and parasite polymorphism data, are available. This is due to the fact that signatures in the host and the parasite are differentially affected

by the various costs (Fig. 5.2).
Inference of the cost of resistance (c_H) works reasonably well if polymorphism data only from the parasite are available. However, this comes at the cost of less accurate inference of the cost of infection (s) as both paramters are affecting the equilibrium frequency in the parasite (Fig. 5.4, Fig. 5.5). While the equilibrium frequency of *INF*-parasites increases with increasing cost of infection, an increase in the cost of resistance for a fixed cost of infection (s) decreases the respective equilibrium frequency. Thus, overestimating the cost of infection (s) can be compensated by overestimating the cost of resistance (c_H) simultaneously. This effect can be seen for low costs of infection (s) if only the information from the parasite polymorphism data is used in Scenario 2 (see Fig. 5.5 c+f). In contrast, inference of the cost of infectivity (c_P) works reasonably well if polymorphism data only from the host are available. This is due to the fact that changing costs of infectivity (c_P) mainly affect the equilibrium frequencies in the host but not in the parasite (Fig. 5.1). Therefore, inference of this parameter does not work if only parasite polymorphism data are available. All the above mentioned effects explain why the simultaneous inference of several cost becomes less accurate with less ($r = 10$) repetitions (Fig 5.A.6).

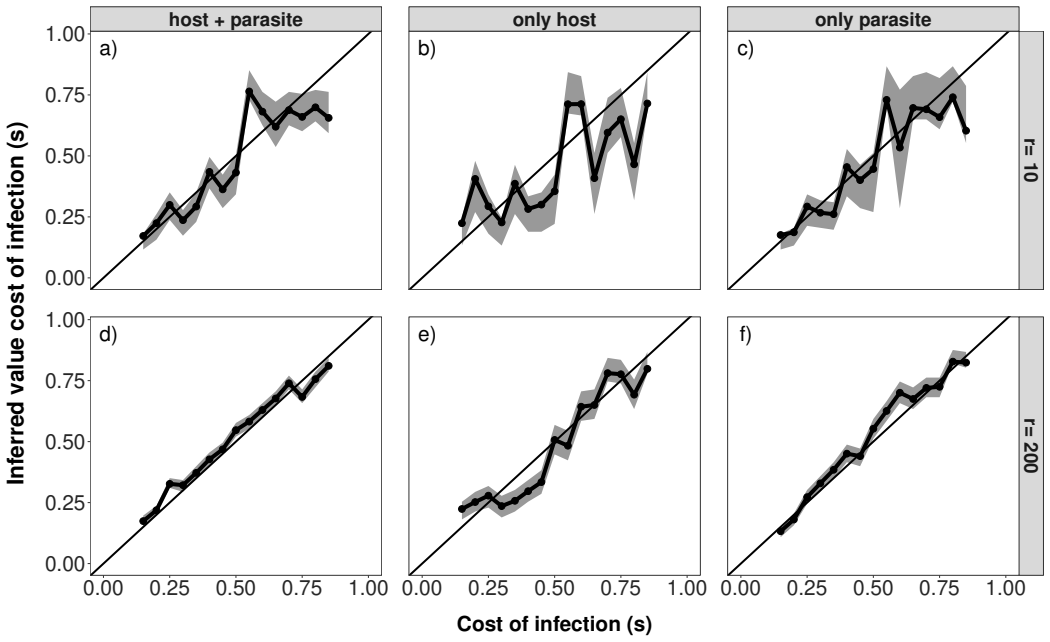


Fig. 5.4: . Median of the posterior distribution (y-axis) for the cost of infection s compared to the true value (x-axis) for $r = 10$ (top, a-c) and $r = 200$ (bottom, d-f). The inference results for scenario 1 based on host and parasite polymorphism data (left, a+d), host polymorphism data only (middle, b+e) and parasite polymorphism data only (right, c+f) are shown. The chosen parameters are: $c_H = 0.05$, $c_P = 0.1$, functional mutation rate 10^{-5}, $c_H = 0.05$, $c_P = 0.1$, $g_{max} = 30,000$, $\theta_H = 5$, $\theta_P = 5$, $n_H = 50$ and $n_P = 50$. Priors have been chosen as follows: N_H log uniform$(2000, 40000)$, N_P log uniform$(2000, 40000)$, s uniform $(0.1, 0.9)$. s, N_H and N_P are inferred simultaneously for all plots.

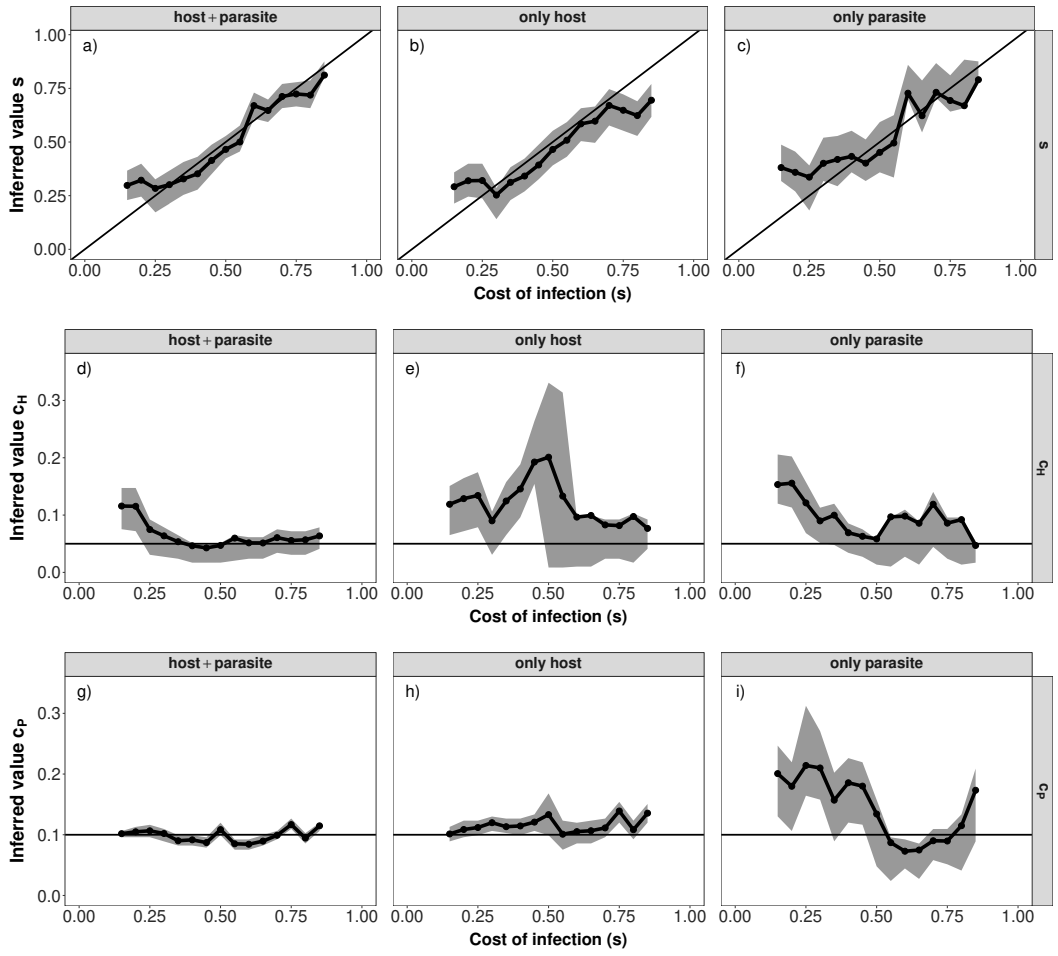


Fig. 5.5: Median of the posterior distribution (y-axis) for the cost of infection s (a-c), cost of resistance c_H (d-f) and cost of infectivity c_P (g-i) compared to the true value (x-axis) for $r = 200$. Inference results for scenario 2 are based on host and parasite polymorphism data (left, a+d+g), host polymorphism data only (middle, b+e+h) and parasite polymorphism data only (right, c+f+i). The chosen parameters are: $N_H = N_P = 10,000$, functional mutation rate 10^{-5}, $g_{max} = 30,000$, $\theta_H = 5$, $\theta_P = 5$, $n_H = 50$ and $n_P = 50$. Priors have been chosen as follows: s uniform$(0.1, 0.9)$, c_H uniform $(0.01, 0.35)$, c_P uniform$(0.01, 0.35)$. s, c_H and c_P are inferred simultaneously for all plots.

5.5 Discussion

We established a link between coevolutionary dynamics (Fig. 5.1), the resulting genomic signatures (Fig. 5.2, Fig. 5.3) and subsequentially the amount of information about the underlying coevolutionary dynamics which can be extracted from genomic signatures at the coevolving loci (Fig. 5.4, Fig. 5.5). Our results indicate that under trench-warfare dynamics the allele frequencies at the non-trivial internal equilibrium point affect the strength of genomic signatures at the coevolving loci in both, the host and parasite. We further could show as a proof of principle that it is possible to infer information about parameters underlying the coevolutionary interaction from polymorphism data at the loci under coevolution if some relevant parameters such as diverse costs (Fig. 5.4) or population sizes (Fig. 5.5) are known. This is due to the fact that various parameter combinations can give rise to similar equilibrium frequencies and thus, result in undistinguishable genomic signatures. In general, inference works best if polymorphism data from both the host and the parasite are available from repeated experiments.

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

The found links between dynamics in infinite population size and genomic signatures in finite population size have several implications. Model-based inference of parameters governing the coevolutionary dynamics is possible if they substantially shift the equilibrium frequencies of the dynamics and thus, the resulting genomic signatures. In cases where different parameters shift the equilibrium frequencies along the same axis, three different inference scenarios are possible. First, it is only possible to infer a compound parameter if there is no *a priori* information on any of the parameters available. This is illustrated by the inference results for scenario 2 when only parasite polymorphism data are available (Fig. 5.5). Here, overestimating the cost of infection s compensates for overestimating

the cost of resistance c_H. Second, if some parameter values are *a priori* known from experiments, the other parameters can be inferred conditional on this information. Third, the parameters have different effects on the equilibrium frequencies in the host and parasite. Thus, combining host and parasite polymorphism data helps to infer the different parameters simulataneously.

For many host-parasite models (including the one used here) it has been shown that the equilibrium frequencies in the host are substantially or exclusively affected by fitness penalties applying to the parasite and vice-versa (Frank 1992; Leonard 1994; Tellier and Brown 2007b). Thus generally speaking, the strength of genomic signatures in either species are presumably most indicative about processes affecting the coevolving partner. We therefore speculate, that the balancing selection signatures which have been found at R-genes in *Arabidopsis thaliana* (Stahl *et al.* 1999; Bakker *et al.* 2006) (Karasov *et al.* 2014), *Solanum sp.* (Rose *et al.* 2007; Hoerger *et al.* 2012; Caicedo and Schaal 2004), *Phaseolus vulgaris* (De Meaux *et al.* 2003), *Capsella* (Gos *et al.* 2012), are indicative about the selective pressure in the coevolving parasite or parasite community. Conversely, the long term maintenance of strains in *P. syringae* (Karasov *et al.* 2018) could reflect fitness costs in *A. thaliana*.

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

In general, our results should not be restricted to the used coevolution model (see Appendix). We acknowledge that we assumed the most simple type of coevolutionary interaction possible. However, understanding possible links between dynamics, signatures and resulting accuracy of inference for this simple scenario is a useful starting point to develop further inference methods where several major loci (Shin and MacCarthy 2015) or quantitative traits (Nuismer and Week 2019) are involved. There are various other coevolution models with respect to the biology of the coevolving species or the ecology of the disease which have been shown to result in trench-warfare dynamics. Nevertheless as long as the coevolutionary interaction is driven by a single bi-allelic functional site in each species, the resulting equilbrium frequencies will be always confined to the 2-dimensional plane and a limited amount of possible genomic signatures (see Fig. 5.A.1, Fig. 5.A.2 and Fig. 5.A.3). Therefore, our findings should also apply to coevolutionary epidemiology models such as in (Ashby and Boots 2017; Gokhale *et al.* 2013).

An important assumption of our model is the absence of intra-locus recombination at the coevolutionary loci. Nevertheless, recombination does occur along the genomes of the host and the parasite, so that the coevolutionary loci evolve independently from other unlinked loci (for example on different chromosomes). This generates two important implications. First, it is possible to estimate the past demographic history based on whole-genome data of both species. So far, we did not take population size changes and the resulting temporal variation in the amount of genetic drift into account. In host-parasite coevolution, population size changes can be due to two different sources: 1) Population size changes which are independent of the coevolutionary interaction and 2) population size changes which arise as an immediate result of coevolutionary interaction, e.g. from epidemiological feedback or any other form of eco-evolutionary feedback. Independently of the particular source, demographic changes always affect all loci in the genome simultaneously, and the resolution of the inference depends on the amplitude and time-scales of the population size fluctuations. A recent study (Živković *et al.* 2019) has shown that fluctuations in population size arising from host-parasite coevolution only leave a signature in the genome-wide parasite site frequency spectrum if they happen at a slow enough time scale. Irrespectively of whether the demographic changes can be resolved from genome-wide data or not, the resulting genomic signatures will be always the result of the allele frequency path at the coevolving loci. Therefore, further studies should focus on the specific effect of eco-evolutionary feedback on the variability of the allele frequency path and the resulting effect of the population size changes on mutation supply at the coevolving loci. Second, our approach can be applied to several pairs of host and parasite coevolving loci, for example a given host species interacting with several parasites species (bacteria, fungi,...). The only requirement is that the coevolutionary dynamics are driven by few major loci in the antagonistic species and no epistasy, pleiotropy or multi-locus phenoytpes are involved. For coevolution due to quantiative traits (Nuismer and Week 2019; Shin and MacCarthy 2015) we expect the signatures to be weaker than in our model (see theory on polygenic selection and polymorphism signatures, (Jain and Stephan 2017)).

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

et al. (2011); Frickel *et al.* (2016)) where coevolutionary interactions can be tracked across several replicates for a reasonable amount of generations. Using data from repeated experiments is one possible attempt to deal with the variability in allele frequency trajectories resulting from the interaction between genetic drift and coevolution. The usage of data from several independent populations or the usage of time-sampled might be possible alternatives. Time samples offer an at least partially time-resolved view on changes in allele frequencies and accordingly, can help to better capture the coevolutionary dynamics.
Our results further show that analyzing both interacting partners in a joint framework rather than analyzing them separately helps to better recover information about the coevolutionary history. This is in line with recent method developments (MacPherson *et al.* 2018; Nuismer *et al.* 2017; Wang *et al.* 2018) which show the value of analyzing hosts and parasite in a joint framework. Additionally, these methods can be promising approaches to identify candidate loci being involved into the coevolutionary interaction on which our approach is based on.

5.6 Conclusion

We investigated here the link between coevolutionary dynamics and resulting genomic signatures and quantify the amount of information available in polymorphism data from the coevolving loci. Although, we started from a very simple coevolutionary interaction we show that model-based inference is possible. With growing availability of highly resolved genome data, even of non-model species, it is important to gain a differentiated and deep understanding of the continuum of possible links between coevolutionary dynamics without or with eco-evolutionary feedbacks and their effect on polymorphism data. Such thorough understanding is the basis for devising appropriate sampling schemes, for using optimal combinations of diverse sources of information and for developing model-based refined inference methods. Our results and the suitability of the ABC approach open the door to further develop inference of past coevolutionary history based on genome-wide data of hosts and parasites from natural populations or controlled experiments.

5.7 Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (TE809/3-1 and TE809/4-1 within the DFG Priority Program SPP1819 "Rapid evolutionary adaptation: Potential and constraints" to AT). We thank Amandine Cornille and Cas Retel for helpful comments on early versions of the manuscript and Lukas Heinrich for performing preliminary analyses.

Supplementary information

5.A Supplementary figures

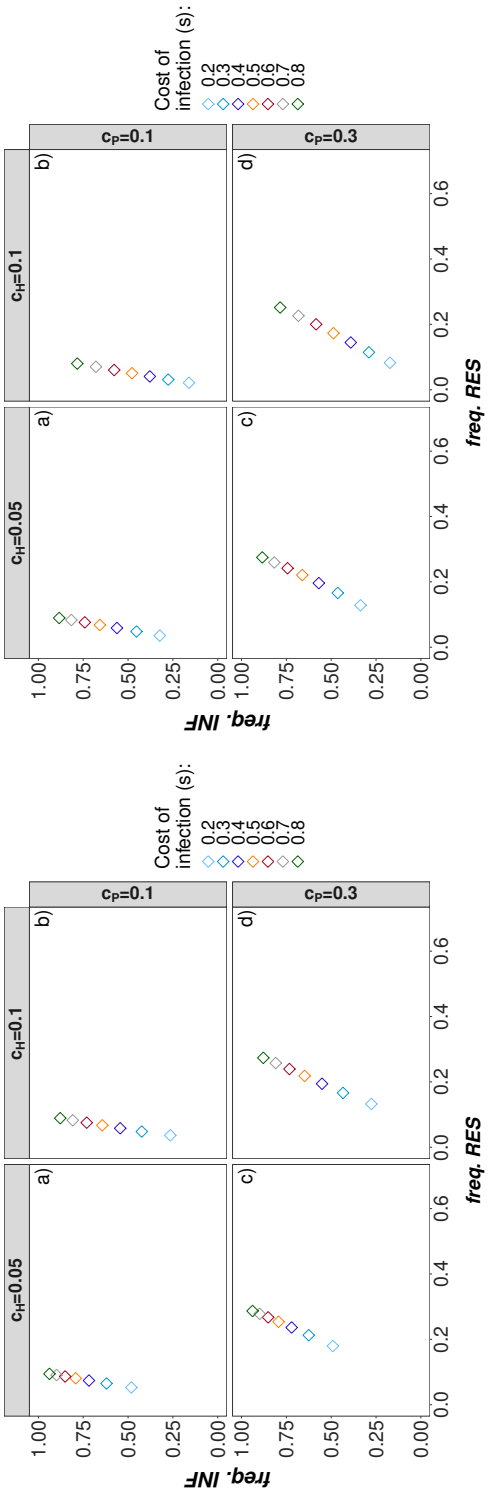


Fig. 5.A.1: Deterministic equilibrium frequencies for **model 5B** for a) $T = 5$ parasite generations (left) and b) $T = 10$ parasite generations (right) per host generation. The equilibrium frequencies for different combinations of cost of resistance $c_H = (0.05, 0.1)$ (columns), cost of infectivity $c_P = (0.1, 0.3)$ (rows) and cost of infection $s = (0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8)$ (color of the squares) are shown. Only combinations with trench-warfare dynamics are shown. Centres of the squares represent the equilibrium frequencies obtained by simulating numerically the recursion equations in Supplementary information models for $g_{max} = 30,000$ host generations starting with an initial frequency of $R_0 = 0.2$ resistant hosts and $a_0 = 0.2$ infective parasites.

Fig. 5.A.2: Deterministic equilibrium frequencies for **model 5C** (auto-allo-infection model) with $T = 2$ parasite generations per host generation for two different autoinfection rates $\psi = 0.75$ (left) and $\psi = 0.95$ (right). The equilibrium frequencies for different combinations of cost of resistance $c_H = (0.05, 0.1)$ (columns), cost of infectivity $c_P = (0.1, 0.3)$ (rows) and cost of infection $s = (0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8)$ (color of the squares) are shown. Only combinations which result in trench-warfare dynamics are plotted. Centres of the squares represent the equilibrium frequencies obtained by simulating numerically the recursion equations in 5.B for $g_{max} = 30,000$ host generations starting with an initial frequency of $R_0 = 0.2$ resistant hosts and $a_0 = 0.2$ infective parasites. Heads of the arrows represent the equilibrium frequencies based on Eq. 5.3 which corresponds to the case $\psi = 1$ (Tellier and Brown 2007b).

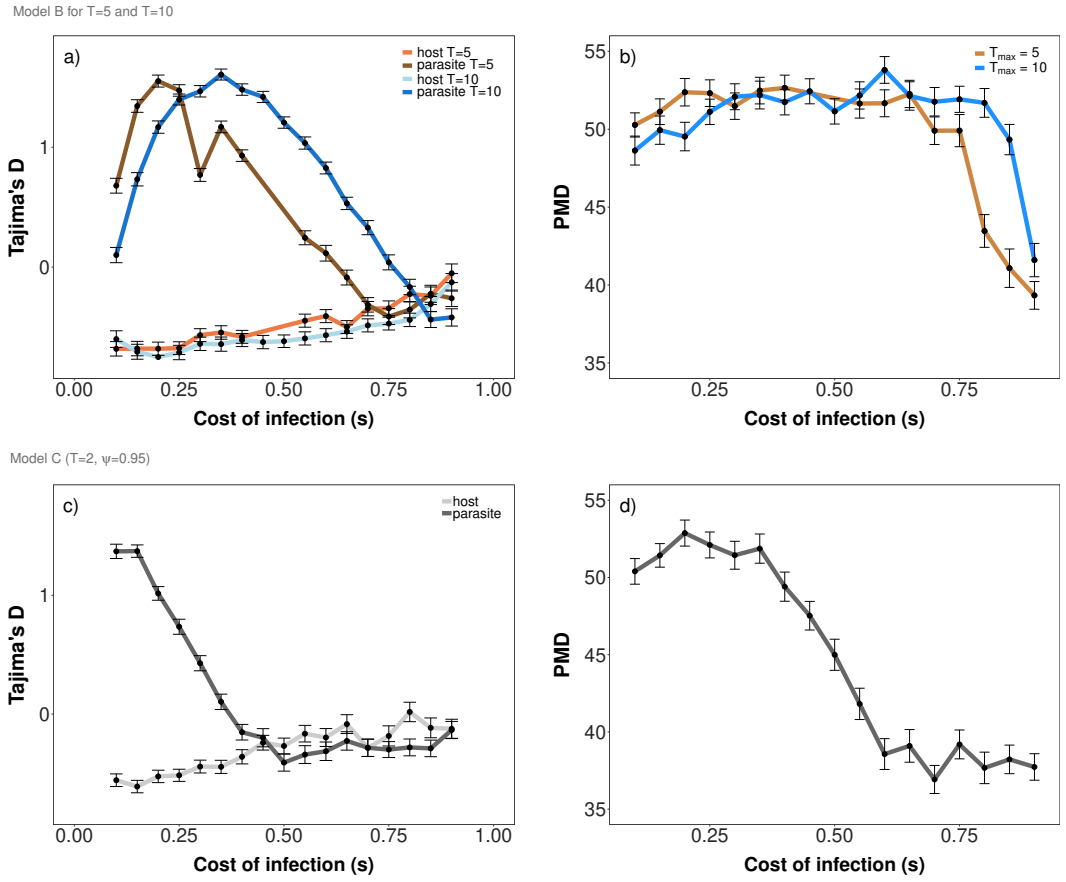


Fig. 5.A.3: Mean and standard error of Tajima's D (a+c) and pairwise manhattan distance (PMD) (b+d) for various costs of infection s (x-axis) and $r = 200$ repetitions. Results for **model 5B** (pure autoinfection model with $T = 5$ and $T = 10$) are shown at the top, results for **model 5C** (auto-allo-infection model with $\psi = 0.95$) are shown at the bottom. The other parameters are fixed to: $c_H = 0.05$ and $c_P = 0.1$. Initial frequencies R_0 and a_0 in a and b are chosen randomly from a uniform distribution between 0 and 1 while $R_0 = a_0 = 0.2$ in c and d.

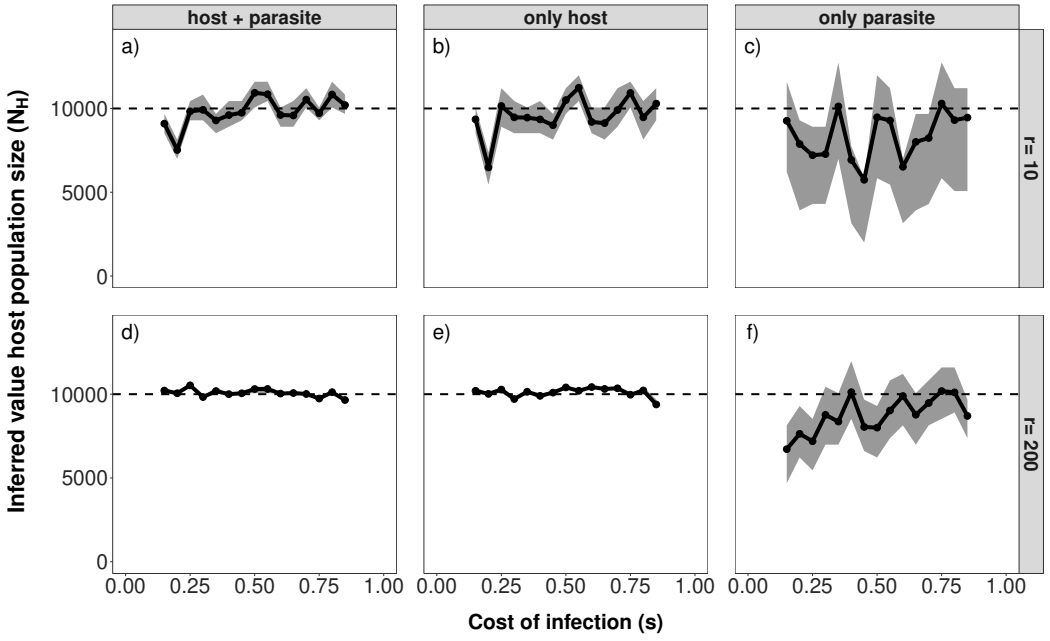


Fig. 5.A.4: Median of the posterior distribution (y-axis) for the host population size N_H against the true cost of infection s for $r = 10$ (top, a-c) and $r = 200$ (bottom, d-f). The inference results for scenario 1 based on host and parasite polymorphism data (left, a+d), host polymorphism data only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

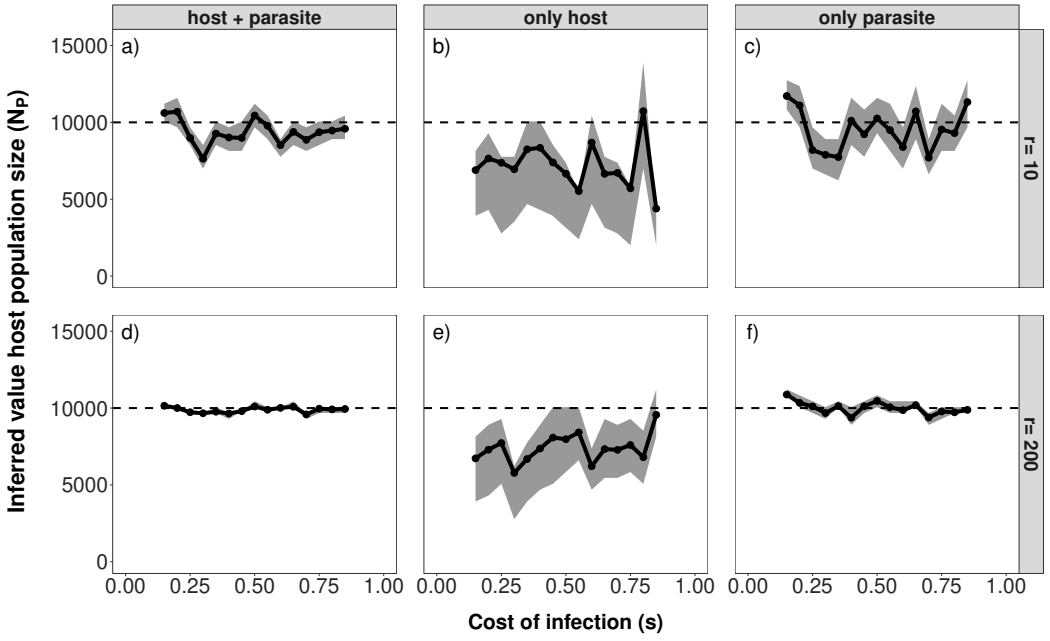


Fig. 5.A.5: Median of the posterior distribution (y-axis) for the parasite population size N_P against the true cost of infection s for $r = 10$ (top, a-c) and $r = 200$ (bottom, d-f). The inference results for scenario 1 based on host and parasite polymorphism data (left, a+d), host polymorphism data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

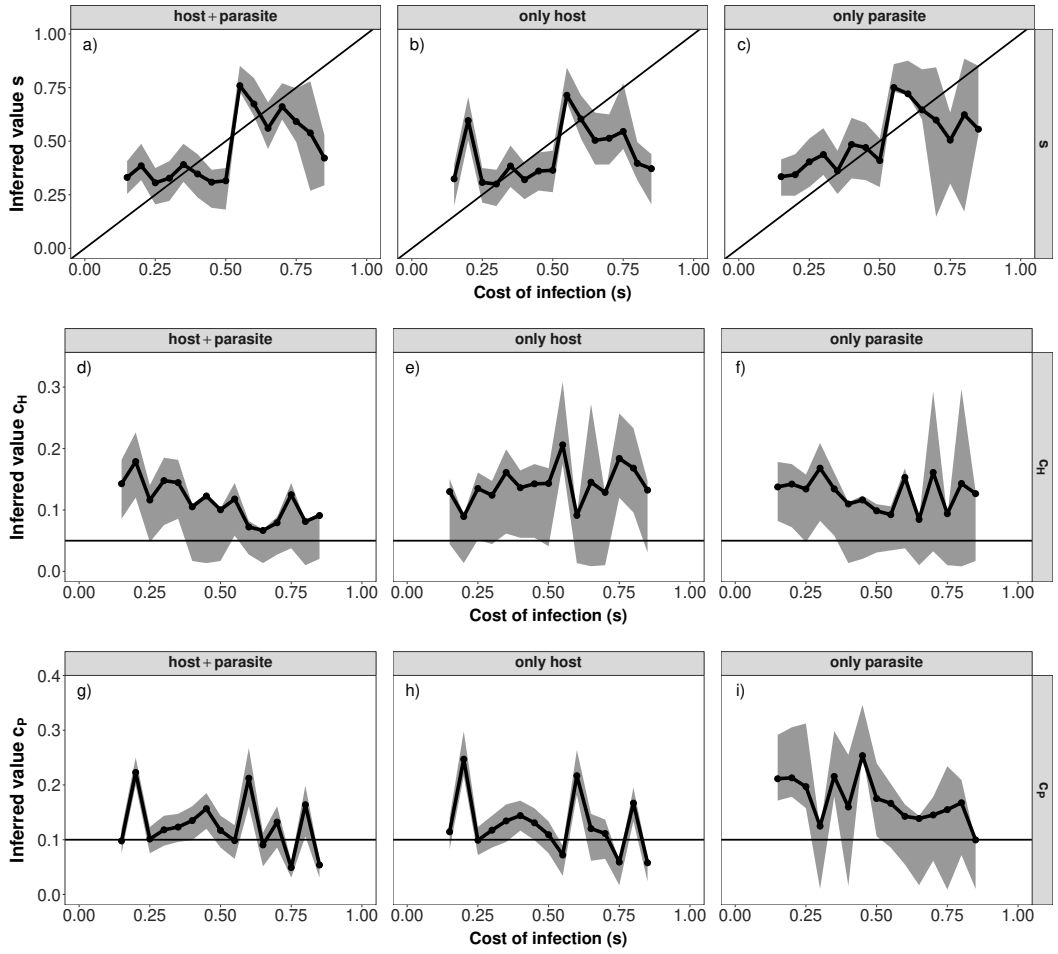


Fig. 5.A.6: Median of the posterior distribution (y-axis) for the cost of infection s (a-c), cost of resistance c_H (d-f) and cost of infectivity c_P (g-i) compared to the true values (x-axis) for $r = 10$. The results are shown for inference based on host and parasite polymorphism data (left, a+d+g), host polymorphism data only (middle, b+e+h) and parasite polymorphism data only (right, c+f+i). The fixed parameters are chosen as: $N_H = N_P = 10,000$, $\mu_{Rtor} = \mu_{ntoI} = \mu_{rtoR} = \mu_{Iton} = 10^{-5}$, $g_{max} = 30,000$, $\theta_H = 5$, $\theta_P = 5$, $n_H = 50$ and $n_P = 50$. Priors have been chosen as follows: s uniform$(0.1, 0.9)$, c_H uniform$(0.01, 0.35)$, c_P uniform$(0.01, 0.35)$. s, c_H and c_P are inferred simultaneously for all plots.

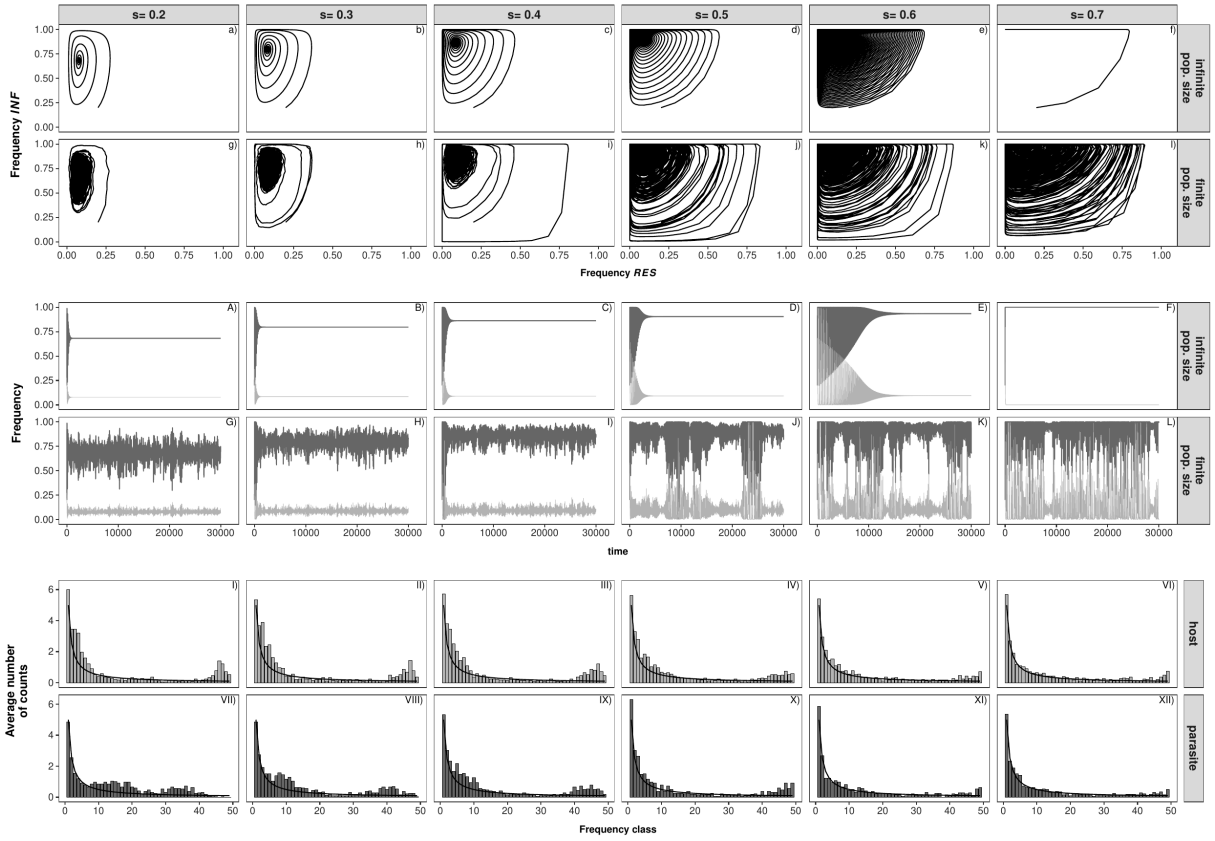


Fig. 5.A.7: Influence of the cost of infection (s) on the coevolutionary dynamics and genomic signatures in **model 5A**. The subfigures show the allele frequency trajectory in infinite population size (a-f, A-F), one exemplary allele frequency path in finite population size which takes genetic drift and functional mutations into account (g-l, G-L), the average unfolded host site frequency spectrum of $r = 200$ repetitions (I-VI) and the average unfolded parasite site frequency spectrum of $r = 200$ repetitions (VII-XII). In subfigures a-l each dot represents the frequency of resistant (RES) hosts (x-axis) and infective (INF) parasites (y-axis) at the beginning of a single host generation g. The same information is displayed in a slightly different way in subfigures A-L. Here, the frequencies of resistant (RES) hosts (light grey) and infective (INF) parasites (dark grey) (y-axis) are plotted over time (x-axis). Costs are fixed to $c_H = 0.05$, $c_P = 0.1$. The results in finite population size are plotted for $N_H = N_P = 10,000$, $\mu_{Rtor} = \mu_{ntoI} = \mu_{rtoR} = \mu_{Iton} = 10^{-5}$. The site frequency spectra are shown for $\theta_P = \theta_H = 5$ and $n_H = n_P = 50$.

5.B Supplementary information models

5.B.1 Model 5A

Detailed description how the allele frequency path is obtained

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

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

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

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

- the number of mutants M_{In} from *INF* to *ninf* is obtained by sampling from a Poisson distribution with rate $\lambda = \mu_{Iton} \cdot N_I$.
- the number of mutants M_{nI} from *ninf* to *INF* is obtained by sampling from a Poisson distribution with rate $\lambda = \mu_{ntoI} \cdot N_n$.

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

$$N_{x,I} = N_I - M_{In} + M_{nI} \tag{5.1}$$

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

$$\frac{N_{x,I}}{N_P} \tag{5.2}$$

The corresponding steps for the host population are as follows.

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

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

$$N_{g+1,R} = N_R - M_{Rr} + M_{rR} \tag{5.3}$$

Tab. 5.B.1: Fitness matrix for **model 5A** capturing the fitness effects of different interactions between host genotypes and parasites genotypes within a single host generation g. R_g (r_g) denotes the frequency of resistant (susceptible) hosts in generation g. $A_{g,1}$ ($A_{g,2}$) denotes the frequency of non-infective parasites and $a_{g,1}$ ($a_{g,2}$) denotes the frequency of infective parasites at the beginning of the first (second) parasite generation $t = 1$ ($t = 2$) within host generation g. The costs are: c_H=cost of resistance, c_P=cost of infectivity, s_1, s_2=cost of infection.

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

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

$$\frac{N_{g+1,R}}{N_H} \tag{5.4}$$

By repeating this procedure for $g_{max} = 3{\cdot}\max(N_H, N_P)$ generations we obtain the so called frequency path which consists of the frequencies of all four alleles at the beginning of each generation g. In order to constrain a modified version of msms (Ewing and Hermisson 2010; Tellier *et al.* 2014) by this frequency path we rescale the generations g in the host to $g^*_H = g/(2N_H)$ and in the parasite to $g^*_P = g/(2N_P)$. Note that msms is in diploid size. As N_H and N_P are haploid population sizes this rescaling is equivalent to rescale time in units of $4 \cdot N_{H,diploid}$ and $4 \cdot N_{P,diploid}$. Based on this time rescaled frequency path we launch msms once for the host and once for the parasite.

Tab. 5.B.2: Approximate frequencies of resistant hosts ($\hat{R}$) and infective parasites ($\hat{a}$) at the non-trivial interal equilbrium point in **model 5A** (Eq. 5.3) for various combinations of cost of resistance (c_H), cost of infectivity (c_P) and cost of infection (s) as plotted in Fig. 5.1.

panel of Fig. 5.1	c_H	c_P	s_1	s_2	$\hat{a}$	$\hat{R}$
a	0.05	0.10	0.20	0.10	0.667	0.081
a	0.05	0.10	0.30	0.15	0.775	0.089
a	0.05	0.10	0.40	0.20	0.835	0.094
a	0.05	0.10	0.50	0.25	0.873	0.097
a	0.05	0.10	0.60	0.30	0.899	0.100
a	0.05	0.10	0.70	0.35	0.919	0.102
a	0.05	0.10	0.80	0.40	0.934	0.104
b	0.10	0.10	0.20	0.10	0.424	0.068
b	0.10	0.10	0.30	0.15	0.603	0.077
b	0.10	0.10	0.40	0.20	0.704	0.084
b	0.10	0.10	0.50	0.25	0.771	0.089
b	0.10	0.10	0.60	0.30	0.818	0.092
b	0.10	0.10	0.70	0.35	0.853	0.096
b	0.10	0.10	0.80	0.40	0.881	0.098
c	0.05	0.30	0.20	0.10	0.667	0.291
c	0.05	0.30	0.30	0.15	0.775	0.324
c	0.05	0.30	0.40	0.20	0.835	0.347
c	0.05	0.30	0.50	0.25	0.873	0.363
c	0.05	0.30	0.60	0.30	0.899	0.375
c	0.05	0.30	0.70	0.35	0.919	0.384
c	0.05	0.30	0.80	0.40	0.934	0.392
d	0.10	0.30	0.20	0.10	0.424	0.235
d	0.10	0.30	0.30	0.15	0.603	0.273
d	0.10	0.30	0.40	0.20	0.704	0.301
d	0.10	0.30	0.50	0.25	0.771	0.323
d	0.10	0.30	0.60	0.30	0.818	0.340
d	0.10	0.30	0.70	0.35	0.853	0.354
d	0.10	0.30	0.80	0.40	0.881	0.366

5.B.2 Model 5B

Model 5B extends the basic model to $T > 2$ parasite generations per host generation. As in the basic model the cost of infection s_t is a function of the parasite generation t in which the host became infected and the maximum cost of infection s, which correspond to the cost of being infected at the first parasite generation t=1 within host generation g. Upon infection a host stays infected until it reproduces and dies from natural death (at the end of the host generation g). An infected host is reinfected by the offspring of the particular parasite for all subsequent parasite generations within host generation g (100% auto-infection). Hosts which have not been infected so far can be attacked by the offspring of any parasite type at the beginning of each parasite generation t. Whether this interaction subsequently results in an infection depends on the infection matrix. The recursion equations for this model are given by:

$$a_{g,t+1} = \frac{(1-c_P)\left[a_{g,1} + \sum_{l=2}^{t} a_{g,l} R_g \prod_{m=1}^{l-1} A_{g,m}\right]}{(1-c_P)\left[a_{g,1} + \sum_{l=2}^{t} a_{g,l} R_g \prod_{m=1}^{l-1} A_{g,m}\right] + A_{g,1} r_g} \tag{5.5a}$$

$$a_{g+1,1} = \frac{(1-c_P)\left[a_{g,1} + \sum_{l=2}^{T} a_{g,l} R_g \prod_{m=1}^{l-1} A_{g,m}\right]}{(1-c_P)\left[a_{g,1} + \sum_{l=2}^{T} a_{g,l} R_g \prod_{m=1}^{l-1} A_{g,m}\right] + A_{g,1} r_g} \tag{5.5b}$$

$$a_{g,2} = \frac{(1-c_P)\cdot a_{g,1}}{(1-c_P) a_{g,1} + A_{g,1} r_g} \tag{5.5c}$$

$$R_{g+1} = \frac{R_g \cdot (1-c_H)\left((1-s_1)\, a_{g,1} + \sum_{t=2}^{T}\left((1-s_t) a_{g,t} \prod_{l=1}^{t-1} A_{g,l}\right) + \prod_{t=1}^{T} A_{g,t}\right)}{R_g \cdot (1-c_H)\left((1-s_1)\, a_{g,1} + \sum_{t=2}^{T}\left((1-s_t) a_{g,t} \prod_{l=1}^{t-1} A_{g,l}\right) + \prod_{t=1}^{T} A_{g,t}\right) + r_g (1-s_1)} \tag{5.5d}$$

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

5.B.3 Model 5C

Model 5C is based on model C in (Tellier and Brown 2007b). As in model 5A, we assume $T = 2$ discrete parasite generations per discrete host generation g and frequency-dependent disease transmission. Parasites of the second ($t = 2$) generation within host generation g infect the same host individual as there parent at rate ψ (auto-infection) or a different host at rate $1-\psi$ (allo-infection). A host which is infected throughout the whole host generation g looses the amount $s_1 = s$ (cost of infection) of its fitness. If it is only infected during a single parasite generation the cost of infection reduces to $s_2 = \frac{s}{2}$. The respective fitness matrix is shown in table 5.B.3 with $A_{g,t}$ ($a_{g,t}$) denoting the frequency of *ninf* (*INF*)-parasites in the t-th parasite generation within host generation g and R_g (r_g) denoting the frequency of *RES* (*res*)-hosts in host generation g.

$$a_{g,2} = \frac{a_{g,1} \cdot (1 - c_P)}{a_{g,1} \cdot (1 - c_P) + A_{g,1} \cdot r_g} \tag{5.6a}$$

$$a_{g+1,1} = \frac{(1 - c_P) \cdot (R_g A_{g,1} a_{g,2} + r_g A_{g,1} a_{g,2}(1 - \psi) + a_{g,1}[\psi + a_{g,2}(1 - \psi)])}{r_g \cdot (\psi A_{g,1} + A_{g,2}(1 - \psi)) + (1 - c_P) \cdot (R_g A_{g,1} a_{g,2} + r_g A_{g,1} a_{g,2}(1 - \psi) + a_{g,1}[\psi + a_{g,2}(1 - \psi)])} \tag{5.6b}$$

$$R_{g+1} = \frac{\begin{array}{r} R_g \cdot (1 - c_H)(A_{g,1}A_{g,2} + (1 - s_2)(A_{g,1}a_{g,2} + a_{g,1}A_{g,2}(1 - \psi)) \\ + (1 - s_1)(a_{g,1}\psi + a_{g,1}a_{g,2}(1 - \psi))) \end{array}}{\begin{array}{r} R_g \cdot (1 - c_H)\,(A_{g,1}A_{g,2} + (1 - s_2)(A_{g,1}a_{g,2} + a_{g,1}A_{g,2}(1 - \psi)) \\ +(1 - s_1)(a_{g,1}\psi + a_{g,1}a_{g,2}(1 - \psi))) + r_g(1 - s_1) \end{array}} \tag{5.6c}$$

The allele frequency path for this model is obtained in the same way as in **model 5A**.

Fitness matrix

Tab. 5.B.3: Fitness matrix for **model 5C** capturing the fitness effects of different interactions between hosts and parasites within a single host generation g. R_g (r_g) denotes the frequency of resistant (susceptible) hosts in generation g. $A_{g,1}$ ($A_{g,2}$) denotes the frequency of non-infective parasites and $a_{g,1}$ ($a_{g,2}$) denotes the frequency of infective parasites at the beginning of the first (second) parasite generation $t = 1$ ($t = 2$) within host generation g. n/a indicates that these hosts have not been infected as *ninf*-parasites fail to infect *RES*-hosts. The costs are: c_H=cost of resistance, c_P=cost of infectivity, s_1, s_2=costs of infection.

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

5.C Supplementary information pairwise manhattan distance

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

$$PMD = \sum_{i=1}^{n-1} |\xi_{H,i} - \xi_{P,i}| \tag{5.1}$$

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

5.D Supplementary tables

Tab. 5.D.1: Overview of all parameters and variables used in this paper.

name	standard value	Description
c_H	0.05	cost of resistance
s	0.3	cost of infection
c_P	0.10	cost of virulence
ψ	1	auto-infection rate
R_0	0.20	initial frequency of *RES*-hosts
a_0	0.20	initial frequency of *INF*-parasites
R_g		frequency of *RES*-hosts in generation g
r_g		frequency of *res*-hosts in generation g
$A_{g,t}$		frequency of *ninf*-parasites at the beginning of the t-th parasite generation within host generation g
$a_{g,t}$		frequency of *INF*-parasites at the beginning of the t-th parasite generation within host generation g
g_{max}	$30,000$	number of host generations to simulate

Tab. 5.D.1 Continued:

T	2	number of parasite generations within host generation g
g	-	counter for host generations
t	-	counter for parasite generations within host generation g
n_H	50	sample size host
n_P	50	sample size parasite
N_H	$10,000$	haploid host population size
N_P	$10,000$	haploid parasite population size
θ_H	5	neutral population mutation rate host
θ_P	5	neutral population mutation rate parasite
μ_{Rtor}	10^{-5}	mutation rate *RES* to *res*
μ_{rtoR}	10^{-5}	mutation rate *res* to *RES*
μ_{ntoI}	10^{-5}	mutation rate *ninf* to *INF*
μ_{Iton}	10^{-5}	mutation rate *INF* to *ninf*
$\mu_{neutral}$	10^{-7}	neutral mutation rate per bp
N_R		number of *RES*-hosts after selection and genetic drift
N_r		number of *res*-hosts after selection and genetic drift
N_n		number of *ninf*-parasites after selection and genetic drift
N_I		number of *INF*-parasites after selection and genetic drift
M_{Rr}		current number of mutants from *RES* to *res*
M_{rR}		current number of mutants from *res* to *RES*
M_{nI}		current number of mutants from *ninf* to *INF*
M_{In}		current number of mutants from *INF* to *ninf*
$N_{g+1,R}$		number of *RES*-hosts in host generation $g+1$
$N_{x,I}$		number of *INF*-parasites in parasite generation x

Tab. 5.D.2: Summary statistics calculated for the pseudo-observed data sets

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

Tab. 5.D.3: Settings which have been used for running Approximate Bayesian Computation for the different scenarios and number of repetitions.

Scenario S	nb repetitions r	information used	'known' parameters	inferred parameters	prior distribution	complex parametes
$S = 1$	200	host + parasite	model 5A, $c_H = 0.05$, $c_P = 0.10$, $T = 2$, $\psi = 1$	s	unif$(0.1, 0.9)$	$\theta_H = N_H/1000$
				N_H	logunif$(2,000, 20000)$	$\theta_P = N_P/1000$
				N_P	logunif$(2,000, 20000)$	$g_{max} = 3 \cdot \max(N_H, N_P)$
$S = 1$	200	host	model 5A, $c_H = 0.05$, $c_P = 0.10$, $T = 2$	s	unif$(0.1, 0.9)$	$\theta_H = N_H/1000$
				N_H	logunif$(2,000, 20000)$	$\theta_P = N_P/1000$
				N_P	logunif$(2,000, 20000)$	$g_{max} = 3 \cdot \max(N_H, N_P)$
$S = 1$	200	parasite	model 5A, $c_H = 0.05$, $c_P = 0.10$, $T = 2$	s	unif$(0.1, 0.9)$	$\theta_H = N_H/1000$
				N_H	logunif$(2,000, 20000)$	$\theta_P = N_P/1000$
				N_P	logunif$(2,000, 20000)$	$g_{max} = 3 \cdot \max(N_H, N_P)$
$S = 1$	10	host + parasite	model 5A, $c_H = 0.05$, $c_P = 0.10$, $T = 2$	s	unif$(0.1, 0.9)$	$\theta_H = N_H/1000$
				N_H	logunif$(2,000, 20000)$	$\theta_P = N_P/1000$
				N_P	logunif$(2,000, 20000)$	$g_{max} = 3 \cdot \max(N_H, N_P)$
$S = 1$	10	host	model 5A, $c_H = 0.05$, $c_P = 0.10$, $T = 2$	s	unif$(0.1, 0.9)$	$\theta_H = N_H/1000$
				N_H	logunif$(2,000, 20000)$	$\theta_P = N_P/1000$
				N_P	logunif$(2,000, 20000)$	$g_{max} = 3 \cdot \max(N_H, N_P)$
$S = 1$	10	parasite	model 5A, $c_H = 0.05$, $c_P = 0.10$, $T = 2$	s	unif$(0.1, 0.9)$	$\theta_H = N_H/1000$
				N_H	logunif$(2,000, 20000)$	$\theta_P = N_P/1000$
				N_P	logunif$(2,000, 20000)$	$g_{max} = 3 \cdot \max(N_H, N_P)$
$S = 2$	200	host + parasite	model 5A, $N_H = 10,000$, $N_P = 10,000$, $T = 2$, $g_{max} = 30,000$	s	unif$(0.1, 0.9)$	-
				c_H	unif$(0.01, 0.35)$	-
				c_P	unif$(0.01, 0.35)$	-
$S = 2$	200	host	model 5A, $N_H = 10,000$, $N_P = 10,000$, $T = 2$, $g_{max} = 30,000$	s	unif$(0.1, 0.9)$	-
				c_H	unif$(0.01, 0.35)$	-
				c_P	unif$(0.01, 0.35)$	-
$S = 2$	200	parasite	model 5A, $N_H = 10,000$, $N_P = 10,000$, $T = 2$, $g_{max} = 30,000$	s	unif$(0.1, 0.9)$	-
				c_H	unif$(0.01, 0.35)$	-
				c_P	unif$(0.01, 0.35)$	-
$S = 2$	10	host + parasite	model 5A, $N_H = 10,000$, $N_P = 10,000$, $T = 2$, $g_{max} = 30,000$	s	unif$(0.1, 0.9)$	-
				c_H	unif$(0.01, 0.35)$	-
				c_P	unif$(0.01, 0.35)$	-
$S = 2$	10	host	model 5A, $N_H = 10,000$, $N_P = 10,000$, $T = 2$, $g_{max} = 30,000$	s	unif$(0.1, 0.9)$	-
				c_H	unif$(0.01, 0.35)$	-
				c_P	unif$(0.01, 0.35)$	-
$S = 2$	10	parasite	model 5A, $N_H = 10,000$, $N_P = 10,000$, $T = 2$, $g_{max} = 30,000$	s	unif$(0.1, 0.9)$	-
				c_H	unif$(0.01, 0.35)$	-
				c_P	unif$(0.01, 0.35)$	-

6
Locus specific and genome wide signatures of host-parasite coevolution under eco-evolutionary feedbacks

6.1 Introduction

During the past decades there is accumulating empirical evidence of ecological and evolutionary dynamics taking place at comparable timescales involving bidirectional feedback loops (Hiltunen and Becks 2014; Fronhofer and Altermatt 2015; Frickel *et al.* 2016). A process often being termed as eco-evolutionary feedback or ecogenetic feedback (Bailey *et al.* 2009; Post and Palkovacs 2009; Schoener 2011; Kokko and López-Sepulcre 2007). Ecological dynamics encompass changes in population size whereas evolutionary changes are mainly concerned with changes at the genic level such allele frequency changes.
It has been shown that eco-evolutionary feedback loops can take place in host-parasite coevolution (Frickel *et al.* 2016). Host-parasite coevolution is defined as reciprocal changes in trait distributions and allele frequencies due to selective pressures two interacting species exert on one another (Janzen 1980). Parasite population sizes can fluctuate across several orders of magnitude resulting from factors such varying success of between host-transmission, seasonally fluctuating epidemiological dynamics and temporal fluctuations in host availability. On the other hand parasites have the potential to severly reduce host population size due diseae induced mortality as for example shown in bats (Frick *et al.* 2010), amphibians (Berger *et al.* 1998) and *Daphnia magna* (Ebert *et al.* 2000). The resulting genetic changes can be rather fast as it has been demonstrated for barley mildew adapting rapidly to newly introduced resistance gene in the UK (Brown 2015) or phage $\phi 2$ rapidly responding to evolution of *Pseudomonas fluorescens* (Paterson *et al.* 2010). Rapid adaptive changes on the genetic level are not limited to the parasite but have been also demonstrated in hosts such as *Drosophila* (Obbard *et al.* 2006).
Coevolutionary dynamics at the underlying loci can be placed into continuum between two extremes being known as arms-race and trench-warfare (Red Queen dynamics). Arms-race dynamics are characterized by recurrent rapid and reciprocal increases of new ben-

eficial host resistance and parasite virulence genes ultimately resulting in their fixation (Woolhouse *et al.* 2002; Dawkins and Krebs 1979). In constrast several alleles exhibit temporally fluctuating frequencies in trench-warfare dynamics (Woolhouse *et al.* 2002; Stahl *et al.* 1999). The classic expectation is that arms-race dynamics result in signatures of selective sweeps such as negative Tajima's D whereas trech-warfare dynamics result in signatures of balancing selection.

Our understanding of host-parasite coevolutionary dynamics and concerning genes (and genomic architecture) underlying host-parasite coevolution has been greatly improved by the increasing availability of sequencing data. So far most of the methods for detecting genes under coevolution have been based on analysing both of the coevolving partners independently using single species selection detection methods such as outlier identification or genome-wide association analysis (GWAS) (Dalman *et al.* 2013; Karasov *et al.* 2014; Hartmann *et al.* 2018; Talas *et al.* 2016). Although, such single species analysis have improved our understanding of host-parasite ceovolution they have particular shortcomings. First, they might discard some additional information contained in the genomic data of the coevolving partner arising from their intermingled evolutionary histories. Only recently methods analyzing both interacting partners simultaneously have been developed and demonstrated their benefits (MacPherson *et al.* 2018; Nuismer *et al.* 2017; Wang *et al.* 2018). Second, outlier detection methods require information on the underlying demographic history in order to choose appropriate thresholds for the detection of outliers. Eco-evolutionary feedbacks resulting from host-parasite coevolution can result in persistent population size fluctuations (henceforward termed co-demographic population size fluctuations). It has been shown that these co-demographic population size fluctuations can result in detectable signatures in genome-wide neutral polymorphism data (Živković *et al.* 2019) if time samples are available. Thus, the appropriateness of establishing the demography based on genome-wide neutral polymorphism data from a single time-point is questionable in such cases. Using time-samples rather than single time point data can be used to improve our understanding of the temporal fluctuations of host and parasite population sizes due to coevolution. For example Foll *et al.* (2014) make use of time-sampled whole-genome data in a two-round ABC to first estimate changes in effective population size of influenza viruses and use the obtained posterior distribution to infer selection coefficients for each locus.

Furthermore, little is known so far about how the signatures at the coevolving loci deviate from the classic predictions (selective sweep or balancing selection) depending on the

amplitude and period of the co-demographic population size fluctuations.
Therefore, our first goal is to understand how the genomic signatures at the coevolving loci are affected by co-demographic population size fluctuations arising from eco-evolutionary feedback. Therefore, we extend an already existing foward-in time simulator (Živković *et al.* 2019) which couples a susceptible-infected-model with a population genetic coevolution model to simulate the genome-wide neutral site frequency spectrum (SFS) over time. We extend it in such a way that we can explicitly keep track of polymorphism patterns and haplotype distributions at the coevolving loci in both species simultaneously over time.
Second, we seek to understand how well whole-genome samples from different time-points in combination with unbiased estimators of genetic drift are suited to infer the demographic history arising from host-parasite eco-evolutionary feedback. In this context, we also aim to devise optimal sampling schemes in terms of sampling intervals and sample sizes. Therefore, our simulator not only simulates the genome-wide neutral SFS over time but also keeps track of allele frequency changes at several independent neutral loci over time.
Third, we are interested to which extend the power to detect coevolutionary loci can be increased when host and parasite whole genome samples from several rather than single time-points are available.
Fourth, we aim to investigate whether biological properties such as disease transmission rate, cost of resistance or selection coefficients for the loci under coevolution can be retrieved from time-sampled whole genome data of the host and the parasite by extending the ABC in chapter 5 to an approach similar to the one in Foll *et al.* (2014) and Foll *et al.* (2015). Therefore, we extend the forward-in-time simulator by Živković *et al.* (2019) in such a way that we can keep track of 1) the coevolutionary dynamics, namely the allele frequency changes at the coevolutionary loci and the host and parasite population size changes, 2) the genome-wide neutral host and neutral parasite SFS, 3) the allele frequency changes at 2000 independent and neutral host and parasite SNPs and 4) the haplotypes at the coevolving host and parasite loci. This simulator will allow us to investigate the short-term and long-term consequences of host-parasite coevolution on the genomic diversity at the coevolving genes and at a genome-wide level.

6.2 Methods and material

In the next few paragraphs, we outline in detail how all these different properties are implemented and in which way they will be used to answer our research questions.

6.2.1 Coevolution model

In order to simultaneously keep track of the epidemiological dynamics (population size changes) and evolutionary dynamcis (allele frequency changes) we use an susceptible-infected (SI) model which is coupled with a population genetics coevolution model (Živković *et al.* 2019). In the used SI-model, the number of individuals in the healthy host compartment increases due to birth of newborns from the healthy and infected class and decreases due to natural death and healthy individuals receiving an infection from already infected individuals. The number of infected individuals increases due to new infections and decreases by natural death and disease induced death (virulence).
We integrate a population genetic component into the SI-model, by assuming that there is a total number of two host and two parasite types. These host and parasite types are determined by a single locus with two alleles each. As we further assume haploid hosts and haploid parasites, the phenotype and the genotype are the same in our model. Whether a given host genotype i ($i \in \{1, 2\}$) can be infected by a given parasite genotype j ($j \in \{1, 2\}$) depends on the so called infection matrix α (see section 3.1.1).
Our SI-model keeps track of the number of susceptible hosts of type i (S_i) and the number of host of type i which are infected by a parasite of type j (I_{ij}). Hence, there is a total of six compartments in our model. The disease is only transmitted horizontally at rate β, the disease transmission rate and disease transmission is density dependent. The fitness of infected hosts decreases by an amount s (the cost of infection or the selection coefficient). Additionally, infected hosts die from the disease at rate γ, the disease induced mortality or virulence (*sensu* animal literature). Further, each host genotype i can be associated with some fitness cost c_{H_i}, such as a cost of resistance (Karasov *et al.* 2014; Bergelson and Purrington 1996; Tian *et al.* 2003). Similarly, each parasite genotype j can be associated with some fitness cost c_{P_j} ($c_{P_j} \in [0, 1]$), such as a cost of infectivity (Thrall and Burdon 2003). All hosts produce healthy offspring at rate b (birth rate) and die at rate d (death

rate). Therefore, the full model (Živković *et al.* 2019) writes as:

$$\begin{aligned}\frac{dH_i}{dt} &= H_i\left[b(1-c_{H_i}) - d - \sum_{j=1}^{2}\alpha_{ij}\beta(1-c_{P_j})\sum_{k=1}^{2}I_{kj}\right] + b(1-c_{H_i})\sum_{j=1}^{2}(1-s)I_{ij} \\ \frac{dI_{ij}}{dt} &= I_{ij}(-d-\gamma) + H_i\left[\alpha_{ij}\beta(1-c_{P_j})\sum_{k=1}^{2}I_{kj}\right]\end{aligned} \tag{6.1}$$

The total number of hosts of type i can be obtained as $H_i = S_i + \sum_j I_{ij}$. We obtain the effective host population size N_H by summing up the numbers in all six compartments. We assume that each host is solely infected by a single parasite. Therefore, the number of parasites of type j can be obtained as $P_j = \sum_i I_{ij}$ and the change in number is given by $\frac{dP_j}{dt} = \sum_i \frac{dI_{ij}}{dt}$. Accordingly, the effective parasite population size is given by $N_P = \sum_{j=1}^{2}\sum_{i=1}^{2} I_{ij}$.

To simulate the coevolutionary dynamics we discretize the continuous time model into small time steps of size $\delta_t = 0.001$. The choice of δ_t is based on previous analysis (Živković *et al.* 2019) and ensures that the discretized dynamics match the continuous-time behavior of the model. Thus, one discrete host generation t consist of $n_{max} = 1/\delta_t = 1000$ small time steps. In each small time step n we update the number of individuals in each compartment based on eq. 6.1. To do so, we rescale the rates $\beta,\ \gamma,\ b,\ d$ by a factor δ_t. Therefore, the change in number of healthy hosts of type i in the n-th small time step within host generation t is given by:

$$\begin{aligned}H_{i,n} - H_{i,n-1} = \delta_t\Bigg(H_{i,n-1}\Bigg[b(1-c_{H_i}) - d - \sum_{j=1}^{2}\alpha_{ij}\beta(1-c_{P_j})\sum_{k=1}^{2}I_{kj,n-1}\Bigg] \\ +b(1-c_{H_i})\sum_{j=1}^{2}(1-s)I_{ij,n-1}\Bigg)\end{aligned} \tag{6.2}$$

and the change in number of hosts of type i infected by parasites of type j is given by:

$$I_{ij,n} - I_{ij,n-1} = \delta_t\left(I_{ij,n-1}(-d-\gamma) + H_{i,n-1}\left[\alpha_{ij}\beta(1-c_{P_j})\sum_{k=1}^{2}I_{kj,n-1}\right]\right) \tag{6.3}$$

In our simulations, we allow for type changing mutations to take place at the coevolutionary loci. Every time a host (parasite) offspring is born by a parent of type i (j), this offspring is born with the other host (parasite) genotype k at rate μ_{f_H} (μ_{f_P}), the functional host (parasite) mutation rate. Thus, we assume (for now) symmetric functional mutation rates in the host and the parasite. But see the discussion in Kirby and Burdon (1997), that these mutation rates between functional types can be also fairly asymmetric. To include these type changing mutations into our simulations, we proceed as follows. For each small time step n we calculate the total number of newborn hosts of type i, $B_{i,n}$, based on the birth terms (terms which include b) in eq. 6.2. As we have discretized the dynamics into small time steps of size δ_t, this results in the birth of fractional individuals ($B_{i,n} < 1$) for most of the δ_t-time steps. Usually one would draw the number of type changing mutants $M_{i,n}$ from a Poisson distribution with $\lambda_{f_H} = \mu_{f_H} \cdot B_{i,n}$ where $B_{i,n}$ is the number of newborn individuals of host type i in the n-th small time-step. However, this gives rise to an numerical issue and a biologically motivated issue. The small mutation rate and the small number of newborn individuals in each single time step δ_t (frequently $B_{i,n} < 1$) are likely to produce numerical issues. In addition, from a biological perspective only a 'full' newborn host individual can mutate. Therefore, we use a little approximation. We count the total number of newborn individuals of type i $B_{i,TT'} = \sum_{n=T}^{T'} B_{i,n}$ in some time interval $T \to T'$ consisting of consecutive δ_t time steps. Once $B_{i,TT'} > 1$, and thus, at least one full host individual has been born by parents of type i, we allow for type changing mutations to take place by drawing a random number $M_{i,n}$ from a Poisson distribution with $\lambda = \mu_{f_H} \cdot B_{i,TT'}$. The number $M_{i,n}$ is the number of type changing mutants among the offspring of parents of type i. Accordingly, we decrease the number of healthy hosts of type i by $M_{i,n}$ and increase the number of healthy hosts of type k by $M_{i,n}$. After the type changing mutations have taken place we reset $B_{i,TT'} = 0$ and set $T = T' + 1$. We repeat this procedure until n_{max} and hence, the end of the generation is reached. Once n_{max} is reached we reset $B_{i,TT'}$ to zero. We proceed in a similar manner for the parasites. Note that new infections correspond to the birth of a parasite. For parasites we also have to keep track of host type i on which a parasite by a parent of type j was born. For example a type changing mutation in an offspring from a parasite of type $j = 1$ which is born on a host of type $i = 1$ decreases the number of I_{11} by one 1 and increases the number of infected hosts of type I_{12} by one. The number of offspring from

parasite parents of type j on hosts of type i at time n is given by:

$$B_{j,n} = \delta_t H_{i,n-1} \left[\alpha_{ij} \beta (1 - c_{P_j}) \sum_{k=1}^{2} I_{kj,n-1} \right] \tag{6.4}$$

6.2.2 Simulation of the neutral whole genome population site frequency spectrum

While simulating the coevolutionary dynamics we keep track of the unfolded neutral whole genome site frequency spectrum for both, the host and the parasite. The whole genome-site frequency spectrum summarizes the allele frequencies at neutral single nucleotide polymorphism (SNP) which are distributed across the whole genome. By definition, these neutral SNPs do not affect the infection outcome. We assume that they evolve independently of each other and independently of the coevolutionary locus as the recombination rate along the genome is high enough. By using the unfolded site frequency spectrum, we assume that the ancestral and derived state are known for each SNP. As neutral SNPs are not determining the infection outcome, we can track and update a single site frequency spectrum for the whole host (parasite) population and must not proceed separately for each host and each parasite types.

We initialize the population site frequency spectrum for each species under the assumptions of an infinite sites mutation model and mutation-drift balance. In an infinite sites models each site can only mutate once, and thus all neutral SNPs are bi-allelic. Given these assumptions, the expected unfolded population site frequency spectrum of species x, where x can either be the host or the parasite, is given by:

$$\xi_{k_x} = \frac{2N_{x,0}\mu_{G_x}}{k} = \frac{\theta_x}{k} \qquad k \in \{1, \ldots, N_x - 1\} \tag{6.5}$$

Here, $\mu_{G,x}$ is the genome-wide mutation rate per generation for species x and θ_x is the population mutation rate of species x. Note that as our model is based on haploid individuals, $N_{x,0}$ corresponds to the initial haploid population size of species x. The genome-wide mutation rate μ_{G_x} is obtained as the product of the mutation rate per site per generation μ_{l_x} and the total genome size (l_x).

$$\mu_n = \mu_{l_x} \cdot l_x \tag{6.6}$$

Accordingly, we start our simulation with a total number of S_x neutral SNPs.

$$S_x = \theta \sum_{k=1}^{N_{x,0}-1} \frac{1}{k} \tag{6.7}$$

The neutral site frequency spectrum of each species is updated after the SI-model has been iterated for n_{max} time-steps and hence, the end of a discrete generation is reached. As our SI-model allows for overlapping generations, namely when $d < 1$, the host and parasite population are composed of individuals which already lived in the previous generation, the overlap $O_{x,t}$, and newborn individuals $B_{x,t}$. This has to be taken into account when updating the SFS. Therefore, we first calculate the number of overlapping individuals in both species. The overlap in the host is obtained as:

$$O_{H,t} = (1 - d)\, N_{H,t-1} \tag{6.8}$$

and the overlap in the parasite is obtained as:

$$O_{P,t} = (1 - (d + \gamma))\, N_{P,t-1} \tag{6.9}$$

We base our calculation of overlap on the previous population size and not on the current population size as only indiviuals which already lived in the previous generation can survive and thus, be overlapping. If the population size rapidly declines between consecutive discrete generations and the number of overlapping individuals would become larger than the current population size ($O_{x,t} > N_{x,t}$), we set $O_{x,t} = N_{x,t}$. Note that in such cases the current population is only composed of overlapping individuals. Based on the number of overlapping individuals $O_{x,t}$ and the current population size, the number of newborn individuals ($B_{x,t}$) is obtained as:

$$B_{x,t} = N_{x,t} - O_{x,t} \tag{6.10}$$

where $N_{x,t}$ is the current population size of species x.
We use these numbers to update the SFS based on the previous SFS (from $t-1$). For each single SNP l which had some derived allele frequency k in the previous generation, the update consists of two sampling steps. First, we randomly sample the alleles of the overlap from a hypergeometric distribution (sampling without replacement). Second, we sample

the alleles of the newborns from a binomial distribution (sampling with replacement). The hypergeometric sampling for the overlap is performed with parameters N_{t-1}, k and $O_{t,x}$. The biological intuition behind is as follows: In the previous generation there have been N_{t-1} individuals among which k individuals had the derived allele at locus l and from which $O_{t,x}$ individuals are still living. As each individual can be only chosen once to be one of the surviving individuals, the sampling has to be performed without replacement and thus is hypergeometric. The random number drawn from the hypergeometric distribution corresponds to the number of overlapping individuals with the derived allele at SNP l. The newborns are sampled from a binomial distribution with parameters k/N_{t-1} and $B_{t,x}$. For neutral SNPs, each parent is equally likely to be the parent of a newborn individual as both alleles have the same fitness by definition (unless they are physically linked to a selected SNP). This is equivalent to let each newborn choose its parent at random from all $N_{x,t-1}$ individuals in the previous generation. Among these $N_{x,t-1}$ parents k parents had the derived allele at SNP l. Thus, the probability for each offspring to be born with the derived allele is $p = k/N_{t-1}$. Note that obligatorily offspring have the same allele as their parent as a consequence of the infinite-sites model assumption, which excludes the possibility of recurrent mutations. As in addition each parent can potentially produce several offspring the sampling is binomial. The random number drawn from the binomial distribution is the number of newborn individuals which have the derived allele at the considered SNP.

Thus, the current derived allele frequency at SNP l is obtained as the sum of the numbers drawn from the hypergeometric and binomial distribution. This procedure is applied to every neutral SNPs of the previous SFS. In addition, new SNPs can arise due to mutation in the $B_{x,t}$ newborn individuals. The number of new SNPs is drawn from a Poisson distribution with mean $\lambda = \mu_{G_x} \cdot B_{x,t}$. Each mutation produces a new SNP with frequency one and thus, increases the number of singletons in the current SFS by one.

Time samples genome wide site frequency spectrum

At each discrete generation t we output the expected sample site frequency spectrum for a sample of size n_x The sample site frequency spectrum can be obtained from the population site frequency spectrum as:

$$\rho_m = \binom{n_x}{m} \sum_{k=1}^{N-1} \xi_{k,t} \left(\frac{k}{N}\right)^m \left(1 - \frac{k}{N}\right)^{n_x - m} \tag{6.11}$$

where n_x is the sample size and ρ_m is the expected number of sites in the sample with minor allele frequency m ($m \in \{1, \ldots, n_x - 1\}$) and $\xi_{k,t}$ is the number of SNPs with minor allele frequency k in the population SFS at time t.

6.2.3 Tracking allele frequency changes at neutral loci

Further, we initialize $\tau_{n,x} = 2000$ neutral SNPs for which we explicitely track the allele frequency changes over time. We use the allele frequency changes at these SNPs to obtain an estimate of the effective population size between consecutive sampling time points by using a drift estimator proposed by Jorde and Ryman (2007). The inital allele frequencies of these neutral SNPs are drawn randomly from the initial population SFS under drift-mutation-equilibrium (see previous section). Accordingly, the probability that the derived allele frequency of the r-th tracked SNP in species x is k, is given by:

$$p(k) = \frac{\xi_k}{\sum_{i=1}^{N_{x,0}-1} \xi_i} \tag{6.12}$$

where $N_{x,0}$ is the initial population size. The allele frequencies for all neutral SNPs are updated at the end of each discrete generation based on the new total host (parasite) population size, the previous total host (parasite population size) and the overlap $(1-d)$. As for the update of the population site frequency spectrum the update for the explicitly tracked SNPs consists of two sampling steps, a binomial sampling for the newborn individuals and a hypergeometric sampling for the overlapping individuals.

Temporal samples for 2000 neutral SNPs

Every ten time steps, we draw a random sample of $n = 100$ individuals for each SNP r from a binomial distribution with parameter $p = \frac{k}{N_{x,t}}$. We acknowledge that for large $n/N_{x,t}$ ratio the assumption of a binomial sampling might be violated. In such cases one should rather sample from a hypergeometric than from a binomial distribution. We then apply the estimator from Jorde and Ryman (2007) to obtain an estimate of the amount of drift between the current and the previous sampling point. This estimator is based on measuring the allele frequency change between two sampling times being g generations

apart from each other. The estimator over all tracked neutral SNPs τ_x is given by:

$$\mathrm{Fs}' = \frac{\mathrm{Fs}(1-\frac{1}{4\tilde{n}})-\frac{1}{\tilde{n}}}{(1+\frac{\mathrm{Fs}}{4})(1-\frac{1}{2n_y})} \quad \text{with} \tag{6.13}$$

$$\mathrm{Fs} = \frac{\sum_{r=1}^{\tau_x}(\kappa_1-\kappa_2)^2}{\sum_{i=r}^{\tau_x} z_r(1-z_r)} \tag{6.14}$$

$$\tag{6.15}$$

where κ_1 is the derived allele frequency at SNP r in the first sample and κ_2 is the the derived allele frequency in the second sample which is g generations apart. z_r is the unweighted mean of the allele frequencies at locus r and is given by: $z_r = \frac{\kappa_1+\kappa_2}{2}$. $\tilde{n} = \frac{2}{\frac{1}{n_1}+\frac{1}{n_2}}$ is the harmonic mean of the diploid sample sizes at the two sampling points and n_2 is the diploid sample size for the second sample. Based on this an estimate of the effective population size between two time points being g generations apart from each other, is obtained as:

$$N_e = \frac{g}{\mathrm{Fs}'} \tag{6.16}$$

6.2.4 Simulation of the coevolutionary loci

As already introduced in the previous section we simulate the coevolution of a single bi-allelic host locus and a single bi-allelic parasite locus. Our coevolutionary locus consists of a so-called functional site which is surrounded by a number of l_{c_H} (l_{c_P}) neutral sites in the host (parasite). The functional site determines the allelic state at the coevolutionary locus and can be a single SNP, an insertion/deletion polymorphism or the absence of presence of a transposable element. We introduce the term background for the neutral sites at the coevolutionary locus. All sites at the coevolutionary locus (functional and neutral) are fully linked. Thus, we assume that recombination is absent or negligible at the coevolutionary locus. Throughout the simulation we keep track of all haplotypes existing at the coevolutionary locus. We initialize the host and parasite coevolutionary locus with two haplotypes each. Namely, one haplotype for each host genotype i and for each parasite genotype j. These two haplotypes only differ in the identity of the functional site. Thus, the initial absolute number of the first (second) host haplotype is H_{1_0} (H_{2_0}) and the initial number of the first (second) parasite haplotype is P_{1_0} (P_{2_0}). We update

the haplotypes at the end of each discrete generation t. In contrast to the update of the genome wide site frequency spectrum we have to update the haplotypes separately for each host genotype i and for each parasite genotype j. Therefore, we proceed as follows.

1. First, we calculate the number of overlapping individuals and newborn individuals for each host type i (parasite type j). The number of overlapping individuals for host type i is obtained as:

$$O_{i,t} = (1 - d)\, H_{i,t-1} \tag{6.17}$$

   Similarly, the number of overlapping individuals of parasite type j is obtained as

$$O_{j,t} = (1 - (d + \gamma))\, P_{j,t-1} \tag{6.18}$$

2. Based on the overlap we calculate the number of newborn hosts of type i as:

$$B_{i,t} = H_{i,t} - O_{i,t} \tag{6.19}$$

   and the number of newborn hosts of type j as

$$B_{j,t} = P_{j,t} - O_{j,t} \tag{6.20}$$

3. To take the functional mutations into account we draw the number of mutant offspring $M_{i,t}$ of host parents of type i from a Poisson distribution with mean $\lambda_{f_H} = \mu_{f_H} \cdot B_{i,t}$ and the number of mutant offspring $P_{j,t}$ of parasite parents of type j from a Poisson distribution with mean $\lambda_{f_P} = \mu_{f_P} \cdot B_{j,t}$

4. We search for all host (parasite) haplotypes with functional site i (j).

5. We sample the haplotypes of the overlapping individuals of host (parasite) type i (j) from all haplotypes with functional site i (j) of the previous generation. Therefore, we perform a hypergeometric sampling based on the absolute frequencies of all haplotypes with functional site i (j) in the previous generation

6. We obtain the haplotypes of the $B_{i,t} - M_{i,t}$ ($B_{j,t} - M_{j,t}$) non-mutated offspring of parents of host (parasite) type i (j). Therefore, we perform a multinomial sampling from all haplotypes with functional site i (j) in generation $t - 1$. Each newborn

individual by default inherits the haplotype of its parent. With probability $l_{c_H} \cdot \mu_{l_H}$ ($l_{c_P} \cdot \mu_{l_P}$) a mutation happens in the coevolutionary background of each newborn individual (Remember that μ_{l_H} (μ_{l_P}) is the host (parasite) per site mutation rate per generation). We assume that these neutral mutations also follow an infinite sites model. Every neutral mutation in the background creates a new haplotype and this haplotype is initially in frequency one.

7. We obtain the haplotypes of the $M_{i,t}$ mutated offspring of parents of host (parasite) type i (j). Therefore, we perform a multinomial sampling from all haplotypes with functional site i (j) in generation $t-1$. Each offspring inherits the background of its parents but the functional site changes to the other host (parasite) type. In contrast to the non-mutated $B_{i,t} - M_{i,t}$ ($B_{j,t} - M_{j,t}$) no mutations happen in the background of the mutated offspring, as we assume that to mutations happen at the coevolutionary locus in a single individual is negligible.

7 Discussion

Advances in sequencing technologies have largely increased the amount of host and parasite genomic data available. There are big hopes that such data provide a very valuable resource to improve our understanding of host-parasite coevolution. This includes elucidating the genomic basis of host-parasite coevolution on the one hand and inferring the past coevolutionary history on the other hand. One central aim of this thesis was to understand how and which information about the genomic basis of host-parasite coevolution and the past coevolutionary interactions can be extracted by jointly analysing host and parasite genomic data. We could show that by means of cross-species genome wide association studies it is possible to pinpoint the loci under coevolution. However, the ability to detect these loci varies over time and depends on the underlying GxG-interaction and the nature of the coevolutionary dynamics. Therefore, we suggested that samples from several time points could be potentially used to gain a deeper understanding about the underlying GxG interaction. We further could show in chapter 5 that under trench-warfare dynamics the genomic signatures at the coevolving genes are indicative about the allelic equilibrium frequencies. In the simple polycyclic model we used they were solely determined by several fitness costs. Thus, shifts in the signatures at the coevolving genes could be directly attributed to changing costs. However, there are distinct processes which can generate direct frequency- dependent selection and thus, favour trench-warfare dynamics. In the model used, direct frequency-dependent selection is created via auto-infections. However, other factors such as seed-banks, overlapping generation or spatially heterogeneous selection have the potential to create direct frequency-dependent selection (Brown and Tellier 2011). Thus, in such cases the genomic data will be only informative about the costs if they are the main determinants in shifting the equilibrium frequencies substantially.
Another aim was to provide additional methods for jointly analysing host and parasite data in order to gain more information on the past coevolutionary history. Therefore, we proposed two indices which could be potentially useful when performing cross-species GWAS. We hypothesize that these indices might be also useful to detect divergent coevolutionary interactions in two populations given that data from several time points are available. One possibility of such divergent selection could be that a host-parasite-locus

pair is coevolving according to trench-warfare dynamics in population A and there is no coevolution any longer in population B. For such cases it is expected that CSA is consistently high in population A but constantly low with some minor fluctuations in population B. If on the other hand the same host-parasite locus pair is coevolving in both populations, but the dynamics are having some time lag due to limited migration, one then expects to see the qualitatively same change in CSA over time.

Besides these indices we proposed a new approach which is based on coupling a coevolutionary model with an Approximate Bayesian Computation method to infer properties of the past coevolutionary history. Using this methods we could show that the genomic signatures at the coevolving genes contain information about host and parasite coevolutionary costs. Thus, this approach is an addition to the limited amount of methods which exist (MacPherson *et al.* 2018; Nuismer *et al.* 2017; Wang *et al.* 2018; Bartoli and Roux 2017; Bartha *et al.* 2013) to analyse jointly host and parasite genomic data and which take the reciprocal nature of coevolutionary interactions into account. Nuismer and Week (2019) have also proposed to use ABC in order to gain a better understanding about host-parasite arms-races. In contrast to our approach they used phenotypic data from several populations in order to estimate the intensity of reciprocal selection. Combining their ABC approach with our ABC approach and the cross-species indices could potentially yield refined insights into the past coevolutionary interaction.

As further aspects of our results have been already discussed at the end of each chapter we will now move to a more general discussion about our current understanding of genomic polymorphism patterns under coevolution and the value of genomic data to gain a better understanding on host-parasite coevolution. Here, we will address the following questions: 1) What is the value of genomic data for understanding host-parasite coevolution? 2) How much do we know about sequence evolution at the coevolutionary loci and at the genome-wide level under coevolution? 3) What is the value of analyzing host and parasite data in a joint framework? 4) How does the complexity of host-parasite coevolutionary interactions likely affect our ability to extract information from genomic data? 5) Why is theory important for analyzing genomic data of host-parasite coevolution? 6) How can existing methods to analyze genomic data of host-parasite coevolution be further improved? 7) How can we improve our understanding of host parasite coevolution in general?

7.1 The value of genomic data for understanding host-parasite coevolution

Analysis of host or parasite genomic data based on single species techniques such as outlier scans or genome-wide association studies have uncovered several host defense genes and parasite attack genes which show either signatures of positive or balancing selection (Schweizer *et al.* 2018; Dalman *et al.* 2013; Caicedo and Schaal 2004; Bergelson *et al.* 2001; Hoerger *et al.* 2012; Rose *et al.* 2007; Krishnan *et al.* 2018). Thus, genomic data have been proven to be a valuable source to identify genes which are potentially involved into coevolutionary interactions. These hits provide useful candidate genes for designing molecular experiments and for further characterisation of potential host-parasite interactions at the molecular level. However, signatures on their own do not contain any conclusive information about whether these genes have been coevolving and/or whether they are still coevolving and the exact nature of the underlying process. What are for example the likely causes of the strong balancing selection signatures at several resistance genes in plants (Karasov *et al.* 2014; Hoerger *et al.* 2012; Rose *et al.* 2007; Caicedo and Schaal 2004)? In a nice study which combined molecular, experimental approaches and sequencing approaches Karasov *et al.* (2014) suggested that the found balancing selection signature at the Rps5-gene of *A. thaliana* is likely the result of diffuse interactions with several parasite strains in their particular example. However in general, there are several distinct processes which can give rise to balancing selection signatures under host parasite coevolution. These include for example seed-banks, overlapping generations or epidemiological feedback loops (Brown and Tellier 2011). Thus, outlier scans as such only provide ways to detect polymorphism patterns which deviate from neutrality but do not give an indication about the underlying process. It is therefore important to keep in mind that outlier scans always need some validation in order to be conclusive.

Further, as we have shown in chapter 4 and 5 our ability to detect genes under coevolution based on genomic signatures is also crucially affected by the underlying dynamics of the coevolutionary interaction (for example the equilibrium frequencies, temporal changes of allele frequencies). For example the signatures of interactions where the equilibrium frequencies are close to the boundaries resemble almost neutral signatures. Therefore, it is important to keep in mind that the absence of strong signatures does not necessarily imply absence of coevolution as such. On the opposite hand, the presence of a strong

signature is not necessarily an indicator that the particular genes has been involved into coevolution (see Nuismer *et al.* (2010); Janzen (1980) for similar arguments concerning correlations).
Further, for devising optimal disease management strategies, it is crucial to not only detect loci under coevolution but also to gain an understanding of the underlying coevolutionary processes. The presented ABC method in chapter 5 is an approach to extract such information based on genomic data of the host and the parasite and thus, extends beyond classic outlier scans. However, the model used in the ABC has to be used with caution and as for outlier scans the results require a careful validation in order to avoid biased interpretations of the results.

7.2 Link between coevolutionary dynamics and sequence evolution at the coevolving loci

The classic prediction is that arms-race dynamics result in signatures of positive selection and trench-warfare dynamics result in signatures of balancing selection. The reasoning behind these simple predictions is straightforward. But taking the multitude of processes interacting with each other in host-parasite coevolution into account (chapter 2), one should expect that the genomic signatures at the coevolving genes rather fall into continuum than map to a strict dichotomy. Such continua of signatures have been for example found in Tellier *et al.* (2014) and chapter 5. Generally speaking, this demonstrates the necessity to gain a refined understanding on how coevolutionary dynamics in finite population size in combination with the biology of the coevolving species and the abiotic and biotic environment link to genomic signatures at the coevolving genes.
As we have shown in chapter 5 the amount of genetic drift, which is intimately connected to the demographic history, can interact with and alter the coevolutionary dynamics at the coevolving genes. This can result in the stochastic extinction of alleles which in turn has subsequent effects on the resulting signatures. So far, most studies have only considered the effect of drift by assuming a Wright-Fisher model (chapter 3 and 5 of this thesis and Tellier and Brown 2007b; Kirby and Burdon 1997; Tellier *et al.* 2014; Salathe *et al.* 2005). However, the Wright-Fisher model is based on the assumption that the number of offspring per individual is Poisson distributed with mean one. But, parasites can be

characterized by highly skewed offspring distributions (Irwin *et al.* 2016). This implies that some parasite individuals can potentially contribute a disproportionately large fraction of offspring and thus, the whole parasite population is composed by the offspring of only a few individuals. It has been shown that the time to fixation of a beneficial mutant (conditional on fixation) is substantially decreased in populations with skewed offspring distributions (Eldon and Stephan 2018). Further, the chance of loosing beneficial mutants due to drift is also increased (Eldon and Stephan 2018). As the changes in host and parasite allele frequencies feedback on each other in host-parasite coevolution it is very likely that such forms of genetic drift have an impact on the coevolutionary dynamics and thus on the genomic signatures at the coevolving loci. Skewed offspring distributions are rather characterised by a multiple merger coalescent then a standard coalescent. In multiple-merger coalescents there is a substantial chance that more than two individuals find a common ancestor in the previous generation. Multiple merger coalescents violate the assumptions of classic outlier scans or coalescent simulators such as msms (Ewing and Hermisson 2010) which should be kept in mind when performing outlier scans.
Besides the effect of skewed offspring distributions on the amount of drift, the effect of eco-evolutionary feedbacks on the signatures at the coevolving loci has not been theoretically investigated so far. However, it is important to understand such effects as host-parasite coevolution can result in eco-evolutionary feedback loops (Frickel *et al.* 2016). Population size changes which arise from eco-evolutionary feedback loops determine the effect of genetic drift and the supply of new mutations via the population mutation rate. Strong population size fluctuations can result in recurrent bottlenecks which are likely to cause a depletion of genetic variation at the coevolving loci and have been shown to affect the genome-wide site frequency spectrum (Živković *et al.* 2019). This effect has the potential to weaken the strength of balancing selection signatures although the alleles might be maintained stabily for a long period of time. In addition strong strong bottlenecks can also cause the random loss of alleles and generate multiple merger coalescent signatures (Tellier and Lemaire 2014).
Thus, besides collecting increasing amounts of genomic data and asking for the development of new inferences methods, it is also crucial to improve our understanding on how the interplay between coevolutionary dynamics, the biology of the species and eco-evolutionary feedbacks shapes the polymorphism data at the coevolving loci. This understanding is the basis for developing new meaningful methods and the correct interpretation of any population genomic analysis.

7.3 Analysing host and parasite data in a joint framework

All chapters in thesis are based on the assumption that host and parasite polymorphism data are jointly analysed. This approach has several advantages compared to pure single species techniques but also introduces some potential difficulties in certain circumstances. Due to the fact that allele frequency changes at the coevolving loci feedback on one another the coevolutionary histories of the host and the parasite locus are inevitably intermingled with each other. Therefore, the polymorphism data at both loci have bits of information about their joint coevolutionary history. Thus, the joint analysis of data from both partners is likely to give a much clearer picture about the past coevolutionary dynamics and histories. This is for example also resembled by the fact that host (parasite) equilibrium frequencies are determined by costs which apply to the coevolving partner and vice-versa (Tellier and Brown 2007b; Frank 1992). Additionally, we have shown in chapter 5 that the signature in either species is presumably most indicative about costs which applied to the coevolutionary partner. This also explained why analysing the host and the parasite together greatly improved the accuracy of our inference approach in chapter 5. Further, Wang *et al.* (2018) and MacPherson *et al.* (2018) could show that the effect size of single species GWAS depends on the allele frequency in the coevolving partner. In addition, Živković *et al.* (2019) could show that the co-demographic changes arising from host-parasite coevolution are detectable in time-samples from the parasite but not in time samples of the host. Thus, the signal of co-demographic changes is missed when only analysing the host. When performing further analysis for chapter 6 we also expect that jointly analysing the host and the parasite over time will give a better indication about the temporal changes in allele frequency and population sizes. Therefore, the inference of co-evolutionary parameters using this simulator should be also largely improved when analysing the host and the parasite together. Based on all these examples, analysing both partners simultaneously is a natural choice as such analysis take the interdependence of both coevolving partners explicitly into account. In the light of decreasing sequencing costs, we should therefore aim to perform more of such co-species approaches (Wang *et al.* 2018; Bartha *et al.* 2013). Nevertheless, joint host-parasite analysis can be also related to some complications. First, multiple comparisons have to be performed which potentially

decrease the statistical power and further, can be computationally extensive (MacPherson and Otto 2018). In addition, co-species analysis method compared to single species analysis methods have to fit the biology of both interacting partners. This requires an even more careful thinking compared to single species analysis.

7.4 How to deal with more complex forms of host-parasite coevolution in inference

All the analysis presented in this thesis assume a strict one by one relationship between the host and the parasite and that the coevolutionary interaction is driven by a single major bi-allelic locus in each species. It is justified to claim that these assumptions are a quite strong simplification of the reality and that such systems are barely or even never found in nature. It is for example common that hosts are co-infected by several parasites (Tollenaere *et al.* 2016). Further, is very likely that the different parasites display a distinct arsenal of effectors and thus, the host is exposed to different selective pressures from different parasites. Karasov *et al.* (2014) therefore suspect that the long-term balancing selection signature at the R-gene Rps5 in *Arabidopsis* might be the results of the interaction with a community of different parasites.
On the other hand, it is known that many parasite species can infect several hosts (Barrett *et al.* 2009). In such cases the parasite is potentially exposed to selective pressures from different hosts. Third, it has been shown that hosts have a large number of resistance gene homologues (Van de Weyer *et al.* 2019) and in the same manner parasite species have a large number of effector genes. Therefore, there might be also coevolution among some host R-genes and the R-gene network will determine the interaction outcome with a single parasite.
However, developing new methods by starting from such a simple systems has several advantages. First, already in such comparatively simple systems, it can be quite hard to track simultaneously all quantities which determine the coevolutionary dynamics and those which are affected by the coevolutionary interaction. There is no standard population genetic simulator available which is by default suited to simultaneously keep track of host and parasite genomic data. In addition understanding the power of new methods for simple scenarios makes it also easier to understand potentially more complex interactions

such as those with several R-genes involved.
Overall, the methods presented in this thesis are easily extensible to systems with several major genes involved as long as these genes are not in linkage and there are no pleiotropic or epistatic interaction among these genes. It has been shown that the conditions to obtain trench-warfare dynamics are less restrictive in mulit-locus compared to single locus systems (Tellier and Brown 2007a). However, the resulting signatures at each single locus might be potentially weaker and thus harder to be detected. This effect should be especially pronounced for quantitative traits (Jain and Stephan 2017).

7.5 The value of theoretical studies for analysing host and parasite genomic data

One could question why the methods presented in this thesis have so far not been applied to real data. Further, it is also justified to ask why the effect of coevolution on genomic signatures and the power of methods are first investigated by means of simulation. Put in a nutshell, real genomic data are noisy. They are not only affected by coevolution but also by other selective pressures such as the abiotic environment, competitive interactions with other species or spatial heterogeneity in resource availability. Therefore, testing new methods on simulated data is quite powerful to understand the full potential power of these methods for a perfect data set. Further, it is exactly known which processes have given rise to simulated data and replications can be easily done. Therefore, it becomes feasible to investigate the relative importance of the different processes. In real data the effect of particular processes will be always intermingled with the background noise. However in the long-run, new methods are only useful when they can be applied to real data. For the ABC in chapter 5 one could think about applying it to data from microcosm experiments (Frickel *et al.* 2016). One further alternative could be to apply it to data from coevolving metapopulations such as found in the *Plantago lanceolata-Podosphaera plantaginis*-system (Jousimo *et al.* 2014) or the *Linum usitatissimum-Melampsora lini*-system (Thrall *et al.* 2012). However, this requires to take the spatial structure into account.

7.6 Possible extensions of existing inference methods

It is important to refine and combine existing methods and to access which additional sources of information can be incorporated to increase the power of such methods. One promising approach to increase the power of inference methods is to incorporate genomic data from several time points as proposed in chapter 6. Time-samples have proven to be a valuable source to estimate the demographic history from allele frequency changes between different sampling times (Forde *et al.* 2008). Based on the established demography appropriate thresholds for detecting loci under selection (Foll *et al.* 2014, 2015) can be obtained. However, the availability of such data will crucially depend on the life-history traits of the coevolving species. For species with short generation times such as used in microcosm experiments (Frickel *et al.* 2016; Lenski 1988) it is straightforward to obtain time samples. Even for annual crop species it is possible to obtain such time-sampled data at least from the host. But obtaining such data can be rather difficult or even impossible for long-lived species.

An other type of data which could be very helpful are genomic samples from different populations. When migration rates are low the coevolutionary dynamics in the different populations can be asynchronous and thus, the different populations are likely to be at different stages of the coevolutionary cycles. Therefore, such data provide time samples in space. There is already a large body of literature how such data are potentially informative about coevolution (Gandon *et al.* 1996; Gandon and Michalakis 2002; Gandon *et al.* 2008; Gandon and Nuismer 2009). However, environmental conditions can differ among populations and it has been shown that abiotic factors can affect coevolutionary interactions (Duncan *et al.* 2017). Further, the parasite community could be different in the different populations and thus, introduce unknown confounding effects. Note in addition, that the spatial structure has to be also taken into account when establishing the demographic history of both species.

One further important question to be addressed is how coevolutionary inference methods for species with very low (or even absent) recombination rates (for example clonal reproducing fungi) can be designed. In such cases the assumption of evolutionary independence between the functional and coevolving loci is clearly violated. This makes it potentially hard to disentangle the co-demographic or demographic history from the effect of coevolution on the resulting genome data.

Conclusions

Given the plethora of host-parasite coevolutionary systems with varying life-history traits and the variety of interacting processes, there will be no single standard for analyzing genomic data from potentially coevolving species. However, in the last years promising approaches have been developed. Therefore, further effort should be put into thinking about how data from different sources, may it be genomic, phenotypic, from from a single or several populations, from a single or several time points, can be efficiently combined. Additionally, it is important to derive optimal sampling schemes, in terms of sampling times, sample sizes and spatial distribution of the samples. This requires a tight integration of theoretical modelling and experimental approaches. Further, we should aim to identify system where such approaches can be easily tested with enough generality. And most importantly, it requires that theoreticians, population geneticists, molecular biologist and phytopathologist force exchanges between disciplines and stop focusing on limitations of single approaches. Rather the field should move towards identifying the strengths of the different approaches. Coevolution is an ongoing process and so are our efforts to gain a better understanding of it. Scientists from different disciplines should aim to make this field of study a mutualistic rather than an antagonistic one.

Bibliography

Agrawal, A. and Lively, C. M. 2002. Infection genetics: gene-for-gene versus matching-alleles models and all points in between. *Evolutionary Ecology Research*, 4(1): 79–90.

Aguileta, G., Refrégier, G., Yockteng, R., Fournier, E., and Giraud, T. 2009. Rapidly evolving genes in pathogens: Methods for detecting positive selection and examples among fungi, bacteria, viruses and protists. *Infection, Genetics and Evolution*, 9(4): 656–670.

Alizon, S., de Roode, J. C., and Michalakis, Y. 2013. Multiple infections and the evolution of virulence. *Ecology Letters*, 16(4): 556–567.

Anderson, R. and May, R. 1979. Population biology of infectious disesases .1. *Nature*, 280(5721): 361–367.

Antonovics, J. 2017. Transmission dynamics: critical questions and challenges. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 372(1719).

Ashby, B. and Boots, M. 2017. Multi-mode fluctuating selection in host-parasite coevolution. *Ecology Letters*, 20(3): 357–365.

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

Bahri, B., Kaltz, O., Leconte, M., de Vallavieille-Pope, C., and Enjalbert, J. 2009. Tracking costs of virulence in natural populations of the wheat pathogen, *Puccinia striiformis* f.sp.*tritici*. *BMC Evolutionary Biology*, 9(1): 26.

Bailey, J. K., Hendry, A. P., Kinnison, M. T., Post, D. M., Palkovacs, E. P., Pelletier, F., Harmon, L. J., and Schweitzer, J. A. 2009. From genes to ecosystems: an emerging synthesis of eco-evolutionary dynamics. *New Phytologist*, 184(4): 746–749.

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

Bank, C., Ewing, G. B., Ferrer-Admettla, A., Foll, M., and Jensen, J. D. 2014. Thinking too positive? Revisiting current methods of population genetic selection inference. *Trends in Genetics*, 30(12): 540–546.

Barrett, L. G., Kniskern, J. M., Bodenhausen, N., Zhang, W., and Bergelson, J. 2009. Continua of specificity and virulence in plant host-pathogen interactions: causes and consequences. *New Phytologist*, 183(3): 513–529.

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., and 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.

Bartoli, C. and Roux, F. 2017. Genome-Wide Association Studies in plant pathosystems: Toward an ecological genomics approach. *Frontiers in Plant Science*, 8.

Barton, N. H. 1998. The effect of hitch-hiking on neutral genealogies. *Genetics Research*, 72(2): 123–133.

Beaumont, M. A. 2010. Approximate Bayesian Computation in Evolution and Ecology. In Futuyma, DJ and Shafer, HB and Simberloff, D, editor, *Annual Review of Ecology, Evolution, and Systematics, VOL 41*, volume 41 of *Annual Review of Ecology Evolution and Systematics*, pages 379–406.

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

Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., and Robert, C. P. 2009. Adaptive approximate Bayesian computation. *Biometrika*, 96(4): 983–990.

Becks, L., Ellner, S. P., Jones, L. E., and Hairston Jr., N. G. 2012. The functional genomics of an eco-evolutionary feedback loop: Linking gene expression, trait evolution, and community dynamics. *Ecology Letters*, 97(15): 492–501.

Bergelson, J. and Purrington, C. 1996. Surveying patterns in the cost of resistance in plants. *American Naturalist*, 148(3): 536–558.

Bergelson, J., Kreitman, M., Stalh, E. A., and Tian, D. 2001. Evolutionary Dynamics of Plant R-Genes. *Science*, 292(5525): 2281–2285.

Berger, L., Speare, R., Daszak, P., Green, D., Cunningham, A., Goggin, C., Slocombe, R., Ragan, M., Hyatt, A., McDonald, K., Hines, H., Lips, K., Marantelli, G., and Parkes, H. 1998. Chytridiomycosis causes amphibian mortality associated with population declines in the rain forests of Australia and Central America. *Proceedings of the National Academy of Sciences of the United States of America*, 95(15): 9031–9036.

Bernatchez, L. and Landry, C. 2003. MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? *Journal of Evolutionary Biology*, 16(3): 363–377.

Berngruber, T. W., Lion, S., and Gandon, S. 2015. Spatial structure, transmission modes and the evolution of viral exploitation strategies. *PLoS Pathogens*, 11(4): 1–13.

Best, A., Ashby, B., White, A., Bowers, R., Buckling, A., Koskella, B., and Boots, M. 2017. Host-parasite fluctuating selection in the absence of specificity. *Proceedings of the Royal Society B-Biological Sciences*, 284(1866).

Betts, A., Kaltz, O., and Hochberg, M. E. 2014. Contrasted coevolutionary dynamics between a bacterial pathogen and its bacteriophages. *Proceedings of the National Academy of Sciences of the United States of America*, 111(42): 15279–15279.

Bohannan, B. and Lenski, R. 2000. Linking genetic change to community evolution: insights from studies of bacteria and bacteriophage. *Ecology Letters*, 3(4): 362–377.

Boots, M. and Haraguchi, Y. 1999. The evolution of costly resistance in host-parasite systems. *American Naturalist*, 153(4): 359–370.

Boots, M., White, A., Best, A., and Bowers, R. 2014. How specificity and epidemiology drive the coevolution of static trait diversity in hosts and parasites. *Evolution*, 68(6): 1594–1606.

Bozick, B. A. and Real, L. A. 2015. The role of human transportation networks in mediating the genetic structure of seasonal influenza in the United States. *PLOS Pathogens*, 11(6): 1–17.

Brown, J. K. M. 2015. Durable resistance of crops to disease: A Darwinian perspective. In VanAlfen, NK, editor, *Annual Review of Phytopathology*, volume 53 of *Annual Review of Phytopathology*, pages 513–539.

Brown, J. K. M. and Tellier, A. 2011. Plant-parasite coevolution: Bridging the gap between genetics and ecology. In VanAlfen, NK and Bruening, G and Leach, JE, editor, *Annual Review of Phytopathology, VOL 49*, volume 49 of *Annual Review of Phytopathology*, pages 345–367.

Brussaard, C. P., Kuipers, B., and Veldhuis, M. J. 2005. A mesocosm study of *Phaeocystis globosa* population dynamics I. Regulatory role of viruses in bloom control. *Harmful Algae*, 4(5): 859–874.

Burmeister, A. R., Lenski, R. E., and Meyer, J. R. 2016. Host coevolution alters the adaptive landscape of a virus. *Proceedings of the Royal Society B: Biological Sciences*, 283(1839): 20161528.

Caicedo, A. and Schaal, B. 2004. Heterogeneous evolutionary processes affect R gene diversity in natural populations of *Solanum pimpinellifolium*. *P Natl Acad Sci USA*, 101(50): 17444–17449.

Castberg, T., Larsen, A., Sandaa, R. A., Brussaard, C. P. D., Egge, J. K., Heldal, M., Thyrhaug, R., Van Hannen, E. J., and Bratbak, G. 2001. Microbial population dynamics and diversity during a bloom of the marine coccolithophorid *Emiliania huxleyi* (Haptophyta). *Marine Ecology Progress Series*, 221: 39–46.

Charlesworth, B. 2009. Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, 10(3): 195–205.

Charlesworth, B. 2013. Background selection 20 years on. *Journal of Heredity*, 104(2): 161–171.

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

Charlesworth, B., Morgan, M. T., and Charlesworth, D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4): 1289–1303.

Charlesworth, D. 2006. Balancing selection and its effects on sequences in nearby genome regions. *PLOS Genetics*, 2(4): 379–384.

Chisholm, S., Coaker, G., Day, B., and Staskawicz, B. 2006. Host-microbe interactions: Shaping the evolution of the plant immune response. *Cell*, 124(4): 803–814.

Clarke, B. C. 1979. The evolution of genetic diversity. *Proceedings of the Royal Society of London B*, 205: 453–474.

Clokie, M. R., Millard, A. D., Letarov, A. V., and Heaphy, S. 2011. Phages in nature. *Bacteriophage*, 1(1): 31–45.

Colson, P., La Scola, B., and Raoult, D. 2017. Giant viruses of amoebae: A journey through innovative research and paradigm changes. *Annual Review of Virology*, 4(1): 61–85.

Crisci, J. L., Poh, Y. P., Bean, A., Simkin, A., and Jensen, J. D. 2012. Recent progress in polymorphism-based population genetic inference. *Journal of Heredity*, 103(2): 287–296.

Csillery, K., Blum, M. G. B., Gaggiotti, O. E., and Francois, O. 2010. Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution*, 25(7): 410–418.

Cvijović, I., Nguyen Ba, A. N., and Desai, M. M. 2018. Experimental studies of evolutionary dynamics in microbes. *Trends in Genetics*, 34(9): 1–11.

Dalman, K., Himmelstrand, K., Olson, A., Lind, M., Brandström-Durling, M., and Stenlid, J. 2013. A genome-wide association study identifies genomic regions for virulence in the non-model organism *Heterobasidion annosum* s.s. *PLOS One*, 8(1): 1–10.

Danovaro, R., Corinaldesi, C., Dell'Anno, A., Fuhrman, J. A., Middelburg, J. J., Noble, R. T., and Suttle, C. A. 2011. Marine viruses and global climate change. *FEMS Microbiology Reviews*, 35(6): 993–1034.

Dawkins, R. and Krebs, J. 1979. Arms races between and within species. *Proceedings of the Royal Society Series B-Biological Sciences*, 205(1161): 489–511.

De Andreazzi, C. S., Guimarães, P. R., and Melián, C. J. 2018. Eco-evolutionary feedbacks promote fluctuating selection and long-term stability of antagonistic networks. *Proceedings of the Royal Society B: Biological Sciences*, 285(1874).

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

Decaestecker, E., Gaba, S., Raeymaekers, J. A., Stoks, R., Van Kerckhoven, L., Ebert, D., and De Meester, L. 2007. Host-parasite 'Red Queen' dynamics archived in pond sediment. *Nature*, 450(7171): 870–873.

Declerck, S. A. J., Winter, C., Shurin, J. B., Suttle, C. A., and Matthews, B. 2013. Effects of patch connectivity and heterogeneity on metacommunity structure of planktonic bacteria and viruses. *ISME Journal*, 7(3): 533–542.

Dennehy, J. J. 2012. What can phages tell us about host-pathogen coevolution? *International Journal of Evolutionary Biology*, pages 1–12.

Der, R., Epstein, C. L., and Plotkin, J. B. 2011. Generalized population models and the nature of genetic drift. *Theoretical Population Biology*, 80(2): 80–99.

Desai, M. M. 2013. Statistical questions in experimental evolution. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(1).

Desai, M. M., Fisher, D. S., and Murray, A. W. 2007. The speed of evolution and maintenance of variation in asexual populations. *Current Biology*, 17(5): 385–394.

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

Dodds, P. N. and Rathjen, J. P. 2010. Plant immunity: towards an integrated view of plant-pathogen interactions. *Nature Reviews Genetics*, 11(8): 539–548.

Donnelly, P. and Tavare, S. 1995. Coalescents and genealogical structure under neutrality. *Annual Review of Genetics*, 29: 401–421.

Drake, J. W., Charlesworth, B., Charlesworth, D., and Crow, J. F. 1998. Rates of spontaneous mutation. *Genetics*, 148(4): 1667–1686.

Duffy, S., Shackelton, L. A., and Holmes, E. C. 2008. Rates of evolutionary change in viruses: Patterns and determinants. *Nature Reviews Genetics*, 9(4): 267–76.

Duncan, A. B., Dusi, E., Jacob, F., Ramsayer, J., Hochberg, M. E., and Kaltz, O. 2017. Hot spots become cold spots: coevolution in variable temperature environments. *Journal of Evolutionary Biology*, 30(1): 55–65.

Ebert, D. 1994. Virulence and local adaptation of a horizontally transmitted parasite. *Science*, 265: 1084–1068.

Ebert, D. 2018. Open questions: what are the genes underlying antagonistic coevolution? *BMC Biology*, 16.

Ebert, D., Lipsitch, M., and Mangin, K. 2000. The effect of parasites on host population density and extinction: Experimental epidemiology with *Daphnia* and six microparasites. *American Naturalist*, 156(5): 459–477.

Ejsmond, M. J. and Radwan, J. 2015. Red Queen processes drive positive selection on Major Histocompatibility Complex (MHC) Genes. *PLOS Computational Biology*, 11(11).

Eldon, B. and Stephan, W. 2018. Evolution of highly fecund haploid populations. *Theoretical Population Biology*, 119: 48 – 56.

Eldon, B. and Wakeley, J. 2009. Coalescence times and Fst under a skewed offspring distribution among individuals in a population. *Genetics*, 181(2): 615–629.

Ellegren, H. and Galtier, N. 2016. Determinants of genetic diversity. *Nature Reviews Genetics*, 17(7): 422–433.

Engelstädter, J. 2015. Host-parasite coevolutionary dynamics with generalized success/-failure infection genetics. *American Naturalist*, 185(5): E117–29.

Engelstaedter, J. 2015. Host-Parasite coevolutionary dynamics with generalized success/-failure infection genetics. *American Naturalist*, 185(5): E117–E129.

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

Ewing, G. B. and Jensen, J. D. 2016. The consequences of not accounting for background selection in demographic inference. *Molecular Ecology*, 25(1): 135–141.

Faria, N. R., Rambaut, A., Suchard, M. A., Baele, G., Bedford, T., Ward, J., Tatem, A. J., Sousa, J. D., Arinaminpathy, N., Pépin, J., Posada, D., Peeters, M., Pybus, O. G., and Lemey, P. 2014. The early spread and epidemic ignition of HIV-1 in human populations. *Science*, 346(6205): 3–8.

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

Felsenstein, J. 1974. The evolutionary advantage of recombination. *Genetics*, 78: 737–756.

Felsenstein, J. 1981. Evolutionary trees from DNA-seqeuences - A maximum-likelihood approach. *Journal of Molecular Evolution*, 17(6): 368–376.

Fenton, A., Antonovics, J., and Brockhurst, M. A. 2009. Inverse-gene-for-gene infection genetics and coevolutionary dynamics. *American Naturalist*, 174(6): E230–E242.

Fijarczyk, A. and Babik, W. 2015. Detecting balancing selection in genomes: Limits and prospects. *Molecular Ecology*, 24(14): 3529–3545.

Flor, H. H. 1956. The complementary genic systems in flax and flax rust. *Advances in Genetics*, 8(C): 29–54.

Flor, H. H. 1971. Current status of the gene for gene concept. *Annual Reviews of Phytopathology*, 9: 275–296.

Foll, M., Poh, Y. P., Renzette, N., Ferrer-Admetlla, A., Bank, C., Shim, H., Malaspinas, A. S., Ewing, G., Liu, P., Wegmann, D., Caffrey, D. R., Zeldovich, K. B., Bolon, D. N., Wang, J. P., Kowalik, T. F., Schiffer, C. A., Finberg, R. W., and Jensen, J. D. 2014.

Influenza virus drug resistance: A time-sampled population genetics perspective. *PLoS Genetics*, 10(2).

Foll, M., Shim, H., and Jensen, J. D. 2015. WFABC: a Wright-Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data. *Molecular Ecology Resources*, 15(1): 87–98.

Forde, S. E., Beardmore, R. E., Gudelj, I., Arkin, S. S., Thompson, J. N., and Hurst, L. D. 2008. Understanding the limits to generalizability of experimental evolutionary models. *Nature*, 455(7210): 220–223.

Frank, S. 1992. Models of plant-pathogen coevolution. *Trends in Genetics*, 8(6): 213–219.

Frick, W. F., Pollock, J. F., Hicks, A. C., Langwig, K. E., Reynolds, D. S., Turner, G. G., Butchkoski, C. M., and Kunz, T. H. 2010. An emerging disease causes regional population collapse of a common North American bat species. *Science*, 329(5992): 679–682.

Frickel, J., Sieber, M., and Becks, L. 2016. Eco-evolutionary dynamics in a coevolving host–virus system. *Ecology Letters*, 19(4): 1–31.

Frickel, J., Theodosiou, L., and Becks, L. 2017. Rapid evolution of hosts begets species diversity at the cost of intraspecific diversity. *Proceedings of the National Academy of Sciences of the United States of America*, 42(114): 11193–11198.

Frickel, J., Feulner, P. G. D., Karakoc, E., and Becks, L. 2018. Population size changes and selection drive patterns of parallel evolution in a host–virus system. *Nature Communications*, 9(1): 1706.

Fronhofer, E. A. and Altermatt, F. 2015. Eco-evolutionary feedbacks during experimental range expansions. *Nature Communications*, 6.

Fu, Y. and Li, W. 1993. Statistical tests of neutrality of mutations. *Genetics*, 133(3): 693–709.

Gandon, S. and 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(3): 451–462.

Gandon, S. and Nuismer, S. L. 2009. Interactions between genetic drift, gene flow, and selection mosaics drive parasite local adaptation. *America Naturalist*, 173(2): 212–224.

Gandon, S., Capowiez, Y., Dubois, Y., Michalakis, Y., and Olivieri, I. 1996. Local adaptation and gene-for-gene coevolution in a metapopulation model. *Proceedings of the Royal Society B-Biological Sciences*, 263(1373): 1003–1009.

Gandon, S., Buckling, A., Decaestecker, E., and Day, T. 2008. Host-parasite coevolution and patterns of adaptation across time and space. *Journal of Evolutionary Biology*, 21(6): 1861–1866.

Gandon, S., Day, T., Metcalf, C. J. E., and Grenfell, B. T. 2016. Forecasting epidemiological and evolutionary dynamics of infectious diseases. *Trends in Ecology & Evolution*, 31(10): 776–788.

Gavrilets, S. and Michalakis, Y. 2008. Effects of environmental heterogeneity on victim-exploiter coevolution. *Evolution*, 62(12): 3100–3116.

Gerrish, P. J. and Lenski, R. E. 1998. The fate of competing beneficial mutations in an asexual population. *Genetica*, 102-103(1-6): 127–144.

Gire, S. K., Goba, A., Andersen, K. G., Sealfon, R. S. G., Park, D. J., Kanneh, L., Jalloh, S., Momoh, M., Fullah, M., Dudas, G., Wohl, S., Moses, L. M., Yozwiak, N. L., Winnicki, S., Matranga, C. B., Malboeuf, C. M., Qu, J., Gladden, A. D., Schaffner, S. F., Yang, X., Jiang, P.-P., Nekoui, M., Colubri, A., Coomber, M. R., Fonnie, M., Moigboi, A., Gbakie, M., Kamara, F. K., Tucker, V., Konuwa, E., Saffa, S., Sellu, J., Jalloh, A. A., Kovoma, A., Koninga, J., Mustapha, I., Kargbo, K., Foday, M., Yillah, M., Kanneh, F., Robert, W., Massally, J. L. B., Chapman, S. B., Bochicchio, J., Murphy, C., Nusbaum, C., Young, S., Birren, B. W., Grant, D. S., Scheiffelin, J. S., Lander, E. S., Happi, C., Gevao, S. M., Gnirke, A., Rambaut, A., Garry, R. F., Khan, S. H., and Sabeti, P. C. 2014. Genomic surveillance elucidates ebola virus origin and transmission during the 2014 outbreak. *Science*, 345(6202): 1369–1372.

Gokhale, C. S., Papkou, A., Traulsen, A., and Schulenburg, H. 2013. Lotka-Volterra dynamics kills the Red Queen: Population size fluctuations and associated stochasticity dramatically change host-parasite coevolution. *BMC Evolutionary Biology*, 13: 254.

Gómez, P. and Buckling, A. 2011. Bacteria-phage antagonistic coevolution in soil. *Science*, 332(6025).

Gómez, P., Ashby, B., Buckling, A., Gomez, P., Ashby, B., and Buckling, A. 2015. Population mixing promotes arms race host-parasite coevolution. *Proceedings of the Royal Society B: Biological Sciences*, 282(1798): 20142297.

Good, B. H., Walczak, A. M., Neher, R. A., and Desai, M. M. 2014. Genetic diversity in the interference selection limit. *PLOS Genetics*, 10(3): 1–14.

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

Govaert, L., Fronhofer, E. A., Lion, S., Eizaguirre, C., Bonte, D., Egas, M., Hendry, A., de Brito Martins, A., Melián, C. J., Raeymaekers, J. A., Ratikainen, I. I., Saether, B., Schweitzer, J. A., and Matthews, B. 2019. Eco-evolutionary feedbacks—Theoretical models and perspectives. *Functional Ecology*, 33(1): 13–30.

Gutierrez, A. P., Bean, T. P., Hooper, C., Stenton, C. A., Sanders, M. B., Paley, R. K., Rastas, P., Bryrom, M., Matika, O., and Houston, R. D. 2018. A Genome-Wide Association Study for Host Resistance to Ostreid Herpesvirus in Pacific Oysters (Crassostrea gigas). *G3: Genes, Genomes, Genetics*, 8(April): g3.200113.2018.

Hairston, N. G., Ellner, S. P., Geber, M. A., Yoshida, T., and Fox, J. A. 2005. Rapid evolution and the convergence of ecological and evolutionary time. *Ecology Letters*, 8(10): 1114–1127.

Hall, A. R., Scanlan, P. D., Morgan, A. D., and Buckling, A. 2011. Host-parasite coevolutionary arms races give way to fluctuating selection. *Ecology Letters*, 14: 635–642.

Hamilton, W. D., Axelrod, R., and Tanese, R. 1990. Sexual reproduction as an adaptation to resist parasites (a review). *Proceedings of the National Academy of Sciences of the United States of America*, 87(9): 3566–73.

Hartmann, F. E., McDonald, B. A., and Croll, D. 2018. Genome-wide evidence for divergent selection between populations of a major agricultural pathogen. *Molecular Ecology*, 27(12): 2725–2741.

Hein, J., Schierup, M., and Wiuf, C. 2004. *Gene genealogies, variation and evolution: A primer in coalescent theory*. Oxford University Press, Oxford, 1 edition.

Hill, W. and Robertson, A. 1966. Effect of linkage on limits to artificial selection. *Genetics Research*, 8(3): 269+.

Hiltunen, T. and Becks, L. 2014. Consumer co-evolution as an important component of the eco-evolutionary feedback. *Nature Communications*, 5.

Hinkley, T., Martins, J., Chappey, C., Haddad, M., Stawiski, E., Whitcomb, J. M., Petropoulos, C. J., and Bonhoeffer, S. 2011. A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. *Nature Genetics*, 43(5): 487–490.

Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., Poss, M. L., Reed, L. K., Storfer, A., and Whitlock, M. C. 2016. Finding the genomic basis of local adaptation: Pitfalls, practical solutions, and future directions. *The American Naturalist*, 188(4): 379–397.

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

Holub, E. 2001. The arms race is ancient history in *Arabidopsis*, the wildflower. *Nature Reviews Genetics*, 2(7): 516–527.

Horas, E. L., Theodosiou, L., and Becks, L. 2018. Why are algal viruses not always successful ? *Viruses*, 10(474).

Hudson, R. 1983. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23(2): 183–201.

Hudson, R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2): 337–338.

Irwin, K. K., Laurent, S., Matuszewski, S., Vuilleumier, S., Ormond, L., Shim, H., Bank, C., and Jensen, J. D. 2016. On the importance of skewed offspring distributions and background selection in virus population genetics. *Heredity*, 117(6): 393–399.

Jain, K. and Stephan, W. 2017. Modes of rapid polygenic adaptation. *Molecular Biology and Evolution*, 34(12): 3169–3175.

Janzen, D. 1980. When is it coevolution. *Evolution*, 34(3): 611–612.

Jayakar, S. 1970. A mathematical model for interaction of gene frequencies in a parasite and its host. *Theoretical Population Biology*, 1(2): 140–164.

Jee, J., Rasouly, A., Shamovsky, I., Akivis, Y., Steinman, S. R., Mishra, B., and Nudler, E. 2016. Rates and mechanisms of bacterial mutagenesis from maximum-depth sequencing. *Nature*, 534(7609): 693–696.

Jensen, J. D., Payseur, B. A., Stephan, W., Aquadro, C. F., Lynch, M., Charlesworth, D., and Charlesworth, B. 2019. The importance of the Neutral Theory in 1968 and 50 years on: A response to Kern and Hahn 2018. *Evolution*, 73(1): 111–114.

Jiang, L., Liu, P., Bank, C., Renzette, N., Prachanronarong, K., Yilmaz, L. S., Caffrey, D. R., Zeldovich, K. B., Schiffer, C. A., Kowalik, T. F., Jensen, J. D., Finberg, R. W., Wang, J. P., and Bolon, D. N. 2016. A balance between inhibitor binding and substrate processing confers influenza drug resistance. *Journal of Molecular Biology*, 428(3): 538–553.

Johannessen, T. V., Larsen, A., Bratbak, G., Pagarete, A., Edvardsen, B., Egge, E. D., and Sandaa, R. A. 2017. Seasonal dynamics of haptophytes and dsDNA algal viruses suggest complex virus-host relationship. *Viruses*, 9(4).

Jones, J. D. G. and Dangl, J. L. 2006. The plant immune system. *Nature*, 444(7117): 323–329.

Jorde, P. E. and Ryman, N. 2007. Unbiased estimator for genetic drift and effective population size. *Genetics*, 177(2): 927–935.

Jousimo, J., Tack, A. J. M., Ovaskainen, O., Mononen, T., Susi, H., Tollenaere, C., and Laine, A.-L. 2014. Ecological and evolutionary effects of fragmentation on infectious disease dynamics. *Science*, 344(6189): 1289–1293.

Jukes, T. and Cantor, C. 1969. Evolution of protein molecules. In H. Munroe, editor, *Mammalian protein metabolism*, volume III, chapter 10, pages 21–231. Academic Press, New York, N.Y., U.S.A.

Kaplan, N. L., Hudson, R. R., and Langley, C. H. 1989. The "hitchhiking effect" revisited. *Genetics*, 123(4): 887–899.

Karasov, T. L., Kniskern, J. M., Gao, L., DeYoung, B. J., Ding, J., Dubiella, U., Lastra, R. O., Nallu, S., Roux, F., Innes, R. W., Barrett, L. G., Hudson, R. R., and Bergelson, J. 2014. The long-term maintenance of a resistance polymorphism through diffuse interactions. *Nature*, 512(7515): 436–U472.

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

Kern, A. D. and Hahn, M. W. 2018. The Neutral Theory in light of natural selection. *Molecular Biology and Evolution*, 35(6): 1366–1371.

Kimura, M. 1962. On the probability of fixation of mutant genes in a population. *Genetics*, 47(391): 713–719.

Kimura, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4): 893–&.

Kimura, M. and Crow, J. 1964. The number of alleles that can be maintained in a finite population. *Genetics*, 49(4): 725–&.

Kimura, M. and Ohta, T. 1969. The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, 61(3): 763–&.

Kingman, J. 1982a. On the genealogy of large populations. *Journal of Applied Probability*, 19(): 27–43.

Kingman, J. 1982b. The coalescent. *Stochastic Processes and their Applications*, 13: 235–248.

Kirby, G. C. and Burdon, J. J. 1997. Effects of mutation and random drift on leonard's gene-for-gene coevolution model. *Phytopathology*, 87(5): 488–493.

Kishino, H. and Hasegawa, M. 1989. Evaluation of the maximum-likelihood estimate of evolutionary tree topologies from DNA-sequence data, and the branching order in Hominoidea. *Journal of Molecular Evolution*, 29(2): 170–179.

Koch, H., Frickel, J., Valiadi, M., and Becks, L. 2014. Why rapid, adaptive evolution matters for community dynamics. *Frontiers in Ecology and Evolution*, 2(May): 1–10.

Koenig, D., Hagmann, J., Li, R., Bemm, F., Slotte, T., Nueffer, B., Wright, I, S., and Weigel, D. 2019. Long-term balancing selection drives evolution of immunity genes in *Capsella*. *eLIFE*, 8.

Kokko, H. and López-Sepulcre, A. 2007. The ecogenetic link between demography and evolution: Can we bridge the gap between theory and data? *Ecology Letters*, 10(9): 773–782.

Koonin, E. V., Senkevich, T. G., and Dolja, V. V. 2006. The ancient virus world and evolution of cells. *Biology Direct*, pages 1–27.

Kosheleva, K. and Desai, M. M. 2013. The dynamics of genetic draft in rapidly adapting populations. *Genetics*, 195(3): 1007–1025.

Koskella, B. and Brockhurst, M. A. 2014. Bacteria-phage coevolution as a driver of ecological and evolutionary processes in microbial communities. *FEMS Microbiology Reviews*, 38(5): 916–931.

Koskella, B. and Meaden, S. 2013. Understanding bacteriophage specificity in natural microbial communities. *Viruses*, 5(3): 806–823.

Kourelis, J. and van der Hoorn, R. A. L. 2018. Defended to the Nines: 25 years of resistance gene cloning identifies nine mechanisms for R protein function. *Plant Cell*, 30(2): 285–299.

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

Krishnan, P., Ma, X., McDonald, B. A., and Brunner, P. C. 2018. Widespread signatures of selection for secreted peptidases in a fungal plant pathogen. *BMC Evolutionary Biology*, 18(1): 7.

Kryazhimskiy, S., Dushoff, J., Bazykin, G. A., and Plotkin, J. B. 2011. Prevalence of epistasis in the evolution of influenza A surface proteins. *PLoS Genetics*, 7(2).

Kwiatkowski, M., Engelstaedter, J., and Vorburger, C. 2012. On genetic specificity in symbiont-mediated host-parasite coevolution. *PLOS Computational Biology*, 8(8).

Laanto, E., Hoikkala, V., Ravantti, J., and Sundberg, L.-R. 2017. Long-term genomic coevolution of host-parasite interaction in the natural environment. *Nature Communications*, 8(111).

Laine, A. 2004. Resistance variation within and among host populations in a plant-pathogen metapopulation: implications for regional pathogen dynamics. *Journal of Ecology*, 92(6): 990–1000.

Lang, G. I. and Murray, A. W. 2008. Estimating the per-base-pair mutation rate in the yeast *Saccharomyces cerevisiae*. *Genetics*, 178(1): 67–82.

Lang, G. I., Botstein, D., and Desai, M. M. 2011. Genetic variation and the fate of beneficial mutations in asexual populations. *Genetics*, 188(July): 647–661.

Lennon, J. T. and Martiny, J. B. H. 2008. Rapid evolution buffers ecosystem impacts of viruses in a microbial food web. *Ecology Letters*, 11: 1178–1188.

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

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

Leuenberger, C. and Wegmann, D. 2010. Bayesian computation and model selection without likelihoods. *Genetics*, 184(1): 243–252.

Lewontin, R. and Kojima, K. 1960. The evolutionary dynamics of complex polymorphisms. *Evolution*, 14(4): 458–472.

Longdon, B., Brockhurst, M. A., Russell, C. A., Welch, J. J., and Jiggins, F. M. 2014. The evolution and genetics of virus host shifts. *PLOS Pathogens*, 10(11).

Luijckx, P., Fienberg, H., Duneau, D., and Ebert, D. 2013. A matching-allele model explains host resistance to parasites. *Current Biology*, 23(12): 1085–1088.

Luo, S. and Koelle, K. 2013. Navigating the devious course of evolution: The importance of mechanistic models for identifying eco-evolutionary dynamics in nature. *The American Naturalist*, 181(S1): 58–75.

MacPherson, A. and Otto, S. P. 2018. Joint coevolutionary – epidemiological models dampen Red Queen cycles and alter conditions for epidemics. *Theoretical Population Biology*, 122: 137–148.

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

Marjoram, P., Molitor, J., Plagnol, V., and Tavare, S. 2003. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciecnes of the United States of America*, 100(26): 15324–15328.

Marston, M. F., Pierciey, F. J., Shepard, A., Gearin, G., Qi, J., Yandava, C., Schuster, S. C., Henn, M. R., and Martiny, J. B. H. 2012. Rapid diversification of coevolving marine *Synechococcus* and a virus. *Proceedings of the National Academy of Sciences of the United States of America*, 109(12): 4544–4549.

Martiny, J. B., Riemann, L., Marston, M. F., and Middelboe, M. 2014. Antagonistic coevolution of marine planktonic viruses and their hosts. *Annual Review of Marine Science*, 6(1): 393–414.

May, R. and Anderson, R. 1983. Epidemiology and genetics in the coevolution of parasites and hosts. *Proceedings of the Royal Society Series B-Biological Sciences*, 219(1216): 281–313.

Maynard Smith, J. and Haigh, J. 1974. The hitch-hiking effect of a favourable gene. *Genetics Research*, 23(1): 23–35.

Meier-Kolthoff, J. P., Uchiyama, J., Yahara, H., Paez-Espino, D., and Yahara, K. 2018. Investigation of recombination-intense viral groups and their genes in the Earth's virome. *Scientific Reports*, 8(1): 1–11.

Messer, P. W., Ellner, S. P., and Hairston, N. G. 2016. Can population genetics adapt to rapid evolution? *Trends in Genetics*, 32(7): 408–418.

Meyer, J. R., Dobias, D. T., Weitz, J. S., Barrick, J. E., Quick, R. T., and Lenski, R. E. 2012. Repeatability and contingency in the evolution of a key innovation in phage lambda. *Science*, 335(January).

Montarry, J., Hamelin, F. M., Glais, I., Corbière, R., and Andrivon, D. 2010. Fitness costs associated with unnecessary virulence factors and life history traits: evolutionary insights from the potato late blight pathogen *Phytophthora infestans*. *BMC Evolutionary Biology*, 10(1): 283.

Muller, H. 1964. The relation of recombination to mutational advance. *Mutation Research*, 1: 2–9.

Mundt, C. C. 2009. Importance of autoinfection to the epidemiology of polycyclic foliar disease. *Phytopathology*, 99(10): 1116–1120.

Neher, R. A. 2013. Genetic draft, selective interference, and population genetics of rapid adaptation. *Annual Review of Ecology, Evolution, and Systematics*.

Nei, M. and Tajima, F. 1981. DNA polymorphism detectable by restriction endonucleases. *Genetics*, 97(1): 145–163.

Nielsen, R. 2005. Molecular signatures of natural selection. *Annual Review of Genetics*, 39: 197–218.

Nuismer, S. L. and Week, B. 2019. Approximate Bayesian estimation of coevolutionary arms races. *PLOS Computational Biology*, 15(4).

Nuismer, S. L., Gomulkiewicz, R., and Ridenhour, B. J. 2010. When is correlation coevolution? *American Naturalist*, 175(5): 525–537.

Nuismer, S. L., Jenkins, C. E., and Dybdahl, M. F. 2017. Identifying coevolving loci using interspecific genetic correlations. *Ecology and Evolution*, 7(17): 6894–6903.

Obbard, D., Jiggins, F., Halligan, D., and Little, T. 2006. Natural selection drives extremely rapid evolution in antiviral RNAi genes. *Current Biology*, 16(6): 580–585.

Obbard, D. J., Jiggins, F. M., Bradshaw, N. J., and Little, T. J. 2011. Recent and recurrent selective sweeps of the antiviral RNAi gene *Argonaute-2* in three species of *Drosophila*. *Molecular Biology and Evolution*, 28(2): 1043–1056.

Otto, S. P. and Day, T. 2007. *A biologist's guide to mathematical modelling in ecology and evolution*, volume 1. Princeton University Press, 41 William Street, Princeton, New Jersey 08540.

Otto, S. P. and Whitlock, M. C. 1997. The probability of fixation in populations of changing size. *Genetics*, 146(1967): 723–733.

Papkou, A., Gokhale, C. S., Traulsen, A., and Schulenburg, H. 2016. Host–parasite coevolution: Why changing population size matters. *Zoology*, 119(4): 330–338.

Papkou, A., Guzella, T., Yang, W., Koepper, S., Pees, B., Schalkowski, R., Barg, M.-C., Rosenstiel, P. C., Teotónio, H., and Schulenburg, H. 2018. The genomic basis of Red Queen dynamics during rapid reciprocal host-pathogen coevolution. *Proceedings of the National Academy of Sciences of the United States of America*, page 201810402.

Park, S.-C. and Krug, J. 2007. Clonal interference in large populations. *Proceedings of the National Academy of Sciences of the United States of America*, 104(46): 18135–18140.

Paterson, S., Vogwill, T., Buckling, A., Benmayor, R., Spiers, A. J., Thomson, N. R., Quail, M., Smith, F., Walker, D., Libberton, B., Fenton, A., Hall, N., and Brockhurst, M. A. 2010. Antagonistic coevolution accelerates molecular evolution. *Nature*, 464(7286): 275–278.

Pennings, P. S., Kryazhimskiy, S., and Wakeley, J. 2014. Loss and recovery of genetic diversity in adapting populations of HIV. *PLOS Genetics*, 10(1).

Pérez-Losada, M., Arenas, M., Galán, J. C., Palero, F., and González-Candelas, F. 2015. Recombination in viruses: mechanisms, methods of study, and evolutionary consequences. *Infection, Genetics and Evolution*, 30: 296–307.

Post, D. M. and Palkovacs, E. P. 2009. Eco-evolutionary feedbacks in community and ecosystem ecology: interactions between the ecological theatre and the evolutionary play. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 364(1523): 1629–1640.

Poullain, V., Gandon, S., Brockhurst, M. A., Buckling, A., and Hochberg, M. E. 2008. The evolution of specificity in evolving and coevolving antagonistic interactions between a bacteria and its phage. *Evolution*, 62(1): 1–11.

Pritchard, J., Seielstad, M., Perez-Lezaun, A., and Feldman, M. 1999. Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12): 1791–1798.

Radford, A. D., Chapman, D., Dixon, L., Chantrey, J., Darby, A. C., and Hall, N. 2012. Application of next-generation sequencing technologies in virology. *Journal of General Virology*, 93: 1853–1868.

Renzette, N., Gibson, L., Bhattacharjee, B., Fisher, D., Schleiss, M. R., Jensen, J. D., and Kowalik, T. F. 2013. Rapid intrahost evolution of human cytomegalovirus is shaped by demography and positive selection. *PLoS Genetics*, 9(9): 1–14.

Renzette, N., Pokalyuk, C., Gibson, L., Bhattacharjee, B., Schleiss, M. R., and Hamprecht, K. 2015. Limits and patterns of cytomegalovirus genomic diversity in humans. *Proceedings of the National Academy of Sciences of the United States of America*, pages 4120–4128.

Renzette, N., Kowalik, T. F., and Jensen, J. D. 2016. On the relative roles of background selection and genetic hitchhiking in shaping human cytomegalovirus genetic diversity. *Molecular Ecology*, 25(1): 403–413.

Rodriguez-Valera, F., Martin-Cuadrado, A.-B., Rodriguez-Brito, B., Thingstad, T. F., Rohwer, F., and Mira, A. 2009. Explaining microbial population genomics through phage predation. *Nature Reviews Microbiology*, 7: 828–836.

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

Rosenzweig, M. and MacArthur, R. 1963. Graphical representation and stability conditions of predator-prey interactions. *American Naturalist*, 97(895): 209–223.

Russell, C. A., Fonville, J. M., Brown, A. E. X., Burke, D. F., Smith, D. L., James, S. L., Herfst, S., van Boheemem, S., Linster, M., Schrauwen, E. J., Katzelnick, L., Mosterin, A., Kuiken, T., Maher, E., Neumann, G., Osterhaus, A. D. M. E., Kawaoka, Y., Fouchier, R. A. M., and Smith, D. J. 2012. The potential for respiratory droplet-transmissible A/H5N1 Influenza virus to evolve in a mammalian host. *Science*, 1541(June): 1541–1548.

Salathe, M., Scherer, A., and Bonhoeffer, S. 2005. Neutral drift and polymorphism in gene-for-gene systems. *Ecology Letters*, 8(9): 925–932.

Sanjuán, R. 2018. Collective properties of viral infectivity. *Current Opinion in Virology*, 33: 1–6.

Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M., and Belshaw, R. 2010. Viral mutation rates. *Journal of Virology*, 84(19): 9733–9748.

Sasaki, A. 2000. Host-parasite coevolution in a multilocus gene-for-gene system. *Proceedings of the Royal Society B-Biological Sciences*, 267(1458): 2183–2188.

Saxenhofer, M., Weber de Melo, V., Ulrich, R. G., and Heckel, G. 2017. Revised time scales of RNA virus evolution based on spatial information. *Proceedings of the Royal Society B: Biological Sciences*, 284(1860).

Scanlan, P. D., Hall, A. R., Lopez-Pascua, L. D. C., and Buckling, A. 2011. Genetic basis of infectivity evolution in a bacteriophage. *Molecular Ecology*, 20(5): 981–989.

Schiffels, S., Szöllosi, G. J., Mustonen, V., and Lässig, M. 2011. Emergent neutrality in adaptive asexual evolution. *Genetics*, 189(4): 1361–1375.

Schoener, T. W. 2011. The newest synthesis: understanding the interplay of evolutionary and ecological dynamics. *Science*, 331(6016): 426–429.

Schweizer, G., Muench, K., Mannhaupt, G., Schirawski, J., Kahmann, R., and Dutheil, J. Y. 2018. Positively selected effector genes and their contribution to virulence in the Smut fungus *Sporisorium reilianum*. *Genome Biology and Evolution*, 10(2): 629–645.

Segarra, J. 2005. Stable polymorphisms in a two-locus gene-for-gene system. *Phytopathology*, 95(7): 728–736.

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(1196): 541–555.

Shi, M., Lin, X. D., Tian, J. H., Chen, L. J., Chen, X., Li, C. X., Qin, X. C., Li, J., Cao, J. P., Eden, J. S., Buchmann, J., Wang, W., Xu, J., Holmes, E. C., and Zhang, Y. Z. 2016. Redefining the invertebrate RNA virosphere. *Nature*, 540(7634): 539–543.

Shin, J. and MacCarthy, T. 2015. Antagonistic coevolution drives Whack-alpha-mole sensitivity in gene regulatory networks. *PLOS Computational Biology*, 11(10).

Simmonds, P. 2018. A clash of ideas - the varying uses of the 'species' term in virology and their utility for classifying viruses in metagenomic datasets. *Journal of General Virology*, 99(3): 277–287.

Simmonds, P., Aiewsakun, P., and Katzourakis, A. 2018. Prisoners of war — host adaptation and its constraints on virus evolution. *Nature Reviews Microbiology*.

Sisson, S. A., Fan, Y., and Tanaka, M. M. 2007. Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, 104(6): 1760–1765.

Song, Y., Gokhale, C. S., Papkou, A., Schulenburg, H., and Traulsen, A. 2015. Host-parasite coevolution in populations of constant and variable size. *BMC Evolutionary Biology*, 15(1): 1–15.

Spanakis, E. and Horne, M. 1987. Co-adaptation of *Escherichia coli* and Coliphage λvir in continuous culture. *Journal of General Microbiology*, 133(2): 353–360.

Staedler, T., Haubold, B., Merino, C., Stephan, W., and Pfaffelhuber, P. 2009. The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics*, 182(1): 205–216.

Stahl, E. and Bishop, J. 2000. Plant-pathogen arms races at the molecular level. *Current Opinion in Plant Biology*, 3(4): 299–304.

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

Sunnaker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., and Dessimoz, C. 2013. Approximate Bayesian Computation. *PLOS Computational Biology*, 9(1).

Suttle, C. A. 2005. Viruses in the sea. *Nature*, 437: 356–361.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3): 585–595.

Talas, F., Kalih, R., Miedaner, T., and McDonald, B. A. 2016. Genome-wide association study identifies novel candidate genes for aggressiveness, deoxynivalenol production, and azole sensitivity in natural field populations of *Fusarium graminearum*. *Molecular Plant-Microbe Interactions*, 29(5): 417–430. PMID: 26959837.

Tamura, K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-tranversion and G+C content biases. *Molecular Biology and Evolution*, 9(4): 678–687.

Tamura, K. and Nei, M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial-DNA in humans and chimpanzees. *Molecular Biology and Evolution*, 10(3): 512–526.

Tellier, A. and Brown, J. K. M. 2007a. Polymorphism in multilocus host-parasite coevolutionary interactions. *Genetics*, 177(3): 1777–1790.

Tellier, A. and Brown, J. K. M. 2007b. Stability of genetic polymorphism in host-parasite interactions. *Proceedings of the Royal Society B-Biological Sciences*, 274(1611): 809–817.

Tellier, A. and Lemaire, C. 2014. Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. *Molecular Ecology*, 23(11): 2637–2652.

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

Thompson, J. N. 2005. *The geographic mosaic of coevolution*. University of Chicago Press, 1 edition.

Thompson, J. N. 2009. Which ecologically important traits are most likely to evolve rapidly? *Oikos*, 118(9): 1281–1283.

Thrall, P. and Burdon, J. 2000. Effect of resistance variation in a natural plant host-pathogen metapopulation on disease dynamics. *Plant Pathology*, 49(6): 767–773.

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

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

Tian, D., Traw, M., Chen, J., Kreitman, M., and Bergelson, J. 2003. Fitness costs of R-gene-mediated resistance in *Arabidopsis thaliana*. *Nature*, 423(6935): 74–77.

Tollenaere, C., Susi, H., and Laine, A.-L. 2016. Evolutionary and epidemiological implications of multiple infection in plants. *Trends in Plant Science*, 21(1): 80–90.

Van de Weyer, A.-L., Monteiro, F., Furzer, O. J., Nishimura, M. T., Cevik, V., Witek, K., Jones, J. D., Dangl, J. L., Weigel, D., and Bemm, F. 2019. The *Arabidopsis thaliana* pan-nlrome. *bioRxiv*.

Van Etten, J., Burbank, D., Xia, Y., and Meints, R. 1983. Growth-cycle of a virus, PBCV-1, that infects *Chorella*-like algae. *Virology*, 126(1): 117–125.

van Velzen, E. and Gaedke, U. 2018. Reversed predator–prey cycles are driven by the amplitude of prey oscillations. *Ecology and Evolution*, 8(12): 6317–6329.

Velzen, E. V. and Gaedke, U. 2017. Disentangling eco-evolutionary dynamics of predator-prey coevolution: the case of antiphase cycles. *Scientific Reports*, 7(1): 17125.

Verin, M. and Tellier, A. 2018. Host-parasite coevolution can promote the evolution of seed banking as a bet-hedging strategy. *Evolution*, 72(7): 1362–1372.

Vitti, J. J., Grossman, S. R., and Sabeti, P. C. 2013. Detecting natural selection in genomic gata. *Annual Review of Genetics*, 47(1): 97–120.

Wakeley, J. 2008. *Coalescent theory: An introduction*. W H Freeman, XY, 1 edition.

Wakeley, J. and Aliacar, N. 2002. Gene genealogies in a metapopulation (vol 159, pg 893, 2001). *Genetics*, 160(3): 1263.

Wang, M., Roux, F., Bartoli, C., Huard-Chauveau, C., Meyer, C., Lee, H., Roby, D., McPeek, M. S., and 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*, page 201710980.

Watterson, G. 1975. Number of segregating sites in genetic models without recombination. *Theoretical Population Biology*, 7(2): 256–276.

Wegmann, D., Leuenberger, C., and Excoffier, L. 2009. Efficient Approximate Bayesian Computation coupled With Markov Chain Monte Carlo without likelihood. *Genetics*, 182(4): 1207–1218.

Wegmann, D., Leuenberger, C., Neuenschwander, S., and Excoffier, L. 2010. ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics*, 11.

Weigand, H. and Leese, F. 2018. Detecting signatures of positive selection in non-model species using genomic data. *Zoological Journal of the Linnean Society*, pages 528–583.

Wilfert, L. and Jiggins, F. M. 2012. The dynamics of reciprocal selective sweeps of host resistance and a parasite counter-adaptation in *Drosophila*. *Evolution*, 67(3): 761–773.

Woolhouse, M. E. J., Webster, J. P., Domingo, E., Charlesworth, B., and Levin, B. R. 2002. Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nature Genetics*, 32(4): 569–577.

Yoshida, M., Yoshida, T., Kashima, A., Takashima, Y., Hosoda, N., Nagasaki, K., and Hiroishi, S. 2008. Ecological dynamics of the toxic bloom-forming cyanobacterium *Microcystis aeruginosa* and its cyanophages in freshwater. *Applied and Environmental Microbiology*, 74(10): 3269–3273.

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

Zhu, Y., Yongky, A., and Yin, J. 2009. Growth of an RNA virus in single cells reveals a broad fitness distribution. *Virology*, 385(1): 39–46.

Živković, D., Steinruecken, M., Song, Y. S., and Stephan, W. 2015. Transition Densities and Sample Frequency Spectra of Diffusion Processes with Selection and Variable Population Size. *Genetics*, 200(2): 601+.

Živković, D., John, S., Verin, M., Stephan, W., and Tellier, A. 2019. Neutral genomic signatures of host-parasite coevolution. *bioRxiv*.