Keeping Pace with the Red Queen: Identifying the Genetic Basis of Susceptibility to Infectious Disease

Ailene MacPherson,*,†,1 Sarah P. Otto,*,† and Scott L. Nuismer‡
*Department of Zoology, and †Biodiversity Research Centre, University of British Columbia, Vancouver, V6T 1Z4, Canada, and ‡Department of Biological Sciences, University of Idaho, Moscow, Idaho 83844
ORCID IDs: 0000-0003-3042-0818 (S.P.O.); 0000-0001-9817-0056 (S.L.N.)

ABSTRACT Genome-wide association studies are widely used to identify "disease genes" conferring resistance/susceptibility to infectious diseases. Using a combination of mathematical models and simulations, we demonstrate that genetic interactions between hosts and parasites [genotype-by-genotype (G × G) interactions] can drastically affect the results of these association scans and hamper our ability to detect genetic variation in susceptibility. When hosts and parasites coevolve, these G × G interactions often make genome-wide association studies unrepeatable over time or across host populations. Reanalyzing previously published data on *Daphnia magna* susceptibility to infection by *Pasteuria ramosa*, we identify genomic regions consistent with G × G interactions. We conclude by outlining possible avenues for designing more powerful and more repeatable association studies.

KEYWORDS host–parasite coevolution; GWAS; *Daphnia magna*; *Pastueria ramosa*; genetic architecture of resistance

INFECTIOUS diseases are pervasive. So pervasive, in fact, that without effective mechanisms of resistance, host populations can be quickly reduced in size or even driven to extinction. For instance, chestnut blight effectively wiped out the American chestnut, which had little if any resistance to this novel pathogen, after its introduction to North America in the early 1900s (Anagnostakis 2000; Anderson *et al.* 2004). Similarly, when Myxoma virus was introduced to Australia in the 1950s, local rabbit populations were almost entirely susceptible, resulting in millions of deaths and the decimation of local populations (Ratcliffe *et al.* 1952). Human populations, too, have been heavily affected by infectious disease in the past, perhaps most notably during the 1918 influenza pandemic that killed >50 million people before fading away in 1920 (Johnson and Mueller 2002; Taubenberger and Morens 2006). Although these examples are striking and demonstrate the impact of unchecked infectious disease, they are far from the norm. More commonly, host populations have effective mechanisms of resistance against pathogens they encounter regularly (Revers *et al.* 2014), with significant variability between populations depending on their history of exposure (Bartholomew 1998; Weatherall and Clegg 2002).

The existence of substantial variation in resistance to infectious disease within host populations has generated hope that it may be possible to identify the genes conferring resistance. Identifying such resistance genes would pave the way for genetic engineering of resistant crops and livestock, focus drug development efforts on likely targets, and open the door to gene therapeutic approaches within human populations. As the genomic revolution has progressed, it has become increasingly common to search for these "resistance genes" using genome-wide association studies (GWAS) (Newport and Finan 2011; Rowell *et al.* 2012). Loosely speaking, these studies compare the marker genotypes of individuals infected with disease and those uninfected and ask which loci predict an individual's infection status. The GWAS approach has now been used to successfully identify a range of candidate genes thought to be important in resistance to infectious disease in plants and animals (Chapman and Hill 2012; Khor and Hibberd 2012; Wang *et al.* 2012; Zila *et al.* 2013; Gurung *et al.* 2014).

Despite the successes of the GWAS approach in some cases, it is becoming increasingly recognized that the approach has

doi: https://doi.org/10.1534/genetics.117.300481
Manuscript received November 6, 2017; accepted for publication December 1, 2017; published Early Online December 8, 2017.
Available freely online through the author-supported open access option.
Supplemental material is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.117.300481/-/DC1.
[1]Corresponding author: Department of Zoology and Biodiversity Research Centre, University of British Columbia, 6270 University Blvd., Vancouver, British Columbia V6T 1Z4, Canada. E-mail: amacp@zoology.ubc.ca

significant limitations. For instance, GWAS are most powerful when resistance depends on common genetic variants with relatively large phenotypic effects (Manolio *et al.* 2009). In addition, which candidate genes are identified by this method may depend on the environment in which the study is conducted (Thomas 2010). These limitations apply to GWAS in general, not just those studies focused on infectious disease, and are widely recognized. When GWAS are used to understand the genetic basis of resistance to infectious disease, however, a potentially more important problem arises. Specifically, the resistance genes identified within the host population may depend on the genetic composition of the infectious disease itself (Newport and Finan 2011). This sensitivity of the GWAS approach to the genetic composition of the infectious disease becomes acute any time genotype-by-genotype (G × G) interactions exist; in other words, when particular combinations of host and pathogen genes yield resistance whereas other combinations lead to susceptibility. These G × G interactions may have drastic effects on the results of genetic association studies and our understanding of disease resistance (Lambrechts 2010), similar to the effects of gene-by-environment interactions. One particularly disconcerting possibility is that rapid pathogen evolution or host–pathogen coevolution will cause the host resistance genes that can be identified by GWAS to fluctuate rapidly over time.

Here we quantitatively explore the performance of GWAS when resistance to infectious disease involves G × G interactions between host and disease. We begin by presenting a general mathematical model of an association study to investigate disease resistance and evaluate the role of G × G interactions for several forms of host–parasite interactions. We then simulate host–pathogen coevolution to illustrate the extent to which G × G interactions may vary across time and/or space. We conclude by reanalyzing published genome-wide association data (Bourgeois *et al.* 2017) of *Daphnia magna* resistance to its *Pasteuria ramosa* pathogen, distinguishing regions of the genome associated with overall health from those involved in resistance specific to a particular *P. ramosa* strain.

Model

We consider a scenario, common in practice, where host resistance is measured as a continuous quantitative trait. This would be the case, for instance, if host resistance is assessed by measuring viral load, duration of infection, or damage to host tissues. Our model assumes that host resistance depends on the value of a quantitative trait in the host, z_H, relative to the value of a quantitative trait in the pathogen, z_P. Specifically, we assume host susceptibility, S, is given by the following function:

$$S = f(z_H - z_P). \tag{1}$$

The function f is sufficiently general to accommodate many commonly observed resistance mechanisms. For instance, in the interaction between the snail *Biomphalaria glabrata* and its trematode parasites, resistance depends on the relative quantities of reactive oxygen molecules in the snail (z_H) and reactive oxygen scavenging molecules produced by the parasite (z_P) (Bayne 2009; Mon *et al.* 2011). In cases like these, the function f may take a sigmoid form which we call the phenotypic-difference model (Figure 1A) (Nuismer *et al.* 2007; Ashby and Boots 2017):

$$f(z_H - z_P) = \frac{1}{1 + e^{\alpha(z_H - z_P)}}. \tag{2}$$

In contrast, in the interaction between the schistosome parasite, *Schistosoma mansoni*, and its snail host, *B. glabrata*, resistance depends on the degree to which the conformation of defensive FREP molecules produced by the snail (z_H) match the conformation of parasite mucin molecules (z_P) and successfully bind to them (Mitta *et al.* 2012). In such cases, the function f may take a Gaussian form which we call a phenotypic-matching model (Figure 1B) (Kopp and Gavrilets 2006):

$$f(z_H - z_P) = e^{-\alpha(z_H - z_P)^2}. \tag{3}$$

To study the effects of genetic interactions on susceptibility to infection, S, we must integrate genetics into our phenotypic model. For a haploid host and pathogen where z_H and z_P depend on n_H and n_P biallelic loci, respectively, we can write general expressions for these phenotypes as functions of alleles present in each species:

$$\begin{aligned} z_H &= b_{H0} + \sum_{i=1}^{n_H} b_{Hi} X_{Hi} + \sum_{\substack{i,j \\ i \neq j}}^{n_H} b_{Hi,Hj} X_{Hi} X_{Hj} \\ &\quad + \sum_{\substack{i,j,k \\ i \neq j \neq k}}^{n_H} b_{Hi,Hj,Hk} X_{Hi} X_{Hj} X_{Hk} + \ldots + \epsilon_H \\ z_P &= b_{P0} + \sum_{i=1}^{n_P} b_{Pi} X_{Pi} + \sum_{\substack{i,j \\ i \neq j}}^{n_P} b_{Pi,Pj} X_{Pi} X_{Pj} \\ &\quad + \sum_{\substack{i,j,k \\ i \neq j \neq k}}^{n_P} b_{Pi,Pj,Pk} X_{Pi} X_{Pj} X_{Pk} + \ldots + \epsilon_P \end{aligned} \tag{4}$$

In these expressions, X_{Mi} is an indicator variable describing the allelic state (0 or 1) of an individual of species M at locus i, b_{M0} is the phenotype of an individual of species M with all "0" alleles, and b_{Mi} is the additive effect of carrying a "1" allele at locus i in species M. The remaining coefficients ($b_{Mi,Mj}$, $b_{Mi,Mj,Mk}$, *etc.*) describe epistatic interactions among loci. Finally, ϵ_M captures an environmental contribution to the phenotype of species M, which is assumed to have mean 0, a constant variance, and be uncorrelated with an individual's phenotype. Substituting Equation 4 into Equation 1 yields a

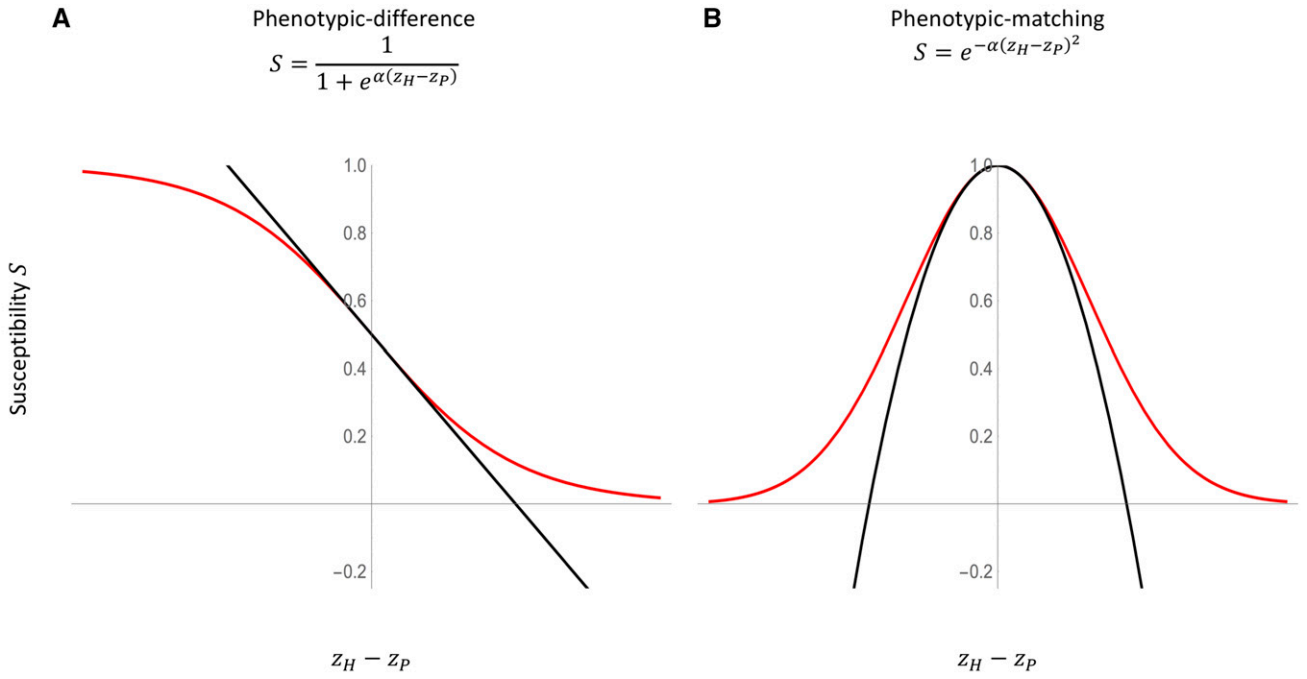

Figure 1 Host–parasite interaction models. Susceptibility to infection as a function of the distance between host and pathogen phenotypes, $z_H - z_P$, for the (A) phenotypic-difference and (B) phenotypic-matching model. Red curves show exact functions whereas black curves are the quadratic approximations.

model of host susceptibility as a function of host and pathogen genotypes.

Our goal now is to use this genetic model to predict the sensitivity of GWAS to the genetic composition of the pathogen population. We will explore both traditional, single-species GWAS approaches and a novel approach that takes genetic information from both host and pathogen into account (co-GWAS). Our investigation will rely on a pair of complementary approaches. First, we will develop and analyze analytical approximations that quantify the sensitivity of GWAS and co-GWAS approaches to changes in pathogen genotype frequencies. These analytical approximations will rely on simplified genotype–phenotype maps and will not explicitly integrate evolution and coevolution. Second, we will develop and analyze simulations that allow us to explore the consequences of rapid pathogen evolution and coevolution between the species on the performance of both GWAS and co-GWAS approaches.

Analytical Approximation

To simplify the genetic model of resistance developed in the previous section sufficiently for mathematical analysis, we begin by considering the case where $n_H = n_P = 2$. In addition, we assume that the phenotypes of host and pathogen are not too far from one another, such that the quantity $z_H - z_P$ is small relative to the extent of phenotypic specificity (α in Equations 2 and 3). Under this assumption, Equation 1 can be approximated by its second order Taylor series expansion. This allows the genetic model of susceptibility to be simplified to the following approximate expression:

$$\begin{aligned} S \approx\, & f(0) + f'(0)[(b_{H0} + b_{H1}X_{H1} + b_{H2}X_{H2} + \epsilon_H) \\ & - (b_{P0} + b_{P1}X_{P1} + b_{P2}X_{P2} + \epsilon_P)] \\ & + \frac{1}{2}f''(0)[(b_{H0} + b_{H1}X_{H1} + b_{H2}X_{H2} + \epsilon_H) \\ & - (b_{P0} + b_{P1}X_{P1} + b_{P2}X_{P2} + \epsilon_P)]^2 \\ & + \mathcal{O}\left[(z_H - z_P)^3\right], \end{aligned} \tag{5}$$

where primes indicate derivatives with respect to the distance between host and pathogen phenotypes. With (5) in hand, we have a model that predicts host resistance as a function of host and pathogen genotypes. In the following two sections, we will use (5) to investigate how the genetic composition of the pathogen population influences the results of GWAS and co-GWAS. Extending these models to complete G × G association studies requires a large number of pathogen loci ($n_P \gg 2$) and thus may be computationally prohibitive. For many pathogens, however, strain type or subtype may be known and capture much of the relevant genetic variation in the pathogen population. In these cases, tracking pathogen types can greatly reduce the effective number of loci, even to $n_P = 2$ as in Equation 5. Such simplifications should allow us to expand beyond two host loci to a whole host genome ($n_H \gg 2$), while avoiding the computational complexity of tracking all possible genetic interactions between the full host genome and the full parasite genome.

Single-species GWAS

We envision a standard GWAS where susceptibility to infection has been measured for some number of host individuals,

each of which has also been genotyped at a large number of marker loci. To focus our model on the effects of species interactions, we will assume this data accurately provides us with the genotype of individuals at the two host resistance loci. Using this data, the goal of the genetic association study is to partition host susceptibility between these genes relative to their effects. This can be done by fitting susceptibility with a linear combination of the genetic indicator variables:

$$S \approx \beta_{H0} + \beta_{H1}X_{H1} + \beta_{H2}X_{H2} + \beta_{H1,H2}X_{H1}X_{H2}, \qquad (6)$$

where the β coefficients can be found using least squares regression. The biological interpretation of this linear model is straightforward. The intercept coefficient, β_{H0}, is the expected host resistance when both 0 host alleles are present. The coefficients β_{H1} and β_{H2} are the inferred additive effects of the 1 alleles at the first and second loci, respectively, and $\beta_{H1,H2}$ captures the epistatic interaction between the two host 1 alleles. Solving for the coefficients in (6) we have (see Supplemental Material, *Mathematica* notebook in File S1):

$$\begin{aligned}
\beta_{H0} &= f(0) + f'(0)\left[\tilde{b}_{H0} - \left(\tilde{b}_{P0} + b_{P1}q_{P1} + b_{P2}q_{P2}\right)\right] \\
&\quad + \frac{1}{2}f''(0)\left\{\left(\tilde{b}_{H0} - \tilde{b}_{P0}\right)^2 + b_{P1}^2 q_{P1} + b_{P2}^2 q_{P2}\right. \\
&\quad + 2\left[(b_{P1}q_{P1} + b_{P2}q_{P2})\left(\tilde{b}_{H0} - \tilde{b}_{P0}\right)\right. \\
&\quad \left.\left. - b_{P1}b_{P2}(q_{P1}q_{P2} + D_P)\right]\right\} \\
\beta_{Hi} &= f'(0)b_{Hi} \\
&\quad + \frac{1}{2}f''(0)\left[b_{Hi}^2 + 2b_{Hi}(b_{H0} - b_{P0} - q_{P1}b_{P1} - q_{P2}b_{P2})\right] \\
\beta_{H1,H2} &= f''(0)b_{H1}b_{H2}, \qquad (7)
\end{aligned}$$

for $i = \{1,2\}$, where $f(0), f'(0)$, and $f''(0)$ are the resistance function and its first and second derivative evaluated at 0 as in Equation 5, and where $\tilde{b}_{H0} = b_{H0} + \epsilon_H$ and $\tilde{b}_{P0} = b_{P0} + \epsilon_P$. Importantly, these expressions for the coefficients depend on the allele frequency at the pathogen loci, q_{P1} and q_{P2}, as well as the linkage disequilibrium between them, D_P. Note that the relevant allele frequencies and linkage disequilibrium are among pathogens to which the host is exposed, which may not be equivalent to the pathogen population as a whole.

As a result of the dependence of the coefficients in (7) on the pathogen allele frequencies and linkage disequilibrium, the allelic effects (β's) inferred by a host-only GWAS can be quite sensitive to the genetic composition of the pathogen population (Figure 2). Changes in pathogen allele frequency can alter the magnitude and sign of the inferred effects. From a practical standpoint, if susceptibility is assayed in two host populations that are exposed to pathogen populations that differ greatly in their allele frequencies, one may find a host allele has a protective effect in one population but increases risk in the other. Similar to hidden host population structure, uncontrolled differences in the pathogen population can greatly alter the inferences of single-species GWAS.

A second result that can be drawn from Equation 7 is that when the resistance function is approximately linear, $f''(0) = 0$, the inferred additive and epistatic effects, β_{H1}, β_{H2}, and $\beta_{H1,H2}$ are independent of the pathogen allele frequencies. For example, in contrast to the nonlinear phenotypic-matching model where the inferred effects vary with pathogen allele frequency, the inferred effects remain constant in the approximately linear phenotypic-difference model (Figure 2). A third conclusion from Equation 7 is that, at least under the assumption that $z_H - z_P$ is small, the epistatic interaction between the host loci, $\beta_{H1,H2}$, is independent of pathogen genetics. We will explore the consequences of this dependence on the pathogen allele frequencies for the stability of GWAS-inferred effects across evolutionary time (See the *Host–Parasite Coevolution* section below).

In addition to identifying the allelic effects on host resistance, an important metric of GWAS performance is the proportion of phenotypic variation explained by the identified causative loci. Given the dependence of the estimated allelic effects on pathogen allele frequencies, we calculated the total phenotypic variation explained by the host loci across the range of pathogen allele frequencies (Figure 2, C and D). When the pathogen population is monomorphic ($q_{P1} = q_{P2} = 0$ or 1), the host loci can explain 100% of the genetic variation in the phenotype. If the pathogen population is polymorphic, however, the host-only approach may explain as little as 10% of the variation. Partitioning the total variation explained into the additive and epistatic contributions demonstrates that, due to changes in the additive effect size b_{Hi}, the relative contribution of additive and epistatic effects also varies with pathogen allele frequency and depends on the form of the host–parasite interaction.

Two-species co-GWAS

The results derived in the previous section demonstrate that traditional single-species GWAS may be sensitive to the genetic composition of the pathogen population at loci involved in host–pathogen specificity. In this section, we attempt to overcome this problem by developing an alternative GWAS design in which both host and pathogen genetics are incorporated. In contrast to the traditional method where only host genotypes are recorded, this design requires that both host and pathogen genotypes are known. As with Equation 6, we now attempt to fit host resistance as a linear function of the allelic indicator variables, but we include pathogen indicators as well as interaction terms between host and pathogen loci:

$$\begin{aligned}
S \approx\ & \beta_0 + \beta_{H1}X_{H1} + \beta_{H2}X_{H2} + \beta_{H1,H2}X_{H1}X_{H2} + \beta_{P1}X_{P1} \\
& + \beta_{P2}X_{P2} + \beta_{P1,P2}X_{P1}X_{P2} + \beta_{H1,P1}X_{H1}X_{P1} \\
& + \beta_{H1,P2}X_{H1}X_{P2} + \beta_{H2,P1}X_{H2}X_{P1} + \beta_{H2,P2}X_{H2}X_{P2}. \qquad (8)
\end{aligned}$$

As with Equation 6, the coefficients of this equation have straightforward biological interpretations. The intercept, β_0, describes the expected host resistance when all host and pathogen loci have

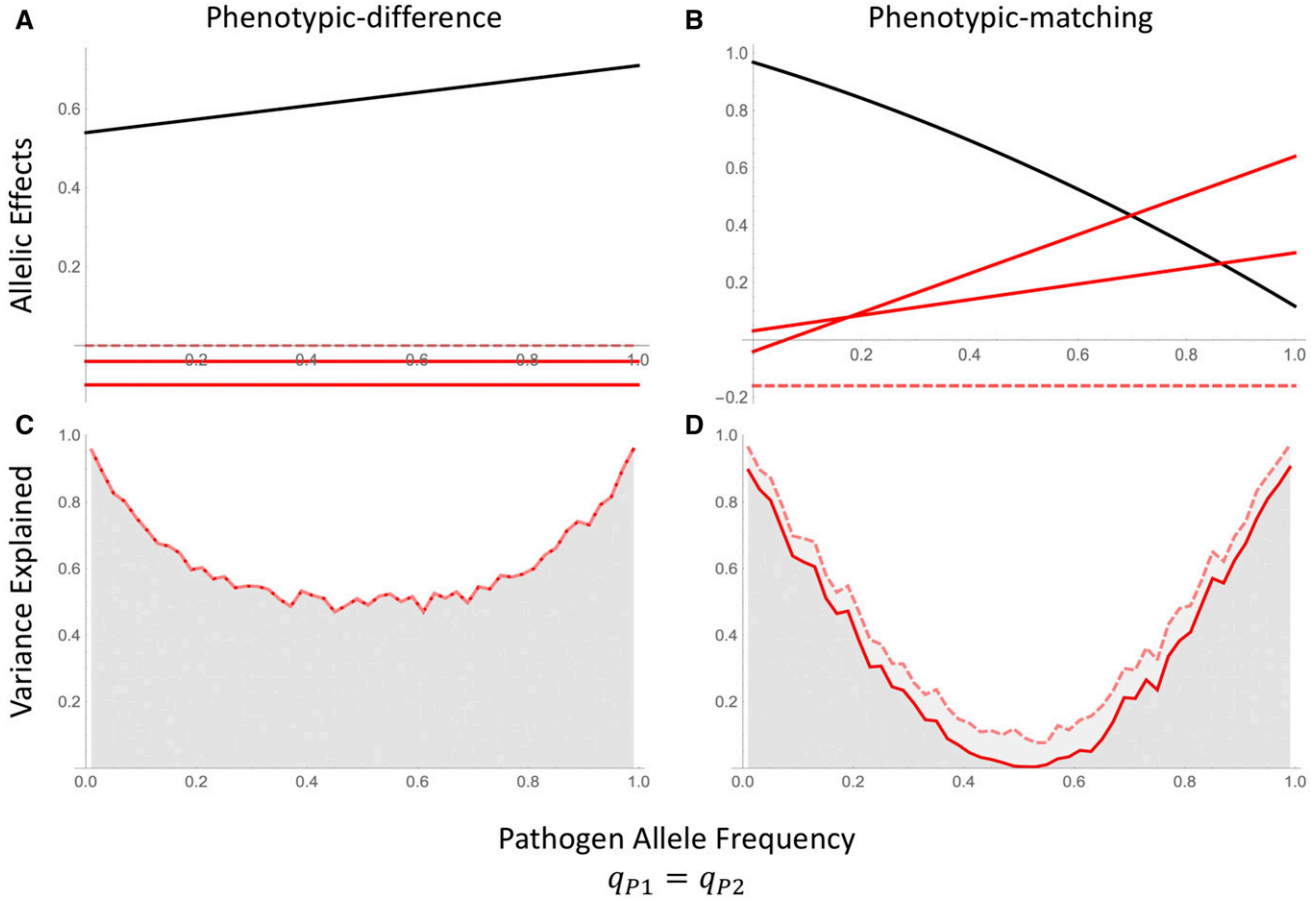

Figure 2 Host-only model with resistance dependent on phenotypic differences (A,C) or phenotypic matching (B,D) between hosts and parasites. (A and B) Allelic effects inferred using the host-only design from Equation 6: β_0 (black), β_{H1} and β_{H2} (solid red lines), $\beta_{H1,H2}$ (dashed red). (C and D) Variation explained by host additive effects only (solid line), and host additive and epistatic effects (dashed line) as given by the host-only model in (6).

0 alleles. Terms 2, 3, 5, and 6 describe the additive effects of each individual host and pathogen 1 allele; and terms 4 and 7 describe the epistatic interactions between loci within the host and pathogen, respectively. The remaining four terms describe the G × G interactions between pairs of host and pathogen loci.

Despite the complexity of Equation 8, and hence the logistical and computational challenges of applying it, the expressions for each of these coefficients in terms of the host and pathogen phenotypic effects are simple (see *Mathematica* notebook in File S1):

$$\beta_{H0} = f(0) + f'(0)\left(\tilde{b}_{H0} - \tilde{b}_{P0}\right) + \frac{1}{2}f''(0)\left(\tilde{b}_{H0} - \tilde{b}_{P0}\right)^2$$

$$\beta_{Hi} = f'(0)b_{Hi} + \frac{1}{2}f''(0)b_{Hi}\left[b_{Hi} + 2\left(\tilde{b}_{H0} - \tilde{b}_{P0}\right)\right] \quad \text{for } i = \{1,2\}$$

$$\beta_{H1,H2} = f''(0)b_{H1}b_{H2}$$

$$\beta_{Pi} = -f'(0)b_{Pi} - \frac{1}{2}f''(0)b_{Pi}\left[b_{Pi} + 2\left(\tilde{b}_{H0} - \tilde{b}_{P0}\right)\right] \quad \text{for } i = \{1,2\}$$

$$\beta_{P1,P2} = f''(0)b_{P1}b_{P2}$$

$$\beta_{Hi,Pj} = -f''(0)b_{Hi}b_{Pj} \quad \text{for } i = \{1,2\},\ j = \{1,2\}. \tag{9}$$

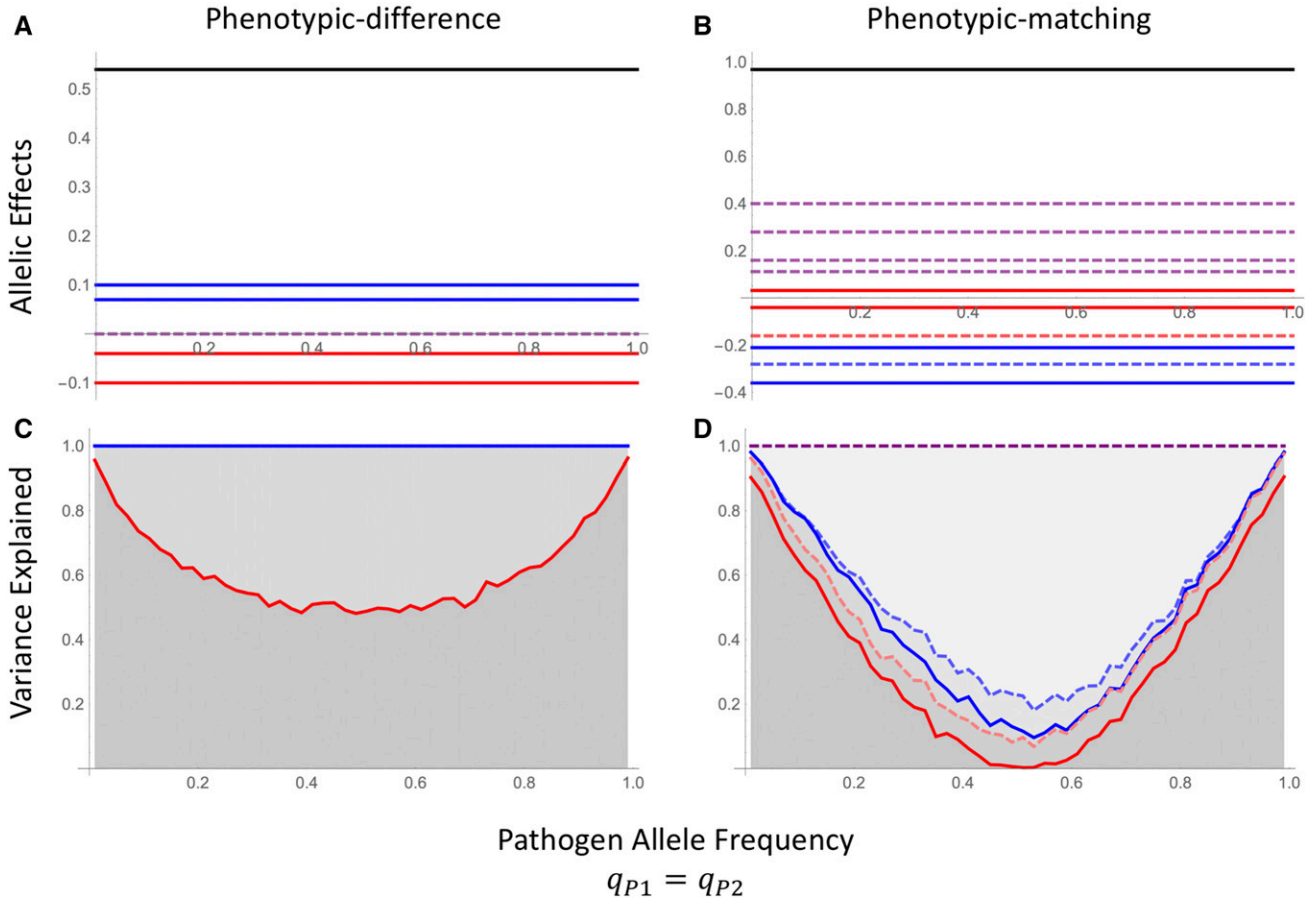

Figure 3 Host–pathogen model with phenotypic-difference (A,C) or phenotypic-matching (B,D) based resistance. (A and B) Allelic effects inferred using the host-parasite design from Equation 8: β_0 (black), β_{Hi} (solid red), $\beta_{H1,H2}$ (dashed red), β_{Pi} (solid blue), $\beta_{P1,P2}$ (dashed blue), and $\beta_{Hi,Pj}$ (dashed purple). (C and D) Variation explained by host additive effects only (solid red), host additive and epistatic effects (dashed red), host and pathogen additive and epistatic effects (dashed blue), and a full host–pathogen model as given in Equation 8.

Comparing the equations in (9) with the coefficients in (7) reveals an important conclusion: the effect sizes no longer depend on the pathogen allele frequencies nor the linkage disequilibrium (Figure 3, A and B). This result suggests that the two-species, co-GWAS approach is more robust to changes in the genetic composition of the pathogen population and thus may be much less sensitive to rapid evolution and spatial genetic structuring within the pathogen population.

In addition to stabilizing the estimated allelic effects across pathogen allele frequencies, the total phenotypic variation explained by the co-GWAS greatly exceeds that of the host-only GWAS. For the two-locus case explored here, the co-GWAS approach can explain 100% of the variation regardless of pathogen allele frequency (Figure 3, C and D). The contributions of additive, epistatic, and G × G interactions do, however, vary with pathogen allele frequency. As with the host-only approach, when the pathogen population is monomorphic the host effects explain all of the observed phenotypic variation. In summary, unlike the host-only model, the effect size coefficients (Equation 9) and the total variation explained, no longer vary with pathogen allele frequency. This contrast between the host-only and co-GWAS approaches is particularly relevant any time the composition of the pathogen population is likely to differ between the sample used for the association study and the population in which the resulting inferences are applied. In the following section we explore how temporal changes in the host and pathogen populations driven by coevolution affects the reproducibility of GWAS over time and, by extension, space.

Host–Parasite Coevolution

To simulate host–parasite coevolution, we envision a system where each host comes into contact with a single parasite each generation. The probability that this contact results in infection is determined by host susceptibility, S, which is a function of the host and parasite genotype. Infected hosts experience a fitness cost ξ_H, whereas their infecting parasites receive a fitness benefit ξ_P. In the absence of infection, both hosts and parasites have a fitness of 1. Together, these assumptions lead to the following fitness of a host with genotype $\{X_{H1}, X_{H2}\}$ that comes into contact with a pathogen with genotype $\{X_{P1}, X_{P2}\}$:

$$W_H = 1 - \xi_H S(X_{H1}, X_{H2}, X_{P1}, X_{P2}); \quad (10)$$

whereas the pathogen has a fitness of

$$W_P = 1 + \xi_P S(X_{H1}, X_{H2}, X_{P1}, X_{P2}). \quad (11)$$

Given these fitnesses, we simulate allele frequencies and linkage disequilibrium over time assuming random mating, a per-locus mutation rate of μ, and a recombination rate r (see *Mathematica* notebook in File S1). We then use Equations 7 and 9 to calculate the inferred allelic effect sizes by using a host-only GWAS or co-GWAS for each generation over the course of coevolution for both the phenotypic-difference and phenotypic-matching models (Figure 4).

As expected, using the host-only GWAS approach, the inferred allelic effects can vary over time but only under the quadratic-shaped, phenotypic-matching model. As noted above, the estimated effects can even change sign, having large positive values when sampled in one generation and large negative values when sampled only a few generations later. In contrast, the inferred effects remain constant in the co-GWAS approach regardless of the coevolutionary model. In terms of the phenotypic variation explained, the host-only approach explains only a portion of genetically determined phenotypic variation, whereas the co-GWAS approach can explain up to 100%. The contribution of different genetic components to the total variation explained remains approximately constant under the phenotypic-difference model but varies rapidly as allele frequency changes in the phenotypic-matching model.

Data availability

The analysis, numerical simulations, and scripts to generate the original figures were coded in Wolfram *Mathematica* 11 (File S1) and are available for download from the Dryad Digital Repository (DOI: https://doi.org/10.5061/dryad.tb25q).

Daphnia–Pasteuria GWAS

Taken together, our analytical model and simulations illustrate that incorporating pathogen genetic information into the search for disease genes can greatly increase the explanatory power and repeatability of genome scans. Testing these theoretical predictions with biological data is a critical step in evaluating the power of the co-GWAS approach relative to a traditional single-species GWAS. Analysis of biological data will include several complications that we ignored above, including finite sample sizes, arbitrary forms of coevolutionary interactions, and complex genomic architectures. Unfortunately, we know of no studies that include full host and parasite genomic data as well as the outcome of infection experiments. Further, the computational tools to perform a co-GWAS in the form of Equation 8 do not yet exist. We can, however, use recently published data by Bourgeois *et al.* (2017) on the susceptibility of *D. magna* to two *P. ramosa* strains, C1 and C19, as a preliminary test of our analytical predictions. In particular, we compare the results of genome scans for C1 and C19 susceptibility analyzed separately to a single genome scan for susceptibility using all the data but ignoring pathogen strain type. Our analytical model predicts that, despite having half the sample size, the separate genome scans for C1 and C19 resistance should reveal loci that determine host–parasite specificity, whereas the full data scan will have lower power to do so. Note that strain type captures almost all of the relevant genetic information in this case, given that the parasite is clonal.

The original data set, provided on Dryad by the authors (Bourgeois *et al.* 2017), sampled 97 *D. magna* clones from three distinct geographic regions—1 site in Germany, 1 in Switzerland, and 11 sites in Finland—and provided the sequence at 6403 SNPs. Host susceptibility (S: susceptible; R: resistant) infection by each *P. ramosa* strain, C1 and C19, was determined by assessing whether fluorescently labeled spores attached to the host's esophagus (Duneau *et al.* 2011). All four possible combinations of susceptibility and resistance to the two strains (SS, SR, RS, and RR) were present. By performing two separate association studies, one for each strain, Bourgeois *et al.* (2017) used this experimental design to identify genomic regions associated with susceptibility to a specific parasite strain. Following the methods in the original work, we compare their results to a third genome scan including all the data, a total of 194 samples, ignoring the *Pasteuria* strain type tested. All genome scans were performed using the R package GenAbel, adjusting for population structure and repeated measures of the same host genotype using the Eigenstat method (Aulchenko *et al.* 2007).

To accurately assess which genomic regions are associated with susceptibility to C1, C19, and/or "overall" susceptibility from the complete data set, we used the *Daphnia* genetic map constructed by Dukić *et al.* (2016) to array the scaffolds into 10 linkage groups. To limit the detection of false positives, we followed an approach analogous to that used in Bourgeois *et al.* (2017) where SNPs were only considered significantly associated with a given susceptibility phenotype if there existed four SNPs in a 10-cM region with a log-likelihood score >2 (Figure 5). Multiple genomic regions are significantly associated with susceptibility to C1, C19, and to disease susceptibility in the complete data set without strain information. Four linkage groups (4, 5, 7, and 9), with a total of 28 significant SNPs, are associated with C1 susceptibility. Three linkage groups (1, 4, and 7) with 38 SNPs are associated with C19 susceptibility, and two linkage groups (4 and 5) with 35 SNPs are associated with susceptibility in the complete data set. Thus, while the complete data set has twice as many measures of disease susceptibility, it has less power to detect genetic regions underlying disease susceptibility because of the lack of parasite information.

The contrast between the associations for C1 and C19 susceptibility to overall susceptibility in the complete data

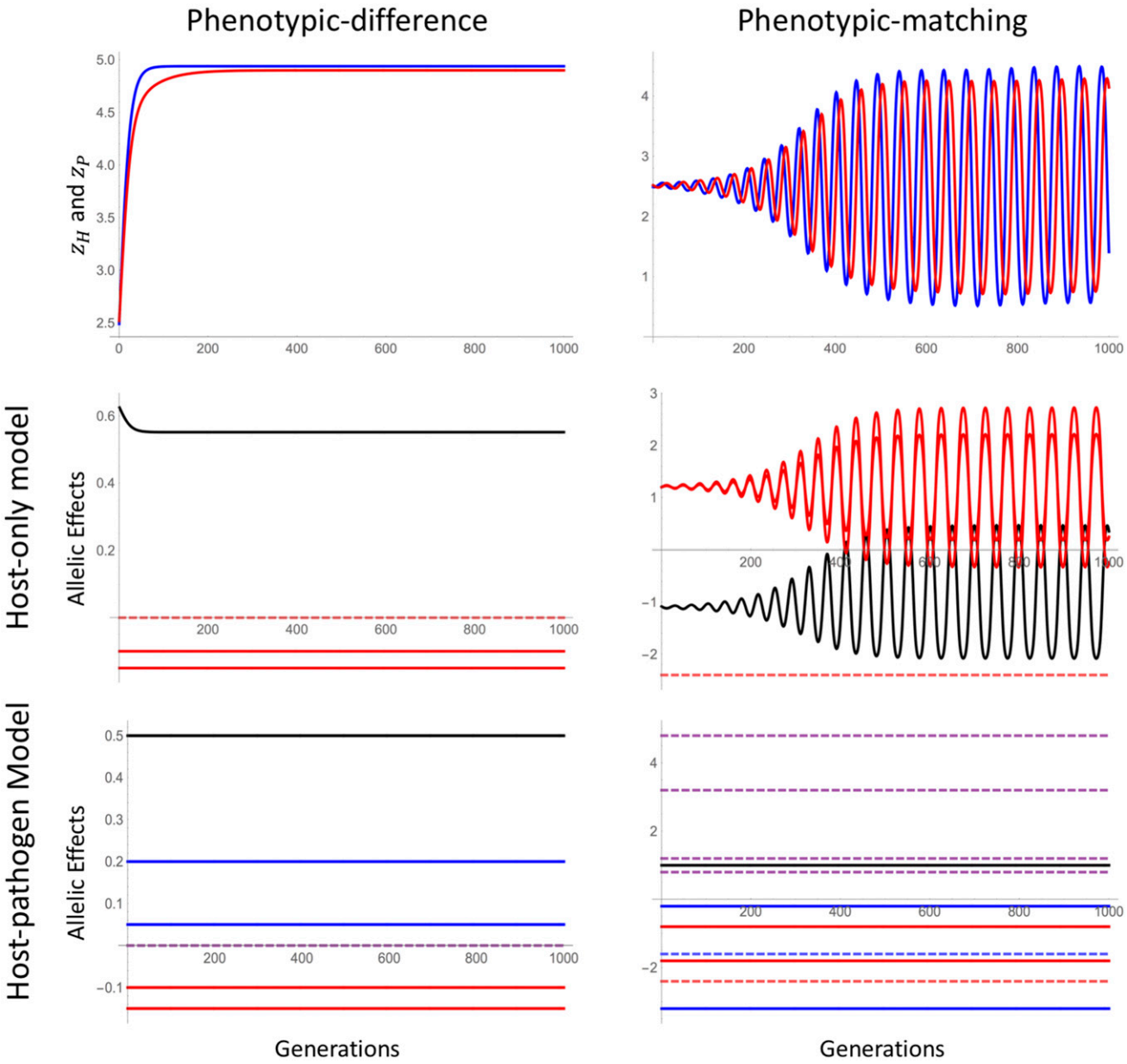

Figure 4 Allelic effects over coevolutionary time. Top row: Phenotypes z_H (red) and z_P (blue) simulated over coevolutionary time in the phenotypic-difference (left) and phenotypic-matching models (right). Middle row: Coefficients estimated under the host-only model (7) (black is β_0, solid red is β_{Hi}, dashed red is $\beta_{H1,H2}$). Bottom row: Coefficients estimated under the host–pathogen model (9) (black is β_0, solid red is β_{Hi}, dashed red is $\beta_{H1,H2}$, blue is β_{Pi}, dashed blue is $\beta_{P1,P2}$, purple dashed is $\beta_{Hi,Pj}$). Because epistatic and G × G interactions are absent in the phenotypic-difference model, their allelic effects all overlap at 0 and hence are not all visible.

set provides additional information about the nature of the genetic basis to resistance. Genomic regions associated with the overall resistance regardless of parasite type, particularly when these regions are also associated with C1 and C19 resistance, provide increased resistance regardless of the parasite strain tested and are consistent with general host health and nonspecific immune response. By contrast, sites that are not associated with overall resistance—despite the data set having twice the size—but are associated with either C1 or C19, are good candidates for loci that act in a parasite-specific manner. Examining Figure 5, we therefore conclude that linkage group 4 and possibly 5 are involved in general health and resistance. In contrast, the regions on the far left and right of linkage group 7 as well as the regions on linkage group 1 and 9, which are associated only with C1 or C19 resistance, are indicative of parasite-specific resistance loci.

These conclusions are in agreement with the hypothesized model and previous molecular work on *Daphnia* resistance to *Pasteuria*. In particular, resistance to *Pasteuria* is hypothesized to be controlled by a three-locus, matching-alleles system. One of these loci (the C locus) determines overall

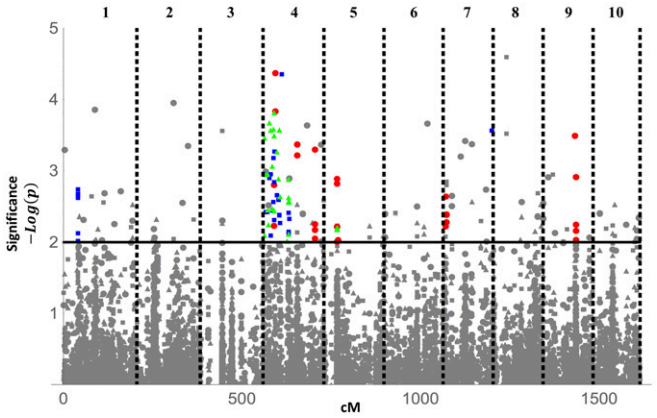

Figure 5 GWAS of *D. magna* susceptibility. Genetic associations of each SNP with C1 (red ●), C19 (blue ■), and overall susceptibility in the complete data set without parasite-type information (green ▲). Hence, each SNP is represented three times, once for each genome-wide scan. Note that closely linked SNPs often overlap with one another and are not all individually visible. Significant SNPs are shown in color while those below the log-likelihood of two threshold or that are not clustered within a 10-cM region of three other significant SNPs are shown in gray. The 10 linkage groups are delineated by vertical dashed lines.

host susceptibility regardless of pathogen strain and is thought to reside on linkage group 4 (Bento *et al.* 2017). In the absence of protection from the C locus, a second "A locus" is thought to confer resistance to C1 when the dominant allele is present. The regions detected on linkage groups 7 and 9 in the hosts exposed to C1 may only be candidates for such C1-specific resistance. Finally, if the C locus and A locus are both homozygous recessive, a third "B locus" determines susceptibility to the C19 strain. Such a locus would likely be hard to detect in a GWAS due to epistasis between the A, B, and C loci; nevertheless, the regions associated with only C19 resistance (on linkage groups 1 and 7) would be candidates for such a B locus. Overall we conclude that significant SNPs obtained without accounting for parasite type may signal general health status. Against this background, a co-GWAS can help identify genes whose regions are likely critical to host–parasite specificity and variation in host susceptibility.

Discussion

Identifying genes that determine a host's susceptibility to infection is a promising frontier with a wide range of applications, including agriculture and human health. Yet, as our mathematical models demonstrate, association studies focusing on identifying genes in a single species without accounting for the genetics of the interacting species can drastically affect our ability to detect disease genes involved in host–pathogen specificity and limit our ability to account for the genetic variation in disease susceptibility. When the genetic composition of the pathogen population varies over time and/or space, this can further lead to inconsistencies in the results of genetic association studies. Finally, using previously published data on *D. magna* resistance to its *Pasteuria* parasite, we illustrate that performing association studies with and without information about pathogen type can be used to distinguish genomic regions affecting general *vs.* specific resistance to pathogens. Consistent with current models for *Daphania–Pasteuria* interactions, we identify one region associated with general health as well as candidate regions more directly involved in mediating host–pathogen specificity.

The mathematical analysis presented above focuses on host–pathogen interactions of a specific form, given by Equation 1. Although we have relied on an approximation that assumes weak phenotype differences, *i.e.*, $z_H - z_P$ is small, we postulate that the power to detect strain-specific resistance genes will be increased whenever parasite information is incorporated, even when genes have major effects and phenotypic differences become large. Similarly, the methods used above can be extended to include alternative interaction types such as a "matching-alleles" interaction (see *Mathematica* notebook in File S1). The expressions for the β coefficients under this interaction model are unruly and difficult to interpret. Using a numerical approach, we observe that once again G × G interactions can explain a significant proportion of the variation in susceptibility (Figure S1 available on Dryad), particularly in highly variable pathogen populations. Unlike the phenotypic-difference and phenotypic-matching models, however, the co-GWAS approach (Equation 8) no longer explains all of the variation in susceptibility and the coefficients vary with pathogen allele frequency. This is a result of higher order interactions not included in our model. Hence, although the co-GWAS approach performs significantly better than a single-species approach, it will not always capture the full genetic basis of infection because of the second order approximation used in Equation 8.

Regardless of the form of the interaction, our analytical models and simulations illustrate that incorporating pathogen genetics into the search for disease genes can greatly increase the explanatory power and repeatability of genome scans. Unfortunately, several logistical and computational challenges preclude applying a full two-species GWAS. Most notably, such a design requires additional genetic data that is not currently available. More specifically, this design requires genotyping all hosts and the pathogens to which they are exposed, not just the host–parasite combinations observed in infected individuals. Future exploration is warranted to determine whether uninfected individuals can simply be treated as unknown with respect to pathogen exposure, and what the consequences of doing so would be for the statistical power of our approach.

The complexity of the two-species design (Equation 8) relative to that of a single-species design (Equation 6) also introduces computational challenges. In addition to requiring larger sample sizes, estimating the effects of the large number of potential G × G interactions in a full host–genome by parasite–genome study is computationally unrealistic. In

addition to the large number of pairwise interactions between hosts and pathogens, depending on the form of the interaction, higher order genetic interactions may be necessary to fully explain the variation in susceptibility. These higher order interactions can be particularly important as the number of loci underlying susceptibility, n_H and n_P, increases. Although incorporating complete pathogen genetic data may be unfeasible, there often exists some form of pathogen typing, which is largely indicative of the pathogen's genotype and may be sufficient for the purposes of a host genome-wide scan. For example, despite its vast diversity, Hepatitis C virus has been subdivided into seven genotypes (Irvine *et al.* 1993; Murphy *et al.* 2015), which may capture much of the relevant variation in host susceptibility.

The *Daphnia–Pasteuria* data set we analyzed provides a valuable test case for a two-species co-GWAS. In this study, we know exactly to which pathogen type individuals have been exposed, which is generally not known in natural populations. This information may have increased the power of the study to detect loci underlying C1 and C19 susceptibility. Despite this increased power, we chose to use the arguably lenient significance threshold of a log-likelihood score >2 plus clustering of four or more SNPs, as in the original article. Requiring more stringent threshold corrections for multiple sampling, such as a Bonferroni correction, does not yield any significant SNPs. Given the correspondence between the GWAS results and those of functional studies (Bento *et al.* 2017), however, many of the observed SNPs are arguably not false positives. Using the log-likelihood of two and clustering threshold, we observe fewer genomic regions associated with overall susceptibility when parasite information is not incorporated than when conducting GWAS with exposure to either C1 or C19, despite the complete data set containing twice the number of data points. As an alternative to analyzing the complete data set, we could hold the sample size constant in a combined analysis by randomly choosing whether the host was exposed to C1 or to C19 for each host genotype (Figure S2 available on Dryad). Interestingly, this "mixed" GWAS not only identifies the same regions on linkage group 4 and 5 but also identifies regions on linkage groups 1, 9, and 10, as were found in the single pathogen-type GWAS. The fact that this mixed analysis picks up some of the potentially parasite-specific loci is likely due to randomly sampling an excess of C1- or C19-tested clones. Consistent with this interpretation, exactly which parasite-specific regions are identified varies with the random sample chosen. Nevertheless, as with the complete data set, a comparison between C1, C19, and mixed susceptibility provides additional information about which genes are involved in general health *vs.* parasite-specific susceptibility.

The results presented here highlight several important avenues for future research. First and foremost, designing genome-wide association methods that allow for G $\times$ G interactions is critically important, as is the collection of genotypic data from hosts and pathogens. This could be approached, for example, by adapting GWAS designs and analyses used to detect gene-by-environment interactions (Winham and Biernacka 2013). Recognizing the importance of host–pathogen genetic interactions is important for understanding the applicability and limitations of single-species association scans. Developing metrics that capture relevant variability in host and pathogen populations may facilitate the application of these results. Finally, incorporating G $\times$ G interactions into our association studies will also enable us to understand what mathematical models of host–parasite interactions best predict the genetic interactions observed in natural systems, allowing for further refinements of the models.

Acknowledgments

We thank Matt Osmond and two anonymous reviewers for their many helpful suggestions that improved this manuscript. This project was supported by a fellowship from the University of British Columbia to A.M., a National Science Foundation grant to S.L.N. (DEB 1450653), and a Natural Sciences and Engineering Research Council of Canada grant to S.P.O. (RGPIN-2016-03711).

Literature Cited

Anagnostakis, S., 2000 Revitalization of the majestic chestnut: chestnut blight disease. APSnet Features. Online. DOI: 10.1094/APSnetFeature-2000-1200 .

Anderson, P. K., A. A. Cunningham, N. G. Patel, F. J. Morales, P. R. Epstein *et al.*, 2004 Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. Trends Ecol. Evol. 19: 535–544.

Ashby, B., and M. Boots, 2017 Multi-mode fluctuating selection in host-parasite coevolution. Ecol. Lett. 20: 357–365.

Aulchenko, Y. S., S. Ripke, A. Isaacs, and C. M. van Duijn, 2007 GenABEL: an R library for genome-wide association analysis. Bioinformatics 23: 1294–1296.

Bartholomew, J. L., 1998 Host resistance to infection by the myxosporean parasite *Ceratomyxa shasta*: a review. J. Aquat. Anim. Health 10: 112–120.

Bayne, C. J., 2009 Successful parasitism of vector snail *Biomphalaria glabrata* by the human blood fluke (trematode) *Schistosoma mansoni*: a 2009 assessment. Mol. Biochem. Parasitol. 165: 8–18.

Bento, G., J. Routtu, P. D. Fields, Y. Bourgeois, L. Du Pasquier *et al.*, 2017 The genetic basis of resistance and matching-allele interactions of a host-parasite system: the *Daphnia magna-Pasteuria ramosa* model. PLoS Genet. 13: e1006596.

Bourgeois, Y., A. C. Roulin, K. Müller, and D. Ebert, 2017 Parasitism drives host genome evolution: insights from the *Pasteuria ramosa-Daphnia magna* system. Evolution. 71: 1106–1113.

Chapman, S. J., and A. V. S. Hill, 2012 Human genetic susceptibility to infectious disease. Nat. Rev. Genet. 13: 175–188.

Dukić, M., D. Berner, M. Roesti, C. R. Haag, and D. Ebert, 2016 A high-density genetic map reveals variation in recombination rate across the genome of *Daphnia magna*. BMC Genet. 17: 137.

Duneau, D., P. Luijckx, F. Ben-Ami, C. Laforsch, and D. Ebert, 2011 Resolving the infection process reveals striking differences in the contribution of environment, genetics and phylogeny to host-parasite interactions. BMC Biol. 9: 11.

Gurung, S., S. Mamidi, J. M. Bonman, M. Xiong, G. Brown-Guedira *et al.*, 2014 Genome-wide association study reveals novel quantitative trait loci associated with resistance to multiple leaf spot diseases of spring wheat. PLoS One 9: e108179.

Irvine, B., P. L. Yap, J. Kolberg, S.-W. Chan, T.-A. Cha *et al.*, 1993 Classification of hepatitis C virus into six major genotypes and a series of subtypes by phylogenetic analysis of the NS-5 region. J. Gen. Virol. 74: 2391–2399.

Johnson, N. P. A. S., and J. Mueller, 2002 Updating the accounts: global mortality of the 1918–1920 "Spanish" influenza pandemic. Bull. Hist. Med. 76: 105–115.

Khor, C. C., and M. L. Hibberd, 2012 Host-pathogen interactions revealed by human genome-wide surveys. Trends Genet. 28: 233–243.

Kopp, M., and S. Gavrilets, 2006 Multilocus genetics and the coevolution of quantitative traits. Evolution 60: 1321–1336.

Lambrechts, L., 2010 Dissecting the genetic architecture of host-pathogen specificity. PLoS Pathog. 6: e1001019.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff *et al.*, 2009 Finding the missing heritability of complex diseases. Nature 461: 747–753.

Mitta, G., C. M. Adema, B. Gourbal, E. S. Loker, and A. Theron, 2012 Compatibility polymorphism in snail/schistosome interactions: from field to theory to molecular mechanisms. Dev. Comp. Immunol. 37: 1–8.

Mon, Y., A.-C. Ribou, C. Cosseau, D. Duval, A. Thron *et al.*, 2011 An example of molecular co-evolution: reactive oxygen species (ROS) and ROS scavenger levels in *Schistosoma mansoni/Biomphalaria glabrata* interactions. Int. J. Parasitol. 41: 721–730.

Murphy, D. G., E. Sablon, J. Chamberland, E. Fournier, R. Dandavino *et al.*, 2015 Hepatitis C virus genotype 7, a new genotype originating from central Africa. J. Clin. Microbiol. 53: 967–972.

Newport, M. J., and C. Finan, 2011 Genome-wide association studies and susceptibility to infectious diseases. Brief. Funct. Genomics 10: 98–107.

Nuismer, S. L., B. J. Ridenhour, and B. P. Oswald, 2007 Antagonistic coevolution mediated by phenotypic differences between quantitative traits. Evolution 61: 1823–1834.

Ratcliffe, F. N., K. Myers, B. V. Fennessy, and J. H. Calaby, 1952 Myxomatosis in Australia: a step towards the biological control of the rabbit. Nature 170: 7–11.

Revers, F., and V. Nicaise, 2014 Plant resistance to infection by viruses, in *Encyclopedia of Life Sciences*. John Wiley & Sons, Ltd, Chichester, UK.

Rowell, J. L., N. F. Dowling, W. Yu, A. Yesupriya, L. Zhang *et al.*, 2012 Trends in population-based studies of human genetics in infectious diseases. PLoS One 7: e25431.

Taubenberger, J. K., and D. M. Morens, 2006 1918 influenza: the mother of all pandemics. Emerg. Infect. Dis. 12: 15–22.

Thomas, D., 2010 Gene-environment-wide association studies: emerging approaches. Nat. Rev. Genet. 11: 259–272.

Wang, M., J. Yan, J. Zhao, W. Song, X. Zhang *et al.*, 2012 Genome-wide association study (GWAS) of resistance to head smut in maize. Plant Sci. 196: 125–131.

Weatherall, D. J., and J. B. Clegg, 2002 Genetic variability in response to infection: malaria and after. Genes Immun. 3: 331–337.

Winham, S. J., and J. M. Biernacka, 2013 Gene-environment interactions in genome-wide association studies: current approaches and new directions. J. Child Psychol. Psychiatry 54: 1120–1134.

Zila, C. T., L. F. Samayoa, R. Santiago, A. Butrón, and J. B. Holland, 2013 A genome-wide association study reveals genes associated with fusarium ear rot resistance in a maize core diversity panel. Genetics 3: 2095–2104.

Communicating editor: W. Stephan