Estimating genetic and non-genetic effects for host susceptibility, infectivity and recoverability using temporal epidemic data

Christopher M. Pooley[1,2§], Stephen C. Bishop[1,3], Andrea B. Doeschl-Wilson[1*], and Glenn Marion[2]*

[1] The Roslin Institute, The University of Edinburgh, Midlothian, EH25 9RG, UK.

[2] Biomathematics and Statistics Scotland, James Clerk Maxwell Building, The King's Buildings, Peter Guthrie Tait Road, Edinburgh, EH9 3FD, UK

[3] Deceased

[§] Corresponding author

*Joint senior authors

Abstract

Hosts differ widely in their response to infection and therefore also in their relative contribution to the spread of infection within and across populations. Three key epidemiological host traits affect infectious disease spread: susceptibility (propensity to acquire infection), infectivity (propensity to transmit infection to others, once infected) and recoverability (propensity to recover quickly). Disease control strategies aimed at reducing disease spread may target improvement in any one of these traits. In this paper we introduce a novel software tool called SIRE (standing for "Susceptibility, Infectivity and Recoverability Estimation"), which allows, for the first time, simultaneous estimation of the genetic effect of a single nucleotide polymorphism (SNP), as well as of environmental and specific non-genetic influences on these three host traits. SIRE implements a Bayesian algorithm which makes use of temporal data (consisting of any combination of recorded individual infection times, recovery times or disease status measurements) from multiple epidemics whose dynamics can be represented by the susceptible-infectious-recovered (SIR) model. Validation of SIRE was achieved through simulation studies. Different data scenarios representing realistic recording schemes were simulated to evaluate the impact on the precision of parameter estimates. This analysis revealed that, for the majority of scenarios, SNP effects associated with recoverability can be estimated with highest precision and accuracy, followed by susceptibility and finally infectivity. In the latter case it was found that many epidemics with few individuals give substantially more statistical power to identify SNP effects than the reverse. Furthermore, precise estimates of SNP effects could be obtained even when only recovery times of individuals are known, albeit requiring around four times as many individuals to give equivalent precision. SIRE represents a new tool for analysing a wide range of experimental and field disease data with the aim of discovering and validating SNPs and other factors controlling infectious disease transmission.

1 Introduction

In the era of rapid expansion of the human population with increasing demands on food security, effective solutions that reduce the incidence and impact of infectious diseases not only in humans, but also in plants and livestock, are urgently needed. Emergence of anti-microbial resistance [1, 2] and escape mutants to viral vaccines [3, 4] demonstrate that infectious diseases cannot be combatted by pharmaceutical interventions alone.

Genetic improvement of host disease resistance to infections has long been considered a viable and long-lasting alternative to conventional disease control in livestock and plant breeding [5, 6]. Indeed, a vast number of genetic studies conducted over the last decades indicate that host genetic variation in response to infection is ubiquitous in most host and pathogen species. Significant breakthroughs in genetic disease control have been particularly expected with the advent of affordable high density single-nucleotide polymorphism (SNP) chip panels, as these may allow identification of individuals with high genetic risk of becoming infected or transmitting infections purely based on their genetic make-up, without the need of being exposed to infectious pathogens [7]. To date, however, these high expectations stand in stark contrast to the relatively limited practical application of genomic tools in infectious disease control in plants and livestock. Generally, the host genetic basis underlying infectious disease transmission is still poorly understood.

Modelling disease transmission in genetically heterogeneous populations is well established (see *e.g.*[8, 9]). Particularly relevant are so-called compartmental models in which individuals are classified as, for example, susceptible to infection, infected and infectious, or recovered (or alternatively dead). Transitions between these states are determined by three key individual traits: susceptibility (which governs the probability a susceptible individual becomes infected as a result of exposure to a pathogen), infectivity (the propensity with which infected individuals, once infected, pass on their infection to others) and recoverability (the rate at which infected individuals recover/die) [10, 11]. As demonstrated by numerous simulation studies, host genetic variation in any one of these traits can be harnessed to reduce infectious disease risk and prevalence through selection [11-14]. However, to date most breeding programmes in plants and livestock only target reduction in host susceptibility [14]. The opportunity of reducing disease spread through genetic selection for low infectivity or recoverability, although widely recognized in the scientific literature [15], has rarely been exploited in practice.

Despite their strong epidemiological importance, the genetic regulation and co-regulation of these three host traits is largely unexplored. Whereas a plethora of studies have identified substantial heritable variation and SNPs associated with host susceptibility [14], remarkably little is known about the genetic regulation of host recoverability and infectivity, despite emerging evidence that genetic variation in these traits exists [16, 17]. In particular, it is currently not known to what extent infectivity is genetically controlled, despite compelling evidence that super-spreaders, defined as a small proportion of individuals responsible for a disproportionally large number of transmissions, are a common phenomenon in epidemics [18-20]. This shortcoming is largely because statistical methods for estimating genetic and also non-genetic (treatment) effects for all three key

epidemiological traits controlling disease transmission from infectious disease data are currently lacking.

In conventional genome wide association studies (GWAS) [21] target traits for genetic improvement are measured directly and so establishing genetic associations is relatively straightforward. In the epidemiological setting, however, the susceptibility, infectivity and recoverability of individuals are not measured directly. Rather their effect is manifested in the infection and recovery times of individuals in the epidemic (or epidemics) as a whole. Furthermore, most conventional GWAS assume that an individual's infection status is controlled by its own genetic susceptibility and environmental effects. From an epidemiological viewpoint however, an individual's disease phenotype (*e.g.* infected or not) is controlled by several different underlying host traits, such as its own susceptibility and recoverability, as well as the infectivity of infected group members. Hence, whether and when an individual becomes infected may not only depend on its own susceptibility genes, but also on the infectiousness of other individuals in the same contact group, *i.e.* their infectivity and recoverability genes [22]. This complex interdependence between underlying and observable traits poses challenges for existing GWA methods.

The motivation behind this paper is to introduce new computational models that utilise this temporal information and trait interdependence to estimate, for the first time, genetic effects for all three underlying epidemiological host traits. This requires combining epidemiological and genetic modelling principles. Analysis of incomplete epidemic data to draw inferences on epidemiological parameters is well established [23, 24]. However, analysing such data to draw joint inferences on both the disease epidemiology and host genetic variation has proved challenging [25]. Recent studies have expanded conventional quantitative genetics threshold models to enable joint genetic evaluation of cattle susceptibility to and recoverability from mastitis [26, 27], which led to the identification of novel SNPs and candidate genes associated with either trait [16]. However, because infectivity acts on group members rather than the focal individual itself, applying these technique to estimate genetic effects for infectivity is problematic. Anacleto *et al.* [28] developed a Bayesian inference approach to produce genetic risk estimates for host susceptibility and infectivity from epidemic time to infection data, assuming that susceptibility and infectivity are under polygenic control[1]. This approach however does not incorporate genetic variation in recoverability, and does not estimate SNP effects. An alternative approach, based on the assumption that susceptibility and infectivity are controlled by two single bi-allelic genetic loci [29, 30], used a generalized linear model (GLM) to estimate the relative allelic effects on host susceptibility and infectivity. Whilst an important contribution, this approach focused on the disease status of individuals at the end of each epidemic (*i.e.* discarding potentially useful information from the infection and recovery times themselves), also did not incorporate variation in recoverability, and relied on a number of simplifying assumptions which were found to produce biased estimates under certain circumstances. A variant of this approach [31] used a GLM to analyse time-series data on individual disease status, but was again subject to the above restriction and bias as a consequence of approximations made.

In this study we present a novel software tool called SIRE (standing for "susceptibility, infectivity and recoverability estimation") that implements a Bayesian inference approach to simultaneously estimate the effects of a single SNP, together with that of other fixed effects (such as *e.g.* sex, breed

[1] *I.e.* they are determined by a large number of genes, each with small effect.

or vaccination status) on host susceptibility, infectivity and recoverability from temporal epidemic data. This approach can be applied to a wide range of epidemic data, collected at the level of individuals, and accounts for different types of uncertainty in a statistically consistent way (*e.g.* censoring of data[2]), and permits the incorporation of prior knowledge. We validate SIRE for a variety of simulated epidemic scenarios, comprising not only the ideal case in which infection and recovery / death times of each individual are known exactly, but also under more realistic scenarios in which epidemics are only partially observed.

2 Material and methods

2.1 Data structure and the underlying genetic-epidemiological model

SIRE applies to individual-level disease data originating from one or more contact groups in which infectious disease is transmitted from infectious to susceptible individuals through effective contact[3]. This data can come from well controlled disease transmission experiments or from much less well controlled field data (which may be less complete, but readily available in larger quantity).

In the context of disease transmission experiments, epidemics are initiated by means of artificially infecting a proportion of "seeder" individuals which go on to transmit their infection to susceptible individuals sharing the same contact group. In field data contact groups may consist of animal herds, or any contemporary group of individuals sharing the same environment such as pasture, pen, cage or pond, and infection is assumed to invade the group by some external, usually unknown, means (*e.g.* by the unintentional spread of infected material, or the introduction of an infected individual from elsewhere). For simplicity it is assumed that throughout the observation period groups are considered closed, *i.e.* no births, migrations, or transmission of disease between groups. This assumption generally holds for experimental studies and also for the common field situations, where a movement ban is imposed after disease notification.

The dynamic spread of disease within a contact group is modelled using a so-called SIR model, as illustrated in Fig. 1(a) [32]. Individuals are classified as being either susceptible to infection (S), infected and infectious (I), or recovered/removed/dead (R). Under the simple SIR model for homogeneous populations, the time-dependent force of infection for a susceptible individual *j* (*i.e.* the probability per unit time of becoming infected) is given by $\lambda_j(\mathrm{t}) = \beta I(\mathrm{t})$, which is the product of an average transmission rate β and the number of infected individuals at time *t*, $I(\mathrm{t})$. To incorporate individual-based variation in host susceptibility and infectivity, this simple expression for $\lambda_j(\mathrm{t})$ is replaced by an individual force of infection (see [28] for a formal derivation)

$$\lambda_j(t) = \beta e^{G_z} e^{g_j} \sum_i e^{f_i}. \tag{1}$$

Here g_j characterises the fractional deviation[4] in individual *j's* susceptibility as compared to that of the population as a whole[5], f_i characterises the corresponding quantity for individual *i's* infectivity, and the sum in Eq.(1) goes over all individuals infected at time *t* sharing the same contact group *z* as

[2] Situations in which the observation period only partially covers the duration of the epidemics.
[3] *I.e.* contact with infectious material shed by an infected individual that results in infection.
[4] *E.g.* g_j=0.1 corresponds to individual *j* being ≃10% more susceptible than the population average.
[5] By definition the average of g_j over the population is zero.

individual j (note, this sum varies as a function of time t). Whilst other link functions can be used, the exponential dependencies in Eq.(1) ensure that λ_j is strictly positive and allow for the possibility that some individuals are much more/less susceptible/infections than others[6].

The term G_z in Eq.(1) accounts for the fractional deviation in disease transmission for group z. This is used to incorporate group-specific factors that influence the overall speed of an epidemic in one contact group relative to another (*e.g.* animals kept in different management conditions, environmental differences, or variation in pathogen strains with different virulence). Whilst variation in G_z may be small for a well-controlled challenge experiment, this may not be the case in real field data. G_z is assumed to be a random effect with standard deviation σ_G.

Whilst in Eq. (1) infection is modelled as a Poisson process with individual infection rates λ_j [18, 20], the recovery process is modelled by assuming that the time taken for individual m to recover after being infected is drawn from a gamma distribution with an individual-based mean w_m and shape parameter k (which for simplicity is assumed to be the same across individuals). This mean recovery time is expressed as

$$w_m = (\gamma e^{r_m})^{-1}, \tag{2}$$

where γ represents the average recovery rate and r_m describes the fractional deviation from this for individual m. This approach is taken to allow the recovery probability distribution to adopt a more biological realistic profile, rather than the exponential distribution usually imposed on it (see electronic supplementary material Appendix A for further details).

The individual-based deviations in susceptibility $\boldsymbol{g}$, infectivity $\boldsymbol{f}$ and recoverability $\boldsymbol{r}$ (which are vectors with elements relating to each individual) are decomposed into the following contributions

$$\begin{aligned} \boldsymbol{g} &= \boldsymbol{g}^{\mathrm{SNP}} + \boldsymbol{X}\boldsymbol{b}_g + \boldsymbol{\varepsilon}_g, \\ \boldsymbol{f} &= \boldsymbol{f}^{\mathrm{SNP}} + \boldsymbol{X}\boldsymbol{b}_f + \boldsymbol{\varepsilon}_f, \\ \boldsymbol{r} &= \boldsymbol{r}^{\mathrm{SNP}} + \boldsymbol{X}\boldsymbol{b}_r + \boldsymbol{\varepsilon}_r. \end{aligned} \tag{3}$$

SNP contribution –Assuming a diploid genomic architecture with biallelic SNP implies three genetic values: *AA, AB* and *BB*. The SNP contribution to the traits for individual j depends on j's genotype in the following way:

$$g_j^{SNP} = \begin{matrix} a_g \\ a_g\Delta_g \\ -a_g \end{matrix}, \quad f_j^{SNP} = \begin{matrix} a_f \\ a_f\Delta_f \\ -a_f \end{matrix}, \quad r_j^{SNP} = \left.\begin{matrix} a_r \\ a_r\Delta_r \\ -a_r \end{matrix}\right\} \begin{matrix} \text{if } j \text{ is } AA \\ \text{if } j \text{ is } AB \\ \text{if } j \text{ is } BB \end{matrix} \tag{4}$$

The parameters a_g, a_f and a_r capture the relative differences in trait values between *AA* and *BB* individuals, and are subsequently referred to as the "SNP effects" for susceptibility[7], infectivity and recoverability, respectively. The scaled dominance factors Δ_g, Δ_f and Δ_r characterise the trait deviations between the heterozygote *AB* individuals and the homozygote mean (a value of 1

[6] Which, in particular, concurs with the empirical observation of super-spreaders.
[7] This means that if a_g is positive, individuals with an *AA* genotype will be more susceptible to disease than those with a *BB* genotype.

corresponds to complete dominance of the *A* allele over the *B* allele and -1 when the reverse is true).

Fixed effects – The design matrix $\boldsymbol{X}$ and fixed effects $\boldsymbol{b}_g$, $\boldsymbol{b}_f$ and $\boldsymbol{b}_r$ in Eq.(3) allow for other known sources of variation to be accounted for (*e.g.* breed, sex or vaccination status)[8].

Residual contributions – Here $\boldsymbol{\varepsilon}=(\boldsymbol{\varepsilon}_g, \boldsymbol{\varepsilon}_f, \boldsymbol{\varepsilon}_r)$ account for all genetic effects not captured by the SNP in consideration, as well as any non-genetic environmental variation. We assume that for each individual in $\boldsymbol{\varepsilon}$ the three trait values are drawn from a multivariate normal distribution with zero mean and 3×3 covariance matrix $\boldsymbol{\Sigma}$. Including these correlations is important because it allows for the possibility that, for example, more susceptible individuals may also, on average, be more infectious and recover at a slower rate. Note that in this study, which focuses on the estimation of SNP effects, there is no explicit distinction between random genetic and environmental effects, although the model could be extended to incorporate estimation of polygenic effects. It is thus assumed that individuals are randomly distributed across the groups with respect to the genetic effects on the epidemiological traits not captured by the SNP. Also note that Eq(3) does not contain random group effects for the individual epidemiological traits. This is because the group effect has already been incorporated in the expression of the individual force of infection in Eq(1). In other words, it is assumed that the group environment more likely affects the speed at which infection spreads within a group rather than on individual's susceptibility, infectivity or recoverability.

2.2 Bayesian inference

Based on the description above, the model contains the following set of parameters: $\theta=(\beta, \gamma, k, a_g, a_f, a_r, \Delta_g, \Delta_f, \Delta_r, \boldsymbol{b}_g, \boldsymbol{b}_f, \boldsymbol{b}_r, \boldsymbol{\varepsilon}, \boldsymbol{\Sigma}, \boldsymbol{G}, \sigma_G)$. We denote the complete set of infection and recovery event times for all individuals as ξ over the observed duration of the epidemics[9]. Typically ξ is not precisely known, and so we consider the general case in which ξ represents a set of latent model variables. The nature of the actual observed data y will be problem dependant. For example, in some instances recovery or removal (*e.g.* due to death) times will be precisely known but infection times completely unknown. In other instances infection and recovery times will both be unknown, but results from disease diagnostic tests provide information regarding disease status at particular points in time. The framework presented in this paper is flexible to these various possibilities.

Application of Bayes' theorem implies that the posterior probability distribution for model parameters and latent variables is given by

$$\pi(\theta,\xi \mid y) \propto \pi\left(y \mid \xi\right) L(\xi \mid \theta)\pi(\theta). \tag{5}$$

These contributions are described as follows:

[8] Following convention an additional fixed effect (with corresponding column in $\boldsymbol{X}$ set to one) is added to account for trait mean. The values for this fixed effect for the three traits are explicitly chosen to ensure the population averages of $\boldsymbol{g}$, $\boldsymbol{f}$ and $\boldsymbol{r}$ are zero (remembering that the average effects are already captured by the parameters β and γ).

[9] For computational convenience the recovery event times even after the observation period ends are also included within ξ.

Observation model $\pi(y|\xi)$ – the probability of the data given a set of event times ξ. In the context of this paper this simply takes the values one or zero depending on whether ξ is consistent with y or not. For example a disease diagnostic test showing that an individual is infected is only consistent with ξ containing an infection event on that individual *prior* to the time of the test and a recovery event *after* the time of the test[10]. Similarly, if data y indicates that an individual becomes infected at a particular point in time, this is only consistent provided ξ also contains this infection event. In summary, the observation model constrains the possible event sequences coming from the model, and this, in turn, informs the model parameters.

Latent process likelihood $L(\xi|\theta)$ – the probability of ξ being sampled from the model given parameters θ. This can be derived from the genetic-epidemiological model described in the previous section [23, 24] (see Appendix B for a details), and is given by

$$L(\xi \mid \theta) = \prod_z \left[\left(\prod_{j \in z} \lambda_j \right) \left(\prod_{e \in E_z} e^{-\Lambda_z(t_e) \times (t_e - t_{e-1})} \right) \times \left(\prod_{m \in z} F_\Gamma(\delta t_m \mid w_m, k) \right) \right]. \tag{6}$$

The functional dependence of $L(\xi|\theta)$ on the parameters θ is expressed in terms of the force of infections λ_j in Eq.(1) and mean recovery times w_m in Eq.(2), which themselves depend in $\boldsymbol{g}$, $\boldsymbol{f}$ and $\boldsymbol{r}$ in Eq.(3). The product z goes over all contact groups and within each contact group: j goes over individuals that become infected *excluding* those which initiate epidemics[11], m goes over individuals that become infected *including* those which initiate epidemics and e goes over both infection and recovery events (with corresponding event times t_e). Here the notation $j \in z$ indicates that j goes over all those individuals j in contact group z, and $e \in E_z$ indicates that e goes over all events E_z. The force of infection λ_j is given by Eq.(1) immediately prior to individual j becoming infected. The gamma distributed probability density function F_Γ for recovery events gives the probability an individual is infected for duration δt_m given a mean duration w_m and shape parameter k. The time dependent total rate of infection events Λ_z in contact group z immediately prior to event time t_e is given by

$$\Lambda_z(t_e) = \sum_s \lambda_s, \tag{7}$$

where the sum in s goes over all susceptible individuals in group z at that time.

An important point to mention is that Eq.(6) is calculated on an unbounded time line. In situations in which data is censored, the observation model restricts events that occur within the observed time window, but other events can exist outside of this observed region[12].

Prior $\pi(\theta)$ – the state of knowledge prior to data y being considered. To account for the prior assumption that residuals $\boldsymbol{\varepsilon}$ in Eq.(3) are multivariate normally distributed and that the vector of group effects $\boldsymbol{G}$ in Eq.(1) are random effects, $\pi(\theta)$ can be decomposed into

$$\pi(\theta) = \pi(\theta_{-\varepsilon,\boldsymbol{G}}) \pi(\boldsymbol{\varepsilon} \mid \boldsymbol{\Sigma}) \pi(\boldsymbol{G} \mid \sigma_G), \tag{8}$$

[10] Note, this approach can readily be extended to include imperfect diagnostic tests, in which case the observation probability will be less than one and depend on the sensitivity and specificity of the test.

[11] In the case of disease transmission experiments this would exclude individuals that seed the infection and in field data it would exclude the first (usually unknown) infected individual(s) within each contact group.

[12] These include infection and recovery events before observations start and recoveries after observations end. To improve computational efficiency infection events after observations end are not included as these have no impact on the posterior.

where $\theta_{-\varepsilon,G}$ includes all parameters with the exception of $\boldsymbol{\varepsilon}$ and $\boldsymbol{G}$ and

$$\pi(\boldsymbol{\varepsilon}\,|\,\Sigma) = \prod_j \frac{1}{\sqrt{2\pi|\Sigma|}} e^{-\frac{1}{2}\boldsymbol{\varepsilon}_j^T \Sigma^{-1} \boldsymbol{\varepsilon}_j},$$
$$\pi(\boldsymbol{G}\,|\,\sigma_G) = \prod_z \frac{1}{\sqrt{2\pi}\sigma_G} e^{-\frac{1}{2\sigma_G^2} G_z^2}. \tag{9}$$

Here j goes over each individual and $\boldsymbol{\varepsilon}_j = (\varepsilon_{g,j}, \varepsilon_{f,j}, \varepsilon_{r,j})^T$ is a three dimensional vector giving the residual contributions to the susceptibility, infectivity and recoverability of j. $\boldsymbol{\Sigma}$ is a 3×3 covariance matrix (which describes not only the overall magnitude of the residual contributions, but also any potential correlations between traits). Finally, the product z in Eq.(9) goes over all contact groups and G_z represents the group-based fractional deviation in transmission rate which is assumed to be normally distributed with standard deviation σ_G.

The default prior for $\theta_{-\varepsilon,G}$ (which can be modified if necessary) is largely uninformative but does place upper and lower bounds on many of the key parameters to stop them straying into biologically unrealistic values (details are given in Appendix C).

Samples for θ and ξ from the posterior are generated by means of an adaptive Markov Chain Monte Carlo (MCMC) schemes which implements optimised random walk Metropolis-Hastings updates for most parameters and posterior-based proposals [33] to aid fast mixing of the residual parameters (details are given in Appendix D).

2.3 SIRE

SIRE is a desktop application that implements the Bayesian algorithm outlined in this section. It is freely available for download in the supplementary material or at www.mkodb.roslin.ed.ac.uk/EAT/SIRE.html (for Windows, Linux and Mac). An easy to used point and click interface is used to import data tables in a variety of formats and graphically output results. It utilises efficient C++ code and allows for parallel running on multiple cores. SIRE takes as input any combination of information about infection times, recovery times, disease diagnostic test results, genotype of SNP or any other fixed effect (Fig. 2(a)), details of which individuals belong to which contact groups, and any prior specifications (Fig. 2(b)). The output from SIRE consists of trace plots for parameters, posterior distributions (Fig. 2(c)), posterior plots for ξ, dynamic population estimates and summary statistics (means and 95% credible intervals for model parameters θ) as well as MCMC diagnostic statistics (Fig. 2(d)). SIRE allows for exporting graphs and files containing samples of θ and ξ for further analysis using other tools. The user guide for SIRE is available in the electronic supplementary material.

3 Assessment of performance and data requirements

In this section we apply SIRE to simulated datasets in order to 1) test the extent to which the inferred posterior parameter distributions agree with their true values, and 2) investigate how the precision of inferred model parameters changes under different data scenarios.

3.1 Example simulation and inference

For this purpose, we begin by investigating a representative set of arbitrary parameters with a large SNP effect, and later go on to look at how results change when things are changed. We assumed that the SNP under investigation is in Hardy-Weinberg equilibrium[13] with an *A* allele frequency of p=0.3. Individuals were randomly assigned into N_{group} different contact groups, with each group assumed to contain G_{size} individuals. For the SNP effects, we used the values a_g=0.4, a_f=0.3 , a_r=-0.4, which represents a case in which the SNP has a relatively large pleiotropic effect (in this example the SNP confers higher susceptibility for *AA* compared to *BB* individuals, as well as slightly higher infectivity and lower recoverability). The choice of Δ_g=0.4, Δ_f=0.1, Δ_r=-0.3 for the scaled dominance factors represents partial, but not strong, dominance of either the *A* or *B* allele. For simplicity we assumed a single fixed effect, *e.g.* sex, of arbitrary moderate size b_{g0}=0.2, b_{f0}=0.3, b_{r0}=-0.2 with individuals in the population randomly selected to be male or female. The residual variances were chosen to be $\Sigma_{gg}=\Sigma_{ff}=\Sigma_{rr}=1$, corresponding to a large variation in traits between individuals (perhaps too large, but here we want to show that inference of the SNP effects is still possible *despite* significant variation in trait values arising from other sources). In line with the direction of the SNP effects, the covariances were chosen to be Σ_{gf}=0.3, Σ_{gr}=-0.4 and Σ_{fr}=-0.2, representing a potential scenario in which more susceptible individuals are also more infectious and recover at a slower rate and *vice-versa*). To accommodate variation in epidemic speed across groups we set the standard deviation in the group effect to σ_G=0.5. Finally, the average contact and recovery rates and were chosen to be $\beta=0.3/G_{size}$ (chosen such that the basic reproductive ratio remain constant when G_{size} is changed) and $\gamma=0.1$ with shape parameter k=5.

Simulated epidemic data were generated by means of a Doob-Gillespie algorithm [34] modified to account for non-Markovian recovery times (details of this procedure are given in Appendix F). A typical output for one simulated epidemic in a single contact group N_{group}=1 with G_{size}=50 individuals is shown in Figure 1(b). Whilst the simulation itself is generated on an individual basis, this graph summarises dynamic variation in the susceptible, infectious and recovered populations, categorised by SNP genotype. The graph reveals the classic epidemic SIR model behaviour, where a single infected individual passes its infection on to others, triggering a rapidly spreading infection process throughout the population until the epidemic eventually dies out as a result of the susceptible population becoming largely exhausted and the remaining infected population recovering. Note that in closed groups not all susceptible individuals become infected. In this particular case some *AB* and *BB* individuals remain uninfected at the end of the epidemic. The absence of *AA* individuals partly stems from natural stochasticity in the system, but also partly from the fact that a_g=0.4 is positive, *i.e. AA* individuals are more susceptible to disease and so on average are less likely to remain uninfected. Consequently we can make a direct link between the genetic composition in the final state of the epidemic and the expected value for a_g (which in this example is more likely positive than negative). Over and above information from the final state, however, there is much to be gained from also accounting for the infection and recovery events themselves. The Bayesian approach adopted in this paper utilises all this information to extract the best available parameter estimates.

To illustrate what a typical output from SIRE looks like, Fig. 3 shows inferred posterior probability distributions obtained from SIRE when it analysed a single simulated dataset consisting of known

[13] This assumes that the expected fraction of individuals in the three genotypes is given by $\{p^2, 2p(1-p), (1-p)^2\}$.

infection and recovery times (realistic situation when these are not know precisely are discuss later in section 3.5) for N_{group}=20 contact groups each containing G_{size}=50 individuals using the parameter set from Fig. 1. The standard deviations (SDs) in these distributions characterise the precision with which parameters can be inferred and the vertical black lines denote the actual parameter values used in the simulation. Whilst these lines consistently lie within regions of high posterior probability, there is a notable variation in precision across parameters.

Whilst the results in Fig. 3 may appear somewhat specific to an arbitrary selection of parameters, the results are actually far more generic. As discussed later, the standard deviations in these distributions are largely independent of the parameter values themselves. This means inspecting the relative SDs sheds light of which parameters in the model can be estimated with a greater degree of precision, and which are not.

From the point of view of this paper the critical distributions are those that determine the SNP effects themselves, as shown in Fig. 3(d-i). In Fig. 3(f) we find that a_r is highly peaked around its true value of -0.4 (SD of 0.075). Importantly this distribution has an extremely low posterior probability at a_r=0. Indeed, since a_r=0 does not lie within the 95% credible interval it can be concluded, to a high degree of certainty, that the SNP *is* associated with recoverability. The same is true for the susceptibility SNP effect a_g in Fig. 3(d), albeit with a wider posterior probability distribution (SD of 0.14). This is for two reasons: firstly the recovery process involves only a_r, whereas the infection process involves both a_g and a_f (leading to potential confounding between these parameters) and secondly the recovery processes is gamma distributed which has a smaller standard deviation than the wide exponentially distributed Poisson process governing infection.

The infectivity SNP effect a_f in Fig. 3(e) exhibits a much wider probability distribution (SD of 0.47) than the other two SNP effects. The fact that zero does lie within the 95% posterior credible interval[14] means that no certain association with infectivity can be inferred from exact infection data from 20 epidemics consisting of 50 individuals each. Figure 3(d-f) illustrates a general principle: SNPs effects associated with recoverability are most precisely estimated, followed by susceptibility, and finally infectivity[15].

Figures 3(g)-(i) show posterior estimates for the scaled dominance parameters. We find that the precision of these is relatively poor compare to the SNP effects themselves and actually reduces as the size of the SNP effects goes down[16].

3.2 Different parameter values

Section 3.1 showed a illustrative example for a particular parameter set. Here we assess prediction accuracy for all the parameters in the model. This is achieved by means of taking a "base" set of parameters

[14] Which goes from -0.35 to 2.1.
[15] The only exception to this is when recovery times are unknown recoverability SNP effects become imprecise.
[16] Which makes sense in the limit of zero SNP effect size, because here no information about dominance is available.

$$\begin{aligned} &\beta = 0.3/N,\ \gamma = 0.1,\ k = 5 \\ &a_g = a_f = a_r = 0, \\ &\Delta_g = \Delta_f = \Delta_r = 0, \end{aligned} \qquad \begin{aligned} &b_{g0} = b_{f0} = b_{r0} = 0, \\ &\sigma_G = 0.5 \end{aligned} \qquad \Sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{10}$$

and then making changes to each parameters separately (whilst fixing all others). Figure 4 shows scatter plots in which each cross represents the posterior mean (with the error bars representing 95% posterior credible intervals) inferred from a simulated dataset using a true parameter value taken from the *x*-axis.

Focussing on Fig. 4(d), the fact that the crosses do not systematically deviate from the diagonal dashed line means that the algorithm generates unbiased parameter estimates (a fact that remains true even for incomplete data, as illustrated in Appendix G). Note, the precisions of parameter estimates are found to be largely independent of the values of the parameters themselves (*i.e.* the error bars on the edges of Fig. 4(d) have almost the same magnitude as in the middle of the graph). Figures 4(e) and (f) reveal the same picture for the infectivity and recoverability SNP effects. Taken together, Fig. 4 implies that a comprehensive assessment of both the precision and accuracy of SNP effect estimates obtained with the algorithm can be made by simply measuring posterior SDs in a_g, a_f and a_r when a_g=a_f=a_r=0. Looking at how these SDs depend on contact group structure and measured data now becomes the focus of the remainder of this paper.

3.3 Different contact group number / size

First we look at how SDs change as a function of the number of individuals within each epidemic G_{size} (where N_{group}=10 contact groups are assumed). The results are represented by the crosses in Fig. 5 (here 50 simulated datasets were generated for each G_{size} with the average SD being shown by the cross and the error bars denoting variation across replicates[17]). Inspecting Fig. 5(a) we find that the SD in a_g reduces as the number of individuals in each contact group G_{size} increases. Importantly this relationship scales as a line of slope -½, corresponding to precision increasing by a factor of two as the number of individuals is increases by a factor of four (note the log scales on this plot).

Based on Fig. 5 we now provide a crude calculation to estimate how many individuals need to be observed in order to be able to make an association with a susceptibility SNP effect of a given size. Suppose, as above, that the true value for a_g is 0.4. Given actual data the posterior distribution will have a mean value which is sometimes above 0.4 and sometimes below 0.4 (depending upon replicate). As a typical case let's look at when it is exactly 0.4. Since the posterior is approximately normally distributed (Fig. 3), and 95% of the area of a normal distribution lies within two SDs of its mean, the SD in the posterior distribution would need to be less than 0.2 to ensure that a_g=0 does not lie within the 95% credible interval. Following the black dashed line in Fig. 5(a) we see that this corresponds to G_{size}=20 individuals per contact group, and so $G_{size}\times N_{group}$=200 individuals would be needed in total. Naturally, if a_g is smaller than 0.4, more individuals would be needed for associations to be made, and *vice-versa*.

Figure 5(c) shows the same scaling relationship for identifying recoverability SNP effects, but this time only $G_{size}\times N_{group}$=10 individuals are needed to make associations for recovery SNP effects (reflecting the fact that a_r can be inferred more precisely, as mentioned previously). A very different

[17] Specifically, 95% of the simulated datasets have posterior means that lie within this interval.

state of affairs, however, is observed in Fig 5(b). Here we see that not only is the infectivity SNP effect a_f poorly estimated, but also its precision does not markedly improve even when the number of individuals in each contact group G_{size} is substantially increased.

Instead of varying G_{size} and fixing the number of contact groups N_{group}, we now fix G_{size}=10 and vary N_{group}. Results for this are shown in Fig. 6 (represented by the crosses). This reveals a similar behaviour to before for the SD in a_g and a_r, but crucially we find the SD in the infectivity SNP effect a_f now also scales with the familiar line of slope -½. The reason for this behaviour lies in the fact that infectivity is an indirect genetic effect, i.e. an individual's infectivity SNP affects the disease phenotype of group members rather than its own disease phenotype [35-37]. More intuitively, this can be explained as follows. Susceptibility and recoverability SNPs of an individual directly affect its own measured disease phenotype (the former affecting its infection time and the latter affecting its recovery time). Therefore the precision with which these two quantities can be inferred is expected to scale with the total number of individuals. On the other hand, as an individual's infectivity SNP acts on all susceptible individuals sharing the same contact group, it affects the epidemic dynamics as a whole. In fact much of the information regarding infectivity comes from the overall speed of epidemics. For example, if those contact groups containing individuals with more *A* alleles consistently experience epidemics which are faster than those with fewer *A* alleles, this provides evidence that the *A* allele confers greater infectivity than the *B* allele[18]. Because information about the infectivity SNP effect comes from epidemic wide behaviour, its precision scales linearly with the number of contact groups N_{group} (Fig. 6(b)), but not with the number of individuals per contact group G_{size} (Fig. 5(b)).

Finally, we investigate the case in which we fix the total number of individuals to $G_{size}\times N_{group}$=1000 whilst simultaneously varying G_{size} and N_{group}, as shown in Fig. 7 (see crosses). In Fig. 7(a) we find very little variation in the precision of a_g . However, the infectivity SNP effect in Fig. 7(b) clearly shows that *larger* numbers of contact groups containing *fewer* individuals help to reduce the SD in a_f. In the case of G_{size}=2 the posterior SDs in a_g and a_f are actually the same due to the symmetry of this particular setup (*i.e.* each group consists of exactly one infected and one susceptible individual). Lastly, Fig. 7(c) shows that the SD in a_r is largely independent of G_{size}. This is because recovery is solely an individual-based process, and so happens independently of others sharing the same contact group[19].

3.4 Different allele frequency

So far we have assumed a fixed *A* allele frequency *p*=0.3 in the population. Figure 8 investigates what happens when this is no longer the case by varying *p*, which in turn changes the Hardy-Weinberg equilibrium frequencies for the three genotypes. We find that the curves are symmetric around the minimum of *p*=0.3 and remain remarkably flat over a large region. They only increase substantially when the minor allele frequency drops below around 10%. This result shows that the statistical power to establish SNP effects dramatically reduces when they are rare, which is consistent with observations from conventional GWAS analyses [38]).

[18] The situation is further complicated by the fact that differences in susceptibility can also cause this behaviour. However the algorithm can independently estimate a_g, so removing this potential confounding.
[19] Although in cases in which R_0 is small, differences may result from variation in the fraction of individuals which actually become infected.

3.5 Different data scenarios

Reflecting real-world datasets we consider five data scenarios (DS) for potential observations made on the system outlined below:

DS1: Infection and recovery times for all individuals exactly known (perfect data/best case scenario)

This represents the best case scenario for inferring parameter values[20] and has been assume in all the preceding result. Although this scenario may rarely apply in practice, it still provides useful insights for software validation and application (in fact, as shown later, the inferred parameter estimates under DS1 are very nearly the same as when infection and recovery times are known to only a large degree of uncertainty).

DS2: Only recovery times known

Often "recovery" in compartmental SIR models (Fig. 1) represents the death of individuals. Consequently DS2 is pertinent to cases in which the only measurable quantity is the time at which individuals die. For example, disease challenge experiments in aquaculture routinely record time of death rather than infection times, which are usually difficult to measure [40]. Naïvely it might expect that because infection times are unknown and the SNP effects a_g and a_f relate to the infection process then nothing can be inferred about these quantities. This section, however, clearly demonstrates this not to be the case. The reason lies in the fact that whilst infection times are latent variables, the distribution from which they are sampled *is* informed by the available recovery data and dynamics of the model through the likelihood in Eq.(6).

The square symbols in Fig. 5(a) denote the posterior SDs in the susceptibility SNP effect a_g under DS2. Compared to the best case scenario DS1, SD in a_g increases as a result of having to infer probable infection times for individuals (as opposed to knowing them exactly). Following the crude calculation from above we see that the number of individuals per group now needed to identify an association for a susceptibility SNP effect of a_g =0.4 is now G_{size}=80 (see dashed purple line in Fig. 5(a)), as opposed to G_{size}=20 in the case of DS1. Consequently to achieve an equivalent precision for a_g under DS2 as that under DS1 requires around 4 times as many individuals. In the case of the infectivity SNP effect a_f, this factor becomes approximately 4.2 (see Fig. 6(b), assuming a large number of contact groups), and for the recoverability it is 1.9 (see Fig. 5(c)). These factors are found to be remarkably consistent across a range of group numbers and sizes.

In summary our analysis of DS2 clearly demonstrates that even when infection times are unknown, accurate inference regarding all SNP effects can be made, given sufficient data.

DS3: Only infection times known

Whilst less common than DS2, in some instances data provides information regarding when individuals become infected but not when they recover. For example in human epidemics, patients may go to the doctor when they become ill, but no records will be kept on when they recover.

The triangles in Figs. 5, 6 and 7 show results under DS3. Here the SDs in the SNP effects for susceptibility a_g and infectivity a_f are found to be almost the same as for DS1 (because uncertainty in recovery times only has a very weak impact on uncertainty in the infection process). However the SD

[20] For example, changes in visual or behavioural signs may indicate the onset of disease, and recovery times are given by the time of death [39].

for the recovery SNP effect a_r is much larger, meaning that little can be inferred regarding SNP-based differences in recoverability. This is because under DS3 the only indirect information regarding recovery times comes from the very early stages of each epidemic (*e.g.* we know that the first infected individual cannot recover before the second individual becomes infected). This explains why SDs for recovery SNP effects decrease at a rate of -1/2 (on the log-scale) as the as the number of contact groups N_{group} increases (*i.e.* the triangles in Fig. 6(c) scale with the black line) but not when the number of individuals per contact group G_{size} is changed (see Fig. 5(c)).

DS4: Disease status periodically checked

DS4 represents the most common scenario for monitoring infectious disease spread in livestock or plant populations, where each individual is periodically checked to establish its disease status. Under DS4 the point at which epidemics start is usually unknown, as well as the infection and recovery times of individuals themselves. However the diagnostic test results place constraints on these quantities. For example, if an individual is found to be uninfected at one sampling time and infected at the next sampling time this means that infection must have occurred at some point in time between the two checks (note here we assume perfect diagnostic tests but even imperfect tests provide some information on infection times). Similarly, recovery times are constrained to occur between consecutive checks.

Figure 9 shows results under DS4 assuming a time interval between checks of Δt. When Δt=0 (as shown on the left of this figure) the DS4 results are the same as in DS1 (because here infection and recovery times are effectively exactly known). On the other hand as checking becomes less and less frequent, the SDs in the SNP effects rise. A surprising feature, however, is that this reduction in statistical power is perhaps less than might be expected. The vertical lines in Fig. 9 represent two key timescales: $\langle t_I \rangle$ is the average infection time as measured from the beginning of the epidemic[21]and $\langle t_R \rangle$ is the average recovery time. We see that statistical power only marginally reduces even when disease diagnostic checking is performed on a similar timescale as the epidemics as a whole. The limit on the right hand side of this diagram shows the situation in which there is no information regarding infection and recovery times (*i.e.* only the initial and final states of the epidemic are observed). Unfortunately it was found to be difficult to probe this regime using SIRE due to mixing problems arising in the MCMC algorithm[22] (principally because the number of possible parameter sets and event sequences consistent with a given final outcome is vast).

The results here emphasise the fact that even relatively infrequent disease status checks provide useful data from which accurate inferences regarding SNP effects can be drawn.

DS5: Time censored data

This data scenario relates to situations in which epidemics are not observed over their entire time period. For example a disease transmission experiment being carried out may be terminated early, due to cost or other factors, even though epidemics have not completely died out. In Fig. 10(a) it is assumed the infection and recovery times are exactly known but only up to some final time t_{end} (subsequent to which no further data is available). We find that very little information is lost when restricting t_{end} to around the average recovery time $\langle t_R \rangle$. This is in part because (based on the choices

[21] Which is found from a large number of simulated replicates.
[22] Mixing relates to the number of MCMC iterations needed to generate a set of samples representative of the posterior.

for θ used in the simulation study[23]) most individuals recover before $\langle t_R \rangle$. Given that $\langle t_R \rangle$ is usually substantially less than the total epidemic time, from a practical point of view terminating disease transmission trials prior to the end of the epidemic when no new infections occur, (and perhaps performing further replicates) may be beneficial, although the effectiveness of this would depend on the variation in recoverability in the populations, which *a priori* may be unknown.

Figure 10(b) shows the opposite scenario, in which contact groups are observed from an initial starting time t_{start} after the start of the epidemic up until its termination. This scenario may apply to field outbreaks, where sampling occurs only after notification of the outbreak. Here again we see a reduction in statistical power with increasing t_{start}, but this reduction is not substantial until around the average infection time. This result is surprising, but it turns out that whilst none of the events before t_{start} are actually measured (which may include a large proportion of total number of infection events), the disease status of all the individuals at t_{start} can be accurately inferred [24] and this encapsulates almost the same amount of information as when the event times are precisely known.

General data scenario

It should be noted that the data scenarios DS1-5 considered above are not comprehensive, and any combination of infection time, recovery time and state data can be used as inputs into SIRE. Furthermore SIRE accounts for additional uncertainties in cases in which data is missing on some individuals

4 Discussion

The availability of dense genome wide SNP panels has revolutionized human medicine and has paved the way for genetic disease control in agriculture. With declining genotyping costs, discovery of new disease susceptibility loci has increased exponentially over the recent years, and evidence for their effective utilization in personalized medicine and livestock and plant breeding programmes continues to emerge. However, there is increasing awareness amongst researchers and policy makers that disease susceptibility is not the only host genetic trait controlling disease incidence and prevalence in populations, and in particular that host genetic infectivity and recoverability may also constitute important improvement targets for reducing disease spread [15, 16, 41, 42]. Yet, genetic loci associated with host recoverability reported in the literature are sparse, and to the best of our knowledge no infectivity SNP has yet been identified to date. This is not surprising given that phenotypic measurements of recoverability and infectivity, such as individuals' recovery or pathogen shedding rates are rarely available in practice and statistical inference methods to accurately infer these from available epidemic data are still in their infancy. In line with the lack of suitable statistical methods, little is known about what type and number of measurements are needed to produce unbiased and precise estimates of SNP effects for these 'new' epidemiological host traits.

In this paper we developed a Bayesian methodology to allow, for the first time, simultaneous estimation of SNP effects for host susceptibility, recovery and infectivity from temporal epidemic

[23] This is a consequence of a small number of individuals having very low recoverability, which arises because of the large assumed residual variance Σ_{rr}=1. These individual take a very long time to recover and so significantly increase $\langle t_R \rangle$ compared to what it would otherwise be.

[24] Because the final state is known and all the subsequent events from t_{start} are also known, the state at t_{start} is exactly specified.

data. The methodology was validated with data from simulated epidemics, which were also used to assess how different data scenarios representing different recording schemes in field or experimental studies may affect the estimates of SNP effects and other parameters influencing the transmission dynamics. The relatively complex Bayesian algorithm outlined in this paper has been implemented into a user-friendly software called SIRE, which allows analyses to be performed by anyone with relevant epidemiological data (as shown in Appendix H, outputs typically takes only a few minutes of CPU time per 1000 individuals).

Our results indicate that it is possible to obtain simultaneous unbiased estimates of SNP effects for all three epidemiological host traits, in addition to that of other fixed or random effects influencing disease transmission, from temporal epidemic data. Across all simulated data scenarios, we found that recoverability SNP effects are generally (with few exceptions) easiest to identify, followed by susceptibility and then infectivity SNP effects. In the latter case a large number of contact groups with few individuals provide much more information than the reverse. Simulations of different data scenarios representing optimal and practically feasible recording schemes produced the following main results: firstly, even when only recovery (or death) times of individuals are known inference of SNP effects is still possible, albeit requiring around four times as many individuals to gain equivalent precision. Secondly, only knowing infection times marginally reduces statistical power to detect SNP effects for susceptibility and infectivity, but recovery SNP effects become difficult to detect. Thirdly, when data consist of periodic measurements of individuals' disease status it was found that even relatively infrequent measurements (*e.g.* on a similar timescale as the entire epidemic) could produce SNP effects with high precision, given sufficient data. Lastly, precise estimates of SNP effects could still be obtained with censored epidemic data.

For the model validation, we chose a complex inter-dependence structure for the model parameters by assuming that the SNP in consideration is associated with all three epidemiological host traits (*i.e.* pleiotropy), but with different allele substitution effects and different modes of dominance. Furthermore, we assumed that the traits are also influenced by other fixed effects, have large residual variance (introducing much noise into the system) and are correlated, and that environmental group effects influence the within-group transmission dynamics. This choice represents a worst case scenario as simpler structures can only improve the quality of the parameter estimates.

The models results from different data scenarios indicate a log-linear relationship between the precision of SNP effect estimates and group size or number of groups [25]. For the majority of the simulations presented here, a moderate total population size of 1000 or less individuals was assumed. The corresponding posterior standard deviations for estimated SNP effects were generally above 0.01, and in the case of infectivity effects, more often above 0.1. This would suggest that for datasets comprising 1000 individuals or less, SIRE is only able to detect SNPs of large effects on the epidemiological host traits, but identification of SNPs of small to moderate effects on this trait requires more data or more appropriate data structures, in particular for infectivity.

We chose a dataset comprising 1000 individuals partly because of computational efficiency but also because generating datasets of this size seems feasible for transmission experiments in plants and most domestic livestock species is feasible, in particular in aquaculture species [17, 44, 45], as well as

[25] This relationship in analytically confirmed in a follow up paper [43].

for most field studies. However, many existing field data, in particular in dairy cattle populations with routine genotyping and frequent recordings of disease phenotypes *e.g.* for mastitis, bovine Tuberculosis, and other infectious diseases [46-48] already exceed this number by several orders of magnitude. As genotyping costs continue to fall and automated recording systems are applied at rapidly increasing frequency in agriculture [49, 50], the possibility of identifying SNPs with small to moderate effects on the epidemiological host traits, and their mode of dominance, which was poorly estimated for the given sample size, would appears well within reach in the near future.

It is widely known that disease traits are mostly polygenic, *i.e.* regulated by many genes each with small effect, and hence that SNPs with large effect on disease phenotypes are the exception rather than the norm [6, 51]. Despite this, it is not unreasonable to assume that SNPs with moderate to large effects on either epidemiological trait, and in particular on host infectivity, may indeed exist. This is partly due to the fact that observed disease phenotypes, such as individuals' binary infection status or infection time are the result of many interacting biological processes, each controlled by different set of genes or genetic pathways. Hence the impact of an individual gene on the disease phenotypic is diluted. In contrast, the relative impact of a particular gene on traits that are more closely related to specific biological processes, such as *e.g.* pathogen entry, replication or shedding affecting susceptibility, recoverability or infectivity, respectively, may be higher. Furthermore, evolutionary theory suggests that alleles that confer low susceptibility to infection or fast recoverability from infection are subject to strong directional selection when individuals are commonly exposed to infection. Hence, such beneficial alleles tend to become fixed within few generations, and consequently, SNPs with large effects on disease susceptibility or recoverability would be expected to occur primarily only in populations that have not experienced strong selection pressure for these traits. This is exemplified in the case of Infectious Pancreatic Necrosis (IPN) in farmed Atlantic salmon that have only undergone few generations of selection, where a single SNP explains most of the variation in mortality of fish exposed to the IPN virus [45, 52]. In contrast, selection pressure on infectivity is assumed to be low, since an individual's infectivity genes affect the disease phenotype of group members rather than its own disease phenotype [35, 53, 54]. Therefore, infectivity SNPs with large effect may indeed exist, and may now be identifiable with the methods presented here.

This methods developed in this study and integrated into SIRE complement and succeed previous studies that aimed to develop statistical methods for estimating genetic effects for the different host epidemiological traits, in addition to the much investigated susceptibility effects. Using mastitis in dairy cattle as a case study, Welderufael et al. already demonstrated that genetic risk estimates for both susceptibility and recoverability can be obtained with the help of bivariate threshold models applied to longitudinal binary individual disease records; however incorporation of infectivity effects into these models is challenging.

Alternative approaches have focused on disentangling susceptibility from infectivity effects, but these ignored genetic variation in recoverability [55, 29, 28, 53]. The key novelty of our approach lies in its ability to extract genetic information for all three epidemiological host traits from temporal epidemic data, even when that data is incomplete.

Applications

Many disease challenge experiments and field studies have identified SNPs with moderate to large effects on measurable disease resistance phenotypes) [56-58]. However, the role of these SNPs on transmission dynamics is often poorly understood. For example, it is generally not known whether individuals that carry the beneficial allele for *e.g.* surviving infectious challenge are less likely to become infected (*i.e.* less susceptible), or more prone to surviving infection (*e.g.* due to better recoverability), and also less prone to transmitting infection, once infected (*i.e.* less infective). From an epidemiological perspective, SNPs with favourable pleiotropic effects on all three host epidemiological traits are highly desirable for preventing or mitigating disease spread [59]. In contrast alleles associated with better survival in existing GWAS would only bring the expected beneficial epidemiological benefits if they don't simultaneously confer greater infectivity. In other words, knowing the SNP effects for all three underlying epidemiological host traits is pertinent for effective employment of genetic disease control. Based on the results of this paper, SIRE can be readily applied to disentangle such SNP effects using data from transmission experiments or field studies.

Furthermore, although this paper focuses on estimating SNP effects, SIRE could also immediately be applied to estimating breed, age, sex or treatment or vaccination effects, or any other factor that may affect disease spread, even if genetic information is absent. Indeed, estimates of fixed effects were relatively precise and robust across all simulated datasets in this study.

Limitations of the current approach and future work

One of the potential practical limitations for accurately estimating infectivity SNP effects is that they require a large number of epidemic groups. Previous work has shown that experimental designs can have a significant impact on the precision and accuracy with which model parameters can be estimated (as demonstrated to some extent in this paper and also investigated for indirect genetic effects in numerous other studies [36]). In particular, theoretical studies indicate that significant improvement in estimates of infectivity effects can be achieved by appropriately grouping of genetically related individuals [53, 29]. Whilst this paper has focused entirely on a fixed *A* allele frequency *p* across groups, a follow up paper will show that appropriate variation in genotypes within and across contact groups can lead to substantial improvements in the precision of the infectivity SNP effect a_f, without the need for large numbers of epidemic contact groups[26].

A tool such as SIRE that can accurately estimate the effects of single SNPs on hitherto inaccessible epidemiological traits presents an important first step towards creating a statistically consistent scheme for performing GWAS on epidemiological traits using potentially incomplete data. GWAS, however, typically contains additional features beyond the scope of the simple single SNP analysis presented here. In particular, the current software focuses on one SNP at a time for estimating genetic effects for susceptibility, infectivity and recovery, but ignores the contributions of other genes on these traits. In the current model design these are incorporated in the residual effects. This simplifying assumption may have little impact for appropriately designed transmission experiments, but may lead to biased estimates of SNP effects if genetically similar individuals are not randomly distributed across groups. Theory also suggests that the required sample size for GWAS increases with the number of loci affecting the trait in consideration [51]. Hence, further model development

[26] Interestingly, the susceptibility and recoverability SNP effects cannot be substantially improved.

is required for enabling GWAS for the three underlying epidemiological host traits. Previous work in our group developed a Bayesian algorithm for estimating polygenic effects for host susceptibility and infectivity from incomplete epidemic data [28]. Combining both approaches may prove a useful way forward to allow estimation of genetic effects for all epidemiological host traits under all realistic genetic architectures and different population structures.

This paper introduces, for the first time, software that can estimate genetic and non-genetic effects for susceptibility, infectivity and recoverability simultaneously. This user-friendly tool can be applied to a range of experimental and field data and will help move genetic disease control significantly forward, beyond the focus on genetic improvement of resistance alone.

Acknowledgements

This research was funded by the Scottish Government through the Strategic Partnership in Animal Science Excellence (SPASE) and the Strategic Research programme of the Scottish Government's Rural and Environment Science and Analytical Services Division (RESAS). ADW's contribution was funded by the BBSRC Institute Strategic Programme Grants (BB/J004235/1 (ISP1), BBS/E/D/20002172 (ISP2.1) & BBS/E/D/30002275 (IPS3.1))

Supporting Information

ESM Appendix. The appendixes provide details of the simulation method, the MCMC approach and additional information referred to in the text.

References

1. Organization WH. WHO global strategy for containment of antimicrobial resistance. 2001.

2. Organization WH. Antimicrobial resistance: 2014 global report on surveillance. World Health Organization; 2014.

3. Sheldon J, Soriano V. Hepatitis B virus escape mutants induced by antiviral therapy. Journal of antimicrobial chemotherapy. 2008;61(4):766-8.

4. Gandon S, Day T. Evidences of parasite evolution after vaccination. Vaccine. 2008;26:C4-C7.

5. Russell GE. Plant breeding for pest and disease resistance: studies in the agricultural and food sciences. Butterworth-Heinemann; 2013.

6. Bishop SC, Axford RF, Nicholas FW, Owen JB. Breeding for disease resistance in farm animals. CABI; 2010.

7. Bishop S, Morris C. Genetics of disease resistance in sheep and goats. Small Ruminant Research. 2007;70(1):48-59.

8. Anderson RM, MAY RM. Spatial, temporal, and genetic heterogeneity in host populations and the design of immunization programmes. Mathematical Medicine and Biology. 1984;1(3):233-66.

9. Hethcote HW, Van Ark JW. Epidemiological models for heterogeneous populations: proportionate mixing, parameter estimation, and immunization programs. Math Biosci. 1987;84(1):85-118.

10. Nath M, Woolliams J, Bishop S. Assessment of the dynamics of microparasite infections in genetically homogeneous and heterogeneous populations using a stochastic epidemic model. J Anim Sci. 2008;86(8):1747-57.

11. Doeschl-Wilson AB, Davidson R, Conington J, Roughsedge T, Hutchings MR, Villanueva B. Implications of host genetic variation on the risk and prevalence of infectious diseases transmitted through the environment. Genetics. 2011;188(3):683-93.

12. Raphaka K, Sánchez-Molano E, Tsairidou S, Anacleto O, Glass EJ, Woolliams JA et al. Impact of genetic selection for increased cattle resistance to bovine tuberculosis on disease transmission dynamics. Frontiers in veterinary science. 2018;5.

13. Lively CM. The effect of host genetic diversity on disease spread. The American Naturalist. 2010;175(6):E149-E52.

14. Tsairidou S, Anacleto O, Woolliams J, Doeschl-Wilson A. Enhancing genetic disease control by selecting for lower host infectivity and susceptibility. Heredity. 2019:1.

15. Brooks-Pollock E, De Jong M, Keeling M, Klinkenberg D, Wood J. Eight challenges in modelling infectious livestock diseases. Epidemics-Neth. 2015;10:1-5.

16. Welderufael BG, Løvendahl P, De Koning D-J, Janss L, Fikse F. Genome-wide Association Study for Susceptibility to-and Recoverability from Mastitis in Danish Holstein Cows. Frontiers in genetics. 2018;9:141.

17. Anacleto O, Cabaleiro S, Villanueva B, Saura M, Houston RD, Woolliams JA et al. Genetic differences in host infectivity affect disease spread and survival in epidemics. Sci Rep-Uk. 2019;9(1):4924.

18. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature. 2005;438(7066):355-9.

19. Wong G, Liu W, Liu Y, Zhou B, Bi Y, Gao GF. MERS, SARS, and Ebola: the role of super-spreaders in infectious disease. Cell host & microbe. 2015;18(4):398-401.

20. O'Hare A, Orton R, Bessell PR, Kao RR. Estimating epidemiological parameters for bovine tuberculosis in British cattle using a Bayesian partial-likelihood approach. Proceedings of the Royal Society of London B: Biological Sciences. 2014;281(1783):20140248.

21. Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. Plos Comput Biol. 2012;8(12).

22. Lipschutz-Powell D, Woolliams JA, Doeschl-Wilson AB. A unifying theory for genetic epidemiological analysis of binary disease data. Genet Sel Evol. 2014;46(1):15.

23. Gibson GJ, Renshaw E. Estimating parameters in stochastic compartmental models using Markov chain methods. Ima J Math Appl Med. 1998;15:19-40.

24. O'Neill PD, Roberts GO. Bayesian inference for partially observed stochastic epidemics. J R Statist Soc A. 1999;162:121-9.

25. Lipschutz-Powell D, Woolliams JA, Doeschl-Wilson AB. A unifying theory for genetic epidemiological analysis of binary disease data. Genet Sel Evol. 2014;46:15.

26. Franzén J, Thorburn D, Urioste JI, Strandberg E. Genetic evaluation of mastitis liability and recovery through longitudinal analysis of transition probabilities. Genet Sel Evol. 2012;44(1):10.

27. Welderufael B, Janss L, de Koning D, Sørensen L, Løvendahl P, Fikse W. Bivariate threshold models for genetic evaluation of susceptibility to and ability to recover from mastitis in Danish Holstein cows. J Dairy Sci. 2017;100(6):4706-20.

28. Anacleto O, Garcia-Cortés LA, Lipschutz-Powell D, Woolliams JA, Doeschl-Wilson AB. A novel statistical model to estimate host genetic effects affecting disease transmission. Genetics. 2015;201(3):871-84.

29. Anche MT, Bijma P, De Jong MCM. Genetic analysis of infectious diseases: estimating gene effects for susceptibility and infectivity. Genet Sel Evol. 2015;47.

30. Anche M, De Jong M, Bijma P. On the definition and utilization of heritable variation among hosts in reproduction ratio R0 for infectious diseases. Heredity. 2014;113(4):364-74.

31. Biemans F, de Jong MCM, Bijma P. A model to estimate effects of SNPs on host susceptibility and infectivity for an endemic infectious disease. Genet Sel Evol. 2017;49(1):53. doi:10.1186/s12711-017-0327-0.

32. Keeling MJ, Rohani P. Modeling infectious diseases in humans and animals. Princeton: Princeton University Press; 2008.

33. Pooley CM, Bishop SC, Doeschl-Wilson A, Marion G. Posterior-based proposals for speeding up Markov chain Monte Carlo. Submitted to JASA. 2018.

34. Gillespie DT. Exact Stochastic Simulation of Coupled Chemical-Reactions. J Phys Chem. 1977;81:2340-61.

35. Lipschutz-Powell D, Woolliams JA, Bijma P, Doeschl-Wilson AB. Indirect genetic effects and the spread of infectious disease: are we capturing the full heritable variation underlying disease prevalence? Plos One. 2012;7(6):e39551.

36. Bijma P. Estimating indirect genetic effects: precision of estimates and optimum designs. Genetics. 2010.

37. Wolf JB, Brodie III ED, Cheverud JM, Moore AJ, Wade MJ. Evolutionary consequences of indirect genetic effects. Trends Ecol Evol. 1998;13(2):64-9.

38. Gondro C, Van der Werf J, Hayes BJ. Genome-wide association studies and genomic prediction. Springer; 2013.

39. al. Ae. Submitted. 2018.

40. Ødegård J, Baranski M, Gjerde B, Gjedrem T. Methodology for genetic evaluation of disease resistance in aquaculture species: challenges and future prospects. Aquaculture Research. 2011;42:103-14.

41. Tsapakis I, Schneider WH, Nichols AP. A Bayesian analysis of the effect of estimating annual average daily traffic for heavy-duty trucks using training and validation data-sets. Transport Plan Techn. 2013;36(2):201-17. doi:10.1080/03081060.2013.770944.

42. Gov.Uk. Bovine TB strategy review: summary and conclusions. . https://wwwgovuk/government/publications/a-strategy-for-achieving-bovine-tuberculosis-free-status-for-england-2018-review/bovine-tb-strategy-review-summary-and-conclusions 2018. .

43. Pooley CM, Marion G, Bishop SC, Doeschl-Wilson A. Analysis and experimental design when estimating SNP effects for host susceptibility, infectivity and recovery from epidemic data. bioRxiv. 2019.

44. Gitterle T, Rye M, Salte R, Cock J, Johansen H, Lozano C et al. Genetic (co) variation in harvest body weight and survival in Penaeus (Litopenaeus) vannamei under standard commercial conditions. Aquaculture. 2005;243(1-4):83-92.

45. Houston RD, Haley CS, Hamilton A, Guy DR, Tinch AE, Taggart JB et al. Major quantitative trait loci affect resistance to infectious pancreatic necrosis in Atlantic salmon (Salmo salar). Genetics. 2008;178(2):1109-15.

46. Heringstad B, Klemetsdal G, Ruane J. Selection for mastitis resistance in dairy cattle: a review with focus on the situation in the Nordic countries. Livestock Production Science. 2000;64(2-3):95-106.

47. Banos G, Winters M, Mrode R, Mitchell A, Bishop S, Woolliams J et al. Genetic evaluation for bovine tuberculosis resistance in dairy cattle. J Dairy Sci. 2017;100(2):1272-81.

48. Biemans F, Bijma P, Boots NM, de Jong MC. Digital Dermatitis in dairy cattle: The contribution of different disease classes to transmission. Epidemics-Neth. 2018;23:76-84.

49. Matthews SG, Miller AL, Clapp J, Plötz T, Kyriazakis I. Early detection of health and welfare compromises through automated detection of behavioural changes in pigs. The Veterinary Journal. 2016;217:43-51.

50. Sellier N, Guettier E, Staub C. A review of methods to measure animal body temperature in precision farming. American Journal of Agricultural Science and Technology. 2014;2(2):74-99.

51. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. Genome research. 2007;17(10):1520-8.

52. Moen T, Baranski M, Sonesson AK, Kjøglum S. Confirmation and fine-mapping of a major QTL for resistance to infectious pancreatic necrosis in Atlantic salmon (Salmo salar): population-level associations between markers and trait. Bmc Genomics. 2009;10(1):368.

53. Biemans F, De Jong MC, Bijma P. A model to estimate effects of SNPs on host susceptibility and infectivity for an endemic infectious disease. Genet Sel Evol. 2017;49(1):53.

54. Tsairidou S, Anacleto O, Raphaka K, Sanchez-Molano E, Banos G, Woolliams J et al., editors. Enhancing genetic disease control by selecting for lower host infectivity. Proceedings of the World Congress on Genetics Applied to Livestock Production (Auckland); 2018.

55. Lipschutz-Powell D, Woolliams JA, Doeschl-Wilson AB. A unifying theory for genetic epidemiological analysis of binary disease data. Genet Sel Evol. 2014;46. doi:10.1186/1297-9686-46-15.

56. Houston RD, Gheyas A, Hamilton A, Guy DR, Tinch AE, Taggart JB et al. Detection and Confirmation of a Major QTL Affecting Resistance to Infectious Pancreatic Necrosis (IPN) in Atlantic Salmon (Salmo Salar). Dev Biologicals. 2008;132:199-204.

57. Li D, Lian L, Qu L, Chen Y, Liu W, Chen S et al. A genome-wide SNP scan reveals two loci associated with the chicken resistance to Marek's disease. Anim Genet. 2013;44(2):217-22.

58. Serão N, Kemp R, Mote B, Harding J, Willson P, Bishop S et al., editors. Whole-genome scan and validation of regions previously associated with PRRS antibody response and growth rate using gilts under health challenge in commercial settings. Proceedings of the 10th world congress of genetics applied to livestock production; 2014.

59. Doeschl-Wilson A, Anacleto O, Nielsen H, Karlsson-Drangsholt T, Lillehammer M, Gjerde B, editors. New opportunities for genetic disease control: beyond disease resistance. Proceedings of the World Congress on Genetics Applied to Livestock Production (Auckland); 2018.

Tables

Parameter	Accuracy	y-intercept	Slope	SD	Description
β	0.833	0.000152	1.08	0.0033	Average transmission rate.
γ	0.982	0.000852	0.999	0.0132	Average recovery rate.
k	0.806	1.81	0.633	1.58	Recovery shape parameter
a_g	0.985	-0.00389	1.02	0.091	SNP effects.
a_f	0.875	-0.0535	0.860	0.287	
a_r	0.995	-0.0199	0.990	0.065	
Δ_g	0.728	-0.029	0.450	0.439	Dominance factors.
Δ_f	0.327	0.071	0.128	0.530	
Δ_r	0.789	-0.082	0.608	0.373	
b_{g0}	0.978	-0.012	1.00	0.105	Fixed effect.
b_{f0}	0.871	-0.035	1.10	0.365	
b_{r0}	0.992	0.008	1.00	0.073	
Σ_{gg}	0.885	0.101	0.903	0.136	Covariance matrix for residuals.
Σ_{ff}	0.691	0.264	0.563	0.203	
Σ_{rr}	0.981	0.027	1.00	0.071	
Σ_{gf}	0.789	-0.022	0.949	0.230	
Σ_{gr}	0.978	0.000	0.959	0.067	
Σ_{fr}	0.862	0.002	0.983	0.144	
σ_G	0.899	0.008	1.071	0.144	SD in group effect

Table 1. Summary statistics taken from Fig. 4.

Figures

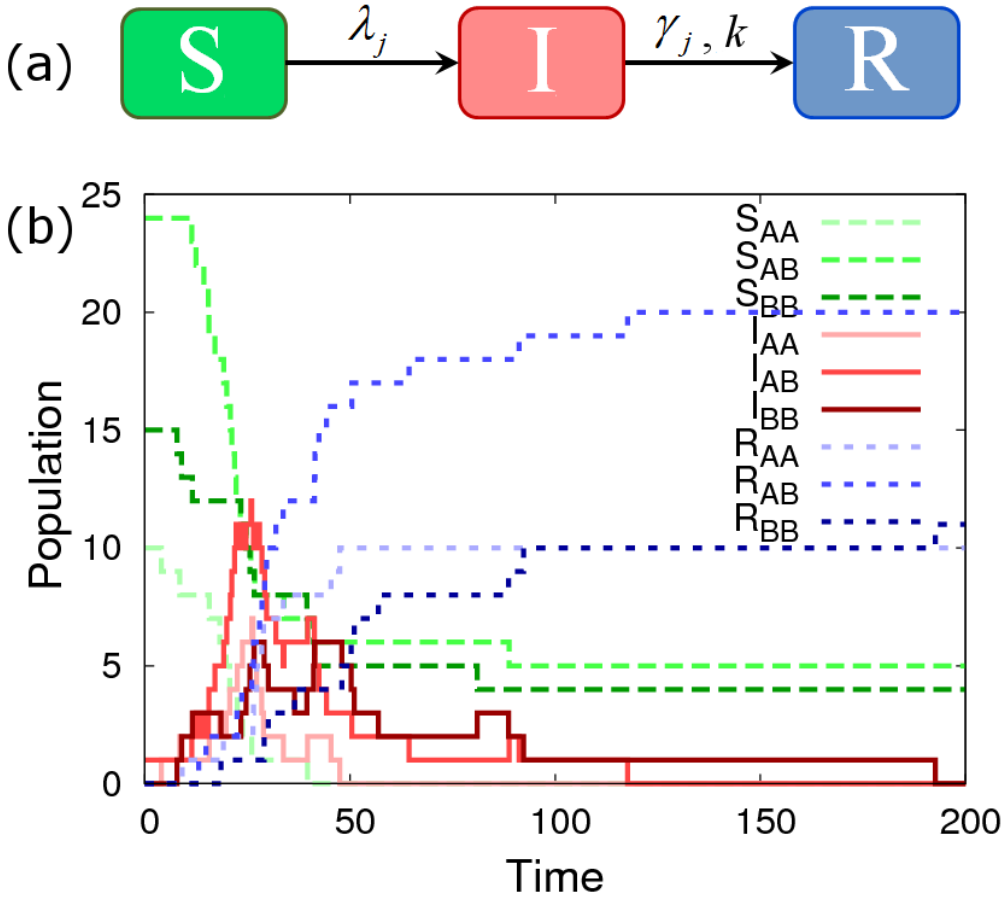

Figure 1. The Model. (a) The individual-based compartmental model, where S, I and R refer to susceptible, infected and recovered classifications. Susceptible individual *j* has a probability per unit time λ_j of becoming infected (see Eq.0.) and the recovery period is assumed to be gamma distributed with mean time γ_j^{-1} (see Eq.(2)) and shape parameter *k*. (b) Whilst both the data and analysis is performed on an individual basis, this graph shows population wide variation in the three SNP genotypes (*i.e. AA*, *AB* or *BB*). This example is simulated using G_{size}=50 individuals, of which one is initially infected. The *A* allele has frequency *p=0.3* in the population with the following parameters used: β=0.006, γ=0.1, *k*=5, a_g=0.4, a_f=0.3 , a_r=-0.4, Δ_g=0.4, Δ_f=0.1, Δ_r=-0.3, b_{g0}=0.2, b_{f0}=0.3, b_{r0}=-0.2, Σ_{gg}=1, Σ_{gf}=0.3, Σ_{gr}=-0.4, Σ_{ff}=1, Σ_{fr}=-0.2, Σ_{rr}=1, σ_G=0.5. Note, the discrete jumps in curves result from discrete disease status transitions in individuals.

Figure 2. SIRE. Illustrative screenshots of the software package: (a) Different data sources are importing by loading user defined data tables (text or cvs files), (b) prior specification can be made on parameters, (c) shows a posterior distribution when inference in being performed, and (d) summary statistics.

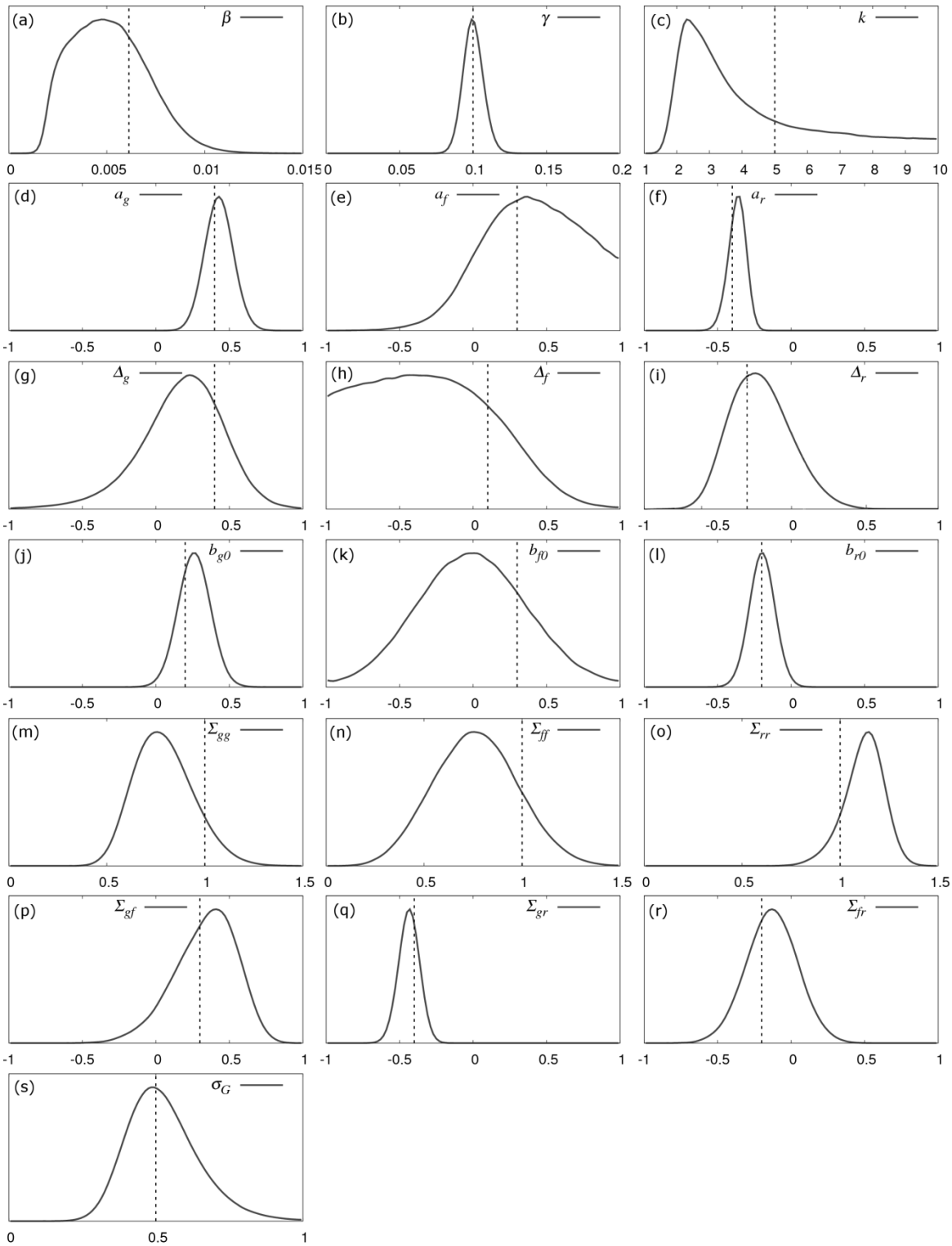

Figure 3. Parameter posterior distributions. Probability distributions for all model parameters inferred from simulated data for N_{group}=20 contact groups each containing G_{size}=50 individuals. The parameter values in Fig. 1 were used for the simulation (denoted by the vertical black lines). Here the data consisted of infection and recovery times for each individual (DS1).

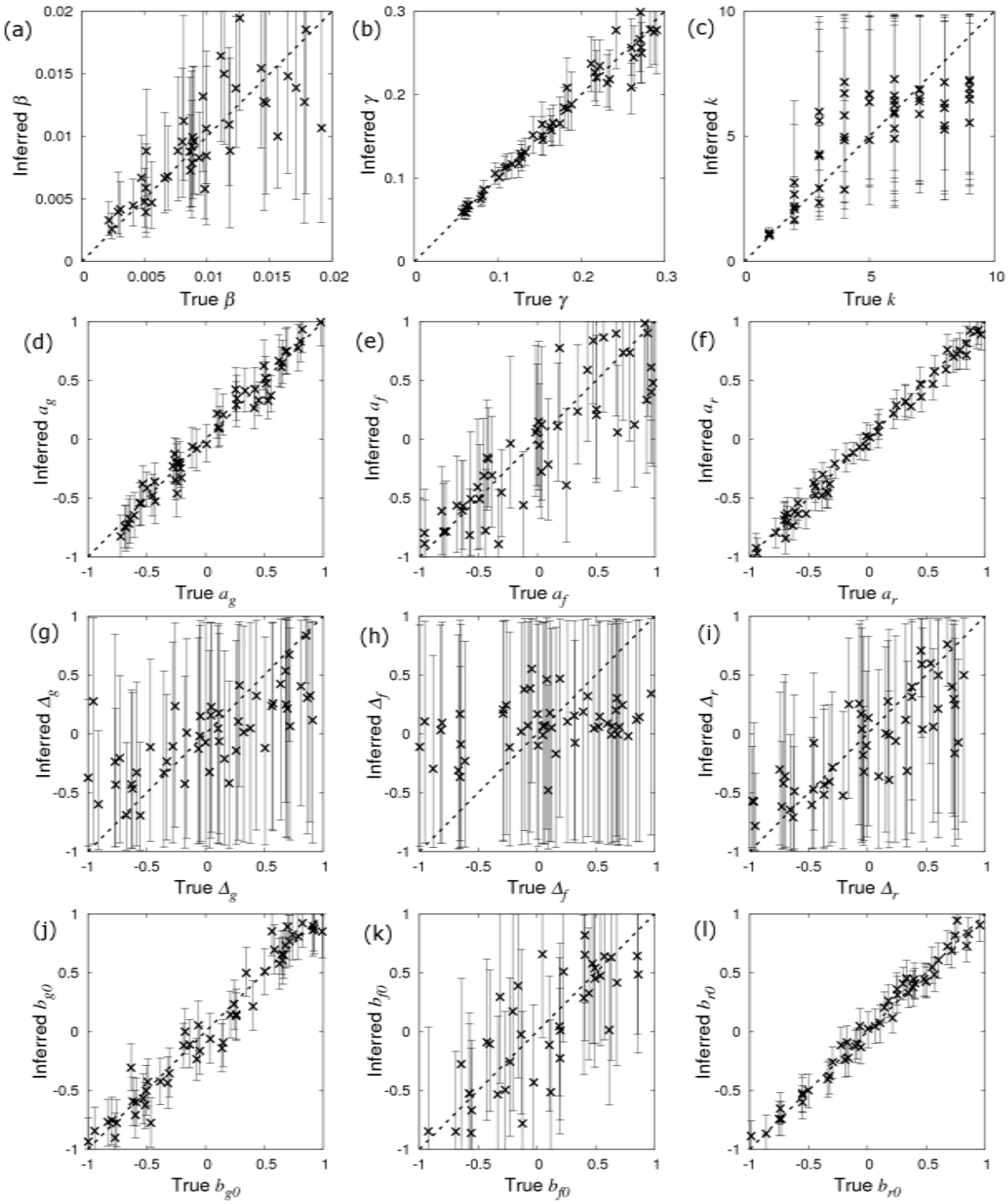

Figure 4. Posterior bias, accuracy and precision. The plots show the inferred values for the SNP effects a_g, a_f and a_r as compared to their true values for 50 simulations (crosses represent posterior means and the error bars indicate 95% credible intervals) with N_{group}=20 contact groups each containing G_{size}=50 individuals for data scenario DS1 (i.e. assuming that all individual infection and recovery times are known). Bias is measured as the regression coefficient between inferred and true parameter value, accuracy is measured by the differences between the crosses and the diagonal line whereas precision is measured by the posterior credible intervals.

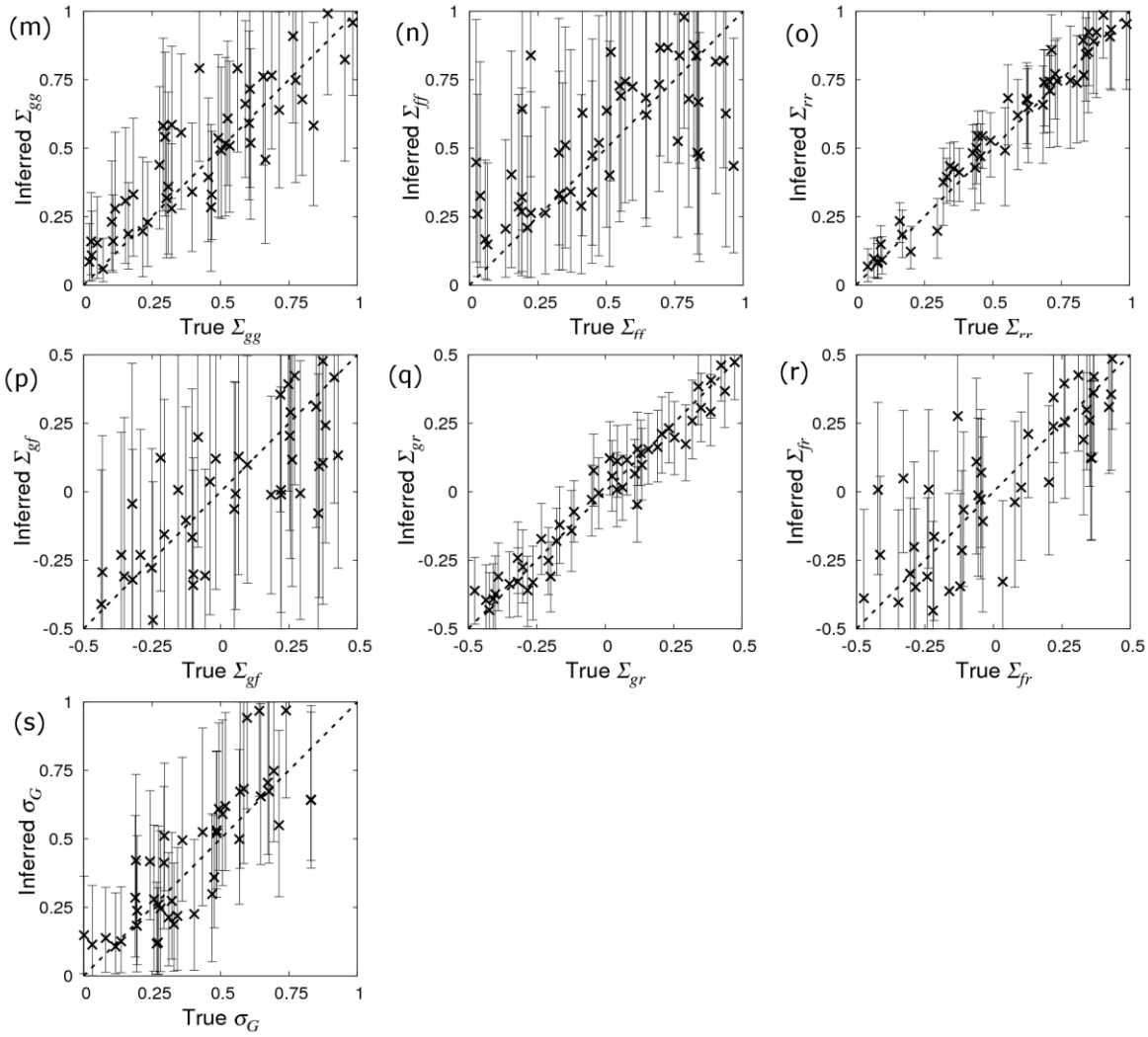

Figure 4 continued.

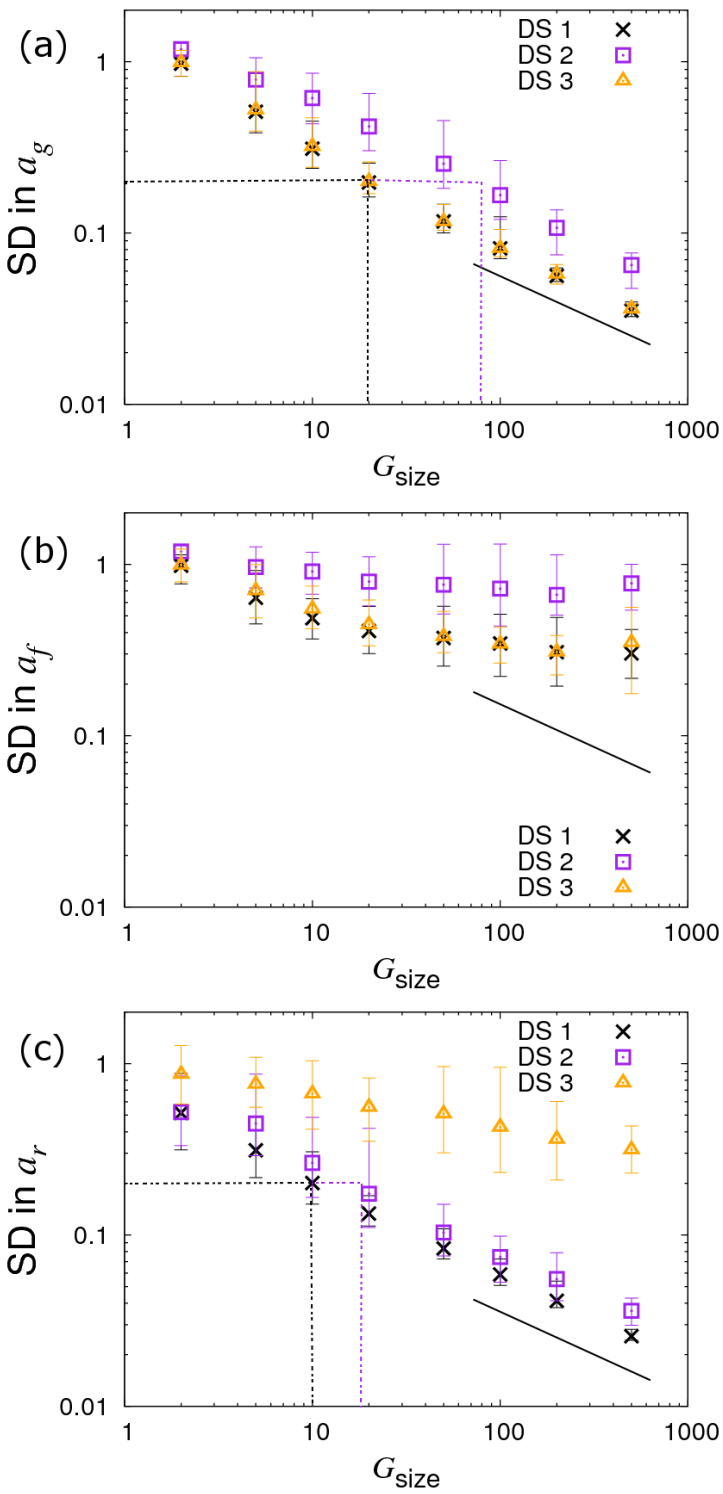

Figure 5. Variation in precision of the SNP effect estimates with group size G_{size}. Posterior standard deviations (SDs) in SNP effects for (a) susceptibility a_g, (b) infectivity a_f and (c) recoverability a_r from simulated data with N_{group}=10 contact groups each containing G_{size} individuals (which is varied). Different symbols represent different data scenarios: DS1) Both the infection and recovery times for individuals are known, DS2) only recovery times are known, and DS3) only infection times are known. Each symbol represents the average in posterior means for 50 simulated data replicates with the error bar denoting the 95% credible interval. The black line indicates a slope of -½ and the dashed black and purple dash lines provide an illustrative example described in the main text. Parameters used are given in Eq.(10).

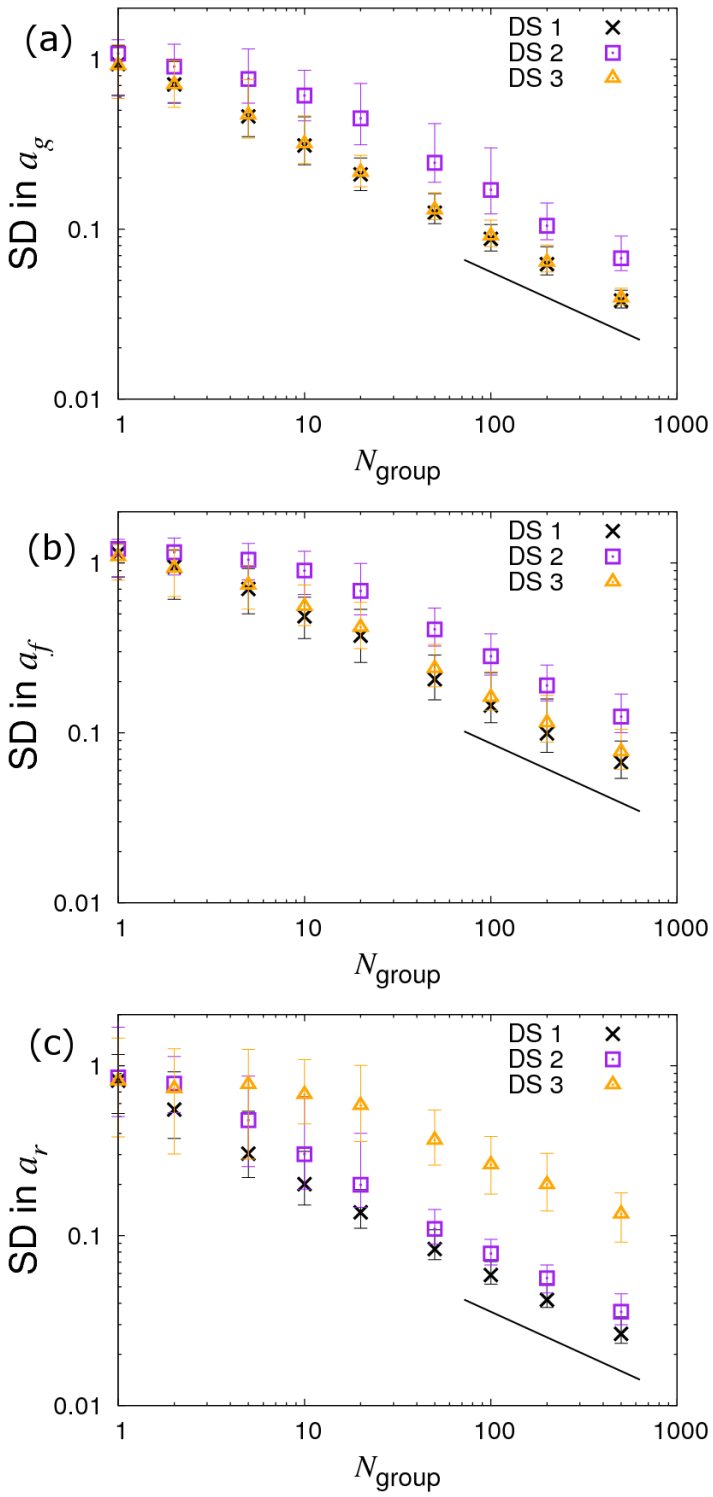

Figure 6. Variation in precision of the SNP effect estimates with number of groups N_{group}. Posterior standard deviations (SDs) in SNP effects for (a) susceptibility a_g, (b) infectivity a_f and (c) recoverability a_r from simulated data with N_{group} contact groups (which is varied) each containing G_{size}=10 individuals. Different symbols represent different data scenarios: DS1) Both the infection and recovery times for individuals are known, DS2) only recovery times are known, and DS3) only infection times are known. Each symbol represents the average in posterior means for 50 simulated data replicates with the error bar denoting the 95% credible interval. The black line indicates a slope of -½. Parameters used are given in Eq.(10).

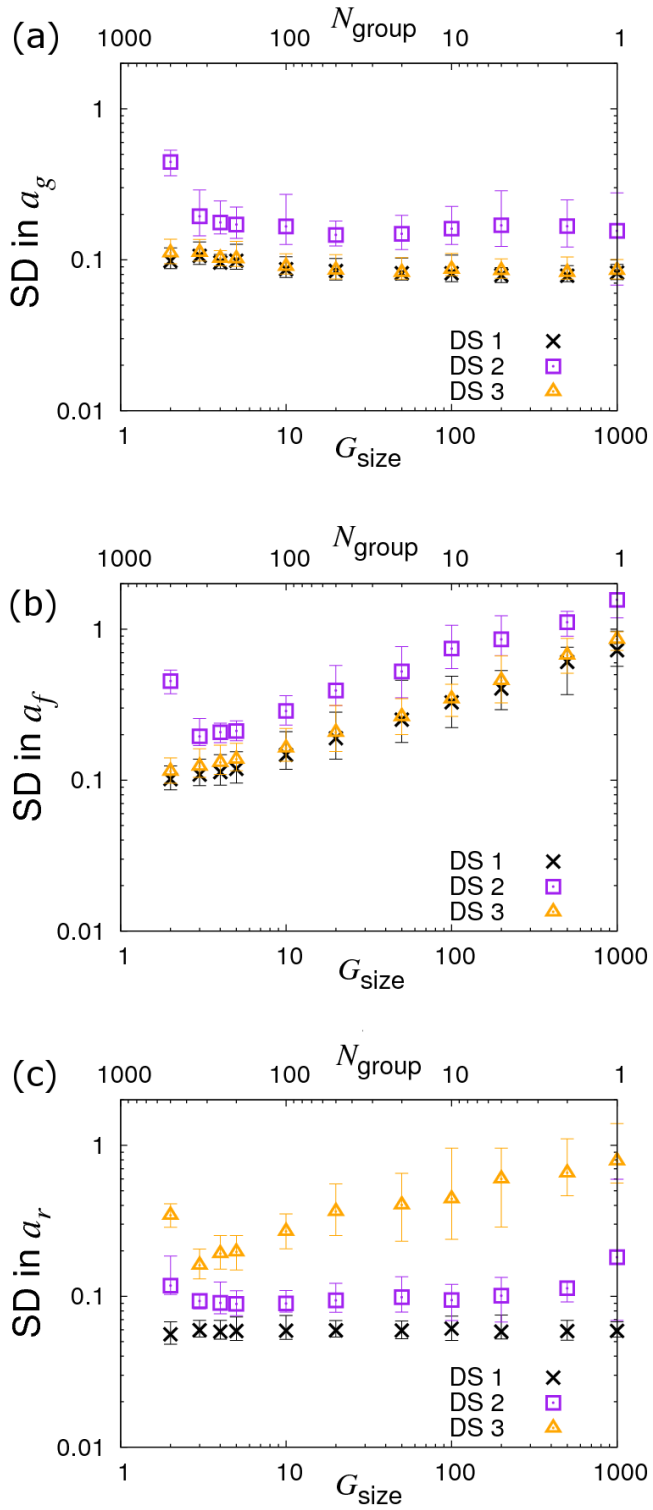

Figure 7. Variation in precision of the SNP effect estimates with partitioning into group. Posterior standard deviations (SDs) in SNP effects for (a) susceptibility a_g, (b) infectivity a_f and (c) recoverability a_r from simulated data with N_{group} contact groups each containing G_{size} individuals, both of which are varied such that the total population $N_{group} \times G_{size}$ is fixed to 1000. Different symbols represent different data scenarios: DS1) Both the infection and recovery times for individuals are known, DS2) only recovery times are known, and DS3) only infection times are known. Each symbol represents the average in posterior means for 50 simulated data replicates with the error bar denoting the 95% credible interval. Parameters used are given in Eq.(10).

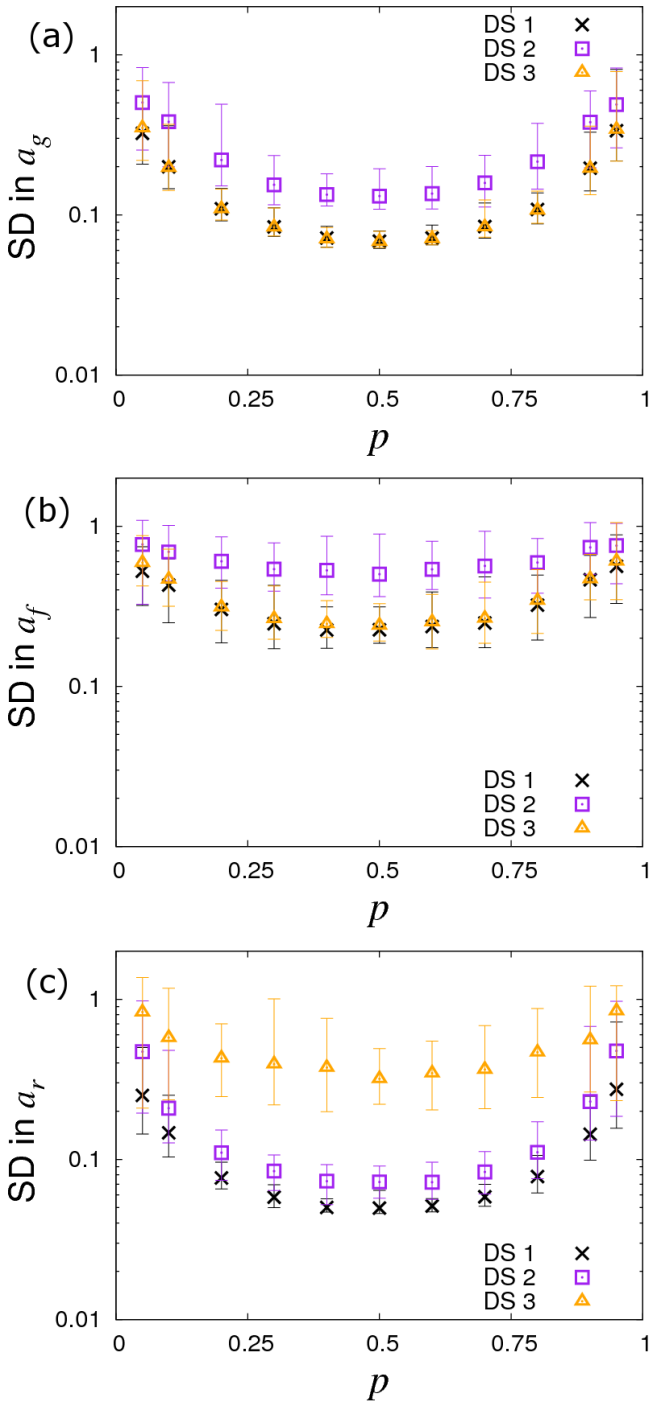

Figure 8. Variation in precision of the SNP effect estimates with allele frequency *p*. Posterior standard deviations (SDs) in SNP effects for (a) susceptibility a_g, (b) infectivity a_f and (c) recoverability a_r from simulated data with N_{group}=20 contact groups each containing G_{size}=50 individuals. Different symbols represent different data scenarios: DS1) Both the infection and recovery times for individuals are known, DS2) only recovery times are known, and DS3) only infection times are known. Each cross represents the average in posterior means for 50 simulated data replicates with the error bar denoting the 95% credible interval**.** Parameters used are given in Eq.(10).

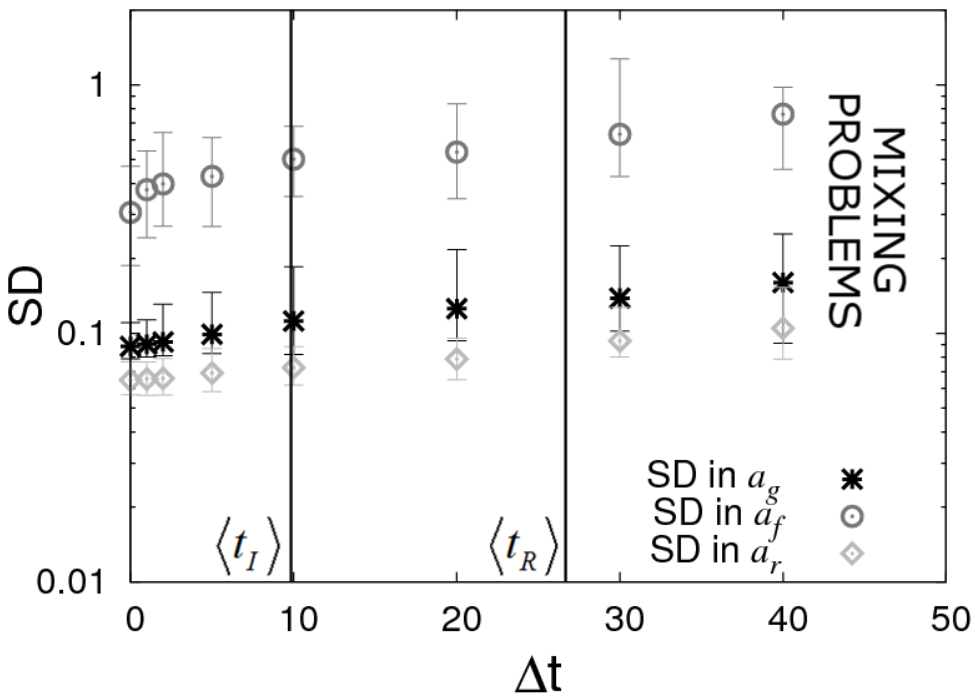

Figure 9. Periodic checking of disease status (DS4). Posterior standard deviations (SDs) in estimated SNP effects a_g, a_f and a_r from simulated data with N_{group}=20 contact groups each containing G_{size}=50 individuals. Here it is assumed that the disease status of individuals is periodically checked with time interval Δt. Each symbol represents the average in posterior means for 50 simulated data replicates (with the checking times randomly shifted across these replicates) with the error bar denoting the 95% credible interval. The vertical lines represent key epidemic times: $\langle t_I \rangle$ is the mean infection time (as averaged over an large number of simulations) and $\langle t_R \rangle$ the mean recovery time. . Parameters used are given in Eq.(10).

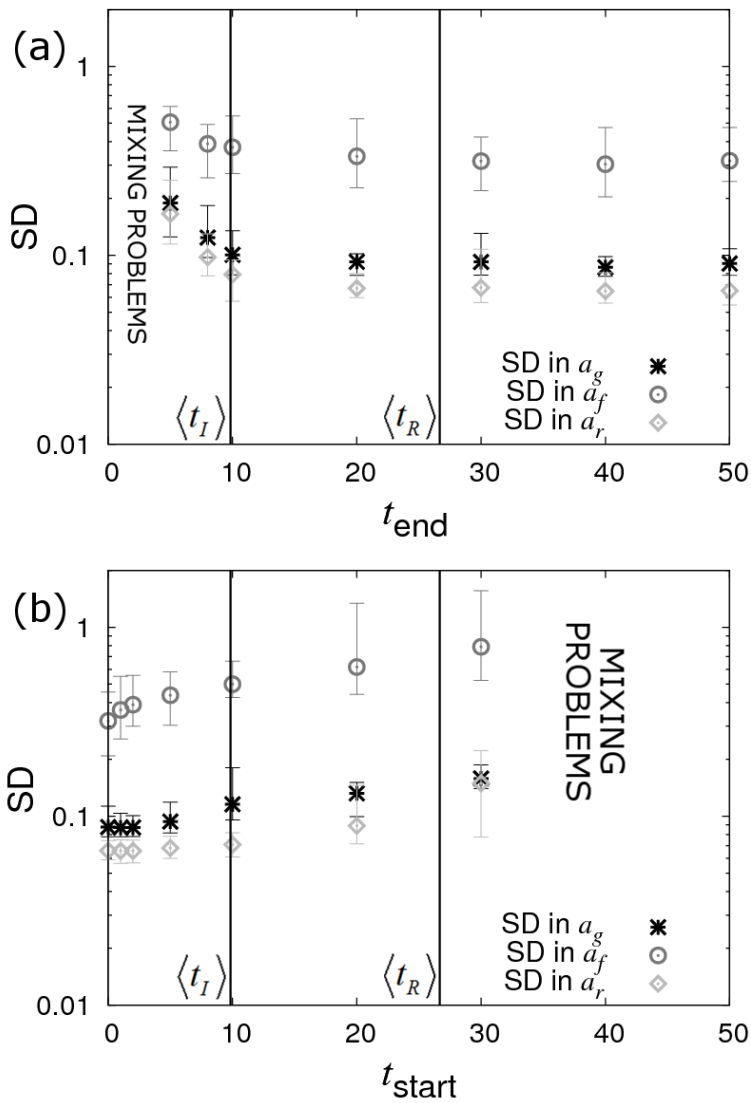

Figure 10. Censoring of data (DS5). Posterior standard deviations (SDs) in SNP effects a_g, a_f and a_r from simulated data with N_{group}=20 contact groups each containing G_{size}=50 individuals. Each symbol represents the average in posterior means for 50 simulated data replicates with the error bar denoting the 95% credible interval. (a) Contact groups are observed until time t_{end}, after which no further data is taken. (b) Contact groups are observed from time t_{start} until the end of all epidemics. The vertical lines represent key epidemic times: $\langle t_I \rangle$ is the mean infection time (as averaged over an large number of simulates) and $\langle t_R \rangle$ the mean recovery time. Parameters used are given in Eq.(10).