bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Estimating individuals' genetic and non-genetic effects underlying infectious disease transmission from temporal epidemic data

Christopher M. Pooley[1,2*], Glenn Marion[2&], Stephen C. Bishop[†], Richard I. Bailey[1] and Andrea B. Doeschl-Wilson[1&]

[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

[†] Deceased

[*] Corresponding author

Email: christopher.pooley@roslin.ed.ac.uk

[&] These authors contributed equally to this work.

Abstract

Individuals differ widely in their 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) and recoverability (propensity to recover quickly). Interventions aiming to reduce disease spread may target improvement in any one of these traits, but the necessary statistical methods for obtaining risk estimates are lacking. In this paper we introduce a novel software tool called *SIRE* (standing for "Susceptibility, Infectivity and Recoverability Estimation"), which allows simultaneous estimation of the genetic effect of a single nucleotide polymorphism (SNP), as well as non-genetic influences on these three unobservable host traits. SIRE implements a flexible Bayesian algorithm which accommodates a wide range of disease surveillance data comprising any combination of recorded individual infection and/or recovery times, or disease status measurements. Different genetic and non-genetic regulations and data scenarios (representing realistic recording schemes) were simulated to validate SIRE and to assess their impact on the precision, accuracy and bias of parameter estimates. This analysis revealed that with few exceptions, SIRE provides unbiased, accurate parameter estimates associated with all three host traits. For most scenarios, SNP effects associated with recoverability can be estimated with highest precision, followed by susceptibility. For infectivity, many epidemics with few individuals give substantially more statistical power to identify SNP effects than the reverse. Importantly, precise estimates of SNP and other effects could be obtained even in the case of incomplete, censored and relatively infrequent measurements of individuals' infection or survival status, albeit requiring more individuals to yield 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.

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1 Introduction

In the era of rapid expansion in the human population resulting in increasing demands on food security, effective solutions that reduce the spread of infectious diseases not only in humans, but also in plants and livestock, are urgently needed. Failure of stringent biosecurity measures [2, 3] and emergence of anti-microbial resistance [4, 5] and escape mutants to viral vaccines [6, 7] demonstrate that infectious diseases cannot be combatted by conventional biosecurity and pharmaceutical interventions alone.

The advent of genome wide high density single-nucleotide polymorphism (SNP) chip panels has already led to a remarkable range of discoveries regarding the genetic regulation and biology of diseases and translation towards innovative therapeutics [8]. In agriculture, SNP chip panels have revolutionized breeding practices by facilitating genomic selection [9, 10]. In the infectious disease context genomic selection may effectively prevent or reduce disease spread by providing a means to identify and select against individuals with high genetic risk of becoming infected or transmitting infections purely based on their genetic make-up, without the need of exposing them to infectious pathogens [11]. However, to date the full host genetic basis underlying infectious disease transmission is still poorly understood.

Epidemiological models are widely used to identify risk factors for disease spread in populations. Indeed, modelling disease transmission in genetically heterogeneous populations is well established (see *e.g.*[12, 13]). 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*, the relative risk of an uninfected individual to become infected when exposed to a typical infectious individual or infectious material excreted from such an individual, *infectivity*, the propensity of an individual, once infected, to transmit infection to a typical (average) susceptible individual, and *recoverability*, the propensity of an individual, once infected, to recover or die) [14, 15]. As demonstrated by numerous simulation studies, host genetic variation in any one of these traits, if correctly identified, could be exploited to reduce infectious disease spread within and across populations [15-18]. However, 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 [18], remarkably little is known about the genetic regulation of host recoverability and infectivity, despite emerging evidence that genetic variation in these traits exists [19, 20]. 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 [21-23]. This shortcoming is largely because appropriate 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 many conventional genome-wide association studies (GWAS) [24], target traits for genetic improvement are measured directly, 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 effects are 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) may not only depend on its own susceptibility

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and recoverability genes, but also on the infectiousness of other individuals in the same contact group, *i.e.* their infectivity and recoverability genes [25]. This complex interdependence between underlying and observable traits poses challenges for existing methods.

The motivation behind this paper is to introduce new statistical and computational methods that utilise information derived from observation of epidemics and trait interdependence to estimate, for the first time, genetic and other systematic effects for all three underlying epidemiological host traits. This requires combining statistical, epidemiological and genetic modelling principles. Analysis of incomplete epidemic data to draw inferences on epidemiological parameters is well established [26, 27]. However, analysing such data to draw joint inferences on both the disease epidemiology and host genetic variation has proven challenging [25, 28]. Recent studies have expanded conventional quantitative genetics threshold models to enable joint genetic evaluation of cattle susceptibility to, and recoverability from, mastitis [29, 30], which led to identification of novel SNPs and candidate genes associated with these traits [19]. However, because infectivity acts on group members rather than the focal individual itself, applying these technique to estimate genetic effects for infectivity is problematic.

Alternative approaches have focused on disentangling susceptibility from infectivity effects. For example, Anacleto *et al.* [31] 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 (*i.e.* they are determined by a large number of genes, each with small effect). 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 [32, 33], used a generalized linear model (GLM) to estimate 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). It also failed to 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 [34], which adopted a GLM to analyse time-series data on individual disease status, illustrated the benefits of longitudinal records of individuals' infection status for improving prediction accuracies of SNP effects, although it still relied on a number of simplifications that may compromise prediction accuracies and lead to unwanted bias. A further shortcoming of previous approaches [32-34] is that they ignore potential pleiotropic effects, *i.e.* SNPs affecting more than one epidemic trait. This seems unrealistic, since, for example, SNPs that control within host pathogen replication may also lower the risk that infection can establish, *i.e.* reduce susceptibility, and simultaneously reduce pathogen shedding and hence infectivity, and speed up recovery.

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 (importantly capturing any pleiotropy), together with that of other fixed effects (such as *e.g.* sex, breed 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 or imperfect diagnostic tests), 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

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

are known exactly, but also under more realistic scenarios in which epidemics are only partially observed.

2 Materials 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 contact. 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 in plants or livestock, 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 group of individuals sharing the same environment such as a 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 closed, *i.e.* no births, migrations, or transmission of disease between groups. This assumption generally holds for experimental studies and also for most common field situations, where a movement ban is imposed after disease notification [35].

The dynamic spread of disease within a contact group is modelled using a so-called SIR model [36]. 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 $I(\mathrm{t})$, the number of infected individuals at time *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 [31] 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 in individual *j's* susceptibility as compared to that of the population as a whole (*e.g.* g_j=0.1 corresponds to individual *j* being ≃10% more susceptible than the population average), 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 individual *j* (note, this sum varies as a function of *t* as individuals become infected and recover). The term G_z in Eq.(1) accounts for the fractional deviation in disease transmission for group *z*. This incorporates 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 differing 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. The exponential dependencies in Eq.(1) ensure that λ_j is strictly positive and allow for the possibility that some groups or individuals are much more/less susceptible/infectious than others, *i.e.* it can accommodate potential super-spreaders.

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 an average recovery rate across the population 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 biologically realistic profile compared with the exponential distribution often assumed (see electronic supplementary material Appendix A for further details).

Following standard quantitative genetics theory [37], the individual-based deviations in susceptibility ***g***, infectivity ***f*** and recoverability ***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 effects – The model assumes that a specific locus defined by a SNP (potentially) plays an important contribution to the trait values (note, repeated analysis can be performed on different SNPs of interest). Assuming a diploid genomic architecture with biallelic SNP implies three SNP genotypes: *AA, AB* and *BB*. The SNP contribution to the traits for individual *j* depends on *j*'s genotype in the following way:

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

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, infectivity and recoverability, respectively (*e.g.* if a_g is positive, individuals with an *AA* genotype will, on average, be more susceptible to disease than those with a *BB* genotype). 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 corresponds to complete dominance of the *A* allele over the *B* allele and -1 when the reverse is true, whereas absence of dominance is represented by a value of 0) [38].

Fixed effects – The design matrix ***X*** and fixed effect vectors $\boldsymbol{b}_g$, $\boldsymbol{b}_f$ 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). Following convention, an additional fixed effect is added to account for trait mean, which is explicitly chosen to ensure the population averages of ***g***, ***f*** and ***r*** are zero (remembering that the average effects are already captured by the parameters β and γ).

Residual contributions – Here $\boldsymbol{\varepsilon}=(\boldsymbol{\varepsilon}_g, \boldsymbol{\varepsilon}_f, \boldsymbol{\varepsilon}_r)$ accounts for all other contributions to the traits (*i.e.* coming from genetic effects not captured by the SNP under consideration, as well as any non-genetic environmental variation). We assume that for each individual the three trait residuals are drawn from a single multivariate normal distribution with zero mean and 3×3 covariance matrix **Σ**.

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 (on top of any correlations which may also arise from the SNP and fixed effects). 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 these 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 is the dominant mechanism affecting the speed at which infection spreads within a group rather than group specific factors affecting individuals' susceptibility, infectivity or recoverability.

2.2 Bayesian inference

Based on the description above, the model contains the following set of parameters: θ=(β, γ, k, a_g, a_f, a_r, Δ_g, Δ_f, Δ_r, $\boldsymbol{b}_g$, $\boldsymbol{b}_f$, $\boldsymbol{b}_r$, $\boldsymbol{\varepsilon}_g$, $\boldsymbol{\varepsilon}_f$, $\boldsymbol{\varepsilon}_r$,, $\boldsymbol{\Sigma}$, $\boldsymbol{G}$, σ_G). We denote the complete set of infection and recovery event times for all individuals as ξ over the observed duration of the epidemics [39]. 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

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

where individual components are defined as follows:

Observation model $\pi(y|\xi)$ – the probability of the data given a set of event times ξ. The expression for the observation model depends on the nature of the data being observed. In many contexts this simply takes the values one or zero depending on whether ξ is consistent with y or not. For example a perfect disease diagnostic test showing that an individual is infected would be 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. 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. When imperfect disease diagnostic test results are available the observation model includes the sensitivity and specificity of the test to account for this uncertainty in the data. In summary, the observation model depends on the data collection process and constrains the possible event sequences ξ, and this, in turn, informs the model parameters θ.

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 [26, 27] (see Appendix B for details), and is given by

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

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 [40], 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 calculated 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 [41].

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

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

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

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

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 independent between groups and normally distributed with standard deviation σ_G.

The default prior for $\theta_{-\varepsilon,\boldsymbol{G}}$ (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 [1] to aid fast mixing of the residual parameters (details are given in Appendix D).

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2.3 SIRE

SIRE is a desktop application that implements the Bayesian algorithm outlined above. It is freely available to download from the supplementary material or at www.mkodb.roslin.ed.ac.uk/EAT/SIRE.html (with versions for Windows, Linux and Mac). An easy to use point and click interface allows for data tables to be imported in a variety of formats and graphical outputs are dynamically displayed as inference is performed. The core of SIRE utilises efficient C++ code and allows for running MCMC chains on multiple CPU cores.

SIRE takes as input any combination of information about infection times, recovery times, disease status measurements, disease diagnostic test results, genotypes of SNPs or any other fixed effects (see screenshot in Fig 1a), details of which individuals belong to which contact groups and any prior specifications (Fig 1b). The output from SIRE consists of posterior trace plots for model parameters θ, distributions (Fig 1c), visualisation of infection and recovery times ξ, dynamic population estimates and summary statistics (means and 95% credible intervals) as well as MCMC diagnostic statistics (Fig 1d). Posterior distribution graphs can be exported from SIRE and also files containing posterior samples of θ and ξ for further analysis using other tools. The user guide for SIRE is available in the electronic supplementary material and on the website.

2.4 Data scenarios

SIRE is flexible to many possible inputs. Reflecting real-world datasets this paper considers five potential data scenarios (DS):

DS1: Infection and recovery times for all individuals exactly known

This represents the best case scenario for inferring parameter values. For example, appearance of symptoms or visual or behavioural signs may indicate the onset of infection, and recovery/removal times are given by the time of death.

DS2: Only recovery times known

Often "recovery" in compartmental SIR models represents the death and removal 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 [42].

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.

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 the intervening period (note here we assume perfect diagnostic tests but SIRE also allows for imperfect diagnostic test results to be used, provided the sensitivity and specificity of the tests are known).

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 (*e.g.* animal welfare), even though epidemics have not completely died out.

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, accuracy and bias of inferred model parameters depends on the type of data available.

Initially the focus of results will be on DS1 (which although rarely applies in practice, still provides useful insights for software validation and application) and later in section 3.5 consideration is given to DS2-5.

3.1 Illustrative example simulation and inference

We first demonstrate the performance of SIRE assuming complete information of individuals' infection and recovery times, for a representative but complex set of parameters with regards to the genetic and non-genetic regulation of the three epidemiological host traits. Subsequently we investigate how these results change under different parameter and data scenarios.

Simulations

Individuals were randomly assigned into N_{group} different contact groups, with each group containing G_{size} individuals. The SNP under investigation was assumed to be in Hardy-Weinberg equilibrium [38] with an *A* allele frequency of p=0.3. For the effect sizes we used the values a_g=0.4, a_f=0.3 , a_r=-0.4, representing a relatively large pleiotropic effect (which confers higher susceptibility for *AA* compared to *BB* individuals, as well as slightly higher infectivity and reduced 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 included only 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 larger than is biologically realistic, but here we want to demonstrate 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 individuals that are more susceptible 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 effects to σ_G=0.5. Finally, the average transmission rate was chosen to be $\beta=0.3/G_{size}$ (selected because it led to a substantial fraction of individuals becoming infected and including G_{size} such that the basic reproductive ratio R_0 remained independent of group size, *i.e.* frequency dependent transmission) and an average recovery rate γ=0.1 with shape parameter k=5 (corresponding to the infection duration being relatively highly peaked around a mean of 10 time units).

Simulated epidemic data was generated by means of a Doob-Gillespie algorithm [43] 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 Fig 2. Whilst the simulation itself is generated on an individual basis, this graph

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

summarises dynamic variation in the susceptible, infectious and recovered populations, categorised by SNP genotype. It reveals classic epidemic SIR model behaviour: 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 less likely to remain uninfected. Consequently we can link the genetic composition in the final state of the epidemic to the expected value for a_g (which, based on this particular dataset, 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 event times themselves. The Bayesian approach adopted in this paper utilises all this information to extract the best available parameter estimates.

The information content from a single epidemic is generally insufficient to estimate the large number of parameters in the model. Therefore we next simulated a more realistic dataset (using the same parameter set as above) made up of 1000 individuals split into N_{group}=20 contact groups, each containing G_{size}=50 individuals. The infection and recovery event times from this simulation were then used as input data into SIRE (scenarios in which infection and recovery times are not known precisely are discussed later in section 3.5).

Parameter estimates

Fig 3 shows the inferred posterior probability distributions for all parameters in θ corresponding to the simulated multi-group scenario described above. The actual parameter values used to generate the data (see vertical black dashed lines in Fig 3) consistently lie within regions of high posterior probability. The standard deviations (SDs) in these distributions characterise the precision with which parameters can be estimated:

Population average parameters (Figs 3a-c) – The recovery rate γ has the greatest precision (smallest relative SD), followed by the transmission rate β. Whilst the distribution for the shape parameter *k* is wide, it is clearly able to discount the possibility of an exponential recovery duration (*i.e.* k=1), which has a very low posterior probability, over a more peaked distribution (*i.e*. *k*>1).

SNP effects (Figs 3d-f) – The estimated recovery SNP effect a_r is highly peaked around its true value of -0.4 (Fig 3f). 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 3d, albeit with a wider posterior probability distribution. This difference 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 more dispersed Poisson process governing infection. The infectivity SNP effect a_f in Fig 3e exhibits a much wider probability distribution than the other two SNP effects. The fact that zero *does* lie within the 95% posterior credible interval (which goes from -0.35 to 2.1) means that no certain association with infectivity can be made in this particular example. Figs 3d-f illustrates a general principle that was common in the vast majority of subsequent simulation scenarios: SNP effects associated with recoverability are most precisely estimated, followed by susceptibility, and finally infectivity [44].

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Scaled dominance factor (Figs 3g–i) – Compared to the SNP effects themselves, precision of the scaled dominance parameters is relatively poor, and actually reduces as the size of the SNP effects goes down (results not shown), which makes sense in the limit of zero SNP effect size, because here no information about dominance is available. Estimating them accurately, therefore, either requires very large SNP effects or substantially more data.

Fixed effects (Figs 3j–l) – Since SNP effects are also a type of fixed effect, the same comments as above also apply for other fixed effects.

Residual covariance matrix and random group effect (Figs 3m–s) – Interestingly, it was possible to obtain relatively good estimates for elements in the residual covariance matrix. Again, the familiar pattern is observed whereby quantities related to recoverability are more precisely estimated than those related to susceptibility, with infectivity the least precise. Finally, the variance of the group effect could be estimated with similar precision as that for susceptibility (Figs 3s & m).

3.2 Dependence on parameter values

The previous section showed an illustrative example for a particular parameter set. Here we assess what happens when different parameters in the model are altered. This was achieved by means of taking the following "base" set of parameters

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

and then changing each parameter separately (fixing all others) [45]. Fig 4 shows scatter plots (each referring to a different selected parameter) of the posterior means (crosses) with corresponding 95% credible intervals inferred from a single simulated dataset using the true selected parameter value on the *x*-axis. Plots in which most crosses lie near to the diagonal line imply that inference is able to accurately capture the true parameter values. Table 1 shows the corresponding prediction accuracy, measured as the correlation between the inferred and true parameter values. Except for Δ_f for which prediction accuracy was only 34%, prediction accuracies for all other parameters ranged from 69-99%. In line with the discussion above, parameters associated with recoverability have generally higher predication accuracies than those associated with susceptibility, which are again higher than those for infectivity.

Bias indicates systematic differences between the true parameter values and those inferred from the data. Bias was measured by fitting regression lines through the posterior means in Fig 4 (as a function of the true parameter value). The corresponding *y*-intercept and slope values are shown in Table 1, where a zero *y*-intercept and a slope of one indicate absence of bias. Whilst the majority of observed *y*-intercepts tended to be very small, the slope for some of the parameters is markedly less than one (most notably for Δ_f). The reason for this is as follows. When Bayesian analysis reveals insufficient information regarding a parameter, its distribution follows that of the prior (which are uniform for all the parameters in this particular study, as described in Appendix C). This behaviour happens irrespective of the parameter's true value, leading to a plot in Fig 4 that would be entirely flat (*i.e.* a slope of zero). Therefore, the slopes of less than one in Fig 4 simply reflect a lack of data, which is essentially another manifestation of a lack of parameter precision. Consequently, bias reduces as the amount of data increases (provided the model being fitted is the correct one).

From the point of view of this paper, the probability distributions which are of greatest interest are the SNP effects. Noting the sizes of the error bars across Figs 4d-f demonstrate that the precisions of

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the parameter estimates are largely independent of the values of the parameters themselves, a result which can be supported analytically [46]. This implies that the precision of SNP effects calculated using the base set of parameters in Eq.(10) is expected to be generally applicable to any other parameter set [47] (*e.g.* the average SDs in Table 1 for the base parameter set are very similar to the SDs shown in Fig 3).

Consequently, the remainder of this paper focuses on investigating how SNP effect estimates are affected by contact group structure and the nature of the measured data using this base set of parameters. We focus first on outlining the behaviour with respect to key design features, *e.g.* group size, number of individuals per group and allele frequency, and then go on to consider how observations of the system influence what can be learned.

3.3 Dependence on the number and size of contact groups

The crosses in Fig 5 shows how SDs in the SNP effects change as a function of the number of individuals G_{size} within each contact group (here N_{group}=10 contact groups are assumed). The SD in a_g reduces as the number of individuals in each contact group G_{size} increases (Fig 5a). Importantly this relationship scales as a line of slope -½ (note the log scales on this plot), corresponding to precision increasing by a factor of two as the number of individuals is increases by a factor of four (in line with what would be expected from central limit theorem). Fig 5a provides insights into how many individuals would need to be observed in order to be able to make an association with a susceptibility SNP effect of a given size. For example, in order to detect an association with a susceptibility SNP of effect size a_g = 0.4, G_{size}=20 individuals per contact group, and so $G_{size}\times N_{group}$=200 individuals in total would be needed to assure that the 95% credibility interval does not contain zero (assuming approximate normality for the posterior distribution), as illustrated by that black dashed line in Fig 5a. Fig 5c shows the same scaling relationship for identifying recoverability SNP effects, but this time only $G_{size}\times N_{group}$=100 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 state of affairs, however, is observed in Fig 5b. 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 as seen 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 [48-50]. 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 information on 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 (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). Because information about the infectivity SNP effect comes from epidemic-wide behaviour, it is expected to

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

scale linearly with the number of contact groups N_{group} (Fig 6b), but not with the number of individuals per contact group G_{size} (Fig 5b).

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 7a we find very little variation in the precision of a_g. Interestingly, the results in Fig 7b clearly demonstrate that *larger* numbers of contact groups containing *fewer* individuals help to reduce the SD in the infectivity SNP effect 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 7c 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 (although in cases in which R_0 is small, differences may result from variation in the fraction of individuals which actually become infected).

3.4 Dependence on allele frequency

So far we have assumed a fixed *A* allele frequency *p*=0.3 in the population. Fig 8 demonstrates 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 a minimum of *p*=0.5 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 [51].

3.5 Different data scenarios

This section shows results from the various data scenarios introduce in section 2.4, in which the infection and recovery times of all individuals are not known precisely:

DS2: Only recovery times known

Since a_g and a_f relate to the infection process, naïvely it might be expected that because infection times are unknown then nothing can be inferred about these SNP effects. 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 through the likelihood in Eq.(6).

The square symbols in Fig 5a denote the posterior SDs in the susceptibility SNP effect a_g under DS2. Compared to the best case scenario DS1, the SD in a_g increases as a result of having to infer probable infection times for individuals (as opposed to knowing them exactly). The number of individuals per group 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 5a), as opposed to G_{size}=20 in the case of DS1. Consequently to achieve an equivalent precision for a_g under DS2 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 6b, assuming a large number of contact groups), and for the recoverability it is 1.9 (see Fig 5c). These factors were found to be remarkably consistent across a broad range of group numbers and sizes (results not shown).

Estimates of prediction accuracies and bias for the case of DS2 were obtained as described in section 3.2, and results are presented in Appendix H. Compared to DS1 (Fig 4 and Table 1), The prediction accuracies tend to be slightly lower (but still above 0.5 in the majority of cases and above 0.9 for some parameters) and the bias slightly higher, reflecting the reduction in data. However, similar patterns with regards to which parameters are associated with lower prediction accuracy and bias emerge as was seen for DS1 (Fig 4).

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

The triangles in Figs 5, 6 and 7 show results under DS3 for different group sizes and group compositions. 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 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 -½ (on the log-scale) as the number of contact groups N_{group} increases (*i.e.* the triangles in Fig 6c scale with the black line) but not when the number of individuals per contact group G_{size} is changed (see Fig 5c).

DS4: Disease status periodically checked

Fig 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 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 and $\langle t_R \rangle$ is the average recovery time (these quantities are found by averaging over a large number of simulated replicates). 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. This means that results assuming DS1 which are either analytical (explored in a follow up paper [46]) or numerical, as looked at in sections 3.1-3.4, remain relevant in realistic data scenarios.

The limit on the right hand side of Fig 9 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 [52] (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

In Fig 10a 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 largely because most individuals recover before $\langle t_R \rangle$ as a consequence of a small number of individuals having very low recoverability (which itself arises because of the large residual variance Σ_{rr}=1 assumed here). Given that $\langle t_R \rangle$ is usually substantially less than the total epidemic time, from a practical point of view terminating disease transmission experiments prior to the end of the epidemic when no new infections occur, (and perhaps performing further replicates) may be beneficial. However, the effectiveness of this approach would depend on a large assumed variation in recoverability in the population, which *a priori* may be unknown.

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Fig 10b 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 the total number of infection events), the disease status of all the individuals at t_{start} can be accurately inferred (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) 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 are not comprehensive. Any combination of infection time, recovery time, disease status data and diagnostic test results can be used as inputs into SIRE. Furthermore SIRE accounts for additional uncertainties in cases in which data is missing on some individuals and where diagnostic tests are imperfect.

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 recent years, and evidence for their effective utilization in personalized medicine and livestock and plant breeding programmes continues to emerge [53-56]. 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 [19-22, 57, 58]. 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. This, perhaps, is unsurprising 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 trait phenotypes.

In this paper we developed a Bayesian methodology to allow simultaneous estimation of SNP effects for host susceptibility, recoverability and infectivity from temporal epidemic data. This methodology was validated with data from simulated epidemics, which were also used to assess how different parameter values and data scenarios representing different recording schemes in field or experimental studies may affect the estimates of SNP effects and other parameters influencing transmission dynamics. The sophisticated Bayesian algorithm outlined in this paper has been implemented into a user-friendly software tool called SIRE, which allows computationally efficient analyses to be performed by anyone with relevant epidemiological data (as shown in Appendix G, outputs typically take 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 simulated data scenarios we found that

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 (perfect and complete data) and practically feasible recording schemes produced the following relevant insights: 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 as for perfect data. 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 consists 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) yields SNP effects with high precision, given sufficient data. Lastly, precise estimates of SNP effects could still be obtained with censored epidemic data.

For model validation, we chose a complex inter-dependence structure for the model parameters by assuming that the SNP under 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 an extremely challenging system in which to estimate SNP effects and in practice most real world examples are likely to be considerably less challenging as simpler structures and reduced variation/better control of variation will improve the quality of the parameter estimates.

The results from different data scenarios indicate a log-log scaling relationship with slope -½ between the precision (as measured by the SD in the posterior) of SNP effect estimates, and group size or number of groups (this relationship in analytically confirmed in a follow up paper [46]). 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 of 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 significantly more data, in particular for infectivity.

We chose a dataset comprising of 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, in particular aquaculture species [20, 59, 60]. 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 [61-63] 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 [64, 65], 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 appear to be well within reach in the near future.
It is widely recognised that disease traits are for the most part 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 [10, 62]. This is partly due to the fact that observed disease

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

phenotypes, such as individuals' binary infection status or infection time are the result of many interacting biological processes, each controlled by a different set of genes or genetic pathways and characteristics of the wider population. Hence the impact of an individual gene on the disease phenotype 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 [66]. Therefore, it is not unreasonable to assume that SNPs with moderate to large effects on these epidemiological traits, and in particular on host infectivity, may indeed exist. 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 [67]. Hence, such beneficial alleles tend to become fixed within only a 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 a few generations of selection, where a single SNP explains most of the variation in mortality of fish exposed to the IPN virus [60, 68]. In contrast, selection pressure on infectivity is expected to be relatively low, since an individual's infectivity genes affect the disease phenotype of group members rather than its own disease phenotype [34, 48, 69]. Therefore, infectivity SNPs with large effect may indeed exist, and may now be identifiable with the methods presented here.

The approach 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 [25, 30-32, 34]. The key novelty of our approach lies in its ability to estimate genetic and non-genetic effects associated with all three epidemiological host traits from a range of 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 [54, 55, 70]. 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 [71]. In contrast alleles associated with better survival in existing GWAS would only bring the expected 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 desirable for effective employment of genetic disease control. Based on the results in this paper, SIRE can readily be applied to disentangle such SNP effects using data from transmission experiments or field studies.

Furthermore, although this paper focused on estimating SNP effects, SIRE could also immediately be applied to estimating breed, age, sex, treatment or vaccination effects, or any other factor that may affect disease spread, even if genetic information is absent.

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

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

estimated (as demonstrated to some extent in this paper and also investigated for indirect genetic effects in numerous other studies [49]). In particular, theoretical studies indicate that significant improvement in estimates of infectivity effects can be achieved by appropriately grouping genetically related individuals [32, 34]. Whilst this paper focused entirely on a fixed *A* allele frequency *p* across groups, a follow up paper [46] 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 (interestingly, the susceptibility and recoverability SNP effects cannot be substantially improved in this way).

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 into 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 under consideration [72]. Hence, further model development 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 [31]. Combining both approaches may prove a useful way forward to allow estimation of genetic effects under all realistic genetic architectures and population structures.

In summary, 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.

Funding

This research was funded by 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)).

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Tables

Parameter	Accuracy	*y*-intercept	Slope	Av. SD	Description
β	0.833	0.000	1.080	0.003	Average transmission rate
γ	0.982	0.001	0.999	0.013	Average recovery rate
k	0.806	1.810	0.633	1.580	Recovery shape parameter
a_g	0.985	-0.004	1.020	0.091	SNP effect for susceptibility
a_f	0.875	-0.054	0.860	0.287	SNP effect for infectivity
a_r	0.995	-0.020	0.990	0.065	SNP effect for recoverability
Δ_g	0.910	-0.038	0.751	0.439	
Δ_f	0.335	0.065	0.133	0.530	Dominance factor (per trait)
Δ_r	0.920	-0.005	0.781	0.373	
b_{g0}	0.978	-0.012	1.000	0.105	
b_{f0}	0.871	-0.035	1.100	0.365	Fixed effect (per trait)
b_{r0}	0.992	0.008	1.000	0.073	
Σ_{gg}	0.885	0.101	0.903	0.136	
Σ_{ff}	0.691	0.264	0.563	0.203	
Σ_{rr}	0.981	0.027	1.000	0.071	Residual covariance matrix
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 of group effects

Table 1. Prediction accuracy, bias and precision for the parameter estimates. Other columns relate to the sub-plots in Fig 4 (see Fig 4 caption for information about the underlying data). Prediction accuracy is defined as the correlation between the inferred and true parameter values. The *y*-intercept and slope were obtained by fitting regression lines through the data points in Fig 4 (a *y*-intercept of zero and slope of one indicates no bias). Av. SD gives the average posterior standard deviation across all data points as an indicator for precision of parameter estimates. Subscripts *g*, *f* and *r* refer to susceptibility, infectivity and recovery, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figures

Fig 1. SIRE software.

Illustrative screenshots of the software package: (a) Different data sources can be imported by loading user defined data tables (text or cvs files), (b) prior specification can be made on parameters, (c) posterior distributions can be visualised as inference in being performed, and (d) summary statistics and MCMC diagnostics.

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

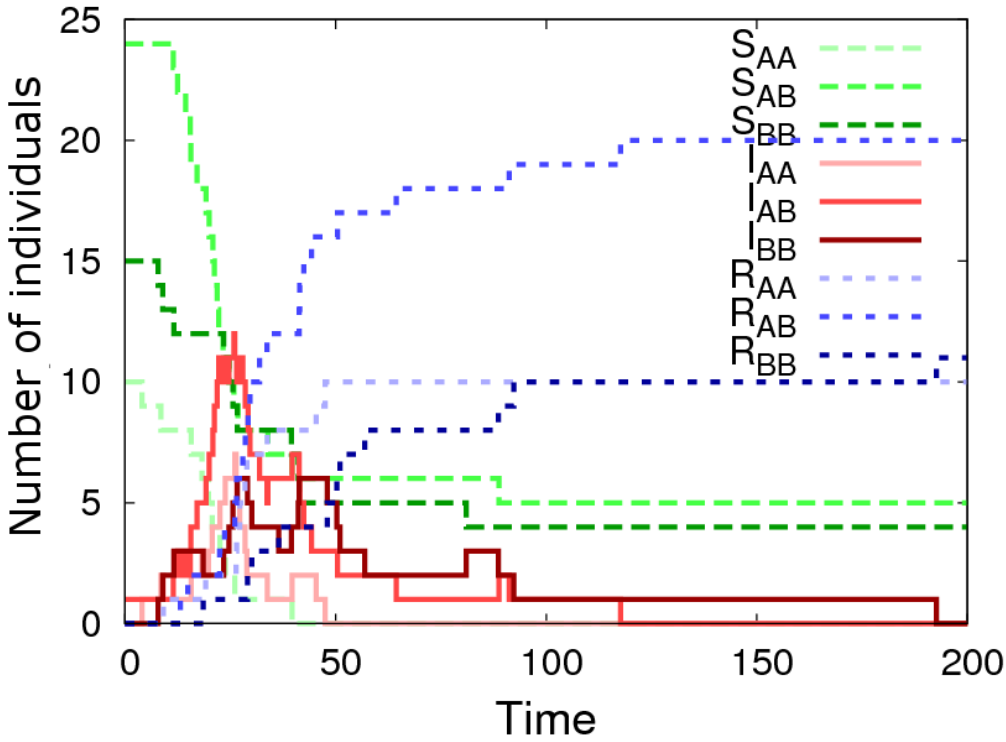

Fig 2. Simulated epidemic profiles.

This graph shows epidemic profiles for the three SNP genotypes (*i.e. AA*, *AB* or *BB*), where S_g, I_g, R_g indicate the number of susceptible, infected and recovered individuals of genotype *g*, respectively. This example is simulated using a single contact group containing G_{size}=50 individuals, of which one is initially infected. The model parameters θ are: 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 and the *A* allele has frequency *p=0.3*. Note, the step jumps in curves result from discrete disease status transitions in individuals.

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

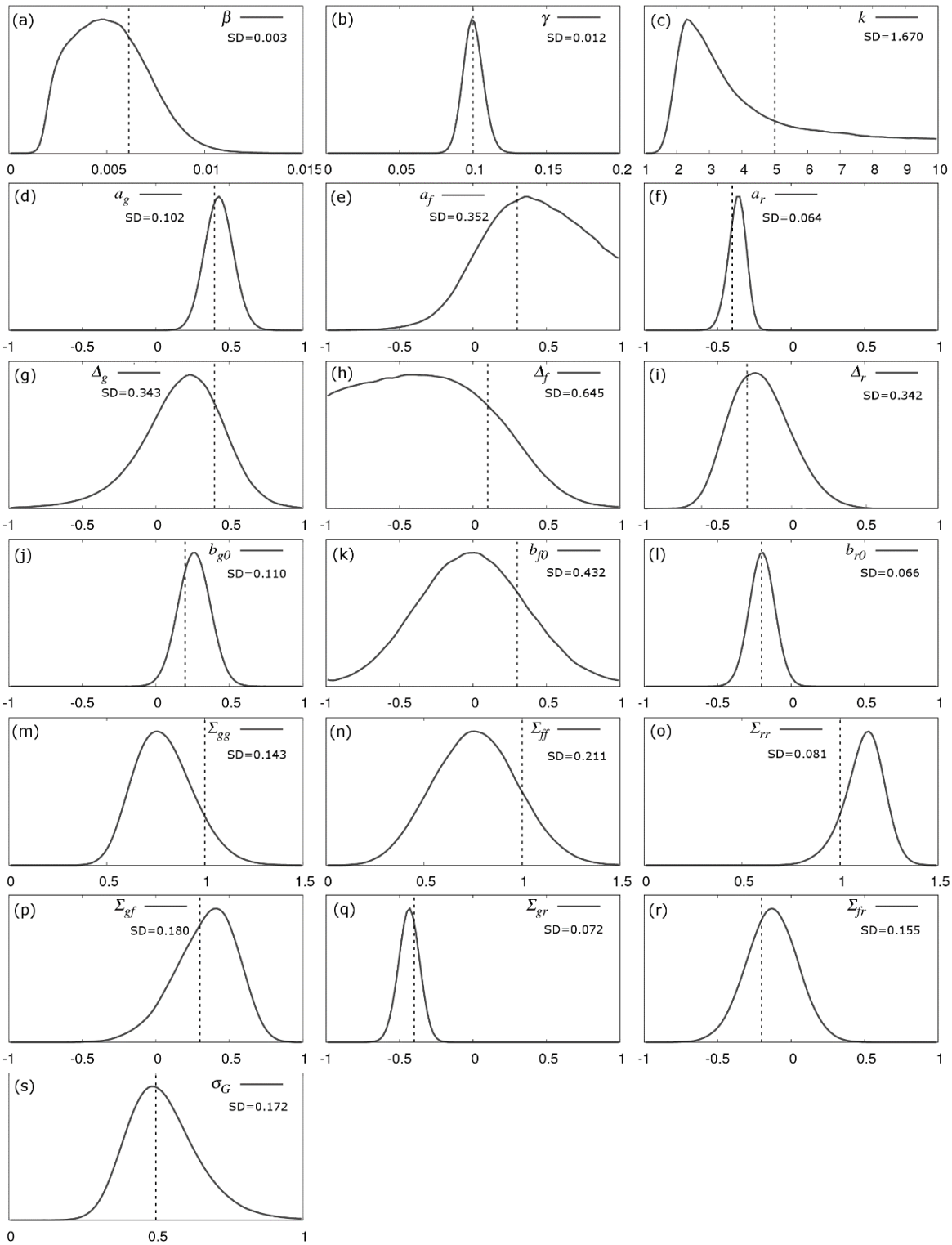

Fig 3. Parameter posterior distributions.

Probability distributions for model parameters inferred from a simulated dataset which consisted of exact infection and recovery times (DS1) 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 dashed lines). The standard deviations (SD) give a measure of precision.

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

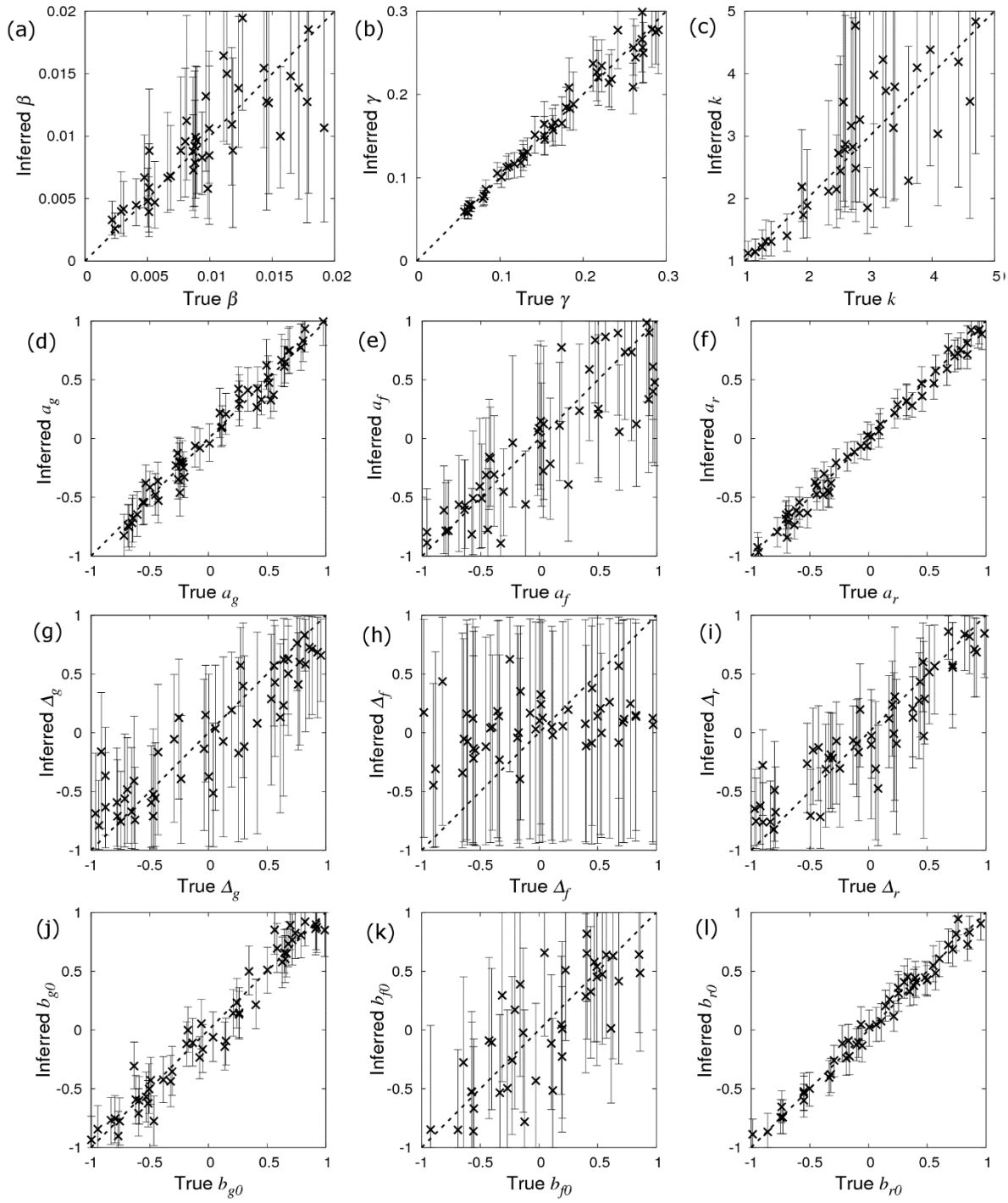

Fig 4. Prediction accuracy and bias.
The inferred posterior distributions for parameters compared to their true value. Simulated data was generated using the base parameter set in Eq.(10) except for a single parameter which was singled out in each of the sub-plots above*. Each cross corresponds to the inferred posterior mean (with error bars indicating 95% credible intervals) of the selected parameter (whose true value is on the *x*-axis) when SIRE is applied to a single simulated dataset consisting of infection and recovery times (DS1) from N_{group}=20 contact groups each containing G_{size}=50 individuals. A description of the model parameters, together with calculated prediction accuracies (correlation between true and inferred value), and bias (represented by intercept and slope of regression lines fitted to the data points), and average standard deviations are given in Table 1. (*Additionally for (g) a_g=0.4, (h) a_f=0.4 and (i) a_r=0.4, such that dominance has an effect).

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

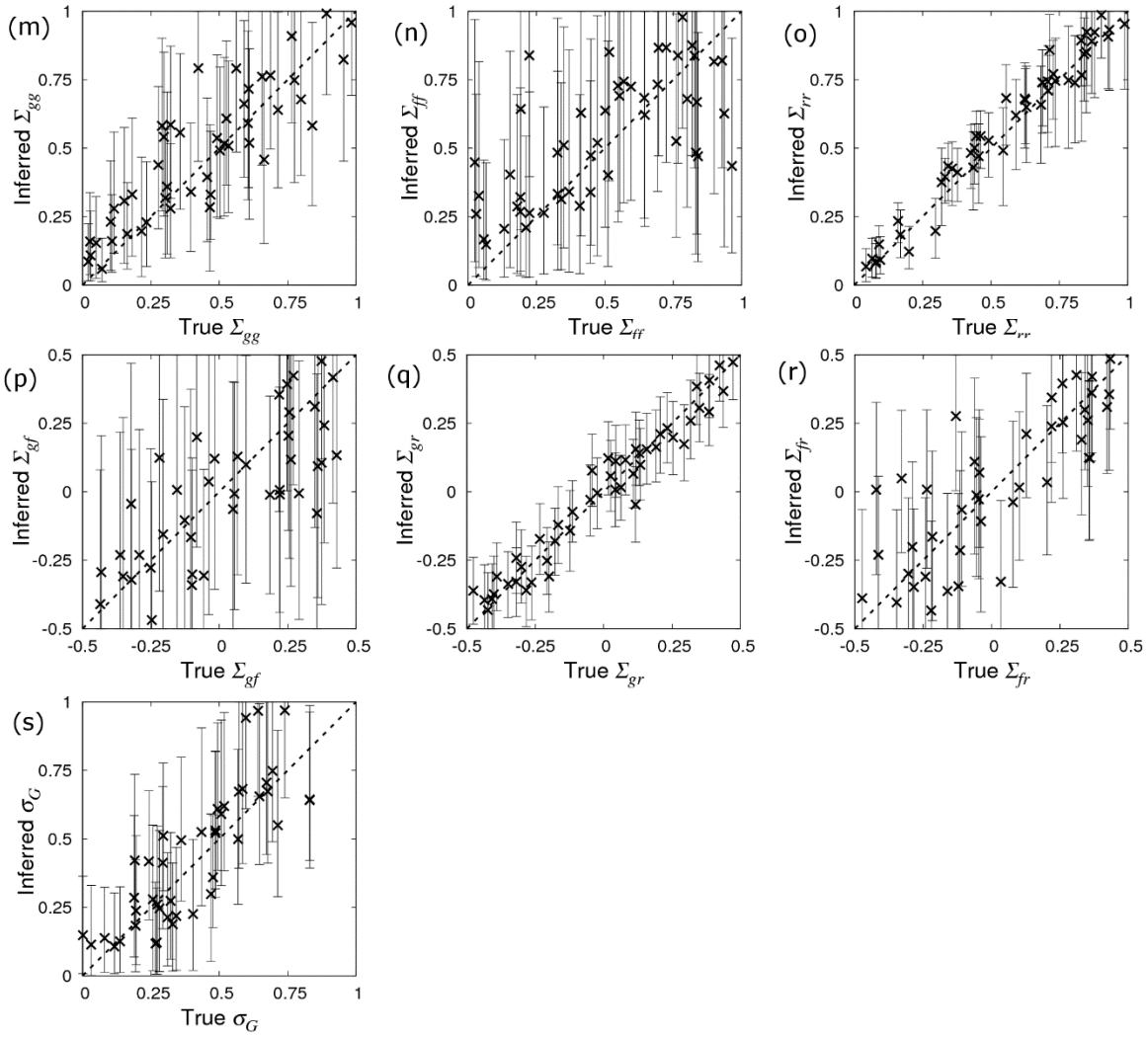

Fig 4 continued.

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

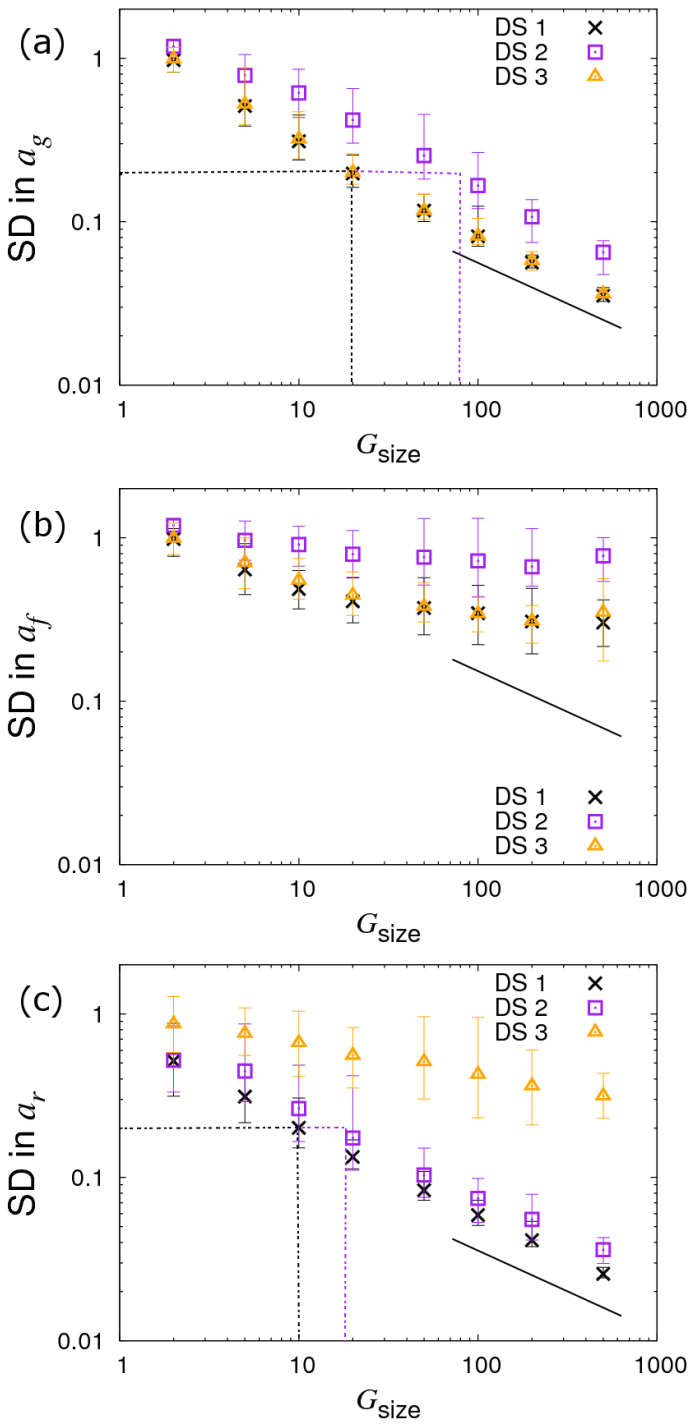

Fig 5. Variation in precision of the SNP effect estimates with group size G_{size}.
Posterior standard deviations (SDs) in SNP effects for (a) susceptibility a_g, (b) infectivity a_f and (c) recoverability a_r from simulated data with N_{group}=10 contact groups each containing G_{size} individuals (which 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 posterior SD over 50 simulated data replicates with the error bar denoting 95% of the stochastic variation about this value, *i.e.* 95% of posterior SDs lie within the interval (note, they *do not* represent posterior credible intervals, as in Fig 4). The black line indicates a slope of -½ and the dashed black and purple dash lines indicate the sample size required for identifying a SNP with effect size 0.4 for the trait under consideration (see main text for further explanation). Parameter values are given in Eq.(10).

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

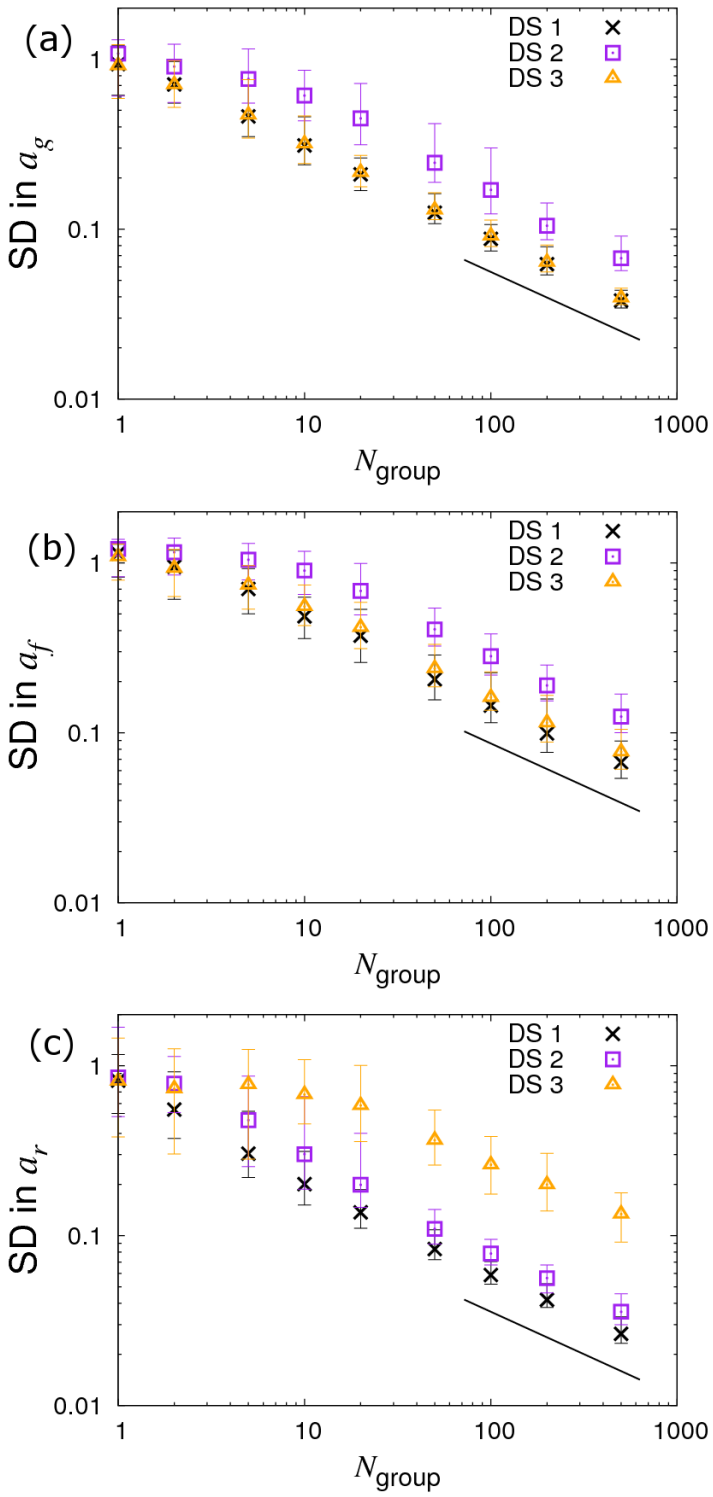

Fig 6. Variation in precision of the SNP effect estimates with number of groups N_{group}. Posterior standard deviations (SDs) in SNP effects for (a) susceptibility a_g, (b) infectivity a_f and (c) recoverability a_r from simulated data with N_{group} contact groups (which is varied) each containing 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 posterior SD over 50 simulated data replicates with the error bar denoting 95% of the stochastic variation about this value. The black line indicates a slope of -½. Parameter values are given in Eq.(10).

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

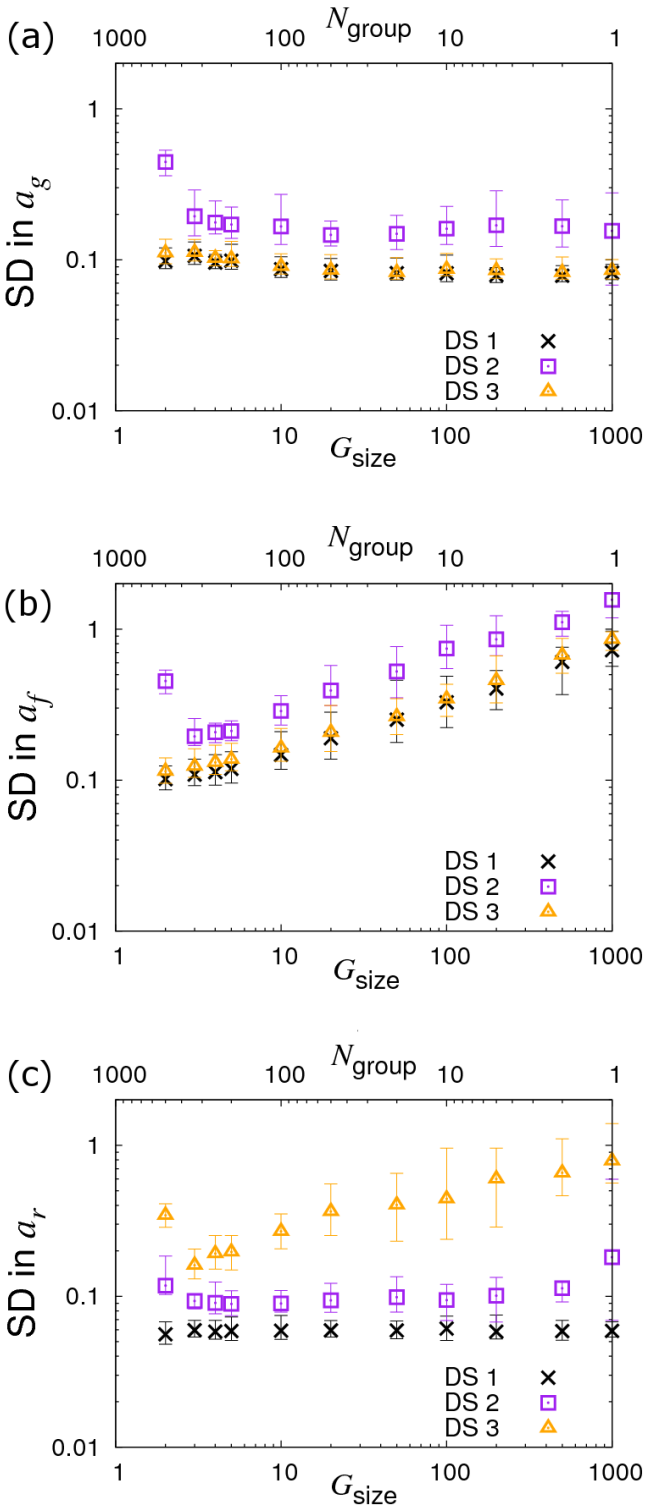

Fig 7. Variation in precision of the SNP effect estimates with partitioning into groups. Posterior standard deviations (SDs) in SNP effects for (a) susceptibility a_g, (b) infectivity a_f and (c) recoverability a_r from simulated data with N_{group} contact groups each containing G_{size} individuals, 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 posterior SD over 50 simulated data replicates with the error bar denoting 95% of the stochastic variation about this value. Parameter values are given in Eq.(10).

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

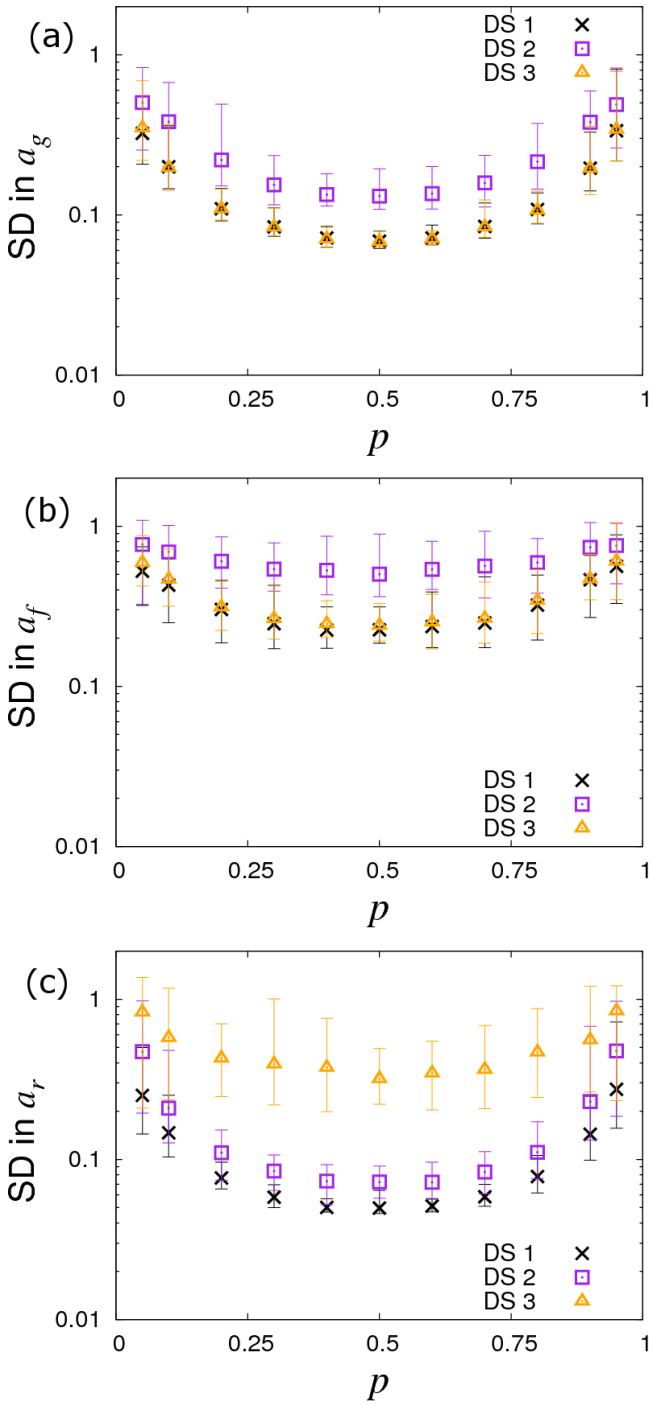

Fig 8. Variation in precision of the SNP effect estimates with allele frequency *p*. Posterior standard deviations (SDs) in SNP effects for (a) susceptibility a_g, (b) infectivity a_f and (c) recoverability a_r from simulated data with N_{group}=20 contact groups each containing G_{size}=50 individuals. 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 symbol represents the average posterior SD over 50 simulated data replicates with the error bar denoting 95% of the stochastic variation about this value**.** Parameters used are given in Eq.(10).

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

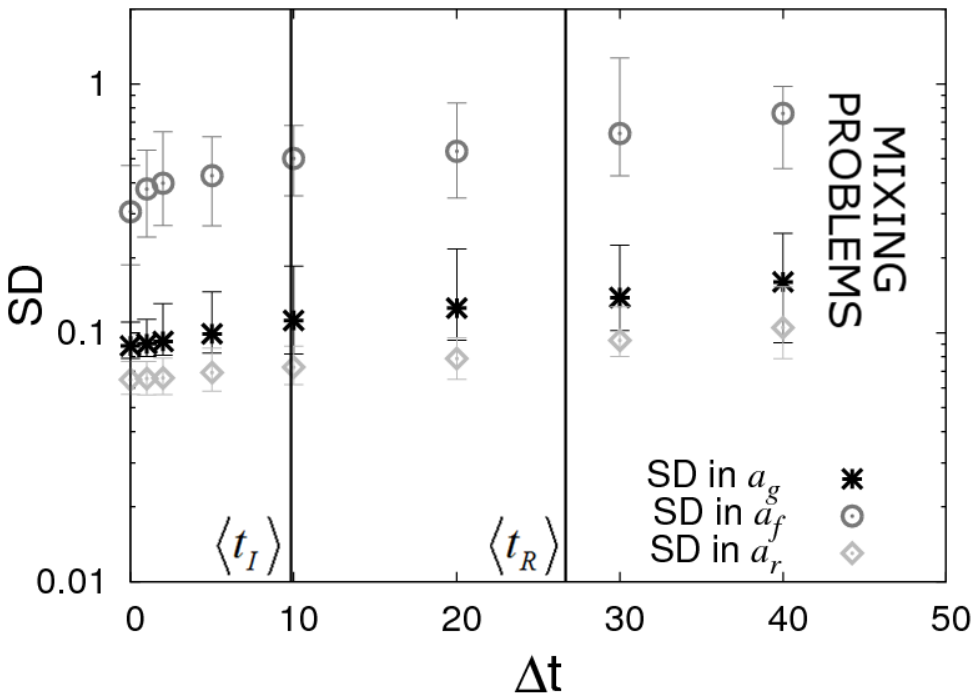

Fig 9. Periodic checking of disease status (DS4). Posterior standard deviations (SDs) in estimated SNP effects a_g, a_f and a_r from simulated data with N_{group}=20 contact groups each containing G_{size}=50 individuals. Here it is assumed that the disease status of individuals is periodically checked with time interval Δt. Each symbol represents the average posterior SD over 50 simulated data replicates with the error bar denoting 95% of the stochastic variation about this value (with the checking times randomly shifted across these replicates) with the error bar denoting stochastic variation in posterior mean. 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. Parameter values given in Eq.(10).

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

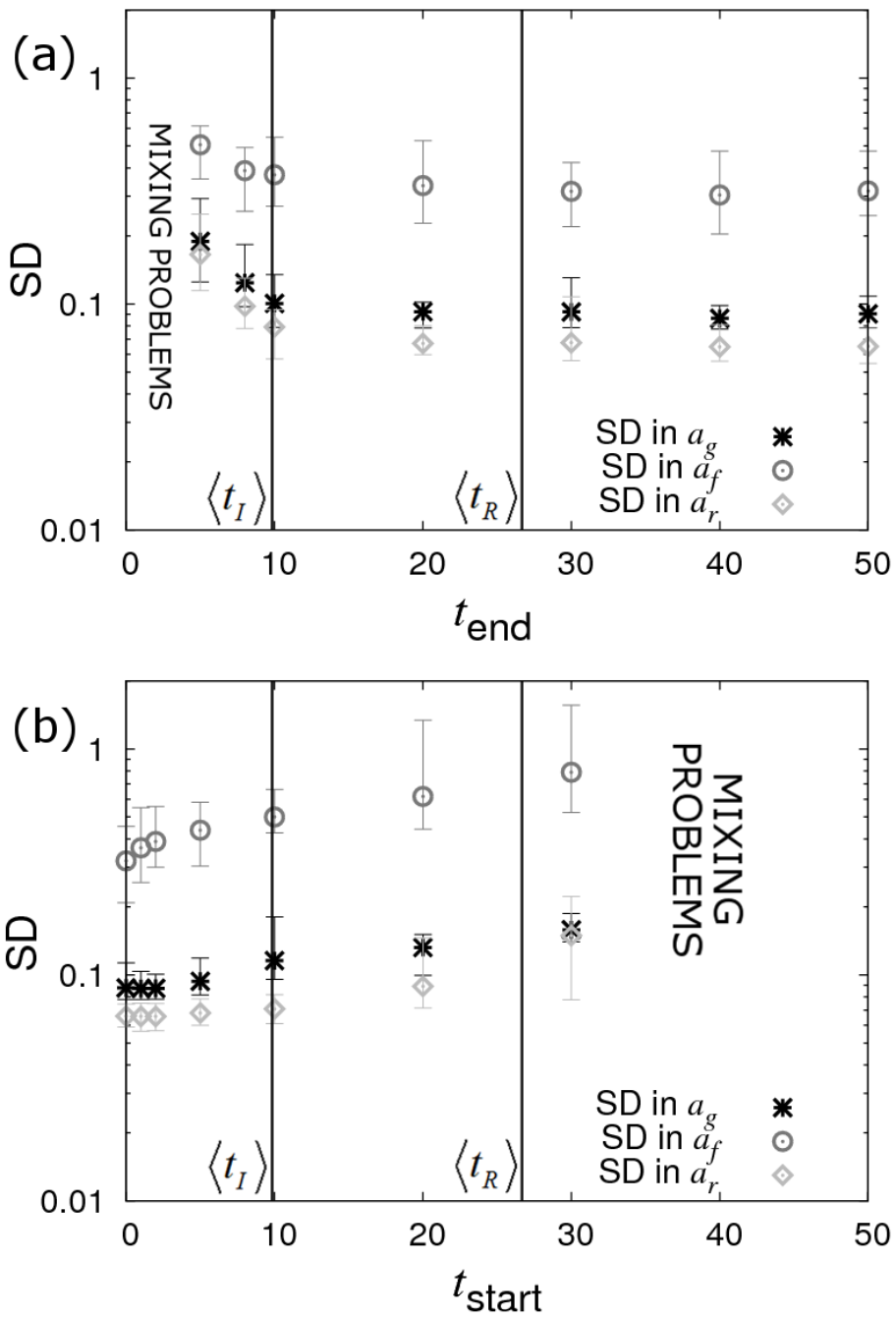

Fig 10. Censoring of data (DS5).
Posterior standard deviations (SDs) in SNP effects a_g, a_f and a_r from simulated data with N_{group}=20 contact groups each containing G_{size}=50 individuals. Each symbol represents the average posterior SD over 50 simulated data replicates with the error bar denoting 95% of the stochastic variation about this value. (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 simulations) and $\langle t_R \rangle$ the mean recovery time. Parameter values given in Eq.(10).

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Appendix A: Recovery dynamics

Fig A shows examples of gamma distributions for three different values of shape parameter *k=1, 2* and *10*. A common approach when modelling epidemics is to assume *k=1*, corresponding to the Markovian assumption of constant recovery probability. From a biological point of view, however, this is rather unrealistic. Typically after infection, the host's immune system takes some time to respond, *e.g.* to generate the appropriate antibodies to fight off the infection. Thus, for many real diseases *k* may be large.

Consequently, incorporation of a gamma distributed infection duration (which is characterised by two parameters, a mean and a shape parameter, instead of just one for the exponential distribution) allows the model to more realistically capture true disease dynamics.

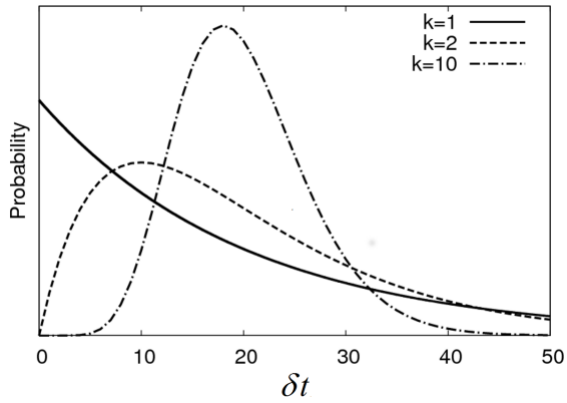

Fig A. Infection duration. Shows the probability distribution for an individual to recover from their infection a time δt after they were infected.

Appendix B: Derivation of the likelihood

The likelihood in Eq.(6) represents the probability the model in section 2.1 generates a certain set of events ξ, assuming a given set of parameters θ. This can be calculated by multiplying the probabilities for each of the individual sampling steps used in the simulation procedure, as described in Appendix F.

During initialisation of a given contact group, the infection duration for the initially infected individual *j* is sampled with probability

$$F_\Gamma(\delta t_j \mid w_j, k). \tag{B1}$$

We now consider each event *e* in turn. In step 1 of Appendix F, the inter-event time Δt is sampled from the exponential distribution $\Lambda_e e^{-\Lambda_e \Delta t}$ (where $\Lambda_e(t_e)$ is the total infection rate Eq.(7) evaluated immediately prior to t_e, the time of event *e*). Two possibilities exist for event *e*:

1. **It is a recovery event**. Considering the algorithm in Appendix F, step 2(a) would have been branched to, and this only happens if the sampled value for Δt is greater than the observed inter-event time t_e-t_{e-1}. The probability of this is calculated from the following integral:

$$\int_{t_e - t_{e-1}}^{\infty} \Lambda_e e^{-\Lambda_e(t_e)\times\Delta t} d\Delta t = e^{-\Lambda_e(t_e)\times(t_e - t_{e-1})}. \tag{B2}$$

2. **It is an infection event**. Again, considering the algorithm in Appendix F, step 2(b) would have been branched to, and this happens with probability $\Lambda_e e^{-\Lambda_e \times (t_e - t_{e-1})}$ (which comes from the exponential distribution above). The individual *j* which becomes infected is selected with probability λ_j / Λ_e. Finally, the infection duration of that individual δt_j is sampled from a gamma distribution with probability density function $F_\Gamma(\delta t_j \mid w_j, k)$. Combining these three contributions gives an overall probability:

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$$\lambda_j e^{-\Lambda_e(t_e)\times(t_e-t_{e-1})} F_\Gamma(\delta t_j \mid w_j, k). \tag{B3}$$

Multiplying the results from Eqs.(B1), (B2) and (B3) for all the infection and recovery events leads to the likelihood for a single contact group of:

$$L(\xi \mid \theta) = \left(\prod_j \lambda_j\right)\left(\prod_e e^{-\Lambda_e(t_e)\times(t_e-t_{e-1})}\right)\left(\prod_m F_\Gamma(\delta t_m \mid w_m, k)\right), \tag{B4}$$

where *j* goes over individuals that become infected (*excluding* those which initiate epidemics), *m* also goes over individuals that become infected but *including* those which initiate epidemics and *e* goes over both infection and recovery events (with corresponding event times t_e).

Since contact groups are assumed to be independent, the likelihood for multiple contact groups is simply the product of each separate one, as shown in Eq.(6).

Appendix C: Prior definitions for model parameters

The default prior in SIRE (which can be altered), which is used for all the results in this paper, is largely uninformative but does place upper and lower bounds on many of the key parameters to stop them straying into biologically unrealistic regimes. Bounding parameters in this way was found to be especially important when considering relatively uninformative data scenarios when unbounded flat priors could lead to improper posterior probability distributions. Specifically, a uniform prior between -2.3 and 2.3 was chosen for a_g. This corresponds to assuming that it is biologically unrealistic for a single SNP to change the susceptibility of individuals by more than a factor of 100 between *AA* and *BB* individuals[1]. An identical uniform prior was also placed on a_f, a_r and on each of the fixed effects in $\boldsymbol{b_g}$, $\boldsymbol{b_f}$ and $\boldsymbol{b_r}$. Similarly, a uniform prior between -3.45 and 3.45 was placed on each of the residuals $\boldsymbol{\varepsilon_g}$, $\boldsymbol{\varepsilon_f}$ and $\boldsymbol{\varepsilon_r}$, This larger range reflects a potential factor of 1000 variation across individuals (chosen to be larger as residual contributions account for all other SNPs as well as non-genetic factors, as opposed to just the effect of the single SNP under analysis).

The scaled dominance factors Δ_g, Δ_f, Δ_r, were chosen to have uniform priors between 1 and -1, *i.e.* going from complete dominance of *A* to complete dominance of *B* [2]. The prior for the shape parameter *k* was chosen to be uniform between 1 and 10, where 1 represents a Poisson random process and 10 represents a situation in which recovery times of individuals are highly concentrated around their mean.

Because parameters β and γ depend on the timescale over which the epidemic is measured[3], which is partly pathogen specific, placing informative priors on β and γ by default is not appropriate (although it can be done). Instead a uniform prior between 0 and 20 was placed on the equivalent basic reproductive ratio R_0 [4], which *is* a dimensionless quantity.

[1] This factor comes from the exponential dependency in Eq.(1) coupled with the result $e^{2\times2.3}$ =100.
[2] If over dominance is considered a possibility, this prior distribution would be extended.
[3] E.*g.* if measurements are made in hours then β and γ would be very different to if they were made in days.
[4] Defined by $R_0=\beta(\langle N\rangle-1)/\gamma$, where ⟨N⟩ is the average population size of contact groups as taken from the data, this represents the number of infections one typically infectious individual generates on average over the course of its infectious period in an otherwise uninfected population.

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Appendix D: MCMC procedure

Markov chain Monte Carlo (MCMC) produces a list of (correlated) parameter θ^q and event time ξ^q samples drawn from the posterior probability distribution in Eq.(5), where q indexes sample number.

1) Initialisation – MCMC is initially started with some set of parameters θ^1 and event times ξ^1 consistent with the data y. In the case of parameters, θ^1 is simply sampled from the prior distribution $\pi(\theta)$. For events, initialisation depends on the available data: DS1) events are defined by the data, DS2) infection times are exponential sampled backwards in time from the first observed recovery time[5], DS3) recovery times are exponentially sampled forwards in time from the final observed infection time[6], DS4) infection and recovery times are sampled uniformly in the time intervals identified between successive diagnostic tests, and DS5) events within the observation period are defined by the data, infection events prior to this time period are exponentially sampled backward in time and similarly recoveries events after this time period are exponentially sampled forward in time.

2**) Iteration** - MCMC operates by proposing changes to the model parameters θ and event times ξ and accepting or rejecting these changes in accordance with a Metropolis-Hastings probability. In this way the underlying dynamics are able to explore all potential possibilities consistent with the observations.

A single MCMC "update" consists of making the following sets of proposals:

MCMC UPDATE

Parameters – Each individual parameter in $\theta=(\beta, \gamma, k, a_g, a_f, a_r, \Delta_g, \Delta_f, \Delta_r, \boldsymbol{b_g}, \boldsymbol{b_f}, \boldsymbol{b_r}, \boldsymbol{\varepsilon_g}, \boldsymbol{\varepsilon_f}, \boldsymbol{\varepsilon_r}, \boldsymbol{\Sigma}, \boldsymbol{G}, \sigma_G)$, denoted by θ_j, is considered in turn. A proposed value is drawn from a normal distribution centred on the parameter's current chain value

$$\theta_j^p \sim Norm(\theta_j^q, J_j^2), \tag{C1}$$

with all other parameters in θ^p remaining the same as in θ^q (if this produces an inconsistent value, *e.g.* β becomes negative, the proposal is immediately rejected). The proposal is accepted with Metropolis-Hastings probability

$$P_{MH} = \min\left(1, \frac{L(\xi^p|\theta^p)\pi(\theta^p)}{L(\xi^q|\theta^q)\pi(\theta^q)}\right). \tag{C2}$$

If accepted, we set $\theta^{q+1}= \theta^p$ else $\theta^{q+1}= \theta^q$ with $\xi^{q+1}= \xi^q$.

Tuning J_j in Eq.(C1) is important. If it is too large, very few proposals will be accepted and if too small, mixing will be slow. Motivated by adaptive MCMC [73, 74], a robust heuristic method for optimising J_j within the burn-in period is as follows. Initially, J_j is set to a small quantity. Each time a proposed change on parameter j is accepted J_j is updated according to

$$J_j^{new} = J_j \times 1.01, \tag{C3}$$

and when rejected

$$J_j^{new} = J_j \times 0.995. \tag{C4}$$

[5] The standard deviation in recovery times is used to set the decay rate.

[6] The standard deviation in infection times is used to set the decay rate.

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

These numerical factors are chosen for two reasons: Firstly, the updates in Eqs.(C3) and (C4) balance each other out when acceptance occurs around 33% of the time, leading to a steady state solution for J_j. Secondly, they are chosen to be sufficiently close to *1* to prevent large fluctuations in J_j, but sufficiently far to allow the steady state solution to be found within the burn-in period.

In addition to the single parameter proposals outlined above, joint proposals are also made on residuals along with their corresponding covariance matrix, as described in Appendix E. The reason these joint updates are necessary is that $\boldsymbol{\Sigma}$ and $\boldsymbol{\varepsilon}$ are often highly correlated within the model (as determined by Eq.(9)). In many cases the data provides little information regarding $\boldsymbol{\Sigma}$ itself, so these correlations can lead to extremely slow mixing.

Events – In situations in which ξ is not precisely known (*i.e.* data scenarios other than DS1), the unknown latent event times must be stochastically changed in accordance with the posterior distribution. Each individual *j* is considered in turn. In the case of proposing changes to the infection time of *j*, the following normal distribution is sampled from:

$$t_j^{\mathrm{I},p} \sim Norm(t_j^{\mathrm{I},q}, J_{I,j}^2), \tag{C5}$$

with all other event times in $\xi^p=(t^{\mathrm{I}},t^{\mathrm{R}})$ remaining the same as in ξ^q. If this proposed infection time exceeds the recovery time $t_j^{R,q}$ the proposal is immediately rejected (infections cannot occur after recoveries). The proposal in Eq.(C5) is accepted with probability

$$P_{MH} = \min\left(1, \frac{\pi\left(y|\xi^p\right)L(\xi^p|\theta^p)}{\pi\left(y|\xi^q\right)L(\xi^q|\theta^q)}\right). \tag{C6}$$

If accepted, we set $\xi^{q+1}= \xi^p$ else $\xi^{q+1}= \xi^q$ with $\theta^{q+1}= \theta^q$.

A similar proposal to Eq.(C5) is used to change recovery times. The jumping parameters $J_{I,j}$ and $J_{R,j}$ are again tuned to give an average acceptance probability of around 33% (using the same method as in Eqs.(C3) and (C4) above). Furthermore, there are data scenarios in which it is not possible to tell whether an individual has become infected or not. In these cases it is necessary to include additional proposals which insert infection/recovery event pairs and, conversely, other proposals which remove them.

Each of the proposals above have been optimised to calculate only those parts of the likelihood that change, *e.g.* if the infection time of individual *j* is altered then only the change in likelihood in the period between then initial and proposed times within *j*'s contact group needs to be calculated. Furthermore, when making some sets of proposals computational speed is substantially increased by pre-calculating terms in the likelihood which remain unchanged.

Appendix E: Posterior-based proposals for residuals

This appendix describes special joint proposals in the residuals ($\boldsymbol{\Sigma},\boldsymbol{\varepsilon}$) which are used to aid MCMC mixing. Specifically three type of proposal are considered: 1) those which randomly make changes to Σ_{gg}, Σ_{gf}, or Σ_{gr} and stochastically alter $\boldsymbol{\varepsilon}_{\mathbf{g}}$ (with everything else kept fixed), 2) those which randomly make changes to Σ_{ff}, Σ_{gf}, or Σ_{rr} and stochastically alter $\boldsymbol{\varepsilon}_f$ (with everything else kept fixed), and 3) those which randomly makes changes to Σ_{rr}, Σ_{gr}, or Σ_{fr} and stochastically alter $\boldsymbol{\varepsilon}_r$ (with everything else kept fixed). Here we describe just one of these possibilities, but others can be found by suitably permuting indices.

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

So-called "posterior-based proposals" (PBPs) [1] consist of three steps:

Step 1 – A proposed is made to one of the parameters in the model. In this example

$$\Sigma_{gg}^{p} \sim Norm(\Sigma_{gg}^{q}, J_{\Sigma_{gg}}^{2}), \tag{C7}$$

with the value of $J_{\Sigma_{gg}}$ tuned to give an average acceptance probability of around 33% (using the same method as in Eqs.(C3) and (C4) above).

Step 2 – Each element $\varepsilon_{g,j}$ in the $\boldsymbol{\varepsilon}_{\mathbf{g}}$ is considered in turn. The values of means $\mu_{q,j}$ and $\mu_{p,j}$ and standard deviations $\sigma_{q,j}$ and $\sigma_{p,j}$ are calculated such that the normal distributions they characterise approximate, to some level of accuracy, the true posterior distributions for $\varepsilon_{g,j}$ in the initial and proposed states (details on how this is done are discussed below). If $\sigma_{p,j}>\sigma_{q,j}$ then for each individual *j* the proposed residuals are sampled using

$$\varepsilon_{g,j}^{p} \sim Norm\left(\mu_{p,j} + \sqrt{\kappa + (1-\kappa)\frac{\sigma_{p,j}^{2}}{\sigma_{q,j}^{2}}}(\varepsilon_{g,j}^{q} - \mu_{q,j}), \kappa(\sigma_{p,j}^{2} - \sigma_{q,j}^{2})\right), \tag{C8}$$

where κ is a tuneable constant set to 0.03, and when $\sigma_p \leq \sigma_q$

$$\varepsilon_{g,j}^{p} \sim Norm\left(\mu_{p,j} + \sqrt{\kappa + (1-\kappa)\frac{\sigma_{q,j}^{2}}{\sigma_{p,j}^{2}}}\frac{\sigma_{p,j}^{2}}{\sigma_{q,j}^{2}}(\varepsilon_{g,j}^{q} - \mu_{q,j}), \kappa\frac{\sigma_{p,j}^{2}}{\sigma_{q,j}^{2}}(\sigma_{q,j}^{2} - \sigma_{p,j}^{2})\right). \tag{C9}$$

The somewhat complicated expressions in Eqs.(C8) and (C9) are specifically designed such that if the posterior *is* truly represented by the normal approximations (as characterised by $\mu_{q,j}$, $\mu_{p,j}$, $\sigma_{q,j}$ and $\sigma_{p,j}$), the proposal is accepted with probability 1.

Step 3 – The proposed combination $\Sigma_{gg}^{p}, \boldsymbol{\varepsilon}_{g}^{p}$ is accepted or rejected with Metropolis-Hastings probability

$$P_{MH} = \min\left\{1, \frac{\pi(y|\xi^{p})\pi(\xi^{p}|\theta^{p})\pi(\theta^{p})}{\pi(y|\xi^{q})\pi(\xi^{q}|\theta^{q})\pi(\theta^{q})}\prod_{j}\frac{f_{norm}(\varepsilon_{g,j}^{q}|\mu_{q,j},\sigma_{q,j})}{f_{norm}(\varepsilon_{g,j}^{p}|\mu_{p,j},\sigma_{p,j})}\right\}, \tag{C10}$$

where *j* goes over all individuals and

$$f_{\text{norm}}(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^{2}}} e^{-\frac{(x-\mu)^{2}}{2\sigma^{2}}} \tag{C11}$$

is the Gaussian probability density function.

Optimisation

For efficient implementation of this approach it is necessary to generate $\mu_{q,j}$, $\mu_{p,j}$, $\sigma_{q,j}$ and $\sigma_{p,j}$ in a suitable manner. One approximation is to set them to values implied by prior in Eq.(9):

$$\begin{aligned} \mu_{q,j} &= -(M_{gf}^{q}\varepsilon_{f,j} + M_{gr}^{q}\varepsilon_{r,j})/M_{gg}^{q}, \qquad \sigma_{q,j} = 1/M_{gg}^{q}, \\ \mu_{\mathrm{p},j} &= -(M_{gf}^{p}\varepsilon_{f,j} + M_{gr}^{p}\varepsilon_{r,j})/M_{gg}^{p}, \qquad \sigma_{p,j} = 1/M_{gg}^{p}, \end{aligned} \tag{C12}$$

where **M** is the inverse of the covariance matrix $\boldsymbol{\Sigma}^{-1}$. This approximation leads to so-called "model-based proposals" (MBP). A more accurate approximation (because it takes into account the observed events, which themselves are informed by the data) is given by

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$$\mu_q = -(M^q_{gf}\varepsilon_{f,j} + M^q_{gr}\varepsilon_{r,j})/M^q_{gg} + \left.\frac{\partial \log\left(L(\xi^q \mid \theta^q)\right)}{\partial \varepsilon^q_{g,j}}\right|_{\varepsilon^q_{g,j}=0} \sigma^2_{q,j}, \qquad \sigma_{q,j} = 1/M^q_{gg},$$

$$\mu_p = -(M^p_{gf}\varepsilon_{f,j} + M^p_{gr}\varepsilon_{r,j})/M^p_{gg} + \left.\frac{\partial \log\left(L(\xi^p \mid \theta^p)\right)}{\partial \varepsilon^p_{g,j}}\right|_{\varepsilon^p_{g,j}=0} \sigma^2_{p,j}, \qquad \sigma_{p,j} = 1/M^p_{gg},, \tag{C13}$$

and this gives a first order "posterior-based proposal" (PBP). Although Eq.(C13) is computationally slower to calculate than Eq.(C12) (because it contains gradients in the log-likelihood), it can lead to larger jumps in parameter space compared to a corresponding MBP resulting in improved mixing. Empirically it was found that applying PBPs to parameters relating to susceptibility and recoverability and MBPs to parameters relating to infectivity was the fastest approach to take.

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Appendix F: Simulation

The Doob-Gillespie algorithm [43] provides a means of taking into account inherent stochasticity in Markovian compartmental models (*i.e.* models for which the transition rates depend solely on the current state of the system). The model used in this paper combines Markovian infection transitions with more realistic non-Markovian recovery dynamics. Below we describe how these recovery events are incorporated into the standard Doob-Gillespie framework.

The purpose of this procedure is to build up a time ordered sequence of infection and recovery event times indexed by event number *e*. The following notation is used: t_e is the event time, x_e is the event type (infection *in.* or recovery *re.*), j_e is the affected individual, and t_j^I and t_j^R are the infection and recovery times for individual *j*, respectively.

Initialization: Each epidemic is assumed to be started by one (or potentially more) initially infected individual *j* at some initial time point t_{init}. The infection duration δt_j for this individual is drawn from a gamma distribution parameterised in terms of a mean and shape parameter:

$$\delta t_j \sim Gamma(w_j, k) \tag{D1}$$

(note, the dependency of w_j on θ is given through Eqs. (2) and (3)). This allows us to set $t_j^I = t_{init}$ and $t_j^R = t_{init} + \delta t_n$. Individual *j* is then placed onto a list $\mathcal{R}$, which represents all currently infected individuals. We set event index to *e=1*.

Step 1: Calculate the time to the next infection event. This is done by first evaluating the total transition rate that any individual becomes infected

$$\Lambda = \sum_s \lambda_s, \tag{D2}$$

where the sum *s* goes over all currently susceptible individuals and the force of infection λ_s (which gives the probability per unit time of *s* becoming infected) is given by Eq.(1). In accordance with a Poisson process, the time to the next infection event is generated by drawing a sample from the exponential distribution $\Lambda e^{-\Lambda \Delta t}$. In practice, this is achieved by selecting an inter-event time using

$$\Delta t = -\frac{\log(u)}{\Lambda}, \tag{D3}$$

where *u* is a (uniform) randomly generated number between *0* and *1*. The new event time is then defined by

$$t^{new} = t_{e-1} + \Delta t. \tag{D4}$$

Step 2: Choosing the event type. For the SIR model two possibilities exist:

a) If t^{new} is greater than the smallest recovery time of all the individuals in $\mathcal{R}$, which we label j_{min}, then we remove j_{min} from $\mathcal{R}$ and set

$$t_e = t_{j_{\min}}^R, \;\; x_e = re., \;\; j_e = j_{\min}. \tag{D5}$$

b) Otherwise, we set

$$t_e = t^{new}, \;\; x_e = in., \tag{D6}$$

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and select the actual individual that becomes infected with probability

$$\mathrm{Prob}(j_e = s) = \frac{\lambda_s}{\Lambda}. \tag{D7}$$

The infection duration δt_{j_e} for j_e is sampled using Eq.(D1), and the infection and recovery times are set to

$$\begin{aligned} t^I_{j_e} &= t_e, \\ t^R_{j_e} &= t_e + \delta t_{n_e}. \end{aligned} \tag{D8}$$

Individual j_e is then placed onto the list $\mathcal{R}$.

Step 3: Increment *e* and jump to step 1 if there are any remaining infected individuals.

End: Insert recovery times for any remaining individuals *j* in $\mathcal{R}$

$$t_e = t^R_j, \;\; x_e = re., \;\; j_e = j, \tag{D9}$$

incrementing *e* after each addition.

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Appendix G: Computational speed estimate for SIRE

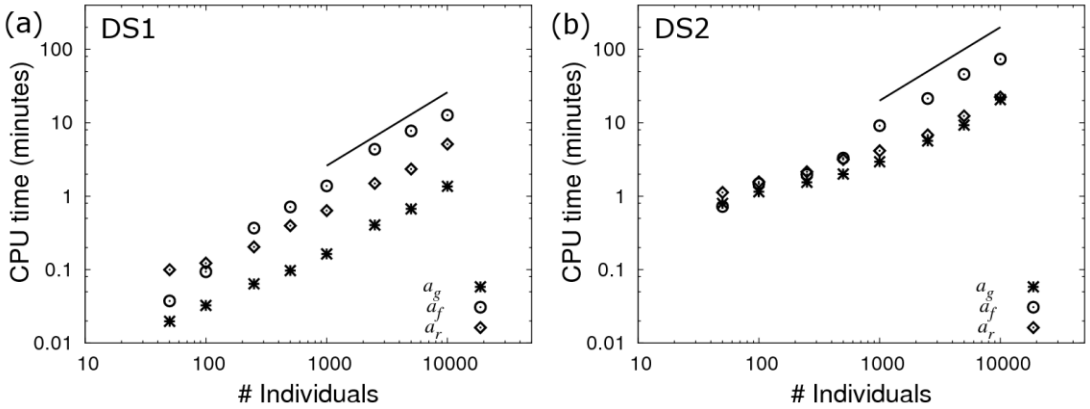

Fig G. SIRE speed. The CPU time taken for SIRE to generate 100 independent posterior samples for the SNP effects as a function of the number of individuals (where the group size is taken to G_{size}=50), as estimated using the effective sample size [1]. Simulated data was generated using the base parameter set in Eq.(10) with (a) known infection and recovery times (DS1) and (b) known recovery times (DS2).

Fig G shows the CPU time SIRE takes to estimate the SNP effects as a function of the total number of individuals (this is based on a single 2GHz core). We find that the SNP effect associated with infectivity takes the longest to accurately estimate. The approximate linear scaling (represented by the solid black lines) means that SIRE is expected to take around one minute per 1000 individuals to generate 100 representative samples from the posterior under DS1 and around 10 minutes for DS2.

Appendix H: Parameter prediction accuracy under DS2

Fig H shows results when repeating the analysis is section 3.2 of the paper, but this time taking the scenario in which infection times are unknown (*i.e.* DS2). The corresponding regression analysis on these points is presented in Table H. The average SDs in Table H are smaller than those in Table 1, reflecting the reduction in statistical power under this data scenario. A factor of two difference in SD corresponds to four times as many individuals required for equivalent accuracy.

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

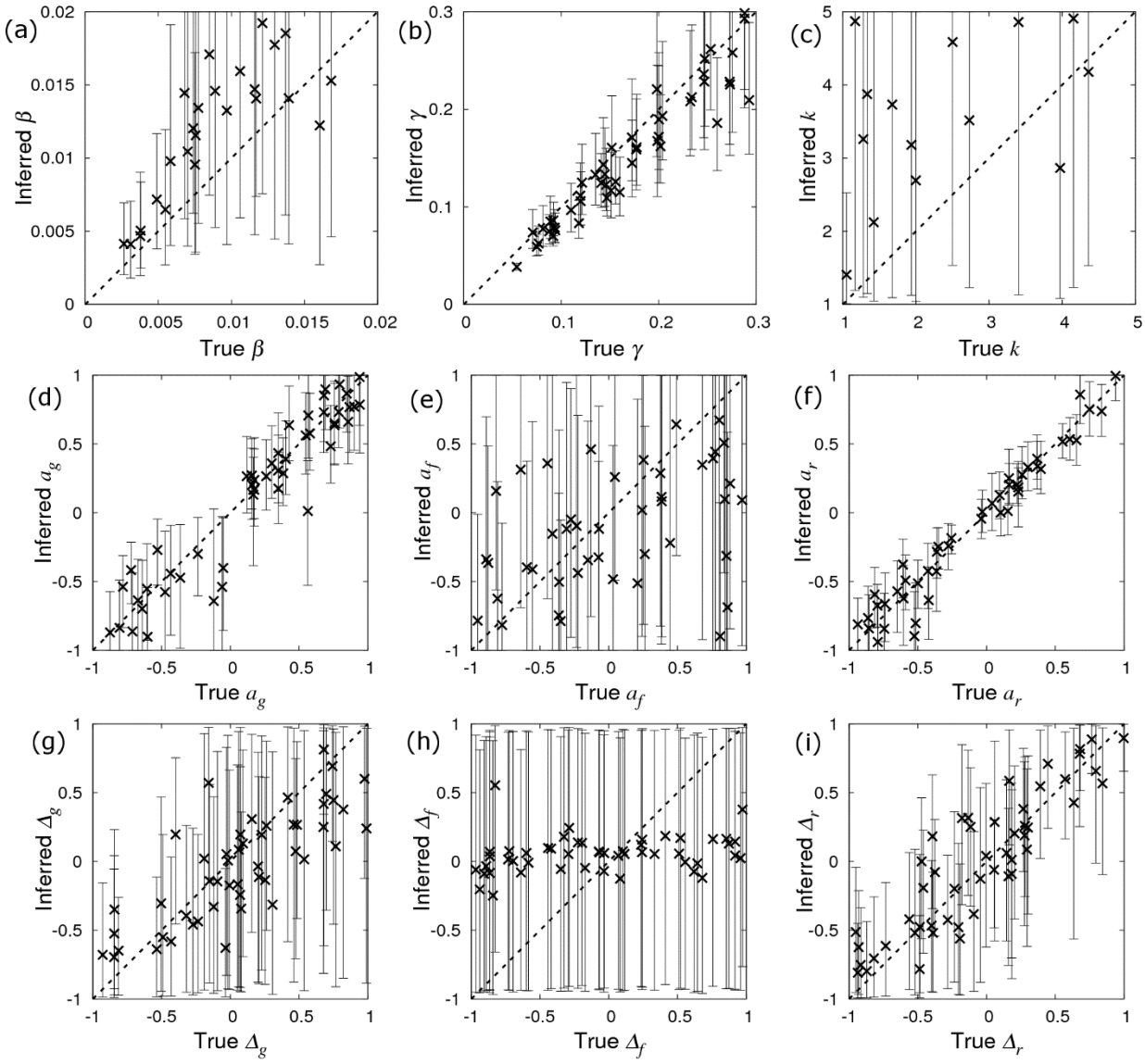

Fig H. Prediction accuracy and bias. These plots summarise inferred posterior distributions for parameters compared to their true value. Simulated data was generated using the base parameter set in Eq.(10) except for a single parameter which was singled out in each of the sub-plots above*. Crosses correspond to the inferred posterior mean (with error bars indicating 95% credible intervals) of the selected parameter (whose true value is on the x-axis) when SIRE is applied to a single simulated data set consisting of recovery times (*i.e.* DS2) from N_{group}=20 contact groups each containing G_{size}=50 individuals. Prediction accuracies, and the intercept and slope of regression lines fitted to the data points are given in Table 1. (*Additionally for (g) a_g=0.4, (h) a_f=0.4 and (i) a_r=0.4, such that dominance has an effect).

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

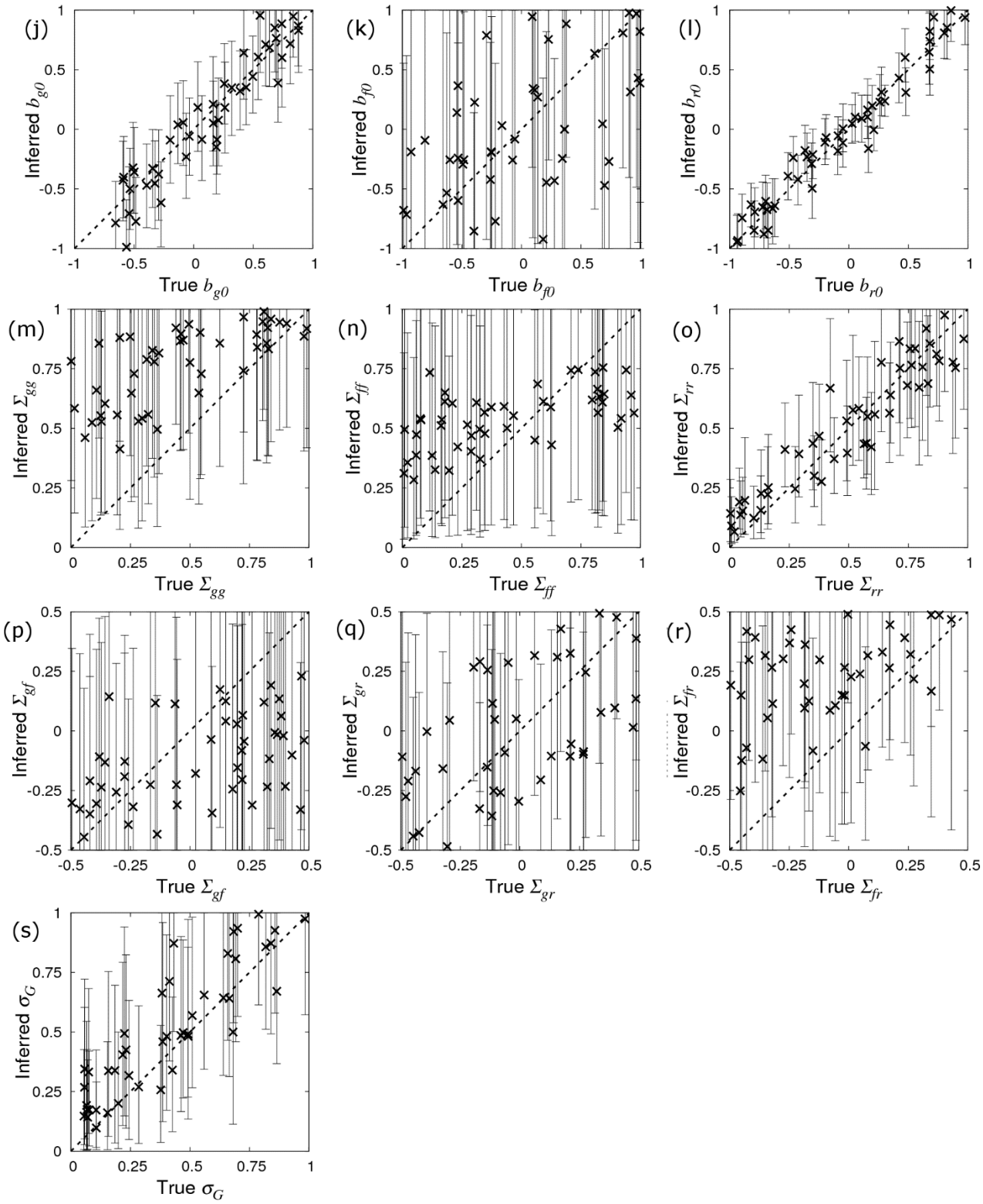

Fig H continued.

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Parameter	Accuracy	*y*-intercept	Slope	Av. SD	Description
β	0.831	0.003	1.191	0.005	Average transmission rate
γ	0.956	-0.008	0.946	0.024	Average recovery rate
k	0.494	3.454	0.516	2.272	Recovery shape parameter
a_g	0.950	-0.004	0.995	0.173	SNP effect for susceptibility
a_f	0.355	-0.019	0.370	0.525	SNP effect for infectivity
a_r	0.979	-0.006	1.022	0.111	SNP effect for recoverability
Δ_g	0.791	-0.079	0.619	0.387	
Δ_f	0.260	0.056	0.054	0.544	Dominance factor (per trait)
Δ_r	0.891	-0.040	0.857	0.249	
b_{g0}	0.961	-0.033	1.083	0.188	
b_{f0}	0.411	-0.004	0.469	0.647	Fixed effect (per trait)
b_{r0}	0.978	0.011	0.980	0.111	
Σ_{gg}	0.723	0.581	0.416	0.236	
Σ_{ff}	0.503	-0.141	0.312	0.330	
Σ_{rr}	0.574	0.014	0.655	0.296	Residual covariance matrix
Σ_{gf}	0.588	0.440	0.233	0.266	
Σ_{gr}	0.537	0.292	0.382	0.248	
Σ_{fr}	0.941	0.101	0.817	0.113	
σ_G	0.906	0.131	0.936	0.182	SD of group effects

Table H. Prediction accuracy and bias under DS2. Prediction accuracy is defined as the correlation between the inferred and true parameter values (a value of one implies perfect inference). The *y*-intercept and slope are taken from regression lines fitted through the data point (a *y*-intercept of zero and slope of one indicates no bias). Av. SD gives the average posterior standard deviation across all datasets.

References

1. Pooley C, Bishop S, Doeschl-Wilson A, Marion G. Posterior-based proposals for speeding up Markov chain Monte Carlo. arXiv preprint arXiv:190310221. 2019.
2. Enticott G. The spaces of biosecurity: prescribing and negotiating solutions to bovine tuberculosis. Environment and Planning A. 2008;40(7):1568-82.
3. Waage J, Mumford J. Agricultural biosecurity. Philosophical Transactions of the Royal Society B: Biological Sciences. 2007;363(1492):863-76.
4. Organization WH. WHO global strategy for containment of antimicrobial resistance. 2001.
5. Organization WH. Antimicrobial resistance: 2014 global report on surveillance: World Health Organization; 2014.
6. Sheldon J, Soriano V. Hepatitis B virus escape mutants induced by antiviral therapy. Journal of antimicrobial chemotherapy. 2008;61:766-8.
7. Gandon S, Day T. Evidences of parasite evolution after vaccination. Vaccine. 2008;26:C4-C7.
8. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 years of GWAS discovery: biology, function, and translation. The American Journal of Human Genetics. 2017;101(1):5-22.

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

9. Heffner EL, Sorrells ME, Jannink J-L. Genomic selection for crop improvement. Crop Science. 2009;49(1):1-12.
10. Goddard M, Hayes B. Genomic selection. Journal of Animal breeding and Genetics. 2007;124(6):323-30.
11. Bishop SC, Axford RF, Nicholas FW, Owen JB. Breeding for disease resistance in farm animals: CABI; 2010.
12. Anderson RM, MAY RM. Spatial, temporal, and genetic heterogeneity in host populations and the design of immunization programmes. Mathematical Medicine and Biology. 1984;1:233-66.
13. Hethcote HW, Van Ark JW. Epidemiological models for heterogeneous populations: proportionate mixing, parameter estimation, and immunization programs. Math Biosci. 1987;84:85-118.
14. 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;86:1747-57.
15. 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;188:683-93.
16. 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.
17. Lively CM. The effect of host genetic diversity on disease spread. The American Naturalist. 2010;175:E149-E52.
18. Tsairidou S, Anacleto O, Woolliams J, Doeschl-Wilson A. Enhancing genetic disease control by selecting for lower host infectivity and susceptibility. Heredity. 2019:1.
19. 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.
20. 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;9:4924.
21. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature. 2005;438:355-9.
22. 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;18:398-401.
23. 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;281:20140248.
24. Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. Plos Comput Biol. 2012;8. PubMed PMID: ISI:000312901500028.
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. PubMed PMID: ISI:000333517400001.
26. Gibson GJ, Renshaw E. Estimating parameters in stochastic compartmental models using Markov chain methods. Ima J Math Appl Med. 1998;15:19-40. PubMed PMID: ISI:000072720200002.
27. O'Neill PD, Roberts GO. Bayesian inference for partially observed stochastic epidemics. J R Statist Soc A. 1999;162:121-9. PubMed PMID: ISI:000078538300009.
28. Biemans F. Transmission of digital dermatitis in dairy cattle: population dtynamics and host quantitative genetics. Wageningen University, PhD Thesis, Chapter 4. 2018.
29. 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;44:10.

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

30. 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;100:4706-20.
31. 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;201:871-84.
32. Anche MT, Bijma P, De Jong MCM. Genetic analysis of infectious diseases: estimating gene effects for susceptibility and infectivity. Genet Sel Evol. 2015;47. PubMed PMID: ISI:000364188700001.
33. 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;113:364-74.
34. 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;49:53.
35. Karolemeas K, McKinley T, Clifton-Hadley R, Goodchild A, Mitchell A, Johnston W, et al. Recurrence of bovine tuberculosis breakdowns in Great Britain: risk factors and prediction. Preventive veterinary medicine. 2011;102(1):22-9.
36. Keeling MJ, Rohani P. Modeling infectious diseases in humans and animals. Princeton: Princeton University Press; 2008.
37. Lynch M, Walsh B. Genetics and analysis of quantitative traits: Sinauer Sunderland, MA; 1998.
38. Falconer D, Mackay T. Introduction to quantitative genetics. Essex. UK: Longman Group. 1996.
39. For computational convenience the recovery event times after the observation period are also included within ξ.
40. 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.
41. 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
42. 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.
43. Gillespie DT. Exact Stochastic Simulation of Coupled Chemical-Reactions. J Phys Chem. 1977;81:2340-61. PubMed PMID: ISI:A1977EE49800008.
44. The only exception to this rule is when recovery times are unknown In this case uncertainty in the recoverability SNP effects actually becomes large (see section 35).
45. The only exception to this is when dominance factors are changed the corresponding SNP effects are turned on (see Fig 4).
46. 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.
47. Within the range of parameter values that actually generate epidemics.
48. 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:e39551.
49. Bijma P. Estimating indirect genetic effects: precision of estimates and optimum designs. Genetics. 2010.
50. Wolf JB, Brodie III ED, Cheverud JM, Moore AJ, Wade MJ. Evolutionary consequences of indirect genetic effects. Trends Ecol Evol. 1998;13:64-9.
51. Gondro C, Van der Werf J, Hayes BJ. Genome-wide association studies and genomic prediction: Springer; 2013.

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

52. Mixing relates to the number of MCMC iterations needed to generate a set of samples representative of the posterior
53. E Laing R, Hess P, Shen Y, Wang J, X Hu S. The role and impact of SNPs in pharmacogenomics and personalized medicine. Current drug metabolism. 2011;12:460-86.
54. Houston R, Gheyas A, Hamilton A, Guy D, Tinch A, Taggart J, et al. Detection and confirmation of a major QTL affecting resistance to infectious pancreatic necrosis (IPN) in Atlantic salmon (Salmo salar). Dev Biologicals. 132: Karger Publishers; 2008. p. 199-204.
55. 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;44:217-22.
56. Sekhwal M, Li P, Lam I, Wang X, Cloutier S, You F. Disease resistance gene analogs (RGAs) in plants. International journal of molecular sciences. 2015;16:19248-90.
57. 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.
58. 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.
59. 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;243:83-92.
60. 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;178:1109-15.
61. 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;64:95-106.
62. 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;100:1272-81.
63. 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.
64. 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.
65. 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;2:74-99.
66. Doeschl-Wilson A, Knap P, Kinghorn B, Van der Steen H. Using mechanistic animal growth models to estimate genetic parameters of biological traits. Animal. 2007;1(4):489-99.
67. Roy B, Kirchner J. Evolutionary dynamics of pathogen resistance and tolerance. Evolution. 2000;54(1):51-63.
68. 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;10:368.
69. 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.
70. 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.
71. 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.

bioRxiv preprint doi: https://doi.org/10.1101/618363; this version posted September 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

72. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. Genome research. 2007;17:1520-8.
73. Andrieu C, Thoms J. A tutorial on adaptive MCMC. Stat Comput. 2008;18:343-73.
74. Roberts GO, Rosenthal JS. Examples of Adaptive MCMC. J Comput Graph Stat. 2009;18:349-67.