ABSTRACT: Data on disease occurrences at field level are valuable resources for quantifying host genetic variation in disease resistance. However, they are often inaccurate due to incomplete information describing exposure, disease prevalence and imperfect diagnostic tests. Quantitative genetic models of disease occurrence data do not typically account for these factors leading to underestimation of the true extent of genetic variation. We propose a framework that integrates genetics and epidemiology including genetic relationships between animals, observed disease state, prevalence of the disease and sensitivity and specificity of diagnostic tests. Bayesian inference allows quantification of host genetic variation accounting for the complexities inherent in field disease data. Prior information, as elicited by expert opinion, is incorporated. Application to simulated data shows this novel approach provides reliable inferences on genetic and epidemiological parameters that are of practical relevance to animal breeders.

Key words: disease genetics; disease diagnosis; Markov chain Monte Carlo (MCMC); heritability.

INTRODUCTION

Evaluating genetic variation in host resistance to infectious disease in animals requires large amounts of phenotypic data to obtain sufficient power. The availability of high density single nucleotide polymorphic (SNP) arrays has enhanced the possibilities for identifying candidate genes contributing to this variation using genome-wide association studies. There is tremendous potential to combine traditional quantitative genetics approaches with modern molecular technologies to augment our understanding of the role of host genetic variation in disease processes. This would allow identification of animals e.g. with reduced susceptibility to infection, and hence designing possible breeding strategies to improve overall herd or flock resistance.

Large-scale designed experiments with challenged animals are expensive, often infeasible and rare. On the other hand, data recorded at field level on populations that have experienced an epidemic are often readily available. Such field disease data, however, are often inaccurate and inconsistent due to uncertainties associated with epidemiological parameters and sensitivity and specificity of the diagnostic test used to determine disease state of individuals. It is imperative that these factors be accounted for when developing models to evaluate the true extent of genetic variation in disease resistance. For example, Bishop and Woolliams (2010) showed that estimated heritability is biased downwards if such factors are not taken into account.

Existing quantitative genetics methods typically use threshold models to analyze disease incidence data and these approaches do not adequately account for uncertainties in exposure or sensitivity and specificity of diagnostic tests (Bishop and Woollams, 2010). Using Bayesian inference, this paper proposes a novel genetic-epidemiological modeling framework to estimate relevant genetic and phenotypic parameters that more fully accounts for the complexities inherent in field disease data than currently available models.

MATERIALS AND METHODS

Data structure. The data consist of the binary disease state of individual animals (infected or healthy) as ascertained by a suitable diagnostic test at a defined time point during an epidemic outbreak. Additionally, pedigree information indicating relationships between animals and measurements on covariates and the levels of factors e.g. herd, year, season, breed, sex etc. (modeled as fixed effects) are assumed to be known.

Model assumptions. The proposed model is implemented in a static disease scenario, in which data are collected at a given time point during an epidemic outbreak and thus epidemic dynamics are not incorporated. The mixing between animals and exposure to infection are assumed to be complete and constant for all individuals. The model does not consider the interaction effects between genetic and epidemiological parameters e.g., it is assumed that the sensitivity and specificity of the diagnostic test are independent of host genotype.

Model description and parameters. A standard quantitative genetic model is implemented in the proposed genetic epidemiological framework, including fixed effects (factors or covariates), random effects of animal and relevant epidemiological parameter (prevalence) and the diagnostic test parameters (sensitivity and specificity). The model can be expressed in terms of two key components:

(i) \( p(y|y^*, Se, Sp) \) describing the observed state of disease \( y = \{y_i\} \) conditional on the true (latent) state of disease \( y^* = \{y^*_i\} \) and diagnostic test sensitivity \( Se \) and specificity \( Sp \), where \( y_i \sim Bernoulli(q_i) \)

\[ q_i = y^*_i \ast Se + (1 - y^*_i) \ast (1 - Sp) \]

Here, \( y_i \) and \( y^*_i \) are, respectively, the observed (presence = 1, absence = 0) and true disease states of the \( i \)-th individual. The quantity \( q_i \) is the mean probability of corresponding Bernoulli distribution. The parameter \( Se \) is the probability that a truly infected individual is classified by the diagnostic test as infected, and \( Sp \) is the probability that a truly healthy individual is classified by the diagnostic test as healthy.

(ii) \( p(y^*|\beta, u, \pi) \) describing the latent disease state in terms of fixed effect \( \beta \), random effect \( u = \{u_i\} \) and disease prevalence \( \pi \), where
\[ y_i^* \sim \text{Bernoulli}(p_i) \]
\[ \text{probit}(p_i) = (x_i \beta_k + z_i u_i) - \tau(\pi) \]

The latent disease state is represented on the so-called liability scale, conditional on the mean probability \( p_i \) of the corresponding Bernoulli distribution. Here, \( u_i \) is the additive effect for susceptibility to a disease for the \( i \)-th individual. Assuming an infinitesimal model, the distribution of \( u \) is multivariate normal and is represented as \( u \sim N(0, \mathbf{A} \sigma_u^2) \) where \( \mathbf{A} \) is the relationship matrix between individuals and \( \sigma_u^2 \) is the additive variance. The known design matrix relating to random animal effect \( u_i \) comprises elements of \( z_i \). The matrix \( \mathbf{A} \) is known and calculated from the pedigree information. The occurrence of disease also depends on a fixed effect with corresponding known design matrix elements \( x_{ik} \) and \( k \)-th coefficient \( \beta_k \). The model is defined such that positive values of additive effects (or coefficients) confer higher susceptibility to a disease for the animal (or reduced resistance to the disease).

Finally, the true disease state also depends on the prevalence of the disease \( \pi \). Considering complete exposure, \( \pi \) defines the mean prevalence of the disease which is expressed as the proportion of animals that would be infected in the population. In this model, \( \pi \) is equivalent to the threshold \( \tau(\pi) \) on the liability scale such that \( F(\tau(\pi)) = 1 - \pi \), where \( F(.) \) is the standard cumulative normal distribution.

A probit link function is used to map the linear predictor to the mean probability of occurrence of disease \( p_i \). In the case of binary disease state observations, as modeled here, the variance of the non-structured random effect \( \sigma_e^2 \) is confounded with the linear predictor, and is not identifiable (Sorensen et al., 1995). Therefore genetic variability is estimated relative to environmental variation and \( \sigma_e^2 \) is taken as 1.

**Bayesian inference.** The parameters to be inferred from data \( \theta = (\beta, u, \sigma_u^2, \pi, Se, Sp) \) can be considered as genetic \( (\beta, u, \sigma_u^2) \) and epidemiological \( (\pi, Se, Sp) \). The joint posterior density for all parameters given the observed disease state is given by:
\[
p(\theta|y) \propto p(y|y^*, Se, Sp)p(y^*|\beta, u, \pi)p(u|\sigma_u^2, A)p(\theta)
\]

Here, \( p(y|y^*, Se, Sp) \), \( p(y^*|\beta, u, \pi) \) and \( p(u|\sigma_u^2, A) \) are respectively the observation process, the latent disease state and the genetic relatedness in the population and we have introduced the prior for all parameters as \( p(\theta) \).

**Prior distributions.** Here we assume independent prior information for each parameter so that \( p(\theta) \) is simply the product of the marginal priors \( p(\beta) \), \( p(\sigma_u^2) \), \( p(\pi) \), \( p(Se) \) and \( p(Sp) \). Incarnation of priors into the model allows previous knowledge or expert knowledge to be taken into account. For the purposes of this study, largely uninformative flat uniform priors are considered for \( p(\beta) \), \( p(\sigma_u^2) \) and \( p(\pi) \), while beta distributions are chosen for \( p(Se) \) and \( p(Sp) \). Beta distributions are chosen to represent prior information for the diagnostic test parameters since they are flexible and the hyperparameters of these beta distributions can easily be elicited based on expert knowledge. For example, it is possible to base the expert opinion on the most-likely (modal) prior value of the parameter and an upper or lower percentile for the parameter (Branscum et al., 2004). Assuming expert opinions on 50% and 97.5% quantiles of \( Se \) and \( Sp \) as (0.88 and 0.97) and (0.93 and 0.98), we elicited hyperparameters \( (a, b) \) of the beta distributions for \( Se \) and \( Sp \) as (22, 99, 3.44) and (35.00, 3.00), respectively.

**Sampling from the posterior.** The posterior distribution described above represents the Bayesian inference on the model parameters \( \theta \) given the model structure, the data and the prior information \( p(\theta) \) (Gelman et al., 2004). A Markov chain Monte Carlo (MCMC) algorithm is implemented to sample parameters from the posterior. A Metropolis-Hastings algorithm is used with proposals designed to ensure good mixing of the samples generated. Each model is run for 500,000 iterations, after an initial burn-in of 20,000 iterations. Convergence of the Markov chain is assessed by standard diagnostic points and statistics (Gelman and Rubin, 1992). Hardware acceleration based on graphical processor unit technology is implemented to speed up the MCMC sampling and as a result, total computation time for each model is completed in less than 10 minutes.

**Simulated Data Scenarios.** To demonstrate the utility of the proposed model, we generate simulated data under scenarios representative of typical field data. Fitting the model to simulated data sets enables performance to be tested, and here we explore the impact of different levels of heritability on our ability to infer genetic and epidemiological parameters.

The simulated data is based on two generations (including base generation) with known parameter values as described below. The base population consists of 100 sires and 1000 dams and the relationships between these animals are assumed to be unknown. The number of progeny from each sire and dam is taken to be 100 and 10, respectively. Matings are assumed to be at random. Three different scenarios are considered, with heritabilities (on the liability scale) of 0.20, 0.40 and 0.60, respectively. A factor with two levels (representing breed effect) is included as a fixed effect and coefficients are assigned as (0, 0.5) on the liability scale. Data on the observed disease states of all individuals are simulated accounting for other epidemiological and diagnostic test parameters \( (\pi = 0.30, Se = 0.85 \text{ and } Sp = 0.95) \).

**RESULTS AND DISCUSSION**

The posterior inferences of model parameters based on simulated data from each of the three scenarios described above show that all genetic and epidemiological parameters are well-estimated in each instance, and 95% credible intervals contain the true parameter values used to simulate the data (Table 1). Moreover, the method is able to achieve such inferences even for reasonably low additive genetic variance; see Scenario 1 for additive variance (and heritability) and the corresponding median and 95% credible interval (Table 1). Additionally, a contour plot (Figure 1) of the variable posterior probability, as a function of the heritability and sensitivity of the diagnostic test, shows that the probability mass concentrates around the true estimates of these parameters, suggesting that the
model reasonably captures the true values. All diagnostic plots and statistics on model convergence support acceptable convergence of the model. Monte Carlo error values suggest acceptable accuracy of posterior estimates.

CONCLUSION

A genetic epidemiological model framework is presented to quantify the genetic variation in disease susceptibility from field disease data. The model integrates data on disease state, genetic relationship between individuals and epidemiological and diagnostic characteristics. Using Bayesian inference, the model provides estimates of the genetic (e.g. heritability of susceptibility to disease), phenotypic and epidemiological (e.g. prevalence of disease) parameters of the population.

The present model serves as a novel and practical approach to extract genetic information from rich but noisy field disease data. This represents an advance in which complexities of field disease data are better accounted for than in current modeling strategies.

ACKNOWLEDGMENTS

We wish to thank the Scottish Government for funding through SPASE and the BBSRC.

LITERATURE CITED


Table 1. Posterior inferences of genetic and epidemiological parameters based on simulated data for three different scenarios.

<table>
<thead>
<tr>
<th>Par</th>
<th>TV</th>
<th>Mean</th>
<th>Med</th>
<th>SD</th>
<th>MCE</th>
<th>95%CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_2$</td>
<td>0.50</td>
<td>0.50</td>
<td>0.49</td>
<td>0.08</td>
<td>0.005</td>
<td>(0.37, 0.68)</td>
</tr>
<tr>
<td>$\sigma_u^2$</td>
<td>0.25</td>
<td>0.31</td>
<td>0.29</td>
<td>0.11</td>
<td>0.008</td>
<td>(0.17, 0.58)</td>
</tr>
<tr>
<td>$Se$</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.07</td>
<td>0.006</td>
<td>(0.69, 0.96)</td>
</tr>
<tr>
<td>$Sp$</td>
<td>0.95</td>
<td>0.94</td>
<td>0.94</td>
<td>0.03</td>
<td>0.003</td>
<td>(0.86, 0.99)</td>
</tr>
<tr>
<td>$\pi$</td>
<td>0.30</td>
<td>0.30</td>
<td>0.30</td>
<td>0.05</td>
<td>0.004</td>
<td>(0.19, 0.39)</td>
</tr>
<tr>
<td>$h^2$</td>
<td>0.20</td>
<td>0.23</td>
<td>0.22</td>
<td>0.06</td>
<td>0.004</td>
<td>(0.14, 0.37)</td>
</tr>
</tbody>
</table>

Scenario 2

| $\beta_2$ | 0.50 | 0.50 | 0.49 | 0.13 | 0.004 | (0.25, 0.77) |
| $\sigma_u^2$ | 0.68 | 0.61 | 0.57 | 0.21 | 0.021 | (0.31, 1.12) |
| $Se$ | 0.85 | 0.87 | 0.88 | 0.06 | 0.001 | (0.72, 0.97) |
| $Sp$ | 0.95 | 0.92 | 0.93 | 0.04 | 0.001 | (0.82, 0.98) |
| $\pi$ | 0.30 | 0.29 | 0.29 | 0.05 | 0.002 | (0.21, 0.37) |
| $h^2$ | 0.41 | 0.37 | 0.36 | 0.07 | 0.008 | (0.24, 0.53) |

Scenario 3

| $\beta_2$ | 0.50 | 0.49 | 0.48 | 0.15 | 0.005 | (0.23, 0.82) |
| $\sigma_u^2$ | 1.50 | 1.31 | 1.19 | 0.57 | 0.049 | (0.62, 2.76) |
| $Se$ | 0.85 | 0.87 | 0.88 | 0.06 | 0.001 | (0.72, 0.97) |
| $Sp$ | 0.95 | 0.92 | 0.93 | 0.04 | 0.001 | (0.82, 0.98) |
| $\pi$ | 0.30 | 0.29 | 0.29 | 0.05 | 0.003 | (0.19, 0.39) |
| $h^2$ | 0.60 | 0.55 | 0.54 | 0.09 | 0.010 | (0.38, 0.73) |

Par: Parameters of the model.
TV: True values of the parameters.
Mean: Mean of MCMC samples.
Med: Median of MCMC samples.
SD: Standard deviation of MCMC samples.
MCE: Monte Carlo error of MCMC samples.
95%CI: 95% credible interval of MCMC samples.
$\beta_2$: Coefficient of the second level of fixed effect.
$\sigma_u^2$: Additive variance.
$Se$: Sensitivity of the diagnostic test.
$Sp$: Specificity of the diagnostic test.
$\pi$: Mean prevalence of disease.
$h^2$: Heritability of disease susceptibility.

Figure 1. The probability surface over the x-y plane where the model outcomes for the heritability of susceptibility to a disease (x-axis) and sensitivity of a diagnostic test (y-axis) are plotted. The model inference is based on a simulated dataset with heritability of 0.20 and diagnostic test sensitivity of 0.85. The red dot shows the intersection of true values of the parameters and the white dot shows the intersection of median estimates of parameters.