Simultaneous Inference of Genetic Parameters Underlying Susceptibility and Infectivity of Livestock from Epidemiological Data

O. Anacleto*, D. Lipschutz-Powell*, L. Garcia-Cortez†, J.A. Woolliams* and A. Doeschl-Wilson*

* The Roslin Institute and R(D)SVS, University of Edinburgh, Roslin, Midlothian EH25 9PS, UK
† SGIT - INIA, Ministerio de Ciencia e Innovación, Carretera de la Coruña, 28040 Madrid, Spain

ABSTRACT: The implementation of animal breeding programs to select for disease resistance can be a very helpful control strategy. However, usual quantitative genetics models cannot cope with the complex features inherent in epidemiological data, mainly because these models fail to accommodate the transmission dynamics from the population undergoing an epidemic. In this paper, we show how parameters related to infectivity and susceptibility traits can be simultaneously estimated by including the disease dynamics into the model. Preliminary results show that the proposed genetic-epidemiological model can be a promising tool to estimate breeding values and covariance matrices of infectious disease traits when taking into account not only variation in susceptibility but also the usually neglected variation in infectivity.
Keywords: Disease genetics; Bayesian statistics; Infectious disease

Introduction

The lack of effective control measures for infectious diseases in livestock not only causes significant economic losses but may also endanger human health and food security. Although genetic selection for disease resistance has been considered as a desirable disease control strategy, current quantitative genetics models cannot cope with the complex features inherent in epidemiological data (Lipschutz-Powell et al. (2012)).

Epidemiological data used in quantitative genetic analyses often come in binary form (infected /not infected). However, epidemiological theory points to two important underlying host traits affecting the binary disease status of individuals: susceptibility and infectivity (Lipschutz-Powell et al. (2012)). Both may harbour considerable genetic (co-) variation. Lipschutz-Powell et al. (2014) derived a probability function that links the binary disease phenotype to both host susceptibility and infectivity of the host group members and incorporates transmission dynamics.

In this paper, we show how to explore this probability function to make inference about genetic parameters for host susceptibility and infectivity. In particular, we derive a likelihood function that takes into account that only infected individuals can express infectivity and that artificially infected individuals may not express differences in susceptibility. We show how a Bayesian framework with a suitable MCMC strategy which considers the incomplete nature of epidemic data can provide a powerful approach to estimate genetic parameters of infectious disease traits.

Materials and Methods

Definitions and assumptions. Consider a population of N related animals distributed over multipleclosed groups of size n. We assume that susceptible individuals may become infected after contact with infectious individuals, and that individuals, once infected become immediately infectious and remain so over the observation time period (i.e. epidemic SI model). The disease status (non-infected / infected) of each animal is observed at discrete sampling times $\mathbf{t}^\top = [t_1 = 0, \ldots, t_L = T]$. Assume further that S infected animals from the population are already infected at time 0 (denoted here as index cases), with $\mathbf{s}^\top = [s_1, \ldots, s_N]$ representing the vector of infection status at time t_1, where $s_i = 0$ if animal i is infected at time 0 and $s_i = 1$ otherwise, $i=1,\ldots N$.

Data. Repeated sampling of individual infection status results in the data vectors $\mathbf{t_B}^\top = [t_{B1}, \ldots, t_{BN}]$ and $\mathbf{t_E}^\top = [t_{E1}, \ldots, t_{EN}]$ where $\mathbf{t_{B_i}}$ and $\mathbf{t_{E_i}}$ represent, respectively, the last sampling time that animal i was observed as susceptible and the first sampling time where i was observed as infected. Also, if animal i was non-infected during the observation period we define $t_{Bi} = t_{Ei} = T$, while, for the index cases, $t_{Bi} = t_{Ei} = 0$.

Modelling infectivity and susceptibility. According to Lipschutz-Powell et al. (2014), the probability of an animal j becoming infected by time τ_j, denoted by the distribution function $F()$, is,

(1)

where $H()$ is the Heaviside step function, g_j is the susceptibility of animal j and f_k is the infectivity of the group members of j, $k=1,\ldots n-1$. Both g_j and f_k are defined as probabilities in (1). Note that the true infection time τ_j is usually unknown and has to be inferred from the disease status of each individual during sampling times, previously defined by the vectors $\mathbf{t_B}$ and $\mathbf{t_E}$.

In this context, we can define the parameter vectors **g** and **f** representing, respectively, the unknown susceptibility and infectivity of the animals from the population. Moreover, the susceptibility of each animal can be modelled as

$$\psi_j = \log\left(\frac{g_j}{1-g_j}\right) = \mu_\psi + a_{\psi,j} + \epsilon_{\psi,j},\ j = S+1,\ldots,N, \quad (2)$$

which represents the assumption that the susceptibility of an animal *j* is a function of a population mean μ_ψ, its susceptibility breeding value $a_{\psi,j}$ and an environmental deviation $\epsilon_{\psi,j}$. Note that the susceptibility is not expressed in artificially infected index cases, as these individuals are already infected prior to the observation period so that their susceptibility cannot be modelled.

Similarly, the infectivity of each animal can be written as

$$\iota_j = \log\left(\frac{f_j}{1-f_j}\right) = \mu_\iota + a_{\iota,j} + \epsilon_{\iota,j},\ j = 1,\ldots,I \quad (3)$$

where *I* is the number of individuals in the population that have become infected by time *T*. Hence, the infectivity of an animal *j* is modelled as a function of a population mean μ_ι, its infectivity breeding value $a_{\iota,j}$ and an environmental deviation $\epsilon_{\iota,j}$. Note that infectivity is only expressed for infected animals, as this trait is not observed in the non-infected individuals. The Bayesian hierarchical structure underlying the model based on Equations (2) and (3) is presented in Doeschl-Wilson, Lipschutz-Powell, Anacleto, et al. (2014).

Bayesian inference. Let $\boldsymbol{\psi}$ and $\boldsymbol{\iota}$ be the vectors of susceptibility and infectivity transformed onto the real line as defined in (2) and (3), with associated vectors of breeding values $\mathbf{a}_\psi$ and $\mathbf{a}_\iota$. Given the hierarchical structure of the model, the joint posterior distribution of the parameter set $\boldsymbol{\theta}^\top = (\boldsymbol{\tau}, \boldsymbol{\psi}, \boldsymbol{\iota}, \mu_\psi, \mu_\iota, \boldsymbol{a}_\psi, \boldsymbol{a}_\iota, \boldsymbol{G}, \boldsymbol{V})$ is given by

$$f(\boldsymbol{\tau}, \boldsymbol{\psi}, \boldsymbol{\iota}, \mu_\psi, \mu_\iota, \boldsymbol{a}_\psi, \boldsymbol{a}_\iota, \boldsymbol{G}, \boldsymbol{V} | \boldsymbol{t}_B, \boldsymbol{t}_E, \boldsymbol{s}, \boldsymbol{A}) =$$

$$f(\boldsymbol{t}_B, \boldsymbol{t}_E | \boldsymbol{\psi}, \boldsymbol{\iota}, \boldsymbol{\tau}, \boldsymbol{s}) f(\boldsymbol{\psi}, \boldsymbol{\iota} | , \mu_\psi, \mu_\iota, \boldsymbol{a}_\psi, \boldsymbol{a}_\iota, \boldsymbol{V})$$
$$f(\boldsymbol{a}_\psi, \boldsymbol{a}_\iota | \boldsymbol{G}, \boldsymbol{A}) f(\boldsymbol{G} | \boldsymbol{A}) f(\boldsymbol{V}) f(\mu_\psi) f(\mu_\iota), \quad (4)$$

where **A** is the relationship matrix. Independent normal priors are defined for each vector $\boldsymbol{\psi}$, $\boldsymbol{\iota}$, $\mathbf{a}_\iota$, $\mathbf{a}_\iota$, and population means μ_ψ and μ_ι. Additionally, non-informative priors are defined for the covariance matrices **G** and **V** using inverse Wishart distributions.

Likelihood. Using the distribution function for infection times shown in Equation (1), the likelihood $f(\boldsymbol{t}_B, \boldsymbol{t}_E | \boldsymbol{\psi}, \boldsymbol{\iota}, \boldsymbol{\tau}, \boldsymbol{s})$ in (4) can be written as

(5)

where *F()* is the probability of *j* being infected up to time τ_j as defined in (1) and

---

is the density function evaluated at τ_j, *j=S+1,…,I*. Note that the infected and non-infected animals contribute differently to the likelihood in (5): while the density function is evaluated at τ_j for each infected individual, the contribution of the non-infected animals to the likelihood is included by evaluating their probability of being infected after the last sampling time, which is *1-F(T)* for each animal. Note also that the contribution of the index cases to the likelihood is only through their infectivity parameters in the likelihood terms of the non-index cases.

Hybrid MCMC algorithm. Bayesian inference on the vector $\boldsymbol{\theta}^\top = (\boldsymbol{\tau}, \boldsymbol{\psi}, \boldsymbol{\iota}, \mu_\psi, \mu_\iota, \boldsymbol{a}_\psi, \boldsymbol{a}_\iota, \boldsymbol{G}, \boldsymbol{V})$ is done by developing an MCMC algorithm where each component of of **θ** is sampled according to its conditional posterior distribution, which can be obtained from Equation (4).

The assumptions of independent normal priors for the breeding values $\mathbf{a}_\psi$, and $\mathbf{a}_\iota$, and population means μ_ψ and μ_ι, together with non-informative Wishart priors for **G** and **V** allow a Gibbs sampling approach for $\mathbf{a}_\psi$, $\mathbf{a}_\iota$, μ_ψ, μ_ι as well as for the genetic and environmental matrices **G** and **V**. Also, due to the non-trivial likelihood function defined in (5), a Metropolis-Hastings (MH) algorithm was used to sample each term of the liability vectors $\boldsymbol{\psi}$ and $\boldsymbol{\iota}$. As the posterior distributions of the $\boldsymbol{\psi}$ and $\boldsymbol{\iota}$ components may differ greatly among each other, an adaptive MH strategy provides a more efficient sampling algorithm for these parameter vectors. Additionally, similar to MCMC algorithm used in Strefaris and Gibson (1998), each unknown infection time τ_j can be sampled from a uniform distribution on $[t_{Bi}, t_{Ei}]$.

Simulations. A population with a half-sib family structure based on 200 sires and 10 dams (*N*=2000) was simulated to evaluate the model performance. Both susceptibility and infectivity of each animal were independently generated according to Equations (2) and (3) and assuming unit variances for both (diagonal) genetic and environmental covariance matrices. Each animal was randomly assigned to one of 200 closed groups of size 10 each, and one animal in each group was defined as an index case. The disease spread in each group was simulated following a stochastic epidemiological SI model according to the methodology outlined in Lipschutz-Powell et al. (2012). It was also assumed that the infection times could be observed in the population.

Results and Discussion

For an initial evaluation of the proposed genetic epidemiological model, the MCMC algorithm was run to obtain posterior estimates of breeding values and environmental and genetic covariance matrices based on the simulated data. Figures (1) and (2) show scatterplots of the posterior mean and true susceptibility and infectivity breeding values for all animals of the population. These scatterplots indicate a good agreement between estimates based on the proposed model and true values from the population. Additionally, Pearson correlation coefficients for the relationship between posterior means and true breeding values were 0.66 for susceptibility and 0.42 for infectivity.

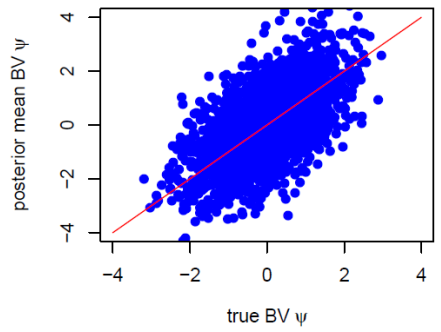

Figure 1: Posterior means *versus* true susceptibility breeding values from the simulated population.

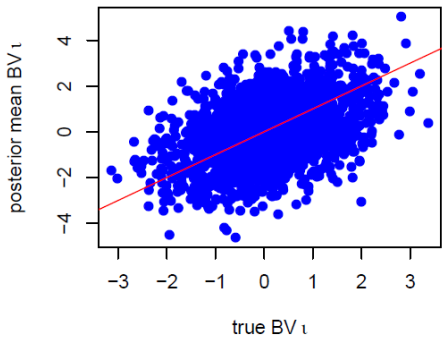

Figure 2: Posterior means *versus* true infectivity breeding values from the simulated population.

Posterior distributions for both genetic and environmental covariance matrices are distributed around the true (unit) values used to simulate the data. However, some indications of mixing problems were found in the analysis of the resulting Markov chains. Hence, further work is required to develop more efficient MCMC algorithms to improve the estimates of genetic and environment covariance matrices and of the liabilities for susceptibility and infectivity.

Conclusion

Preliminary results presented in this paper show that the proposed genetic-epidemiological model is a promising tool to estimate genetic parameters of infectious disease traits when taking into account not only variation in susceptibility but also the usually neglected variation in infectivity. As with most of the Bayesian models used in epidemiology, great care must be taken to develop an MCMC algorithm which can efficiently provide accurate estimates for the high number of required parameters.

The simulation considered for the model evaluation assumed known infection times, which is usually uncommon in practice. However, as shown in Strefaris and Gibson (2004), experimental designs with a high number of sampling times can provide results very similar to the case when infection times are known.

Literature Cited

Doeschl-Wilson A.B, Lipschutz-Powell D., Anacleto O., et al. (2014). *Proceedings of the 10th WCGALP 2014*.

Lipschutz-Powell D., Woolliams J.A., Bijma P. and Doeschl-Wilson A.B. (2012). *PLoS ONE*, 7, e39551.

Lipschutz-Powell D., Woolliams J.A., and, Doeschl-Wilson A.B. (2014). *Genet. Sel. Evol.* 46(15).

Strefaris G. and Gibson G.J. (2004). *Stat. Model.*, 4, 63-75.