Why genetic selection to reduce the prevalence of infectious diseases is way more promising than currently believed

Andries D. Hulst,[1,2,*] Mart C.M. de Jong [iD],[1] and Piter Bijma[2]

[1]Quantitative Veterinary Epidemiology, Wageningen University & Research, 6700AH, Wageningen, The Netherlands
[2]Animal Breeding and Genomics Group, Wageningen University & Research, 6700AH, Wageningen, The Netherlands

*Corresponding author: Quantitative Veterinary Epidemiology, Wageningen University & Research, P.O. Box 338, 6700AH, Wageningen, The Netherlands. dries.hulst@wur.nl

Abstract

Genetic selection for improved disease resistance is an important part of strategies to combat infectious diseases in agriculture. Quantitative genetic analyses of binary disease status, however, indicate low heritability for most diseases, which restricts the rate of genetic reduction in disease prevalence. Moreover, the common liability threshold model suggests that eradication of an infectious disease via genetic selection is impossible because the observed-scale heritability goes to zero when the prevalence approaches zero. From infectious disease epidemiology, however, we know that eradication of infectious diseases is possible, both in theory and practice, because of positive feedback mechanisms leading to the phenomenon known as herd immunity. The common quantitative genetic models, however, ignore these feedback mechanisms. Here, we integrate quantitative genetic analysis of binary disease status with epidemiological models of transmission, aiming to identify the potential response to selection for reducing the prevalence of endemic infectious diseases. The results show that typical heritability values of binary disease status correspond to a very substantial genetic variation in disease susceptibility among individuals. Moreover, our results show that eradication of infectious diseases by genetic selection is possible in principle. These findings strongly disagree with predictions based on common quantitative genetic models, which ignore the positive feedback effects that occur when reducing the transmission of infectious diseases. Those feedback effects are a specific kind of Indirect Genetic Effects; they contribute substantially to the response to selection and the development of herd immunity (*i.e.*, an effective reproduction ratio less than one).

Keywords: infectious disease; epidemiology; quantitative genetics; heritability; response to selection; indirect genetic effect

Introduction

Infectious diseases are of great concern in agriculture, because they cause considerable economic damage in terms of production losses and costs of treatment. In livestock, moreover, infectious diseases harm animal health and welfare, and cause indirect cost through trade restrictions and impact on human health in the case of zoonoses (Bennett 2003). Although combatting infectious diseases has always been a challenge, the increase in antibiotic resistance, followed by restrictions on the use of antibiotics has intensified the need for additional solutions (Speksnijder *et al.* 2015).

Genetic selection for improved disease resistance is a common strategy to combat, especially endemic, infectious diseases, as an alternative or additional intervention to classical control measures such as antibiotic treatment (Bishop *et al.* 2010). In livestock, genetic variation in disease resistance seems to be present in virtually all species, and for a wide range of diseases (Bishop *et al.* 2010). While this observation implies that genetic improvement of infectious disease traits is possible *in principle*, the rate of genetic improvement is restricted by the limited heritability of the disease traits.

Infectious disease traits are often measured as the binary disease status of the individual (healthy/diseased = 0/1), and heritability estimates for disease status are typically below 10%. For mastitis in dairy cattle, for instance, observed-scale heritability estimates range from 0.01 to 0.09, with corresponding prevalence ranging from 0.09 to 0.39 (Martin *et al.* 2018). Although the heritability of binary traits is sensitive to prevalence, with the highest values at intermediate prevalence, the above estimates are fairly low. Given the population average value for binary disease status (p), the additive genetic variance in binary disease status is fully determined by heritability; $\sigma^2_{A_{obs}} = h^2_{obs}p(1-p)$. Thus these low heritabilities also indicate limited genetic variation in binary disease status. Because the mean value of individual disease status is equal to the prevalence of the disease, the formula for the additive genetic variance also seems to imply that the potential response to selection for reduction of disease prevalence is limited.

Higher heritabilities are found with threshold models, which assume a continuous liability underlying the observed binary disease status. (Dempster and Lerner 1950; Robertson 1950). However, while threshold models are statistically much more

appropriate than linear models (Gianola 1982), a high heritability on the liability scale does not imply a large response to selection, because the response in binary disease status depends on the heritability on the observed binary scale (Robertson 1950).

An important expectation following from classical quantitative genetic models is the impossibility to eradicate a disease via selection. Eradication seems impossible because additive genetic variance on the observed binary scale goes to zero when the prevalence approaches zero (Robertson 1950). Hence, for breeders, this implies that genetic reduction of disease prevalence becomes increasingly difficult and ultimately impossible at lower prevalence. In epidemiology, however, successful eradication of several infectious diseases has been achieved using vaccination, with vaccines being neither 100% effective nor successfully applied to everyone (Halloran *et al.* 1992). A well-known example in livestock is the worldwide eradication of rinderpest, which has been successful without the need to vaccinate all the animals. In Somalia, for example, only about 50% of the animals had vaccination immunity for rinderpest at eradication (Mariner *et al.* 2012).

The eradication of infectious diseases using vaccination relates to a phenomenon known as "herd immunity" (Fine 1993). When a sufficient proportion of the population is immune to the disease, the disease can no longer spread in the population and thus dies out. Hence, the mechanisms determining the prevalence of an infectious disease are fundamentally different from those determining the prevalence of a noncommunicable disease, such as, say, heart failure. This difference is relevant not only for the case of vaccination, but also for the results of selection for disease resistance; a reduction in disease susceptibility due to genetic selection would have the same effect on disease prevalence as a corresponding reduction in susceptibility achieved with vaccination. However, despite this fundamental difference, mainstream theory and methods in quantitative genetics and livestock genetic improvement completely disregard the key role of transmission between individuals.

Whereas the previous paragraph stresses the importance of transmission dynamics, other explanations for the low heritability of infectious diseases have also been proposed. The low heritability of individual disease status, even at intermediate prevalence, seems to disagree with the moderate to high heritability estimates for immune response traits (Knap and Bishop 2000; Henryon *et al.* 2006; Thompson-Crispi *et al.* 2012). To explain this discrepancy, Bishop and Woolliams (2010) argue that incomplete exposure of individuals to the infectious agent, low test specificity and sensitivity, and incomplete data recording can cause underestimation of the heritability of binary disease status. In addition, Lipschutz-Powell *et al.* (2012) point to the potential importance of genetic variation in infectivity, defined as the propensity of infected individuals to infect others, which is not captured by the current models. Hence, these mechanisms might also partly explain why not all genetic variation is captured with the current breeding methods.

Nevertheless, while more accurate disease data and accounting for genetic variation in infectivity might indeed improve the response to selection (Tsairidou *et al.* 2019), these factors cannot explain why epidemiological models that account for transmission dynamics often suggest a substantially greater response to selection than genetic models, even in the absence of measurement errors and when genetic variation is in disease resistance only (*i.e.*, in the absence of genetic variation in infectivity). Using an epidemiological model tailored to foot rot in sheep, for example, Nieuwhof *et al.* (2009) have shown that selection for

resistance reduces prevalence faster than predicted by common quantitative genetic models of disease status. This suggests that the ordinary breeding values for individual disease status do not correctly predict prevalence in the offspring generation. Empirical examples of a relatively large reduction in disease prevalence resulting from genetic selection were found for infectious pancreatic necrosis (IPN) in Atlantic salmon and for clinical mastitis in dairy cows. The salmon industry managed to decrease IPN mortality by over 70% in 2 years, using marker-assisted selection for an IPN-resistance QTL (AquaGen 2012). In a selection experiment, Heringstad *et al.* (2007) observed a phenotypic decrease in the prevalence of clinical mastitis of 15% after 5 generations of single trait selection against clinical mastitis, while the reduction predicted by estimated breeding values was only 8%. Although this phenotypic response could be explained by environmental changes (*e.g.*, improvements in management), a correlated response in clinical mastitis from single trait selection on higher protein yield observed in the same experiment contradicts this explanation. Selection for higher protein yield resulted in a phenotypic increase in clinical mastitis prevalence from 10% to 25%, while the increase predicted by EBVs was only 2%. Hence, an environmental decreasing effect on clinical mastitis is highly unlikely, because this disagrees with the much higher phenotypic increase observed after selection for higher protein yield. Results of Heringstad *et al.* (2007), therefore, indicate greater genetic response than predicted by EBVs from classical quantitative genetic models.

The discrepancy between predictions based on epidemiological *vs.* genetic models illustrates that the relationship between the heritability of binary disease status and the amount of genetic variation that can actually be used to reduce disease prevalence is still unclear, even for the most basic and well-established epidemiological models. As a consequence, prospects of genetic selection to reduce infectious disease prevalence are not correctly predicted by current quantitative genetic models.

Here, we show that the low heritabilities of binary disease status are fully consistent with a large genetic variation in disease resistance. Heritability estimates from common quantitative genetic models by no means represent the variation available to breeders for genetic improvement, even for the simplest epidemiological model and with genetic variation in disease resistance only. Consequently, the potential for genetic selection to reduce the prevalence of infectious diseases will be considerably larger than currently believed, even when the conventional breeding values for disease status are used as a selection criterion. This occurs because the effect of genetic selection on the disease status of individuals acts in two ways: Both directly via a reduced susceptibility of the individuals, and indirectly via reduced exposure of individuals to infected herd mates, because of a lower susceptibility of those herd mates.

Material and methods

To quantify the relationship between the heritability of binary disease status and the potential response to genetic selection for lower prevalence, we will first find the genetic variation in disease susceptibility required to reproduce the common heritability values of disease status. For this purpose, we will use simulation of an endemic infectious disease with an epidemiological model that is appropriate for many endemic infectious diseases at the farm level, while keeping the mathematics as simple as possible. We aim to find the epidemiological parameters and genetic variances that result in observed-scale heritabilities for binary

Table 1 Notation key and overview of input values

SYMBOL	MEANING	SIMULATED VALUE(S)
β	Transmission rate parameter (t^{-1})	$0.03\,\mathrm{day}^{-1}$
α	Recovery rate parameter (t^{-1})	$0.02\,\mathrm{day}^{-1}$
S	Susceptible state of an individual	–
I	Infected state of an individual	–
S (in italics)	Number of susceptible individuals ()	–
I (in italics)	Number of infected individuals ()	–
N	Herd size ()	102
$\frac{I}{N}$	Prevalence	0.33
β_i	Transmission rate parameter for individual i (t^{-1})	$\beta\gamma_i\,\mathrm{day}^{-1}$
γ_i	Individual susceptibility	$e^{A_{\gamma_i}+E_{\gamma_i}}$
A_{γ_i}	Individual additive genetic effect (breeding value) for the logarithm of susceptibility	$\sim N(0,\ \sigma^2_{A,\gamma})$
E_{γ_i}	Individual permanent environmental effect for the logarithm of susceptibility	$\sim N(0,\ \sigma^2_{E,\gamma})$
$\sigma^2_{A,\gamma}$	Simulated additive genetic variance for the logarithm of susceptibility	–
$\sigma^2_{E,\gamma}$	Simulated permanent environmental variance for the logarithm of susceptibility	–
a^2_P	Chosen additive genetic proportion of the full permanent variance in log-susceptibility among individuals.	0.25; 0.5; 0.75; 1
$h^2_{\gamma,\ liab}$	Liability-scale heritability for the logarithm of susceptibility	–
h^2_{Obs}	Target observed-scale heritability of binary disease status	0.02; 0.05; 0.1

disease status of 0.02, 0.05, and 0.10, using a common prevalence of endemic infectious diseases in livestock. After identifying the epidemiological parameters and genetic variances corresponding to those heritabilities, we will estimate the potential response to selection using those parameters and variances as input.

Because the common linear models capture genetic variation in disease resistance only, we simulated individuals to differ only in disease resistance. For reasons of mathematical convenience and consistency with epidemiological models, we in fact modeled disease susceptibility instead of resistance. Susceptibility is the (relative) probability an individual becomes infected given exposure (Doeschl-Wilson and Kyriazakis 2012), and is the opposite of disease resistance; individuals with a high susceptibility have a low resistance, and *vice versa*.

This section starts with a description of the epidemiological model we used for simulation of the endemic disease. This description initially ignores variation in susceptibility, to make the model more easily understandable. Next, we describe how individual genetic variation in susceptibility can be introduced into the epidemiological model. Then, we describe how we simulated a population with variation in susceptibility, followed by a description of the implementation and the scenarios used to quantify response to selection. Table 1 provides a notation key.

Susceptible-Infectious-Susceptible model without genetic variation

To illustrate our point with a minimum of mathematical detail, we used one of the simplest epidemiological models for endemic infectious diseases, the so-called Susceptible-Infectious-Susceptible (SIS) model (Hethcote 1989). Although the model is simple, it provides a realistic representation of the transmission of several endemic infectious diseases in livestock populations and is well-established in veterinary epidemiology for that reason. Note that in the context of an SIS-model, "susceptible" merely means that an individual is not infected and can in principle become infected; it does not indicate high susceptibility or low disease resistance.

In the SIS-model, individuals can be in one of two states: susceptible (S) or infected and then also infectious (I; Figure 1). In epidemiology, the symbols S and I are generally used to indicate both the state of an individual and the total number of individuals in that state. To prevent confusion, we will use S and I (in

$$S \underset{\alpha I}{\overset{\beta S \frac{I}{N}}{\rightleftarrows}} I$$
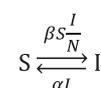
Figure 1 Susceptible-Infected-Susceptible-Model.

italics) to indicate *the number* of susceptible and infected individuals in a herd, and S and I for *the state* of an individual. Thus, S indicates the number of individuals with disease state S in a herd. The infection is endemic within the separate herds, and we assume that transmission can occur only between herd mates. N is the total number of individuals in a herd, which is equal to the sum of the number of susceptible and infected individuals in the SIS-model (S + I = N).

Transitions in an individual state, so-called events, are possible in both directions (from S to I and from I to S). Susceptible individuals can become infected through contacts with infected herd mates, while infected individuals can recover and thus become susceptible again. The types of events (*i.e.*, transmission or recovery) and the bookkeeping of the individuals involved in these events define our stochastic model. The average number of susceptible individuals that becomes infected per unit of time (*e.g.*, per day) depends on the total number of susceptible individuals in the herd at that moment, the fraction infected among their herd mates ($\frac{I}{N}$, *i.e.*, the prevalence of the disease), and the transmission rate parameter of the disease, β (Diekmann *et al.* 2012):

$$Rate\ S \rightarrow I = \beta S \frac{I}{N} \qquad (1)$$

For a population of $N = 100$ individuals in the endemic state with, for example, a prevalence of 0.1 ($I = 10$, $S = 90$), and a β of 0.03, on average $0.03 \cdot 90 \cdot 0.1 = 0.27$ susceptible individuals will become infected each day.

The number of recovering individuals per unit of time depends on the number of infected individuals in the herd at that moment (I) and on the recovery rate parameter, α:

$$Rate\ I \rightarrow S = \alpha I \qquad (2)$$

The recovery rate parameter α also determines the average duration of the infectious period, which is equal to $\frac{1}{\alpha}$. Continuing

the previous example of a population of 100 individuals and a prevalence of 0.1 ($I = 10$), with an α of 0.02, each day on average $0.02 \cdot 10 = 0.20$ infected individuals will recover and thus become susceptible again.

Combining transmission and recovery in the above example shows that in this situation the average number of infected individuals increases, and the number of susceptible individuals decreases (i.e., $0.27 > 0.20$). This continues until the prevalence reaches 0.33; after which the average number of infected and susceptible individuals remains constant ($\beta S \frac{I}{N} = 0.03 \cdot 67 \cdot 0.33 = 0.66$, $\alpha I = 0.02 \cdot 33 = 0.66$). The endemic disease is then said to be in equilibrium, the so-called pseudo-steady state in stochastic models. The SIS-model tends to such a dynamic equilibrium, in which the average number of infections is equal to the average number of recoveries. The actual number of susceptible and infective individuals fluctuates around this equilibrium because we have a stochastic model in a finite population. The expected prevalence of the endemic disease is equal to the average proportion of infected individuals in the equilibrium.

Thus, at equilibrium, the transmission rate equals the recovery rate:

$$\beta S \frac{I}{N} = \alpha I \tag{3}$$

The equilibrium prevalence, therefore, follows from solving Equation (3) for $\frac{I}{N}$. Using the fact that $S = N - I$, yields:

$$P_{equilibrium} = \frac{I}{N} = 1 - \frac{\alpha}{\beta} = 1 - \frac{1}{R_0} \tag{4}$$

where $R_0 = \beta / \alpha$, which is the so-called basic reproduction ratio (Diekmann et al. 1990). The prevalence is thus only a function of the ratio of α and β and can therefore also be expressed as a function of the basic reproduction ratio. With variation among individuals in susceptibility (introduced in the next paragraph), Equation (4) is no longer exactly equal to the average prevalence (Biemans et al. 2017).

R_0 is a key parameter in epidemiology. It represents the average number of susceptible individuals that get infected by a single average infectious individual in a totally susceptible population. Hence, R_0 has a threshold value of 1: if $R_0 > 1$ a single infected individual will infect initially more than one new individual, so that the infection can spread in the infection-free population after introduction. On the other hand, if $R_0 < 1$, infectious disease will die out. Hence, also for an endemic disease to persist, R_0 must be greater than one. Moreover, any measure that reduces R_0 to a value smaller than 1 results in eradication of the infectious disease. Obviously, Equation (4) does not apply when $R_0 < 1$. In that case, the disease is absent and prevalence at equilibrium is zero.

Introducing variation in susceptibility into the SIS-model

Here, we describe how variation in disease susceptibility can be introduced among individuals. With equal exposure to infected herd mates, individuals with higher susceptibility are more likely to become infected. Because we will only simulate variation in susceptibility, we define an individual transmission rate parameter for the recipient animal, following (Anche et al. 2014):

$$\beta_i = \beta \gamma_i \tag{5}$$

where β is the transmission rate parameter for the average susceptible individual (infectiousness being the same for all individuals), and γ_i the (relative) susceptibility of individual i. For the average individual, $\gamma_i = 1$, so that β_i is equal to β. Individuals with above-average susceptibility have $\gamma_i > 1$, such that their β_i is greater than β, which means that they have a higher than average probability to become infected (given equal exposure). Accordingly, individuals with below-average susceptibility have $\gamma_i < 1$, $\beta_i < \beta$, and a lower than average probability to become infected. Note that γ_i cannot be lower than 0, since this would result in a negative transmission rate.

Replacing β by β_i in Equation (1), and using $S = 1$ representing the single individual i, yields the transmission rate for individual i exposed to all individuals in the population, of which a fraction $\frac{I}{N}$ is infectious:

$$Rate\ S_i \rightarrow I_i = \beta_i \frac{I}{N} = \beta \gamma_i \frac{I}{N}. \tag{6}$$

We know from the previous paragraph that individual susceptibility (γ_i) must be nonnegative and have a value of 1 for the average individual. For this reason, we simulated γ_i from a log-normal distribution. The log-normal distribution is nonnegative and has positive-skewness, which is often observed for disease traits (Lloyd-Smith et al. 2005). Thus, following Anacleto et al. (2015), we simulated an additive model with normally distributed random effects for the logarithm of susceptibility, so that susceptibility of individual i is:

$$\gamma_i = e^{A_{\gamma_i} + E_{\gamma_i}} \tag{7}$$

where $A_{\gamma i}$ denotes the individual additive genetic effect, or breeding value for the logarithm of susceptibility, and E_{γ_i} the permanent individual environmental effect for the logarithm of susceptibility, with $A_\gamma \sim N(0, \sigma_{A_\gamma}^2)$ and $E_\gamma \sim N(0, \sigma_{E_\gamma}^2)$. Thus $\overline{A_\gamma} = \overline{E_\gamma} = 0$, as common in quantitative genetics. Consequently, $\gamma_i = 1$ for an individual with the mean A_γ and E_γ (but $\overline{\gamma}$ is not exactly 1, because of the positive skewness).

For brevity, we will refer to A_{γ_i} as the "breeding value for susceptibility" or just "breeding value" and to E_{γ_i} as the "individual permanent environmental value for susceptibility" or just "permanent environmental value". Beware though that A_{γ_i} and E_{γ_i} refer to the logarithm of individual susceptibility.

We included a permanent environmental effect for susceptibility, E_γ, to allow for a similarity between repeated records on the same individual that is not purely genetic, but also environmental (e.g., because of developmental effects). This relates to repeatability models as discussed in e.g., Falconer and MacKay (1996). We use the symbol a_P^2 to denote the additive genetic proportion of the full permanent variance in log-susceptibility among individuals,

$$a_P^2 = \frac{\sigma_{A_\gamma}^2}{\sigma_{A_\gamma}^2 + \sigma_{E_\gamma}^2} \tag{8}$$

Simulation of a population with variation in susceptibility

We simulated a population with individual variation in susceptibility. The simulated population consisted of two generations;

parents and offspring. We simulated offspring breeding values according to the infinitesimal model (Fisher 1919):

$$A_{\gamma_i} = \frac{1}{2} A_{\gamma_{sire_i}} + \frac{1}{2} A_{\gamma_{dam_i}} + MS_{\gamma_i} \tag{9}$$

in which $A_{\gamma_{sire}}$ and $A_{\gamma_{dam}}$ were simulated from Normal distributions as shown above, and MS_{γ_i} denotes the Mendelian sampling term, sampled from a normal distribution $MS_{\gamma_i} \sim N\left(0, \frac{1}{2}\sigma^2_{A_\gamma}\right)$. The additive genetic variance for susceptibility was obtained using an iterative procedure tuned to the desired value of observed heritability (see subsection Implementation). The permanent environmental variance in susceptibility was calculated from the additive genetic variance and the chosen additive proportion (a_P^2; Equation 8), $\sigma^2_{E_\gamma} = \sigma^2_{A_\gamma}(1 - a_P^2)/a_P^2$.

Implementation

Our simulations consisted of the iteration of four consecutive steps, aiming to find the additive genetic standard deviation in susceptibility (σ_{A_γ}) that is required to arrive at the desired heritability value of disease status at the observed scale:

1) Stochastic simulation of 20 replicates of a population with a certain genetic standard deviation (σ_{A_γ}) and relative magnitude of genetic permanent effects (a_P^2) for the logarithm of susceptibility.
2) Stochastic dynamic simulation of an endemic infectious disease in herds of the population and recording of binary disease status data for all animals in all replicates.
3) Estimation of the heritability of disease status (h^2_{Obs}) from the recorded disease data with a linear mixed model for all replicates.
4) Change σ_{A_γ} by 0.05 and rerun the entire simulation starting at step 1 if the mean h^2_{Obs} of the 20 replicates is not within two standard errors from either 0.02, 0.05, or 0.1.

Supplementary Figure S2 provides a schematic overview of the above steps.

R (R Core Team 2020) was used for simulation of the population and the infectious disease (steps 1 and 2). The code is available in supplemental files S3 (population simulation) and S4 (infectious disease simulation). The endemic disease was simulated using the Gillespie algorithm (Gillespie 1976), which is the standard procedure for stochastic simulation of infectious diseases (Keeling and Ross 2008). The algorithm simulates the successive infection and recovery events (see section SIS-model without genetic variation) in two steps:

a) Monte Carlo sampling of the time to the next event,
b) Monte Carlo sampling of the type of event (either infection or recovery) and the individual involved.

The probability of sampling an event and the individual involved depends on the individual transmission rates (β_i) and the population recovery rate (α) (see section Introducing variation into the SIS-model). More details are provided in Appendix II.

We started the simulations by infection of a fraction of randomly chosen individuals in each herd. To speed up the simulations and to prevent early dying out of the disease by chance (this occurs when no infected individuals are left in a herd, $I = 0$), we started the simulations near the equilibrium prevalence. To prevent an effect of the initially infected individuals on the parameter estimation, we simulated a burn-in period of 500 days without recording of data for parameter estimation. The recording period was 1,000 days, such that the total simulated period was

1,500 days. We chose this relatively long period to allow for repeated records on disease status, which are essential for the estimation of genetic parameters, given the highly dynamic nature of infectious diseases (Doeschl-Wilson *et al.* 2014; Biemans *et al.* 2017). For the parameter estimation, the disease status of all individuals in each herd was recorded monthly, resulting in 33 records per individual.

In step 3, we estimated the heritability of binary disease status with a linear mixed model including a random animal effect with a pedigree-relationship matrix (ID_i), fitted with ASReml 4.1 (Gilmour *et al.* 2015). Next to the random genetic effect, the model contained a random herd effect ($Herd_k$) and a random nongenetic animal effect (a so-called permanent effect) to account for repeated records on the same individual (IDR_i),

$$y_{ikt} = \mu + Herd_k + ID_i + IDR_i + e_{ikt}, \tag{10}$$

where y_{ikt} is the tth binary disease record (0/1) of individual i present in herd k, μ is the overall mean prevalence of the disease in the population, e_{ikt} is the residual variance at the observed scale.

Input values

In our simulations, the prevalence ($\frac{I}{N}$) and the mean duration of the infectious period ($\frac{1}{\alpha}$) were set to 0.02 day^{-1}, so that the average duration of the infectious period was 50 days, consistent with Biemans *et al.* (2018). This was done to reflect a common endemic disease in livestock: digital dermatitis (DD), a bacterial infectious claw disorder in cattle. The prevalence was set to 0.33, representing herds in the Netherlands that suffer considerably from DD (Holzhauer *et al.* 2006). To obtain a prevalence of 0.33, β was set to 0.03, based on Equation (4). Note that R_0 was thus 1.5, as can be calculated from the prevalence or from $\frac{\beta}{\alpha}$.

In part of the scenarios with a high desired value for h^2_{Obs}, a large variance in susceptibility was needed. In these scenarios, the positive skewness of the lognormal distribution caused the population mean susceptibility (γ_i) to be (much) larger than 1. Since γ acts as a scaling factor on β, $\bar{\gamma_i} > 1$ resulted in $\bar{\beta_i} > \beta$ (Equation 5), which in turn resulted in a prevalence higher than the desired value of 0.33 (Equations 4 and 6). To avoid inconsistencies in prevalence between scenarios, the β was iteratively corrected for each combination of h^2_{Obs} and a_P^2, such that the prevalence was 0.33 in all scenarios. This issue is further addressed in the discussion.

The simulated population consisted of a parental generation of 102 sires, each mated to 102 dams, with one offspring per mating, resulting in a total of 10,404 offspring with a half-sib family structure. The offspring generation was kept in 102 herds of size 102, each consisting of offspring of six randomly sampled sires, each sire contributing 17 offspring. This structure is somewhat similar to a dairy cattle population, but more balanced. The disease was simulated in the offspring generation only.

The relative magnitude of genetic permanent effects (a_P^2) was set to 0.25, 0.50, 0.75, or 1. Since practically nothing is known about the relative magnitude of genetic and nongenetic permanent effects on susceptibility, we simulated values of a_P^2 in the range of 0.25 to 1. Repeatabilities thereby range from equal to the heritability to the 4-fold of the heritability.

Estimation of response to selection

The above procedure results in a genetic standard deviation in susceptibility (σ_{A_γ}) that corresponds to heritabilities of the disease status of 0.02, 0.05, or 0.10. Next, we used this σ_{A_γ} and the same epidemiological model to investigate the potential response

to selection in disease prevalence for this range of observed heritabilities. To do so, we performed additional simulations in which we selected the six best out of the 102 sires, based on their true (*i.e.*, simulated) breeding value (TBV) for susceptibility. Each of these six sires was mated to 1,734 randomly chosen dams, resulting in a total of 10,404 offspring, which were allocated randomly to 102 herds. Thus, there was a selection in sires only. To better illustrate the effect of selection, the starting prevalence in the off-spring generation was kept at 0.33, even though this was higher than the expected prevalence in that generation (since $\gamma_i^- < 1$ after selection). To show trends in the number of infected individuals over time, we used daily records of the entire 1,500 days. Hence, the simulations will illustrate the gradual decrease of prevalence from the initial 0.33 to the new equilibrium value.

Heritability on the underlying scale

The heritability of binary disease status can be defined on two different scales; the observed scale of the binary record, and the underlying scale of the linear additive model for the genetics of susceptibility. The latter is the so-called liability scale (Dempster and Lerner 1950; Robertson 1950). In our simulations, we estimated the observed-scale heritability for binary disease status, but the heritability of the underlying susceptibility was unknown [note that we did not simulate residual variance in susceptibility (Equation 7)]. To provide a heritability value for susceptibility, we calculate heritability on the liability scale in this section. Because our model is additive for log-susceptibility, the liability scale refers to the logarithm of susceptibility. Note, we do not describe new simulations or analyses in this section, but merely motivate the calculation of a liability-scale heritability for our model.

Binary data (y) can be generated from an underlying linear model in two different ways: using a threshold model and using a generalized linear (mixed) model (GLMM). First, in the threshold model, a linear model is used to specify an individual liability, including a residual. Individuals with a liability value greater than a predefined threshold have y = 1; the others have y = 0. Hence, given the threshold, an individual's liability fully determines its observed binary record. Second, in the GLMM, a linear model without a residual is combined with a link function to specify the probability that y = 1. Subsequently, the binary records are sampled from a Bernoulli distribution with this probability. Hence, in this approach, the linear model specifies the expectation of the binary record, $E(y) = P(y = 1)$; not the actual record. Both models are fully equivalent when the link function of the GLMM is the cumulative density function (cdf) that corresponds to the probability density function (pdf) of the residual of the threshold model (De Villemereuil *et al.* 2016). In the classical threshold model, for example, the residual on the liability scale follows a standard normal pdf, while the probit link of a GLMM corresponds to the standard normal cdf. Thus the classical threshold model is equivalent to a GLMM with a probit link [see *e.g.*, Supplementary Information A in De Villemereuil *et al.* (2016)]. The link function in a GLMM "replaces" the liability residual in the threshold model. For this reason, the residual liability variance is also known as the "link variance" (De Villemereuil *et al.* 2016). However, a liability-scale heritability can be defined only based on the threshold model, because the underlying linear model in a GLMM has no residual. (In other words, a GLMM has no "phenotypic" variance on the underlying scale, so the denominator of heritability cannot be determined).

We simulated binary disease status data by specifying the probability of events (infection vs. recovery; see step 2 and Appendix II) using the rates for these events as the rates of a Poisson process (for the infection event this is $1 - e^{-\beta \gamma_i^- \frac{i}{N}}$), which is the GLMM approach. Thus, to find the heritability on the liability scale, we have to translate our model into the corresponding threshold model. GLMM comes with different link functions, each relating to a different pdf of the residual of a corresponding threshold model. As our binary disease data are generated by a Poisson process, the appropriate link function is the complementary log-log. [This is a standard result for GLMM (McCullagh 2019), and motivated in detail for binary disease data in Anche *et al.* (2015)]. Thus, we have to find the residual liability variance of the threshold model that corresponds to a GLMM with a clog-log link function. The link variance of a GLMM with a clog-log link follows from the Gumbel distribution and equals $\frac{\pi^2}{6}$ [Appendix I, Nakagawa *et al.* (2017)].

Therefore, we define the liability for the jth record of individual i as

$$l_{ij} = A_{\gamma_i} + E_{\gamma_i} + e_{l_{i,j}}$$

where $var(e_{l_{i,j}}) = \frac{\pi^2}{6} \approx 1.64$, and heritability on the underlying liability scale of log-susceptibility equals:

$$h^2_{\gamma,liab} = \frac{\sigma^2_{A_\gamma}}{\sigma^2_{A_\gamma} + \sigma^2_{E_\gamma} + \frac{\pi^2}{6}} \tag{11}$$

Data availability

Supplemental Material available at figshare: https://doi.org/10.25386/genetics.13090076. Supplementary Table S1 contains heritability estimates for different genetic standard deviations in the logarithm of susceptibility. Supplementary Figure S2 provides a schematic overview of our methodology. Supplemental File S3 contains the R-code for simulation of a population with genetic variation in the logarithm of susceptibility. Supplemental File S4 contains the R-code for simulation of the endemic disease.

Results

This section starts with the genetic variances and corresponding liability heritabilities for the logarithm of susceptibility that correspond to observed heritabilities of 0.02, 0.05, and 0.10 for disease status. Subsequently, we show to what extent individual breeding values for susceptibility are reflected in individual disease status. Next, we show the response to selection for each value of the observed heritability. Finally, we illustrate the mechanisms underlying the observed response to selection, using simulations of herds consisting of individuals with similar susceptibility.

Genetic variation in susceptibility

Table 2 shows the genetic standard deviations in the logarithm of susceptibility (σ_{A_γ}) and liability heritabilities ($h^2_{\gamma, liab}$) that correspond to observed scale heritabilities of binary disease status (h^2_{obs}) of 0.02, 0.05 and 0.1. Results are shown for additive genetic proportions of the permanent variance (a^2_P) of 1, 0.75, 0.5 and 0.25. To illustrate the magnitude of the genetic variation in susceptibility, Table 2 also shows the genetic values for R_0 for the 10% of individuals with the highest breeding values and the 10% individuals with the lowest breeding values for susceptibility. These R_0-values are relevant for the prevalence of the infectious disease (see Equation 4) and give an indication of the potential response to selection. In Supplementary Table S1 actual h^2_{obs}

Table 2 Genetic standard deviation in susceptibility required for observed scale heritability of disease status of 0.02, 0.05, and 0.10

h^2_{obs}	0.02				0.05				0.10			
a^2_P	σ_{A_γ}	$h^2_{\gamma,liab}$	$R_{0_{-10\%}}$	$R_{0_{+10\%}}$	σ_{A_γ}	$h^2_{\gamma,liab}$	$R_{0_{-10\%}}$	$R_{0_{+10\%}}$	σ_{A_γ}	$h^2_{\gamma,liab}$	$R_{0_{-10\%}}$	$R_{0_{+10\%}}$
1.0	0.30	0.05	0.88	2.54	0.50	0.13	0.62	3.62	0.75	0.25	0.40	5.62
0.75	0.30	0.05	0.88	2.54	0.50	0.13	0.62	3.62	0.75	0.23	0.40	5.62
0.50	0.30	0.05	0.88	2.54	0.50	0.12	0.62	3.62	0.85	0.23	0.34	6.70
0.25	0.30	0.04	0.88	2.54	0.55	0.11	0.57	3.95	1.10	0.19	0.22	10.4

Required genetic standard deviation in the logarithm of susceptibility (σ_{A_γ}), liability heritability for susceptibility $h^2_{\gamma,liab}$ and R_0 for the 10% of individuals with the highest ($R_{0_{+10\%}}$) and 10% with the lowest ($R_{0_{-10\%}}$) values for susceptibility for each combination of a^2_P(rows) and h^2_{obs} (columns).

estimates for the different σ_{A_γ} are given with corresponding standard errors.

Table 2 shows that common observed scale heritability values for binary disease status correspond to a substantial genetic variation in the logarithm of susceptibility (σ_{A_γ}). An observed scale heritability of 0.02, for example, corresponds to a genetic standard deviation in susceptibility of 0.3. With a σ_{A_γ} of 0.3, the mean breeding value of the 10% of individuals with the lowest breeding values for susceptibility (A^-_γ) is –0.53, accordingly, the mean breeding value of the 10% of individuals with the highest breeding values (A^+_γ) is 0.53. Corresponding genetic susceptibility values ($\gamma^- = e^{A^-_\gamma}$, $\gamma^+ = e^{A^+_\gamma}$) for these individuals are 0.59 and 1.70, which results in values for R_0 of 0.59*1.5 = 0.88 and 1.7*1.5 = 2.54 respectively. There is thus an almost 3-fold genetic difference in R_0.between the highest and lowest 10% of individuals. This difference is substantial in itself, but more importantly, R_0 for the 10% individuals with the lowest susceptibility are below 1, which means that an individual with such susceptibility, on average, infects less than 1 herd mate. Implying that an infectious disease will die out. For higher values of observed heritability, the corresponding genetic standard deviation increased considerably, resulting in values for R_0 even further below 1 for a h^2_{obs} of 0.05 and 0.10.

The relative magnitude of genetic versus environmental permanent effects on susceptibility (a^2_p) only had a minor effect on the observed heritability, as illustrated by the similar values of the genetic standard deviation in susceptibility needed to obtain a certain observed heritability. The effect of a^2_p was substantial only for an observed scale heritability of 0.10. In this case, we needed a considerably higher genetic standard deviation in susceptibility for lower values of a^2_p, especially for a $a^2_p = 0.25$. Similarly, the liability heritability is quite consistently, approximately twice the value of h^2_{obs}, across the different values of a^2_p.

In conclusion, a remarkably large genetic variation in susceptibility is needed to reproduce common observed scale heritabilities for disease status. As a consequence, a considerable proportion of the population had genetic R_0 values smaller than one.

The relationship between susceptibility and individual disease status

Figure 2 shows the relationship between breeding value for susceptibility and disease prevalence (over all herds) when herds are composed of offspring of random sires. The figure shows the prevalence over time (mean of 20 replicates) in the 10% most and 10% least susceptible individuals in the population, for $a^2_p = 1$. Results were very similar for other values of a^2_p, as expected given the minor effect shown above.

The distance between two lines increases with h^2_{obs}, because a higher h^2_{obs} corresponds to a greater genetic variation in susceptibility. For an observed heritability of 0.02, the prevalence in the
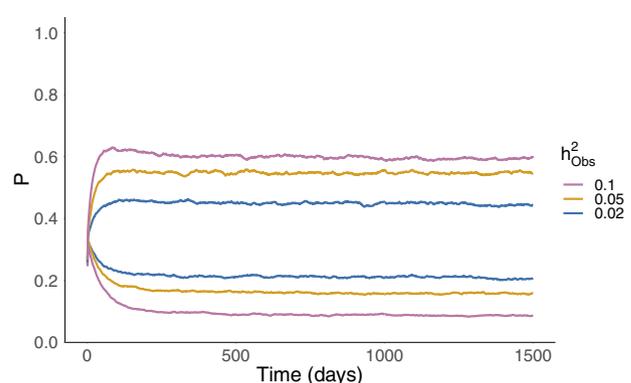

Figure 2 Population prevalence with random herds. Prevalence (P) of the infectious disease in the total population (over all herds) for a^2_p of 1 and different h^2_{obs}. Lines indicate the 10% of individuals with highest and the 10% with lowest breeding value for susceptibility for the respective h^2_{obs}, herds formed by random allocation of sires.
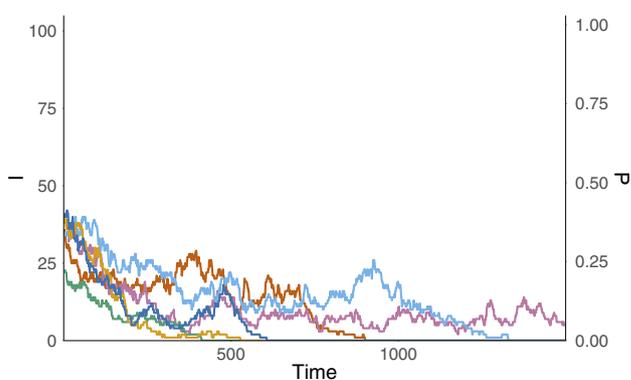

Figure 3 Prevalence in six herds after sire selection. Number of infected individuals (I) and prevalence (P) over time in six representative herds, represented by the different colors, consisting of offspring from six sires selected for low susceptibility. h^2_{obs} of 0.05, a^2_p of 1. Each herd consists of 102 individuals.

10% most susceptible individuals is about a factor of 2 higher than in the 10% least susceptible individuals (25% vs. 50%). This difference of about 25% points is considerable given the very low heritability. For an observed heritability of 0.10, the distance increases up to a factor of 4 (15% vs. 60%). Hence, though a heritability of 0.10 is only a moderate value, the corresponding difference in prevalence is large.

A comparison of Figure 2 and Table 2 shows that the actual prevalence in the 10% most and least susceptible individuals (Figure 2) is considerably different from the prevalence that is expected based on the values for R_0 (Table 2). For a heritability of 0.10, for example, the actual prevalence is about 0.6 in the 10%
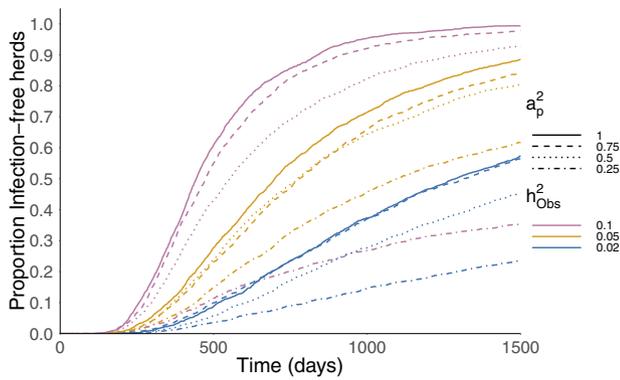
Figure 4 Infection-free herds after sire selection. Proportion of infection-free herds over time for all combinations of h_{obs}^2 and a_P^2, herds consisting of offspring from six sires selected for low susceptibility. Different colors represent different h_{obs}^2, different line types represent different a_P^2. The initial prevalence was 0.33 in each herd.
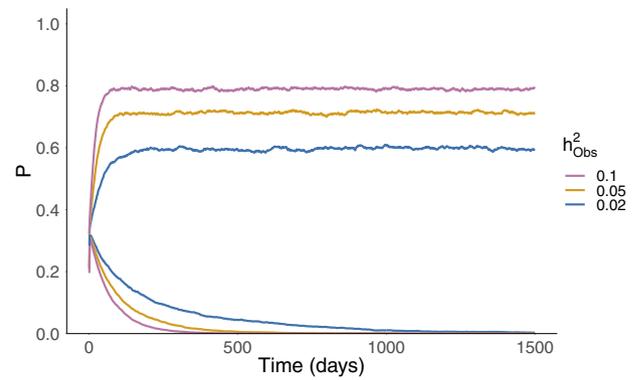

Figure 5 Population prevalence with nonrandom herds. Prevalence (P) of the infectious disease in the total population (over all herds) for a_P^2 of 1 and different h_{obs}^2 Lines indicate the 10% of individuals with highest and the 10% with lowest breeding value for susceptibility for the respective h_{obs}^2, herds formed by allocation of individuals based on breeding value for susceptibility.

most susceptible and about 0.15 in the 10% least susceptible individuals, while the expected prevalence based on R_0 is 0.82 and 0, respectively (see Table 2 and Equation 4). This indicates that the genetic susceptibility values are not completely reflected in the disease status of individuals. This phenomenon occurs because individual disease status depends not only on individual susceptibility, but also on exposure to infectious herd mates. With randomly composed herds, individuals experience similar exposure to infectious herd mates, irrespective of their susceptibility. For this reason, differences in prevalence between subgroups in the population were smaller than expected based on their genetic value for R_0.

Response to selection

To investigate the effects of selection for lower susceptibility on the prevalence of the disease, the six sires with the lowest true breeding values for susceptibility were selected to produce all offspring. Figure 3 illustrates the effect of this single generation of sire selection on the prevalence in six herds for the scenario with a $a_P^2 = 1$, ($h_{\gamma,liab}^2 = 0.13$) and a $h_{obs}^2 = 0.05$. Herds were selected such that they represent the different prevalence patterns we observed for this scenario in the simulations. The figure shows that the prevalence in all herds fluctuates around a decreasing trend from the starting prevalence of 0.33. Eventually, the infection "dies out" in some herds, which means that there are no infectious individuals anymore to sustain infection.

For a more extensive investigation of response to selection, Figure 4 shows the proportion of infection-free herds over time (mean of 20 replicates), for all 12 scenarios. For each scenario, we selected the six sires with the lowest true breeding value for susceptibility to produce all offspring. The figure clearly shows that the disease disappeared from a substantial proportion of herds. Even when the observed heritability was only 0.02, 20 to 50 of the 102 herds became infection-free after a single generation of sire selection.

The response was larger with higher h_{obs}^2; for the highest value ($h_{obs}^2 = 0.10$, $a_P^2 \neq 0.25$), the infection was eliminated from practically all herds. Again the relative magnitude of genetic permanent effects (a_P^2) only had a considerable effect when it was 0.25. The lines for the other values of a_P^2 are very close to each other for a given h_{obs}^2. The response for $a_P^2 = 0.25$ and $h_{obs}^2 = 0.10$ seems not in accordance with this general pattern. For example, it is lower than the response for $a_P^2 = 0.25$ and $h_{obs}^2 = 0.05$. This is most likely a consequence of the very large genetic variance in

susceptibility required to reproduce a heritability of disease status of 0.10 for that scenario, together with the lognormal distribution of γ (see Discussion).

The mechanism underlying response to selection

Results in Figure 4 show that the responses to selection are considerably greater than the differences in disease status between the 10% most and 10% least susceptible individuals within a generation (Figure 2). We hypothesize that this difference originates from positive feedback effects in the transmission dynamics, resulting in some degree of herd immunity in the results in Figure 4. To investigate this hypothesis and clarify the mechanism underlying these unexpectedly large responses to selection, we simulated a population where individuals were grouped into herds based on their breeding value for susceptibility. The first herd consisted of the 102 least susceptible individuals, the second herd of the 102 individuals with the second lowest breeding values for susceptibility, and so forth, until the last herd, consisting of the 102 most susceptible individuals. The herds consisting of the least susceptible individuals resemble the herds after selection, since herds consist entirely of individuals with low susceptibility in both cases. Hence, similar to the case with selection, individuals with low susceptibility are accompanied by herd mates with low susceptibility, which reduces their exposure to the infectious agent.

Figure 5 shows the prevalence in herds composed of the 10% most and 10% least susceptible individuals based on breeding value for susceptibility. Comparison to Figure 2 shows that this formation of herds substantially increases the difference in prevalence between the most and least susceptible individuals. For an h_{obs}^2 of 0.02, for example, the difference increased from 25% versus 50% in random herds to 5% versus 60% in sorted herds. For an h_{obs}^2 of 0.10, it increased from 15% versus 60% to $\sim$ 0% versus 80%. The prevalence in the top and bottom individuals in Figure 5 is more in accordance with the expected prevalence based on the values for R_0 (Table 2). The bottom lines in Figure 5, which represent the herds consisting of the least susceptible individuals, show a clear decreasing prevalence over time, and reach zero for all three observed heritabilities. This fully agrees with the R_0 values below 1 for these individuals.

This observation clearly shows that the susceptibility of herd mates has a considerable effect on the disease status of an individual. An individual is more often infected when its herd mates

have a higher than average susceptibility, and less often when its herd mates have a lower than average susceptibility, on top of the effect of the susceptibility of the individual itself. This mechanism explains the larger than expected response to selection; when all individuals in a herd descend from superior parents, not only the individual itself will be less susceptible, but it will also be accompanied by herd mates that are less often infected. This reduces the exposure of the individual to the infectious agent. These results show that susceptibility not only has a direct effect on the disease status of the individual itself, but also an Indirect Genetic Effect (IGE) on the disease status of its herd mates (see also Anche *et al.* 2014).

Discussion

Here, we showed that low-heritability estimates for infectious disease status (0/1 = healthy/diseased) are fully consistent with a large amount of genetic variation in disease susceptibility. The genetic variation needed to reproduce an observed-scale heritability of only 0.02 roughly corresponds to R_0-values of 2.5 and 0.9 in the 10% top and 10% bottom-ranking individuals. This large difference in R_0 corresponded to a large reduction in prevalence of the disease after selection, and even eradication of the disease occurred after a single generation of sire selection in our simulations.

The possibility to arrive at a prevalence of zero using selection, *i.e.*, herd-level eradication, is an important result. It contradicts predictions based on classical quantitative genetic models for binary traits that do not account for the transmission of infection between individuals, such as the classical threshold model by Dempster and Lerner (1950). For binary traits, the classical model shows that the observed-scale heritability approaches zero when prevalence approaches zero or one (Robertson 1950). Hence, continued selection against an infectious disease will reduce the heritability in the threshold model, so that response to selection will approach zero as well, and it is impossible to reach a prevalence of zero.

The difference between our findings and predictions based on classical models originates from positive feedback effects occurring in the transmission of infectious diseases. These feedback effects entail that individuals with a low susceptibility are not only less likely to become infected themselves, but they also infect fewer herd mates, just because they are less likely to be in the infectious state. This shows that genetic variation in disease susceptibility leads to so-called Indirect Genetic Effects (IGE). In general, IGE is an effect of the genotype of an individual on the trait values of other individuals (Griffing 1967; Moore *et al.* 1997; Muir 2005; Bijma 2014). The fact that genetic variation in susceptibility inevitably leads to IGE was also shown by (Anche *et al.* 2014) and (Bijma 2020). The larger difference in prevalence between the least and most susceptible individuals when allocation to herds was based on breeding value for susceptibility, compared to allocation at random, illustrates this IGE of susceptibility.

The response to selection we found can also be placed in the more general framework of the Price equation (Price 1970). The Price equation states that response to selection in a trait is the sum of two components. The first component represents the contribution directly attributable to selection. The second term represents the effect due to "incomplete fidelity of transmission of the trait value to the next generation" (Gardner 2020). This transmission term defines nonselective effects, for example, a difference in nonadditive effects between the parent and offspring

generation, due to a change in allele frequency. The response to selection in prevalence can be described in two ways using the Price equation, depending on whether the IGE of susceptibility is incorporated into the selection term or into the transmission term (Bijma 2020). The common breeding values for disease status from the linear animal model do not capture the IGE of susceptibility and therefore represent the latter case, where the IGE is considered a nonselective effect arising from a change in the environment (less exposure to infectious individuals in the population). However, one can also define a so-called total breeding value for prevalence, including both the direct and indirect genetic effect of susceptibility (Bijma 2011, 2020). In this approach, the indirect effect of susceptibility is shifted into the selection term. The latter approach makes sense, because the IGE due to genetic variation in susceptibility is a special kind, which arises entirely via the direct genetic effect on the disease status of the individual itself. Hence, the direct and indirect genetic effects due to susceptibility are fully correlated, such that selection on susceptibility is automatically on the indirect effect as well. This complete correlation also increases the total genetic variation in disease status (Bijma 2010). Consequently, the breeding value for prevalence predicts a larger response in prevalence because of selection on susceptibility than the ordinary breeding values for disease status. Especially at lower prevalence, the selection differential in susceptibility required to eliminate a disease is much smaller than expected based on the classical breeding value for disease status (Bijma 2020).

The positive feedback mechanism described above is well known in epidemiology, with herd immunity ($R_0 < 1$) as the most prominent example (Fine 1993). As illustrated by the eradication of rinderpest in cattle, it is not necessary that all individuals are fully resistant to infection to reach herd immunity. If a sufficient fraction of the population is vaccinated, a disease will have no possibility to transmit, because there are not enough sufficiently susceptible individuals in the population to sustain transmission. For herd immunity (*i.e.*, for R_0), it is in principle irrelevant whether a certain reduction in susceptibility is obtained by genetic selection or by vaccination causing incomplete immunity, because the positive feedback mechanisms underlying herd immunity result from the reduction in susceptibility, independent of how the reduction is obtained, and are therefore equally present in both cases. Thus, as with vaccination, a sufficient fraction of the population should have a sufficiently low susceptibility to reach herd immunity using genetic selection.

In our model, we simulated genetic variation in susceptibility only. Next to susceptibility, the transmission and prevalence of infectious diseases is affected by two other host traits: infectivity and the duration of the infectious period (Doeschl-Wilson *et al.* 2011). Infectivity is the propensity of an infected individual to transmit the disease to a susceptible individual per unit of time. It has an impact only on the disease status of other individuals, not on the disease status of the individual itself. In other words, infectivity only has an indirect genetic effect, no direct genetic effect. Consequently, the common genetic analysis of the binary disease status of the individual does not capture genetic variation in infectivity (Lipschutz-Powell *et al.* 2012). Variation in infectivity thereby does not contribute to the established values of the heritability of disease status, which were the starting point of our analysis. Incorporating variation in infectivity in our simulations would therefore not change the genetic variation in susceptibility needed to arrive at the target observed heritabilities. However, if host infectivity shows genetic variation and it is incorporated in selection, this would increase the potential of the population to

respond to selection for lower disease prevalence (*e.g.*, Tsairidou *et al.* 2019).

The second trait, the duration of the infectious period, relates to the ability of an individual to recover from infection, and determines the time an individual stays in the infectious state. In contrast to infectivity, variation in duration of the infectious period is directly reflected in an individual's disease status. In fact, it is the reverse of susceptibility, since it determines the rate at which individuals change from the infected to the susceptible state. In an earlier analysis with the same SIS-model, we found that the effect of variation in the duration of the infectious period on the heritability of disease status is comparable to that of susceptibility. Moreover, just like susceptibility, the duration of the infectious period clearly has an indirect effect as well. If individuals have a short infectious period, they are less likely to infect others, just because they are in the infectious state for a shorter period of time. Because the effects of the duration of the infectious period are similar to those of susceptibility, we, therefore, chose to simulate the genetic variation in susceptibility only.

In some scenarios, we needed to simulate a large variance in susceptibility to reproduce the desired heritability of disease status. A large variance in susceptibility had two counteracting effects on the prevalence of the disease in our simulations: (1) an increase in prevalence because $\bar{\gamma}$ is much larger than 1 leading to a higher R_0; (2) a decrease in prevalence because increasing heterogeneity among individuals reduces the prevalence (Springbett *et al.* 2003). With a small variance these two effects balanced each other, but with increasing variation the first effect became dominant, such that the actual prevalence in some of the simulations was considerably higher than the desired value of 0.33. To prevent inconsistencies in the estimation of heritability and response to selection, we needed to correct for this higher prevalence, to obtain a prevalence of 0.33 in all scenarios. A prevalence much lower or higher than 0.33, close to 0 or 1, would have the effect that much higher genetic variances are needed to reach our target observed heritabilities. On first sight, the correction should ensure that the property $\beta \bar{\gamma_i}$ is equal to 0.03, either by decreasing β to $0.03 \bar{\gamma_i}$ or by setting $\bar{\gamma_i}$ to 1 (via the introduction of an extra term in Equation 7). However, such a correction of either β or $\bar{\gamma_i}$ resulted in a prevalence (much) lower than 0.33 because of the decreasing effects of heterogeneity on prevalence. We, therefore, chose to iteratively correct β until actual prevalence was 0.33 in all scenarios.

We observed a considerably lower response to selection for $a_p^2 = 0.25$ and $h_{obs}^2 = 0.10$ than for the other scenarios. The total (genetic + environmental) permanent susceptibility variance needed to reproduce $h_{obs}^2 = 0.10$ was exceptionally large for this scenario. It could even be considered unrealistically high, since it corresponds to a coefficient of variation of 111% which is above maximum values observed in literature (*e.g.*, Houle 1992). Mean population susceptibility (γ_i) was much larger than 1 for this variance, because of the skewness of the lognormal distribution of γ. We prevented effects of this higher mean susceptibility on the prevalence by the correction of β described in the previous paragraph. In case of selection, however, (extremely) high individual dam and environmental effects, resulting from the highly skewed distribution of susceptibility, cause a much higher average susceptibility in the offspring than would be expected from the mean breeding value of the selected sires. This higher susceptibility in the offspring resulted in a larger than expected $\beta \bar{\gamma_i}$, and consequently in a lower proportion of infection-free herds compared to the other scenarios. Because the combination of $a_p^2 = 0.25$ and

$h_{obs}^2 = 0.10$ leads to unrealistically high variation in susceptibility, we feel this issue is not very relevant.

To illustrate the potential response to selection, we selected sires on their true (simulated) breeding values for susceptibility. We took this approach to reveal the additive genetic variance available for genetic improvement and to clarify the mechanisms underlying response to selection, without interference of the accuracy of breeding value estimation. However, the correlations between the estimated breeding values for disease status from the linear model and the true breeding values for log-susceptibility of sires were between 0.75 (h_{obs}^2 of 0.02) and 0.90 (h_{obs}^2 of 0.10). This indicates that differences due to selecting sires based on their EBV for disease status instead of selecting them on their TBV for susceptibility would be relatively small. These high accuracies likely result from the large number of offspring per sire in our simulations and the intermediate prevalence of the disease, since individual differences in susceptibility are best visible at intermediate prevalence. If the prevalence is close to zero, the accuracy of EBVs will be smaller.

Even though our results clearly show that selection against infectious diseases should be much more promising than the common low heritability for disease status suggests, questions may arise as to whether our conclusions are achievable in practice and are not too optimistic. These questions might be motivated by the limited availability of empirical examples of large response to selection in disease prevalence. Even though the common quantitative models ignore indirect genetic effects, the previous paragraph shows that they still relatively accurately ranked the individuals on their susceptibility, which suggests that more examples of large response might be available than the two we found and mentioned in the introduction. In the next paragraphs, we will identify important aspects related to the visibility and validity of our results in real data examples.

A first important point is that the feedback effects underlying our response are only effective if the entire herd is selected for low susceptibility. If only part of a herd is selected for low susceptibility, the offspring of selected parents are still exposed to (the offspring of) nonselected individuals with higher susceptibility. In practice, this might for instance be the case in dairy cattle herds, where individuals usually belong to different generations. Another complicating factor is that herd-year effects in models for breeding value estimation may hide the feedback effects. When the model contains the direct genetic effect only, the IGE-component of genetic differences in the prevalence of the disease between herds, or between consecutive years of the same herd, will end up in the herd-year effect, and thus be attributed to differences in herd management. Hence, response due to positive feedback (*i.e.*, due to IGE) is difficult to see in classical quantitative genetic analysis.

A key assumption underlying our results is that the pathogen can replicate only in the host individual, meaning that a reduction in individual host susceptibility fully translates in reduced exposure of the host population to the pathogen. When a pathogen is able to replicate outside the host population, for instance in the environment or another species, the impact of the feedback effects on transmission will be smaller. An example of such a case is Bovine Tuberculosis, where badgers are an external reservoir in which the pathogen can replicate (Böhm *et al.* 2009). Because the key mechanism underlying our results is the positive feedback of selection for lower susceptibility, we expect that as long as no external replication occurs, our conclusions remain valid. For instance, in case where the transmission process is delayed because of a latency period between the moment of

infection and the time an individual becomes infectious to others, or when the pathogen can survive but not replicate in the environment. The feedback effects in transmission are still present in those cases; only the total duration of the infection-cycle is prolonged (Ma and Earn 2006).

A general problem in relation to management of infectious diseases is the evolutionary response of pathogens to the applied interventions, such that these interventions become less effective. Pathogens are known to be able to adapt to every type of intervention, with the widespread antibiotic resistance as probably the best-known example (Davies and Davies 2010; Kennedy and Read 2017). Here we discuss briefly what breeders could do to minimize the risk of pathogen adaptation. Importantly, evolution of pathogen resistance can only occur when there is transmission. Consequently, the most promising interventions with respect to prevention of resistance development are those that prevent transmission to occur ($R_0 < 1$) as soon as possible. The higher the selection pressure put on the pathogen population, especially when targeting the pathogen in multiple ways, the lower the probability of development of resistance (REX Consortium 2013). A well-known example is combination therapy for HIV, which is relatively successful in preventing pathogen resistance development by targeting the pathogen in multiple ways (Kennedy and Read 2017).

With respect to genetic selection, the above paragraph implies that strong selection for polygenic resistance should minimize the risk of pathogen adaptation. Plant breeders, for example, recognize that breeding for broad-spectrum resistance is the most sustainable way to manage infectious diseases such as potato blight (Vleeshouwers *et al.* 2011). In livestock populations, results of Genome-Wide Association Studies show that important endemic infectious diseases, such as mastitis and digital dermatitis are highly polygenic traits (Tiezzi *et al.* 2015; Biemans *et al.* 2019). This would argue for a few generations of strong selection for lower susceptibility of the host population. However, such selection is currently uncommon in practice because of undesired correlated responses in other traits in the breeding goal, such as yield traits. For mastitis, for instance, the correlation of disease status with milk yield is positive, with most estimates between 0.20 and 0.55 (Rupp and Foucras 2010). Selection solely for a lower prevalence of mastitis would thus result in a negative (correlated) response in yield. Nevertheless, many questions about the adaption of pathogens to artificial selection in livestock are still unanswered.

The results of our simulations provide a new perspective on genetic selection for lower infectious disease prevalence, and show that this might lead to much better results than current quantitative genetic models predict. Because of the complicating factors described in the above paragraphs, carefully designed selection experiments are needed to confirm our results with real data examples. Next to confirmation of our results with real data, work into the genetic background of the disease traits is needed as well, particularly on the presence of genetic variation in infectivity (Lipschutz-Powell *et al.* 2012). The use of a method for statistical genetic analysis that accounts for the dynamic nature of transmission and for differences in infection exposure between individuals is essential here, and might also increase the response to selection. Examples of such methods are models using Bayesian inference (Pooley *et al.* 2020) and Generalized linear mixed models (Biemans *et al.* 2019).

In this study, we used simulations of an established epidemiological model to demonstrate that genetic selection against infectious diseases is much more promising than expected based on commonly used quantitative genetic models. As most models of biological systems, it will be needed to tailor our model to fit certain specific cases. However, by incorporating the positive feedback effects that have been demonstrated time and again in the field of epidemiology, our model provides a fundamentally better description of infectious disease transmission than the classical quantitative genetic models for binary traits, which do not account for feedback effects. The main implication for breeders of this work is that a low ordinary additive genetic variance in binary disease status, as often observed in practice, should not be interpreted as a limiting factor for potential response to selection. Instead, we showed that these low values are fully consistent with a large amount of available genetic variation in disease susceptibility, which in turn translates into a larger than expected response to selection in the prevalence of the disease. Feedback effects occurring in transmission are of crucial importance for this response, and make it possible to eradicate infectious disease, at least in theory.

Acknowledgments

Literature cited

Anacleto O, Garcia-Cortés LA, Lipschutz-Powell D, Woolliams JA, Doeschl-Wilson AB. 2015. A novel statistical model to estimate host genetic effects affecting disease transmission. Genetics 201: 871–884.

Anche M, De Jong M, Bijma P. 2014. On the definition and utilization of heritable variation among hosts in reproduction ratio R 0 for infectious diseases. Heredity 113:364–374.

Anche MT, Bijma P, De Jong MC. 2015. Genetic analysis of infectious diseases: estimating gene effects for susceptibility and infectivity. Genet Sel Evol. 47:85.

AquaGen 2012. Field Results of QTL-innOva IPN – From Egg To Harvest. Trondheim: Aqua Gen AS.

Bennett R. 2003. The 'direct costs' of livestock disease: the development of a system of models for the analysis of 30 endemic livestock diseases in Great Britain. J Agric Economics 54:55–71.

Biemans F, Bijma P, Boots NM, de Jong MC. 2018. Digital dermatitis in dairy cattle: the contribution of different disease classes to transmission. Epidemics 23:76–84.

Biemans F, de Jong M, Bijma P. 2019. A genome-wide association study for susceptibility and infectivity of Holstein Friesian dairy cattle to digital dermatitis. J Dairy Sci. 102:6248–6262.

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

Bijma P. 2010. Estimating indirect genetic effects: precision of estimates and optimum designs. Genetics 186:1013–1028.

Bijma P. 2011. A general definition of the heritable variation that determines the potential of a population to respond to selection. Genetics 189:1347–1359.

Bijma P. 2014. The quantitative genetics of indirect genetic effects: a selective review of modelling issues. Heredity 112:61–69.

Bijma P. 2020. The Price equation as a bridge between animal breeding and evolutionary biology. Philos Trans R Soc Lond B Biol Sci. 375:20190360.

Bishop SC, Axford RF, Nicholas FW, Owen JB. 2010. Breeding for Disease Resistance in Farm Animals. Wallingford: CABI.

Bishop SC, Woolliams JA. 2010. On the genetic interpretation of disease data. PLoS ONE 5:e8940.

Böhm M, Hutchings MR, White PCL. 2009. Contact networks in a wildlife-livestock host community: identifying high-risk individuals in the transmission of bovine TB among badgers and cattle. PLoS ONE 4:e5016.

Consortium REX. 2013. Heterogeneity of selection and the evolution of resistance. Trends Ecol Evol. 28:110–118.

Core Team R. 2020. R: A Language and Environment for Statistical Computing. Vienna: Austria: R Foundation for Statistical Computing.

Davies J, Davies D. 2010. Origins and evolution of antibiotic resistance. Microbiol Mol Biol Rev. 74:417–433.

De Villemereuil P, Schielzeth H, Nakagawa S, Morrissey M. 2016. General methods for evolutionary quantitative genetic inference from generalized mixed models. Genetics 204:1281–1294.

Dempster ER, Lerner IM. 1950. Heritability of threshold characters. Genetics 35:212–236.

Diekmann O, Heesterbeek H, Britton T. 2012. Mathematical Tools for Understanding Infectious Disease Dynamics. Princeton, NJ: Princeton University Press.

Diekmann O, Heesterbeek JAP, Metz JA. 1990. On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. J Math Biol. 28: 365–382.

Doeschl-Wilson A, Lipschutz-Powell, O. Anacleto, G D, Lough, Lengeling A. 2014. New methods for capturing unidentified genetic variation underlying infectious disease in livestock populations. Proceedings of 10th World Congress on Genetics Applied to Livestock Production. Vancouver, BC, Canada: WCGALP.

Doeschl-Wilson AB, Davidson R, Conington J, Roughsedge T, Hutchings MR, et al. 2011. Implications of host genetic variation on the risk and prevalence of infectious diseases transmitted through the environment. Genetics 188:683–693.

Doeschl-Wilson AB, Kyriazakis I. 2012. Should we aim for genetic improvement in host resistance or tolerance to infectious pathogens? Front Genet. 3:272.

Falconer DS, MacKay TFC. 1996. Introduction to Quantitative Genetics. Harlow: Pearson Education Limited.

Fine PE. 1993. Herd immunity: history, theory, practice. Epidemiol Rev. 15:265–302.

Fisher RA. 1919. The correlation between relatives on the supposition of mendelian inheritance. Trans R Soc Edinb. 52:399–433.

Gardner A. 2020. Price's equation made clear. Philos Trans R Soc Lond B Biol Sci. 375:20190361.

Gianola D. 1982. Theory and analysis of threshold characters. J Anim Sci. 54:1079–1096.

Gillespie DT. 1976. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J Comp Physics 22:403–434.

Gilmour A, Gogel B, Cullis B, Welham S, Thompson R. 2015. ASReml User Guide Release 4.1 Structural Specification. Hemel hempstead: VSN international ltd.

Griffing B. 1967. Selection in reference to biological groups I. Individual and group selection applied to populations of unordered groups. Aust J Biol Sci. 20:127–140.

Halloran ME, Haber M, Longini IM, Jr, 1992. Interpretation and estimation of vaccine efficacy under heterogeneity. Am J Epidemiol. 136:328–343.

Henryon M, Heegaard PM, Nielsen J, Berg P, Juul-Madsen HR. 2006. Immunological traits have the potential to improve selection of pigs for resistance to clinical and subclinical disease. Anim Sci. 82:597–606.

Heringstad B, Klemetsdal G, Steine T. 2007. Selection responses for disease resistance in two selection experiments with Norwegian red cows. J Dairy Sci. 90:2419–2426.

Hethcote HW. 1989. Three basic epidemiological models. In: Levin SA, Hallam TG, Gross LJ editors. Applied Mathematical Ecology. Berlin: Springer, pp. 119–144.

Holzhauer M, Hardenberg C, Bartels C, Frankena K. 2006. Herd-and cow-level prevalence of digital dermatitis in the Netherlands and associated risk factors. J Dairy Sci. 89:580–588.

Houle D. 1992. Comparing evolvability and variability of quantitative traits. Genetics 130:195–204.

Keeling MJ, Ross JV. 2008. On methods for studying stochastic disease dynamics. J R Soc Interface 5:171–181.

Kennedy DA, Read AF. 2017. Why does drug resistance readily evolve but vaccine resistance does not? Proc Biol Sci. 284:20162562.

Knap PW, Bishop SC. 2000. Relationships between genetic change and infectious disease in domestic livestock. BSAP Occas Publ. 27: 65–80.

Lipschutz-Powell D, Woolliams JA, Bijma P, Doeschl-Wilson AB. 2012. Indirect genetic effects and the spread of infectious disease: are we capturing the full heritable variation underlying disease prevalence? PLoS ONE 7:e39551.

Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. 2005. Superspreading and the effect of individual variation on disease emergence. Nature 438:355–359.

Ma J, Earn DJ. 2006. Generality of the final size formula for an epidemic of a newly invading infectious disease. Bull Math Biol. 68: 679–702.

Mariner JC, House JA, Mebus CA, Sollod AE, Chibeu D, et al. 2012. Rinderpest eradication: appropriate technology and social innovations. Science 337:1309–1312.

Martin P, Barkema H, Brito L, Narayana S, Miglior F. 2018. Symposium review: novel strategies to genetically improve mastitis resistance in dairy cattle. J Dairy Sci. 101:2724–2736.

McCullagh P. 2019. Generalized Linear Models. London: Routledge.

Moore AJ, Brodie, ED IIIWolf JB. 1997. Interacting phenotypes and the evolutionary process: I. Direct and indirect genetic effects of social interactions. Evolution 51:1352–1362.

Muir WM. 2005. Incorporation of competitive effects in forest tree or animal breeding programs. Genetics 170:1247–1259.

Nakagawa S, Johnson PC, Schielzeth H. 2017. The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. J R Soc Interface 14:20170213.

Nieuwhof GJ, Conington J, Bishop SC. 2009. A genetic epidemiological model to describe resistance to an endemic bacterial disease in livestock: application to footrot in sheep. Genet Sel Evol. 41:19.

Pooley CM, Marion G, Bishop SC, Bailey RI, Doeschl-Wilson AB. 2020. Estimating individuals' genetic and non-genetic effects underlying infectious disease transmission from temporal epidemic data. PLOS Computational Biology 16: e1008447.

Price GR. 1970. Selection and covariance. Nature 227:520–521.

Robertson A. 1950. Proof that the additive heritability on the P scale is given by the expression Z2hx2/pq. Genetics 35:234–236.

Rupp R, Foucras G. 2010. Genetics of mastitis in dairy ruminants. In: SC Bishop, RF Axford, FW Nicholas, Owen JB editors. Breeding for Disease Resistance in Farm Animals, 3rd edn. Wallingford: CABI. pp. 183–212.

Speksnijder D, Mevius D, Bruschke C, Wagenaar J. 2015. Reduction of veterinary antimicrobial use in the Netherlands. The Dutch success model. Zoonoses Public Health 62:79–87.

Springbett A, MacKenzie K, Woolliams J, Bishop S. 2003. The contribution of genetic diversity to the spread of infectious diseases in livestock populations. Genetics 165:1465–1474.

Thompson-Crispi K, Sewalem A, Miglior F, Mallard B. 2012. Genetic parameters of adaptive immune response traits in Canadian Holsteins. J Dairy Sci. 95:401–409.

Tiezzi F, Parker-Gaddis KL, Cole JB, Clay JS, Maltecca C. 2015. A genome-wide association study for clinical mastitis in first parity US Holstein cows using single-step approach and genomic matrix re-weighting procedure. PLoS ONE 10:e0114919.

Tsairidou S, Anacleto O, Woolliams J, Doeschl-Wilson A. 2019. Enhancing genetic disease control by selecting for lower host infectivity and susceptibility. Heredity 122:742–758.

Vleeshouwers VG, Raffaele S, Vossen JH, Champouret N, Oliva R, *et al.* 2011. Understanding and exploiting late blight resistance in the age of effectors. Annu Rev Phytopathol. 49:507–531.

Appendix

I. Derivation of link variance for the complementary log-log link function

The link variance for the liability model as presented in Equation (11) in the main text, can be derived by calculation of the variance belonging to the pdf of the inverse of the clog-log link function. This pdf is obtained by differentiation of the inverse of the complementary log-log function.

The formula for the complementary log-log link function is:

$$x(p) = \log(-\log(1-p))$$

The inverse of this function is:

$$p(x) = 1 - e^{-e^x}$$

Differentiation of the inverse results in the formula for the pdf:

$$f(x) = e^{x - e^x}$$

The variance of this probability density function can be obtained by using the general formula for the variance of a function:

$$VAR(X) = E(X^2) - E(X)^2$$

in which E denotes the expected value. The expected values are calculated by integration of $f(x)$:

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x) x = \gamma^2 + \frac{\pi^2}{6}$$

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx = -\gamma$$

in which γ denotes the Euler-Mascheroni constant. $VAR(X)$ becomes:

$$VAR(X) = \gamma^2 + \frac{\pi^2}{6} - \left(-\gamma\right)^2 = \frac{\pi^2}{6}$$

In fact, the inverse of the clog-log and its corresponding pdf are (mirrored) standard Gumbel distributions, the variance of the standard Gumbel distribution is indeed equal to $\frac{\pi^2}{6}$.

II. Gillespie-algorithm

In the simulation of the infectious disease, the time to the next event is sampled from an exponential distribution, which requires calculation of the total rate parameter r, defined as the sum of all individual transmission and recovery rates:

$$r = \sum_i \beta_i \frac{I}{N} + \alpha I$$

After sampling a time to the next event based on r, the type of event and the individual involved in the event is sampled based on the contribution of each individual to r, such that the probability of infection for a certain susceptible individual i is $\frac{\beta_i \frac{I}{N}}{r}$ and the probability of recovery for a certain infected individual j is $\frac{\alpha}{r}$ (this α is the same for each individual in our simulations). The effect of individual susceptibility on the probability of infection is thus multiplicative: An individual with a susceptibility of two has a two times higher β_i than an average individual with a susceptibility of one. Consequently, given same exposure to infectious individuals, its share in r, and thus its probability of infection, is two times higher as well.

Communicating editor: M. Sillanpää