Estimating host genetic effects on susceptibility and infectivity to infectious diseases and their contribution to response to selection

Mahlet Teka Anche

Thesis committee

Promotor
Prof. Dr M. C. M. De Jong
Professor of Quantitative Veterinary Epidemiology
Wageningen University

Co- Promotor
Dr P. Bijma
Assistant Professor, Animal Breeding and Genomics Centre
Wageningen University

Other members
Prof. Dr B. J. Zwaan - Wageningen University, Wageningen, The Netherlands
Prof. Dr J. A. Woolliams - University of Edinburgh, Edinburgh, Scotland
Prof. Dr G. van Schaik – Monitoring and surveillance of farm animal health
Gezondheidsdienst voor dieren, Utrecht, The Netherlands
Dr E. P. C. Koenen - CRV, Arnhem, The Netherlands

This research was conducted under the auspices of the Graduate School of Wageningen Institute of Animal Sciences (WIAS).

Estimating host genetic effects on susceptibility and infectivity to infectious diseases and their contribution to response to selection

Mahlet Teka Anche

Thesis

submitted in fulfillment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr A. P. J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Friday June 15 2016
at 1.30 p.m. in the Aula.

Mahlet Teka Anche
Estimating host genetic effects on susceptibility and infectivity to infectious diseases and their contribution to response to selection

PhD thesis, Wageningen University, Wageningen, NL (2016)
With references, with summary in English

ISBN 978-94-6257-744-2

Abstract

Mahlet Teka Anche. (2016). Estimating host genetic effects on susceptibility and infectivity to infectious diseases and their contribution to response to selection. PhD thesis, Wageningen University, the Netherlands

Genetic approaches aiming to reduce the prevalence of an infection in a population usually focus on improving host susceptibility to an infection. The prevalence of an infection, however, is also affected by the infectivity of individuals. Studies reported that there exists among host (genetic/phenotypic) variation in susceptibility and infectivity to infectious diseases. The effect of host genetic variation in susceptibility and infectivity on the prevalence and risk of an infection is usually measured by the value of the basic reproduction ratio, R_0. R_0 is an important epidemiological parameter that determines the risk and prevalence of an infection. It has a threshold value of 1, where major disease outbreak can occur when $R_0 > 1$ and the disease will die out when $R_0 < 1$. Due to this threshold property, genetic improvements aiming to reduce the prevalence of an infection should focus on reducing R_0 to a value below 1. The overall aim of this thesis was to develop methodologies that allow us to investigate the genetic effects of host susceptibility and infectivity on the prevalence of an infection, which is measured by the value of R_0. Moreover, we also aim to investigating the effect of relatedness among groupmates on the utilization of among host genetic variation in susceptibility and infectivity so as to reduce the prevalence of infectious diseases. The theory of direct-indirect genetic effects and epidemiological concepts were combined to develop methodologies. In addition, a simulation study was performed to validate the methodologies developed and examine the effect of relatedness on the utilization of genetic variation in susceptibility and infectivity. It was shown that an individual's genetic effect on its susceptibility and infectivity affect the prevalence of an infection and that an individual's breeding value for R_0 can be defined as a function of its own allele frequencies for susceptibility and infectivity and of population average susceptibility and infectivity. Moreover, simulation results show that, not only an individual's infectivity but also an individual's susceptibility represents an indirect genetic effect on the disease status of individuals and on the prevalence of an infection in a population. It was shown that having related groupmates allows breeders to utilize the genetic variation in susceptibility and infectivity, so as to reduce the prevalence of an infection.

By the strength of The One

Contents

1

General introduction

1.1 Introduction

Infectious diseases impose a worldwide concern to the sustainability of livestock production, particularly due to their impact on the welfare and productivity of livestock. In addition to this, the fact that infectious diseases impose a threat to human health due to their zoonotic effect has raised the need to reduce the threat imposed by infectious diseases. In the past few decades, the existence of heritable variation among individuals in their response to different infectious diseases has been reported by studies on quantitative genetics of livestock diseases (Nicholas, 2005). These findings have, therefore opened the door for animal breeders to use selective breeding for livestock with an improved response to infectious diseases as a complementary method to existing disease control strategies in order to reduce the impact of infectious diseases.

Among others, individual susceptibility and infectivity are important disease-related traits that influence the transmission of an infection in a population. Individual susceptibility is the probability of an individual to become infected given it is exposed to a typical (average) infectious individual, whereas individual infectivity is the rate at which an individual transmits the infection to a typical susceptible individual. It is clear that there is phenotypic variation among individuals for these disease-related traits, which will impact the transmission and prevalence of an infection in the population. These traits might have genetic basis and it is therefore likely, that there exists among-individual genetic variation. Understanding the impact of genetic variation in these disease-related traits on the transmission of an infection, however, requires modelling of the disease dynamics in such a heterogeneous population.

1.2 Epidemiology of infectious diseases

Epidemiological modelling of disease dynamics involves the study of the mathematics underlying the change in number of infected individuals over time. A classical model used in these studies is the SIR model, where S stands for **S**usceptible, I for **I**nfected, and R for **R**ecovered. The SIR model is one of the variants of compartmental models that can be used to model disease dynamics in a population. The models can be implemented either deterministically or stochastically (Addy et al., 1991; Kermack and McKendrick, 1991a, b, c; Velthuis et al., 2007). In the classical SIR model, individuals move through the states in the order S → I → R. With stochasticity, these transmission events, i.e. S → I and I → R, occur with a certain rate (probability per unit of time) that is specified by the model parameters. These rates are the transmission rate $\beta SI/N$ for S → I with a

transmission rate parameter β, and the recovery rate αI for I → R with a recovery rate parameter α. Note that the symbols S, I and R denote both the disease status and the number of individuals with that disease status. The transmission rate parameter β is the probability per unit of time that a typical infectious individual infects another individual in a totally susceptible population (Diekmann et al., 1990; Anderson et al., 1992). The recovery rate parameter α is the probability per unit of time for an infective individual to recover from an infection. In other words, for constant α, the infectious period is exponentially distributed with a mean duration of α^{-1} time units.

To facilitate the understanding of the basic SIR model, we use the deterministic equivalent of the stochastic SIR model that can be formulated in terms of ordinary differential equation as follows:

$$\frac{dS}{dt} = -\beta SI/N$$
$$\frac{dI}{dt} = \beta SI/N - I\alpha$$
$$\frac{dR}{dt} = \alpha I$$

where N is the total population size and, N = S + I + R.

In the basic SIR-model, an individual begins the transmission process as a susceptible individual, which can become infected by another individual that has been infected some time ago. The first equation describes the change in the number of susceptible individuals (S) through time. The probability that a random contact of a susceptible is with an infectious individual is I/N. Therefore, the rate at which infections occur is the product of the number of susceptible individuals (S), the probability that a random contact of a susceptible individual is with an infectious individual I/N, and the transmission rate parameter β. Thus, the rate of change in the number of susceptibles is given by $-\beta SI/N$. The second equation describes the change in the number of infected individuals (I) through time. This number increases due to susceptibles becoming infected, at a rate $\beta SI/N$, and decreases either by complete recovery or death with a rate of recovery αI. The last equation describes the change in the number of recovered. Figure 1.1 shows the change in the number of susceptible and infected individuals through time.

In epidemiology, an important population parameter that determines the risk and severity of an infection in the population is the basic reproduction ratio, R_0. R_0 is the average number of new infected individuals (cases) produced by a typical infectious individual during its entire infectious lifetime in an otherwise naïve population. R_0 has a threshold value of 1, which determines whether a major

disease outbreak can occur or whether the endemic equilibrium can exist. When R_0 < 1, only minor outbreaks can occur and the disease will die out. On the other hand, when R_0 > 1, major outbreaks can occur and affect a larger fraction of the population, or an endemic equilibrium can exist.

In epidemiology, an infection is said to be endemic when the infection persists in the population with a certain fraction of individuals being infected all the time. Hence, all the time new infections will occur. These new infections could be due to the loss (lack) of immunity of the recovered individuals or due to the introduction of new susceptible individuals to the population. A steady state or an equilibrium exists when every infected individual passes the infection on to a single other individual on average. Thus, the average number of cases that an infectious individual produces, which is the effective reproduction ratio R_E must be 1. In this case, the disease will neither die out nor increase exponentially. For endemic diseases, we thus have $R_E = \frac{S}{N}R_0 = 1$, where the fraction of individuals that is infected is given by $1 - \frac{S}{N} = 1 - 1/R_0$ (where N is the total population size). Figure 1.2 shows the relationship between the fraction infected and the basic reproduction ratio, which shows that the fraction that gets infected $1 - \frac{S}{N}$ increases with increasing R_0.

For epidemic diseases, the number of individuals that gets infected increases initially exponentially, but eventually only a fraction of individuals gets infected. This fraction is known as the final size, $1 - s_\infty$. The final size is also a function of R_0, and is given as the solution of the final size equation, $\ln s_\infty = R_0(s_\infty - 1)$ (Kermack and McKendrick, 1991a). Figure 1.2 shows the relationship between the value of R_0 and the fraction of individuals that gets infected by the end of an epidemic, $1 - s_\infty$. For different values of R_0, the final size of the epidemic varies, increasing with increasing the value of the R_0. Note that the change in outcome (Figure 1.2) is the steepest near $R_0 = 1$.

Thus, for both endemic and epidemic diseases, a breeding strategy to reduce the prevalence of an infection should reduce the value of the reproduction ratio R_0, preferably to a value below 1.

Breeding for reduced R_0, however, will involve a conceptual difference between quantitative genetics and epidemiology. In epidemiology on the one hand, R_0 is a parameter referring to the whole population. In quantitative genetics on the other hand, breeding values are used which are properties of single individuals. Thus, breeding for reduced R_0 requires defining the breeding value of all individuals for R_0 and from that the heritable variation for R_0.

Moreover, even though breeding for lower R_0 would be an obvious goal for an epidemiologist who aims to reduce the prevalence of an infection, it might not be obvious for animal breeders. For animal breeders, using individual disease status (0/1) as a selection criterion would be more common. As mentioned above, however, the fraction of infected individuals, for both endemic and epidemic diseases, is coupled with the value of R_0. Thus, breeding for reduced R_0 will reduce the fraction of individuals that gets infected, which in turn reduces the disease incidence and prevalence in the population.

R_0 is an emergent trait that arises when different individuals (susceptible and infectious) interact. As mentioned above, however, breeding for reduced R_0 requires defining individual breeding values for R_0. Bijma (2011) has shown that results from the field of indirect genetic effects (IGEs) can be used to define individual breeding values for traits that are a property of the population, such as R_0 (which is discussed in the 2nd chapter of this thesis). In the next section of this chapter, I will, therefore, briefly discuss what IGEs are and their role in the transmission of an infection in a population.

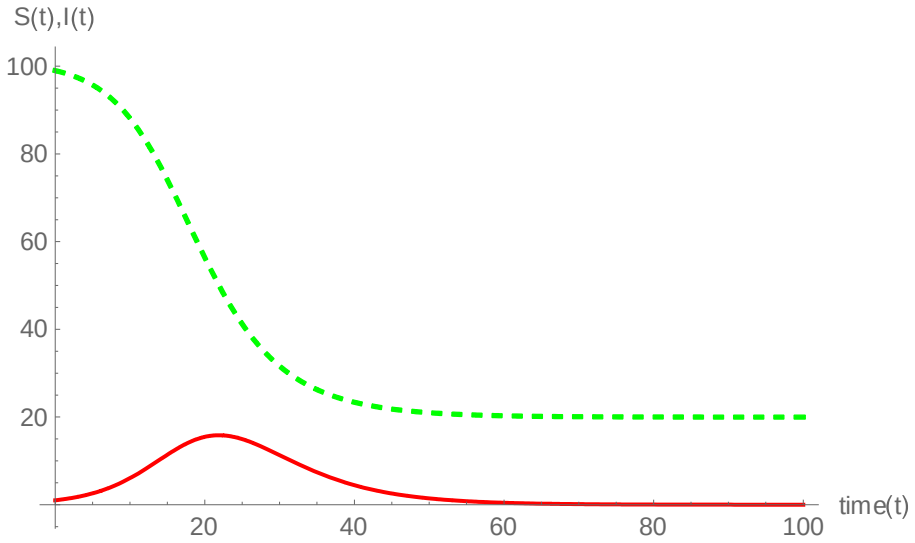

Figure 1.1. Change in the number of susceptible S(t) (green dotted line) and infected individuals I(t) (red continuous line) through time, t.

1.3 Indirect genetic effects (IGEs)

In classical quantitative genetics, the phenotypic value of an individual is decomposed into a heritable component A_i, known as the breeding value and non-heritable environmental component E_i (Lynch and Walsh, 1998). Thus, the phenotypic value can be written as:

$$P_i = A_i + E_i \quad [1]$$

In this equation, an individual's breeding value A_i is the sum of the average effect of genes carried by the individual on its own trait value. Because the breeding value affects the trait value of the individual itself, it is known as a direct genetic effect (DGEs).

In the presence of (social) interactions, however, an individual's phenotypic value is also affected by the genes of its $n-1$ (where n denotes group size) groupmates. These effects on the phenotype are known as indirect genetic effects (IGEs) (Griffing, 1967). IGEs, which are also known as social or associative effects, are heritable effects of an individual on the phenotypic value of other individuals (Griffing, 1967, 1976, 1981; Moore et al., 1997; Wolf et al., 1998).

Thus, in the presence of (social) interaction, the phenotypic value of an individual is modelled as:

$$P_i = A_{D,i} + E_{D,i} + \sum_{j=1}^{n-1} A_{I,j} + \sum_{j=1}^{n-1} E_{I,j} \quad [2]$$

where P_i is phenotypic value of individual *i*, $A_{D,i}$ is the direct genetic effect of an individual's genes on its own trait value, $E_{D,i}$ is the non-heritable direct effect, $A_{I,j}$ is indirect genetic effect of all genes arising from individual j (j≠i) which is one of the $(n-1)$ groupmates of individual *i*, and $E_{I,j}$ is the non-heritable indirect effect. When groupmates are unrelated, phenotypic variance is given by,

$$\sigma_P^2 = \sigma_{A_D}^2 + (n-1)\sigma_{A_I}^2 + \sigma_e^2 \quad [3]$$

In the presence of (social) interaction, we can define the total breeding value of an individual $A_{T,i}$, which is the heritable effect of an individual on the population mean. It combines the individual's direct and indirect genetic effect as follows:

$$A_{T,i} = A_{D,i} + (n-1)A_{I,i} \quad [5]$$

where $A_{I,i}$ is indirect genetic effect of individual *i* on the trait values of its $(n-1)$ groupmates. Note that, in contrast to the phenotypic value (Equation 2), the total breeding value of an individual originates entirely from the focal individual i. As a result, variance in total breeding value which is the heritable variation that is available for response to selection will be:

$$\sigma_{A_T}^2 = \sigma_{A_D}^2 + 2(n-1)\sigma_{A_{D,I}} + (n-1)^2\sigma_{A_I}^2 \quad [6]$$

where $\sigma_{A_T}^2$ is variance in total breeding value, $\sigma_{A_D}^2$ is variance in DGE, $\sigma_{A_{D,I}}$ is covariance between DGE and IGE and $\sigma_{A_I}^2$ is heritable variance in IGEs. Thus, in the

presence of (social) interaction, IGEs may increase the total heritable variance ($\sigma^2_{A_T}$ > σ^2_P) (Griffing, 1967; Bijma et al., 2007).

IGEs are a common phenomenon in both plants and animals (Frank, 2007). Even though IGEs are often considered to be associated with behaviour traits (Muir and Craig, 1998; Muir, 2005), they may also work in other ways, for example through the exposure to infections. An individual's infectivity is the propensity of an infected individual to infect other susceptible individuals in its proximity. Hence, in the context of exposure to infections, individual infectivity can be regarded as an IGE of the individual.

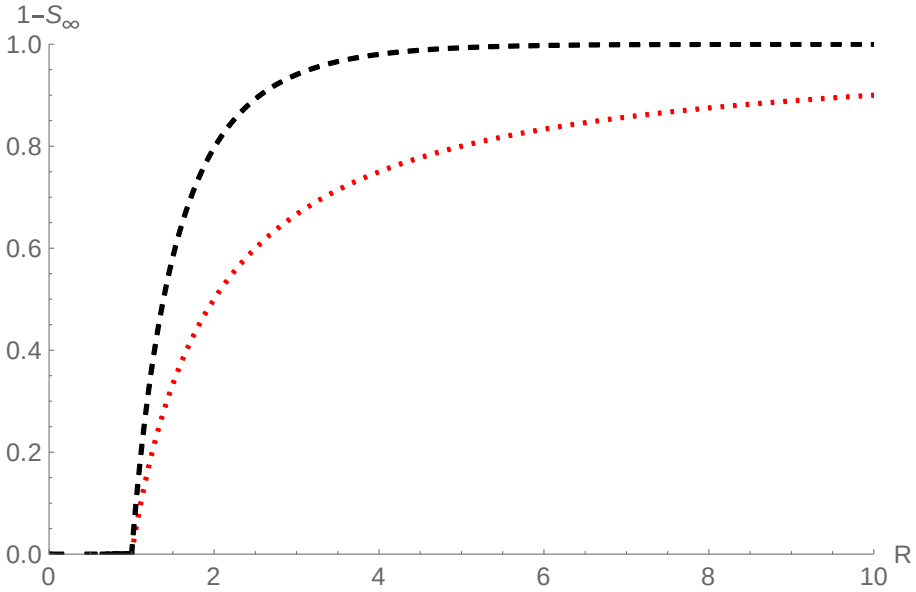

Figure 1.2. Relationship between the basic reproduction ratio R_0 and the fraction of individual that gets infected ($\mathbf{1} - \boldsymbol{s}_\infty$), for both endemic (red dotted line) and epidemic diseases (black dotted line).

Both natural and artificial selection will work to exhaust the existing heritable variation in traits that are part of individual fitness, such as individual's susceptibility. An individual's infectivity, on the other hand, is not part of individual fitness. This would prevent natural selection to exhaust heritable variation that may be present in infectivity. As a result, evolutionary theory predicts that a relatively larger heritable variation may be present in infectivity than in susceptibility. This indicates that there may accumulate a significant amount of heritable variation in infectivity, which can contribute to the total heritable

variation that reflects the potential of a population to respond to selection. In the next section of this chapter, I will discuss what appears to be lacking in the classical quantitative genetic analysis of infectious diseases and the aims of this thesis in an attempt to fill this gap.

1.4 The gap

Classical quantitative genetics analysis of disease-related traits is usually based on binary data, that is, data which solely indicate whether an individual became infected or not. In such analysis, only a direct genetic effect of the individual itself is usually fitted. Thus, it is implicitly assumed that an individual's disease status is a function only of its own genes, which can be considered as an individual's direct genetic effect (DGE) for susceptibility. Because individuals may infect each other, however, the prevalence and dynamics of an infection also depend on indirect genetic effects. Accumulating evidence on the existence of "superspreaders" in the transmission of an infection, especially in transmission of bacterial infections, suggests that there exists among host (phenotypic) variation in infectivity, which might have a genetic basis and affect the dynamics and prevalence of an infection (Diekmann and Heesterbeek, 2000; Lloyd-Smith et al., 2005).

As mentioned above, the existence of variation among individuals for different disease-related traits can be seen as an opportunity for animal breeders to use selective breeding for improved response to infectious diseases, as a complementary method to the existing disease control strategies. Selective breeding for reduced impact of infectious diseases, however, has proven difficult due to lower heritability estimates reported for the disease-related traits under selection (Bishop and Woolliams, 2010). One of the reasons for such low heritability estimates could be the failure of the conventional statistical methods used in parameter estimation to reflect the true genetic variance present in disease-related traits (Lipschutz-Powell et al., 2012a).

The standard linear mixed models used in quantitative genetic analysis do not capture genetic variation present in IGEs, such as in infectivity. This is because they connect the disease status of an individual to its own pedigree. Individual infectivity, on the other hand, is observed in the disease status of other individuals than the one carrying the gene. Thus standard analysis will overlook the heritable variation in infectivity that, when present, may contribute to the total heritable variation in the population (Lipschutz-Powell et al., 2012b).

Moreover, estimating breeding values and genetic variation in individual susceptibility and infectivity from data on individual infection status is

methodologically challenging. This is because the linear mixed models that are used in classical quantitative genetic analysis of infectious diseases do not take the non-linear stochastic nature of infection dynamics into account. Thus, the fact that classical quantitative genetic analysis of infectious diseases fails to take the IGEs of infectivity and the stochastic nature of infection dynamics into account may have caused seemingly low heritability estimates for disease traits (Bishop and Woolliams, 2010).

In this thesis, we aim to fill this gap by developing methodology that takes the stochastic nature of an infection and the IGEs of infectivity into account in order to achieve the following goals. First, we aim to define breeding values and heritable variation for the basic reproduction ratio, R_0 (chapter 2). Moreover, studies have shown that for traits affected by IGEs, such as individual disease status, group selection and relatedness among interacting individuals, increase response to selection (Griffing, 1967, 1976, 1981; Bijma and Wade, 2008). In the second chapter, we will also investigate selection mechanisms that affect utilization of heritable variation in R_0. In the 3rd chapter, we will develop a statistical model that allows us to estimate gene effects for loci affecting susceptibility and infectivity of an individual. In this chapter, we will also investigate factors that affect the quality of estimates for the gene effects. In chapter 4, we will estimate the effect of major histocompatibility complex (MHC) polymorphisms on individual susceptibility and infectivity to nematode infection in a population of Scottish Blackface sheep. In chapter 5, we will develop a methodology to estimate breeding values and variance components for susceptibility and infectivity, and also investigate the effect of relatedness on the quality of the estimates. In chapter 6, the general discussion, I will discuss three main points in a broader perspective. First, I will discuss the breeding value for the basic reproduction ratio R_0, and its relation to susceptibility and infectivity of an individual. Second, I will discuss selection strategies that can be used for reducing R_0. Finally, I will discuss the practical implications of the findings of this thesis.

1.5 References

Addy, C. L., I. M. Longini Jr, and M. Haber. 1991. A generalized stochastic model for the analysis of infectious disease final size data. Biometrics: 961-974.

Anderson, R. M., R. M. May, and B. Anderson. 1992. Infectious diseases of humans: dynamics and control. Wiley Online Library.

Bijma, P. 2011. A general definition of the heritable variation that determines the potential of a population to respond to selection. Genetics: genetics. 111.130617.

Bijma, P., W. M. Muir, and J. A. Van Arendonk. 2007. Multilevel selection 1: quantitative genetics of inheritance and response to selection. Genetics 175: 277-288.

Bijma, P., and M. Wade. 2008. The joint effects of kin, multilevel selection and indirect genetic effects on response to genetic selection. Journal of evolutionary biology 21: 1175-1188.

Bishop, S. C., and J. A. Woolliams. 2010. On the Genetic Interpretation of Disease Data. Plos One 5.

Diekmann, O., and J. Heesterbeek. 2000. Mathematical epidemiology of infectious diseases. Wiley, New York.

Diekmann, O., J. Heesterbeek, and J. A. Metz. 1990. On the definition and the computation of the basic reproduction ratio R 0 in models for infectious diseases in heterogeneous populations. Journal of mathematical biology 28: 365-382.

Frank, S. A. 2007. All of life is social. Curr Biol 17: R648-R650.

Griffing, B. 1967. Selection in Reference to Biological Groups .I. Individual and Group Selection Applied to Populations of Unordered Groups. Australian Journal of Biological Sciences 20: 127-&.

Griffing, B. 1976. Selection in Reference to Biological Groups .5. Analysis of Full-Sib Groups. Genetics 82: 703-722.

Griffing, B. 1981. A Theory of Natural-Selection Incorporating Interaction among Individuals .2. Use of Related Groups. J Theor Biol 89: 659-677.

Kermack, W., and A. McKendrick. 1991a. Contributions to the mathematical theory of epidemics—I. Bulletin of mathematical biology 53: 33-55.

Kermack, W., and A. McKendrick. 1991b. Contributions to the mathematical theory of epidemics—II. The problem of endemicity. Bulletin of mathematical biology 53: 57-87.

Kermack, W., and A. McKendrick. 1991c. Contributions to the mathematical theory of epidemics—III. Further studies of the problem of endemicity. Bulletin of mathematical biology 53: 89-118.

Lipschutz-Powell, D., J. A. Woolliams, P. Bijma, and A. B. Doeschl-Wilson. 2012a. Indirect genetic effects and the spread of infectious disease: are we capturing the full heritable variation underlying disease prevalence?

Lipschutz-Powell, D., J. A. Woolliams, P. Bijma, and A. B. Doeschl-Wilson. 2012b. 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, J. O., S. J. Schreiber, P. E. Kopp, and W. M. Getz. 2005. Superspreading and the effect of individual variation on disease emergence. Nature 438: 355-359.

Lynch, M., and B. Walsh. 1998. Genetics and analysis of quantitative traits. Sinauer Sunderland, MA.

Moore, A. J., E. D. Brodie, and J. B. Wolf. 1997. Interacting phenotypes and the evolutionary process .1. Direct and indirect genetic effects of social interactions. Evolution 51: 1352-1362.

Muir, W. M. 2005. Incorporation of competitive effects in forest tree or animal breeding programs. Genetics 170: 1247-1259.

Muir, W. M., and J. V. Craig. 1998. Improving animal well-being through genetic selection. Poultry Sci 77: 1781-1788.

Nicholas, F. W. 2005. Animal breeding and disease. Philosophical Transactions of the Royal Society B: Biological Sciences 360: 1529-1536.

Velthuis, A., A. Bouma, W. Katsma, G. Nodelijk, and M. De Jong. 2007. Design and analysis of small-scale transmission experiments with animals. Epidemiology and Infection 135: 202-217.

Wolf, J. B., E. D. Brodie III, J. M. Cheverud, A. J. Moore, and M. J. Wade. 1998. Evolutionary consequences of indirect genetic effects. Trends in Ecology & Evolution 13: 64-69.

2

On the definition and utilization of heritable variation among hosts in basic reproduction ratio R_0 for infectious diseases

Mahlet T. Anche[1,2], Mart C. M. de Jong[2], P. Bijma[1]

[1] Animal breding and Genomics Centre, Wageningen Institute of Animal Sciences (WIAS), Wageningen University, The Netherlands; [2] Quantitative Veterinary Epidemiology Group, Wageningen Institute of Animal Sciences (WIAS), Wageningen University, The Netherlands

Heredity (2014):1-11

Abstract

Infectious diseases have a major role in evolution by natural selection and pose a worldwide concern in livestock. Understanding quantitative genetics of infectious diseases, therefore, is essential both for understanding the consequences of natural selection and for designing artificial selection schemes in agriculture. The basic reproduction ratio, R_0, is the key parameter determining risk and severity of infectious diseases. Genetic improvement for control of infectious diseases in host populations should therefore aim at reducing R_0. This requires definitions of breeding value and heritable variation for R_0, and understanding of mechanisms determining response to selection. This is challenging, as R_0 is an emergent trait arising from interactions among individuals in the population. Here we show how to define breeding value and heritable variation for R_0 for genetically heterogeneous host populations. Furthermore, we identify mechanisms determining utilization of heritable variation for R_0. Using indirect genetic effects, next-generation matrices and a SIR (Susceptible, Infected and Recovered) model, we show that an individual's breeding value for R_0 is a function of its own allele frequencies for susceptibility and infectivity and of population average susceptibility and infectivity. When interacting individuals are unrelated, selection for individual disease status captures heritable variation in susceptibility only, yielding limited response in R_0. With related individuals, however, there is a secondary selection process, which also captures heritable variation in infectivity and additional variation in susceptibility, yielding substantially greater response. This shows that genetic variation in susceptibility represents an indirect genetic effect. As a consequence, response in R_0 increased substantially when interacting individuals were genetically related.

Key words: Reproduction ratio R_0, indirect genetic effect, emergent trait, breeding values, heritable variation, kin selection

2.1 Introduction

Infectious diseases are widespread in humans, animals and plants. In natural populations, infectious diseases have a major role in the process of evolution by natural selection (Haldane, 1949; O'Brien and Evermann, 1988). In domestic populations, particularly in livestock, infectious diseases are imposing a worldwide concern owing to their impact on the welfare and productivity of livestock, and in the case of zoonosis, also because of the threat for human health. To contain the threat imposed by infectious diseases, different control strategies such as vaccination, antibiotic treatments and management practices have been implemented widely. However, the evolution of resistance to antibiotics by bacteria, evolution of resistance to vaccines by viruses and undesirable environmental impacts of antibiotic treatment put these strategies under question (Gibson and Bishop, 2005). Thus, there is a need to investigate additional control strategies, so as to extend the repertoire of possible interventions. A greater repertoire is favourable (1) because it allows for a change in approach when certain control measures fail and (2) because the use of combinations of control measures make emergence of resistance against control more difficult.

Several studies have demonstrated the existence of genetic variation for different disease traits for a wide variety of infectious diseases. Examples are clinical mastitis and *Mycobactrium bovis* infections in dairy cattle (Heringstad et al., 2005). Such studies usually focus on estimating the genetic variance in individual disease status. As this approach connects an individual's own disease status to its own pedigree, it only captures heritable variation in susceptibility (or resistance) to disease (Lipschutz-Powell et al., 2012). However, host genetic variation may be present also in other traits that affect the dynamics of infectious diseases in populations. Thus, to use a general term for such other traits, infectivity will also have an impact on the transmission of infectious diseases. There clearly exists (phenotypic) variation in infectivity as it can be seen from the occurrence of superspreaders (Lloyd-Smith et al., 2005). Thus, it is most likely that the classical quantitative genetic analysis based on individual disease status captures only part of the possible heritable variation in the host underlying infectious disease dynamics (Lipschutz-Powell et al., 2012).

The ultimate goal of selective breeding for disease traits is to reduce the risk of an epidemic and/or to reduce the level of the endemic equilibrium. In epidemiology, the key parameter determining the risk and size of an epidemic and/or the level of the endemic equilibrium is the basic reproduction ratio, R_0. R_0 is the average

number of secondary cases produced by a typical infectious individual during its entire infectious life time, in an otherwise naïve population (Diekmann et al., 1990). R_0 has a threshold value of 1, which determines whether a major disease outbreak can occur or whether the endemic equilibrium exists. When $R_0 < 1$, the epidemic will die out. On the other hand, when $R_0 > 1$ major outbreaks or an endemic equilibrium (persistence) can occur. Hence, breeding strategies to reduce the risk and prevalence of an infectious disease should aim at reducing R_0, preferably to below a value of 1.

Breeding to reduce R_0 raises a conceptual difference between quantitative genetics and epidemiology: R_0 is an epidemiological parameter referring to an entire population, whereas quantitative genetics rests on the concept of breeding value, which refers to a single individual. It is clear that in a genetically heterogeneous population, R_0 is a function of individual genotypes in the population, which in turn are a function of allele frequencies. Moreover, a change in allele frequencies will change R_0, indicating R_0 can respond to selection. Genetic improvement aiming to reduce R_0 should ideally be based on the effects of an individual's genes on R_0, which would require defining individual breeding values for R_0. Moreover, defining a breeding value for R_0 would also allow defining heritable variation in R_0, that is, the variation in individual breeding values for R_0, which would give an indication of the prospects for genetic improvement with respect to R_0.

For domestic populations, the subsequent question would be how to design breeding programs, so as to utilize optimally heritable variation in R_0 and achieve the greatest possible rate of reduction in R_0. The equivalent issue for natural populations would be what ecological conditions are favourable for efficient reduction of R_0 by natural selection. For emergent traits that depend on multiple individuals, research in the field of indirect genetic effects (IGEs) suggests that group selection and relatedness among interacting individuals ('kin selection') can be used to increase response to selection (Griffing, 1967; Anderson and May, 1992; Andreasen, 2011; Bijma, 2011). This suggests that relatedness and group selection may be important mechanisms affecting the utilisation of heritable variation in R_0, either by natural or artificial selection.

Here we show how to define breeding value and heritable variation for R_0 for a genetically heterogeneous host population, where individuals differ for susceptibility and infectivity. For that purpose, we have adapted the theory of IGEs commonly applied to socially affected traits, using the epidemiological concept of next-generation matrices (NGMs) (Diekmann et al., 1990; Diekmann et al., 2010). Furthermore, we examine the mechanisms determining the utilization of heritable

variation in R_0, focussing on the effects of kin selection on response in R_0, and in susceptibility and infectivity.

2.2 Method

2.2.1 Dynamic model of infection

In a completely naïve population where a microparasitic infection is introduced, the disease dynamics can be modelled with a basic compartmental stochastic SIR (Susceptible, Infected and Recovered) model. In this model, individuals move through the states in the order $\mathrm{S} \rightarrow \mathrm{I} \rightarrow \mathrm{R}$ (Anderson et al., 1992). Therefore, the possible events that an individual may encounter are infection and recovery. With stochasticity, these events occur randomly at a certain rate (probability per unit of time) specified by the model parameters. In the SIR-model, these parameters are the transmission rate parameter (β) for $\mathrm{S} \rightarrow \mathrm{I}$, and the recovery rate parameter (α) for $\mathrm{I} \rightarrow \mathrm{R}$. The transmission rate parameter β is the probability per unit of time that a typical infected individual infects another individual in a totally susceptible population (Diekmann et al., 1990; Anderson et al., 1992). When constant population density is assumed, the rate at which the susceptible population becomes infected is $\beta SI/N$, where S denotes the number of susceptible individuals, I the number of infectious individuals, and N the total number of individuals in the population (Kermack and McKendrick, 1991). The recovery rate parameter α is the probability per unit of time for an infective to recover from an infection. In other words, for constant α, the infectious period is exponentially distributed with a mean duration of α^{-1} time units.

The transmission rate parameter, β, depends on the infectivity of infectious individuals and on the susceptibility of uninfected recipient individuals. Thus, in a homogeneous population where all individuals have the same level of infectivity and susceptibility, there is a single β that applies to the whole population, which can be defined as a function of these parameters,

$$\beta = \gamma\varphi c, \qquad (1)$$

where γ is susceptibility, φ is infectivity and c is average number of contacts an infectious individual makes per unit of time (See Table 2.1 for a notation key).

Table 2.1. Notation key

Symbol	Meaning
γ_G	Effect of G allele at susceptibility locus
γ_g	Effect of g allele at susceptibility locus
φ_F	Effect of F allele at infectivity locus
φ_f	Effect of f allele at infectivity locus
p_g	Frequency of the g allele for susceptibility
p_f	Frequency of the f allele for infectivity
$\bar{\gamma}$	Average individual susceptibility
$\bar{\varphi}$	Average individual infectivity
r_γ	Relatedness at susceptibility locus
r_φ	Relatedness at infectivity locus
β_{ij}	Pairwise transmission rate parameter between susceptible individual i and infective individual j
α	Rate of recovery parameter
C	Contact rate
R_0	Basic reproduction ratio
$A_{R_0,i}$	Breeding value for R_0 of individual i
σ_{A_T}	Additive standard deviation in total breeding value
D	Measure of linkage disequilibrium
F_{IS}	Measure of deviation from Hardy Weinberg Equilibrium

2.2.2 Dynamic model of infection with genetic heterogeneity

In a genetically heterogeneous population, however, the transmission rate parameter β may vary among pairs of individuals. This pairwise transmission rate will depend on the infectivity genotype of the infectious individual, and on the susceptibility genotype of the recipient susceptible individual. The assumption that transmission depends on the infectivity of only the infectious individual and on the susceptibility of only the recipient individual is known as separable mixing (Diekmann et al., 1990). Thus, we may define the pairwise transmission rate parameter β_{ij} from an infectious individual *j* to a susceptible individual *i* as

$$\beta_{ij} = \gamma_i \varphi_j c, \qquad (2)$$

where γ_i denotes susceptibility of susceptible individual *i*, and φ_j denotes infectivity of infectious individual *j*. In Equation (2), *c* represents the average contact rate; any variation in contact rate among susceptible and infectious individuals is included in γ_i and φ_i because of the assumption of separable mixing.

In the following, we model genetic heterogeneity in a diploid population using two bi-allelic loci, one locus for susceptibility effect (γ), and the other locus for infectivity effect (φ). The susceptibility locus has alleles G and g, with susceptibility values γ_G and γ_g respectively, and the infectivity locus has alleles F and f, with infectivity values φ_F and φ_f, respectively. Furthermore, both loci are assumed to have additive allelic effects without dominance. Thus, genotypic values are given by $\gamma_{GG} = \gamma_G + \gamma_G = 2\gamma_G$, $\gamma_{gg} = \gamma_g + \gamma_g = 2\gamma_g$ and $\gamma_{Gg} = \gamma_{gG} = \gamma_G + \gamma_g$, for susceptibility, and $\varphi_{FF} = \varphi_F + \varphi_F = 2\varphi_F; \varphi_{ff} = \varphi_f + \varphi_f = 2\varphi_f$ and $\varphi_{Ff} = \varphi_{fF} = \varphi_F + \varphi_f = 2\varphi_F$ for infectivity. As we assumed additive gene action, average susceptibility in the population is given by

$$\bar{\gamma} = 2p_g\gamma_g + 2(1 - p_g)\gamma_G, \qquad (3)$$

and average infectivity is given by

$$\bar{\varphi} = 2p_f\varphi_f + 2(1 - p_f)\varphi_F, \qquad (4)$$

where p_f is the frequency of the *f* allele, and p_g the frequency of the *g* allele, and the "2" arises because each individual carries two alleles. Note that $\bar{\gamma}$ and $\bar{\varphi}$ are average susceptibility and average infectivity over individuals, not average of allele effects. In a population as define here, there are nine genotypes of individuals because of the combinations of their genotype for susceptibility and infectivity.

For this heterogeneous population, we can now construct the NGM. The NGM describes the number of infectious individual of each type in the next generation of the epidemic, produced by infectious individuals of each type in the current generation. Then, we can calculate R_0 as the dominant eigenvalue of the NGM. Under the assumption of separable mixing, the dominant eigenvalue equals the trace of a matrix, and thus R_0 can be obtained as the trace of the NGM (Diekmann et al., 2010).

Appendix 1 shows the NGM for the population with linkage equilibrium and in Hardy-Weinberg Equilibrium (HWE) described by Equations (2)-(4). R_0 is given by the trace of the NGM:

$$R_0 = \bar{\gamma}\,\bar{\varphi}\,c/\alpha\,, \qquad (5)$$

where α is the recovery rate, which is assumed to be the same for all individuals in the population.

The NGM was also constructed for the more general case of a population that deviates from HWE and linkage equilibrium. For that case, R_0 is given by (Appendix 2):

$$R_0 = \left(\bar{\gamma}\,\bar{\varphi} + D\,\frac{(1+F_{IS})}{2}\,\frac{(2\gamma_g-\bar{\gamma})(2\varphi_f-\bar{\varphi})}{(1-p_g)(1-p_f)}\right)\frac{c}{\alpha}\,, \qquad (6)$$

where F_{IS} is the inbreeding coefficient and measures deviation of the population from HWE. It is a function of observed heterozygosity (H_o) and expected heterozygosity (H_e) in the population,

$F_{IS} = 1 - \frac{H_o}{H_e}$.

The D measures the deviation of the population from linkage equilibrium and expresses the excess of coupling phase haplotypes(Falconer and Mackay, 1996),

$D = p_{gf}p_{GF} - p_{Gf}p_{gF}$.

The second term in brackets in Equation 6 is the covariance between susceptibility and infectivity of individuals in the population. When either (i) $D = 0$, or $(ii) F_{IS} = -1$, that is, full dis-assortative ordering of alleles over diploid organisms ($H_o = 2H_e = 1$, which requires p = ½), or (iii) there is no variance in either of the two traits ($\bar{\gamma} = 2\gamma_g$ or $\bar{\varphi} = 2\varphi_f$), then there is no covariance between the two traits and R_0 is given by Equation (5).

Table 2.2. Scenarios and parameter values

Parameters	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Allele effect at infectivity locus				
φ_f	0.6	0.6	1	2.4
φ_F	1	1	1	0.6
Variation at				
Susceptibility locus	Yes	Yes	Yes	Yes
Infectivity locus	Yes	Yes	No	Yes
Relatedness $\boldsymbol{r}$	0	0 - 1	0 - 1	0 or 0.1
Linkage Disequilibrium, *D*	0	0	0	-0.20
Recombination rate ϑ	0.5	0.5	0.5	0

NB: Throughout the four scenarios, contact rate, c = 2, recovery rate $\alpha = 0.5$ and allele effect at susceptibility locus $\gamma_g = 1$ and $\gamma_G = 0.6$ was used. Allele frequencies at both loci were set at 0.5. The r^2 statistic corresponding to D = - 0.20 equals 0.64.

2.2.3 Individual breeding value for R_0

Equation (5) gives R_0, which is an emergent trait of the population, that is. a trait that arises when the different individuals (susceptible and infectious) interact (Dawkins, 2006). The objective here, however, is to define individual breeding values for R_0. We use results from the field of IGEs to define breeding value for R_0. An IGE is heritable effect of an individual on the trait value of another individual (Griffing, 1967; Griffing, 1976; Griffing, 1981; Moore et al., 1997; Wolf et al., 1998; Muir, 2005). Hence, infectivity is an IGE, as an individual's infectivity affects the disease status of its contacts. Moore et al. (1997) and (Bijma et al., 2007) show how breeding value and genetic variance can be defined for such traits. Bijma (2011) shows how the approach can be generalized to any trait, including traits that are an emerging property of a population, such as R_0. They propose a (total) breeding value that follows from the genetic mean of the population, rather than from individual trait values.

In classical quantitative genetics, breeding value is the sum of the average effects of an individual's alleles on its trait value, where the average effects equal the partial regression coefficients of individual trait values on individual allele count (Fisher, 1919; Lynch and Walsh, 1998). For traits affected by IGEs, the total breeding value is the sum of the average effects of an individual's alleles on the mean trait value of the population (Bijma, 2011). For an emergent trait, however,

there is only a single trait value for the entire population, and the average effects of alleles on that trait follow from the partial derivatives of the trait value with respect to allele frequency, rather than from partial regression of individual trait values on allele count. This is analogous to the derivation of economic values in livestock genetic improvement. Applying this approach to R_0 (Equation (5)) with linkage equilibrium and HWE, average effect of the *g*-allele equals

$$\frac{\partial R_0}{\partial p_g} = 2\bar{\varphi}(\gamma_g - \gamma_G)\frac{c}{\alpha}, \tag{7a}$$

and the average effect of the *f*-allele on R_0 equals

$$\frac{\partial R_0}{\partial p_f} = 2\bar{\gamma}(\varphi_f - \varphi_F)\frac{c}{\alpha}, \tag{7b}$$

Consequently, the individual breeding value for R_0 is given by

$$A_{R_0,i} = 2[\bar{\varphi}(\gamma_g - \gamma_G)p_{g,i} + \bar{\gamma}(\varphi_f - \varphi_F)p_{f,i}]\frac{c}{\alpha}, \tag{7c}$$

where $p_{g,i}$ and $p_{f,i}$ refer to the allele frequencies in individual *i*, thus taking values of 0, ½ or 1. The equation for $A_{R_{0,i}}$ for the population that deviates from HWE and with LD is presented in Appendix 2.

In the following, we will refer to $A_{R_0,i}$ as the breeding value for R_0 of individual *i*. Note that, in contrast to the pairwise transmission rate parameter β_{ij}, an individual's breeding value for R_0 is entirely a function of its own genes. This is because an individual transmits its own genes to its offspring, which may differ from the genes affecting its own disease phenotype.

The relationship between the breeding values of the individuals in a population of *n* individuals and R_0 of that population is:

$$R_0 = 4\gamma_G\varphi_F\frac{c}{\alpha} + \frac{\sum_{i=1}^{n} A_{R_0,i}}{n} - 4(\gamma_g - \gamma_G)(\varphi_f - \varphi_F)p_g p_f\frac{c}{\alpha} \tag{8}$$

The first term in Equation (8) is the intercept that determines the magnitude of R_0, but it does not depend on the allele frequencies and is not needed in the breeding value. The last term is there because of the nonlinear relationship between R_0 (Equation (5)) and susceptibility and infectivity. From Equation (8), it can be seen that changes in breeding value for R_0 will lead to corresponding changes (in magnitude and direction) in R_0 itself. Only when also the frequencies in whole populations (p_g, p_f) are changing, the change in R_0 will be more than the change in breeding values due to this last term. In that case, selection that reduces both susceptibility and infectivity will lead to a greater reduction in R_0 than predicted by the breeding values. Response to selection in R_0 will equal the change in average individual breeding value for R_0,

$$dR_0 = d\overline{A_{R_0}}. \tag{9}$$

Hence, a (small) change in average individual breeding value for R_0 due to selection will generate the same change in R_0. Thus, just as with an ordinary breeding value (Fisher, 1919; Lynch and Walsh, 1998), for a small change in allele frequency, the change in mean breeding value for R_0 equals response to selection in R_0.

2.2.4 Heritable variation in R_0

Response to selection in any trait, including emergent traits such as R_0, can be expressed as the product of intensity of selection ι, accuracy of selection ρ_T, and total genetic standard deviation for that trait σ_{A_T} (Bijma, 2011),

$$R = \iota \rho_T \sigma_{A_T} \quad (10)$$

In the above equation, response to selection R is change in mean trait value from one generation to the next. The selection intensity ι is the selection differential expressed in standard deviation units. Accuracy of selection ρ_T is the correlation between the total breeding value and the selection criterion in the candidates for selection, and σ_{A_T} is the standard deviation in total breeding value for the trait in the candidates for selection. Selection intensity and accuracy of selection are scale free parameters and do not include any information about the heritable variance in the trait. Standard deviation in total breeding value, on the other hand, reflects the potential of the population to response to selection. Note that heritable variation in the context of Equation (10) strictly refers to the potential of a population to respond to selection, and may differ from the classical additive genetic variance in a trait. R_0, for example, has no classical additive genetic variance, as there exist no individual phenotypes for R_0. Thus, in the following, heritable variation in R_0 will refer to the potential for genetic change in R_0, and not to the additive genetic component of phenotypic variation in R_0 among individuals. This conceptual difference is discussed in detail in Bijma (2011).

From the above, it follows that heritable variation in R_0 equals the variance in breeding value for R_0 among individuals in the population. We drop the prefix "total" from breeding value and heritable variation, since R_0 has no classical breeding value. Taking the variance of Equation (7c), assuming linkage equilibrium, shows that heritable variation in R_0 equals

$$var(A_{R_0}) = 2\left(p_g(1-p_G)\bar{\varphi}^2\left(\gamma_g - \gamma_G\right)^2 + p_f(1-p_f)\bar{\gamma}^2\left(\varphi_f - \varphi_F\right)^2\right)\left(\frac{c}{\alpha}\right)^2 \quad (11)$$

where $var(A_{R_0})$ is the variance among individuals in breeding value for R_0. Hence, Equation (11) shows how heritable variation in R_0 depends on the susceptibility and infectivity effects of alleles and on the allele frequencies in the population.

The expression in Equation (11) may be recognized as the sum of the additive genetic variances at two independent loci. Additive genetic variance at a single locus is traditionally written as $\mathbf{2}\boldsymbol{p}(\mathbf{1}-\boldsymbol{p})\boldsymbol{\alpha}^{\mathbf{2}}$, $\boldsymbol{\alpha}$ denoting the average effect of an allele substitution(Falconer and Mackay, 1996). In Equation 11, the average effect

at the susceptibility locus equals $\overline{\boldsymbol{\varphi}}(\boldsymbol{\gamma}_{\boldsymbol{g}}-\boldsymbol{\gamma}_{\boldsymbol{G}})\frac{\boldsymbol{c}}{\boldsymbol{\alpha}}$, and average effect at the infectivity locus equals $\overline{\boldsymbol{\gamma}}(\boldsymbol{\varphi}_{\boldsymbol{f}}-\boldsymbol{\varphi}_{\boldsymbol{F}})\frac{\boldsymbol{c}}{\boldsymbol{\alpha}}$ (see also Equation (7a-c)).

2.2.5 Utilization of Heritable Variation in R_0

Efficient reduction of R_0 by means of selective breeding requires selection schemes that optimally utilize the heritable variation in R_0. Because an individual's infectivity represents an IGE, that is, a heritable effect of the individual on the disease status of other individual within the same epidemiological unit, optimal breeding schemes for traits affected by IGEs may provide a clue for the design of optimal schemes for reducing R_0. For traits affected by IGEs, group selection and relatedness among interacting individuals ('kin selection') increase response to selection (Griffing, 1967; Griffing, 1976; Bijma and Wade, 2008). Moreover, Bijma (2011) shows that relatedness among interacting individuals in general tends to increase response to selection for traits that have an IGE. We, therefore, considered a group-structured population, where group mates can be genetically related. The objective of this section is not to precisely quantify or predict response to selection, but to identify and illustrate important factors affecting it.

To investigate mechanisms affecting response in R_0, a simulation study was performed on a population with discrete generations. The genetic model was the same as described above. The population was sub-divided into 100 groups of 100 individuals each. In each group, an epidemic was started by a single randomly infected individual. After the end of an epidemic, selection was based on individual disease status (0/1), where only those that escaped the infection were selected from each group to be parent of the next generation. For the next generation, selected parents were mated randomly and offspring genotypes were randomly sampled based on the parental genotypes. The size and the number of groups were kept constant throughout the generations.

Each group in the population was set up in such a way that group mates showed a certain degree of genetic similarity, which we refer to as "relatedness", *r*, here. The term "relatedness" has different meanings in different scientific disciplines. In animal breeding, for example, relatedness is implicitly understood as "pedigree relatedness". In sociobiology, such as in studies on the evolution of

altruism, on the other hand, relatedness is interpreted as a more general measure of genetic similarity, irrespective of the cause of that similarity; for example as a genetic regression coefficient (Hamilton, 1970); see also (Frank, 1998). Here we define relatedness as the correlation between the allele count of group mates, irrespective of the cause of that correlation. This definition agrees with the use of relatedness in animal breeding applications, such as selection index theory and genomic relationship matrices, where the current population is treated as the base population (Falconer and Mackay, 1996).

Relatedness at the susceptibility locus, r_γ, and at the infectivity locus, r_φ, were allowed to differ. To achieve a certain relatedness among group mates, a fraction f of fully related individuals was added to each group, supplemented by a fraction 1-f of randomly selected individuals. We did not consider negative values for relatedness, because the lower bound for relatedness is practically zero when group size equals 100 individuals, ($r_{min} = -1/99$). Appendix 3 shows that the required fraction equals the square root of relatedness. Thus, a fraction $\sqrt{r_\gamma}$ of individuals that were fully related to each other at the susceptibility locus, and a fraction $\sqrt{r_\varphi}$ of individuals that were fully related to each other at the infectivity locus were added to each group. As each individual carries both loci, these additions cannot be done independently; details of the strategy to jointly make those additions are given in Appendix 4.

The simulation was further extended to allow for a certain degree of LD between both loci. However, for a given LD in the population, there exists an upper and lower bound for r_γ given r_φ and vice versa. For example, when both loci are in strong positive LD and relatedness is zero at the susceptibility locus, then it is not possible to have very high relatedness at the infectivity locus. Appendix 5 provides expressions for those bounds.

Four different scenarios were simulated (Table 2.2). First, a scenario with heritable variation at both the susceptibility and the infectivity locus and groups created randomly with respect to relatedness r among group mates. No LD and a recombination rate θ of 0.5 between both loci were further assumed. Second, varying degrees of relatedness were used, which were the same at both loci. Third, to investigate a potential effect of relatedness on response in susceptibility, heritable variation was simulated at the susceptibility locus only, for varying degrees of relatedness among group mates. Finally, to investigate the potential effect of relatedness on response in R_0 in the case where there is strong negative LD between both loci and no recombination, a scenario with a relatedness of either 0 or 0.1 at both loci was simulated.

2.3 Simulation results

In the first scenario, which had unrelated group mates, a response to selection was observed only at the susceptibility locus, where the G-allele became fixed after an average of 100 generations. At the infectivity locus, in contrast, only a random fluctuation of allele frequency was observed (Figure 2.1). Thus, with groups composed at random with respect to relatedness, no response was observed at the infectivity locus. As a result, in the final generation, the response in R_0 was limited.

In the second scenario, which had related group mates, response to selection was observed at both loci, and the population became fixed for the G-allele at susceptibility locus and for F-the allele at the infectivity locus (Figure 2.2 and 2.3). In this case, selection resulted in a greater reduction of R_0 than in the first scenario (Figure 2.4 *vs.* Figure 2.1). As relatedness among group mates increased, response was much faster in all three traits. As it was also faster on the susceptibility locus, this suggested that also the susceptibility locus showed an IGE.

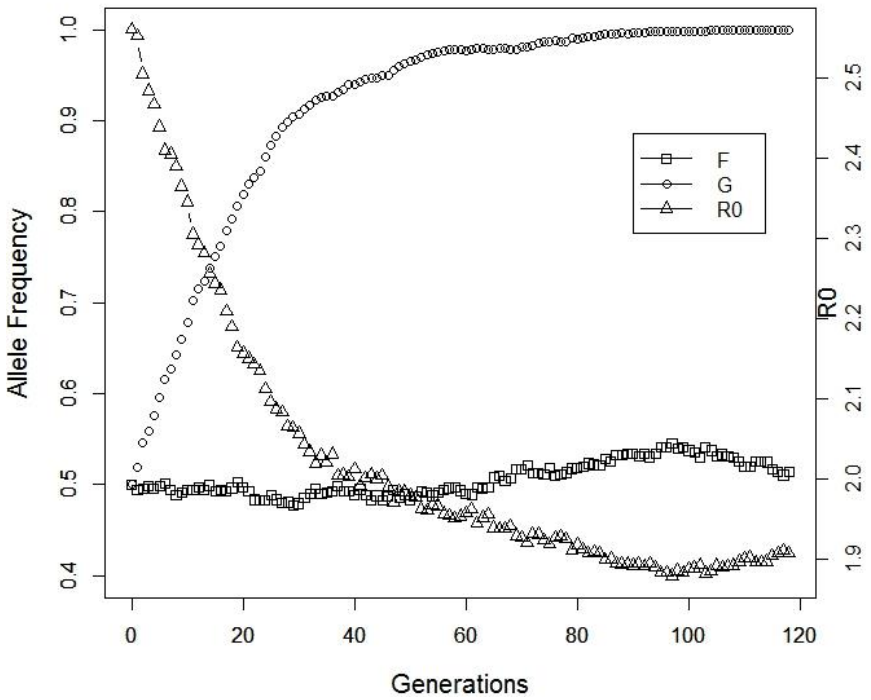

Figure 2.1 Allele frequency (F) at infectivity locus, allele frequency (G) at susceptibility locus and R_0 when there is no relatedness among group mates (Scenario 1, Table 2.2). Results are from one representative replicate.

To verify this IGE in susceptibility in the third scenario, we chose to have variation in the susceptibility only. Also in this case, the response at the susceptibility locus increased substantially when relatedness among group mates increased (Figure 2.5). For selection on individual phenotype, it is known that relatedness increases response in the IGEs, but not in the direct genetic effects (Griffing, 1976; Bijma and Wade, 2008). Thus this result suggests that (1) susceptibility not only has a direct genetic effect on the disease status of the individual itself but also has an IGE on the disease status of its groups mates, and (2) this indirect genetic variance is utilized by kin selection (see discussion), even in the absence of genetic variance in infectivity.

In the fourth scenario, which had strong negative LD and no recombination, the direction of response in R_0 depended on the relatedness among group mates. Without relatedness, selection fixed the G-allele irrespective of the linked allele at the infectivity locus. As a consequence, selection increased the frequency of *f*-allele yielding an increase rather than decrease of R_0. When relatedness $r_\gamma = r_\varphi = 0.1$ was used, however, selection caused fixation of GF haplotype, resulting in a decrease in R_0 (Figure 2.6). This result shows that kin-selection can prevent a maladaptive response to selection.

2.4 Discussion

The aim of this study was to define the breeding value and heritable variation for R_0. This was done for a diploid host population with genetic variation for susceptibility and infectivity. Breeding values of individuals were derived by finding the R_0, linearizing this value in the allele frequencies and substituting the individual's allele frequencies. The heritable variation that measures the potential for response in R_0 can then be found by taking the variance of the breeding values in the population. We applied this approach to a simple SIR-model with genetic variation in susceptibility and infectivity, and assuming separable mixing.

The second focus of this paper was to investigate the mechanisms that affect response in R_0. Since genetic relatedness between interacting individuals is expected to increase response in the general case (Bijma, 2011), we hypothesised that this result would extend to R_0 and considered a group-structured population with related group members. Our results show that, with unrelated group members and no LD between both loci, selection based on individual disease status yields response in susceptibility only. In the absence of relatedness, response in

infectivity depends entirely on the correlation with susceptibility, which was zero in the absence of LD.

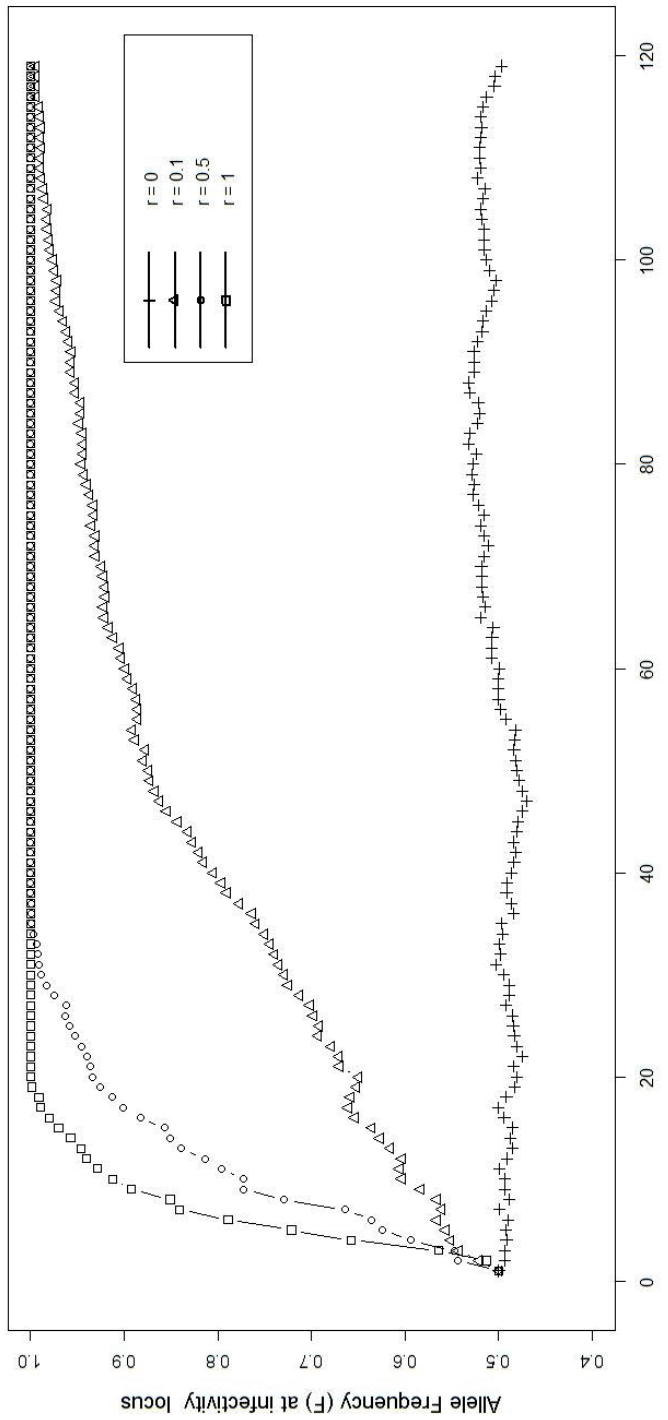

Figure 2.2 Allele frequency (F) at infectivity locus as relatedness among group mates increases from 0 to 1 (Scenario 2,Table 2.2). Results are from one representative replicate.

Relatedness among group members increased response in R_0 in two ways. First, with related group members, selection for individual disease status captures the heritable variation in infectivity. This occurs because an individual that carries the favourable allele for infectivity has group mates with a below-average infectivity, which increases its probability of escaping the epidemic, and thus being selected. Second, relatedness among group mates increases response in susceptibility. This occurs because an individual that carries the favourable allele for susceptibility on an average has fewer infected group mates, which increases its probability of escaping the epidemic and being selected. These results show that not only infectivity, but also susceptibility exhibits an IGE; at the same level of infectivity, individuals with lower susceptibility have a reduced chance of infecting others simply because they have a lower chance of being infected themselves. The net result of both mechanisms is a strong increase in response to selection in R_0 when relatedness increases. To quantify the impact of relatedness on the accuracy of selection for R_0, we calculated the correlation between the selection criteria (healthy/infected) and the breeding value for R_0.

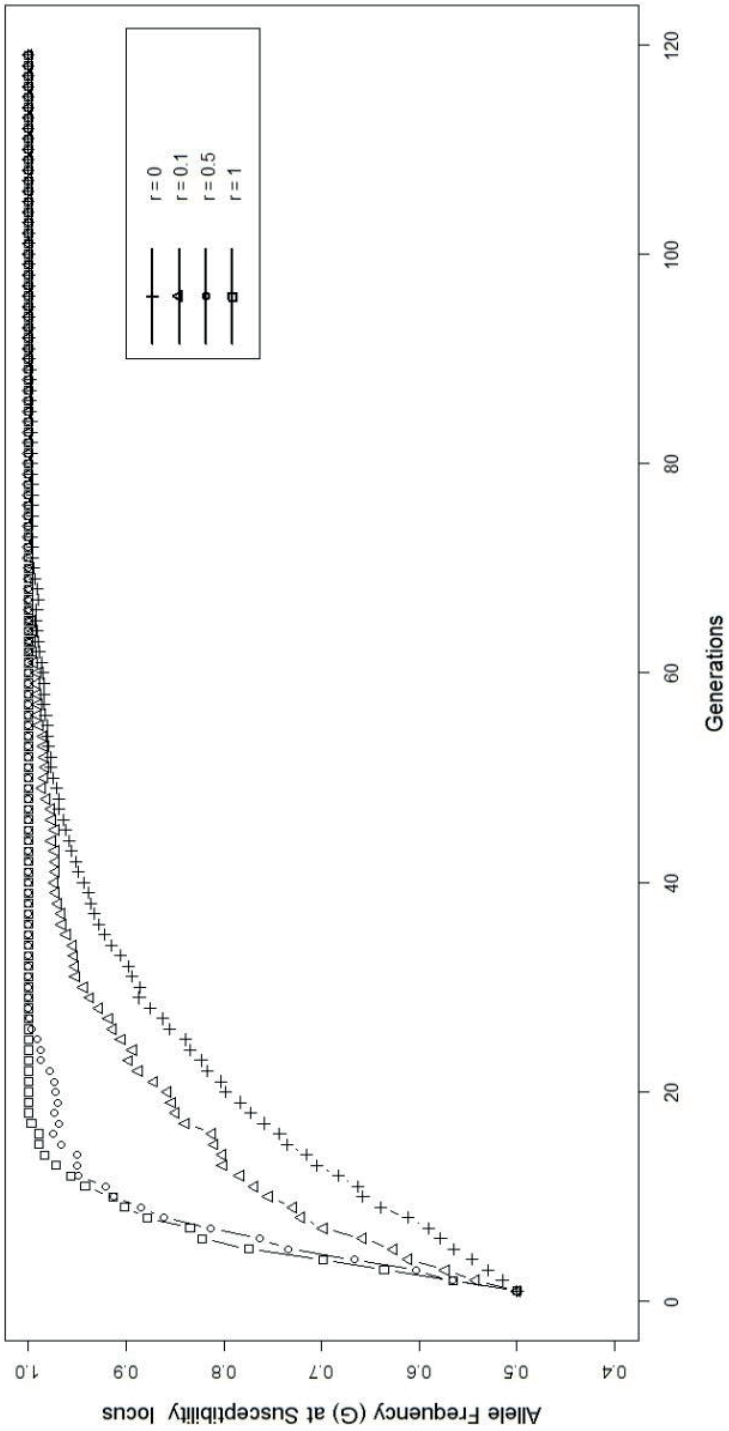

Figure 2.3. Allele frequency (G) at susceptibility locus as relatedness among group mates increases from 0 to 1 (Scenario 2, Table 2.2). Results are from one representative replicate.

Using the parameter values presented in scenario 2, Table 2.2, accuracy of selection increased from 0.05 to 0.24 when relatedness increased from 0 to 1. Thus, our study further supports the claim of Bijma (2011) that relatedness is an important factor in utilization of heritable variation in traits affected by IGEs.

Our results suggest that relatedness among interacting individuals can be used in livestock breeding programs aiming to reduce disease incidence. In current breeding strategies in livestock, data on individual disease status is connected to the pedigree of individuals to estimate breeding values. When interacting individuals are unrelated, those breeding values capture only the direct genetic effect, that is, the direct genetic part of susceptibility. Breeding values can be improved by also considering IGEs, for example, by fitting direct-indirect genetic effects models to data on disease status (Lipschutz-Powell et al., 2012). However, estimating direct and indirect breeding values for disease status is methodologically challenging because the linear mixed models traditionally used in quantitative genetics do not fit the nonlinear dynamics of infectious diseases (Lipschutz-Powell et al., 2012). The use of related group members may offer a low-tech solution, for capturing more of the heritable variation in R_0 without the need to explicitly model IGEs.

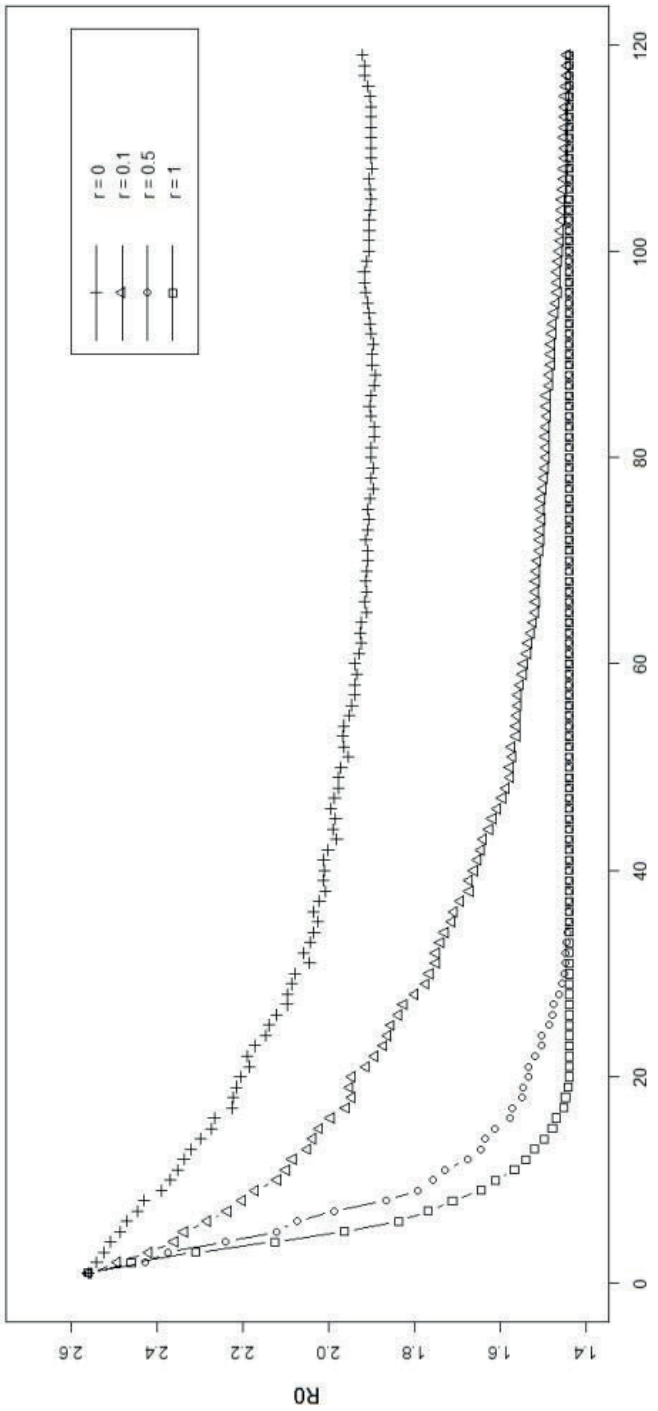

Figure 2.4. R_0 when relatedness among group members increases from 0 to 1 (Scenario 2, Table 2.2). Results are from one representative replicate.

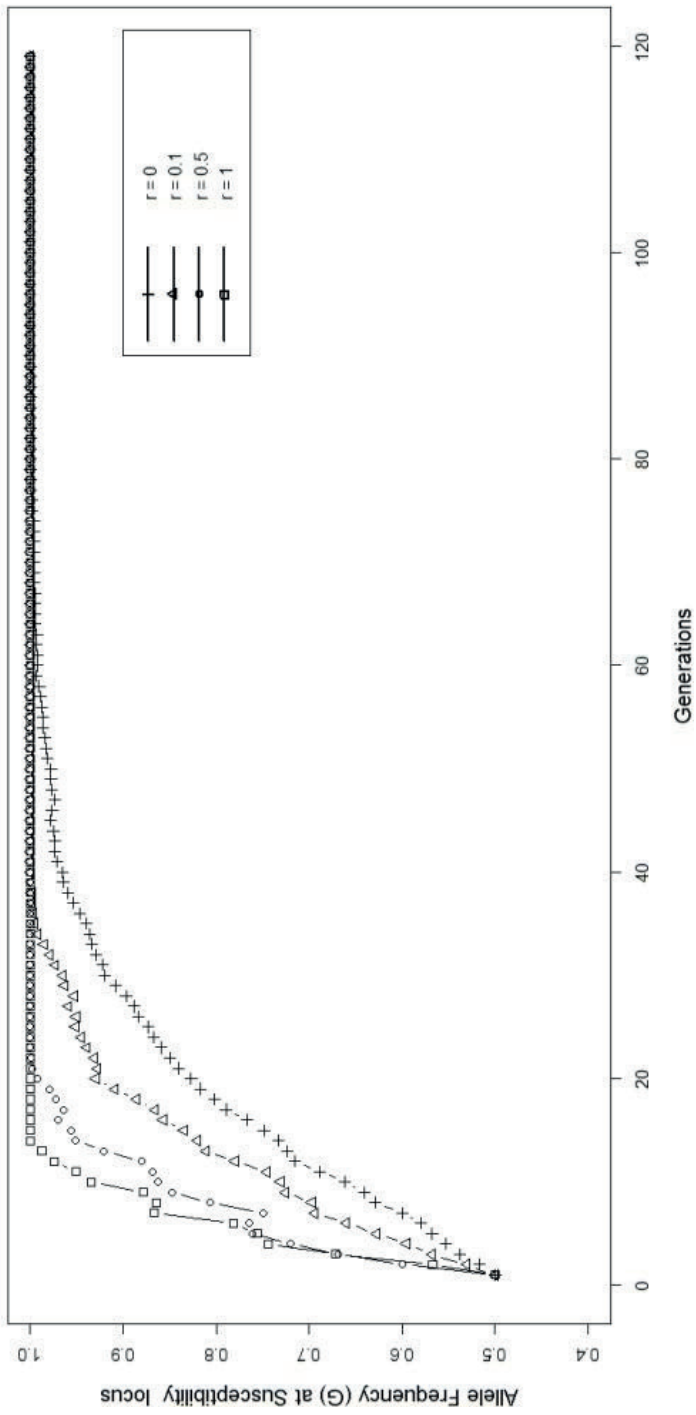

Figure 2.5. Allele frequency (G) at susceptibility locus as relatedness increases from 0 to 1 in the population with heritable variation at susceptibility locus only (Scenario 3, Table 2.2). Results are from one representative replicate.

In this work, we have assumed that the selection objective is to reduce R_0. While this is probably the obvious choice for epidemiologists, it may be unexpected for breeders who are not very familiar with R_0. For breeders, reducing disease incidence might be the more common objective. For example, in the context of our two-locus model, breeders might specify an objective $H_i = v_\gamma p_{g,i} + v_\varphi p_{f,i}$, where v_γ and v_φ are the so-called economic values for susceptibility and infectivity, respectively, which would be the partial derivatives of disease incidence with respect to the population allele frequencies p_g andp_f. However, both objectives are very similar, both for epidemic and endemic diseases. For epidemic diseases, the ultimately affected fraction of the population, known as the final size $1 - s(\infty)$, is determined by R_0 , as is shown by the final size equation: $\ln s(\infty) = R_0(s(\infty) - 1)$(Kermack and McKendrick, 1991). For endemic diseases, the equilibrium-affected fraction is given by: $1 - s(\infty) = 1 - 1/R_0$. Hence, the relationship between disease incidence and allele frequency occurs entirely via R_0, both for epidemic and endemic diseases. Thus, when the objective is to decrease incidence, the economic values for any disease trait, say x, that is, the partial derivatives of incidence with respect to that trait, can be written as

$$v_x = \frac{\partial i}{\partial x} = \frac{\partial i}{\partial R_0}\frac{\partial R_0}{\partial x}.$$

In this expression, the $\partial i/\partial R_0$ is a constant that is the same for all individuals in the population, and is independent of the disease trait considered (e.g. susceptibility or infectivity). Thus, the ranking of individuals will be the same, irrespective of whether they are ranked on breeding value for incidence or on breeding value for R_0.

Beware that breeding for incidence is not the same as breeding for susceptibility. When comparing breeding for susceptibility to breeding for R_0 or incidence, the latter is to be preferred because it also covers the heritable variation originating from infectivity (e.g. Figure 2.4 *vs*. 2.1).

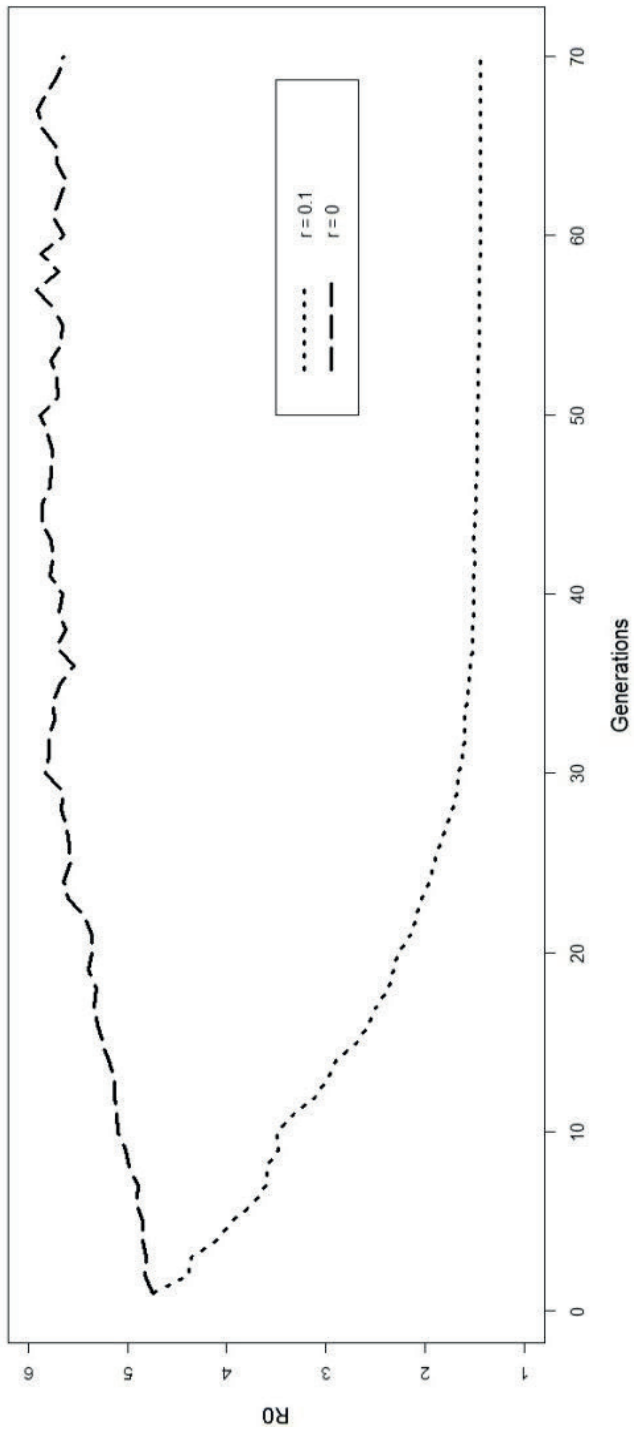

Figure 2.6. The effect of relatedness on response in mean R_0 in a population with strong negative Linkage Disequilibrium (D = -0.20, $\rightarrow r^2 = 0.64$) and no recombination between susceptibility and infectivity locus. For φ_f = 2.4 and φ_F = 0.6. Initial R_0 was 4.7. (Scenario 4a and 4b, Table 2.2). Results are from one representative replicate.

With respect to the evolution of parasite virulence, also the key role of kin selection has been recognized (Levin and Pimentel, 1981; Frank, 1996; Galvani, 2003). Much less attention has been given to the potential for kin selection acting on the host population. Using Monte Carlo simulation, Fix (1984) showed that the presence of kin groups in a small-scale human population considerably accelerated the increase in frequency of a resistance allele. Schliekelman (2007) seems to be the first who used rigorous mathematical modelling to investigate the impact of kin selection on the frequency of mutant alleles conferring resistance to the host. Moreover, despite the evidence of heterogeneity in infectivity (Woolhouse et al., 1997; Lloyd-Smith et al., 2005; Doeschl-Wilson et al., 2011), little attention has been given to the effect of kin selection on the frequency of alleles affecting infectivity in the host population. Our simulations show that, at least in theory, kin selection can greatly accelerate the evolution of R_0, because it utilizes the indirect genetic variance in both susceptibility and infectivity in the host population. For any actual case, the potential impact of kin selection will of course depend critically on the magnitude of this indirect genetic variance. Particularly, the component due to genetic variation in infectivity is unknown at present, but first steps towards estimating this component have recently been made (Lipschutz-Powell et al., 2012).

2.5 Acknowledgement

This study was financially supported by EU Marie Curie NematodeSystemHealth (ITN-2012-264639). The contribution of PB was supported by the foundation for applied sciences (STW) of the Dutch science council (NWO).

2.6 References

Anderson, R. M., and R. M. May. 1992. Infectious Diseases of Humans Dynamics and Control Oxfor Science Publications, USA.

Andreasen, V. 2011. The final size of an epidemic and its relation to the basic reproduction number. Bulletin of Mathematical Biology 73: 2305-2321.

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., W. M. Muir, and J. A. Van Arendonk. 2007. Multilevel selection 1: quantitative genetics of inheritance and response to selection. Genetics 175: 277-288.

Bijma, P., and M. Wade. 2008. The joint effects of kin, multilevel selection and indirect genetic effects on response to genetic selection. Journal of evolutionary biology 21: 1175-1188.

Dawkins, R. 2006. The selfish gene. Oxford university press.

Diekmann, O., J. Heesterbeek, and J. A. Metz. 1990. On the definition and the computation of the basic reproduction ratio R 0 in models for infectious diseases in heterogeneous populations. Journal of mathematical biology 28: 365-382.

Diekmann, O., J. Heesterbeek, and M. Roberts. 2010. The construction of next-generation matrices for compartmental epidemic models. Journal of The Royal Society Interface 7: 873-885.

Doeschl-Wilson, A. B. et al. 2011. Implications of host genetic variation on the risk and prevalence of infectious diseases transmitted through the environment. Genetics 188: 683-693.

Falconer, D., and T. Mackay. 1996. Introduction to quantitative genetics. 4.

Fisher, R. A. 1919. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. Transactions of the Royal Society of Edinburgh 52: 399-433.

Fix, A. G. 1984. Kin groups and trait groups: Population structure and epidemic disease selection. American Journal of Physical Anthropology 65: 201-212.

Frank, S. A. 1996. Models of parasite virulence. Quarterly Review of Biology: 37-78.

Frank, S. A. 1998. Foundations of social evolution. Princeton University Press.

Galvani, A. P. 2003. Epidemiology meets evolutionary ecology. Trends in ecology & evolution (Personal edition) 18: 132-139.

Gibson, J. R., and S. C. Bishop. 2005. Use of molecular markers to enhance resistance of livestock to disease: a global approach. Revue Scientifique Et Technique-Office International Des Epizooties 24: 343-353.

Griffing, B. 1967. Selection in Reference to Biological Groups .I. Individual and Group Selection Applied to Populations of Unordered Groups. Australian Journal of Biological Sciences 20: 127-&.

Griffing, B. 1976. Selection in reference to biological groups. V. Analysis of full-sib groups. Genetics 82: 703-722.

Griffing, B. 1981. A theory of natural selection incorporating interaction among individuals. II. Use of related groups. Journal of Theoretical Biology 89: 659-677.

Haldane, J. B. 1949. Disease and Evolution. La Ricerca Scientifica 19: 8.

Hamilton, W. D. 1970. Selfish and spiteful behaviour in an evolutionary model.

Heringstad, B., Y. M. Chang, D. Gianola, and G. Klemetsdal. 2005. Genetic Association Between Susceptibility to Clinical Mastitis and Protein Yield in Norwegian Dairy Cattle. Journal of Dairy Science 88: 1509-1514.

Kermack, W., and A. McKendrick. 1991. Contributions to the mathematical theory of epidemics—I. Bulletin of mathematical biology 53: 33-55.

Levin, S., and D. Pimentel. 1981. Selection of intermediate rates of increase in parasite-host systems. American Naturalist: 308-315.

Lipschutz-Powell, D., J. A. Woolliams, P. Bijma, and A. B. Doeschl-Wilson. 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, J. O., S. J. Schreiber, P. E. Kopp, and W. M. Getz. 2005. Superspreading and the effect of individual variation on disease emergence. Nature 438: 355-359.

Lynch, M., and B. Walsh. 1998. Genetics and analysis of quantitative traits. Sinauer, Sunderland, Mass.

Moore, A. J., E. D. Brodie III, and J. B. Wolf. 1997. Interacting phenotypes and the evolutionary process: I. Direct and indirect genetic effects of social interactions. Evolution: 1352-1362.

Muir, W. M. 2005. Incorporation of competitive effects in forest tree or animal breeding programs. Genetics 170: 1247-1259.

O'Brien, S. J., and J. F. Evermann. 1988. Interactive influence of infectious disease and genetic diversity in natural populations. Trends in Ecology & Evolution 3: 254-259.

Schliekelman, P. 2007. Kin selection and evolution of infectious disease resistance. Evolution 61: 1277-1288.

Wolf, J. B., E. D. Brodie Iii, J. M. Cheverud, A. J. Moore, and M. J. Wade. 1998. Evolutionary consequences of indirect genetic effects. Trends in Ecology & Evolution 13: 64-69.

Woolhouse, M. E. J. et al. 1997. Heterogeneities in the transmission of infectious agents: Implications for the design of control programs. Proceedings of the National Academy of Sciences 94: 338-342.

Appendix 1

This Appendix shows the construction of the Next Generation Matrix (NGM) (Diekmann et al., 2010) and R_0 for a diploid population where there is no Linkage Disequilibrium between the locus a

ffecting susceptibility and the locus affecting infectivity. In such population, we have nine types of individuals for the combination of their genotype for susceptibility (gg, gG, GG) and infectivity (ff, fF, FF). Thus the NGM has 9 rows and 9 columns. The column of the matrix represents the contributions to the next generation by infectious individuals of the genotype written above the column ("cause"). The rows indicate the genotypes of the susceptible individuals that become infected ("consequence"). In the following we present the NGM on three rows; the first row gives columns 1 through 3, the second columns 4 through 6, and the final row columns 7 through 9. The NGM uses the transmission rate parameters between genotypes, which are given by

$\beta_1 = \gamma_{gg}\varphi_{ff}c$ $\beta_2 = \gamma_{gg}\varphi_{fF}c$ $\beta_3 = \gamma_{gg}\varphi_{FF}c$

$\beta_4 = \gamma_{gG}\varphi_{ff}c$ $\beta_5 = \gamma_{gG}\varphi_{fF}c$ $\beta_6 = \gamma_{gG}\varphi_{FF}c$

$\beta_7 = \gamma_{GG}\varphi_{ff}c$ $\beta_8 = \gamma_{GG}\varphi_{fF}c$ $\beta_9 = \gamma_{GG}\varphi_{FF}c$

	$ggff$	$ggfF$	$ggFF$
$ggff$	$p_g^2 \mathrm{p}_f^2 \beta_1$	$p_g^2 \mathrm{p}_f^2 \beta_2$	$p_g^2 \mathrm{p}_f^2 \beta_3$
$ggfF$	$p_g^2 2\, p_f(1-p_f)\beta_1$	$p_g^2 2\, p_f(1-p_f)\beta_2$	$p_g^2 2\, p_f(1-p_f)\beta_3$
$ggFF$	$p_g^2(1-p_f)^2\beta_1$	$p_g^2(1-p_f)^2\beta_2$	$p_g^2(1-p_f)^2\beta_3$
$gGff$	$2p_g(1-p_g)\mathrm{p}_f^2\beta_4$	$2p_g(1-p_g)\mathrm{p}_f^2\beta_5$	$2p_g(1-p_g)\mathrm{p}_f^2\beta_6$
$gGfF$	$2p_g(1-p_g)2\, p_f(1-p_f)\beta_4$	$2p_g(1-p_g)2\, p_f(1-p_f)\beta_5$	$2p_g(1-p_g)2\, p_f(1-p_f)\beta_6$
gGFF	$2p_g(1-p_g)(1-p_f)^2\beta_4$	$2p_g(1-p_g)(1-p_f)^2\beta_5$	$2p_g(1-p_g)(1-p_f)^2\beta_6$
$GGff$	$(1-p_g)^2\,\mathrm{p}_f^2\beta_7$	$(1-p_g)^2\,\mathrm{p}_f^2\beta_8$	$(1-p_g)^2\,\mathrm{p}_f^2\beta_9$
GGfF	$(1-p_g)^2 2\, p_f(1-p_f)\beta_7$	$(1-p_g)^2 2\, p_f(1-p_f)\beta_8$	$(1-p_g)^2 2\, p_f(1-p_f)\beta_9$
GGFF	$(1-p_g)^2(1-p_f)^2\beta_7$	$(1-p_g)^2(1-p_f)^2\beta_8$	$(1-p_g)^2(1-p_f)^2\beta_9$

	$gGff$	$gGFF$	$gGFF$
$ggff$	$p_g^2\, p_f^2\, \beta_1$	$p_g^2\, p_f^2\, \beta_2$	$p_g^2\, p_f^2 \beta_3$
$ggfF$	$p_g^2 2\, p_f(1-p_f)\, \beta_1$	$p_g^2 2\, p_f(1-p_f)\, \beta_2$	$p_g^2 2\, p_f(1-p_f)\beta_3$
$ggFF$	$p_g^2(1-p_f)^2\, \beta_1$	$p_g^2(1-p_f)^2 \beta_2$	$p_g^2(1-p_f)^2\, \beta_3$
$gGff$	$2p_g(1-p_g)p_f^2\, \beta_4$	$2p_g(1-p_g)p_f^2\, \beta_5$	$2p_g(1-p_g)p_f^2\, \beta_6$
$gGfF$	$2p_g(1-p_g)2\, p_f(1-p_f)p_f^2\, \beta_4$	$2p_g(1-p_g)2\, p_f(1-p_f)\, \beta_5$	$2p_g(1-p_g)2\, p_f(1-p_f)\beta_6$
$gGFF$	$2p_g(1-p_g)(1-p_f)^2\, \beta_4$	$2p_g(1-p_g)(1-p_f)^2\, \beta_5$	$2p_g(1-p_g)(1-p_f)^2\, \beta_6$
$GGff$	$(1-p_g)^2\, p_f^2\, \beta_7$	$(1-p_g)^2\, p_f^2\, \beta_8$	$(1-p_g)^2 p_f^2 \beta_9$
$GGfF$	$(1-p_g)^2 2\, p_f(1-p_f)\, \beta_7$	$(1-p_g)^2 2\, p_f(1-p_f)\, \beta_8$	$(1-p_g)^2 2\, p_f(1-p_f)\beta_9$
$GGFF$	$(1-p_g)^2(1-p_f)^2\, \beta_7$	$(1-p_g)^2(1-p_f)^2\, \beta_8$	$(1-p_g)^2(1-p_f)^2\, \beta_9$

	$GGff$	$GGfF$	$GGFF$
$ggff$	$p_g^2\, p_f^2 \beta_1$	$p_g^2\, p_f^2\, \beta_2$	$p_g^2\, p_f^2\, \beta_3$
$ggfF$	$p_g^2 2\, p_f(1-p_f)\beta_1$	$p_g^2 2\, p_f(1-p_f)\, \beta_2$	$p_g^2 2\, p_f(1-p_f)\, \beta_3$
$ggFF$	$p_g^2(1-p_f)^2\, \beta_1$	$p_g^2(1-p_f)^2\, \beta_2$	$p_g^2(1-p_f)^2\, \beta_3$
$gGff$	$2p_g(1-p_g)p_f^2 \beta_4$	$2p_g(1-p_g)p_f^2\, \beta_5$	$2p_g(1-p_g)\, p_f^2\, \beta_6$
$gGfF$	$2p_g(1-p_g)2\, p_f(1-p_f)\beta_4$	$2p_g(1-p_g)2\, p_f(1-p_f)\, \beta_5$	$2p_g(1-p_g)2\, p_f(1-p_f)\, \beta_6$
$gGFF$	$2p_g(1-p_g)(1-p_f)^2 \beta_4$	$2p_g(1-p_g)(1-p_f)^2 \beta_5$	$2p_g(1-p_g)(1-p_f)^2\, \beta_6$
$GGff$	$(1-p_g)^2\, p_f^2 \beta_7$	$(1-p_g)^2\, p_f^2\, \beta_8$	$(1-p_g)^2\, p_f^2\, \beta_9$
$GGfF$	$(1-p_g)^2 2\, p_f(1-p_f)\beta_7$	$(1-p_g)^2 2\, p_f(1-p_f)\, \beta_8$	$(1-p_g)^2 2\, p_f(1-p_f)\, \beta_9$
$GGFF$	$(1-p_g)^2(1-p_f)^2 \beta_7$	$(1-p_g)^2(1-p_f)^2 \beta_8$	$(1-p_g)^2(1-p_f)^2 \beta_9$

R_0 is the dominant eigenvalue of the NGM. Since we have so-called separable mixing, where elements of the NGM are products of the rows and columns, the NGM has a single eigenvalue only, which therefore equals the trace of the NGM. Thus R_0 is the sum of the diagonal elements of the NGM (given in bold above),

$$R_0 = \{ p_g^2 p_f^2 \beta_1 + p_g^2 2\, p_f(1 - p_f)\beta_2 + p_g^2(1 - p_f)^2 \beta_3$$
$$+ 2p_g(1 - p_g)p_f^2 \beta_4 + 2p_g(1 - p_g)2\, p_f(1 - p_f)\beta_5 + 2p_g(1 - p_g)(1 - p_f)^2 \beta_6$$
$$+ (1 - p_g)^2\, p_f^2 \beta_7 + (1 - p_g)^2 2\, p_f(1 - p_f)\beta_8 + (1 - p_g)^2(1 - p_f)^2 \beta_9 \}\frac{c}{\alpha}$$
$$= \{(p_g^2 \gamma_{gg} + 2\, p_g(1 - p_g)\gamma_{gG} + (1 - p_g)^2 \gamma_{GG})(p_f^2 \varphi_{ff} + 2\, p_f(1 - p_f)\varphi_{fF} +$$
$$(1 - p_f)^2 \varphi_{FF})\frac{c}{\alpha}$$
$$= \{(p_g^2\, 2\gamma_g + 2\, p_g(1 - p_g)\gamma_g + 2\, p_g(1 - p_g)\gamma_G + (1 - p_g)^2 2\gamma_G)(p_f^2\, 2\varphi_f +$$
$$2\, p_f(1 - p_f)\varphi_f + 2\, p_f(1 - p_f)\varphi_F + (1 - p_f)^2 2\varphi_F\}\frac{c}{\alpha}$$
$$=$$
$$\{[p_g(p_g 2\gamma_g + 2(1 - p_g)\gamma_g) + (1 - p_g)(p_g 2\gamma_G + (1 - p_g)2\gamma_G][p_f(\, p_f 2\varphi_f +$$
$$(1 - p_f)2\varphi_f + (1 - p_f)(p_f 2\varphi_F + (1 - p_f)2\varphi_F]\}\frac{c}{\alpha}$$
$$= \{[2p_g\gamma_g + 2(1 - p_g)\gamma_G][2p_f\varphi_f + 2(1 - p_f)\varphi_F]\}\frac{c}{\alpha}$$
$$R_0 = \bar{\gamma}\bar{\varphi}\frac{c}{\alpha} \qquad \text{(A1)}$$

in which $\bar{\gamma} = 2p_g\gamma_g + 2(1 - p_g)\gamma_G$, and $\bar{\varphi} = 2p_f\varphi_f + 2(1 - p_f)\varphi_F$

Appendix 2

The NGM was also constructed for a the population that deviates from linkage equilibrium (LD) and Hardy Weinberg Equilibrium (HWE). Because of LD, the genotype *gGfF* has to be partitioned into the two possible haplotypes for this genotype, *gfGF* and *gFGf*. Hence, when accounting for LD, the NGM includes 10 distinct genotypes, rather than the 9 considered in Appendix 1 (Table A2-1).

Table A2-1. Possible haplotypes and genotypes

Haplotypes	gf	gF	Gf	GF
gf	$\boldsymbol{gfgf}$	$gfgF$	$gfGf$	$gfGF$
gF		$\boldsymbol{gFgF}$	$gFGf$	$gFGF$
Gf			$\boldsymbol{GfGf}$	$GfGF$
GF				$\boldsymbol{GFGF}$

To avoid over presentation of results, we only give the trace of the NGM, which equals R_0 because of the separable mixing assumption,

$$R_0 = \{p_{gfgf}\,\beta_{gfgf} + p_{gfgF}\,\beta_{gfgF} + p_{gfGf}\,\beta_{gfGf} + p_{gfGF}\,\beta_{gfGF} + p_{gFgF}\,\beta_{gFgF} + p_{gFGf}\,\beta_{gFGf} + p_{gFGF}\,\beta_{gFGF} + p_{GfGf}\beta_{GfGf} + p_{GfGF}\,\beta_{GfGF} + p_{GFGF}\,\beta_{GFGF}\}\,1/\alpha \quad \text{(A2-1)}$$

Here β_{vwxy} represents the transmission rate parameter within a genotype, *i.e.*, from genotype *vwxy* to genotype *vwxy*,

$\beta_{vwxy} = \gamma_{vx}\varphi_{wy}c$.

For example, $\beta_{gFGF} = \gamma_{gG}\varphi_{FF}c$.

Haplotype frequencies are ,

$f_{gf} = p_g p_f + D$

$f_{gF} = p_g\,(1 - p_f) - D$

$f_{Gf} = (1 - p_g)p_f - D$

$f_{GF} = (1 - p_g)(1 - p_f) + D$

Where *D* is the usual measure of linkage disequilibrium (see main text).

Genotypes frequencies are

$p_{gfgf} = f_{gf}(f_{gf} + (1 - f_{gf})F_{IS})$

$p_{gfgF} = 2\,f_{gf}\,f_{gF}(1 - F_{IS})$

$p_{gfGf} = 2\,f_{gf}\,f_{Gf}(1 - F_{IS})$

$p_{gfGF} = 2\,f_{gf}f_{GF}(1 - F_{IS})$

$p_{gFgF} = f_{gF}(f_{gF} + (1 - f_{gF})F_{IS})$

$$p_{gFGf} = 2\,f_{gF}\,f_{Gf}(1 - F_{IS})$$
$$p_{gFGF} = 2\,f_{gF}\,f_{GF}(1 - F_{IS})$$
$$p_{GfGf} = f_{Gf}(f_{Gf} + (1 - f_{Gf})F_{IS})$$
$$p_{GfGF} = 2\,f_{Gf}\,f_{GF}(1 - F_{IS})$$
$$p_{GFGF} = f_{GF}(f_{GF} + (1 - f_{GF})F_{IS})$$

After few steps of algebraic manipulation Equation A2-1will reduce to,

$$R_0 = \left(\bar{\gamma}\,\bar{\varphi} + D\,\frac{(1+F_{IS})}{2}\,\frac{(2\gamma_g - \bar{\gamma})(2\varphi_f - \bar{\varphi})}{(1-p_g)(1-p_f)}\right)\frac{c}{\alpha},$$

Individual breeding values for R_0 were obtained by linearizing R_0 in the allele frequencies, using partial first derivatives, and subsequently substituting individual allele frequencies (i.e. 0, ½ or 1),

$$A_{R_0,i} = \frac{\partial R_0}{\partial p_g}\left(p_{g,i} - \bar{p}_g\right) + \frac{\partial R_0}{\partial p_f}\left(p_{f,i} - \bar{p}_f\right). \qquad \text{(A2-3)}$$

$$A_{R_0,i} = 2\left(\bar{\varphi}\left\{\left(\gamma_g - \gamma_G\right) - D\,\frac{(1+F_{IS})}{2}\,\frac{\gamma_g}{(1-p_f)(1-p_g)^2}\right\}p_{g,i} + \bar{\gamma}\left\{\left(\varphi_f - \varphi_F\right) - D\,\frac{(1+F_{IS})}{2}\,\frac{\varphi_f}{(1-p_g)(1-p_f)^2}\right\}p_{f,i}\right) \qquad \text{(A2-4)}$$

Appendix 3

As mentioned in the main text, relatedness at the susceptibility locus, r_γ and at the infectivity locus, r_φ were allowed to be different. With a single bi-allelic locus, pairwise relatedness between individuals takes only three discrete values. However, our interest is in a continuum of the average relatedness among the individuals that together make up a group. To achieve a certain average relatedness among group mates, a fraction *f* of fully related individuals was added to each group, supplemented by a fraction 1-*f* of randomly selected individuals. In this appendix we show that the required fraction equals the square root of relatedness at each locus, that is a fraction $f_\gamma = \sqrt{r_\gamma}$ of random individuals will be replaced by individuals that were fully related to each other at the susceptibility locus, and for the infectivity locus this is a fraction $f_\varphi = \sqrt{r_\varphi}$. We defined

relatedness as the correlation between the genotypes of two group mates, say *x* and *y*,

$$r = corr(x,y) = \frac{cov(x,y)}{\sqrt{var(x)\ var(y)}} \quad \text{(A3-1)}$$

Since the same theory applies to both loci, we will show the derivation for the susceptibility locus only.

Because the addition strategy should not change allele frequency in the population, nor affect the Hardy-Weinberg equilibrium, the population needs to have three types of groups. The first type has gg-individuals added to the group. The second type has gG -individuals added, and the third type has GG-individuals added. The number of groups of the first type equals $no.groups * p^2$, the number of groups of the second type equals $no.groups * 2p(1-p)$, and the number of groups of the third type equals $no.groups * (1-p)^2$, where p is the frequency of the g-allele. The frequency of g in the three types of groups is then,

$p_1 = (f + (1-f)p)$ (A3-2)

$p_2 = (0.5\,f + (1-f)p)$ (A3-3)

$p_3 = ((1-f)p)$ (A3-4)

To derive the correlation, we first derive the covariance between genotypic values of group members,

$cov(x,y) = E(xy) - E(y)E(x)$

$E(xy) = p^2E(xy|1) + 2\,p(1-p)E(xy|2) + (1-p)^2E(xy|3)$

where, for example, $E(xy|1)$ denotes the expectation of the product of the genotypic values of two group members in a group of type 1. To simplify the derivation, without loss of generality, g was given an effect of 1, and G an effect of 0. Since we are interested in additive genetic relationship, resulting genotypic values are 2 for gg, 1 for gG and 0 for GG. Thus *x* and *y* denote genotypic values, taking values of either 0, 1 or 2. The possible genotypes of two individuals and the corresponding values for $E(xy|\text{group type})$ are presented in the table below. Since the genotypic value for GG equals zero, any pair of individuals involving at least one GG-individual has E(*xy*) = 0, and is therefore left out of the table.

Table A3-1. Possible genotypes and the expectation of the product of the genotypic values of two group members in a group of type 1, 2 and 3.

Possible genotypes	$E(xy\|1)$	$E(xy\|2)$	$E(xy\|3)$
gg/gg	$[2(p_1)^2][2(p_1)^2]$	$[2(p_2)^2]\,[2(p_2)^2]$	$[2(p_3)^2]\,[2(p_3)^2]$
gg/gG	$[2(p_1)^2][1\,(2p_1(1-\ p_1))]$	$[2(p_2)^2]\,[1(2p_2(1-p_{g/2}))]$	$[2(p_3)^2][\,1(2p_3(1-p_3))]$
gG/gg	$[1\,(2\,p_1(1-\ p_1))][2(p_1)^2]$	$[1(2p_2(1-p_2))]\,[2(p_2)^2]$	$[\,1(2p_3(1-p_3))][2(p_3)^2]$
gG/gG	$[1\,(2\,p_1(1-\ p_1))][1\,(2\,p_1(1\ p_1))]$	$[1(2p_2(1-p_2))]\,[1(2p_2(1-p_2))]$	$[\,1(2p_3(1-p_3))][1(2p_3(1-p_3))]$

If we insert Equations A4-2 for p_1 and Equation A4-3 and A4-4 for p_2 and p_3 respectively and sum up all the elements in each of the three column for $E(xy)$, we find,

$p^2E(xy/1) = 4f^2p^2 + 8fp^3 - 8f^2p^3 + 4p^4 - 8fp^4 + 4f^2p^4$

$2\,p(1-p)E(xy/2) = 2f^2p + 8fp^2 - 10f^2p^2 + 8p^3 - 24fp^3 + 16f^2p^3 - 8p^4 + 16fp^4 - 8f^2p^4$

$(1-p)^2E(xy/3) = 4p^2 - 8fp^2 + 4f^2p^2 - 8p^3 + 16fp^3 - 8f^2p^3 + 4p^4 - 8fp^4 + 4f^2p^4$

And since

$E(xy) = p^2.E(xy/1) + 2\,p(1-p)E(xy/2) + (1-p)^2E(xy/3)$,

Then

$E(xy) = 4f^2p^2 + 8fp^3 - 8f^2p^3 + 4p^4 - 8fp^4 + 4f^2p^4 + 2f^2p + 8fp^2 - 10f^2p^2 + 8p^3 - 24fp^3 + 16f^2p^3 - 8p^4 + 16fp^4 - 8f^2p^4 + 4p^2 - 8fp^2 + 4f^2p^2 - 8p^3 + 16fp^3 - 8f^2p^3 + 4p^4 - 8fp^4 + 4f^2p^4$

$E(xy) = 2f^2p + 4p^2 - 2f^2p^2$

Next we need to calculate $E(x)$ and $E(y)$:

$E(x) = E(y) = [(2\,p^2) + 1\,(2p(1-p))]$

Then,

$E(x).E(y) = 4p^2$

Then covariance will be

$cov(x,y) = 2f^2p - 2f^2p^2$

Next, the variances are given by

$Var(x) = Var(y) = 2p(1-p)$

Then equation (A3-1) becomes,

$$corr(x,y) = \frac{2f^2p - 2f^2p^2}{\sqrt{2p(1-p)\sqrt{2p(1-p)}}}$$

$$corr(x,y) = \frac{2f^2p}{\sqrt{4p^2-8p^3+4p^4}} - \frac{2f^2p^2}{\sqrt{4p^2-8p^3+4p^4}}$$

Simplifying this expression yields

$corr(x, y) = r = f^2$. (A3-5)

Thus to achieve a certain relatedness, a fraction $f = \sqrt{r}$ of fully related individuals should be added to each group.

Appendix 4

This Appendix contains an example demonstrating the strategy to make additions in each group, so as to achieve a certain relatedness for susceptibility and infectivity among group mates. We considered 100 groups, each with 100 individuals. Let us assume that Linkage Disequilibrium $(D) = 0.15$, and that the allele frequency at susceptibility locus is 0.5 and allele frequency at infectivity locus equals 0.6. Thus, $p_g = 0.5$ and $p_f = 0.6$. The $r_\gamma = 0.75$ and $r_\varphi = 0.6$. It is assumed that the population is in Hardy-Weinberg Equilibrium. The haplotype frequencies will be,

$f_{gf} = p_g p_f + D$

$f_{gF} = (1 - p_g)p_f - D$

$f_{Gf} = p_g(1 - p_f) - D$

$f_{GF} = (1 - p_g)(1 - p_f) + D$

Table A4-1. Haplotype and genotype type frequencies assuming HWE.

Haplotypes	f_{gf}	f_{gF}	f_{Gf}	f_{GF}
f_{gf}	${f_{gf}}^2$	$2f_{gf}f_{gF}$	$2f_{gf}f_{Gf}$	$2f_{gf}f_{GF}$
f_{gF}		${f_{gF}}^2$	$2f_{gF}f_{Gf}$	$2f_{gF}f_{GF}$
f_{Gf}			${f_{Gf}}^2$	$2f_{Gf}f_{GF}$
f_{GF}				${f_{GF}}^2$

Since $r = f^2$, then the fraction f_γ of individuals that are fully related at their susceptibility locus will be $\sqrt{r_\gamma} = \sqrt{0.75} = 0.87$. And the fraction f_φ of individuals that are fully related at their infectivity locus will be $\sqrt{r_\varphi} = \sqrt{0.6} = 0.77$.

Because the required fraction is lowest for the infectivity locus, we start with the infectivity locus. Thus, in each of the 100 groups we added $\sqrt{r_\varphi} \times 100 = 77$ individuals that are fully related at their susceptibility and infectivity locus. The first $100 \times {f_{gf}}^2$ groups will contain 77 individuals with $gfgf$ genotype, $100 \times 2f_{gf}f_{gF}$ groups will contain 77 individuals with $gfgF$ genotype, $100 \times 2f_{gf}f_{Gf}$ groups will

contain 77 individuals with $gfGf$ genotype, $100 \times 2f_{gf}f_{GF}$ groups will contain 77 individuals with $gfGF$ genotype, $100 \times {f_{gF}}^2$ groups contain 77 individuals with

$gFgF$ genotype, $100 \times 2f_{gF}f_{Gf}$ groups will contain 77 individuals with $gFGf$ genotype, $100 \times 2f_{gF}f_{GF}$ groups contain 77 individuals $gFGF$ genotype, $100 \times {f_{Gf}}^2$ groups will contain 77 individuals with $GfGf$ genotype, $100 \times 2f_{Gf}f_{GF}$ groups will contain 77 individuals with $GfGF$ genotype and finally, $100 \times {f_{GF}}^2$ groups will contain 77 individuals with $GFGF$.

With respect to the infectivity locus, there are ${p_f}^2 \times 100 = 36$ groups that contain a fraction of individuals that are of *ff*, $2p_f(1-p_f) \times 100 = 48$ groups that contain a fraction of individuals that are of *fF* genotype and $(1-p_f)^2 \times 100 = 16$ groups that contain a fraction of individuals that are of *FF* genotype at their infectivity locus. Thus the desired additions for the infectivity locus are achieved.

With respect to the susceptibility locus, we have ${p_g}^2 \times 100 = 25$ groups that contain 77 individuals that are of *gg*, $2p_g(1-p_g) \times 100 = 50$ groups that contain 77 individuals that are of *gG* genotype and $(1-p_g)^2 \times 100 = 25$ groups that contain 77 individuals that are of *FF* genotype at their infectivity locus. For the susceptibility locus, however, the required number of individuals to be added equals $n \times \sqrt{r_\gamma} = 100 \times 0.87 = 87$. Since we have already added 77 individuals that are fully related at their susceptibility locus, what is left to add to the group is $87 - 77 = 10$ individuals Thus, the next addition will be 10 individuals that are fully related at their susceptibility locus, but taken at random with respect to their infectivity locus (so that relatedness as the infectivity locus is not affected). Therefore, for those groups that already have a fraction of individuals with gg genotype, we will add 10 more individuals that are off *gg* genotype. Analogously, to groups that already have a fraction of individuals with a certain genotype, 10 more individuals with that genotype are added. Since the groups size is assumed to be 100, the rest of the group, which are 100-87=13 individuals, will be assigned randomly.

Appendix 5

In this appendix we presented the lower (min) and upper (max) bound for r_γ given r_φ and *vice versa* for a given linkage disequilibrium, *D*. These bounds follow from the fraction of available individuals for the second addition step (see Appendix 4), which depends on the allele frequencies, *D*, and relatedness at the locus in the first addition step.

When $D > 0$

$$\max(r_\gamma \mid r_\varphi, D>0) = \left(\min\left[1 - \frac{D(1-\sqrt{r_\varphi})}{\min[p_g(1-p_f);(1-p_g)p_f]} ; 1 \right] \right)^2 \quad \text{(A5-1)}$$

$$\min(r_\varphi \mid r_\gamma, D>0) = \left(\max\left[1 + \frac{\min[p_g(1-p_f);(1-p_g)p_f](\sqrt{r_\gamma}-1)}{D} ; 0 \right] \right)^2 \quad \text{(A5-2)}$$

When $D < 0$

$$\max(r_\gamma \mid r_\varphi, D<0) = \left(\min\left[1 + \frac{D(1-\sqrt{r_\varphi})}{\min[p_g p_f ; (1-p_g)(1-p_f)]} ; 1 \right] \right)^2 \quad \text{(A5-3)}$$

$$\min(r_\varphi \mid r_\gamma, D<0) = \left(\max\left[1 - \frac{\min[p_g p_f ; (1-p_g)(1-p_f)](\sqrt{r_g}-1)}{D} ; 0 \right] \right)^2 \quad \text{(A5-4)}$$

When $D = \text{Max}(D) = \pm 0.25$,

$$\max(r_\gamma \mid r_\varphi, D = \pm 0.25) = r_\varphi \quad \text{(A5-5)}$$

3

Genetic analysis of infectious diseases: estimating gene effects in susceptibility and infectivity

Mahlet T. Anche[1,2], Piter B.[1], Mart C.M. de Jong[3]

[1] Animal Breeding and Genomics Centre, Wageningen University, 6700 AH, Wageningen, The Netherlands; [2] Quantitative Veterinary Epidemiology group, Wageningen University, 6700 AH, Wageningen, The Netherlands

Genet. Sel. Evol. (2015) 47:85

Abstract

Genetic selection of livestock against infectious diseases can complement existing interventions to control infectious diseases. Most genetic approaches that aim at reducing disease prevalence assume that individual disease status (infected/not-infected) is solely a function of its susceptibility to a particular pathogen. However, individual infectivity also affects the risk and prevalence of an infection in a population. Variation in susceptibility and infectivity between hosts affects transmission of an infection in the population, which is usually measured by the value of the basic reproduction ratio R_0. R_0 is an important epidemiological parameter that determines the risk and prevalence of infectious diseases. An individual's breeding value for R_0 is a function of its genes that influence both susceptibility and infectivity. Thus, to estimate the effects of genes on R_0, we need to estimate the effects of genes on individual susceptibility and infectivity. To that end, we developed a generalized linear model (GLM) to estimate relative effects of genes for susceptibility and infectivity. A simulation was performed to investigate bias and precision of the estimates, the effect of R_0, the size of the effects of genes for susceptibility and infectivity, and relatedness among group mates on bias and precision. We considered two bi-allelic loci that affect, respectively, the individuals' susceptibility only and individuals' infectivity only. A GLM with complementary log-log link function can be used to estimate the relative effects of genes on the individual's susceptibility and infectivity. The model was developed from an equation that describes the probability of an individual to become infected as a function of its own susceptibility genotype and infectivity genotypes of all its infected group mates. Results show that bias is smaller when R_0 ranges approximately from 1.8 to 3.1 and relatedness among group mates is higher. With larger effects, both absolute and relative standard deviations become clearly smaller, but the relative bias remains the same. We developed a GLM to estimate the relative effect of genes that affect individual susceptibility and infectivity. This model can be used in genome-wide association studies that aim at identifying genes that influence the prevalence of infectious diseases.

3.1 Background

New and existing infectious diseases represent a major and increasing threat to domestic plants and animals, and to humans. Infectious diseases of animals are a worldwide concern, particularly because of their effects on the productivity and welfare of livestock and also because of their zoonotic threats to human health. In spite of the availability of antibiotic and vaccine treatments, the undesirable environmental impact of antibiotic treatments, the rapid evolution of bacteria to develop resistance to antibiotics and of viruses to escape vaccine protection illustrate the need for additional control strategies that can provide a useful complement to the currently used interventions to control disease (Bishop et al., 2002).

Host susceptibility and tolerance are two of the ways that individuals respond to pathogens. Several studies on the genetics of diseases in animals have shown that the host's susceptibility and tolerance to infectious diseases have a genetic basis, and thus that genotypic differences exist between individuals regarding their susceptibility and tolerance to infectious challenges (Axford, 2000). A number of genome-wide association studies (GWAS) have reported single nucleotide polymorphisms (SNPs) associated with susceptibility to various infectious diseases (Kirkpatrick et al., 2011; Bermingham et al., 2014).

Most genetic approaches that aim at reducing the prevalence of an infection assume that an individual's disease status (infected/not-infected) is solely a function of its own genes and of non-genetic factors (Axford, 2000). Hence, these methods capture only the genetic variation in susceptibility or tolerance (strictly, this latter statement is restricted to the measurement of disease occurrence in groups of unrelated individuals (Anche et al., 2014)). However, the prevalence and dynamics of an infection depend also on the infectivity of infected individuals in the population. Moreover, accumulating evidence on the existence of "superspreaders" in the outbreaks of epidemics suggests that (phenotypic) variation in infectivity exists among hosts (Lloyd-Smith et al., 2005). Thus, the classical quantitative genetic approach of disease analysis based on individual disease status will capture only part of the heritable variation that is present in the host population and affects the dynamics of infectious diseases (Lipschutz-Powell et al., 2012).

Between-host variation in susceptibility and infectivity affects the transmission of an infection in the population. This effect is measured by the value of the basic reproduction ratio R_0. R_0 is defined as the average number of secondary cases produced by one typical infectious individual during its entire infectious lifetime, in an otherwise naïve population (Diekmann et al., 1990a). R_0

has a threshold value of 1, which implies that a major disease outbreak or a stable endemic equilibrium can only occur when R_0 is greater than 1. When R_0 is less than 1, the epidemic will die out. Thus, in order to reduce disease incidence and therewith prevalence, breeding strategies should aim at reducing R_0, preferably to a value less than 1.

Genetic improvement that aims at reducing R_0 should be based on individual breeding values for R_0. An individual's breeding value for R_0 is the sum of the average effects of its alleles on R_0 (Anche et al., 2014), which means that investigating the effects of genes on R_0 is relevant. Anche et al. (2014) showed that an individual's breeding value for R_0 is a function of its genotype for susceptibility and infectivity, and of the population's average susceptibility and infectivity. Thus, in order to estimate effects of genes on R_0, the susceptibility and infectivity effects of the different alleles must be estimated.

Disease data are often available only in binary form (0/1) that is, the value indicates whether an individual has become infected or not. Hence, methods for genetic analyses of disease traits have to be tailored to such data. Generalized linear models (GLM) are commonly used to analyse binary data, where the expected value of the binary response variable is linked to the explanatory variables (traits) by a linear equation after applying a link function (Velthuis et al., 2003a). Velthuis et al. (2003a) showed that the effect of susceptibility and infectivity of hosts on the transmission rate parameter β can be estimated by fitting a GLM with a complementary log-log link function to binary disease data. Lipschutz-Powell et al. (2014a) showed that a GLM with a complementary log-log link function can be used to link the probability of an individual to be infected to the susceptibility genotype of the individual itself and the infectivity genotypes of its infectious contacts. However, they observed that the infectivity component of the model was non-linear, and did not provide an explicit GLM or investigate the quality of estimates resulting from such a GLM.

In this study, we developed a GLM to estimate the relative effects of genes on individual susceptibility and infectivity, and investigated the quality of the resulting estimates in terms of bias and precision. We also investigated the effect of R_0, different sizes of the effects of susceptibility and infectivity genes and population structure with respect to relatedness on bias and precision of the estimates. The GLM was fitted to binary disease data (0/1) recorded at the end of the epidemic. Thus, the data analysed were counts of infected individuals of different genotypes. These data were obtained from a simulated genetically heterogeneous population in which individuals differed in susceptibility and infectivity.

3.2 Methods

3.2.1 Population structure

We assumed a diploid population with between-host genetic heterogeneity in susceptibility and infectivity. We modelled genetic heterogeneity in this population using two bi-allelic loci, one locus for the susceptibility effect (γ) with alleles G and g and susceptibility values γ_G and γ_g, and one locus for the infectivity effect (φ) with alleles F and f and infectivity values φ_F and φ_f, respectively. Both loci were assumed to have multiplicative allelic effects and the reason for this assumption is explained in the section "Generalized linear models".

3.2.2 Epidemiological model of disease dynamics

Disease dynamics that are caused by a microparasitic infection can be modelled with a basic compartmental stochastic Susceptible, Infected and Recovered (SIR) model. In this model, two possible events can occur: infection of a susceptible individual, and recovery of an infectious individual (Kermark and McKendrick, 1927). With stochasticity, these events occur randomly at a certain rate (probability per unit of time) specified by the model parameters and the state variables. In the SIR-model, these parameters are the transmission rate parameter (β) for $\mathrm{S} \rightarrow \mathrm{I}$ with rate $\beta \frac{SI}{N}$, and the recovery rate parameter (α) for $\mathrm{I} \rightarrow \mathrm{R}$ with rate αI, where *N* denotes population size, *S* the number of susceptible individuals and I the number of infectious individuals (in this study, we assumed that an individual will be infectious once it is infected, thus the terms infectious and infected will be used interchangeably; hence, the symbols S, I and R are used to denote both the disease status and the number of individuals with that disease status). The transmission rate parameter β describes the probability per unit of time for one infected individual to infect any other individual in a totally susceptible population (Diekmann et al., 1990a; Anderson et al., 1992) (this can be seen from the transmission rate $dS/dt = -\beta SI/N$, for I = 1 and S = N).

In the following, we will consider binary data at the end of an epidemic, which indicates for each individual whether it has become infected or not. Thus, binomial count data were available to quantify the occurrence of infected individuals according to genotype. As a step towards the GLM, first we derive the probability of an individual to become infected.

In a genetically heterogeneous population, the transmission rate parameter β varies between pairs of individuals, and in addition to the contact rate (c), it will depend on the infectivity genotype of the infectious individual, and on the susceptibility genotype of the recipient susceptible individual. The assumption

that the transmission rate depends only on the infectivity of the infectious individual and the susceptibility of the recipient individual, and not on the combination of these two traits, is known as separable mixing (Diekmann et al., 1990b). In other words, the two individuals that are in contact influence the transmission rate independently. Thus, the transmission rate of a specific susceptible individual with susceptibility genotype *i* from being susceptible to being infected when exposed to a single infectious individual with infectivity genotype *j* can be defined as:

$$\beta_{ij}\frac{1}{N} = \gamma_i \varphi_j c \frac{1}{N}, \qquad (1)$$

where γ_i denotes the susceptibility of the susceptible individual, and φ_j denotes the infectivity of the infectious individual. Note that the transmission rate in Eq. (1) refers to a single specific susceptible individual, whereas the transmission rate parameter β defined above, refers to any susceptible individual among the *N* candidates. Hence, they differ by a factor of *N*. In Eq. (1), *c* represents the average contact rate between any pair of individuals and thus c/N is the average contact rate of a susceptible with a single infectious individual in a group of size *N* (this assumes faecal-oral transmission or similar routes, where $1/N$ of the infectious material ends up with the sender itself). Any variation in contact rate among different types of susceptible and infectious individuals is included in γ_i and φ_j because of the assumption of separable mixing.

When one susceptible individual with susceptibility genotype i is exposed to one infectious individual with infectivity genotype *j*, the expected number of transmissions is the product of the transmission rate and the average length of the infectious period, and is equal to $\gamma_i \varphi_j c \frac{1}{N}\frac{1}{\alpha}$, where $1/\alpha$ is the average length of the infectious period. The probability P_{ij} that the individual escapes infection follows from the zero term of the Poisson distribution, and is equal to:

$$P_{ij} = e^{-\beta_{ij}\frac{1}{N}} = e^{-\gamma_i \varphi_j \frac{c}{\alpha}\frac{1}{N}}. \qquad (2a)$$

Here, it is assumed that the transmission rate parameter β (and thus also γ, φ, and c/α) is constant over time so that there is no over-dispersion and the Poisson distribution can be used.

At the end of the epidemic, the individual with susceptibility genotype i has been exposed not to only one but to all infectious group mates (strictly speaking this is true for the individuals escaping infection only). These group mates can be categorized by their infectivity genotype, *j*. Let I_j denote the number of infected

individuals with infectivity genotype *j* that have become infected during the epidemic and have infectivity φ_j. Then the probability P_i that the individual escapes all infection exposures by individuals of infectivity genotype *j* and still be susceptible by the end of the epidemic is equal to:

$$P_{i,I_j} = \prod_{I_j} e^{-\gamma_i \varphi_j \frac{c}{\alpha}\frac{1}{N}} = e^{-\gamma_i I_j \varphi_j \frac{c}{\alpha}\frac{1}{N}}. \quad (2b)$$

Thus, the probability P_i that the individual with susceptibility genotype i escapes all infection exposures from all genotypes and still be susceptible by the end of an epidemic is equal to the product of all the probabilities that it escapes infection exposures from its infectious group mates of each genotype:

$$P_i = \prod_{j=1}^{n} e^{-\gamma_i I_j \varphi_j \frac{c}{\alpha}\frac{1}{N}} = e^{-\gamma_i \frac{c}{\alpha}\frac{1}{N}\sum_{j=1}^{n} I_j \varphi_j}, \quad (3)$$

where the summation is over the *n* infectivity genotypes; *n* = 3 for a single bi-allelic locus in a diploid population.

In Eq. (3), we can replace I_j by $I * f_j$, where I is the total number of individuals that have been infected at the end of the epidemic and f_j is the fraction of infected individuals of genotype *j*. This yields:

$$P_i = e^{-\gamma_i \frac{c}{\alpha}\frac{I}{N}\sum_{j=1}^{n} f_j \varphi_j}. \quad (4)$$

From Eq. (4), the probability that a susceptible individual with susceptibility genotype i has been infected by the end of the epidemic is equal to:

$$1 - P_i = 1 - e^{-\gamma_i \frac{c}{\alpha}\frac{I}{N}\sum_{j=1}^{n} f_j \varphi_j}. \quad (5)$$

Thus, the probability that a susceptible individual has been infected depends on its own susceptibility, γ_i, and on the arithmetic mean infectiousness $\sum_{j=1}^{n} f_j \varphi_j$ of its I infectious group mates with different infectivity values φ_j, with *j* = 1,… *n*.

In Andreasen (2011), equation 10, which is equivalent to our Equation (5), was presented as the final size equation for a population that is heterogeneous for susceptibility and infectivity (in epidemiology, the so-called final size equation gives the fraction of infected individuals of each type by the end of an epidemic). Our Equation 5 and 14 in Lipschutz-Powell et al. (2014b) follow a similar derivation but, in our case, the equation is applied to the end of the epidemic.

3.2.3 Generalized linear model (GLM)

A GLM, in its simplest form, specifies a linear relationship between a function of the mean of the observed variable y, and a set of observed predictor variables, x:

$$\phi\big(E(y)\big) = c_0 + c_1 x_1 + \cdots c_n x_n,$$

where ϕ is the so-called link function, c_0 is the intercept and the c_i are the regression coefficients for the explanatory variables x_i, for *i* = 1, … *n*. The aim is to estimate c_i coefficients.

For binomial data where the probability of failure (to escape an infection) P is equal to the zero term of a Poisson distribution, as in the above Eq. (4), the complementary log-log link function is the default link function to connect explanatory variables x_i with the observed variable y of the linear model (McCullagh and Nelder, 1989). Applying the complementary log-log link function to $1 - P_i$ based on Eq. (4), yields:

$$\text{cloglog}(1 - \mathrm{P}_i) = \log\left(-\log(P_i)\right) = \log\left(\frac{c}{\alpha}\right) + \log(\gamma_i) + \log\left(\frac{I}{N}\right) + \log\sum_{j=1}^{n} f_j\varphi_j \quad (6)$$

Thus, the dependent variables have now become the fraction of each i type of individual that did become infected (see below).

The model in Eq. (6) is linear in log of susceptibility (γ_i) but not for infectivity (φ_j), since the logarithm of a sum does not equal the sum of the logarithms, as also observed by (Lipschutz-Powell et al., 2014b). In Eq. (6), the term $\sum_{j=1}^{n} f_j\varphi_j$ can be recognized as the arithmetic mean, since $\sum_{j=1}^{n} f_j$ = 1. In order to further linearize Eq. (6), the arithmetic mean was approximated by a geometric mean, using the substitution $\sum_{j=1}^{n} f_j\varphi_j \approx \prod_{j=1}^{n} {\varphi_j}^{f_j}$. This yields:

$$\log\left(-\log(P_i)\right) \approx \log\left(\frac{c}{\alpha}\right) + \log(\gamma_i) + \log\left(\frac{I}{N}\right) + \sum_{j=1}^{n} f_j \log(\varphi_j)\,. \quad (7)$$

The approximation of the arithmetic mean regression by a geometric mean regression was investigated separately as explained in the 'Appendix'.

In Eq. (7), the expectation of the response variable, $cloglog(1 - P_i)$ is a linear expression of $\log(\gamma_i)$ and $\log(\varphi_j)$.

Equation (7) is linear in the log of susceptibility (γ_i) and the log of infectivity (φ_j). To be able to formulate the model in terms of allele counts within individuals, rather than in terms of individual genotypes, it was assumed that the two alleles that make up the genotype within an individual act multiplicatively, so that their effects are additive on the log-scale.

Therefore, the genotypic values will be $\gamma_{GG} = \gamma_G \times \gamma_G = {\gamma_G}^2$, $\gamma_{gg} = \gamma_g \times \gamma_g = {\gamma_g}^2$ and $\gamma_{Gg} = \gamma_{gG} = \gamma_G \times \gamma_g$, for susceptibility, and $\varphi_{FF} = \varphi_F \times \varphi_F = \varphi_F^2$; $\varphi_{ff} = \varphi_f \times \varphi_f = \varphi_f^2$ and $\varphi_{fF} = \varphi_{Ff} = \varphi_f \times \varphi_F$ for infectivity. Furthermore, the effects of the g and f alleles were set to a value of 1, $\gamma_g = \varphi_f$ = 1, so that $\log(\gamma_g)$ = $\log(\varphi_f)$ = 0. This is done without loss of generality, because the interest lies in the

relative effect of one allele to the other, that is the effect of γ_G relative to γ_g and the effect of φ_F relative to φ_f [note that this does not affect the estimates of relative allele effects since the absolute scale of the model is accounted for by the log(c/α)-term]. Using Eq. (7), the GLM for the diploid genetic model becomes:

$$\text{cloglog}\, E\left[\frac{y_i}{n_i}\right] = c_0 + c_1 index_{G,i} + c_2 Num_F + log(\frac{I}{N}), \qquad (8)$$

where individuals are aggregated by their genotype, i. The cloglog is applied to the expectation of $\frac{y_i}{n_i}$, which is the fraction of infected individuals of genotype *i*, by the end of the epidemic and y_i follows a binomial distribution, c_0 is the intercept measuring log(c/α), and c_1 is the regression coefficient for the $index_G$, where $index_{G,i}$ = 0, 1 or 2 is the number of G alleles at the susceptibility locus of individuals of genotype i. The c_2 is the regression coefficient for Num_F, which is the average of the number of F-alleles per individual at the infectivity locus in the infected group mates of the individuals of genotype *i*. It is calculated as $2 * Frac_{FF} + 1 * Frac_{fF/Ff}$ where $Frac_{FF}$ is the fraction of infected individuals with genotype "FF" and $Frac_{fF/Ff}$ is the fraction of infected individuals with genotype "fF" or "Ff". The "2" arises because individuals with the "FF" genotype carry two F alleles, while those with the "fF" or "Ff" genotype carry only one F allele. The $\log(\frac{I}{N})$ corresponds to the total fraction of infected individuals in the group, which is used as an offset in the GLM. Hence, estimates of c_1 and c_2 refer to the effect of a single allele, and represent the so-called average effect of an allele substitution on the log-scale (Falconer and Mackay, 1996). When fitting the model to binomial count data of those individuals of each genotype that are infected and estimating c_0, c_1 and c_2, the effects of alleles G and F relative to $\gamma_g = \varphi_f = 1$ can be calculated as $\hat{\gamma}_G = e^{\widehat{c_1}}$ and $\hat{\varphi}_F = e^{\widehat{c_2}}$, respectively.

3.2.4 Simulation

To investigate the bias and precision of the $\hat{\gamma}_G$ and $\hat{\varphi}_F$, one generation of a diploid population was simulated based on the above assumptions with respect to the effect of alleles at both loci. These two loci were the only genetic effects simulated. Furthermore, it was assumed that allele frequencies at both loci were equal to 0.5, that is, $p_g = p_f = 0.5$. The population was sub-divided into 100 groups of 100 individuals each. Each group was set up in such a way that group mates showed a certain genetic relatedness, *r*, at both loci. Here, relatedness is defined as the correlation of allele counts between group mates, irrespective of what causes the correlation. To limit the number of scenarios to be tested, relatedness at the susceptibility locus, r_γ, and at the infectivity locus, r_φ, were assumed to be the

same (note that relatedness at both loci is expected to be the same when the loci are not under selection). In order to have a certain degree of relatedness among group mates, a fraction of fully related individuals was added to each group, supplemented by randomly selected individuals. Since each individual carries both the susceptibility and the infectivity locus, these additions were done jointly (see Appendix 4 in (Anche et al., 2014) for a detailed description of the strategy to make these additions jointly).

A basic stochastic SIR-model as described above was used to simulate the disease dynamics (Kermark and McKendrick, 1927). In each group, the epidemic began by one randomly infected individual. Then, the next event which could be either infection of a susceptible individual or recovery of infected individual was determined using Gillespie's direct algorithm (Gillespie, 1977). The type of event, i.e. either infection or recovery, was decided by drawing a random number v_1, from a uniform distribution, $v_1 \sim U(0,1)$. The next event was an infection of a susceptible individual if the random number $v_1 < \frac{\sum_i \sum_j \beta_{ij} \frac{S_i I_j}{N}}{\sum_i \sum_j \beta_{ij} \frac{S_i I_j}{N} + I\alpha}$, otherwise it was recovery of an infected individual. The numerator of this ratio represents the total infection rate, and the denominator the total rate, *i.e.*, the sum of the infection and recovery rates. The sampling of the specific individual that became infected depended on individual susceptibility. The probability that a susceptible individual of genotype *i* became infected was proportional to $\frac{\sum_i \sum_j \beta_{ij} \frac{S_i I_j}{N}}{\sum_i \sum_j \beta_{ij} \frac{S_i I_j}{N} + I\alpha}$. Hence, the transmission rates were updated based on the numbers of susceptible and infected individuals of each genotype, while the transmission rate parameter β_{ij} remained constant. The epidemic ended when there was no more infectious individual in the population or when there was no susceptible individual left to be infected. By the end of the epidemic, the number of individuals that got infected together with their genotypes for susceptibility and infectivity were recorded. The fraction of individuals of each genotype that got infected was the dependent variable in the analysis.

We hypothesized that different epidemiological and genetic factors will affect the quality of the estimates, as measured by the bias and precision of $\hat{\gamma}_G$ and $\hat{\varphi}_F$. For that purpose, we simulated different scenarios that are described below. The biases of the estimates were calculated by taking the difference between the 'true' and estimated values and the precision of the estimates were calculated using the standard deviation (SD) of the estimates.

First, we simulated a basic scenario (scenario 1; Table 3.1), in which groups were created randomly with respect to relatedness among group mates. We calculated R_0 using (Anche et al., 2014):

$R_0 = \bar{\gamma}\bar{\varphi}\, c/\alpha$,

where $\bar{\gamma} = p_g^2\gamma_{gg} + 2p_g(1-p_g)\gamma_{gG} + (1-p_g)^2\gamma_{GG}$,

and $\bar{\varphi} = p_f^2\varphi_{ff} + 2p_f(1-p_f)\varphi_{fF} + (1-p_f)^2\varphi_{FF}$.

Population parameters are in Table 3.1. In the basic scenario, R_0 was set to 1.2.

Table 3.1. Simulated scenarios

Parameters	Scenario 1	Scenario 2	Scenario 3
Contact rate, *c*	1.5	0.75-7.5	1.5
Recovery rate α	0.5	0.5	0.5
γ_G	0.6	0.6	0.97, 0.6 and 0.37
φ_F	0.6	0.6	0.3, 0.6 and 0.9
Relatedness *r*	0-1	0-1	0-1
R_0	1.2	0.6-6.1	1.2

For all scenarios, $\gamma_g = \varphi_f = 1$ and $p_g = p_f = 0.5$

Second, to investigate the effect of R_0 on the quality of $\hat{\gamma}_G$ and $\hat{\varphi}_F$, we simulated scenarios with different values of R_0. We varied the contact rate *c*, so that R_0 for a population consisting of groups with unrelated individuals varied from 0.6 (for which no major outbreaks can occur) to 6.1 (for which major outbreaks can occur; Table 3.1, scenario 2).

Third, to investigate the impact of the size of effects of the genes for susceptibility and infectivity on the quality of $\hat{\gamma}_G$ and $\hat{\varphi}_F$, we simulated scenarios with different effect sizes for a constant value of R_0 = 1.2. We simulated all combinations of low, moderate and high values for γ_G and φ_F (Table 3.1, scenario 3).

Furthermore, in all of the above-mentioned scenarios, relatedness between group mates was varied between 0 and 1 to investigate the effect of population structure with respect to relatedness on the quality of $\hat{\gamma}_G$ and $\hat{\varphi}_F$. Relatedness was assumed to be the same at both loci (see (Anche et al., 2014) for details). We used R software to fit the model with a *glm* function and a binomial distribution.

3.3 Results

All estimates presented in this section are averages from 2000 replicates, except for Figure 3.1 which shows the results of all replicates. The black straight line in all Figures represents the true difference between γ_g and γ_G and between φ_f and φ_F, and the bars indicate the standard deviation of these estimates among replicates.

In the basic scenario, in which groups were created randomly with respect to relatedness, *r* = 0, we found that the susceptibility effect was slightly underestimated ($1 - \hat{\gamma}_G$ in Figure 3.2) but the infectivity effect was considerably overestimated ($1 - \hat{\varphi}_F$ in Figure 3.2). When the degree of relatedness among group mates increased, the bias of both estimates decreased, however, the effect of relatedness was more pronounced for infectivity (Figure 3.2). The error in $\hat{\varphi}_F$, that is caused by the geometric mean approximation was quantified and found to be small (Table 3.3, Appendix 1). Moreover, the standard deviation of the estimated susceptibility effect increased only slightly, whereas the standard deviation of the estimated infectivity effect increased considerably as the degree of relatedness increased.

A scatter plot for $(1 - \hat{\gamma}_G)$ and $(1 - \hat{\varphi}_F)$ of the 2000 replicates for the basic scenario where *r* = 0 shows that the estimated differences are uniformly distributed over their range without any pattern (Figure 3.1).

This plot also shows that $(1 - \hat{\varphi}_F)$ is more often underestimated than overestimated, which agrees with the underestimation in Figure 3.2 for *r* = 0.

In the second set of scenarios, where R_0 was varied from 0.6 to 6.1, susceptibility and infectivity effects were also underestimated. Bias in $\hat{\gamma}_G$ and $\hat{\varphi}_F$ was smallest for values of R_0 that ranged approximately from 1.8 to 3.1. Higher values of R_0 increased bias in $\hat{\gamma}_G$ but had little effect on bias in $\hat{\varphi}_F$ when group mates were unrelated (Figure 3.3, panel a). Bias in $\hat{\varphi}_F$ and $\hat{\gamma}_G$ decreased with increasing relatedness among group mates, except for $\hat{\varphi}_F$ at high values of R_0 (Figure 3.3, panels b, c and d). In contrast to the result for the unrelated groups, bias in $\hat{\varphi}_F$ was larger at high values of R_0 when related groups were used (Figure 3.3, panel a *vs.* panels b, c and d). For fully-related groups, i.e. *r* = 1, estimates for $\hat{\varphi}_F$ and $\hat{\gamma}_G$ and their standard deviation were nearly identical (Figure 3.3, panel d). For this scenario, the error in $\hat{\varphi}_F$ as a result of the geometric mean approximation was also quantified and only a small error was found (Appendix, Table 3.4).

For all values of R_0, standard deviations of estimates were greater for infectivity effect than for susceptibility effect, except for *r* = 1 for which they were nearly identical. Standard deviations decreased considerably as relatedness among group mates increased, particularly for infectivity effect. For both susceptibility and

infectivity effects, standard deviations were smaller for values of R_0 for which the bias in $\hat{\gamma}_G$ and $\hat{\varphi}_F$ was smallest, i.e. when R_0 ranged approximately from 1.8 to 3.1.

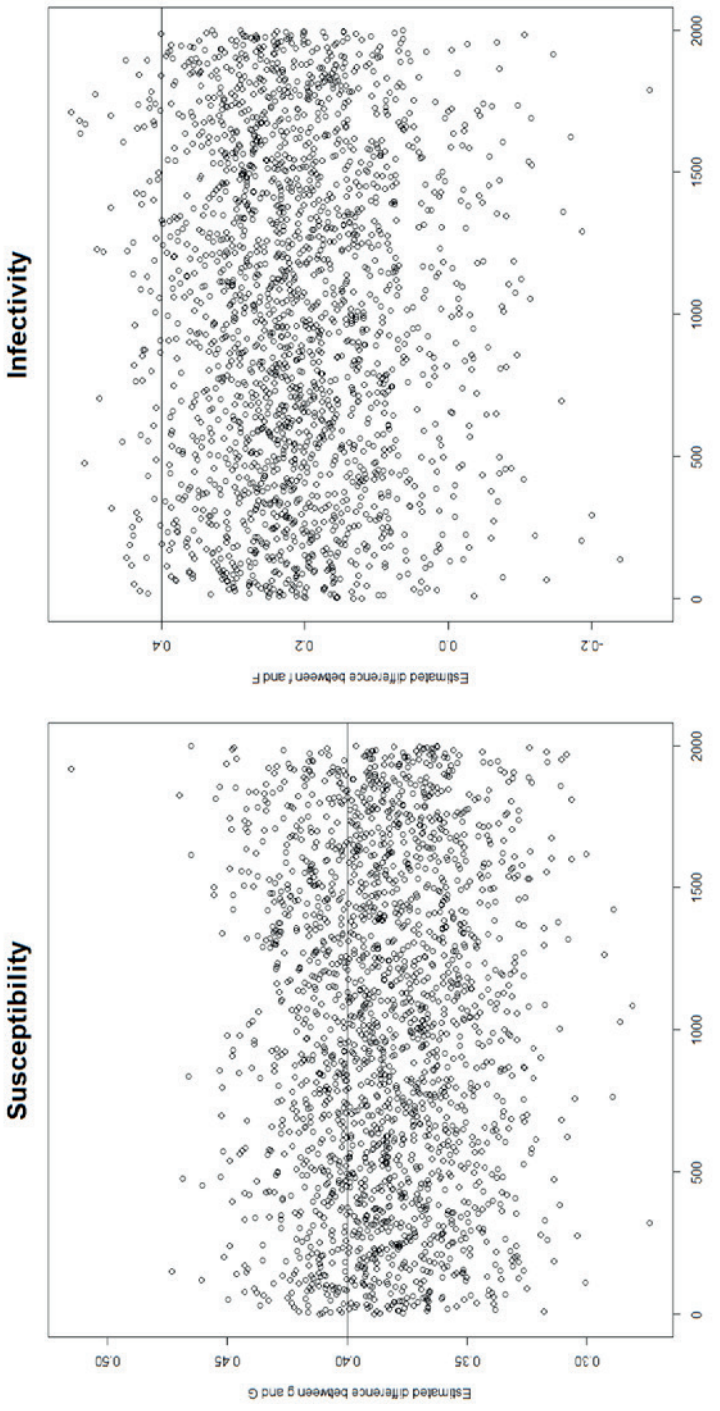

Figure 3.1. Scatter plots for $(1 - \hat{\gamma}_G)$ and infectivity $(1 - \hat{\varphi}_F)$. For the scenario where relatedness between group mates $r = 0$ and $R_0 = 1$.

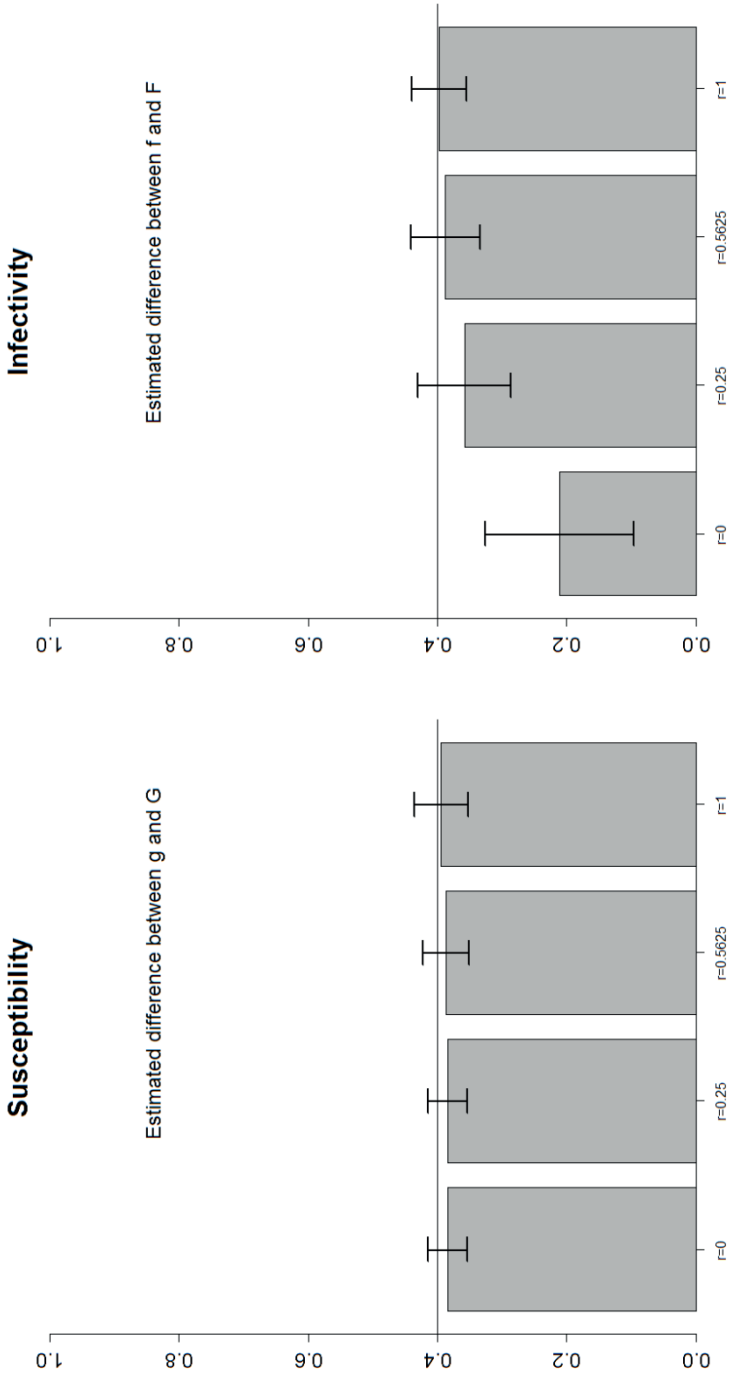

Figure 3.2. Difference between γ_g and $\hat{\gamma}_G$ and between φ_f and $\hat{\varphi}_F$. For the scenario with different values of relatedness *r* between group mates, and $\gamma_g = \varphi_f = 1$.

In the third set of scenarios, different sizes of the effects of γ_G and φ_F were simulated. For both estimates, the relative bias did not change regardless of the size of the effect considered (Figures 3.4 and 3.5). In these scenarios also, both susceptibility and infectivity effects were underestimated regardless of the size of the effects considered, except when there was a large difference in infectivity effect and $r = 1$, there was a small overestimation ($1 - \hat{\varphi}_F$ in Figure 3.5). Moreover, smaller relative standard deviations were found for both susceptibility and infectivity effects when effect sizes were larger, which indicates that the effects are better estimated when they are larger. For this scenario, the error in $\hat{\varphi}_F$ as a result of the geometric mean approximation was also quantified and only a small error was found (Table 3.5, Appendix).

3.4 Discussion

In this work, a generalized linear model with a complementary log-log link function was developed to estimate the relative effects of genes on individual susceptibility and infectivity. The model was developed from an equation that describes the probability of an individual to become infected as a function of its own susceptibility genotype and of the infectivity genotypes of its infected group mates. This GLM was developed following Velthuis et al. (2003b) who developed a GLM for binary data on a transmission trial to estimate the effect of susceptibility and infectivity of hosts on the transmission rate parameter β. A simulation study was performed to investigate the quality of the GLM. From the statistical analysis of the simulated data, we obtained fairly precise estimates, except for some scenarios for which estimates were more biased, particularly for infectivity. The best estimates were found for schemes with intermediate R_0 and related group members. For all the scenarios investigated, the sizes of the effects at both loci were underestimated.

The main objective of this study was to develop a methodology to estimate gene effects and also to investigate its quality in terms of bias and precision of the estimates. To test the methodology without introducing additional assumptions that may contribute to estimation error, we assumed additive allele effects on the log-scale for both susceptibility and infectivity. Thus, allelic effects were simulated multiplicatively on the original scale. This was done for two reasons. First, we wanted to formulate the model in terms of allele counts within individuals, rather than in terms of individual genotypes. In other words, we did not intend to estimate dominance effects. Whether allele effects are more likely to be additive on the log-scale than on the original scale is unknown at present. Second, since the objective of this study was to investigate the quality of the model rather

than the assumptions on the genetic architecture, the data were simulated under a model that agreed with the assumptions of the statistical model. Bias and standard deviation of the estimates were smallest for R_0 that ranged approximately from 1.8 to 3.1.

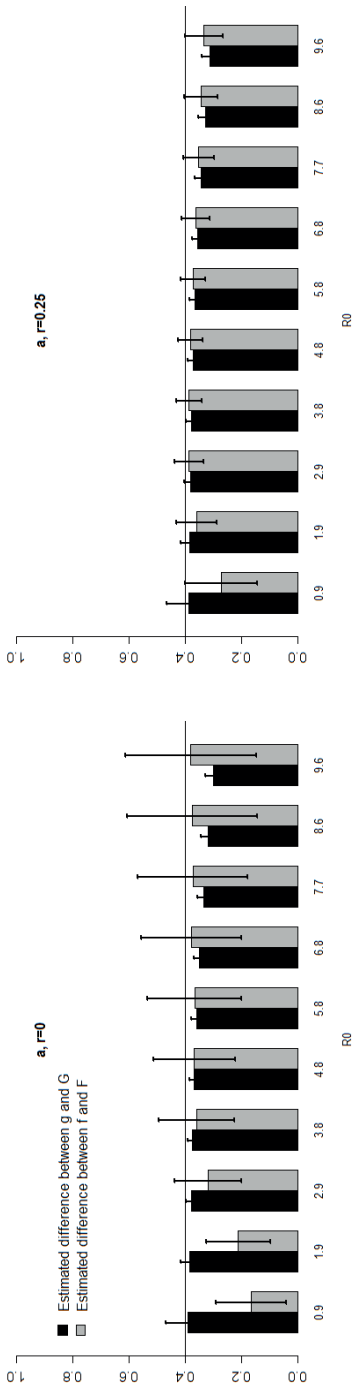

Figure 3.3. Difference between γ_g and $\hat{\gamma}_G$, and between φ_f and $\hat{\varphi}_F$. For the scenario with different values of R_0 and degrees of relatedness *r* between group mates. ($\gamma_g = \varphi_f = 1$).

The basic reproduction ratio R_0 is an important factor that affects the size of an epidemic in a population, i.e. the fraction of individuals that are found to be infected by the end of an epidemic. When R_0 is greater than 1 but near 1 in a group, there will be virtually no individuals infected and thus, there is hardly any variation in disease status, which results in inaccurate estimates of gene effects. Conversely, when R_0 is much greater than 1, nearly all individuals will be infected, which again results in very little variation in disease status. (Table 3.2 indicates the fraction of infected individual for different values of R_0 and relatedness among group mates). Thus, the relationship between R_0 and the fraction of individuals infected affects the estimation of the effect on susceptibility and infectivity, since data on the final size of an epidemic were used for our estimation. This mechanism may explain why the estimated effect on susceptibility is best for intermediate R_0. The effect of infectivity is more difficult to estimate and the bias is larger.

For each scenario, more relatedness between individuals resulted in better estimates for both traits. This is because more relatedness creates more variation between groups, which results in groups with below or above average susceptibility and/or infectivity. This occurs because an individual with a lower susceptibility will also have related group mates with below average susceptibility, and *vice versa*. The same applies for infectivity. However, since we assumed absence of linkage disequilibrium (LD) between the susceptibility and infectivity loci, groups with below average susceptibility will not always have below average infectivity as well. Thus, only those groups with above average susceptibility and above average infectivity will have epidemics with a greater final size, i.e. the fraction of individuals that gets infected by the end of the epidemic, while those with below average susceptibility and infectivity will a lower final size. This variation improved estimates of the effects of susceptibility and infectivity.

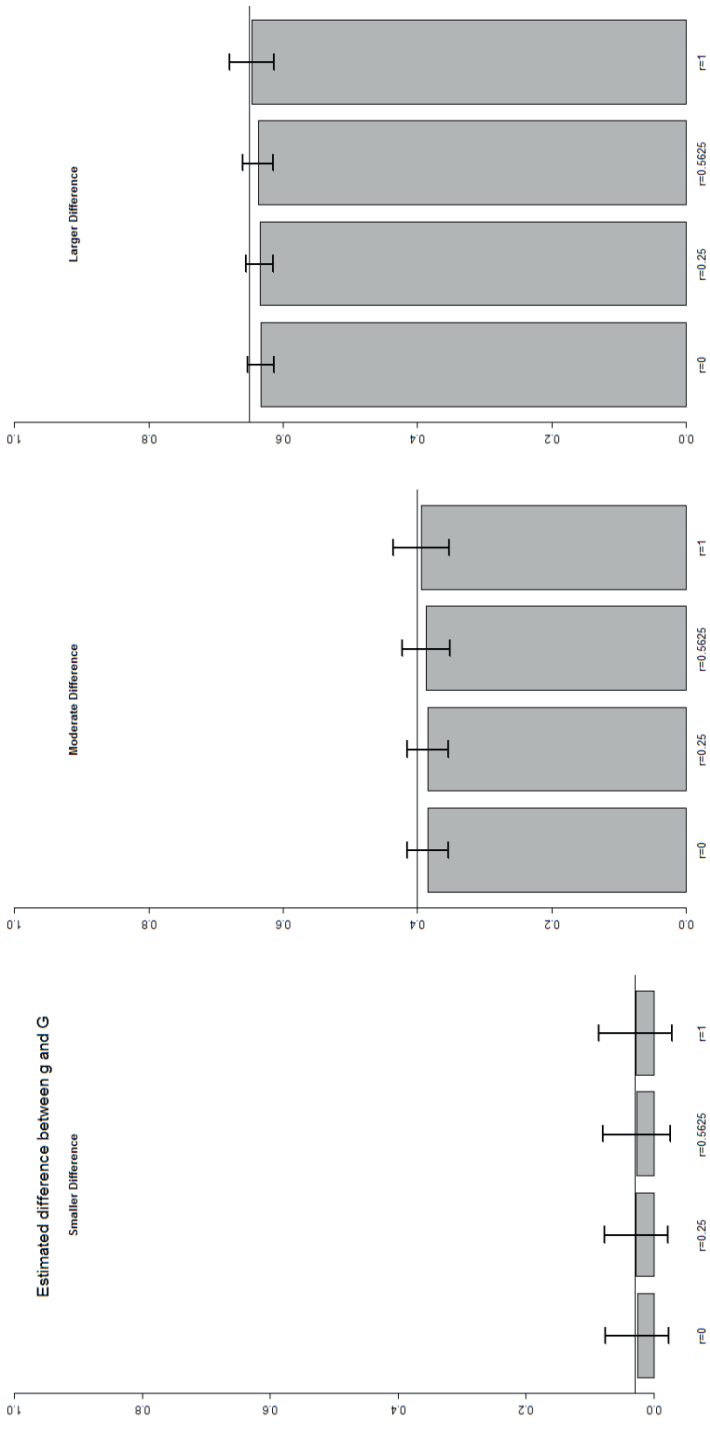

Figure 3.4. Difference between γ_g and $\hat{\gamma}_G$. For the scenario with different values of the (true) difference between γ_g and γ_G, and for different values of relatedness *r* between group mates, and $R_0 = 1.2$. ($\gamma_g = \varphi_f = 1$).

We have made a number of assumptions in building our methodology. In the derivation of Equation (5), we assumed that all individuals that escaped the infection had been exposed to all infected individuals. Of course, this assumption is true for the simulations done here. To what extent, this will be true for real data remains to be seen. It seems reasonable to assume that individuals in relatively small and well-defined groups get mixed up over space and time as is often the case in animal husbandry: for example, in fattening pigs with group sizes of 10 to 30 individuals. The assumption is less reasonable for groups with a spatial structure, for example in tie stalls or when epidemics occur within barns subdivided into multiple groups. In such cases, data should be collected separately for different groups. We also assumed that epidemics could be completely recorded, so that the final disease status of all individuals is known, and all individuals that have escaped the infection have been exposed to all infected individuals. However, for reasons of, *e.g.*, animal welfare and productivity, interventions are often carried out to limit the size of an epidemic. Hence, individuals may not have had the full potential to express their susceptibility and infectivity. For incomplete epidemics, the probability that an individual becomes infected follows from Equation (5) when only the infected individuals to which the focal individual has been exposed are considered (see also (Lipschutz-Powell et al., 2014b)). Thus, extension to incompletely observed epidemics is straightforward (see also application in (Velthuis et al., 2003b) and subsequent papers citing (Velthuis et al., 2003b)).

Bias and precision of estimates may be improved when data are recorded within shorter time intervals. This may be particularly helpful for cases with high R_0. In such cases, each interval forms an incompletely observed epidemic, which can be analysed with the same GLM statistical approach (Velthuis et al., 2003b). When data are collected in sufficiently short time intervals, only a fraction of individuals will become infected in a single interval, even when R_0 is high. This will contribute to accuracy of the estimates. Moreover, collecting data in short time intervals also provide information on the order of infections, *i.e.*, which animal has infected which animal. This will increase the accuracy of estimated gene effects, particularly for infectivity (Pooley et al., 2014). Thus, using data from short time intervals can be complementary to using groups composed of related individuals and data from multiple epidemics. The derivation and resulting model for such cases is very similar to the one presented here, since the probability that an individual escapes infection follows from the zero-term of the Poisson distribution (see also [11, 9]). The key step is to identify the infectious individuals to which the focal individual has been exposed in a time period.

Lipschutz-Powell et al. (2014b) showed that, when there is genetic variation in susceptibility only, a complementary log-log link function can be used to link an equation that describes the probability of an individual to become infected to a linear model that includes the individual's genotype for susceptibility. They also suggested that, when there is genetic variation in infectivity, a Taylor-series expansion of the model term for infectivity can be used to further linearize the model in infectivity. In our study, we obtained a linear model for infectivity by approximating the arithmetic mean by a geometric mean. We quantified the error due to this approximation and found only negligible errors in the estimates (Appendix). Thus, this approximation can be ruled out as the cause of the observed bias. This suggests that, for cases for which there is variation in infectivity, the geometric mean approximation is suitable to obtain a linear combination of the parameters of interest. A full investigation of the causes of the bias is beyond the scope of this study. However, the fact that a population of finite size, *i.e.*, 100 individuals in each group, was used to estimate gene effects can be one of the reasons for the observed underestimation.

Table 3.2. Fraction of individuals infected at the end of the epidemic

	$r = 0$	$r = 0.25$	$r = 0.5625$	$r = 1$
R_0=0.6	0.02	0.03	0.03	0.04
R_0=1.2	0.10	0.12	0.14	0.16
R_0=1.8	0.30	0.30	0.30	0.30
R_0=2.5	0.46	0.45	0.44	0.43
R_0=3.1	0.58	0.57	0.55	0.53
R_0=3.7	0.66	0.65	0.63	0.61
R_0=4.3	0.71	0.70	0.69	0.67
R_0=4.9	0.75	0.75	0.73	0.71
R_0=5.5	0.79	0.78	0.77	0.75
R_0=6.1	0.81	0.80	0.80	0.78

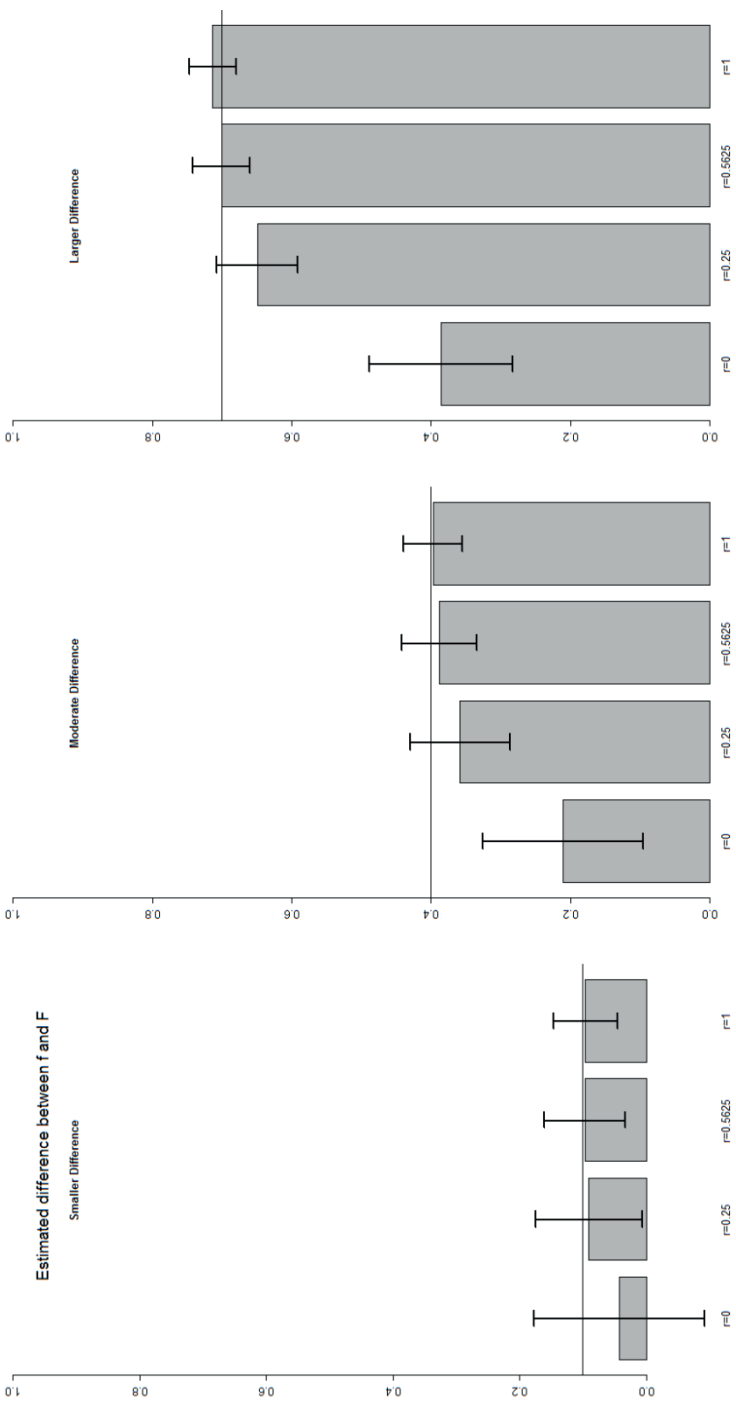

Figure 3.5. Difference between φ_f and $\hat{\varphi}_F$. For the scenario with different values of the (true) difference between φ_f and φ_F, and for different values of relatedness r between group mates and $R_0 = 1.2$. ($\gamma_g = \varphi_f = 1$).

Anche et al. (2014) defined breeding value and heritable variation in R_0. They showed that an individual's breeding value for R_0 is a function of the population's average susceptibility and infectivity, of the gene frequencies within the individual and of average effects of the alleles at both loci (Equation 7c in (Anche et al., 2014)). However, Anche et al. (2014) assumed that effects of alleles at both loci were additive, whereas here we assumed that effects are multiplicative (so that they are additive on the log scale). Multiplicative effects introduce dominance. Hence, before applying the expressions for breeding value and heritable variation of (Anche et al., 2014) to estimates obtained from the methods proposed here, they need to be translated into average effects of alleles (Falconer and Mackay, 1996). Using the common notation for the one-locus model (Falconer and Mackay, 1996), the additive effect is half the difference in genotypic value between both homogyzotes, $a_\gamma = \left(\gamma_g^2 - \gamma_G^2\right)/2$ and $a_\varphi = \left(\varphi_f^2 - \gamma_F^2\right)/2$, the dominance deviation is the difference between the heterozygote and the average of both homozygotes, $d_\gamma = \gamma_g\gamma_G - \left(\gamma_g^2 + \gamma_G^2\right)/2$ and $d_\varphi = \varphi_f\varphi_F - \left(\varphi_f^2 + \varphi_F^2\right)/2$, and the average effects of alleles are given by $\alpha_\gamma = a_\gamma + \left(p_G - p_g\right)d_\gamma$ and $\alpha_\varphi = a_\varphi + \left(p_F - p_f\right)d_\varphi$, where p denotes allele frequency (Falconer and Mackay, 1996). Hence, in Equations 7 and 11 of (Anche et al., 2014), $\gamma_g - \gamma_G$ should be replaced by α_γ, and $\varphi_f - \varphi_F$ should be replaced by α_φ. For example, for $\gamma_g = 1$ and $\gamma_G = 0.6$, genotypic values are $\gamma_{gg} = 1$, $\gamma_{gG} = 0.6$ and $\gamma_{GG} = 0.36$, the additive effect is $a_\gamma = (1 - 0.36)/2 = 0.32$, the dominance deviation is $d_\gamma = 0.6 - (1 + 0.36)/2 = -0.08$, and the average effect is $\alpha_\gamma = 0.32 - 0.08(p_G - p_g)$.

In this study, we assumed a model with two bi-allelic loci, i.e. one locus that affects individual susceptibility and one locus that affects individual infectivity. Furthermore, we assumed that which locus affects infectivity and which locus affects susceptibility, are known. This may be the case with candidate gene approaches which include only the genes for which the function is related to the trait of interest. The effect of the putative causative gene is then examined by association study. In such studies, the GLM developed here can be applied to estimate and confirm the effect of the candidate gene on the trait of interest. However, applying a candidate gene approach is limited because it relies on knowing the functional relation between the genes and the trait of interest. The recent advances in molecular genomics allow us to genotype individuals for thousands of SNPs, and to perform GWAS in which all SNPs are examined for their association with the trait of interest. The GLM developed here can also be used in

GWAS that aim at identifying genes associated with susceptibility and/or infectivity. In such studies, it is not known whether a SNP affects infectivity and/or susceptibility. Hence, this has to be inferred from the significance of the estimated effects. To avoid the need to test all combinations of two SNPs, one could first screen SNPs for susceptibility effects, and then fit only the significant loci for susceptibility effects, together with all other loci for infectivity effects. Moreover, when modified so that gene effects are estimated as random effects, our model can probably be used for polygenic traits, for example in genomic prediction, for which effects of all genes are estimated simultaneously and the interest lies in predicting the breeding value of entire genotypes (Meuwissen et al., 2001).

3.5 Conclusions

We have developed a generalized linear model to estimate the relative effects of genes on individual susceptibility and infectivity. This model may be used in genome-wide association studies that aim at identifying genes that are involved in the prevalence of infectious diseases.

3.6 Acknowledgements

This study was financially supported by the Marie Curie Nematode Health project. The contribution of PB was supported by the foundation for applied sciences (STW) of the Dutch science council (NWO).

3.7 References

Anche, M., M. de Jong, and P. Bijma. 2014. On the definition and utilization of heritable variation among hosts in reproduction ratio R0 for infectious diseases. Heredity.

Anderson, R. M., R. M. May, and B. Anderson. 1992. Infectious diseases of humans: dynamics and control. Wiley Online Library.

Andreasen, V. 2011. The final size of an epidemic and its relation to the basic reproduction number. Bulletin of mathematical biology 73: 2305-2321.

Axford, R. F. E. 2000. Breeding for disease resistance in farm animals. 2nd ed. CABI Pub., Wallingford, Oxon, UK ; New York.

Bermingham, M. L. et al. 2014. Genome-wide association study identifies novel loci associated with resistance to bovine tuberculosis. 112: 543-551.

Bishop, S. C., M. C. M. d. Jong, and D. G. 2002. Commison on genetic resources for food and agriculture.

Diekmann, O., J. Heesterbeek, and J. A. Metz. 1990a. On the definition and the computation of the basic reproduction ratio R 0 in models for infectious diseases in heterogeneous populations. Journal of mathematical biology 28: 365-382.

Diekmann, O., J. A. P. Heesterbeek, and J. A. J. Metz. 1990b. On the Definition and the Computation of the Basic Reproduction Ratio R0 in Models for Infectious-Diseases in Heterogeneous Populations. J Math Biol 28: 365-382.

Falconer, D., and T. Mackay. C. 1996. Introduction to quantitative genetics 4.

Gillespie, D. T. 1977. Exact stochastic simulation of coupled chemical reactions. The journal of physical chemistry 81: 2340-2361.

Kermark, M., and A. McKendrick. 1927. Contributions to the mathematical theory of epidemics. Part I. In: Proc. R. Soc. A. p 700-721.

Kirkpatrick, B. W., X. Shi, G. E. Shook, and M. T. Collins. 2011. Whole-Genome association analysis of susceptibility to paratuberculosis in Holstein cattle. Animal Genetics 42: 149-160.

Lipschutz-Powell, D., J. A. Woolliams, P. Bijma, and A. B. Doeschl-Wilson. 2012. Indirect Genetic Effects and the Spread of Infectious Disease: Are We Capturing the Full Heritable Variation Underlying Disease Prevalence? Plos One 7: e39551.

Lipschutz-Powell, D., J. A. Woolliams, and A. B. Doeschl-Wilson. 2014. A unifying theory for genetic epidemiological analysis of binary disease data. Genet Sel Evol 46: 1-12.

Lloyd-Smith, J. O., S. J. Schreiber, P. E. Kopp, and W. M. Getz. 2005. Superspreading and the effect of individual variation on disease emergence. Nature 438: 355-359.

McCullagh, P., and J. A. Nelder. 1989. Generalized linear models.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819-1829.

Pooley, C., S. Bishop, and G. Marion. 2014. Estimation of single locus effects on susceptibility, infectivity and recovery rates in an epidemic using temporal data. Proceedings, 10th World Congress of Genetics Applied to Livestock Production.

Velthuis, A., M. De Jong, E. Kamp, N. Stockhofe, and J. Verheijden. 2003. Design and analysis of an< i> Actinobacillus pleuropneumoniae</i> transmission experiment. Preventive veterinary medicine 60: 53-68.

Appendix

Geometric mean versus arithmetic mean in the estimation of gene effects on infectivity

In this Appendix, we address one issue regarding the quality of the two estimators, which we use to recover the genetic parameters. In general, one would like to have estimators that give consistent estimates of the parameters. This implies that both the variance and the bias of the estimators for a sufficiently large dataset (size n) can be brought arbitrarily close to zero. Expressed in formulas for the relative infectivity and the relative susceptibility, which are the two parameters that we want to estimate, these requirements look like this:

$$\lim_{n\to\infty}\left(\widehat{\frac{\varphi_F}{\varphi_f}}\right)-\left(\frac{\varphi_F}{\varphi_f}\right)=0,$$

$$\lim_{n\to\infty}\left(\widehat{\frac{\gamma_G}{\gamma_g}}\right)-\left(\frac{\gamma_G}{\gamma_g}\right)=0,$$

$$\lim_{n\to\infty}\operatorname{var}\left(\widehat{\frac{\varphi_F}{\varphi_f}}\right)=0,$$

$$\lim_{n\to\infty}\operatorname{var}\left(\widehat{\frac{\gamma_G}{\gamma_g}}\right)=0.$$

In addition, one would like to know how fast the estimators approach these limits. That analysis is presented in the main text and is done by comparing simulations to the true values. There is, however, an issue with the asymptotic unbiasedness of the effect on infectivity (the first equation): the estimator for the effect of the relative infectivity is not unbiased, but instead we will show below that:

$$\lim_{n\to\infty}\frac{\widehat{Log\left(\frac{\varphi_F}{\varphi_f}\right)}}{Log\left(\frac{\varphi_F}{\varphi_f}\right)}=m\left(\frac{\varphi_F}{\varphi_f}\right),$$

and we will derive the expression for the function m(.) and will show that it is close to 1 and always smaller or equal to 1. Note that m(.)=1 means no bias and m(.)<1 means underestimation of the effect.

As explained in the main text, the transmission rate parameter (β) is the product of the contact rate (c), susceptibility (γ) and infectivity (φ). Applying the complementary log-log link function results in Log(β) being in the expression for the expected value of the dependent variable. Thus, to see whether a linear relation is obtained between the explanatory variables to explain the expected value of the dependent variable, we can write that:

$$\mathrm{Log}(\beta)=\mathrm{Log}(c)+\mathrm{Log}(\gamma)+\mathrm{Log}(\varphi).$$

The heterogeneity in Log(γ) is straightforwardly incorporated in the model since each recipient counted in the dependent variable is only one type of susceptible individual. Thus, take $\gamma_g = 1$ and the other type has γ_G, then:

$Log(\beta) = Log(c) + Index_G\ Log(\gamma_G) + Log(\varphi)$,

where $Index_G$ is equal to 1 if the recipient is G and 0 when the recipient is of type g, with additional modification for the three genotypes as explained in the main text. Thus, the estimated parameter is asymptotically unbiased using the GLM method.

For heterogeneity in φ, it is not straightforward because we are dealing with the arithmetic mean (φ_{AM}) across all types of infectious individuals in the populations as was derived in the main text. Let us again look at the case with only two types of infectious individuals:

$Log(\beta) = Log(c) + Index_G\ Log(\gamma_G) + Log(\varphi_F p_F + \varphi_f p_f)$,

where p_F is the explanatory variable ($p_f = 1 - p_F$).

In order to obtain linearity in the explanatory variable for infectivity, φ_{AM} is replaced by geometric mean φ_{GM} with $\varphi_{GM} = \prod_{j=1}^{n} {\varphi_j}^{p_j}$. The equation with two types of infectious individuals becomes:

$Log(\beta) = Log(c) + Index_G\ Log(\gamma_G) + p_F Log(\varphi_F) + p_f Log(\varphi_f)$.

This is a linear equation in p_F, the explanatory variable, because $p_f = 1 - p_F$.

Now, we calculate the systematic error (bias) made by the approximation of the arithmetic mean (φ_{AM}) by a geometric mean (φ_{GM}). For a bi-allelic genetic model, where there are two alleles, i.e. φ_F and φ_f, with a frequency p_F and $(1 - p_F)$, respectively, the $\mathrm{Log}(\varphi_{AM})$ expression for the two alleles can be written as:

$\mathrm{Log}(\varphi_{AM}) = \mathrm{Log}(p_F\varphi_F + (1 - p_F)\varphi_f)$,

$\mathrm{Log}(\varphi_{AM}) = \mathrm{Log}((\varphi_F - \varphi_f)p_F + \varphi_f)$. (A1)

Thus, the effect of φ_F compared to φ_f is measured by the coefficient of p_F, *i.e.,* the slope of the linear expression within the logarithm, but this is not a linear model. Note that if the number of data points (n) becomes larger and larger, the expected (average) observed values, i.e. the number of cases (y) among the number of susceptible (S), will after applying the cloglog link function become arbitrary close to the expression A1, or:

$\lim_{n\to\infty} cloglog(E\frac{y}{S}) = C_0 + \mathrm{Log}((\varphi_F - \varphi_f)p_F + \varphi_f)$.

To obtain a linear model, we take the $\mathrm{Log}(\varphi_{GM})$ expression for the two alleles which can be written as:

$\mathrm{Log}(\varphi_{GM}) = \mathrm{Log}(\varphi_F^{p_F}\varphi_f^{1-p_F})$,

$\mathrm{Log}(\varphi_{GM}) = p_F\,\mathrm{Log}(\varphi_F) + (1 - p_F)\,\mathrm{Log}(\varphi_f)$,

$$\mathrm{Log}(\varphi_{GM}) = p_F \,\mathrm{Log}\left(\frac{\varphi_F}{\varphi_f}\right) + \mathrm{Log}(\varphi_f). \qquad \text{(A2)}$$

Now the effect of the allele φ_F compared to φ_f is measured by the ratio of the two values instead of the difference as in expression A1. This ratio can thus be calculated as the antilog of the regression coefficient of p_F which is the explanatory variable. In other words, from the GLM in Equation (8) in this paper, the estimated Log of the ratio of φ_F over φ_f is obtained from the regression coefficient c_2.
Now the next issue that we address in this Appendix is to fit a linear equation for the $\mathrm{Log}(\varphi)$ as a function of allele frequency (p_F) which is:

$$\log(\varphi_{LIN}) = A \cdot p_F + B. \qquad \text{(A3)}$$

If, in fact, the transmission depended on the φ_{GM} rather than on φ_{AM}, we would have:

$A = \log\left(\frac{\varphi_F}{\varphi_f}\right)$ and $B = \log(\varphi_f)$.

However, since the data come from a process where the observed φ is in fact the $\varphi_{AM,}$ the resulting linear relationship will not (necessarily) have $A = \log\left(\frac{\varphi_F}{\varphi_f}\right)$.

As we are interested in the allele effects, we need to estimate the slope of the line, i.e. the regression coefficient (A) of p_F, and compare it to $\log\left(\frac{\varphi_F}{\varphi_f}\right)$. To determine A, we need to find the best fitting linear relationship for $\mathrm{Log}(\varphi_{LIN})$ from $\mathrm{Log}(\varphi_{AM})$ data (Figure A1). This was done and we showed that this estimated A is very close to $\mathrm{Log}\left(\frac{\varphi_F}{\varphi_f}\right)$, and we were able give an explicit expression for the bias with respect to this true value.

Derivation of the expression for fitted line through Log(φ_{AM})

The following shows how the A and B for the linear model in Equation (A3) can be obtained when this linear model is fitted to data generated by the non-linear relation between the explanatory variable (p_F) and the observed effect $\mathrm{Log}(\varphi_{AM})$. For each value of p_F, we observe a corresponding value for $\mathrm{Log}(\varphi_{AM})$, which gives a nonlinear relationship (Equation (A1) and Figure A1). Thus, in order to obtain a linear relationship between the parameter of interest (p_F) and the dependent variable, we fit a line through this nonlinear relationship from which we estimate the effect (in this case, the Log of the effect of φ_F compared to φ_f). To fit a line through the true relationship $\mathrm{Log}(\varphi_{AM})$, we sample random values for p_F from a uniform distribution from 0 to 1 and calculate corresponding values of $\mathrm{Log}(\varphi_{AM})$ from Equation (A1). If we draw a least squares regression line through the random

numbers drawn from these (p_F, $\mathrm{Log}(\varphi_{AM})$) pairs, the line passes through the average values sampled: $\bar{p}_F$ and $\overline{\mathrm{Log}(\varphi_{AM})}$.

This allows us to find *B*, since we know that $\bar{p}_F = {}^1/_2$, and $\overline{\mathrm{Log}(\varphi_{AM})}$ is:

$\overline{\mathrm{Log}(\varphi_{AM})} = \int_0^1 \log\left[\left(\varphi_F - \varphi_f\right)p_F + \varphi_f\right] dp_F,$

$\overline{\mathrm{Log}(\varphi_{AM})} = \frac{(\varphi_f - \varphi_F) + \varphi_F \log \varphi_F - \varphi_f \log \varphi_f}{\varphi_F - \varphi_f}.$

Hence, since $\overline{\mathrm{Log}(\varphi_{AM})} = A \cdot \bar{p}_F + B$, and $\bar{p}_F = {}^1/_2$,

$B = \overline{\mathrm{Log}(\varphi_{AM})} - \frac{1}{2} A$, and thus

$\mathrm{Log}(\varphi_{LIN}) = A \cdot p_F + \overline{\mathrm{Log}(\varphi_{AM})} - \frac{1}{2} \cdot A.$

Now we have an equation with only one unknown (A) and the solution for A, denoted A_{min}, can be found by taking the least squares optimization. This means that we can find the minimum solution for the squared difference between the $\mathrm{Log}(\varphi_{AM})$ and $\mathrm{Log}(\varphi_{LIN})$ derived above.

$$A_{min} = MIN_A \int_0^1 \left(A \cdot p_F + \overline{\log(\varphi_{AM})} - \frac{1}{2} \cdot A - \mathrm{Log}\left(\left(\varphi_F - \varphi_f\right) p_F + \varphi_f\right)\right)^2 d\, p_F.$$

This integral was evaluated with symbolic computer algebra using Mathematica. This is a straightforward evaluation but, at first, the expressions appear to be very big. Thus, we undertook some simplifications to find the A for which the minimum of the expression is attained. As the part between brackets is a linear expression in A, the result of the above equation is a quadratic equation in A, and thus can be written as:

$MIN_A\ (K_2 A^2 + K_1 A + K_0).$

The equation between brackets is for an upward open parabola (if K_2>0) and the minimum of this parabola is attained for:

$A_{min} = \frac{-K_1}{2K_2}$, where (when $\varphi_F \neq \varphi_f$):

$K_1 = \frac{-6\,\varphi_F(\varphi_F - \varphi_f) - 6(\varphi_F - \varphi_f)\varphi_f + 12\varphi_F\varphi_f \mathrm{Log}(\frac{\varphi_F}{\varphi_f})}{12(\varphi_F - \varphi_f)^2} = \frac{-\varphi_F^2 + \varphi_f^2 + 2\varphi_F\varphi_f \mathrm{Log}\left(\frac{\varphi_F}{\varphi_f}\right)}{2\left(\varphi_F - \varphi_f\right)^2}$, and

$K_2 = \frac{\varphi_f(\varphi_f - \varphi_F) + \varphi_F(\varphi_F - \varphi_f)}{12(\varphi_f - \varphi_F)^2} = \frac{1}{12}$, thus

$A_{min} = \frac{3{\varphi_F}^2 - 3{\varphi_f}^2 - 6\varphi_F\varphi_f \mathrm{Log}[\frac{\varphi_F}{\varphi_f}]}{(\varphi_f - \varphi_F)^2}.$

Then, both the numerator and denominator of the above equation were divided by φ_f^2 and this resulted in:

$A_{min} = \frac{3\left(\frac{\varphi_F}{\varphi_f}\right)^2 - 3 - 6\left(\frac{\varphi_F}{\varphi_f}\right)\mathrm{Log}[\frac{\varphi_F}{\varphi_f}]}{(1 - \left(\frac{\varphi_F}{\varphi_f}\right))^2}.$

Thus, A_{min} is a function of $\frac{\varphi_F}{\varphi_f}$ only. Note the similarity with Equation (A2) where $A = \log\left(\frac{\varphi_F}{\varphi_f}\right)$. It should be noted that the A_{min} is the estimate (C_2) that will be obtained asymptotically (i.e. when n→∞) from the GLM in Equation (8). Thus, we investigated the relation of this estimated value to the true expected value $\mathrm{Log}\left[\frac{\varphi_F}{\varphi_f}\right]$, to quantify the bias due to our approach.

Let us assume that $\frac{\varphi_F}{\varphi_f} = x$, thus the above equation can be simplified as:

$A_{min} = \frac{3(x^2-1-2x\mathrm{Log}(x))}{(1-x)^2}$.

Still, $A_{min} \neq \log\left(\frac{\varphi_F}{\varphi_f}\right) \neq \log(x)$, hence there is a non-zero bias. However, we can now define m(x) by $A_{min} = m(x) \cdot \log(x)$. The value of *m* quantifies the amount of relative bias that is obtained as a result of the geometric approximation; a value *m* = 1 indicates a zero bias. Thus:

$$m(x) = \frac{A_{min}}{\mathrm{Log}(x)} = \frac{3(x^2-1-2x\mathrm{Log}(x))}{(x-1)^2\mathrm{Log}(x)}. \qquad \text{(A4)}$$

Equation (A4) quantifies the amount of bias, the magnitude of which is numerically investigated below. However, first it is necessary to check Equation (A4) using some relationships that are known to hold for the underlying problem, for example:

$m(x) = m\left(\frac{1}{x}\right)$, since it should not matter which allele is coded *F* or *f*.

$\lim_{x\to 1} m(x) = 1$, since the arithmetic and geometric mean are identical when $\varphi_F = \varphi_f$.

$\lim_{x\to 0} m(x) = 0$, and $\lim_{x\to\infty} m(x) = 0$, since we always underestimate the effect because 0 ≤ *m*(*x*) ≤1 and thus it seems that *m*(*x*) will have to approach zero when the real effect becomes infinitely large (*i.e.*, either *x* = 0 or *x*→∞). As a result, we will estimate a finite value for the effect even when the effect is infinite and, thus, we make an infinitely large error, i.e. *m*(*x*) = 0. All conditions hold as it can be checked using Equation (A4).

Going back to the manuscript, we now look at Equation (8), which is,

$\mathrm{cloglog}\, E\left[\frac{y_i}{n_i}\right] = c_0 + c_1 index_{G,i} + c_2 Num_F + log(\frac{I}{N})$,

where c_2, is the regression coefficient that we estimate. In other words, when applying the geometric mean approximation, we assume $Est(\log(x)) = \hat{c}_2$,

whereas in fact, $\mathrm{Log}(x) = \frac{\hat{c}_2}{m(x)}$, when we correct for the geometric mean approximation.

Since we assumed that $\varphi_f = 1$, then $\mathrm{Log}(x) = \mathrm{Log}\left(\frac{\varphi_F}{\varphi_f}\right) = \mathrm{Log}(\varphi_F)$. Thus:

$$\mathrm{Log}(\varphi_F) = \frac{\hat{c}_2}{m(\varphi_F)}, \qquad \text{(A5)}$$

Where, $m(\varphi_F) = \frac{3((\varphi_F)^2 - 1 - 2(\varphi_F)\,\mathrm{Log}[\varphi_F])}{((\varphi_F)-1)^2\,\mathrm{Log}[\varphi_F]}$. (A6)

The result from Equation (A6) quantifies the amount of error that was obtained as a result of the geometric approximation. An $m(\varphi_F)$ = 1 indicates no error, an $m(\varphi_F)$ < 1 indicates underestimation, while an $m(\varphi_F)$ > 1 indicates overestimation of φ_F. As 0 < $m(\varphi_F)$< 1, the estimated value is always too small. Hence, the geometric mean approximation is conservative. Furthermore, m(φ_F) = $m(\frac{1}{\varphi_F})$ for all φ_F and m(1) = 1, the further φ_F is away from 1 (the larger effect), the higher the error. Roughly speaking for values of φ_F between 0.333 and 3, the error is smaller than 5%; *i.e.*, 0.95 < *m* < 1.

Now that we have quantified the amount of bias (Equations (A5) and (A6)), we can obtain the correct value. Note that in Equation (A5), the (true) value of φ_F appears on both sides of the equation. Thus, we need an iterative procedure to obtain the real value. First, $\hat{\varphi}_F$ is calculated by taking the exponential of c_2 from the GLM analysis. Then, the error $m(\hat{\varphi}_F)$ (Equation (A6)) followed by the new value for in $\log(\varphi_F)$ in Equation (A5) are estimated. $\hat{\varphi}_F$ is then again calculated by taking the exponential of $\log(\varphi_F)$. This iteration process is then allowed to continue until there is no change in $\hat{\varphi}_F$.

In the tables below, the biases obtained as a result of the geometric mean approximation are presented for the different scenarios investigated in the main text. This bias is calculated as the difference between $\hat{\varphi}_F$ that is obtained after accounting for the error as a result of geometric mean approximation and $\hat{\varphi}_F$ that is obtained when the error is not accounted for. Note that there is additional bias with respect to the true value which is of course known from the simulations. This bias is also small but larger than asymptotically expected from the GM approximation.

Table 3.3. Biases in estimated φ_F for scenario 1.

r=0	*r*=0.25	*r*=0.5625	*r*=1
0.000674	0.002126	0.002569	0.002698

Table 3.4. Biases in estimated φ_F for scenario 2

	r=0	*r*=0.25	*r*=0.5625	*r*=1
R_0=0.6	0.000452	0.001309	0.002122	0.002706
R_0=1.2	0.000674	0.002126	0.002569	0.002698
R_0=1.8	0.001889	0.002498	0.002612	0.002671
R_0=2.5	0.002774	0.002467	0.002531	0.002565
R_0=3.1	0.003111	0.002364	0.002364	0.002492
R_0=3.7	0.003425	0.002181	0.002201	0.002375
R_0=4.3	0.003891	0.002032	0.002062	0.002223
R_0=4.9	0.004184	0.001891	0.001866	0.00208
R_0=5.5	0.004707	0.001798	0.001705	0.001903
R_0=6.1	0.005194	0.00167	0.001572	0.001743

Table 3.5. Biases in estimated φ_F for scenario 3

	r=0	*r*=0.25	*r*=0.5625	*r*=1
Small difference	8.60436E-05	9.26451E-05	7.50199E-05	5.75224E-05
Moderate difference	0.000674	0.002126	0.002569	0.002698
Large difference	0.002963	0.013645	0.017404	0.018417

Figure A1. Arithmetic mean (AM), geometric mean (GM) and linear approximation (Lin Apprx) of the best fitted line as a function of allele frequency p_F (input values used for φ_f and φ_F were 1 and 0.6, respectively).

4

The effect of polymorphisms in the major-histocompatibility complex (MHC) of Scottish Blackface sheep on individual susceptibility and infectivity to nematode infection

Mahlet T. Anche[1,3], P. Bijma[1], Michael J. stear[2], Mart M.C. de Jong[3]

[1]Animal Breeding and Genomics Centre, Wageningen University, 6700 AH, Wageningen, The Netherlands; [2] The Boyd Orr Centre for Population and Health, Institute of Biodiversity, Animal Health & Comparative Medicine, College of Medical and College of Medicine, Veterinary and life Sciences, University of Glasgow, Bearsden Road, G61 1QH, Glasgow, United Kingdom; [3] Quantitative veterinary Epidemiological group, Wageningen University, 6700 AH, Wageningen, The Netherlands

To be submitted

Abstract

Gastrointestinal nematode infections are one of the major diseases of domestic sheep, causing severe morbidity and loss of productivity. Loci located in or around the class II region of the Major HistoCompatability complex (MHC) on chromosome 20 have been associated with susceptibility to nematode infection in sheep exposed to the species mixture of nematodes that occurs in Scotland, i.e. predominantly *Teladorsagia circumcincta*. To the best of our knowledge, however, there are no studies that have estimated the effect of MHC-genes on the ability of an individual to infect others (individual *infectivity*) with respect to this nematode infection. Thus, the aim of this study was to estimate the effect of polymorphisms in the MHC-genes on individual susceptibility, and also to get an indication of whether MHC-genes also have an effect on individual infectivity. For that purpose, data on six loci of the MHC with 7 to 22 alleles each were analysed with a Generalized Linear Model (GLM). The effect of each allele for each locus on individual susceptibility was separately estimated. However, for infectivity, the number of separate effects that can be estimated is restricted by the number of independent groups, and therefore principal component analysis was used to reduce the number of separate effects to be estimated. The first two principal components that explained most of the variance were used for deriving an explanatory variable that was used in the GLM. At each locus, several alleles were found to have a significant effect on individual susceptibility. Moreover, at each locus either both or one of principle component based variables was found to have a significant effect on individual infectivity. The later suggests that there is at least one allele from the positive-weighted subset of alleles that has the same effect or larger when compared to any of the allele in the negative-weighted subset. Thus, our result shows that, in addition to their effect on individual susceptibility, the MHC-genes have an effect on individual infectivity for nematode infections. This result suggests that studies on disease genetics should also consider the effect of MHC-genes on individual infectivity, in addition to their effect on individual susceptibility.

Key words: Major histocompatibility complex, susceptibility, infectivity, principal component analysis, generalized linear model

4.1 Introduction

Infectious diseases are one of the major causes of livestock mortality and reduced productivity, worldwide (Bishop et al., 2002). Moreover, the zoonotic nature of some infectious diseases poses a threat to human health. Different disease control and preventive measures, such as antibiotic treatments, vaccines and management practices are implemented to control the threat imposed by pathogens and parasites. However, the development of resistance to antibiotics and vaccines has reduced the efficiency of these control methods. This has led to a demand for additional methods that complement existing disease control strategies.

One of these strategies is breeding for reduced disease prevalence, which is possible because individual animals differ genetically in their response to infectious pathogens (Axford et al., 2000). This variation is present in different types of disease traits, one of which is any difference in host's susceptibility to a given pathogen. As a result of research in quantitative genetics of livestock diseases, it is now common knowledge that individual animals vary genetically in their susceptibility to several infectious diseases (Bishop et al., 1996). Moreover, studies indicate that genetic variation in susceptibility to infectious pathogens plays an important role in the dynamics and prevalence of an infection in the population (Dwyer et al., 1997; Springbett et al., 2003).

The dynamics and prevalence of an infection in a population, however, is not only affected by genetic variation in individual susceptibility. Variation in other traits, such as individual infectivity, which is the ability of an individual to infect other (susceptible) individuals, may also exist, as can be seen (at least phenotypically) from the existence of 'superspreaders' (Lloyd-Smith et al., 2005). Such variation can also affect the dynamics and prevalence of an infection in the population (Diekmann and Heesterbeek, 2000).

Since an individual's susceptibility is a component of an individual's fitness, natural (and straightforward artificial) selection will work to exhaust genetic variation in susceptibility. Unlike individual susceptibility, however, individual infectivity is not part of an individual's fitness. Consequently a large amount of genetic variation in infectivity could persist even under natural selection (Lipschutz-Powell et al., 2012; Anche et al., 2014). Such genetic variation can potentially be utilized through selective breeding, so as to reduce disease prevalence in the population.

The reproduction ratio, R_0, is an important epidemiological parameter that is a measure for the disease risk. It is a measure for epidemic size, and for the prevalence of an infection when endemic. R_0 is defined as the average number of

secondary cases produced by a typical infectious individual during its entire infectious life time, in an otherwise naive population. R_0 has a threshold value of 1, which implies that a major disease outbreak or a stable endemic equilibrium can only occur when R_0>1. When R_0 < 1 the disease will die out. Thus, in order to reduce disease prevalence in a population, breeding strategies should aim at reducing R_0, preferably to a value below 1. We (Anche et al., 2014) showed that an individual's breeding value for R_0 is a function of its breeding values for susceptibility and infectivity, and of the population average susceptibility and average infectivity. Thus, in order to estimate effects of genes on R_0, we need to estimate the effects of genes on both susceptibility and infectivity.

Gastrointestinal nematode infections are one of the major diseases of (domestic) sheep, causing severe morbidity and consequential loss of productivity (Coop et al., 1977). Two quantitative trait loci (QTL) are associated with resistance to the mixed, predominantly *Teladorsagia circumcincta* nematode infections in sheep in Scotland. One QTL is located in or around the class II region of the Major Histocompatability complex (MHC) on chromosome 20, and the other is located in or around the interferon-γ gene on chromosome 3 (Davies et al., 2006). One of the loci within the MHC that is found to be associated with faecal egg count (FEC), which is used as an indicator of host susceptibility to nematode infections, is the DRB1 locus (Buitkamp et al., 1995; Schwaiger et al., 1995; Buitkamp et al., 1996; Stear et al., 1996). To the best of our knowledge, however, there are no studies that have estimated the effect of MHC-genes on individual infectivity. Although FEC may also be seen as a measure of each individual host's infectivity as it may not only be a measure for how often the host has become colonised, but that also the MHC-genes may affect an individual's infectivity in other ways that may (e.g. survival in the host, egg-productivity of the worms in the host) or may not (viability of eggs) affect FEC. In turn like susceptibility also this infectivity may affect the transmission and dynamics of nematode infections in the population. Thus, in this study we aim to jointly estimate the effect of polymorphisms in MHC-genes on individual susceptibility and individual infectivity. Relative susceptibility is defined as the (relative) chance that a host with a certain MHC genotype is infected, i.e. the chance that a host with that MHC genotype had more than the threshold FEC. Relative differences in infectivity are estimated by using the difference between years in the type of MHC genes in the infected hosts to explain between year differences in infection rate (number of cases). Anche et al. (2015) have developed a Generalized Linear Model (GLM) that estimates the relative effects of an individual' genes on its susceptibility and infectivity. Here, we use that model to

estimate the relative susceptibility and infectivity effects of MHC-polymorphisms on nematode infection in Scottish Blackface sheep.

4.2 Materials and Methods

4.2.1 Materials

Approximately 1000 Scottish Blackface lambs from 38 rams and 492 ewes were used in this analysis. The data includes lambs that were born in the years 1992 – 1996 (Davies et al., 2006). The lambs were kept on three separate fields together with their mothers until weaning, which occurred when they were about 3 or 4 months of age. The lambs were then separated from their mothers and were allowed to graze all on the same pasture and were continuously exposed to the same natural mixture of nematodes on that pasture, predominantly *Teladorsagia circumcinta* as excreted by the lambs themselves. The lambs were treated with anthelmintics every 4 weeks, from 4 to 20 weeks of age. Faecal samples were collected from the rectum of the lambs at 4 weeks of age and thereafter at 4-week intervals until 20 weeks of age. Thus, during part of the recording period, lambs were together with their dams in the same flock. Faecal egg count (FEC) per gram of faeces was then recorded for each lamb by a modified McMaster test. At six or seven weeks after the final anthelmintic treatment, lambs were slaughtered, at that time they were about 6-7 months old.

All animals were from the same commercial flock. All animals were kept on the same pasture after weaning. For our analysis, we use the measurement after weaning at 20 weeks of age. In this study, therefore, animals born in the same year were considered as one flock. A flock defines the individuals that take part in the same endemic, *i.e.,* the individuals that can potentially infect each other. Since the data include animals that were born in the years 1992 – 1996, we have data on 5 flocks. In other words, we observed five endemics, infections that were considered independent. Moreover, the data were organized in such a way that individuals with a FEC greater than 100/(gram of faeces) were considered to be infected, while the rest was considered as non-infected and susceptible. This was done assuming that the threshold FEC can be used to demonstrate the purpose of this study, and thus we used a FEC of 100/gram of faeces to discriminate the two sub-classes. As a result, we have binary data on infection status, where individuals are classified as infected or not-infected (susceptible) based on their FEC.

In addition to FEC, each individual was genotyped at 6 MHC loci at the class II region of the MHC. These were the DQA1, DQA2, DQB1, DQB2, DRB1 and DQA2like locus. Genotyping was done using PCR amplification and sequencing of exon 2 which carried most of the polymorphisms (Stear et al., 2005). At each locus,

there were multiple alleles, ranging from 7 to 22. Table 1 shows the frequency distribution of the markers (alleles) at each locus. After removal of animals with missing genotype or FEC, 828 lambs were used in this study.

4.2.2 Methods

Anche et al. (2015) developed a GLM to estimate the relative effect of genes on an individual's susceptibility and infectivity. The model was developed from the so-called final size equation (Andreasen, 2011), i.e. an equation that describes the probability of an individual to become infected, as a function of its susceptibility genotype and the infectivity genotypes of its infectious group mates. However, here we will apply the method to the data of a single measurement of an SIS infection that is endemic in a herd; again the number and type of susceptible and the number and type of infectious determine the chances of infection and the same equation and estimation procedure applies as in case of the final size. Again, as in case of the final size, we estimate here directly the reproduction ratio. The model yields estimates that are the log of the relative susceptibility and infectivity effect of alleles as compared to one reference allele. Hence, estimates of relative susceptibility and infectivity effects of alleles follow from the exponential of the estimated regression coefficients. The detailed steps taken to develop the GLM can be found in (Anche et al., 2015b).

To estimate the effect of MHC-genes on individual susceptibility and infectivity, each locus was analysed separately. Individual records on infection status and allele types provide information on the allele's susceptibility effect. Since we have many more records than the number of unknown allele effects to be estimated (7 to 22 alleles at each locus, and thus 6 to 21 relative effects), we can fit all the alleles at a locus in the GLM and estimate their effect on individual susceptibility.

To estimate the effects of MHC-alleles on infectivity, we have to consider the frequency of these alleles in the infected individuals in the flock. This is because the allele frequencies in the infected sheep affect the infection probability for all the recipient sheep in one flock in the same way. Note that for these data, a flock is the same as one year and because there are only 5 years (i.e. flocks) in the data set, there are only 4 informative contrasts that can be estimated. This implies that we can use only a small part of the possible variation in allele frequencies to estimate any infectivity effects. Therefore, we reduced the number of explanatory variables for infectivity by using principle component analysis (PCA) on the allele frequencies, for each locus separately. The first principle component (PC1) is the vector of weights for each allele that explains most of the variance in allele frequencies (between years), and PC2 explains most of the remaining variance, etc.

Typically, the elements of the PC have positive (w^p) and negative weights (w^n), and we used for the first two PC (PC1 and PC2) the following explanatory variable in the analysis:

$$PC_{pw} = \frac{\sum_{i=1}^{m} w_i^p f_i^p}{\sum_{i=1}^{m} w_i^p f_i^p + \sum_{i=1}^{n} w_i^n f_i^n}$$

which is the sum of the positive weights divided by the absolute values of all the weights. This yields a value between 0 and 1. PC_{pw} stands for the PC's that are positively weighted. The f^p and f^n are the frequencies in each year of the positive-weighted and negative-weighted alleles respectively. The regression coefficient of this explanatory variable obtained from the GLM is an estimate for the (log) effect on infectivity of the weighted average of positive-weighted subsets of alleles compared to the weighted average of negative-weighted subsets of alleles. This means for the estimated effect using this explanatory variable (PC_{pw}), that there is at least one allele in the positive-weighted subset that has the same effect or larger when compared to the allele with the most different effect in the negative-weighted set of alleles.

For each locus, a separate analysis was performed. All the alleles of the individual itself for susceptibility, and the PC_{pw} from the first two PC of its infected flock mates for infectivity were fitted simultaneously in the GLM. In a second analysis, we also fitted a fixed year effect in the GLM together with all the alleles of the individual itself. This was done to investigate whether year also has an effect on individual FEC, in addition to the effect of MHC- alleles (genes).

An allele with a high frequency in the population was set as a reference allele. Thus, susceptibility effects of all the other alleles at this locus were estimated relative to this reference allele. In other words, the effect of the reference allele was set to a value of 1 so that the logarithm of the effect is zero. The exponential of the estimated regression coefficients for the other alleles were then taken to obtain the relative effects of the alleles on individual susceptibility and of the sets of alleles for relative infectivity. This was done because the susceptibility and infectivity effects enter the GLM at the logarithmic scale (Equation 7 in (Anche et al., 2015)). The ingredients of a GLM analysis (see McCullagh and Nelder (1989)) are the dependent variable (here: cases/total number), explanatory variables (here: alleles of the individual and the PC of the alleles of the infected herd members as explained below), a link function (here: complementary loglog), and a distribution of the error term (here: binomial). The final GLM (Equation 8 in (Anche et al., 2015)) used in this study is:

$$\text{cloglog}(\,\mathrm{E}\left[\frac{y_{i,j}}{m_{i,j}}\right]) = c_0 + c_2 index_{2_j,i} + c_3 index_{3_j,i} .. + c_n index_{n_j,i} + c_a PC_{pw1} + c_b PC_{pw2} + \log(\frac{I}{N}) \quad (1)$$

where the cloglog is applied to the expectation of $\frac{y_{i,j}}{m_{i,j}}$, which in this case is the fraction of infected individuals with genotype *i* at locus *j* (where $y_{i,j}$ is the total number of infected individuals with genotype *i* at locus *j*, and $m_{i,j}$ is the total number of individuals with genotype *i* at locus *j*). c_0 is the intercept measuring the logarithm of the transmission rate parameter R_0. (c_1 is not present in Equation 1 since it is the regression coefficient for the reference allele, which was set to an effect of zero. We formulate the model in terms of allele count within individuals rather than individual genotypes, and thus assumed the two alleles that make an individual genotype act multiplicatively, so that their effects act additively on the logarithmic scale): c_2,...c_n are regression coefficients for $index_{2_j,i}$ through $index_{n_j,i}$, where $index_{2_j,i} = 0, 1\ or\ 2$, indicating the number of non-reference alleles at locus *j* of individuals with genotype *i*, and the same applies for the rest of the *index* variables in the model. c_a is the regression coefficient for the PC_{pw1} (again this applies to the allele frequencies so PC_{pw1} is a value between 0-2) and c_b is the regression coefficient for PC_{pw2}. The $\log(\frac{I}{N})$ is the total fraction of infected individuals in each group (year), and was used as an offset in the GLM.

The c_2, c_3,... c_n estimates represent the log of the allele substitution effects of allele 2 till allele *n* at locus *j* on individual susceptibility. The c_a and c_b estimates represent the log of the relative allele substitution effects of the positively-weighted alleles on individual infectivity. The estimates c_2 through c_n and c_a and c_b are on the log scale, and are relative to the reference allele. Thus, when sus_{2_j}, sus_{3_j}, ..., sus_{n_j} represent the susceptibility effects of alleles 2 till *n* at locus *j,* these can be calculated as $\widehat{sus}_{2..n_j} = e^{\hat{c}_{2..n}}$, and are relative to the reference allele number 1. Analogously, $inf_{PC_{pw1,j}}$ and $inf_{PC_{pw2,j}}$ were used to represent the infectivity effect of the two subsets of positively-weighted alleles at locus *j*. Thus, infectivity effects of positively-weighted alleles were calculated as $\widehat{inf}_{PC_{pw1,j}} = e^{\hat{c}_a}$ and $\widehat{inf}_{PC_{pw2,j}} = e^{\hat{c}_b}$, respectively.

Variables that did not have a significant effect, except those that were found to have a confounding effect with the rest of the effects, were removed from the model, starting from the variable with the highest p-value. The effect of a variable

was considered to be confounded if the change in the estimated regression coefficients of other variables was >25% when removing this variable from the model. A variable was also kept in the model if the Akaike's Information Criterion (AIC) went up by more than 2.0 instead of down when removing the explanatory variable from the model.

4.3 Results and Discussion

4.3.1 Susceptibility

In this study, we have used an individual FEC of 100/gram of faeces as a cut-off point to discriminate between those sheep that are infected and those that are not, which thus measures (indicates) whether or not the individual is "infected". This measure allows the estimation of the effects of alleles on susceptibility. Within those individuals that are classified as infected, however, we have individuals with different FEC, which in turn could be a measure of the level of an individual's infectivity. However, we have not included this in our analysis because the observed FEC may not reflect the average FEC over the whole infected period. Estimated susceptibility effects of all the alleles that were found to have a significant effect, including those alleles that were found to have confounding effect (printed in italics), are presented in Table 4.2. The number of alleles with a significant effect relative to the reference allele varied among the loci (Table 4.2). At all the loci estimated effects of most of the alleles were smaller than one, suggesting a favourable effect of the alleles on individual susceptibility compared to the reference allele. Only four alleles, that is, one allele at locus DQA1, one allele at locus DQA2like and two alleles at locus DRB1, were found to have an estimated effect greater than one, indicating greater susceptibility than the reference allele. Allele $\mathrm{AY265308}$ at locus DQA1, allele $\mathrm{GU191459}$ at locus DQB1, allele $*0308$ at locus DRB1 and allele $AY312396$ at locus DQA2like were found to have a confounding effect with the rest of the alleles, and thus were kept in the GLM even though their effects were not significant.

One of the most polymorphic loci within the MHC that is known to have an association with susceptibility to nematode infections is the DRB1 locus. At this locus, allele *1* is most frequent and this allele has been used as reference allele in previous studies. Also in this study, we used allele *1* as a reference. In this study, *G2* allele was found to have an effect less than one, indicating favourable effect on individual susceptibility relative to the reference allele. This suggests that allele *G2* is associated with reduced individual susceptibility.

This result agrees with previous studies (Schwaiger et al., 1995; Stear et al., 1996) that have used the same population of Scottish Blackface Sheep in their

analysis and found that allele *G2* at the DRB1 locus is associated with reduced FEC. In addition to allele *G2*, 4 more alleles were found to have a significant effect on individual susceptibility compared to allele *I* at this locus (Table 4.2). For all the loci, the final model with all the significant variables fitted the in the GLM is given the appendix.

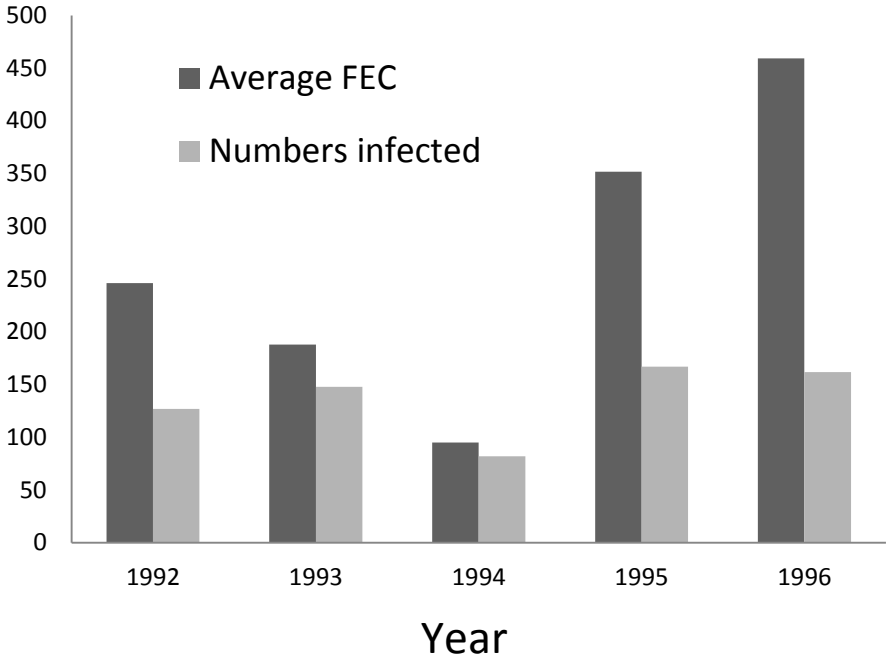

Figure 4.2. Distribution of faecal egg count (FEC) across years and number of individuals that are infected.

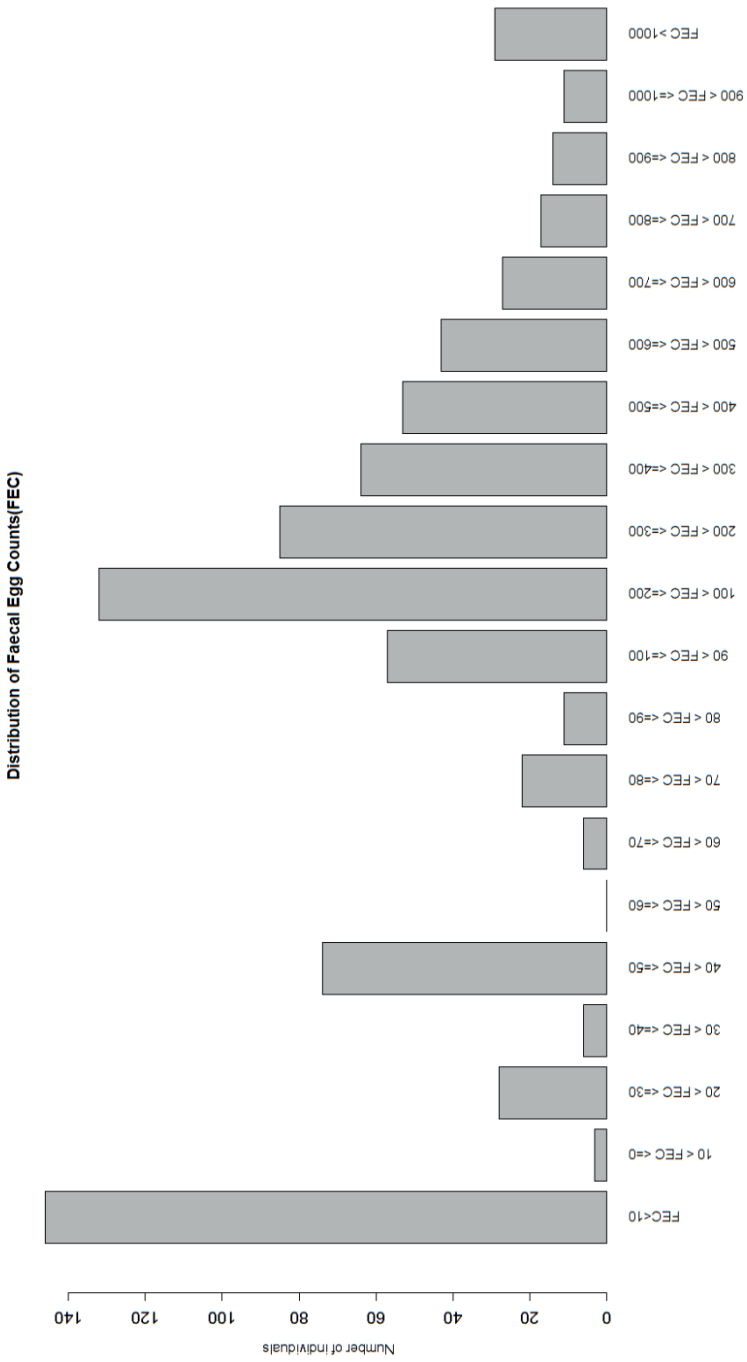

Figure 4.1. Distribution of faecal egg count (FEC) in the population.

4.3.2 Infectivity

Infectivity effects of the two subsets of positively-weighted alleles (PC_{pw1} and PC_{pw2}) were estimated, and at locus DQA1 and DQA2, both subsets of alleles were found to have a significant effect on individual infectivity. At the rest of the loci, only one of the two subsets of alleles was found to have significant effect on individual infectivity (Table 2). In both cases, however, we are not able to pinpoint to the allele (s) with a significant effect. This is because we have fewer records than the number of variables to be estimated. However, estimated infectivity effects were comparable to estimated gene effect of MHC-genes on individual susceptibility (Table 2).

Correlations between estimated regression coefficients of the two subsets of positively-weighted alleles (PC_{pw1} and PC_{pw2}) were calculated for each locus. Only at locus DQA1 considerable correlations were observed (-0.17). One of the reasons for this correlation could be due to the fact that the two subsets share allele(s) that have a marked effect on infectivity.

Moreover, as mentioned in the Method section, year was fitted in the GLM as a fixed effect in order to investigate whether year has an effect on individual FEC in addition to the effect of MHC-genes. In this case, the last two years were found to have a significant effect on individual FEC for each locus analysed separately (Table 4.3). Preliminary analysis of FEC supports this result where the last two years were found to have highest FEC than the first three years (Figure 4.2). Moreover, a confounding effect was observed between year effect and the susceptibility effects of MHC-genes at locus DQB1, DQB2 and at locus DQA2like. At locus DQB1, the effect of allele $AJ238938$ and $GU191460$ that was found to be significant without fixed year effect disappeared when the fixed year effect of year was fitted in the model. At locus DQB2, on the other hand, in addition to the alleles that were found to have significant effect, one more allele, allele $AJ238946$, was also found to have a significant susceptibility effect. At locus DQA2like, the effect of allele $AY312385 + AY312397$ was found to be not significant when fixed year effect was fitted in the model. At all the three loci, however, the change in estimated susceptibility effect of the rest of the alleles was not significant.

Table 4.1. Allele type and allele frequency at all the loci

Locus DQA1		Locus DQA2		Locus DQB1		Locus DQB2		Locus DRB1		Locus DQA2like	
Type	Frequency	Type	Frequency	Type	Frequency	Type	Frequency	Type	Frequency	Type	Frequency
92.y085	0.0789	***0101**	**0.3941**	2032	0.0031	92.y099	0.0064	*0308	0.0015	AY312378	0.0005
94.b089	0.0016	*0102	0.0005	94.b076	0.0041	93.o055	0.0048	*0801	0.0010	AY312385	0.0488
95.y042	0.0005	*0103	0.1100	94.y089	0.0021	94.b089	0.0016	*1301	0.0041	AY312394	0.3909
AY265308	0.0052	*0201	0.0472	94.y164	0.0109	94.y145	0.0331	*1901	0.0041	AY312396	0.0346
CCC55881	0.0280	*0601	0.0330	AJ238936	0.0471	AJ238931	0.0315	A	0.0560	GU191459	0.0005
CCC55882	0.0031	*0602	0.2065	AJ238938	0.0528	AJ238932	0.0235	B	0.0154	AY312395+AY312397	0.0005
HQ728659	0.0426	*0901	0.0288	AJ238939	0.0549	AJ238933	0.0817	C	0.0139	**Null**	**0.5236**
M33304	0.0675	*1001	0.0561	AJ238941	0.2098	AJ238935	0.0855	D	0.1880		
Z28418	0.2066	*1101	0.0838	AJ238942	0.0316	AJ238937	0.0630	F	0.0473		
Z28518	0.0883	*1201	0.0031	AJ238945	0.0777	AJ238940	0.0043	G1	0.0272		
Null	**0.4777**	Null	0.0372	GU191453	0.0083	AJ238942	0.0005	G2	0.1315		
				GU191454	0.0440	AJ238944	0.0459	H1	0.0051		
				GU191459	0.0016	AJ238946	0.1341	H2	0.0092		
				GU191460	0.0187	N_all4	0.0475	**I**	**0.3114**		
				HQ728667	0.0632	N_all5	0.0288	L	0.0801		
				Null	**0.3674**	N_all7	0.0074	M	0.0529		
						N_all9	0.0064	N	0.0319		
						U07030	0.0011	PQR	0.0179		

			Null	0.0010
	U07032	0.0021		
	U07033	0.0267		
	Z28424	0.0011		
	Null	**0.3616**		

When a fixed year effect was fitted together with the PC_{pw} from the first two PC, we found that infectivity effects of MHC-genes were fully confounded with year effects. As a result, infectivity effects of the MHC-genes were not estimable. This suggests that variation between groups in the number of cases could be due to variation between years or due to the difference in infectivity effect of MHC-genes. Both could have these effects through changing the number of infective larvae in the pasture. However, since we have only one group per year, we cannot make a claim that either are the causes for the difference in the number of cases among groups.

As mentioned above, the effect of MHC-genes on individual infectivity and year effect were fully confounded and thus we were not able to disentangle the infectivity effect of MHC-genes and year on the number of cases. The main reason for that shortcoming is the fact that the data comes from only one flock per year (which were considered as groups in this study). Having multiple flocks within each year would have allowed to exclude the extra variation (noise) coming from the years and thus enables better understanding of the effect of MHC-genes on individuals infectivity. Another consequence of having few epidemiological groups is that we were unable to estimate the relative infectivity effects of each. This is because information about individual infectivity is inferred from comparison of different epidemiological groups. Thus, having observation on a large number of flocks would have also allowed us to estimate the relative infectivity effect of all MHC-genes at each locus.

As mentioned in the Material section, the lambs were kept with their dams for part of the recording period. Hence, the dams could also have been infected and could contribute to the disease status of individual lambs. Since the data did not include observation (FEC) on the dams, we were unable to correct for the contribution of the dams to the disease status of their lambs in our analysis. Moreover, it is possible that the same dam to have contributed infected lambs in different years. In this case, we cannot assume independency of epidemics between the different groups (years). This is because the same dam has contributed to the prevalence of infection in the different groups via the infected lambs.

Table 4.2 Final model from the GLM analysis the overall reproduction ratio estimate from the intercept and the effects of MHC-genes (alleles) on individual susceptibility and infectivity

Locus DQA1	Regression coefficient (RC)	$exp^{(RC)}$	P-value
$\boldsymbol{sus_{Null}}$	0	1	
$\boldsymbol{sus_{\mathrm{AY265308}}}$	***-0.790***	***0.454***	***0.052***
sus_{CCC55881}	-0.618	0.539	0.001
inf_{WEV1}	0.993	2.699	7.95e-05
inf_{WEV2}	-0.946	0.388	3.65e-05
Locus DQA2			
sus_{*0101}	0	1	
sus_{*0201}	-0.316	0.729	0.009
sus_{*0901}	-0.580	0.559	0.001
inf_{WEV1}	-0.366	0.694	0.026
inf_{WEV2}	-0.546	0.579	0.017
Locus DQB1			
sus_{Null}	0	1	
sus_{GU191459}	***3.268***	***26.261***	***0.956***
$sus_{\mathrm{94.b076}}$	0.839	2.314	0.013
sus_{AJ238936}	-0.351	0.704	0.004
sus_{AJ238938}	-0.247	0.781	0.036
sus_{GU191460}	-0.554	0.575	0.007
inf_{WEV2}	-0.472	0.624	0.024
Locus DQB2			
sus_{Null}	0	1	
sus_{AJ238944}	-0.271	0.763	0.032
$sus_{\mathrm{N\_all5}}$	-0.471	0.624	0.004
inf_{WEV2}	-0.433	0.648	0.001
Locus DRB1			
sus_{I}	0	1	
$\boldsymbol{sus_{*0308}}$	***3.401***	***29.979***	***0.945***
sus_{*1901}	0.847	2.333	0.021
sus_{D}	0.162	1.176	0.010
sus_{G1}	-0.621	0.537	0.0002
sus_{G2}	-0.206	0.814	0.008
sus_{H1}	-0.767	0.464	0.020
inf_{WEV1}	-0.799	0.449	5.09e-10
Locus DQA2like			

sus_{Null}	0	1	
$sus_{AY312396}$	***3.663***	***38.963***	***0.975***
sus_{AY312385}	-0.426	0.653	0.006
$sus_{\text{AY312385+AY312397}}$	0.440	1.553	0.021
inf_{WEV2}	1.724	5.607	0.046

Note: Those alleles that are in bold and italic are alleles that have confounding effect with the other alleles with significant effect. For each locus, the first allele with a regression coefficient (RC) of zero is the reference allele for that locus. The effect of this allele was set at zero, thus no P-value is shown.

The aim of this study was to estimate the relative effect of MHC-genes on individual susceptibility and also to indicate whether they also have an effect on individual infectivity. Even though the relative effect of specific alleles on individual infectivity could not be estimated, we were able to find significant association between MHC-genes and individual infectivity. This result supports our hypothesis that there will be genes which have effects on individual infectivity since both natural and artificial selection will not exhaust the genetic variation that may present in this trait. However, alternatively these differences can also be explained by differences in between years, which alternatively were found to have significant effect on FEC (Table 4.3).

Significant associations between alleles at the MHC-genes and FEC, which is used as an indicator trait to measure individual susceptibility to nematode infection have been reported in a number of studies (Schwaiger et al., 1995; Buitkamp et al., 1996). On the contrary, the association between MHC-genes and their effect on individual infectivity has not received much attention. One of the main reasons for studies to focus on the association between individual susceptibility and MHC-genes could be due to individual's disease status (infected/not infected) being merely a function of its own susceptibility and the non-genetic environmental factors only. This assumption makes sense since individual's susceptibility reflects the probability an individual to be infected upon exposure to an infectious agent. However, studies have indicated that individual's (disease) phenotype, is not only affected by its own susceptibility but also affected by infectivity (infectiousness) of infected individual in its proximity.

The theory of direct-indirect genetic effect defines heritable effect of an individual on the trait value of the individual itself as direct genetic effect and heritable effect of an individual on the trait value of another individual as indirect (associative)

genetic effects (Griffing, 1967, 1976, 1981). It was showed by a number of studies that genetic variation present in direct and indirect genetic effects of an individual will contribute to the total genetic variation in the traits and thus may affect response to selection, both in magnitude and direction (Muir, 2005; Bijma et al., 2007; Bijma, 2011).

Table 4.3. Estimated year effect for all the loci analysed

Years	Regression Coefficient (RC)	$exp^{(RC)}$	P-value
Locus DQA1			
1995	0.543	1.721	8.36e-09
1996	0.594	1.758	3.01e-10
Locus DQA			
1995	0.488	1.629	9.73e-08
1996	0.562	1.754	3.80e-09
Locus DQB1			
1995	0.521	1.684	1.47e-08
1996	0.581	1.788	4.81e-10
Locus DQB2			
1995	0.494	1.649	5.47e-08
1996	0.582	1.790	9.03e-10
Locus DRB1			
1995	0.556	1.744	2.63e-10
1996	0.653	1.921	4.98e-13
Locus DQA2like			
1995	0.453	1. 573	8.93e-05
1996	0.536	1.709	9.15e-07

Based on this theory, individual's susceptibility can be considered as direct genetic effect since it affects the disease status of the individual itself. An individual's infectivity, on the other hand, is the rate at which an individual transmits the infection to a typical susceptible individual and thus it can be regarded as indirect genetic effect. In fact, not only infectivity, but also susceptibility has an indirect genetic effect component as anybody in a population or in a group with more susceptible individuals has a higher chance of becoming infected (Anche et al., 2014).

In the case of macroparasite infections, such as nematode infections, individual's FEC, which indicates the amount of infectious material excreted by the

individual, can also be seen as a measure of individual's infectivity. Thus, genetic variation that may be present in infectivity will affect the prevalence of an infection in the population. Therefore, studies on the disease genetics of nematode infections should consider the impact of individual's infectivity since genetic variation, when present, can be utilized to affect the prevalence of an infection in the population. Therefore, there should be more focus on designing experiments in such a way that data could be recorded in multiple epidemiological groups in the population, so as to better capture genetic effect of MHC-genes on individual infectivity.

4.4 Conclusion

Our results suggest that MHC-genes may have a genetic effect on individual infectivity in addition to their effect on individual susceptibility to nematode infection. Thus, studies on the disease genetics of sheep should also focus on the effect of MHC-genes on individual infectivity, since variation in infectivity can be used to select for a lower prevalence of infection.

4.5 References

Anche, M., P. Bijma, and M. De Jong. 2015a. Genetic analysis of infectious diseases: estimating gene effects for susceptibility and infectivity. Genetics Selection Evolution 47: 85.

Anche, M., M. de Jong, and P. Bijma. 2014. On the definition and utilization of heritable variation among hosts in reproduction ratio R0 for infectious diseases. Heredity 113: 364-374.

Anche, M. T., P. Bijma, and M. De Jong. 2015b. Genetic analysis of infectious diseases: estimating gene effects for susceptibility and infectivity. Genet Sel Evol 47: 1-15.

Andreasen, V. 2011. The final size of an epidemic and its relation to the basic reproduction number. Bulletin of Mathematical Biology 73: 2305-2321.

Axford, R., S. Bishop, F. Nicholas, and J. Owen. 2000. Breeding for disease resistance in farm animals. CABI publishing.

Bijma, P. 2011. A general definition of the heritable variation that determines the potential of a population to respond to selection. Genetics: genetics. 111.130617.

Bijma, P., W. M. Muir, and J. A. Van Arendonk. 2007. Multilevel selection 1: quantitative genetics of inheritance and response to selection. Genetics 175: 277-288.

Bishop, S., K. Bairden, Q. McKellar, M. Park, and M. Stear. 1996. Genetic parameters for faecal egg count following mixed, natural, predominantly Ostertagia circumcincta infection and relationships with live weight in young lambs. Animal Science 63: 423-428.

Bishop, S. C., M. C. M. d. Jong, and D. G. 2002. Commison on genetic resources for food and agriculture.

Buitkamp, J., P. Filmether, M. J. Stear, and J. T. Epplen. 1996. Class I and class II major histocompatibility complex alleles are associated with faecal egg counts following natural, predominantly Ostertagia circumcincta infection. Parasitology research 82: 693-696.

Buitkamp, J., D. Gostomski, F. Schwaiger, J. Epplen, and M. Stear. 1995. Association between the ovine major histocompatibility complex DRB1 gene and resistance to Ostertagia circumcincta infection. Zentralblatt fuer Bakteriologie.

Coop, R., A. Sykes, and K. Angus. 1977. The effect of a daily intake of Ostertagia circumcincta larvae on body weight, food intake and concentration of serum constituents in sheep. Research in Veterinary Science 23: 76-83.

Davies, G. et al. 2006. Quantitative trait loci associated with parasitic infection in Scottish blackface sheep. Heredity 96: 252-258.

Diekmann, O., and J. Heesterbeek. 2000. Mathematical epidemiology of infectious diseases. Wiley, New York.

Dwyer, G., J. S. Elkinton, and J. P. Buonaccorsi. 1997. Host heterogeneity in susceptibility and disease dynamics: tests of a mathematical model. The American Naturalist 150: 685-707.

Griffing, B. 1967. Selection in Reference to Biological Groups .I. Individual and Group Selection Applied to Populations of Unordered Groups. Australian Journal of Biological Sciences 20: 127-&.

Griffing, B. 1976. Selection in Reference to Biological Groups .5. Analysis of Full-Sib Groups. Genetics 82: 703-722.

Griffing, B. 1981. A Theory of Natural-Selection Incorporating Interaction among Individuals .2. Use of Related Groups. J Theor Biol 89: 659-677.

Lipschutz-Powell, D., J. A. Woolliams, P. Bijma, and A. B. Doeschl-Wilson. 2012. Indirect genetic effects and the spread of infectious disease: are we capturing the full heritable variation underlying disease prevalence?

Lloyd-Smith, J. O., S. J. Schreiber, P. E. Kopp, and W. M. Getz. 2005. Superspreading and the effect of individual variation on disease emergence. Nature 438: 355-359.

McCullagh, P., and J. A. Nelder. 1989. Generalized linear models. CRC press.

Muir, W. M. 2005. Incorporation of competitive effects in forest tree or animal breeding programs. Genetics 170: 1247-1259.

Schwaiger, F.-W. et al. 1995. An ovine major histocompatibility complex DRB1 allele is associated with low faecal egg counts following natural, predominantly Ostertagia circumcincta infection. International journal for parasitology 25: 815-822.

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

Stear, M. et al. 1996. An ovine lymphocyte antigen is associated with reduced faecal egg counts in four-month-old lambs following natural, predominantly Ostertagia circumcincta infection. International journal for parasitology 26: 423-428.

Stear, M., G. Innocent, and J. Buitkamp. 2005. The evolution and maintenance of polymorphism in the major histocompatibility complex. Veterinary immunology and immunopathology 108: 53-57.

Appendix

Using the final size equation (Andreasen, 2011), we previously showed that the transmission rate parameters for a discrete number of types for a SIR type dynamics can be estimated from a single (cross sectional) observation after the outbreak is over (Anche, Bijma, and De Jong, 2015). The number of cases for each type observed to have ever been infected during the outbreak can only have been caused by the infected individuals during the outbreak, which are the same individuals as the cases. Except for the index case, which is infected from outside the population, all other individuals have infected each other. Although this seems a story like the one by the baron von Münchhausen, who pulled himself out of the swamp by his own moustache, it does really work as shown by comparing outcomes of simulations (with known parameters) to the estimated parameters (Anche et al., 2015).

The same approach can be applied to estimating the transmission rate parameters from a population observed at its endemic equilibrium caused by SIS dynamics. The SIS assumption implies that the host will become susceptible again after recovery, i.e. when its infectious period ends. At the endemic equilibrium for the SIS dynamics, a constant distribution over types is observed with random fluctuations around the endemic steady state. In a single cross sectional, we observe the number of infected/infectious individuals and their genotypes. So this provides an estimate of the steady state in terms of numbers infected for the different genotypes. Moreover, as we observe in a single cross sectional all individuals that were infected between 0 and T days ago, where T is the duration of the infected/infectious period, we also observe the number and genotypes of new cases during a time period T. Because of the equilibrium, these have been caused by infected individual that come from the same distribution as observed in the cross sectional. This leads to the same relationship between cases C and infecteds I as for the final size situation, where here the transmission rate parameters are estimated for a time period equal to the infectious period T:

$\frac{C_i}{N_i}$**=1-** $e^{-\gamma_i \varphi_j \frac{c}{\alpha} \frac{1}{N}}$**.**

5

Estimating genetic co(variances) and breeding values for host susceptibility and infectivity from the final disease status of hosts exposed to epidemics in group-structured populations

Mahlet T. Anche[1,2], M.C.M. De Jong[1], P. Bijma[1]

[1]Animal Breeding and Genomics Centre, Wageningen University, 6700 AH Wageningen, The Netherlands

[2]Quantitative Veterinary Epidemiological group, Wageningen University, 6700 AH, Wageningen, The Netherlands

To be submitted

Abstract

The basic reproduction ratio R_0 is an important epidemiological parameter that determines the risk and prevalence of an infection. It has a threshold value of 1, where major disease outbreaks can occur when $R_0 > 1$. When $R_0 < 1$, only minor outbreaks can occur and the disease will die out. Thus, any breeding strategy aiming to reduce the risk and prevalence of an infection should reduce R_0 ideally, in combination with other control measures, to a value below 1. Breeding for reduced R_0, however, requires estimating genetic parameters (*i.e.*,(co)variances) and breeding values for R_0. An individual's breeding value for R_0 is a function of its gene effects on susceptibility and infectivity. Thus, genetic variance present in individual susceptibility and infectivity will contribute the heritable variation in R_0. Because of methodological challenges, such as including the nonlinear and stochastic nature of disease dynamics into linear mixed models, estimating genetic variances in susceptibility and infectivity has proven difficult. Thus, the aim of this study was to develop a methodology to estimate genetic (co)variances and breeding values for susceptibility and infectivity, by taking the nonlinear and stochastic nature of disease dynamics into account. To that end, we have developed a generalized linear mixed model (GLMM). A simulation study was performed to investigate the accuracy of estimated (co)variances and breeding values. We included the effect of relatedness among group mates in this investigation. While the estimated genetic variance in susceptibility closely resembled the simulated value, the genetic variance in infectivity was severely over-estimated, except when groups consisted of a few families. Estimates of genetic parameters and breeding values for infectivity improved when groups consisted of a few families. Accuracies of estimated breeding values for both traits increased when traits were positively correlated. On the other hand, for both traits, smaller biases were found when negative correlation was assumed.

Key words: generalized linear mixed model, susceptibility, infectivity, breeding value, genetic variance, relatedness

5.1 Introduction

Infectious diseases contribute an important set of problems to the sustainability of the livestock industry, mainly due to their impact on the productivity and welfare of livestock. In addition to that, due to their zoonotic nature and/or their ability to transfer antibiotic resistance to human pathogens, several infectious diseases, such as bovine tuberculosis and MRSA, pose a threat to human health. Different disease control strategies, such as vaccination and antibiotic treatments, are used to control the threats imposed by infectious diseases. However, the evolution of resistance by bacteria to the antibiotics and of viruses to escape the vaccines has led to an increased demand for alternative disease control strategies that can complement the existing methods.

In the past few decades, a number of studies have reported the existence of heritable variation among individuals in their response to different infectious diseases (Nicholas, 2005). As a result, these findings have opened the door for animal breeders to use selective breeding for livestock with an improved response to infectious diseases.

Among others, individual host susceptibility and infectivity are two disease-related traits that play an important role in the transmission of an infection in a population. Individual susceptibility is the probability of an individual to become infected upon exposure to a typical infectious individual. Individual infectivity is the probability that an infected individual transmits the infection to a typical susceptible individual, given contact (Anderson et al., 1992). Studies have reported phenotypic variation among individuals in their susceptibility to infectious diseases (Axford et al., 2000). Moreover, it is clear that susceptibility to infectious diseases has a genetic basis and that there exists genetic variation among individuals (Nicholas, 2005). Such heterogeneity influences the prevalence of an infection in a population (Springbett et al., 2003).

Quantitative genetic analyses of livestock diseases tend to focus on genetic variation in individual susceptibility to an infection. This is because these analyses implicitly assume that the only genetic effect affecting an individual's disease status (0/1) comes from its own genes. As a result, these studies capture genetic variation present in susceptibility only. The existence of 'superspreaders' in disease outbreaks, however, suggests that there exists (phenotypic) variation in infectivity among individuals that could affect the prevalence of an infection in a population (Lloyd-Smith et al., 2005). Evolutionary genetic arguments also suggest that a relatively large genetic variation may exist in infectivity, as opposed to susceptibility (Anche et al. 2014, Chapter 2). This is because both natural and artificial selection work to exhaust genetic variation present only in those traits

that are part of individual fitness, such as susceptibility (Denison et al., 2003). Unlike individual susceptibility, however, individual infectivity does not affect the disease status of the individual itself, and is therefore not part of an individual's fitness. Thus, in the absence of group or kin selection (Bijma and Wade, 2008), natural or artificial selection is hindered from exhausting genetic variation that may be present in infectivity.

The risk and severity of an epidemic, and the level of an endemic equilibrium in a population, is measured by R_0, which is a central parameter in epidemiology. R_0 is the average number of new cases produced by one typical infectious individual in a completely susceptible population. R_0 has a threshold value of 1. Major disease outbreaks or an endemic equilibrium can occur only when $R_0 > 1$, while the disease will die out when $R_0 < 1$. Due to this threshold property, any breeding strategy that aims to reduce disease prevalence should ideally aim to reduce R_0 to a value below 1, in combination with other control measures.

Selection for reduced R_0 is ideally based on individual estimated breeding values for R_0. Anche et al. (2014) showed that individual breeding value for R_0 can be defined combining results from the field of indirect-genetic effects with the epidemiological concept of the next generation matrix. It was shown that individual breeding value for R_0 is a function of gene effects on susceptibility and infectivity. Subsequently, heritable variation in R_0, which reflects the potential of a population to respond to selection, equals the variance in breeding values for R_0 among individuals. The results show that heritable variation in susceptibility and infectivity contribute to the heritable variation R_0.

Estimating heritable variation in infectivity has been difficult, since an individual's infectivity does not surface in its own disease status, but in the disease status of other individuals that come into contact with the focal individual. Recent advances in the field of quantitative genetics have, however, opened the door to consider the genetic effect of an individual on trait values of other individuals, which is known as an indirect genetic effect (IGEs). An IGE is a heritable effect of an individual on the trait value of another individual. With regard to the transmission of infectious diseases, infectivity, which is an individual's propensity to infect its contacts, can be considered as an indirect genetic effect. Anche et al. (2014) also found that, not only infectivity, but also susceptibility exhibits an indirect genetic effect. This occurs because a highly susceptible individual will have a higher probability to be infected, which increases the probability that highly susceptible individuals will infect others.

Estimation of genetic parameters for infectivity is more complicated than for ordinary traits affected by IGE (Lipschutz-Powell et al., 2012) This is because an

individual's infectivity is expressed only when the individual is infected itself, and because infectious diseases show non-linear dynamics over time. The ordinary linear mixed models used for traits affected by IGE in quantitative genetics do not take the non-linear and stochastic nature of infectious disease dynamics into account, and ignore the conditional expression of IGE (Muir 2005; Bijma et al. 2007b). Thus those models need to be extended.

Anacleto et al. (2015) developed Bayesian methodology to estimate genetic (co)variances and breeding values based on disease status data recorded over time. However, time-series data on disease status may also not always be available. Hence, there is a need for methods that can be implemented using a single record on an individual's disease status. (Velthuis et al., 2003; Lipschutz-Powell et al., 2012; Anche et al., 2015) have shown that binary disease-status data can be analysed using a generalized linear model with a complementary log-log link function. However, they have not investigated the estimation of genetic variance components and breeding values with such models.

Thus, the aim of this study is to develop methodology for estimating individual breeding values and variance components for susceptibility and infectivity. We will develop a generalized linear mixed model to estimate those variance components and breeding values from the final disease status of individuals in populations that have undergone an epidemic.

Furthermore, previous studies have shown that relatedness among group mates affects the accuracy of estimating breeding values and variance components for traits affected by IGE (Griffing, 1967, 1976, 1981; Bijma and Wade, 2008; Bijma, 2010; Wade et al., 2010). Hence, we also aim to investigate the effect of relatedness among groupmates on the bias and accuracy of estimated breeding values and variance components. For that purpose, we simulated epidemics in genetically heterogeneous populations, where individuals differ genetically in their susceptibility and infectivity, and recorded the final disease status (0/1) of individuals.

5.2 Material and methods

5.2.1 Simulated population and scenarios

We simulated a paternal half-sib family structure, where the parents were unrelated. The population consisted of N = 10,000 individuals, from 100 sires with 100 offspring per sire, where each offspring had a unique dam. The population was sub-divided into 100 groups, each group consisting of 100 individuals.

We hypothesized that the degree of relatedness among group mates will affect the accuracy of estimated variance components and breeding values for susceptibility

and infectivity. To test this hypothesis, we simulated scenarios where the average degree of relatedness among group mates varied. This was done by varying the number of sires that contributed offspring to each group. We simulated six scenarios, where individuals were either allocated to groups at random, or the number of sires contributing to a group was 10, 5, 4, 2 or 1 (Table 5.1).

Parental transmitting abilities for susceptibility (γ) and infectivity (φ), which are half of the breeding values, were drawn from a bivariate normal distribution with mean zero and variance = ¼*$\begin{bmatrix} \sigma^2_{A_\gamma} & \sigma_{A_\gamma,A_\varphi} \\ \sigma_{A_\gamma,A_\varphi} & \sigma^2_{A_\varphi} \end{bmatrix}$, where $\sigma^2_{A_\gamma} = 0.04$ is variance in breeding values for susceptibility (A_γ), $\sigma^2_{A_\varphi} = 0.04$ is variance in breeding values in infectivity (A_φ), and $\sigma_{A_\gamma,A_\varphi}$ is the covariance between breeding values for susceptibility and infectivity. Offspring phenotypes for susceptibility and infectivity were obtained by adding a Mendelian sampling deviation to the transmitting ability of both parents, and subsequently adding the population mean susceptibility and infectivity, $\bar{\gamma} = \bar{\phi} = 1.6.$ With these inputs, the average susceptibility and infectivity were 4 standard deviations away from zero. Hence, occasionally it occurred that negative values for susceptibility and infectivity were sampled. Whenever this happened, those values were set to zero. Non-genetic variance for susceptibility and infectivity was not simulated, because the sampling of the event (infection or recovery, see below) introduces the required noise.

To examine the impact of the genetic correlation between susceptibility and infectivity on the accuracy of estimated breeding values and variance components, we simulated scenarios where the genetic correlation between susceptibility and infectivity was either zero, 0.5, or -0.5. This was done for all the scenarios considered (See Table 1).

Table 5.1. Input values and scenarios

Scenarios	Number of sires
1st scenario	Random[1]
2nd scenario	10 sires
3rd scenario	5 sires
4th scenario	4 sires
5th scenario	2 sires
6th scenario	1 sire

Note: scenarios indicate the number of sires contributing offspring to each group in the population and for all the scenarios, variance in susceptibility and infectivity was 0.04. Contact rate c = 2 and recovery rate α = 2 was also used for all the scenarios. 1 With random allocation of sires to groups, the expected number of distinct sires per group equals 63.2.

5.2.2 Epidemiological model

The dynamics of an infection in the population were simulated using the stochastic SIR-model. In this model, the possible events that can occur are infection of a susceptible individual or recovery of infected individual. With stochasticity, these events occur with a certain rate defined by two model parameters. The first parameter is the transmission rate parameter β , for the transmission of a susceptible individual to become infected, S → I, with a rate $\beta \frac{SI}{N}$, where *N* denotes population size, *S* the number of susceptible individuals and *I* the number of infectious individuals. Hence, symbols *S*, *I* and *R* denote both the disease status and the number of individuals with this disease status, and $S + I + R = N$. The other parameter is the recovery rate parameter for the transition of an infected individual to the recovery state, I → R, with a rate αI.

The transmission rate parameter is the probability per unit of time for one infected individual to infect any other individual in a totally susceptible population. When we have a genetically heterogeneous population, where individuals differ in their susceptibility and infectivity, the transmission rate parameter varies between pairs of individuals. For a pair of individuals, the transmission rate will depend on the breeding value for susceptibility of the susceptible individual, the breeding value for infectivity of the infectious individual and the average contact rate. The assumption that transmission rate depends only on susceptibility of the susceptible individual and infectivity of the infectious individual, but not on the combination of both traits, is known as separable mixing (Diekmann et al., 1990). Thus, the transmission rate of a specific susceptible individual *i* with breeding value for

susceptibility $A_{\gamma,i}$ from being susceptible to being infected when exposed to a single infectious individual *j* with breeding value for infectivity $A_{\varphi,j}$ can be defined as:

$$\beta_{ij}\frac{1}{N} = A_{\gamma,i}A_{\varphi,j}c\frac{1}{N} \qquad [1]$$

where *c* is the average contact rate of an individual with an arbitrary other individual. Thus $\frac{c}{N}$ is the average contact rate of a susceptible individual with a single specific infectious individual in a group of size *N*. We did not simulate genetic heterogeneity in the recovery rate. Hence, there was a single recovery rate parameters, α, that applied to all individuals.

In each group, an epidemic was started by one randomly infected individual. The type of the next event was then determined by using Gillespie's direct algorithm (Gillespie, 1977). The type of event, i.e. either infection or recovery, was decided by drawing a random number v_1, from a uniform distribution, $v_1 \sim U(0,1)$. The next event was an infection of a susceptible individual if the random number

$$v_1 < \frac{\frac{1}{N}\sum_{i=}^{S}\sum_{j=1}^{I}\beta_{ij}}{\frac{1}{N}\sum_{i=}^{S}\sum_{j=1}^{I}\beta_{ij} + \alpha I}$$

otherwise it was a recovery of a random infected individual. The numerator of this ratio represents the total infection rate, and the denominator the total rate, i.e., the sum of the infection and recovery rates. The sampling of the specific individual that became infected depended on individual susceptibility. The probability that susceptible individual i became infected was proportional to its breeding value for susceptibility $\boldsymbol{A_{\gamma,i}}$. The number of susceptible S, and infectious I, individuals changes through the progression of the epidemic and thus the rate is updated accordingly. The epidemic was allowed to run until there was either no infectious individual anymore or no susceptible individual left to be infected in the population. In each group, each individual was assigned a phenotype of 1, if it had been infected, and 0 otherwise. As a result, we have binary data by the end of the epidemic. The simulation of the population and the epidemic process was done in R-statistical software.

5.2.3. Statistical Model

To estimate the relative effects of single genes on susceptibility and infectivity, Anche et al. (2015) presented a generalized linear model with a complementary-log-log link function, where gene effects were estimated as fixed effect. Here, we

develop the analogous mixed model, where we have random genetic effects for entire animals, and a fixed overall mean effect.

Consider the probability P_{ij} that individual *i* escapes infection when exposed to infectious individual j. From the zero term of the Poisson distribution,

$$P_{ij} = e^{-\gamma_i \varphi_j \frac{c}{\alpha} \frac{1}{N}}. \qquad [2]$$

where γ_i is susceptibility of focal individual *i*, and φ_j is infectivity of its infectious group mate *j*. The probability that individual *i* is still susceptible at the end of the epidemic, is the product of all the probabilities that it escapes infection exposures from each of its I infectious group mates,

$$P_{i,I_j} = \prod_{j=1}^{i} e^{-\gamma_i \varphi_j \frac{c}{\alpha} \frac{1}{N}} = e^{-\gamma_i \frac{c}{\alpha} \frac{1}{N} \sum_{j=1}^{I} \varphi_j} \qquad [3]$$

where the summation is over the I infected group mates of focal individual *i*.

This result is analogous to equation 3 in Anche et al. (2015), who considered three categories of infectious individuals in a population, each corresponding to a genotype at a single bi-allelic locus. In this study, however, each individual has a unique breeding value. Thus we cannot categorize individuals, and the summation is over all infectious group mates of an individual. The complementary log-log of (1-P_i) for the P_i given by Equation 3 is

$$\log(-\log(P_i)) = \log\left(\frac{c}{\alpha}\right) + \log(\gamma_i) + \log\left(\frac{1}{N}\right) + \log\left(\sum_{j=1}^{I} \varphi_j\right). \qquad [4]$$

The objective is to implement Equation 4, treating susceptibility and infectivity as random variables. In ordinary random-effect models, the expectation of a random variable is zero (Lynch and Walsh, 1998). In Equation 4, however, γ and φ include the average susceptibility and infectivity, which violates the model assumptions. Thus, to make the expectation of random effects zero, we need to partition γ and φ into a population average and a term that has expectation zero. This can be achieved by defining breeding values, say a_γ and a_φ, on a multiplicative scale, so that

$$\gamma_i = \bar{\gamma}(1 + a_{\gamma,i}) \qquad (5a)$$

and

$$\varphi_i = \bar{\varphi}(1 + a_{\varphi,i}), \qquad (5b)$$

so that $E[a_\gamma] = E[a_\varphi] = 0.$ Hence, the relationship between the (ordinary) additive breeding values and the multiplicative breeding values is as follows, $A_{\gamma,i} = \bar{\gamma} a_{\gamma,i}$, and $A_{\varphi,i} = \bar{\varphi} a_{\varphi,i}$.

Thus,

$$\log(\gamma_i) = \log(\bar{\gamma}(1 + a_{\gamma,i})) = \log(\bar{\gamma}) + \log(1 + a_{\gamma,i}) \approx \log(\bar{\gamma}) + a_{\gamma,i} \quad [6]$$

The last step assumes that *a* is relatively small compared to 1.

Analogously:

$$\log\left(\sum_{j=1}^{I} \varphi_j\right) = \log\left(\sum_{j=1}^{I} \bar{\varphi}\,(1 + a_{\varphi,j})\right) = \log(\bar{\varphi}\,) + \log\left(\sum_{j=1}^{I}(1 + a_{\varphi,j})\right), \quad [7]$$

and:

$$\log\left(\sum_{j=1}^{I}(1 + a_{\varphi,j})\right) = \log\left(I\left(1 + \bar{a}_{\varphi}\right)\right) = \log(I) + \log\left(1 + \bar{a}_{\varphi}\right) \approx \log(I) + \bar{a}_{\varphi} \quad [8]$$

where $\bar{a}_{\varphi}$ is average (multiplicative) breeding value over the infected group mates of individual *i*, $\bar{a}_{\varphi} = \frac{1}{I}\sum_{j=1}^{I} a_{\varphi,j}$

Finally, substituting Equations 6 and 7 into Equation 4 yields

$$\log(-\log(P_i)) = \log\left(\frac{c}{\alpha}\right) + \left\{\log(\bar{\gamma}) + a_{\gamma,i}\right\} + \left\{\log(\bar{\varphi}\,) + \log(I) + \frac{1}{I}\sum_{j=1}^{I} a_{\varphi,j}\right\} + \log\left(\frac{1}{N}\right) \quad [9]$$

Rearranging Equation 8 gives,

$$\log(-\log(P_i)) = \log\left(\frac{c}{\alpha}\bar{\gamma}\bar{\varphi}\right) + a_{\gamma,i} + \frac{1}{I}\sum_{j=1}^{I} a_{\varphi,j} + \log\left(\frac{I}{N}\right) \quad [10]$$

Equation 9 indicates that (co)variances for susceptibility and infectivity can be estimated using a generalized linear mixed model, with a fixed intercept, random genetic effects for susceptibility and infectivity, and an offset equal to $\log\left(\frac{I}{N}\right)$.

Thus, the following generalized linear mixed model (GLMM) with random genetic effects for susceptibility and infectivity was fitted,

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z_1}\boldsymbol{a_\gamma} + \mathbf{Z_2}\boldsymbol{a_\varphi} + \mathbf{Vg} + \mathbf{offset} + \mathbf{e} \quad [11]$$

where $\mathbf{y}$ is vector of observations on disease status, containing 0s for individuals that are still susceptible at the end of the epidemic, and 1s for individuals that have been infected in the course of the epidemic. $\mathbf{1}$ is a vector of ones for the mean μ, which is an estimator of $\log(\frac{c}{\alpha}\bar{\gamma}\bar{\varphi}\,)$. Since the basic reproduction ratio for an SIR model with heterogeneity in susceptibility and infectivity is given by $R_0 = \frac{c}{\alpha}\bar{\gamma}\,\bar{\varphi}$ (Anche et al. 2014), $e^{\hat{\mu}}$ is an estimate of R_0. $\boldsymbol{Z_1}$ is a diagonal incidence matrix for the random genetic effect of susceptibility (a_γ)and $\boldsymbol{Z_2}$ is an incidence matrix for the random genetic effects for infectivity (a_φ) of the infected group mates of a focal individual. In the row of focal individual *i*, $\boldsymbol{Z_2}$ has off-diagonal elements equal to $1/I$ in the columns for each infected group mate of *i*. Thus, in contrast to ordinary IGE-models (Muir, 2005), $\mathbf{Z}_2$ has entries only for those group mates that have been infected. $\boldsymbol{a_\gamma}$ is a vector of random

genetic effects for susceptibility, and $\boldsymbol{a_\gamma}$ is a vector of random genetic effects for infectivity. In addition to random genetic effects, the GLM contained a random group effect, denoted by **Vg**, where $\mathbf{V}$ is an incidence matrix linking records to groups and $\mathbf{g}$ is a vector of iid (identical and independently distributed) random group effects. The random group-effect was included to account for covariance among group mates due to sampling, which is not completely accounted for by the offset log(I/N).

The (co)variance structure of the random genetic terms was $Var\begin{bmatrix} a_\gamma \\ a_\varphi \end{bmatrix} = \mathbf{C} \otimes \mathbf{A}$ where $\otimes$ indicates the Kronecker product of both matrices, $\mathbf{C} = \begin{bmatrix} \sigma^2_{a_\gamma} & \sigma_{a_\gamma, a_\varphi} \\ \sigma_{a_\gamma, a_\varphi} & \sigma^2_{a_\varphi} \end{bmatrix}$ and **A** is the additive genetic relationship matrix (Lynch and Walsh, 1998).

The GLMM was fitted with a binomial distribution and complementary log-log link function in the ASReml 3.0 statistical software (Gilmour et al., 2009).

Since the first randomly chosen individual that started the epidemic does not express its susceptibility, this individual was excluded from the analysis as dependent variable but not as explanatory variable.

The above mixed model produces estimates of breeding values and variances that refer to the multiplicative breeding value, *a*, rather than the simulated additive breeding value, *A*. Thus, to investigate the quality of the estimates in terms of their accuracies and biases, estimated breeding values and variance components were back-transformed to the additive scale. Estimated additive breeding values were

$$\hat{A}_{\gamma,i} = \overline{\gamma}\,\hat{a}_{\gamma,i}\,, \qquad [12]$$

and

$$\hat{A}_{\varphi,i} = \overline{\varphi}\,\hat{a}_{\varphi,i}\,. \qquad [13]$$

Analogously, estimated additive genetic variances were

$$\hat{\sigma}^2_{A_\gamma} = \overline{\gamma}^2 \hat{\sigma}^2_{a_\gamma},$$

and

$$\hat{\sigma}^2_{A_\varphi} = \overline{\varphi}^2 \hat{\sigma}^2_{a_\varphi}.$$

5.3 Results

5.3.1 Bias and precision of estimated genetic variances

Results presented here are from 1000 replicates. Biases in estimated genetic variances were calculated as the difference between the true value and the average estimated value.

Figure 5.1A shows the bias in the estimated variance in susceptibility for the case where the correlation between susceptibility and infectivity was zero. Results show a little over-estimation, except for the case where all offspring in a group descend from the same sire, in which case the variance in susceptibility is was not estimable. Hence, relatedness, as measured by the number of sires contributing offspring to a group, had little effect on the bias.

When the genetic correlation between susceptibility and infectivity was positive, the bias in the estimated variance in susceptibility was a little higher than when no correlation was assumed, but still relatively small (Figure 5.1B). When the genetic correlation between susceptibility and infectivity was negative, the direction of the bias depended on relatedness between group mates; over-estimation was observed for groups composed at random with respect to relatedness, while underestimation was observed for groups composed of 5 or fewer sires (Figure 5.1C).

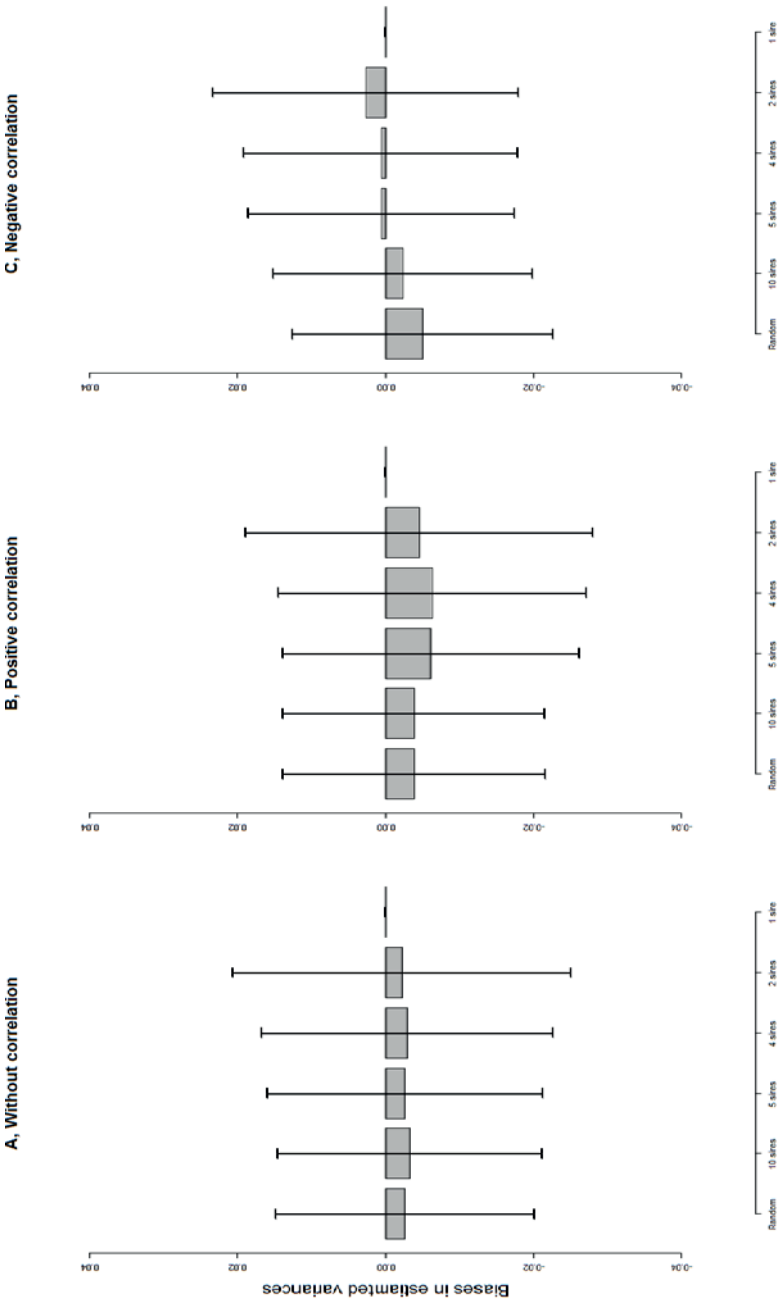
A, Without correlation
B, Positive correlation
C, Negative correlation
Biases in estiamted vatiances
0.04
0.02
0.00
-0.02
-0.04
Random
10 sires
5 sires
4 sires
2 sires
1 sire

Figure 5.1. Box-whisker plots of bias in estimated variance in susceptibility, for an increasing degree of relatedness among group mates, and for different genetic correlations (0, 0.5 and -0.5) between susceptibility and infectivity. Values are true values minus estimates, so positive values indicate under-estimation. The true variance was 0.04. Error-bars indicate the standard deviation of the estimate among replicates. In the scenario with one sire per group, the variance in susceptibility was not identifiable. Hence, the result from this scenario is presented as not estimable (NE).

Bias was substantially greater for infectivity than for susceptibility (Figure 5.1 vs 5.2). For most of the scenarios simulated, variance in infectivity was overestimated (Figure 5.2). When all group mates descended from a single sire, the variance in infectivity was not identifiable. For groups consisting of offspring of at least two sires, the bias showed a strong relationship with the relatedness among group mates. The genetic variance in infectivity was severely over-estimated when group mates were unrelated, irrespective of the genetic correlation between susceptibility and infectivity. When the genetic correlation between susceptibility and infectivity was zero, the bias decreased with increasing relatedness, and became near zero for groups composed of offspring of two sires only (Figure 5.2A). When the genetic correlation between susceptibility and infectivity was positive, the bias also decreased with relatedness, but considerable overestimation remained for groups descending from only two sires (Figure 5.2B). When the genetic correlation between susceptibility and infectivity was negative, the bias changed sign when relatedness increased, resulting in underestimation for groups descending from only two sires (Figure 5.2C).

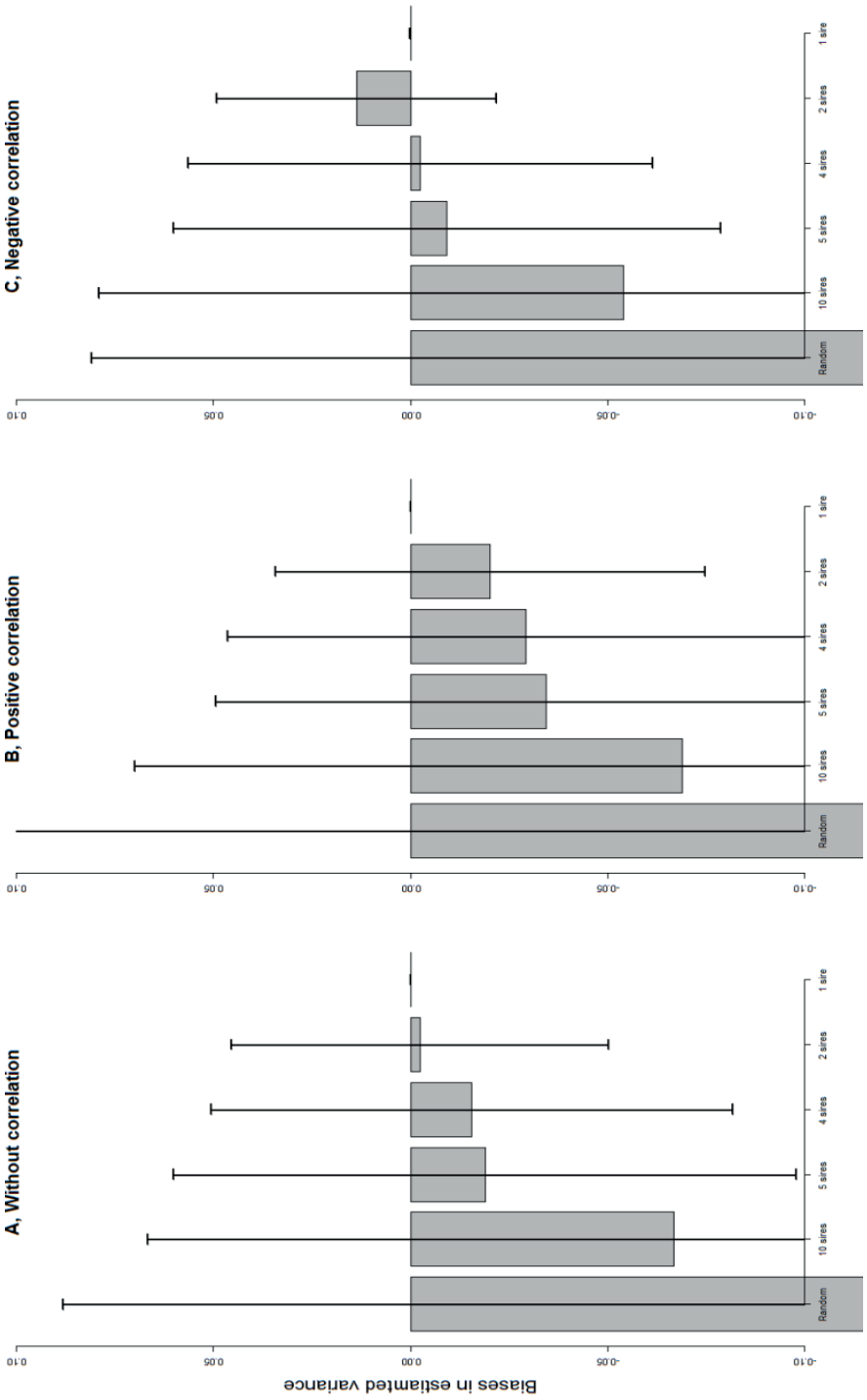
C, Negative correlation
B, Positive correlation
A, Without correlation
1 sire
2 sires
4 sires
5 sires
10 sires
Random
-0.10
-0.05
0.00
0.05
0.10
Biases in estiamted variance

Figure 5.2. Box-whisker plots of bias in estimated variance in infectivity, for an increasing degree of relatedness among group mates, and for different genetic correlations (0, 0.5 and -0.5) between susceptibility and infectivity. Values are true values minus estimates, so positive values indicate under-estimation. The true variance was 0.04. Error-bars indicate the standard deviation of the estimate among replicates. In the scenario with one sire per group, the variance in infectivity was not estimable (NE). The arrow for the random group composition serves to indicate that the bar extends much further down.

5.3.2 Bias and accuracy of estimated breeding values

Results presented in this section are from 50 replicates. The bias of estimated breeding values (EBV) was calculated as the regression coefficient of the true breeding values on the estimated breeding values, where a value of one indicates absence of bias. With Best Linear Unbiased Prediction, the regression coefficient of the true values on the estimates should equal one for random effects; (Henderson, 1975). The accuracies of estimated breeding values were calculated as the correlation between true and estimated breeding values. For the estimation of breeding values, the variances and co-variances of susceptibility and infectivity were fixed to their true (i.e. used in the simulation) values. Hence, results show the quality of EBV for the case where the genetic (co)variances are known, rather than estimated.

Figure 5.3 shows the regression coefficient of the true breeding value for susceptibility on the estimated breeding value. Regression coefficients are systematically smaller than one, indicating that the variance in EBV is too large. In other words, EBV over-predict the differences among sires in true breeding values for susceptibility. Regression coefficients did not show a clear relationship with relatedness among group mates, nor with the genetic correlation between susceptibility and infectivity.

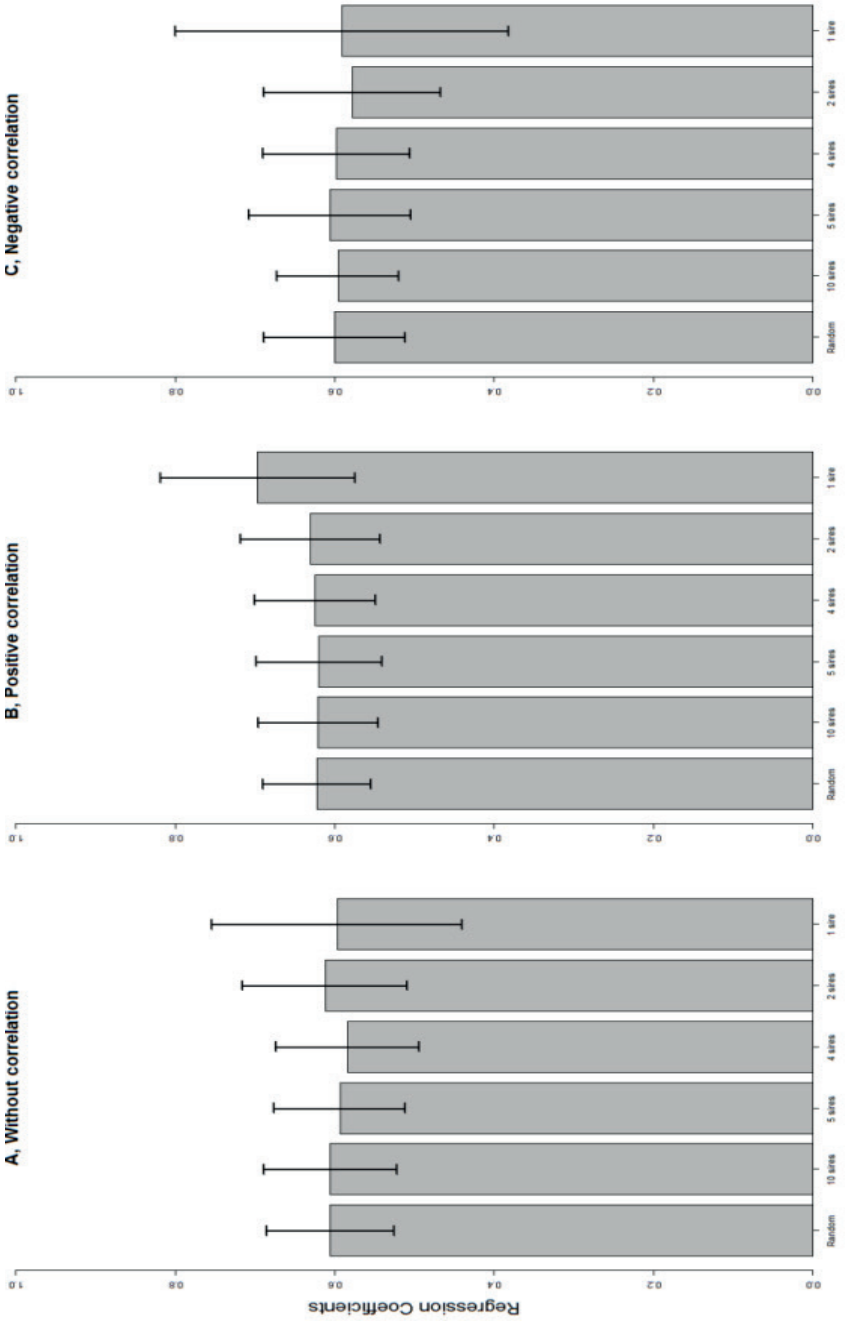
C, Negative correlation
B, Positive correlation
A, Without correlation
Regression Coefficients
0.0
0.2
0.4
0.6
0.8
1.0
1 sire
2 sires
4 sires
5 sires
10 sires
Random

Figure 5.3. Box and whisker plots for the regression coefficients of true breeding values of sires on their EBV, for susceptibility, for an increasing degree of relatedness among group mates, and for different genetic correlations (0, 0.5 and -0.5) between susceptibility and infectivity. Error bars denote standard deviations among the estimated regression coefficients.

Figure 5.4 shows the corresponding results for infectivity. Also for infectivity, regression coefficients are systematically smaller than one, and tend to be lower than those for susceptibility. Thus, EBV for infectivity over-predict the differences among sires in true breeding values, and somewhat more than for susceptibility. When the genetic correlation between susceptibility and infectivity was zero, the over-prediction became smaller when relatedness among group mates increased, whereas the opposite was observed for a negative correlation between susceptibility and infectivity.

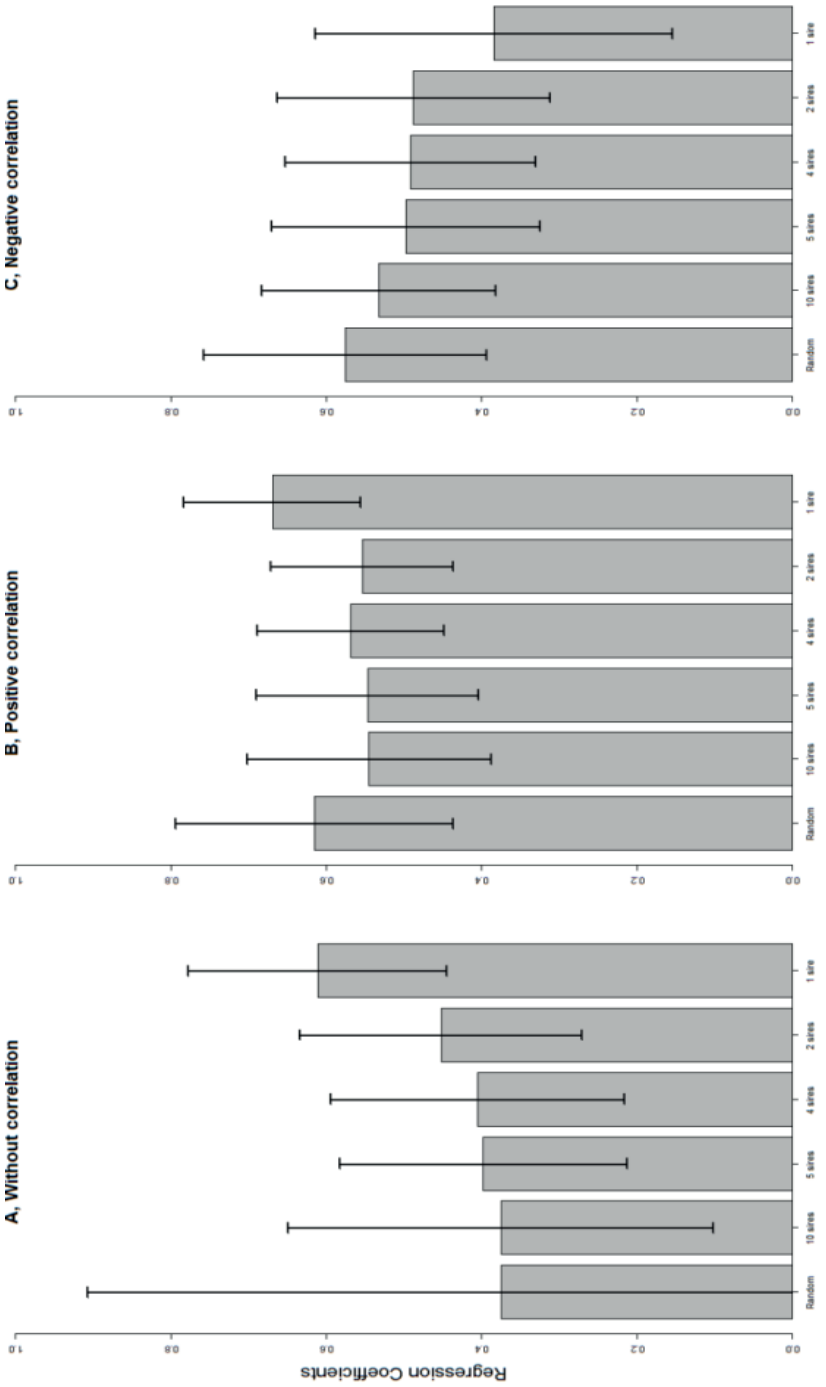
A, Without correlation
B, Positive correlation
C, Negative correlation
Regression Coefficients
0.0
0.2
0.4
0.6
0.8
1.0
Random
10 sires
5 sires
4 sires
2 sires
1 sire

Figure 5.4 Box and whisker plots for the regression coefficients of true breeding values of sires on their EBV, for infectivity, for an increasing degree of relatedness among group mates, and for different genetic correlations between susceptibility and infectivity. Error bars denote standard deviation among the estimated regression coefficients.

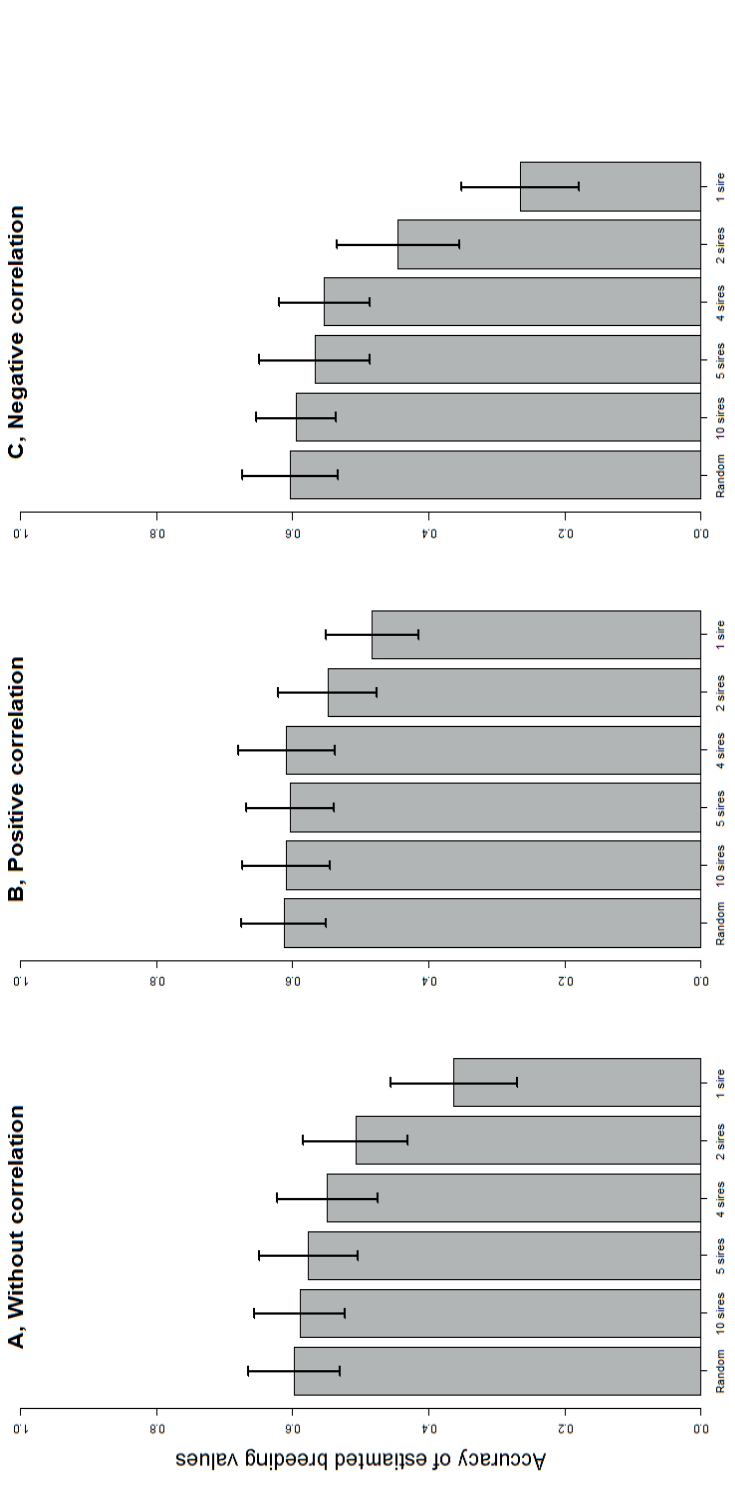

Figure 5.5. Box and whisker plots for the accuracy of EBVs for susceptibility, for an increasing degree of relatedness among group mates, and for different genetic correlations between susceptibility and infectivity. Error bars denote the standard deviations of the estimated accuracies.

For all scenarios, the accuracy of EBVs was greater for susceptibility than for infectivity (Figure 5.5 *vs.* 5.6). Accuracies of EBVs for susceptibility were moderate, and decreased with relatedness among group mates, particularly when the number of sires per group became small (Figure 5.5). The decrease of accuracy with relatedness was strongest when the correlation between susceptibility and infectivity was negative.

Accuracies of EBVs for infectivity showed the opposite trend, and increased with relatedness among group mates in most cases (Figure 5.6). When the genetic correlation between susceptibility and infectivity was zero, accuracy was near zero when group members were unrelated, and increased to moderate values when group mates descended from a single sire (Figure 5.6A). When the genetic correlation was positive, accuracies were higher than for a correlation of zero, and increased only a little with relatedness (Figure 5.6B). When the genetic correlation was negative, accuracies were relatively independent of relatedness, except when the number of sires per group was small (Figure 5.6C).

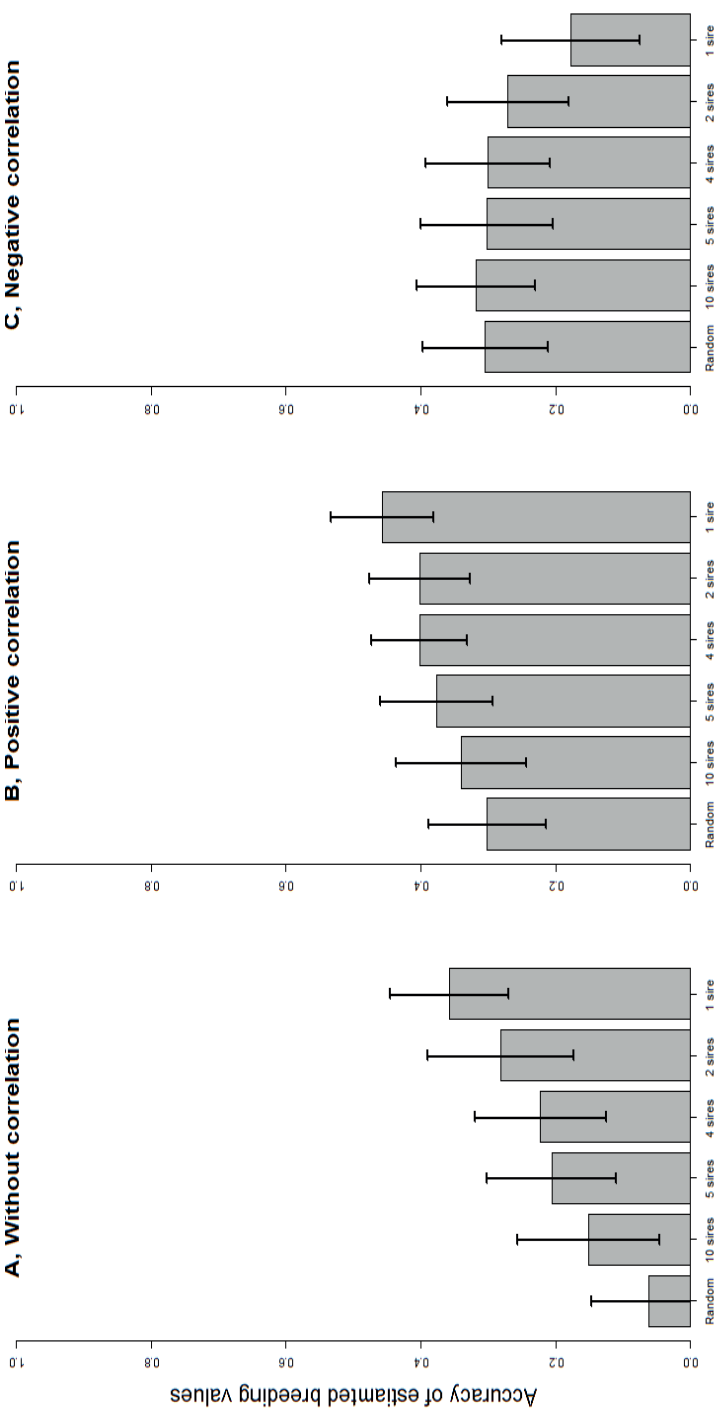

Figure 5.6. Accuracy of EBVs for infectivity, for an increasing degree of relatedness among group mates, and for different genetic correlations between susceptibility and infectivity. Error bars denote standard deviations of the accuracies.

5.4 Discussion

In this study, a generalized linear mixed model (GLMM) was developed to estimate genetic (co)variances and breeding values for susceptibility and infectivity from data on the final disease status of individuals at the end of epidemics. The model was developed from an equation that describes the probability that an individual has been infected as a function of its own breeding value for susceptibility and the breeding values for infectivity of its infected group mates. This model was developed following (Anche et al., 2015), who presented a generalized linear model (GLM) where genetic effects for susceptibility and infectivity were treated as fixed effect (See also Velthuis et al. (2003) and Lipschutz-Powell et al. (2014)). In Anche et al. 2015, it was shown that the estimation of the genetic effects as fixed effects could be done satisfactorily with the GLM as proposed in that paper.

However, breeding values should ideally be estimated as random effects, and thus a simulation study was performed to investigate the quality of the GLMM in terms of the bias and precision of the estimated genetic co(variances), and the bias and accuracy of the predicted breeding values. We also investigated the effect of relatedness among group mates on those quality measures. Results showed a small bias in the estimated genetic variance for susceptibility, except for groups consisting of a single family for which the variance in susceptibility clearly cannot be estimated when a (random) group-effect is included in the model.

In contrast, the genetic variance in infectivity was severely overestimated, particularly when relatedness among group mates was absent or low. When the true genetic (co)variances were used in the breeding value estimation, EBV for susceptibility showed moderate accuracy, while EBV for infectivity showed low to moderate accuracy. When the true genetic (co)variances were used, regression coefficients of true breeding values on EBV were considerably smaller than one, indicating that EBV over-predict the differences in the true breeding values.

We investigated the cause of the severe overestimation of the genetic variance in infectivity, and the reason why this overestimation decreased so substantially when relatedness among group mates increased (See Figure 5.2).

First we analysed the fit of the model in the absence of genetic variation. Given the values of c, α, $\bar{\gamma}$ and $\overline{\varphi}$, we numerically solved the so-called final size equation (Andreasen, 2011) to obtain the probability that an individual escapes infection. Then we simulated the final disease status (0,1) of individuals from a Bernoulli distribution with this probability. Subsequently, we observed the fraction of individuals infected during the outbreak in each group, I/N, and calculated the

probability, $P_{y|\frac{I}{N}}$, that an individual escapes infection using Equation 3. Note that Equation 3 uses the *observed* fraction of infecteds in each group. Finally, we calculated the variance in diseases status around its expected value, $P_{y|\frac{I}{N}}$. From the Bernoulli distribution, the expected variance in disease status equals $P_{y|\frac{I}{N}}(1-P_{y|\frac{I}{N}})$. However, the observed variance was only around half this value. This suggest that the model over-fits the data when the observed fraction of infecteds in a group is used as an explanatory variable.

Thus, we hypothesized that the over-fit also occurs in the linear model (Equation 10), because the offset, $\log(I/N)$, is taken from the observed data, which leads to an excessive correlation between the dependent variable and the offset in the prediction equation. In other words, while the model intends to specify the *probability* that an individual gets infected, the *realized* fraction of infected individuals is used as offset in the model. This inflates the predictive ability of the model.

The previous argument has focussed on the use of the offset and the effect on the variance of the residual. However, a similar phenomenon may occur with the infectivity term in the model, because the incidence matrix for infectivity is based on the realized number of infected group mates of an individual, i.e., the $\mathbf{Z}_2$ matrix in Equation 11 has an entry for each infected group mate. Thus, having established that the GLMM procedure as used in this paper overestimates the genetic variance in infectivity (see Results), we tested whether this was due to the fact that the explanatory variables for infectivity showed an excessive correlation with the observed disease status. We did this by artificially constructing a data set where this excessive correlation was not present. (Note that this cannot occur in any real dataset). In short, we tested whether the use of the observed number of infecteds as explanatory variable in the model causes the over-estimation of the variance in infectivity. For this purpose, we simulated data with genetic variation, but sampled the number and identity of the infectious individuals in each group at random. Subsequently, for each individual, we calculated the probability that it escapes infection from Equation 3. Thus, an individual's probability to escape infection depended on its own susceptibility, and on the infectivity of its randomly sampled infectious group mates. Then we sampled the disease status (0,1) of each individual according to this probability. We used this latter disease status as the dependent variable, rather than the initial disease status that was sampled at random. Finally, we analysed the data with the model in Equation 10, using the infected individuals

that were *a priori* sampled at random to create the incidence matrix for infectivity ($\mathbf{Z}_2$) and the offset, while using the disease status sampled from Equation 3 as the dependent variable. Hence, with this set-up, the excessive correlation between the dependent variable and the explanatory variables in Equation 11 is removed; the disease status of group members is sampled depending on I/N, but the resulting fraction of infecteds in the group may deviate by chance from this I/N. The resulting estimates of the genetic (co)variances showed no bias (results not shown), and the estimated variance of the group effect was practically zero. Hence, this result illustrates that the over-estimation of the genetic variance in infectivity occurs because the number of infected group mates in the explanatory variable for infectivity is taken from the dependent variable. This leads to an excessive correlation between the dependent variable and explanatory variable for infectivity, which causes the infectivity term in the model to absorb too much variance.

The above does not yet explain why the over-estimation of the genetic variance in infectivity decreases so sharply with relatedness among group mates. However, consider the variance due to the infectivity term in the linear model (Equation 10), for simplicity assuming a uniform relatedness (*r*) between group mates. This yields

$$\mathrm{Var}\left(\tfrac{1}{I}\textstyle\sum_{j=1}^{I} a_{\varphi,j}\right) = \sigma^2_{a_\varphi}\left(1+(I-1)r\right)/I$$

Because I was large in our simulations (~90), the $(1+(I-1)r)$ term increases very strongly with relatedness. Now suppose that the excessive-correlation phenomenon explained above creates a certain covariance, say *C*, between the dependent variable and the infectivity-term. In other words,

$$\mathrm{Var}\left(\frac{1}{I}\sum_{j=1}^{I} a_{\varphi,j}\right) = C$$

The "required" value of $\sigma^2_{a_\varphi}$ to accommodate this covariance equals $CI/(1+(I-1)r)$, which decreases strongly with *r*. In other words, when group mates are more related, a smaller $\sigma^2_{a_\varphi}$ is needed to account for a certain covariance. We think this explains why the over-estimation of the variance in infectivity decreases so sharply with relatedness among group mates.

As mentioned in the results, variance in susceptibility and infectivity was not estimable for the scenario when offspring are from one sire. In this case, the sire variance is completely confounded with the group variance. When the correlation between susceptibility and infectivity was positive, accuracies of estimated

breeding values were higher for both traits. The same result was found by (Lipschutz-Powell et al., 2012). This gain in accuracy could be due to the same reason as given by (Lipschutz-Powell et al., 2012), namely that the covariance contributes to the accuracy of estimated breeding values.

Anacleto et al. (2015) developed a Bayesian method to estimate genetic (co)variances for susceptibility and infectivity from time-interval data on binary disease status. In their disease model, infected individuals did not recover, but stayed infectious for the rest of their life (SI-Model). They found that genetic (co)variances for susceptibility and infectivity can be estimated accurately. Having a time-interval disease data improves the accuracy of estimating genetic effect for susceptibility and infectivity as it gives information on the order of infection, that is, on who infected whom (Pooley et al., 2014; Anche et al., 2015). For a short interval, the fraction of infectious individuals during the interval (explanatory variable) is not correlated with the fraction of new cases within that interval (dependent variable). In the statistical analysis, this is achieved by using the infected individuals at the beginning of the interval as explanatory variable and the new cases during the interval as dependent variable. If the interval is longer, the cases that occur during the interval will also contribute to the new cases within the interval and thus that information should ideally be used also in the explanatory variable. However, this introduces the excessive correlation phenomenon explained above, at least to some degree. The final size observation as used here is the extreme case where the whole epidemic is contained within the interval, and where the information about the first infected individual is not very informative as explanatory variable for the observed cases during the epidemic. Thus we used the infected individuals observed at the end of the interval as explanatory variables in the analysis. The latter approach seems to be ok for the estimation of fixed effects of genes (Anche et al. 2015), but not for the estimation of the random genetic effects for infectivity.

In summary, we considered the case where only the final disease status is known at the end of an epidemic. We believe this is the major cause of the difference in the quality of estimated genetic (co)variances and breeding values between both studies, rather than the difference in methodology (i.e. a Bayesian framework vs. a GLMM). This suggests that time-interval disease data provide better estimates of genetic parameters and breeding values and thus should be considered in the collection of data on infectious diseases.

5.6 References

Anacleto, O., L. A. Garcia-Cortés, D. Lipschutz-Powell, J. A. Woolliams, and A. B. Doeschl-Wilson. 2015. A Novel Statistical Model to Estimate Host Genetic Effects Affecting Disease Transmission. Genetics 201: 871-884.

Anche, M., M. de Jong, and P. Bijma. 2014. On the definition and utilization of heritable variation among hosts in reproduction ratio R0 for infectious diseases. Heredity 113: 364-374.

Anche, M. T., P. Bijma, and M. De Jong. 2015. Genetic analysis of infectious diseases: estimating gene effects for susceptibility and infectivity. Genet Sel Evol 47: 1-15.

Anderson, R. M., R. M. May, and B. Anderson. 1992. Infectious diseases of humans: dynamics and control. Wiley Online Library.

Andreasen, V. 2011. The final size of an epidemic and its relation to the basic reproduction number. Bulletin of mathematical biology 73: 2305-2321.

Axford, R., S. Bishop, F. Nicholas, and J. Owen. 2000. Breeding for disease resistance in farm animals. CABI publishing.

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

Bijma, P., and M. Wade. 2008. The joint effects of kin, multilevel selection and indirect genetic effects on response to genetic selection. Journal of evolutionary biology 21: 1175-1188.

Denison, R. F., E. T. Kiers, and S. A. West. 2003. Darwinian agriculture: when can humans find solutions beyond the reach of natural selection? The quarterly review of biology 78: 145-168.

Diekmann, O., J. Heesterbeek, and J. A. Metz. 1990. On the definition and the computation of the basic reproduction ratio R 0 in models for infectious diseases in heterogeneous populations. Journal of mathematical biology 28: 365-382.

Gillespie, D. T. 1977. Exact Stochastic Simulation of Coupled Chemical-Reactions. J Phys Chem-Us 81: 2340-2361.

Gilmour, A. R., B. Gogel, B. Cullis, R. Thompson, and D. Butler. 2009. ASReml user guide release 3.0. VSN International Ltd, Hemel Hempstead, UK.

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-&.

Griffing, B. 1976. Selection in Reference to Biological Groups .5. Analysis of Full-Sib Groups. Genetics 82: 703-722.

Griffing, B. 1981. A Theory of Natural-Selection Incorporating Interaction among Individuals .2. Use of Related Groups. J Theor Biol 89: 659-677.

Henderson, C. R. 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics: 423-447.

Lipschutz-Powell, D. et al. 2012. Bias, accuracy, and impact of indirect genetic effects in infectious diseases. Frontiers in genetics 3.

Lipschutz-Powell, D., J. A. Woolliams, and A. B. Doeschl-Wilson. 2014. A unifying theory for genetic epidemiological analysis of binary disease data. Genet Sel Evol 46: 1-12.

Lloyd-Smith, J. O., S. J. Schreiber, P. E. Kopp, and W. M. Getz. 2005. Superspreading and the effect of individual variation on disease emergence. Nature 438: 355-359.

Lynch, M., and B. Walsh. 1998. Genetics and analysis of quantitative traits. Sinauer Sunderland, MA.

Nicholas, F. W. 2005. Animal breeding and disease. Philosophical Transactions of the Royal Society B: Biological Sciences 360: 1529-1536.

Pooley, C., S. Bishop, and G. Marion. 2014. Estimation of single locus effects on susceptibility, infectivity and recovery rates in an epidemic using temporal data. In: Proceedings of the 10th World Congress on Genetics Applied to Livestock Production. Vancouver, Canada

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

Velthuis, A., M. De Jong, E. Kamp, N. Stockhofe, and J. Verheijden. 2003. Design and analysis of an Actinobacillus pleuropneumoniae transmission experiment. Preventive veterinary medicine 60: 53-68.

Wade, M. J., P. Bijma, E. D. Ellen, and W. Muir. 2010. Group selection and social evolution in domesticated animals. Evolutionary Applications 3: 453-465.

6

General discussion

6.1 Introduction

The basic reproduction ratio, R_0 is an important epidemiological parameter that determines the risk and severity of an epidemic and the existence of an endemic equilibrium in a population. R_0 is the average number of new cases that a typical infectious individual produces during its entire infectious lifetime in a completely susceptible population. It has a threshold value of 1. When $R_0 < 1$, only a minor disease outbreak can occur and the disease will die out. When $R_0 > 1$, also major disease outbreaks can occur and the disease may become endemic. Because of this threshold property, any breeding strategy that aims to reduce the prevalence of an infection should reduce the value of R_0 preferable to a value below 1. The reduction to below 1 can be achieved of course in combination with other mitigating measures.

An important concept in quantitative genetics is the breeding value, which is a property of a single individual. On the contrary, R_0 is a parameter that refers to the entire population. Reducing R_0 via breeding, therefore, requires defining individual breeding values and heritable variances for R_0. In the 2nd chapter, we showed that individual breeding value and heritable variance for R_0 can be defined by combining the indirect genetic effects theory of quantitative genetics and the epidemiological concept of the next generation matrix.

In the 3rd chapter of this thesis, we showed that gene effects for susceptibility and infectivity can be estimated using a Generalized Linear Model (GLM), which was developed from an equation that describes the probability of an individual to be infected as a function of its own susceptibility and the average infectivity of its infectious group mates. This was done in order to show that the relative effects of genes on individual susceptibility and infectivity can be estimated using a model that takes the stochastic nature of an epidemic into account. Furthermore, in this chapter, we showed the effect of the value of R_0 and the degree of relatedness among groupmates on the bias and precision of the estimated gene effects in susceptibility and infectivity.

In the 4th chapter, we showed that a GLM developed in the 3rd chapter can be used to estimate the effect of Major Histocompatibility Complex (MHC) polymorphisms on individual susceptibility and infectivity for nematode infections in sheep. In this chapter, the MHC polymorphisms were found to have an effect on individual infectivity in addition to their effect on individual susceptibility. In the 5th chapter we showed that breeding values and additive genetic (co)variances for susceptibility and infectivity can be estimated from binary data with a Generalized Linear Mixed Model (GLMM)). However, estimates for infectivity showed considerable bias and reasons for this bias are discussed.

This thesis is among the first few studies (see also Lipschutz-Powell et al., 2012b; Lipschutz-Powell et al., 2012c) towards understanding the effects of genetic heterogeneity in susceptibility and infectivity on the risk and severity of an epidemic, which is determined by the value of R_0. Thus, with the aim of indicating the possible extensions and practical implications of this thesis, in this last chapter of the thesis, I will discuss three points in broader perspective. In the first section of this chapter, I will discuss how to define individual breeding values for the basic reproduction ratio, R_0 and its relation to susceptibility and infectivity traits of an individual. In the second section of this chapter, I will discuss selection strategies that could be used for increasing selection response in R_0. Finally, I will discuss the practical implications of the findings of this thesis.

6.2 Breeding value for R_0

The aim of animal breeding is to genetically improve livestock populations, that is, to have a next generation that performs better than the current generation. These genetic improvements are achieved by selecting those individuals that are genetically superior than the rest of the population and using them as parents of the next generation. The genetic superiority of an individual animal is reflected by its breeding value. The individual breeding value is twice the expected deviation of the phenotypic value of its offspring from the population mean (Falconer and Mackay, 1996). Since we cannot observe the breeding value of an individual, they need to be estimated (predicted) from already existing phenotypic information. In classical quantitative genetics, different phenotypic information sources, such as own performance and offspring/sib-information are used to estimate individual breeding values (Lynch and Walsh, 1998). With recent advances made in high-throughput genotyping technologies, genomic information is used to estimate (predict) individual (genomic) breeding values (Meuwissen et al., 2001). Once we have estimated individual breeding values, individual animals can be ranked based on their estimated breeding values and we can select the individuals with the best estimated breeding values to be parents of the next generation.

With regard to infectious diseases, the ultimate goal of selective breeding would be to have a population with reduced risk and size of an epidemic and/or reduced equilibrium prevalence of an endemic disease. The risk and size of an epidemic and/or the level of the endemic equilibrium depends on the value of R_0 (Diekmann et al., 1990). The larger R_0 is, the larger is the risk and size of an epidemic and the level of the endemic equilibrium, which is reflected by the fraction of individuals that gets infected (Figure 2 in chapter 1). Thus, breeding strategies aiming to reduce the risk and size of an epidemic and/or the level of endemic equilibrium should reduce R_0.

Classical livestock genetic improvement focuses on the genetic merit of an individual that is transmitted to its offspring. As mentioned above, the genetic merit of an individual is reflected by its breeding value. Genetic improvement for reduced R_0, should therefore be based on average effects of individual genes on R_0. This would then require to define and/or estimate individual breeding values for R_0. R_0 is an emergent trait of a population; i.e. it cannot be attributed to a single individual. The parameter R_0 is determined by how susceptible and infectious individuals interact in a population. Bijma (2011) has, however, shown that the result from the field of indirect genetic effects, which states that the trait value of an individual depends on multiple individuals, can be extended to traits that are an

emergent property of the population. In the 2nd chapter of this thesis we have shown that the approach proposed by Bijma (2011) and the epidemiological concept of the next generation matrix (Diekmann et al., 2010), can be used to define individual breeding value for R_0. As a result, we have shown that individual breeding value for R_0 can be defined as average gene effects at susceptibility and infectivity locus, and of population's average susceptibility and infectivity.

In the 3rd chapter of this thesis, we have shown that gene effects for susceptibility and infectivity can be estimated using a generalized linear model (GLM). Together with the average susceptibility and infectivity that can also be obtained from the model, these estimated gene effects thus can be used to formulate individual breeding values for R_0.

In the 2nd and 3rd chapter of this thesis, we have assumed that susceptibility is affected by one locus and infectivity is affected by another locus, each locus with two alleles. This is of course a simplified assumption, chosen to illustrate the effect of genetic heterogeneity on R_0. In reality, however, it is highly unlikely that we will find genes with such major effects. In fact, a common assumption in quantitative genetics is that quantitative traits are determined by a large number of loci each locus with a small effect (Fisher, 1918). It is most likely that this assumption holds also for susceptibility. The presence of 'Superspreaders', however, suggest that there could be few major genes with bigger effect and many with small effect on individual infectivity. In any of cases, the effect of the alleles is summed over all the loci within an individual, and the sum is known as the breeding value of an individual (Falconer and Mackay, 1996). Thus, to account for this fact, the gene effects in equation 7c of the 2nd chapter of this thesis need to be formulated in terms of breeding values.

Using equation 7c presented in chapter 2 and translating the average gene effects for susceptibility and infectivity into their respective breeding values, breeding value for R_0 can be formulated as follows:

$$A_{R_0,i} = [\bar{\varphi}\, A_{\gamma,i} + \bar{\gamma}\, A_{\varphi,i}] \qquad [1]$$

where $\bar{\varphi}$ is population average infectivity, $\bar{\gamma}$ is population average susceptibility, $A_{\gamma,i}$ is individual breeding value for susceptibility and $A_{\varphi,i}$ is individual breeding value for infectivity.

Based on this equation, breeding value for R_0 is the weighted sum of breeding values in susceptibility and infectivity, population average infectivity and susceptibility being the weighting factors. This can be seen as a selection index where the selection index is the sum of estimated individual breeding values of an individual for different traits weighted by their respective "economic" weights.

In the 5th chapter of this thesis, we have made an attempt towards estimating breeding values and (co)variances in susceptibility and infectivity using a generalized linear mixed model (GLMM) from a binary data. We fitted a direct-indirect genetic effects model with complementary log-log link function where the dependent variable was assumed to have a binomial distribution. The estimated breeding values together with the population means for susceptibility and infectivity can thus be used to define/assign individual breeding value for R_0. Once we have individual estimated breeding values for R_0, we can then select those with the best breeding value for R_0 to be parents of the next generation and breed for reduced R_0.

In this thesis, I try to show how animal breeders can use breeding for reduced R_0 as a complementary method to reduce the risk and size of an epidemic and the equilibrium prevalence of an endemic infection. Even though breeding for reduced R_0 might be the obvious choice for epidemiologists, it may not be the expected choice for animal breeders. For animal breeders, breeding for reduced disease incidence/prevalence and risk might be an obvious choice. Thus, one might ask why animal breeders should want to breed for reduced R_0 while they can breed for reduced disease incidence/prevalence and risk. As mentioned above, the risk and size of an epidemic and the level of endemic equilibrium is determined by the value of R_0. In addition to that, it was shown that individual breeding value for R_0 is a function of individual breeding values and population averages in susceptibility and infectivity (Equation 1). With this knowledge in mind, when the aim is to reduce R_0, the amount of selection response that can be obtained depends on the structure of the data, which is usually binary indicating the disease status (infected/not-infected) of the individual from which individual breeding values for R_0 is estimated. When the data comes from a population where there is no family structure, the disease status of an individual captures part of the genetic variation present in its susceptibility. Thus, when individuals are selected based on their estimated breeding values, it will result limited selection response in R_0. When the data is from a population where there exists relatedness among individuals, disease status of an individual captures the genetic variation present in susceptibility and infectivity. Thus, selection based on individual breeding value for R_0 will bring about greater selection response in R_0 and thus greater reduction in the risk and size of an epidemic and the level of endemic equilibrium.

On the other hand, when the objective is to breed for reduced disease incidence/prevalence, breeder may use disease status of individuals as a selection criterion. In this case, the amount of selection response that can be obtained depends on the type of selection strategy applied. When breeders selects among

unrelated individuals, breeding for reduce disease incidence will result in limited response. This is because individual selection among unrelated individuals captures the genetic variation present in susceptibility only (see section 6.3.2 for details). Furthermore, due to the nonlinear relationship between the value of R_0 and the incidence/prevalence of an infection, breeding for reduced disease incidence/prevalence brings about a smaller response in reducing the risk and size of an epidemic.

6.3 Selection response in R_0

Responses to selection in any given trait, including R_0, can be expressed as the product of selection intensity *i*, accuracy of selection ρ_T and total genetic standard deviation σ_T (Bijma, 2011).

$$R = \iota \, \rho_T \, \sigma_T \qquad [2]$$

Response to selection, R, refers to the change in mean value of the trait from one generation to the next generation. Selection intensity is selection differential expressed in standard units. Accuracy of selection is the correlation between the total breeding value and the selection criterion in the selection candidates. Both selection intensity and accuracy of selection are scale-free parameters and thus provide no information about the heritable variance in the trait. Total genetic standard deviation, however, is the parameter that provides information about the potential of a population to respond to selection. With regard to R_0, this total genetic standard deviation refers to the total genetic standard deviation in R_0.

Bijma et al. (2007b) have shown that three factors determine the amount of response to selection that can be obtained in traits that are affected by direct and indirect genetic effects of individuals, such as individual disease status. These are: (1) the amount of genetic variation present in the direct and indirect genetic effects and covariance between the traits, which is the total genetic variance in the trait (2) the type of selection strategy used, that is, individual versus group selection and (3) the degree of relatedness among interacting groupmates. The last two factors can be considered as factors that determine the utilization of the heritable variance present in a trait. Out of the three factors which are outlined to be determinants of response to selection, the amount of genetic variation in susceptibility and infectivity and the degree of relatedness among groupmates, were observed to have an effect on the amount of selection response that can be obtained in R_0 when individuals are selected based on their disease status (in the 2nd chapter of this thesis). In the next sections, I will discuss the relationship between the three factors and selection response in R_0 when the selection criterion is individual disease status.

6.3.1 Heritable variance in susceptibility and infectivity

In the 2nd chapter of this thesis, it was shown that that heritable variance in R_0 can defined by taking variance in breeding values for R_0. As a result, it was shown that heritable variance present in susceptibility and infectivity will contribute to the heritable variance in R_0. Individual susceptibility can be defined as the probability of an individual to be infected upon exposure to an infection. Individual infectivity, on the other hand, is the probability of an individual to infect an average susceptible individual given contact.

The theory of direct-indirect genetic effects states that, an individual has two unobserved genetic effects: an effect expressed on the trait value of the individual itself, which is known as direct genetic effect (DGE), and an effect expressed on the trait value of another individual in its proximity, which is known as indirect genetic effect (IGE) (Griffing, 1967, 1976, 1981; Moore et al., 1997; Wolf et al., 1998). IGEs, which are also known as associative effects, can be considered as a heritable component in the environment that an individual experiences. Based on the direct-indirect genetic effects theory, the average effect of individual's genes on its susceptibility to an infection can be considered as a DGE, and the average effect of individual's genes on its infectivity can be considered as an IGE. It was shown by Bijma et al.(2007b) that in the presence of interaction, heritable variance in IGEs can contribute to the total heritable variance that reflects the potential of a population to respond to selection. This suggests that genetic variation present in the direct effect of susceptibility and indirect effect of individual's infectivity contribute to the total heritable variation in R_0.

A number of genome-wide association studies have reported the existence of single nucleotide polymorphisms (SNPs) that are associated with susceptibility to different infectious diseases (Pant et al., 2010; Kirkpatrick et al., 2011; Sherlock et al., 2013; Bermingham et al., 2014). Heritability estimates for susceptibility to infectious diseases also indicate the presence of exploitable genetic variation in susceptibility to infectious diseases (Heringstad et al., 2005; Gonda et al., 2006). In addition to that, in the 2nd chapter of this thesis it was found that susceptibility has an indirect genetic effect, which is expressed on the disease status of other individuals in its proximity. That is to say, an individual that is in a group with highly susceptible group mates will have a higher probability to be infected than an individual that is in a group with less susceptible individuals. A similar phenomenon was observed by Bishop and Stear (1997), who showed that selection for resistance yielded greater response than predicted from classical quantitative genetics, due to changes in the epidemiology of the disease. This suggests that a reasonable

heritable variance may still be available that can be utilized by artificial or natural selection.

Unlike traits that are part of an individual's fitness, such as the direct effect of an individual's susceptibility, an individual's infectivity is not part of its fitness. As a result, both natural and artificial selection do not exhaust the heritable variance that may be present in infectivity (at least in the absence of kin-selection). The presence of "superspreaders" also suggests the presence of genetic variation among hosts in their infectivity (Woolhouse et al., 1997; Lloyd-Smith et al., 2005).

Summing all, reasonable variance may exist in susceptibility and as well as in infectivity, that will contribute to the heritable variance in R_0. Thus, the next question would be what selection strategies could be used to utilize the heritable variance present in these traits, and thus provide increased response in R_0. In the next sections, I will discuss possible selection strategies and their potential in the utilization of heritable variation in R_0 and selection response for R_0.

6.3.2 Classical individual selection versus group selection

For a given trait, response to selection can also be predicted as the regression coefficient of the total breeding value of an individual on the selection criterion, multiplied by the selection differential. As a result, an expression for selection response with any level of selection measured by g and with relatedness measured by r among groupmates in a group of size n was provided by Bijma et al. (2007b) as:

$$R = \left\{g[(n-1)r+1]\sigma^2_{TBV} + (1-g)\sigma_{P,TBV}\right\}\frac{\iota}{\sigma_C} \qquad [3]$$

where σ^2_{TBV} is the heritable variance in total breeding value in the population and $\sigma_{P,TBV}$ is the covariance between the phenotype of an individual and its total breeding value. $\frac{\iota}{\sigma_C}$ is the selection gradient, which is the regression coefficient of fitness on the selection criterion (Falconer and Mackay, 1996).

In equation [3], the factor g measures the level on which selection acts, that is, individual *versus* group selection. When $g = 0$, selection is made based on individual trait value, that is, individual (mass) selection. On the other hand, when $g = 1$, it represents selection on the sum of the trait values of the entire group, which is selection among groups. A $g = -1/(n-1)$, corresponds to selection of individuals on the deviation of their trait value from the mean trait value of their group (Bijma and Wade, 2008).

Individual selection: In the absence of relatedness among groupmates, the above equation for response to individual selection becomes:

$$R = \left\{\sigma_{P,TBV}\right\}\frac{\iota}{\sigma_C} = \left\{\sigma^2_{A_D} + (n-1)\sigma_{A_{DS}}\right\}\frac{\iota}{\sigma_C} \qquad [4]$$

This expression suggests that in the presence of indirect genetic effects, response to individual selection depends on the variance in the direct genetic effect $\sigma^2_{A_D}$, and the genetic correlation between the direct and indirect genetic effects $\sigma_{A_{DS}}$. Thus, individual selection will ignore the heritable variation present in the indirect genetic effect of an individual and thus provides limited response to selection.

With regard to infectious diseases, individual selection refers to selection of individuals that are not-infected are selected to be parents of the next generation. When considering breeding for reduced R_0, and when breeders would have to rely on individual's disease status (infected/not-infected) as a selection criterion, the amount of selection response that can be obtained depends on the structure of the population. When unrelated individuals are housed together, the disease status of an individual captures the direct genetic effect of the individual itself only. This phenomenon was observed in the 2nd chapter of this thesis, where it was shown that when groups are composed of random individuals, individual selection based on their disease status captures the heritable variation present in the direct genetic effect of susceptibility only and thus led to limited response in R_0. Moreover, based on the above expression, response to individual selection in the absence of relatedness among individuals could also lead to negative response to selection (Bijma et al., 2007b). This happens for the case where genetic covariance between the direct and indirect genetic effects, $\sigma_{A_{DS}} < 0$, which could result in $|(n-1)\sigma_{A_{DS}}|$ to exceed the variance in the direct genetic effect, $\sigma^2_{A_D}$. With regard to infectious disease, this applies to the case where individuals with lower susceptibility are more infectious. Simulation results from the 2nd chapter of this thesis support this claim, where it was shown that when there exists relatively strong negative linkage disequilibrium between susceptibility and infectivity, which was considered as the only measure of correlation, selection based on individuals' diseases status results in selection response in R_0 in the opposite direction, i.e. increased R_0.

Group selection: In this case, individuals in a group with an average phenotypic value above a certain threshold are selected to be parents of the next generation, that is, $g = 1$ (Equation 3). Based on the expression provided by Bijma et al., (2007a), group selection makes the total heritable variance in the trait to be the factor that determines selection response. As a result, group selection always leads to positive response to selection.

The expression for group selection given by Bijma et al. (2007a) is:

$$R = [(n-1)r + 1]\sigma^2_{TBV}\left(\frac{\iota}{\sigma_C}\right) \qquad [5]$$

The later equation also suggests that selection response is greater when selection is made between family groups ($r > 0$) than between groups with unrelated individuals.

When breeders consider groups as selection units in their aim to breed for reduced R_0, their selection criterion might be to select for those groups with fewer/no individuals that gets infected. The probability of the group to have fewer/no individuals that gets infected is higher when a group is composed of individuals with lower than average susceptibility and infectivity. Thus, when selecting those groups with fewer/no infected individuals to contribute parents of the next generation, we will capture heritable variance in both susceptibility and infectivity and thus, increase response in R_0.

In Bijma et al. (2007a), the effect of group selection on response to selection was reported to be more pronounced when selection is made between groups composed of families. This also applies for increasing selection response in R_0 in a way that the probability for a group to be composed of individuals with lower than average susceptibility and infectivity is higher because of relatedness. Thus, the probability for the group to have fewer/no infected individual is increased. As a result, genetic variation in susceptibility and infectivity is captured, which leads to increased response in R_0. Thus, breeders should consider keeping families in groups when the aim is to reduce R_0 by selective breeding.

6.3.3 Relatedness among interacting groupmates

Relatedness: In this case, selection is made between relatives within a group. It was shown by (Bijma et al., 2007b) that relatedness converts covariance between individual phenotype and the total breeding value into total heritable variance, which is always positive and thus leads to positive selection response. The expression for response to individual selection when relatedness among interacting individuals is *r*=1 is provided by (Bijma et al., 2007a)*:*

$$R = \sigma^2_{TBV}\left(\frac{\iota}{\sigma_C}\right) \qquad [6]$$

With regard to breeding for reduced R_0, it was shown in the 2nd chapter of this thesis that considerable increased selection response in R_0 was obtained when selecting among individuals that were present in groups of related individuals. This happens, because relatedness allows us to capture the genetic variation present in the indirect genetic effect of susceptibility and infectivity. This occurs, because individuals with less susceptibility and infectivity will, on average, have groupmates with below average susceptibility and infectivity (since they are related). As a result, they will have higher probability to escape the infection and be selected as parents of the next generation. Thus, when relatives are kept in a group, individual selection captures the genetic variation present in the indirect genetic effect of susceptibility and infectivity and provides increased response in R_0.

In the 3rd chapter of this thesis, it was also shown that relatedness has an effect on the bias of the estimated gene effects on individual susceptibility and infectivity. It was also shown that when the degree of relatedness among groupmates is higher, the bias in estimated gene effects is smaller.

6.4 Practical implication

Throughout this thesis, we focused on developing methodologies that elucidate the impact of genetic variation in two disease-related traits, namely susceptibility and infectivity on the risk and final size of an infection in a population, which is determined by the value of R_0. R_0, being a parameter that determines the risk and final size of an infection, we suggest that breeding for reduced R_0 should be considered as a complementary method to the existing disease control strategies that aim to reduce the risk and final size of an infection in a population. Moreover, we have shown that individual breeding value for R_0 is a function of individual breeding value and population averages for susceptibility and infectivity. This suggests that, in order to have individual breeding values for R_0, we need to estimate individual breeding values for susceptibility and infectivity.

Advancements in the field of quantitative genetics have made estimation of the indirect genetic effect (IGE) of an individual possible through the use of indirect-genetic effect (IGE) model. The IGE-models can also be used to estimate individual breeding values for susceptibility and infectivity from phenotypic disease data, which are usually binary. Lipschutz-Powell et al. (2012b) have made the first attempt to estimate the variance in susceptibility and infectivity using the IGE-model. Results from (Lipschutz-Powell et al., 2012c), however, indicated that the linear IGE-model has a shortcoming in estimating breeding values for the indirect genetic effect of infectivity using binary disease data accurately. This shortcoming was also observed in chapter 5 of this thesis, where a generalized linear mixed model (GLMM) was used to estimate breeding values and variances in susceptibility and infectivity. One reason for this shortcoming could be the fact that the model captures the variation in disease dynamics among the different groups, which cannot be explained by random group effect or by an offset as variance in infectivity. This has led to an overestimation of genetic variance in infectivity. Moreover, it was observed that not only breeding values and variances infectivity are estimated less accurately, but also population average susceptibility and infectivity were not estimated accurately. However, the advancement made in this thesis (chapter 5) and by (Lipschutz-Powell et al., 2012a), can be considered as a first step towards estimating breeding values and variance in susceptibility and infectivity from disease data and for animal breeders to consider breeding for reduced R_0 in their breeding program.

Recent advances in molecular genetics have made genotyping of individuals for thousands of genetic markers, such as single nucleotide polymorphisms (SNPs) across the genome feasible. SNPs, which are the most common source of genetic variation that have been used in genome-wide association studies (GWAS) in order

to identify genes that are associated with a number of quantitative traits (Cochran et al., 2013; Mancini et al., 2013; Duchemin et al., 2014). GWAS have also been applied to identify genes that are associated with susceptibility to various infectious diseases (Pant et al., 2010; Kirkpatrick et al., 2011; Sherlock et al., 2013; Bermingham et al., 2014; LaRose et al., 2015). In addition to their application to identify genes associated with individual susceptibility to infectious diseases, GWAS can also be used to identify genes that are associated with individual infectivity, from which the gene effects on R_0 can be estimated.

As mentioned in the first section of this chapter, it is likely that individual susceptibility is affected by many genes each with small effect and infectivity by fewer major genes with big effect and large number of genes with small effects. Thus, in order to pick up the effect of all the possible genetic variations, we need to genotype individuals for thousands of dense SNP markers from which the gene effects are estimated to predict individual breeding values for susceptibility and infectivity.

As an alternative to genome-wide SNP genotyping, one could densely genotype chromosomal regions (candidate genome regions) that are known to have association with susceptibility to infection, such as the major histocompatibility complex (MHC). The major histocompatibility complex (MHC) is one of the major genes that affect disease susceptibility/resistance against different infectious diseases (Lamont, 1989; Schwaiger et al., 1995; Grimholt et al., 2003). In fact, in chapter 4, the MHC was also found to have an effect on the infectivity of individuals for nematode infection in the Scottish Blackface sheep. Through the application of candidate gene approach, these results will help us to select for individuals with genes that have desirable effects on individual susceptibility and infectivity to infectious diseases.

The advancements made in the molecular genetics not only allows the discovery of genes that are associated with different quantitative/diseases traits, but also open the door for animal breeders and the livestock industry to utilize these genetic variations to select and breed for improved livestock population, through the application of marker assisted selection (MAS) and genomic selection.

Genomic selection: Genomic selection is a variant of MAS that uses predicted breeding value from large number of estimated SNP effects across the genome to select individuals to be parents of the next generation (Meuwissen et al., 2001; Goddard and Hayes, 2007). The key property of this approach is that thousands of markers that are assumed to be in linkage disequilibrium with the actual quantitative trait loci (QTL) are used to cover the whole genome and predict

breeding values. It was shown that an accuracy > 50% can be attained in predicting breeding values using high-density SNP markers (Hayes et al., 2009).

The success made in the application of genomic selection to genetically improve different quantitative traits show that with the availability of marker information and appropriate statistical tools, genomic selection can also be applied for improvement of disease traits and reduction of disease prevalence in a population.
It was mentioned in Pooley et al. (2014) that information from multiple epidemic (groups) are needed to estimate gene effect for infectivity accurately. The availability of records from multiple epidemic groups is challenging, since in real situations, animals are often kept in one big stable rather than in multiple small groups. Alternative to that, one can use data from shorter time-intervals where each interval is considered as incomplete epidemic in order to predict individual breeding value for infectivity with reliable accuracy (Anacleto et al., 2015). This is because time-interval disease data allows us to observe the order of an infection, that is, on who infects who and thus provides better information about infectiousness of individuals. On the other hand, when time-interval disease data is coupled with records on a number of groups, it will allow us to capture the genetic variation present among individual in their susceptibility and infectivity to a larger extent.

In chapter 3, we have developed a generalized linear model (GLM) that can allow us to estimate gene effects in susceptibility and infectivity. With the availability of genomic and phenotypic data, the GLM that was developed in chapter 3 can also be used in genomic predictions of breeding values for susceptibility and infectivity, which can thus be used to predict individual breeding values for R_0, from which genomic selection can be applied for reduced R_0.

6.5 References

Anacleto, O., L. A. Garcia-Cortés, D. Lipschutz-Powell, J. A. Woolliams, and A. B. Doeschl-Wilson. 2015. A Novel Statistical Model to Estimate Host Genetic Effects Affecting Disease Transmission. Genetics 201: 871-884.

Bermingham, M. et al. 2014. Genome-wide association study identifies novel loci associated with resistance to bovine tuberculosis. Heredity 112: 543-551.

Bijma, P. 2011. A general definition of the heritable variation that determines the potential of a population to respond to selection. Genetics: genetics. 111.130617.

Bijma, P., W. M. Muir, E. D. Ellen, J. B. Wolf, and J. A. M. Van Arendonk. 2007a. Multilevel selection 2: Estimating the genetic parameters determining inheritance and response to selection. Genetics 175: 289-299.

Bijma, P., W. M. Muir, and J. A. Van Arendonk. 2007b. Multilevel selection 1: quantitative genetics of inheritance and response to selection. Genetics 175: 277-288.

Bijma, P., and M. Wade. 2008. The joint effects of kin, multilevel selection and indirect genetic effects on response to genetic selection. Journal of evolutionary biology 21: 1175-1188.

Bishop, S., and M. Stear. 1997. Modelling responses to selection for resistance to gastro-intestinal parasites in sheep. Animal Science 64: 469-478.

Cochran, S. D., J. B. Cole, D. J. Null, and P. J. Hansen. 2013. Discovery of single nucleotide polymorphisms in candidate genes associated with fertility and production traits in Holstein cattle. BMC genetics 14: 49.

Diekmann, O., J. Heesterbeek, and J. A. Metz. 1990. On the definition and the computation of the basic reproduction ratio R 0 in models for infectious diseases in heterogeneous populations. Journal of mathematical biology 28: 365-382.

Diekmann, O., J. A. P. Heesterbeek, and M. G. Roberts. 2010. The construction of next-generation matrices for compartmental epidemic models. J R Soc Interface 7: 873-885.

Duchemin, S., M. Visker, J. Van Arendonk, and H. Bovenhuis. 2014. Fine-mapping of a candidate region associated with milk-fat composition on Bos Taurus Autosome 17. In: 10th World Congress on Genetics Applied to Livestock Production

Falconer, and Mackay. 1996. Introduction to quantitative genetics. 4th ed. Harlow: Pearson Education Limited; .

Fisher, R. A. 1918. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. Transactions of the royal society of Edinburgh 52: 399-433.

Goddard, M. E., and B. Hayes. 2007. Genomic selection. Journal of Animal Breeding and Genetics 124: 323-330.

Gonda, M., Y. Chang, G. Shook, M. Collins, and B. Kirkpatrick. 2006. Genetic variation of Mycobacterium avium ssp. paratuberculosis infection in US Holsteins. Journal of dairy science 89: 1804-1812.

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-&.

Griffing, B. 1976. Selection in Reference to Biological Groups .5. Analysis of Full-Sib Groups. Genetics 82: 703-722.

Griffing, B. 1981. A Theory of Natural-Selection Incorporating Interaction among Individuals .2. Use of Related Groups. J Theor Biol 89: 659-677.

Grimholt, U. et al. 2003. MHC polymorphism and disease resistance in Atlantic salmon (Salmo salar); facing pathogens with single expressed major histocompatibility class I and class II loci. Immunogenetics 55: 210-219.

Hayes, B., P. Bowman, A. Chamberlain, K. Verbyla, and M. Goddard. 2009. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet Sel Evol 41: 51.

Heringstad, B., Y. Chang, D. Gianola, and G. Klemetsdal. 2005. Genetic association between susceptibility to clinical mastitis and protein yield in Norwegian dairy cattle. Journal of dairy science 88: 1509-1514.

Kirkpatrick, B., X. Shi, G. Shook, and M. Collins. 2011. Whole-Genome association analysis of susceptibility to paratuberculosis in Holstein cattle. Animal genetics 42: 149-160.

Lamont, S. 1989. The chicken major histocompatibility complex in disease resistance and poultry breeding. Journal of dairy science 72: 1328-1333.

LaRose, J., D. Wilson, and K. Rood. 2015. Identification of single nucleotide polymorphisms associated with mastitis resistance in dairy cows.

Lipschutz-Powell, D., J. A. Woolliams, P. Bijma, and A. B. Doeschl-Wilson. 2012a. Indirect Genetic Effects and the Spread of Infectious Disease: Are We Capturing the Full Heritable Variation Underlying Disease Prevalence? Plos One 7: e39551.

Lipschutz-Powell, D., J. A. Woolliams, P. Bijma, and A. B. Doeschl-Wilson. 2012b. Indirect genetic effects and the spread of infectious disease: are we capturing the full heritable variation underlying disease prevalence?

Lipschutz-Powell, D. et al. 2012c. Bias, accuracy, and impact of indirect genetic effects in infectious diseases. Frontiers in genetics 3.

Lloyd-Smith, J. O., S. J. Schreiber, P. E. Kopp, and W. M. Getz. 2005. Superspreading and the effect of individual variation on disease emergence. Nature 438: 355-359.

Lynch, M., and B. Walsh. 1998. Genetics and analysis of quantitative traits. Sinauer Sunderland, MA.

Mancini, G. et al. 2013. Association between single nucleotide polymorphisms (SNPs) and milk production traits in Italian Brown cattle. Livestock Science 157: 93-99.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819-1829.

Moore, A. J., E. D. Brodie, and J. B. Wolf. 1997. Interacting phenotypes and the evolutionary process .1. Direct and indirect genetic effects of social interactions. Evolution 51: 1352-1362.

Pant, S. D. et al. 2010. A principal component regression based genome wide analysis approach reveals the presence of a novel QTL on BTA7 for MAP resistance in holstein cattle. Genomics 95: 176-182.

Pooley, C., S. Bishop, and G. Marion. 2014. Estimation of single locus effects on susceptibility, infectivity and recovery rates in an epidemic using temporal data. In: Proceedings of the 10th World Congress on Genetics Applied to Livestock Production. Vancouver, Canada

Schwaiger, F.-W. et al. 1995. An ovine major histocompatibility complex DRB1 allele is associated with low faecal egg counts following natural, predominantly Ostertagia circumcincta infection. International journal for parasitology 25: 815-822.

Sherlock, R. et al. 2013. Whole genome association analysis of susceptibility to paratuberculosis in New Zealand dairy cattle. In: Proceedings of the Twentieth Conference of the Association for the Advancement of Animal Breeding and Genetics, Translating Science into Action, Napier, New Zealand, 20th-23rd October 2013. p 195-198.

Wolf, J. B., E. D. Brodie III, J. M. Cheverud, A. J. Moore, and M. J. Wade. 1998. Evolutionary consequences of indirect genetic effects. Trends in Ecology & Evolution 13: 64-69.

Woolhouse, M. E. et al. 1997. Heterogeneities in the transmission of infectious agents: implications for the design of control programs. Proceedings of the National Academy of Sciences 94: 338-342.

Summary

Summary

Infectious diseases of animals are a major concern to the livestock industry, particularly due to their effect on the welfare and productivity of livestock. In addition to that, the zoonotic nature of some infectious diseases poses a threat to public health.

Studies have reported that there exists among host genetic variation for different disease related traits. These findings suggest that breeders can implement selective breeding as a complementary method to the existing disease control strategies to genetically improve host populations in order to decrease the prevalence of infectious diseases in livestock populations.

Genetic approaches aiming to reduce the prevalence of an infection usually focus on reducing individual susceptibility to an infection. The prevalence of an infection, however, is affected also by the infectivity of infectious individuals. Host genetic variation in susceptibility and infectivity affects the transmission of an infection in a population, the total effect of which is measured by the basic reproduction ratio, R_0. R_0 is the average number of new cases produced by a typical infectious individual during its entire infectious lifetime in a completely susceptible population. It is an important epidemiological parameter that determines the prevalence and risk of an infection in a population. Moreover, R_0 has a threshold value of 1, where a major disease outbreak can occur only when $R_0 > 1$. When $R_0 < 1$, only minor disease outbreaks can occur and the disease will die out. Thus, breeding strategies that aim to reduce the prevalence of an infection should reduce the value of R_0 below 1. Since the theory of response to selection rests on the concepts of breeding value and heritable variance, genetic strategies to reduce R_0 requires the definition and estimation of breeding values for and heritable variance in R_0.

In Chapter 2, we show how to define individual breeding values and heritable variation for R_0, by combining the quantitative genetic theory of indirect genetic effects, the epidemiological concept of the next generation matrix, and a Susceptible-Infectious-Recovered (SIR) model of an infectious disease. As a result, the individual breeding value was defined as a function of an individual's allele frequencies for susceptibility and infectivity and the population average susceptibility and infectivity. Moreover, we show that, when interacting individuals are unrelated, selection of individuals based on their own disease status captures genetic variation in susceptibility only, resulting in a limited response in R_0. When interacting individual are related, however, selection of individuals based on their disease status also captures genetic variation in infectivity and additional variation in susceptibility, resulting in substantially greater response in R_0. This result shows

that, not only infectivity, but also susceptibility to infectious diseases has an indirect genetic effect on the disease status of individuals.

In Chapter 3, we developed a generalized linear model (GLM) to estimate the relative effects of genes on individual susceptibility and infectivity. The GLM was developed from an equation that describes the probability of an individual to become infected as a function of its own susceptibility genotype and the infectivity genotypes of its infectious contacts. We investigated the quality of the GLM in terms of the bias and precision of the estimates. The bias was smaller when R_0 was between 1.8 and 3.1, and when relatedness among group mates was higher.

In Chapter 4, we used the GLM developed in Chapter 3 to estimate the effects of polymorphisms in the MHC-genes on individual susceptibility and infectivity for nematode infection in sheep. We found that, in addition to their effect on individual susceptibility, the MHC-genes have an effect on individual infectivity for nematode infection.

In Chapter 5, we developed a generalized linear mixed model (GLMM) to estimate individual breeding values and genetic (co)variances in susceptibility and infectivity from binary disease data. For susceptibility, additive genetic variance and breeding values were estimated with acceptable accuracy. For infectivity, however, the estimated additive genetic variance showed a very large bias, and estimated breeding values showed low accuracy. Relatedness among group mates significantly reduced the bias in the estimated genetic variance for infectivity, and increased the accuracy of estimated breeding values for infectivity.

In the General Discussion (Chapter 6) I discuss three points: First, I discuss the individual breeding value for R_0, and its relation to susceptibility and infectivity of the individual. Second, I discuss selection strategies that can be used for reducing R_0. Finally, I discuss the practical implications of the findings of this thesis. I conclude that, with advancements made in statistical methods and quantitative and molecular genetics, breeders should consider breeding for reduced R_0 in their breeding goal when the aim is to reduce the prevalence and risk of an infection.

Curriculum Vitae

About the author

Mahlet Teka Anche is born on 1st November 1985 in Addis Abeba, Ethiopia. When she came of school age, she joined a nearby public elementary school and completed her high school education in Addis Abeba in 2004. Mahlet followed her BSc study in animal and range sciences in Hawassa University, in Hawassa college of agriculture. After finishing her BSC in 2007, Mahlet worked as a graduate assistance for 2 years at Hawassa college of Agriculture. In the year 2009, Mahlet received a scholarship to pursue her masters studies in Animal breeding and genetics. In 2009, she join Wageningen University, animal breeding and genomics centre for her first year masters study and then moved to the University of natural resources and life sciences, Boku, Vienna. After finishing her masters in 2011, she then started her PhD research entitled estimating host genetic effects on susceptibility and infectivity to infectious diseases and their contribution to response to selection. The results of this thesis were presented in international conferences and published in journals. Since February 2016, Mahlet started working as a postdoctoral researcher at Aarhus University on multi-trait genomic prediction.

Publications

Gledler, B., **Anche, M.T.**, Schwarzenbacher, H, Egger-Danner, C. and Solkner, J., *Accuracy of genomic prediction using subsets on SNP markers in dual purpose Fleckvieh Cattle*. Book of Abstracts of the 62nd Annual meeting of the European Federation of Animal Science (EAAP). 2011.

Anche M.T., de Jong M., Bijma P., *On the definition and utilization of heritable variation among hosts in reproduction ratio R0 for infectious diseases*, Heredity. (2014)

Anche M.T., Bijma P., de Jong M., *Genetic analysis of infectious diseases: Estimating gene effect for susceptibility and infectivity*. Genetics Selection Evolution.

Anche M. T., Bijma P., Stear M. J., de Jong M. The effect of polymorphisms in major histocompatibility complex (MHC) on individual susceptibility and infectivity to nematode infection in Scottish Blackface sheep. To be submitted

Anche M.T., de Jong M., Bijma P., Estimating genetic co(variances) and breeding values for host susceptibility and infectivity from the final disease status of hosts exposed to epidemics in group-structured populations. To be submitted

Conference preceedings

Anche M.T., de Jong M., Bijma P., *Identifying factors underlying heritable variation and response to selection in R0*: simulation study. Book of Abstracts of the 64th Annual meeting of the European Federation of Animal Science (EAAP). 2013.

Anche M.T., de Jong M., Bijma P., *Definition and utilization of among hosts heritable variation in reproduction ratio R_0 for infectious diseases.* Proceedings, 10th World Congress of Genetics Applied to Livestock Production. 2014.

Training and education

Training and education

EDUCATION AND TRAINING (minimum 30, maximum 60 credits)		
The Basic Package (minimum 3 credits)	year	credits *
WIAS Introduction Course (mandatory, 1.5 credits)	2011	1,5
Course on philosophy of science and/or ethics (mandatory, 1.5 credits)	2012	1,5
Introduction interview with WIAS scientific director and secretary: 20.10.2011		-
Introduction interview with WIAS education coordinator: 27.10.2011		-
Introduction interview with WIAS PhD students confidant: 19.11.2011		-
Subtotal Basic Package		**3**
Scientific Exposure (conferences, seminars and presentations, minimum 8 credits)	year	credits
International conferences *(minimum 3 credits)*		
10th World Congress on Genetics applied to Livestock Production (WCGALP)	2014	1,5
64th EAAP Annual Meeting	2013	1,2
4th International Conference on Quantitative Genetics (ICQG), Edinburgh (Scotland),	2012	1,5
66th EAAP Annual Meeting, (planned)	2015	1,5
Seminars and workshops		
Seminar 'Wageningen evolution and ecology'	2012	0,2
WIAS Science Day	2012-2015	1,2
Presentations *(minimum 4 original presentations of which at least 1 oral, 1 credit each)*		
10th World Congress on Genetics applied to Livestock Production (WCGALP), - (oral presentation)	2014	1,0
64th EAAP Annual Meeting, (oral presentation)	2013	1,0
66th EAAP Anuual meeting, (oral presentation)	2015	1,0
WIAS Science day 2015 (oral presentation)	2015	1,0
Subtotal Scientific Exposure		**11**
In-Depth Studies (minimum 6 credits, of which minimum 4 at PhD level)	year	credits
Disciplinary and interdisciplinary courses		
Mixed models in quantitative genetics Bruce Walsh	2012	0,9
Genetics of competition by Piter Bijma	2012	0,9
Advanced methods and algorithms in animal breeding with focus on genomic selection	2012	1,5
Mathematical modelling of infectious diseases	2012	3,0
Genetic analysis using ASReml4.0	2014	1,5
Advanced statistics courses *(optional)*		
PhD students' discussion groups *(optional)*		
Quantitative Genetics Discussion Group (weekly meetings), ABGC	2011-2015	2,0
MSc level courses *(only in case of deficiencies)*		
Subtotal In-Depth Studies		**10**
Statutory Courses	year	credits
Use of Laboratory Animals (mandatory when working with animals)		
Laboratory Use of Isotopes (mandatory when working with radio isotopes)		
Subtotal Statutory Courses		**0**
Professional Skills Support Courses (minimum 3 credits)	year	credits
WGS Course: Techniques for Writing and Presenting a Scientific Paper	2014	1,2
WGS Course: Project and Time Management	2011	1,0
Effective behaviour in your professional surrounding	2015	1,3
Subtotal Professional Skills Support Courses		**4**
Research Skills Training (optional)	year	credits
Preparing own PhD research proposal (maximum 6 credits)	2011	6,0
Introduction to R for statistical analysis	2013	0,6
Subtotal Research Skills Training		**7**
Education and Training Total (minimum 30, maximum 60 credits)		**34**

Acknowledgements

It seems that my time with ABGC has actually come to an end. The past few years at ABGC were both challenging and enjoyable. I am so fortunate for being part of this amazing group of people. My journey at ABGC was made possible and enjoyable with the help of some colleagues and friends and families. I would like to take this moment to thank all the people that have helped and supported me along the way.

First of all, I would like to thank God, The One who provided all the strength and the help all the times and who was there to pick me up from all my discouragements and disappointments. I give praise to Him for everything and in everything!

I count myself as lucky person to have worked with two amazing supervisors, Piter and Mart. Piter, this PhD would not have been a success without your tireless help! You thought me so many scientific and personal qualities that a successful researcher should have. You were more of a mentor rather a supervisor for me. You're deeply respected for all your help! I thank you! Mart, my deepest respect is for you too. I'm very grateful for your tireless effort to teach me the mathematical concepts and grounds in this thesis. You two were always three to four steps ahead of me in the thinking process, but with a great help from both of you, I somehow catch up. I was sometimes scared and confused with the discussion you two have in the weekly meeting but they were highly useful and fun. I guess you two also had fun with my silly questions and confused face☺. All in all, I am very thankful for all your help dear supervisors!

I'm very grateful for all my friends and colleagues at ABGC. It feels like home with you all. I laughed and talked and ate and partied with you, which I enjoyed very much. Sandrine, you always have a solution for every problem. I'm lucky to have you as friend during my stay at WUR. You're a great person who wants to exhaust what life has to offer. My response to that is, go girl!!! Britteeeee, oh how I enjoyed your company! You're an amazing, open hearted, a very busyyyy person. I'm so glad that I know you and thank you for being my paranymph! Nancy, you're my inspiration. You're a kind of person that some would say 'I want to be like you when I grow up'. Keep on being an inspiration for many! Ewa, Katrijn, Naomi, Marcos, André, Gabriel, Hadi, Marzieh, Hamed, you all have made a permanent, very positive environmental effect on me, which I received gladly and am going to transmit to the next generation☺. Ilse, you are such an amazing person! I thank you for everything you have done for me and also for my dear friend in Ethiopia!

Acknowledgements

You have brought about a difference in the life of one family! Ada and Lisette, you two have a caring eye and heart over the students at ABGC, I'm grateful for that.

My Ethiopian family, Banchi, Kaye, Mahdi, Sabi, Tsiye, Abiyo, Eskew, Ednaye, Noli…I have no word to tell you how fortunate I'm to have you all as my family. You all have been my docking place to crush my frustration and to multiply my happiness. I was strengthened by your care and thoughtful advices. I was cheered by your jokes and conversations. I was so comfortable by your non-judgmental acceptance of who I'm, as a friend and sister. Biruk, I thank you for being there for me when I needed you and for giving me the best advices. You're the best!
ሁላችሁንም፡የምወደው፡አምላኬ፡በህይወታችሁ፡ሁሉ፡ይባርካቹ!

Finally, my family/ ቤተሰቦቼ
እማማዬ፡አባባዬ፡ሁልጊዜ፡እኔን፡የተሻለ፡ደረጃ፡ላይ፡ለማየት፡ያላችሁ፡ጉጉት፡ለጉጉዬ፡ትልቅ፡ስንቅ፡ነበር።አማማ ዬ፡አንቺ፡የእኔ፡ብቻ፡ጀግና፡ነሽ።የምወደው፡እግዚአብሔር፡ከሁሉ፡በሚበልጠው፡በረከት፡ይባርክሽ፡አባባዬ፡የእኔ ፡ስኬት፡ላንተ፡ስኳር፡መድሃኒት፡መሆኑን፡ስትነግረኝ፡ምን፡ያህል፡እንደምበረታ፡አትጠይቀኝ፡አንተንም፡የምወደ ው፡አምላኬ፡ከሁሉ፡በሚበልጠው፡በረከት፡ይባርክህ፡ላላዬ፡እህት፡አለም፡የኔ፡ኮረፊማ፡እወድሻለሁ፡ሞጅዬ፡የኔ፡ ጎበዝ፡አንቺንም፡እወድሻለሁ፡አዩሻዬ፡የሁላችን፡አሳቢ፡አንቺንም፡በብዙ፡እወድሻለሁ፡አቢዬ፡እህታለም፡የእኔ፡ጀግ ና፡እወድሻለሁ፡ናናዬ፡የእኔ፡ተስፉ፡ነሽ፡አብልጬ፡እወድሻለሁ።ረዱ፣ዊንታና፣ቲምላ፣በላይቱ፣ኢዛና፣ዘውዱ፣ሊያ ዬ፣ተንሳይ፣በላይ፣ኮላዬ፣ፍቅሬ፡ሁላችሁንም፡እወዳችኋለሁ።፤በመንገዴ፡ሁሉ፡ብርታቴ፡ናችሁ።ሁላችሁንም፡የም ወደው፡የምወደው፡አምላኬ፡ከሁሉ፡በሚበልጠው፡በረከት፡ይባርካችሁ።

Colophon

The research described in this thesis was financially supported by EU Marie Curie NematodeSystemHealth (ITN-2012-264639). The data used in this thesis was provided by the Univesity of Glasgow, Institute of Biodiversity, Animal Health & Comparative Medicine, College of Medical and College of Medicine, Veterinary and life Sciences.

The cover of this thesis was designed by Roos Marina Zaalberg.

The thesis was printed by Digiforce| Proefschriftmaken.nl, De Limiet 26, 4131 NC, Vianen, The Netherlands