The quantitative genetics of the endemic prevalence of infectious diseases: Indirect Genetic Effects dominate heritable variation and response to selection

Piter Bijma[1], Andries Hulst[1,2], Mart C. M. de Jong[2]

[1]Animal Breeding and Genomics, Wageningen University and Research, Wageningen, the Netherlands

[2]Quantitative Veterinary Epidemiology, Wageningen University and Research, Wageningen, the Netherlands

Abstract

Pathogens have profound effects on life on earth, both in nature and agriculture. Despite the availability of well-established epidemiological theory, however, a quantitative genetic theory of the host population for the endemic prevalence of infectious diseases is almost entirely lacking. While several studies have demonstrated the relevance of the transmission dynamics of infectious diseases for heritable variation and response to selection of the host population, our current theoretical framework of quantitative genetics does not include these dynamics. As a consequence, we do not know which genetic effects of the host population determine the prevalence of an infectious disease, and have no concepts of breeding value and heritable variation for endemic prevalence.

Here we propose a quantitative genetic theory for the endemic prevalence of infectious diseases. We first identify the genetic factors that determine the prevalence of an infectious disease, using an approach founded in epidemiological theory. Subsequently we investigate the population level effects of individual genetic variation on R_0 and on the endemic prevalence. Next, we present expressions for the breeding value and heritable variation, for both prevalence and individual binary disease status, and show how these parameters depend on the endemic prevalence. Results show that heritable variation for endemic prevalence is substantially greater than currently believed, and increases when prevalence approaches zero, while heritability of individual disease status goes to zero. We show that response of prevalence to selection accelerates considerably when prevalence goes down, in contrast to predictions based on classical genetic models. Finally, we show that most of the heritable variation in the endemic prevalence of the infection is due to indirect genetic effects, suggestion a key role for kin-group selection both in the evolutionary history of current populations and for genetic improvement strategies in animals and plants.

Introduction

Pathogens have profound effects on life on earth, both in nature and agriculture, and also in the human population (Schrag and Wiener, 1995; Russel, 2013). In livestock, for example, the annual cost of fighting and controlling epidemic and endemic infectious diseases is substantially greater than the annual value of genetic improvement (Rushton, 1990; Knap and Doeschl-Wilson, 2020). Moreover, while antimicrobials have revolutionized medicine, the rapid appearance of resistant strains has resulted in a global health problem, both in the human population and in livestock (EFSA 2012; Thanner *et al.* 2016). Thus there is an urgent need for additional methods and tools to combat infectious diseases. For livestock and crop production, artificial genetic selection of (host) populations for infectious disease traits may provide such a tool. To quantify and optimize the potential benefits of such selection, we need to understand the quantitative genetics of infectious disease traits.

Current approaches to select against infectious diseases in livestock and crops are entirely based on the individual host response, ignoring transmission of the infection in the population. Despite the availability of well-established epidemiological theory (*e.g.*, Diekmann *et al.* 2013), quantitative genetic theory of the host population for the endemic prevalence of infectious diseases is almost entirely lacking. Infections for which recovery does not confer any long-lasting immunity typically show endemic behaviour, where the infection remains present in the population. For such infections, the endemic prevalence is defined as the expected fraction of the population that is infected. While several studies have demonstrated the relevance of the transmission dynamics of infectious diseases for heritable variation and response to selection in the host population (Lipschutz-Powell *et al.*, 2012; Anche *et al.*, 2014; Tsairidou *et al.*, 2019; Hulst *et al.*, 2021), mostly using stochastic simulations, the current theoretical framework of quantitative genetics does not include these dynamics. As a consequence, we do not know which genetic effects of the host population determine the prevalence of an infectious disease, and have no concepts of breeding value and heritable variation for endemic prevalence. Hence, we do not understand response to genetic selection in the prevalence of infectious diseases at present. Moreover, we lack general expressions for the genetic variance in key epidemiological parameters, in particular the basic reproduction number R_0, even though such parameters may have a genetic basis. R_0 is defined as the average number of individuals that gets infected by a typical infected individual in an otherwise non-infected population, and is the main parameter determining the prevalence of endemic infections and the size of epidemic infections. In this manuscript we will propose a quantitative genetic framework for heritable variation and response to selection for R_0 and the endemic prevalence of infectious diseases.

Individual phenotypes for infectious diseases are often recorded as the binary disease status of an individual, zero indicating non-infected and one indicating infected. The prevalence of a disease is then defined as the fraction of individuals that is infected and thus shows disease status $y = 1$. Because the average value of individual binary disease status is equal to the fraction of individuals infected, response to selection in binary disease status is identical to response in prevalence, and *vice versa*. Binary disease status (0/1) typically shows low heritability, which suggests limited response to selection (Bishop and Woolliams 2010; Bishop *et al*., 2012; Martin *et al*., 2018).

Geneticists have long realized that the categorical distribution of binary traits does not agree well with quantitative genetic models for polygenic traits, such as the infinitesimal model (Fisher 1918). For this reason, models have been developed that link an underlying normally-distributed trait to the observed binary phenotype, such as the threshold model (Dempster & Lerner, 1950; Gianola, 1982) and the equivalent generalized linear mixed model with a probit link function (*e.g.*, de Villemereuil *et al*. 2016). In such models, the underlying scale is interpreted as causal, and genetic parameters are assumed to represent "biological constants" on this scale. The genetic parameters on the observed scale, in contrast, depend on the prevalence of the trait, and thus change with the trait mean, even when the change in allele frequencies at causal loci is infinitesimally small. In a landmark paper, Robertson (1950) showed that the observed-scale heritability of binary traits reaches a maximum at a prevalence of 0.5, and approaches zero when the prevalence is close to 0 or 1. Hence, observed-scale heritability vanishes when artificial selection moves prevalence close to zero, hampering further genetic change.

Infectious disease status, however, differs fundamentally from binary phenotypes for non-communicable traits, such as, say, heart failure. Because pathogens can be transmitted between host individuals, either directly or via the environment, the infection status of an individual depends on the status of its contact individuals. This suggests that Indirect Genetic Effects (IGE) may play a role, which would fundamentally alter heritable variation and response to selection (Griffing 1967; Moore *et al*., 1997; Wolf *et al*., 1998; Bijma 2011; Bijma and Wade 2008). Results of simulation studies indeed suggest that response in the prevalence of infectious diseases may differ qualitatively from response in non-communicable traits (Nieuwhof *et al*., 2009; Doeschl-Wilson *et al*., 2011; Anche *et al*., 2014; Hulst *et al*. 2021), and this has also been observed in an actual population (Heringstad *et al*. 2007). Results of Hulst *et al*. (2021), for example, show that genetic selection may result in the eradication of an infectious disease via the mechanism of herd immunity, just like with vaccination (FINE 1993). This result contradicts predictions based on observed-scale heritability for non-communicable binary traits, where heritability vanishes when prevalence approaches zero (Robertson, 1950).

While quantitative geneticists and breeders typically focus on individual disease status and (implicitly) interpret prevalence as an average of individual trait values, epidemiologists interpret the endemic prevalence of an infectious disease as the result of a population level process of disease transmission (Kermack and McKendrick, 1927; Keeling and Rohani, 2011; Diekmann *et al.* 2013). In the latter perspective, the prevalence is an emergent trait of a population, similar to the size of a termite colony or the number of prey caught by a hunting pack, rather than an average of individual trait values. Because such emergent traits do not belong to single individuals, we cannot apply the common partitioning of individual phenotypic values into individual additive genetic values (breeding values) and non-heritable residuals ("environment"). Nevertheless, the genetic effects that determine the response to selection in an emergent trait and the heritable variation for an emergent trait can be defined by using an approach based on the so-called total heritable variation (Bijma 2011). The total heritable variation in a trait is based on the individual genetic effects on the level of the emergent trait, rather than on a decomposition of individual trait values into heritable and residual effects. This suggests we can develop a quantitative genetic theory for the endemic prevalence of infectious diseases by combining epidemiological theory with the total heritable variation approach.

Here we propose a quantitative genetic theory for the endemic prevalence of infectious diseases. We first identify the genetic factors that determine the prevalence of an infectious disease. Similar to the threshold model, we will assume an underlying additive infinitesimal model for these genetic factors. However, the link between the underlying additive scale and the observed endemic prevalence will be founded in epidemiological theory, with a key role for the basic reproduction number (R_0). Subsequently we investigate the population level effects of genetic variation in individual disease traits on the level of R_0 and on the endemic prevalence in the population. Next, we move to the individual level, and derive expressions for the breeding value and heritable variation, for both prevalence and individual binary disease status, and show how these parameters depend on the endemic prevalence. Results will show that heritable variation for endemic prevalence increases when prevalence approaches zero, while heritability of individual disease status goes to zero. Then we investigate response to selection, and show that response of prevalence to selection accelerates considerably when prevalence goes down. Finally, we partition the breeding value for prevalence into direct and indirect genetic effects, and show that most of the heritable variation in the endemic prevalence of the infection is indirect, and thus hidden to classical genetic analysis and selection. We focus solely on the development of quantitative genetic theory, and do not consider the statistical estimation of the genetic effects underlying prevalence. Such methods have been developed elsewhere (Anacleto *et al.* 2015; Biemans *et al.* 2017; Pooley *et al.* 2020).

Theory and Results

1. The genetic factors that determine prevalence and R₀.

We consider an endemic infectious disease, where individuals can either be susceptible (*i.e.*, non-infected), denoted by S, or infected, denoted by I. We use corresponding symbols in *italics* to denote the number of individuals with that status. Thus, with a total of N individuals in the population in which the endemic takes place, S denotes the number of susceptible individuals, I the number of infected individuals, and $S + I = N$ (see Table 1 for a notation key). We will assume that infected individuals are also infectious, and can thus infect others (see Discussion). This model is known as the SIS compartmental model (Hethcote, 1989), and was first discussed by Weiss and Dishon (1971).

The prevalence (P) of an endemic infection is defined as the fraction of the population infected (Diekmann *et al.* 2013),

$$P = \frac{I}{N} \tag{1}$$

When individual disease status is coded in a binary fashion, using $y = 0$ for non-infected individuals and $y = 1$ for infected individuals, the prevalence is equal to the average individual disease status in the population,

$$P = \bar{y}. \tag{2}$$

The prevalence of an infectious disease depends on the so-called basic reproduction number of the disease (R_0). The R_0 is defined as the average number of individuals that gets infected by a typical infected individual in an otherwise non-infected population, and is a property of the population (Kermack and McKendrick, 1927; Anderson and May, 1979; Diekmann *et al.* 1990). When $R_0 > 1$, an average infected individual on average infects more than one new individual, and the infection can persist in the population.

The prevalence of an endemic infection reaches an equilibrium value, known as the endemic prevalence, when a single infected individual on average infects one other individual ($R = 1$). This occurs when the product of R_0 and the fraction of contact individuals that is susceptible is equal to 1; $R_0\,(1 - P) = 1$. For example, when $R_0 = 3$, an infected individual could in principle infect three other individuals. However, when only one third of its contact individuals is susceptible (*i.e.*, not infected), meaning $1–P = 1/3$, then the effective reproduction number equals $3 \times 1/3 = 1$. Hence, when $1–P =$

1/3, an infected individual is on average replaced by a single newly infected individual, so that an equilibrium occurs at $P = 1 - 1/3 = 2/3$. The endemic prevalence, therefore, is given by (Weiss and Dishon 1971).

$$P = 1 - 1/R_0. \tag{3}$$

Throughout, we will use the symbol P to denote prevalence in the endemic equilibrium, unless stated otherwise, and we will refer to P as the endemic prevalence. The actual prevalence tends to fluctuate around the equilibrium value because of random perturbations and transient effects, for example when new animals replace some of the resident animals. Equation 3 is an approximation when there is variation among individuals, which will be addressed in section 4 below.

Figure 1 illustrates the relationship between the endemic prevalence and R_0. When R_0 is smaller than one the endemic prevalence is zero (the infection is not present in the long run), and Equation 3 does not apply. For large R_0 the endemic prevalence asymptotes to 1. Note that the curve is steeper the closer R_0 is to 1. This pattern will have considerable consequences for the relationship between the heritable variation in the endemic prevalence and the level of the endemic prevalence, as will be shown in section 6 of this manuscript.

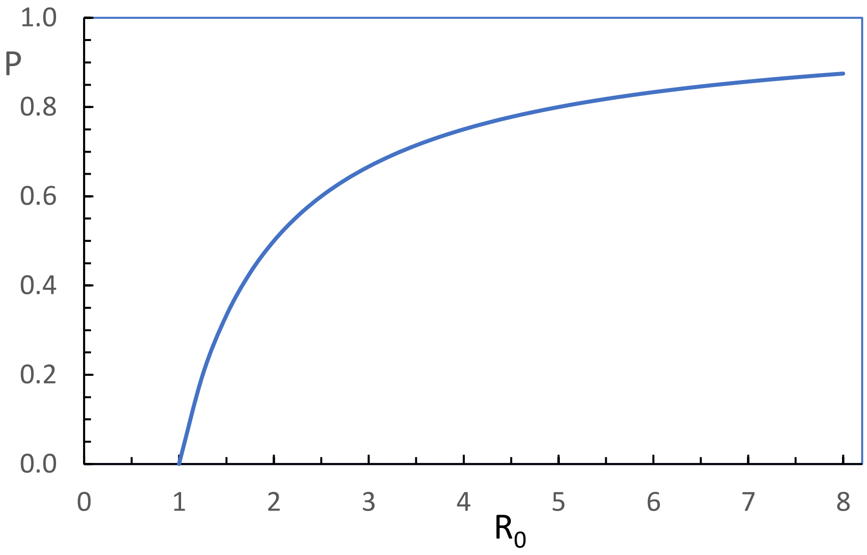


Figure 1 – The relationship between the endemic prevalence (P) and the basic reproduction number (R_0). From Equation 3.

Because the endemic prevalence is determined by R_0 (Equation 3), the response of prevalence to selection, *i.e.*, the genetic change in the endemic prevalence from one (host) generation to the next, follows from the genetic change in R_0. Thus, to measure the value of an individual with respect to response to selection, we should base this measure on the genetic impact of the individual on R_0. In other words, the definition of an individual breeding value for endemic prevalence should be based on R_0. The next step, therefore, is to find the individual genetic factors underlying R_0.

In the absence of variation among individuals (commonly referred to as "heterogeneity" in the epidemiological literature), R_0 is the product of the transmission rate parameter (β) and the mean duration of the infectious period (1/α; Kermack and McKendrick, 1927; Diekmann *et al*., 1990),

$$R_0 = \beta/\alpha. \tag{4}$$

where α is the recovery rate parameter. The β is the average number of individuals infected per unit of time by a single infected individual when all its contact individuals are susceptible, and α is the probability per unit of time for an infected individual to recover. With heterogeneity, Equation 4 is an approximation (Diekmann *et al*., 1990), which we will address in section 3 of this manuscript.

With heterogeneity, the transmission rate parameter may vary among pairs of individuals. The transmission rate parameter between infectious individual *j* and susceptible individual *i* may be modelled as the product of an overall effective contact rate (*c*), the susceptibility (γ) of recipient individual *i* and the infectivity (φ) of donor individual *j* (*e.g.*, Lipschutz-Powell *et al.*, 2014; Anacleto *et al.* 2015; Biemans *et al.* 2017),

$$\beta_{ij} = c\gamma_i\varphi_j \tag{5}$$

Hence, β_{ij} refers to transmission from individual *j* to *i*, and may differ from β_{ji}.

Equations 3 through 5 show that the factors underlying the endemic prevalence of an infection are the contact rate *c*, the susceptibility, γ, the infectivity, φ, and the recovery rate α. The *c* is a fixed parameter for the population (or, for example, for a sex, herd or age class combination), whereas γ, φ and α may show random variation among individuals. To fix the scale of Equations 4 and 5, it is convenient to include the scale in *c*, and to express γ, φ and α relative to a value of 1. Hence, with this parameterisation, the *c* is on the scale of R_0. In the absence of heterogeneity, R_0 and *c* are identical. With heterogeneity R_0 may deviate a bit from *c*.

2. Genetic models for susceptibility, infectivity, recovery and R_0

Genetic variation is potentially present in susceptibility, infectivity and the recovery rate. In this section we propose a genetic model for these traits, which subsequently leads to a genetic model for R_0.

We assume that susceptibility, infectivity and the recovery rate are affected by a large number of loci, each of small effect, so that genetic effects approximately follow a Normal distribution. However, as γ, φ, and α represent rates, *i.e.*, probabilities per unit of time, their values are strictly positive. For this reason, following Anacleto *et al.* (2015), we define normally distributed additive genetic effects for the logarithm of the rates, so that the rates themselves follow a log-normal distribution,

$$\gamma_i = e^{A_{l\gamma,i}} \tag{6a}$$

$$\varphi_i = e^{A_{l\varphi,i}} \tag{6b}$$

$$\alpha_i = e^{A_{l\alpha,i}} \tag{6c}$$

where $A_{l..,i}$ denotes the Normally distributed breeding value for the logarithm of the corresponding rate for individual i, and has a mean of zero,

$$\begin{bmatrix} A_{l\gamma} \\ A_{l\varphi} \\ A_{l\alpha} \end{bmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{A_{l\gamma}} & \sigma_{A_{l\gamma}A_{l\varphi}} & \sigma_{A_{l\gamma}A_{l\alpha}} \\ \sigma_{A_{l\gamma}A_{l\varphi}} & \sigma^2_{A_{l\varphi}} & \sigma_{A_{l\varphi}A_{l\alpha}} \\ \sigma_{A_{l\gamma}A_{l\alpha}} & \sigma_{A_{l\varphi}A_{l\alpha}} & \sigma^2_{A_{l\alpha}} \end{pmatrix}\right]. \tag{7}$$

Throughout, we use subscript l to denote the natural logarithm. Thus, the breeding values for $\log(\gamma)$, $\log(\varphi)$ and $\log(\alpha)$ follow a multivariate normal distribution, as common in quantitative genetics. Moreover, for the average individual the $A_{l..} = 0$, so that its rates are equal to one ($\gamma = \varphi = \alpha = 1$). Hence, those rates should be interpreted relative to a value of 1. An individual with $\gamma = 2$, for example, is twice as susceptible as the average individual.

The breeding values on the log-scale can approximately be interpreted as a relative change of the corresponding rate. For example, since $e^{0.1} \approx 1.1$, and $A_{l\gamma}$ of 0.1 corresponds approximately to a 10% greater than average susceptibility ($\gamma \approx 1.1$). Similarly, and $A_{l\gamma}$ of -0.1 corresponds approximately to a 10% smaller than average susceptibility ($\gamma \approx 0.9$). Realistic values for the genetic variances on the log-scale are probably smaller than ~0.5^2 (Hulst *et al*., 2021). For example, with $\sigma^2_{A_{l\gamma}} = 0.5^2$, the 10% least susceptible individuals have $\bar{\gamma} = e^{-0.88} = 0.42$, while the 10% most susceptible individuals have $\bar{y} = e^{0.88} = 2.40$. Thus average susceptibilities of these top and bottom 10% of individuals differ by a factor of 5.7, which is substantial. Therefore, we will consider additive genetic variances on the log-scale no greater than 0.5^2. With a prevalence of 0.3, this value corresponds to an observed-scale heritability of individual binary disease status of about 0.05 (Hulst *et al*., 2021).

Genotypic value and breeding value for R_0: Based on Equations 4 and 5, we may define an individual genotypic value for R_0 (see also Anche *et al.* 2014 and Biemans *et al.* 2019),

$$G_{R_0,i} = c\gamma_i\varphi_i/\alpha_i \tag{8}$$

In contrast to the pair-wise transmission rate parameter β_{ij} in Equation 5, an individual's genotypic value for R_0 is entirely a function of its own rates, as can be seen from the index i on all elements of Equation 8. This is because $G_{R_0,i}$ refers to the genetic effects that originate from the individual, rather than to those that affects its trait value. Hence, $G_{R_0,i}$ represents a total genotypic value, including both direct and indirect genetic effects (Bijma *et al*., 2007a; Bijma 2011). We focus on the total genotypic value, because our ultimate interest is in response to selection (Section 7). In section 3 of this manuscript, we will show that R_0 is indeed the simple average of G_{R_0}.

From Equations 6 and 8,

$$\begin{aligned} G_{R_0,i} &= c\, e^{A_{l\gamma,i}}\, e^{A_{l\varphi,i}}/e^{A_{l\alpha,i}} \\ &= e^{\ln(c) + A_{l\gamma,i} + A_{l\varphi,i} - A_{l\alpha,i}} \\ &= e^{\ln(c) + A_{lR_0,i}}, \end{aligned} \tag{9}$$

where $A_{lR_0,i}$ is a Normally distributed additive genetic effect (breeding value) for the logarithm of R_0,

$$A_{lR_0,i} = A_{l\gamma,i} + A_{l\varphi,i} - A_{l\alpha,i} \tag{10a}$$

$$A_{lR_0} \sim N(0, \sigma^2_{A_{lR_0}}) \tag{10b}$$

$$\sigma^2_{A_{lR_0}} = \sigma^2_{A_{l\gamma}} + 2\sigma_{A_{l\gamma}A_{l\varphi}} - 2\sigma_{A_{l\gamma}A_{l\alpha}} + \sigma^2_{A_{l\varphi}} - 2\sigma_{A_{l\varphi}A_{l\alpha}} + \sigma^2_{A_{l\alpha}} \tag{10c}$$

Hence, our model of the genotypic value for R_0 is additive with Normally distributed effects on the log-scale. The breeding value for the logarithm or R_0 will play a central role in the remainder of this manuscript.

It follows that the genotypic value for R_0, as defined in Equations 8 and 9, follows a log-normal distribution,

$$G_{R_0} \sim logN\left(\mu = \ln(c), \sigma^2 = \sigma^2_{A_{lR_0}}\right). \tag{11}$$

The genotypic value for R_0 for the average individual, which has $A_{lR_0} = 0$, is equal to the contact rate, c. Hence, the genotypic value is defined including its average, it is not expressed as a deviation from the mean. Moreover, we refer to G_{R_0} as a genotypic value, rather than a breeding value, because the $e^{A_{lR_0}}$ in Equation 9 is a non-linear function, so that G_{R_0} will show some non-additive genetic variance even though A_{lR_0} is additive.

The log-normal distribution of G_{R_0} agrees with the infinitesimal model and the strictly positive values for R_0 (Anacleto *et al.*, 2015), and is also convenient because the mean and variance of G_{R_0} follow from the known properties of the log-normal distribution,

$$\mathrm{E}\left(G_{R_0}\right) = c\, e^{\frac{1}{2}\sigma^2_{A_{lR_0}}} \tag{12}$$

$$\mathrm{var}\left(G_{R_0}\right) = c^2\left(e^{2\sigma^2_{A_{lR_0}}} - e^{\sigma^2_{A_{lR_0}}}\right) \tag{13}$$

With realistic levels of heterogeneity, the mean genotypic value is close to the contract rate, c. For example, for $\sigma^2_{A_{lR_0}} = 0.5^2$, $\mathrm{E}\left(G_{R_0}\right) \approx 1.13c$.

Equations 12 and 13 show that a log-normal distribution for susceptibility, infectivity and recovery results in a positive mean-variance relationship for G_{R_0}. Figure 2 illustrates this relationship, for $\sigma^2_{A_{lR_0}} = 0.3^2$ and genetic variation in susceptibility only. The *x*-axis shows the contact rate, which is equal to the genotypic value for R_0 of the average individual. Hence, the *x*-axis reflects the level of R_0. The small circle represents a population with a prevalence of ~0.33, for which observed-scale heritability of binary disease status is ~0.02 (Hulst *et al.*, 2021). For that population, R_0 is ~1.5, and the genetic standard deviation in R_0 is ~0.48. Hence, despite the small observed-scale heritability, R_0 has considerable genetic variation and some individuals will have a genotypic value smaller than 1, which agrees with the findings of Hulst *et al.* (2021). In the context of artificial selection against infectious diseases, the positive mean-variance relationship resulting from our model may be interpreted as conservative, because it implies a reduction of the genetic variance in R_0 with continued selection for lower prevalence.

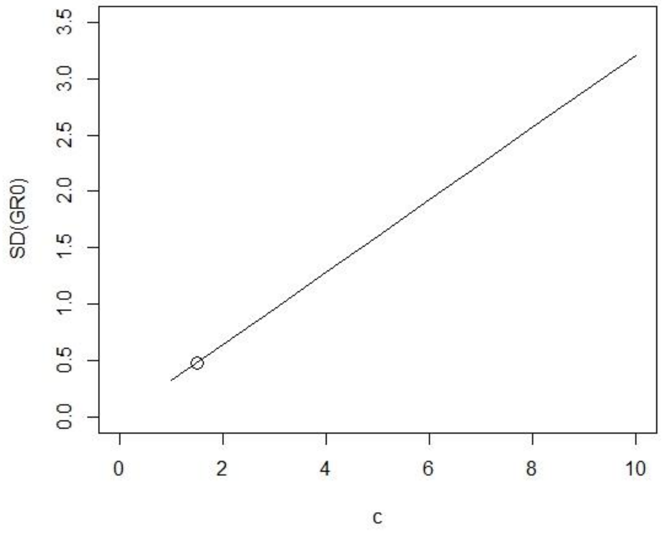


Figure 2 - Genetic standard deviation in R_0 as a function of the level of R_0 (as measured by the contact rate *c*). From Equation 13, for $\sigma^2_{A_{IR_0}} = \sigma^2_{A_{I\gamma}} = 0.3^2$, and no variation in φ and α. The circle represents a population with a prevalence of ~0.33, for which observed-scale heritability of binary disease status is ~0.02 (Hulst *et al*., 2021).

In summary, this section has presented a genetic model for susceptibility, infectivity and recovery, leading to expressions for the genotypic value and genetic variance in R_0. In the next two sections, we will focus on the population-level consequences of genetic heterogeneity, and investigate the impact of genetic variation in R_0 on the level of R_0 and on the endemic prevalence. In the classical quantitative genetic model, genetic variation only affects the variation among individuals; not the population average trait value. However, because R_0 and the endemic prevalence are the result of a transmission process in the population, genetic variation among individuals may affect their value (Diekmann *et al*., 1990).

3. The impact of genetic heterogeneity on *R₀*

R_0 is a key parameter for infectious diseases, because infectious diseases can persist in a population only when R_0 is greater than one (Kermack & McKendrik,1927; Diekmann *et al*. 1990). In other words, an endemic equilibrium can exist only when R_0 is greater than one. Conversely, eradication of an infectious disease, either by vaccination or other measures such as genetic selection of the host population, requires that R_0 is reduced to a value smaller than one. Here we address the consequences of genetic variation in susceptibility, infectivity and recovery for the value of R_0. Note that R_0 is strictly defined for the disease free state of the population (*i.e.*, where the infected fraction is infinitesimally small). Hence, in this section we consider the disease free state, while the endemic equilibrium will be addressed in section 4.

As stated above, R_0 is the average number of individuals that gets infected by a *typical* infected individual in an otherwise non-infected population (Kermack and McKendrick, 1927; Diekmann *et al.*, 1990). The expression for R_0 given in Equation 4 ignores the “typical” term in the definition of R_0, and is therefore an approximation with heterogeneity.

The properties of the “typical infected individual” will depend on the magnitude and nature of the heterogeneity among the individuals in the population. In contrast to the conclusion of Springbett *et al.* (2003), therefore, genetic heterogeneity may affect R_0 (Diekmann *et al.* 1990, 2013). Suppose, for example, that individuals differ in both susceptibility and infectivity, and that susceptibility is positively correlated to infectivity. Because individuals with greater susceptibility are more likely to become infected, the typical infected individual will have an above-average susceptibility. Moreover, because of the positive correlation with infectivity, this will also translate into an above average infectivity of the typical infected individual, leading to higher R_0. Hence, variation among individuals together with a positive (negative) correlation between susceptibility and infectivity results in an increase (decrease) in R_0 (Diekmann *et al.* 1990). For this reason, R_0 may deviate from the right-hand side of Equation 4.

In Appendix 1, we derive the relationship between R_0 and the genetic parameters for susceptibility, infectivity and recovery. The first step is the derivation of the infectivity of the typical infected individual. The result shows that the life-time infectivity (ϕ) of the typical infected individual equals

$$\bar{\phi}_{typ} = \bar{\phi} e^{\sigma_{A_{l\gamma}A_{l\phi}}} \tag{14}$$

where $\bar{\phi}$ is the simple average of life-time infectivity, and $\sigma_{A_{l\gamma}A_{l\phi}}$ the covariance between the breeding values for the logarithms of susceptibility and life-time infectivity. Life-time infectivity is the total infectivity of an individual, aggregated over its infectious period, and is the product of its infectivity per unit of time (φ, Equations 5 & 6) and the mean duration of its infectious period, $1/\alpha_i$,

$$\phi_i = \varphi_i/\alpha_i. \tag{15}$$

Equation 14 shows that the typical infected individual has an above (below) average life-time infectivity when the covariance between susceptibility and life-time infectivity is positive (negative), as argued verbally in the previous paragraph.

From the definition of R_0 and Equations 4, 5 and 14, it follows that

$$R_0 = c\,\bar{\gamma}\,\bar{\phi}\,e^{\sigma_{A_{l\gamma}A_{l\phi}}} \tag{16}$$

where $\bar{\gamma}$ is the simple population average value of susceptibility. The last term of this expression shows that a positive covariance between susceptibility and life-time infectivity indeed increases R_0.

Equation 16 can be simplified by substituting the expression for $\bar{\gamma}$ and $\bar{\phi}$, which follow from the log-normal distribution,

$$\bar{\gamma} = e^{\frac{1}{2}\sigma^2_{A_{l\gamma}}} \tag{17a}$$

$$\bar{\phi} = e^{\frac{1}{2}\sigma^2_{A_{l\phi}}} \tag{17b}$$

Substituting Equations 17a and b into Equation 16, and expressing the genetic variance of life-time infectivity in terms of infectivity per unit of time and recovery, reveals that R_0 is equal to the simple average individual genotypic value for R_0 (see Appendix 1 and Equation 12),

$$R_0 = c\, e^{\frac{1}{2}\sigma^2_{A_{lR_0}}} = \mathrm{E}(G_{R_0}) \tag{18}$$

Thus, while a positive covariance between susceptibility and infectivity indeed increases R_0, this effect is fully captured by the variance in the breeding values for the logarithm of R_0 (the $e^{\frac{1}{2}\sigma^2_{A_{lR_0}}}$ term in Equation 18). As shown in Equation 10c, a positive covariance between susceptibility and infectivity increases $\sigma^2_{A_{lR_0}}$, and this effect fully accounts for the effect of the $e^{\sigma_{A_{l\gamma}A_{l\phi}}}$ term in Equation 16. Thus, R_0 is equal to the simple average of individual genotypic values for R_0 as long as those genotypic values follow a log-normal distribution.

Note that, while R_0 is equal to the simple average of genotypic values for R_0, it may still differ from the product of the simple averages of the rates; $R_0 \neq c\,\bar{\gamma}\,\bar{\varphi}\,/\bar{\alpha}$ when susceptibility, infectivity and/or recovery are correlated. Moreover, R_0 may also differ from $\bar{\beta}/\bar{\alpha}$ with heterogeneity. A numerical investigation of the $e^{\sigma_{A_{l\gamma}A_{l\phi}}}$ term in Equation 16, however, shows that the effect of a correlation between susceptibility and life time infectivity on R_0 is unlikely to be greater than ~ ±25% when those traits are polygenic. For example, for $\sigma^2_{A_{l\gamma}} = \sigma^2_{A_{l\phi}} = 0.5^2$, and a correlation $r_{A_\gamma A_\phi} = 0.8$, R_0 is only 22% greater than $c\,\bar{\gamma}\,\bar{\varphi}\,/\bar{\alpha}$. Thus, for polygenic traits, the effect of a correlation between susceptibility and life-time infectivity on R_0 is expected to be limited.

In summary, this section has shown that heterogeneity and a positive correlation between susceptibly and life-time infectivity lead to an increase of R_0, and thus increase the probability that an infectious disease persists in the population. However, when genotypic values for R_0 follow a log-normal distribution, R_0 is still equal to the simple average of those genotypic values.

4. The impact of genetic variation on the equilibrium prevalence

In this section we present an expression for the endemic prevalence in a population with genetic variation in susceptibility, infectivity and recovery, and also briefly investigate the consequences of such variation for the endemic prevalence. Figure 1 and Equation 3 show the relationship between R_0 and the endemic prevalence for a homogeneous population. With variation among individuals, however, highly susceptible individuals are likely to be in the infected state in the endemic equilibrium. For this reason, the mean susceptibility of the remaining non-infected individuals will be lower than the population average susceptibility. This in turn translates into a prevalence that is lower than the value given by Equation 3 (See Springbett *et al.*, 2003 for epidemic infections). Similar arguments can be used to show that prevalence depends on the variation in the recovery rate and on the correlation of infectivity with susceptibility and recovery. Thus Equation 3 is exact only in the absence of heterogeneity.

The equilibrium prevalence in a heterogeneous population can be derived by realizing that the prevalence must have reached an equilibrium value for each type of individual (Biemans *et al.*, 2017; Aznar *et al.*, 2018). Suppose, for example, that susceptibility, infectivity, and recovery would be governed by a single bi-allelic locus in a diploid organism. Then, for the entire population to be in equilibrium, each of the three genotypic classes should be in equilibrium as well. In other words, the prevalence should have reached an equilibrium value within each genotypic class, but this value may differ among the three classes. Here we adapt this approach to continuous variation in polygenic traits.

In the endemic equilibrium, the number of susceptible individuals of each type, say *i*, should not change over time (apart from random fluctuation). Thus the number of newly infected susceptibles should be equal to the number of recovering infecteds for each type *i*,

$$\frac{dS_i}{dt} = -c\gamma_i\,\bar{\varphi}_{\text{inf}}\,S_i\frac{I}{N} + \alpha_i I_i = 0 \tag{19}$$

where S_i is the number of susceptible individuals of type *i*, *t* denotes time, *c* the contact rate, γ_i the susceptibility of type *i*, $\bar{\varphi}_{\text{inf}}$ the mean infectivity of the infected individuals in the endemic equilibrium, *I* the total number of infected individuals, *N* total population size, α_i the recovery rate for type *i*, and I_i the number of infected individuals of type *i*. Above, we used *i* to index individuals. Here we also use *i* to index classes, since each individual will be genetically unique with polygenic traits, so that a type corresponds to an individual (S_i and I_i may be treated as non-integer). The first term in Equation 19 represents the decrease in the number of susceptibles due to transmission (infection), while the second term represents the increase of the number of susceptibles due to recovery of infected individuals. Note that the mean infectivity of the infected individuals in the endemic equilibrium ($\bar{\varphi}_{\text{inf}}$) will differ from the simple population average of infectivity ($\bar{\varphi}$) when infectivity is correlated to susceptibility and/or recovery.

Equation 19 can be solved for the endemic prevalence in type *i*, $P_i = I_i/N_i$, N_i denoting the total number of individuals of type *i* in the population,

$$P_i = \frac{\mathcal{R}_{0,i}P}{\mathcal{R}_{0,i}P+1} \tag{20a}$$

where *P* denotes the overall endemic prevalence in the population (Equation 1), and

$$\mathcal{R}_{0,i} = \frac{c\gamma_i\bar{\varphi}_{\text{inf}}}{\alpha_i} \tag{20b}$$

Equations 20a&b make no assumptions on the distribution of γ, φ and α, and are thus not restricted to log-normal distributions. Although Equation 20b is similar to Equation 8, note that $\mathcal{R}_{0,i}$ differs from the genotypic value for R_0 ($G_{R_0,i}$; We use a symbol slightly different from *R* to highlight this difference). The $\mathcal{R}_{0,i}$ is a function of the mean infectivity of the infected individuals in the endemic equilibrium ($\bar{\varphi}_{\text{inf}}$), while $G_{R_0,i}$ is a function of the infectivity of the individual itself (φ_i). Our interest here is in the prevalence for an individual with susceptibility γ_i and recovery rate α_i in the endemic equilibrium, where *i* is exposed to the mean infectivity of the infected individuals. Hence the $\bar{\varphi}_{\text{inf}}$ term in $\mathcal{R}_{0,i}$. The $G_{R_0,i}$, in contrast, defines the contribution of an individual to R_0 (Equation 18), and will be relevant for response to selection (section 7).

To find the endemic prevalence, we need to solve Equations 20a and b for *P*. While we found an analytical solution for the case without (correlated) genetic variation in infectivity, the resulting expression is very complex (not shown). We therefore used a numerical solution, which is easily obtained (see Appendix 2 for methods, and Supplementary Material 1 for an R-code). We validated the numerically obtained solution using full stochastic simulation of actual endemics, following standard methods in epidemiology. Results of these simulations confirmed the numerically obtained solutions (see Appendix 3 for methods).

The solutions of Equations 20a and b show two opposing effects of heterogeneity on the endemic prevalence. First, with a log-normal distribution of the rates, heterogeneity in susceptibility, infectivity and/or recovery increases R_0 (Equations 10c and 18), which in turn increases the equilibrium prevalence. Second, at the same R_0, variation in susceptibility and/or recovery reduces the equilibrium prevalence (Equations 20a and b), as argued in the first paragraph of this section. Hence, with variation in susceptibility and/or recovery, prevalence is lower than predicted by Equation 3. Genetic variation in infectivity has no effect on the prevalence beyond its effect on R_0, as long as infectivity is not correlated to susceptibility and/or recovery.

Moreover, the effects of genetic variation on the prevalence are identical for susceptibility and recovery. It follows from Equation 20b that the $\mathcal{R}_{0,i}$ of an individual depends on the difference between its breeding values for log-susceptibility and log-recovery, $A_{l\gamma,i} - A_{l\alpha,i}$,

$$\mathcal{R}_{0,i} = c\bar{\varphi}_{\text{inf}}\, e^{A_{\gamma_i} - A_{\alpha_i}} \tag{21}$$

Hence, in the absence of correlated variation in infectivity, the equilibrium prevalence depends only on the variance of this difference,

$$var\left(A_{l\gamma} - A_{l\alpha}\right) = \sigma^2_{A_{l\gamma}} - 2\sigma_{A_{l\gamma}A_{l\alpha}} + \sigma^2_{A_{l\alpha}}. \tag{22}$$

Finally, when $c = 2$ and there is no variation in infectivity, prevalence is always equal to $1 - 1/c = 0.5$, irrespective of the genetic variation in susceptibility and infectivity. This occurs because the two effects mentioned at the beginning of this paragraph exactly cancel each other. When $c < 2$, prevalence is (much) higher than $1 - 1/c$, while prevalence is only a little lower than $1 - 1/c$ when $c > 2$. Figure 3 illustrates the impact of heterogeneity on the endemic prevalence for a limited number of scenarios. More detailed results can be found in Supplementary Material 2.

In conclusion, for a contact rate greater than ~1.7 ($P \approx 0.4$), the endemic prevalence is very similar to 1-1/c. For small values of the contact rate, the endemic prevalence is larger than 1-1/c.

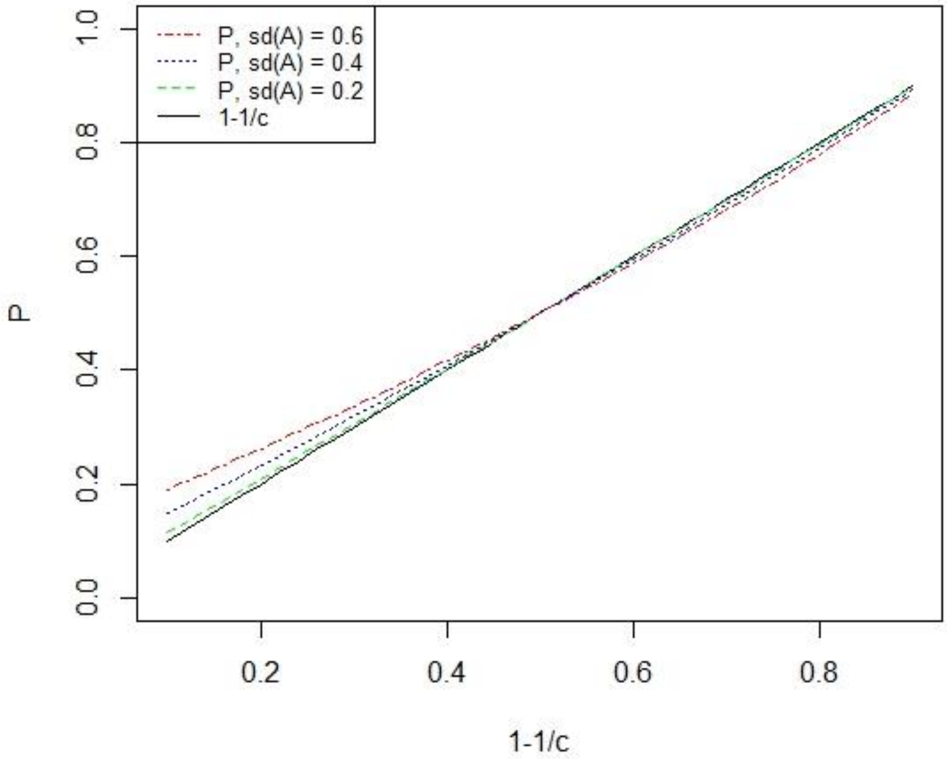


Figure 3 - The impact of heterogeneity on the endemic prevalence. Solid lines show the true prevalence, and the dashed lines the prediction based on Equation 3 using the true R_0. The dotted line uses Equation 3 assuming $R_0 = c$. For three levels of the additive genetic standard deviation in log susceptibility ($\sigma_{A_{l\gamma}}$) and no genetic variation in infectivity or recovery. Note that identical results would have been obtained with the same amount of heterogeneity in the recovery rate ($\sigma_{A_{l\alpha}}$) instead of the susceptibility.

5. Genotypic value for individual binary disease status

In the previous two sections, we have considered the population-level effects of genetic heterogeneity. In the next two sections, we move to the individual level. Section 5 focusses on the effects of an individual's genes on its own diseases status, while section 6 focusses on the effect of an individual's genes on the prevalence in the population.

By definition, the genotypic value for binary disease status is the expected disease status of an individual given its genotype,

$$G_{y,i} = E(y_i | A_{\gamma,i}, A_{\alpha,i}) \tag{23}$$

Thus, the G_y represents the direct genetic effect (DGE) on the own phenotypic value (including the mean, $\bar{y}$, here). The genotypic value of an individual is not a function of its breeding value for log-infectivity, since an individual's infectivity does not affect its own disease status. Hence, Equation 23 does not condition on $A_{\varphi,i}$.

In the previous section, we used Equations 20a and b to investigate the effect of heterogeneity on the endemic prevalence in the population. Equation 20a shows the expected prevalence of an individual of type i. However, since prevalence is simply the mean of binary disease status, Equation 20a may also be interpreted as the expected phenotypic value for disease status (y = 0,1) of an individual, given its genotype (specifically, the γ_i and α_i components of $\mathcal{R}_{0,i}$). Hence, Equation 20a also represents the genotypic value for binary disease status,

$$G_{y,i} = \frac{\mathcal{R}_{0,i}P}{\mathcal{R}_{0,i}P+1} \tag{24}$$

where $\mathcal{R}_{0,i}$ follows from Equation 20b, and P here denotes the prevalence i is exposed to. (Strictly, this does not need to be the prevalence in the endemic equilibrium). The same result was found by Bijma (2020), but based on a different approach. (Note the distinction between subscript y, indicating individual binary disease status, and γ, indicating susceptibility). Thus the G_y refers to the expected binary disease status of individual i in a population with prevalence P, conditional on its genotype. Equation 21 and 24 imply that susceptibility and recovery are equally important for the disease status of an individual. For example, an individual with $A_{l\gamma} = -0.1$ has the same expected disease status as an individual with $A_{l\alpha} = +0.1$.

Calculation of G_y from Equation 24 requires knowledge of the endemic prevalence P. In the previous section we used a numerical approach to find P, because our interest was in the effects of heterogeneity on P. In applied breeding, however, breeders may often have a reasonable idea of realistic values for P, and a numerical solution may not be needed, or have little added value.

Validation: We used stochastic simulation of endemics, following standard epidemiological methodology (see Appendix 3 for methods), to validate the expression for the genotypic value for individual disease status (Equation 24). Figures 4a-c show the mean observed disease status of individuals as a function of their genotypic value G_y. For all three panels in Figure 4, regression coefficients were very close to 1, showing that G_y is an unbiased linear predictor of individual disease status.

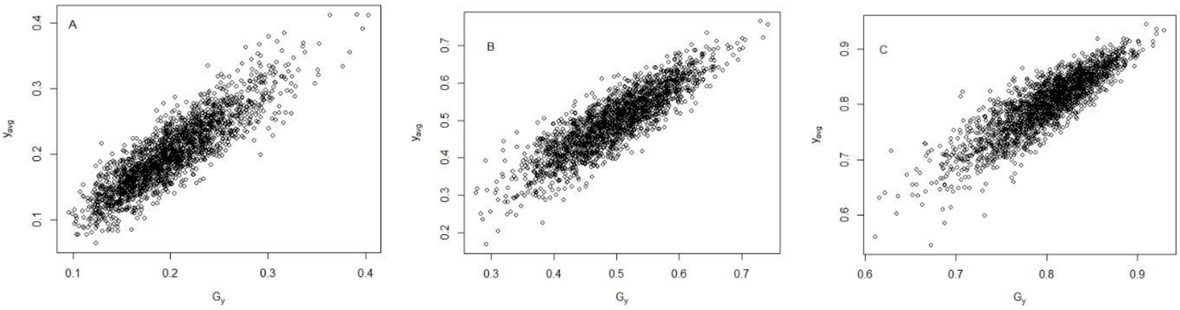


Figure 4. Validation of the genotypic value for individual binary disease status. Panels show a scatter plot of the mean disease status of individuals (y-axis) as a function of their genotypic value for disease status (G_y, x-axis, Equation 24). For genetic variation in susceptibility only, with $\sigma^2_{A_\gamma} = 0.3^2$, N = 2000 individuals, and a total of

300,000 events (sum of recoveries and infections). Panel A: P = 0.2; $\hat{b}_{y,G_\gamma} = 0.998$. Panel B: P = 0.5, $\hat{b}_{y,G_\gamma} = 1.006$. Panel C: P = 0.8, $\hat{b}_{y,G_\gamma} = 1.005$.

We numerically investigated the relative amount of non-additive genetic variance in G_y. Results (not shown) revealed only little non-additive genetic variance. For example, for P = 0.2 and $\sigma^2_{A_\gamma} = 0.5^2$, more than 96% of the genotypic variance in y was additive. Thus the breeding value for own disease status is very similar to the genotypic value,

$$A_y \approx G_y - P, \tag{25}$$

where the "$-P$" term simply reflects subtraction of the average, $\overline{G_y} = P$, so that the mean breeding value is zero by definition. We defer further investigation of the breeding value and the additive genetic variance for individual disease status to the next section, to facilitate comparison with the corresponding measures for prevalence.

6. Heritable variation for the endemic prevalence

The previous section focussed on the genetic effects of individuals on their own disease status. In this section we will consider the genetic effects of individuals on the endemic prevalence in the population. In other words, the previous section focussed on the contribution of genetic effects to the variation in disease status among individuals, while this section considers the genetic effects that are relevant for response to selection. We will present expressions for the genotypic value, breeding value and additive genetic variance for the endemic prevalence. The genotypic value will reflect the full genetic effect of an individual on the endemic prevalence in the population, while the breeding value reflects the additive component thereof. The last part of this section contains a comparison of the breeding value for endemic prevalence and that for individual disease status.

The relationship between R_0 and the endemic prevalence (Equation 3) suggests we can translate the individual genotypic value for R_0 (Equations 8 and 9) to the scale of prevalence, using

$$G_{P,i} = 1 - \frac{1}{G_{R_0,i}} \tag{26a}$$

where the G_P may be interpreted as an individual's genotypic value for prevalence. Substituting Equation 9 yields an expression for the genotypic value of an individual for the endemic prevalence, in terms of its breeding value for the logarithm of R_0,

$$G_{P,i} = 1 - e^{-\ln(c) - A_{lR_0,i}} \tag{26b}$$

Because the term in the exponent is Normally distributed, $1 - G_P$ follows a log-normal distribution,

$$(1 - G_P) \sim logN\left[\mu = -\ln(c)\,, \sigma^2 = \sigma^2_{A_{lR_0}}\right] \tag{27}$$

The mean and variance of the genotypic values for prevalence, therefore, follow from the properties of the log-normal distribution,

$$\mathrm{E}(G_P) = 1 - c^{-1}\, e^{\frac{1}{2}\sigma^2_{A_{lR_0}}} \tag{28}$$

$$\mathrm{var}(G_P) = c^{-2}\left(e^{2\sigma^2_{A_{lR_0}}} - e^{\sigma^2_{A_{lR_0}}}\right) \tag{29}$$

To enhance interpretation of Equation 29, we can express it as a function of R_0 or of the endemic prevalence. With limited heterogeneity, the contact rate c is approximately equal to R_0 and to $1/(1-P)$. Substituting these relationships into Equation (29) yields an expression for the genetic variance in prevalence as a function of R_0 and of the endemic prevalence, respectively,

$$\mathrm{var}(G_P) \approx \frac{1}{R_0^2}\left(e^{2\sigma^2_{A_{lR_0}}} - e^{\sigma^2_{A_{lR_0}}}\right) \tag{30a}$$

$$\mathrm{var}(G_P) \approx (1-P)^2\left(e^{2\sigma^2_{A_{lR_0}}} - e^{\sigma^2_{A_{lR_0}}}\right) \tag{30b}$$

Equations 30a and b show how the genotypic variance in the endemic prevalence changes with a change in R_0 or equivalently, in the endemic prevalence. Hence, in contrast to ordinary additive genetic traits, the genetic variance in endemic prevalence is a function of the level of the endemic prevalence.

Figures 5a and b illustrate that the standard deviation in genetic values for endemic prevalence is considerably larger at lower R_0, or equivalently, at lower prevalence. Hence, even though the genetic variance in R_0 decreases with the level of R_0 (Figure 2), the genetic variance in prevalence increases strongly when R_0 decreases. This result originates from the increasing slope of the relationship between prevalence and R_0 at lower R_0 (Figure 1). In other words, an equal change in R_0 has much greater impact on prevalence at low R_0 than at high R_0, which is well-known in epidemiology (*e.g.*, Bolker and Grenfell, 1996). Hence, for a constant genetic variance of the logarithm of R_0, the genetic variance in endemic prevalence is much greater at lower prevalence, and genetic selection for lower prevalence leads to an increase in the genetic variance in prevalence.

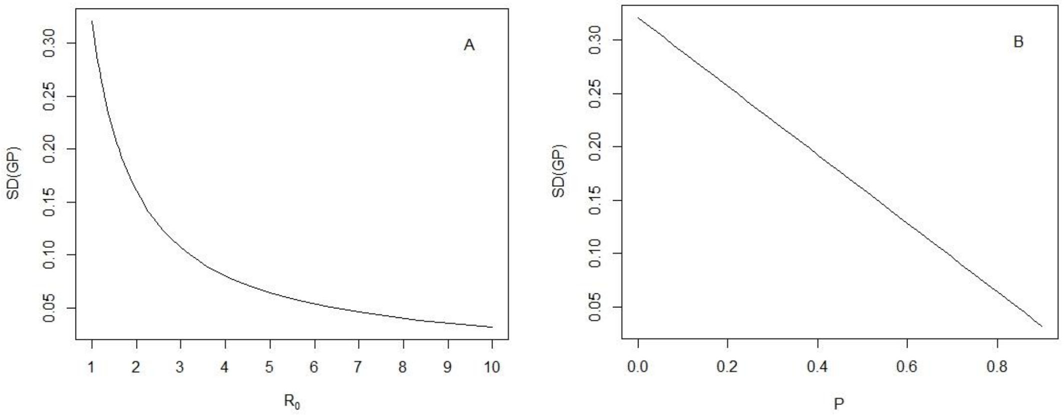


Figure 5 – Genetic standard deviation in prevalence as a function of R_0 (panel A), and as a function of the endemic prevalence (panel B). From Equations 30a and b. For $\sigma^2_{A_{lR_0}} = 0.3^2$. In panel A, *x*-axis values below $R_0 = 1$ are excluded, because equilibrium prevalence is zero (the infection is not present) and Equation 30 does not apply.

Figure 6 shows some examples of the distribution of the genotypic value for endemic prevalence, for different values of R_0 and the corresponding endemic prevalence. For the scenarios in Figure 6, the observed scale heritability of individual disease status does not exceed 0.022 (see Figure 7 below). The panels illustrate that the genotypic standard deviation in the endemic prevalence is relatively large, particularly when prevalence is small. For example, for R_0 = 1.67 (P = 0.4), the standard deviation in genotypic values for prevalence is around 0.19 (See also Figure 5), and values between ~0 and ~0.7 are quite probable. Hence, despite the low observed scale heritability of individual disease status, the probable values of G_P span as much as 70% of the full 0-1 range of endemic prevalence.

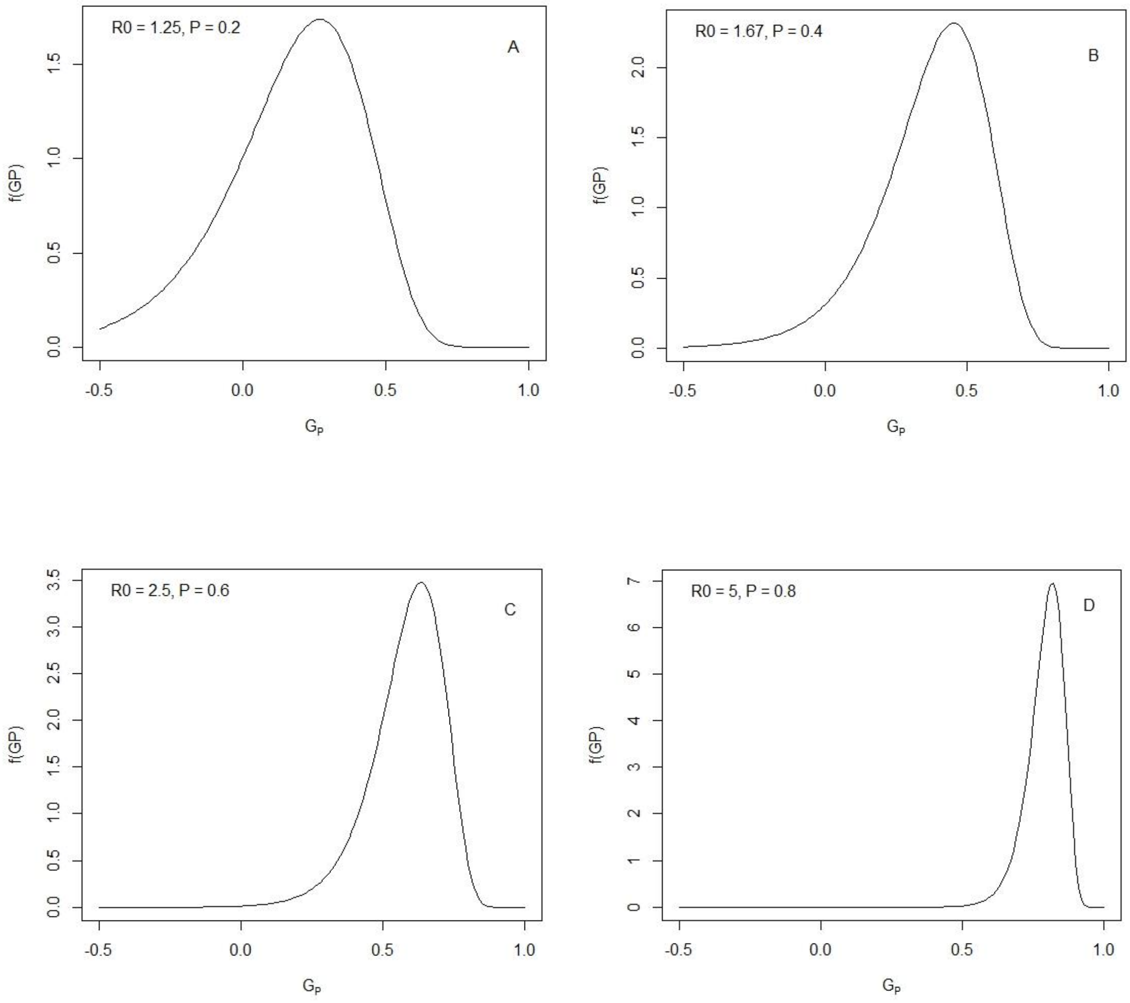


Figure 6 – Distribution of individual genotypic values for prevalence (G_P), for different values of R_0, or equivalently, different values of the endemic prevalence. The distribution is given by $f(G_P) = \frac{1}{(1-G_P)\sigma_{A_{IR_0}}\sqrt{2\pi}} exp\left(-\frac{(\log(1-G_P)+\log(c))^2}{2\sigma^2_{A_{IR_0}}}\right)$, with domain $G_P = (-\infty, 1)$. For $\sigma^2_{A_{IR_0}} = 0.3^2$. Note that G_P can take negative values while prevalence cannot. This is because G_P reflects the genetic effect of an individual on the prevalence of the population, not the expected value of its own disease status. Thus, negative values for G_P are possible, as long as P is positive. Note that P is very close to the average of the distributions shown.

Breeding value and additive genetic variance in prevalence: The genotypic value for prevalence is not identical to the additive genetic value (*i.e.*, breeding value) for prevalence, because the exponential function in Equation 26b is non-linear, so that G_P contains a non-additive component. Appendix 4 shows that the linear regression coefficient of G_P on A_{IR_0} is equal to c^{-1} exp $(\frac{1}{2}\sigma^2_{A_{IR_0}})$. Therefore, the breeding value for prevalence is given by

$$A_{P,i} = \frac{1}{c}\, e^{\frac{1}{2}\sigma^2_{A_{IR_0}}}\ A_{IR_0,i} \tag{31a}$$

With limited heterogeneity, this result is approximately equal to

$$A_{P,i} \approx \frac{1}{R_0}\, A_{IR_0,i} \tag{31b}$$

Thus the additive genetic variance in prevalence equals

$$\sigma^2_{A_P} = \frac{1}{c^2}\, e^{\sigma^2_{A_{IR_0}}}\, \sigma^2_{A_{IR_0}} \tag{32a}$$

and, with limited heterogeneity,

$$\sigma^2_{A_P} \approx \frac{1}{R_0^2}\, \sigma^2_{A_{IR_0}} \tag{32b}$$

or, expressed as a function of endemic prevalence,

$$\sigma^2_{A_P} \approx (1-P)^2\, \sigma^2_{A_{IR_0}} \tag{32c}$$

Equation 32c show that additive genetic variance in the endemic prevalence increases strongly when prevalence decreases, similar to the relationship between genotypic variance and prevalence (Figure 5). This result suggest that response of endemic prevalence to selection will be greater at lower levels of the prevalence.

The relative amount of non-additive genetic variance in the endemic prevalence is determined by the magnitude of $\sigma^2_{A_{IR_0}}$ (Appendix 4). For realistic values of $\sigma^2_{A_{IR_0}}$, the vast majority of the genotypic variance in prevalence is additive. For example, for $\sigma^2_{A_{IR_0}}$ = 0.5^2, 88% of the variance in G_P is additive. Hence, the distinction between the breeding value for prevalence (A_P) and the genotypic value for prevalence (G_P) seems of minor importance, and results in Figures 5 and 6 closely resemble those for the additive genetic effects.

Breeding value and heritability for disease status *vs.* prevalence: Appendix 5 shows that, in the absence of genetic variation in infectivity, the breeding value for prevalence is a factor 1/P greater than the breeding value for individual disease status,

$$A_{P,i} \approx \frac{1}{P} A_{y,i} \tag{33}$$

Note, in contrast to genotypic values, breeding values are expressed as a deviation from their mean. The A_y is the ordinary observed scale breeding value for binary disease status that breeders are familiar with.

This result implies that the impact of an individual's genes on response to selection is considerably larger than their impact on the disease status of the individual itself, particularly when endemic prevalence is small. Consider, for example, an individual with $A_{y,i} = -0.02$ in a population with a prevalence of P = 20%. The expected disease status of this individual in the current population equals $0.20 - 0.02 = 0.18$. Hence, on average, this individual will be infected 18% of the time. However, its breeding value for prevalence equals $A_{P,i} = -0.02/0.2 = -0.10$. Hence, if we select individuals with $A_{y,i} = -0.02$ as parents of the next generation, then the endemic prevalence will go down to 0.20 – 0.10 = 0.10. In other words, response to selection will be fivefold greater than suggested by the ordinary breeding values for individual disease status (1/P = 1/0.2 = 5). We will validate this result in section 7, which focusses on response to selection.

The relationship between the breeding value for prevalence and the breeding value for own disease status shown in Equation 33 suggests a relatively simple expressions for A_y. On combining Equations 31 and 33, and assuming limited heterogeneity, so that $e^{\frac{1}{2}\sigma^2_{A_{lR_0}}} \approx 1$ and $1/c \approx (1-P)$, the breeding value for individual disease status becomes

$$A_{y,i} \approx P(1-P)A_{lR_0,i} \tag{34}$$

The breeding value for individual disease status may also be expressed in terms of R_0 rather than P, by substituting P in Equation 34 by Equation 3. We used stochastic simulation to validate this expression and investigate its precision. Results show that Equation 34 closely matches the regression of individual binary disease status on the breeding value for the logarithm of R_0 for realistic levels of heterogeneity ($\sigma^2_{A_{lR_0}} \leq 0.5^2$; see Supplementary Material 3 for results). Hence, Equation 34 is sufficiently precise for practical application. Note that, since infectivity does not affect the disease status of an individual itself, a potential component due to infectivity has to be left out of the $A_{lR_0,i}$ term when calculating Equation 34. In other words, in Equation 34 the $A_{lR_0,i}$ should include only the breeding values for the logarithm of susceptibility and recovery (see Equation 10a).

It follows from Equation 34 that the additive genetic variance in individual binary disease status equals

$$\sigma^2_{A_y} \approx P^2(1-P)^2\sigma^2_{A_{lR_0}} \tag{35}$$

Observed-scale heritability of binary disease status follows from dividing Equation 35 by the phenotypic variance of binary disease status, $\sigma^2_y = P(1-P)$, giving

$$h^2_y \approx P(1-P)\,\sigma^2_{A_{lR_0}}. \tag{36}$$

Hence, observed-scale heritability for binary disease status has a maximum at a prevalence of 0.5, and goes to zero at a prevalence of zero or one, similar to the heritability of binary phenotypes for non-communicable polygenic traits (Robertson, 1950; Figure 7A; assuming a constant $\sigma^2_{A_{lR_0}}$).

The ratio of additive genetic variance in prevalence over phenotypic variance in binary disease status is given by,

$$T^2_P = \frac{\sigma^2_{A_P}}{P(1-P)} \tag{37}$$

with $\sigma^2_{A_P}$ taken from Equation 32. The T_P^2 is an analogy of heritability, but the numerator represents the additive genetic variance relevant for response to selection in endemic prevalence, rather than for individual binary disease status. The T_P^2, therefore, reflects the genetic variance that can be used for response to selection, whereas h_y^2 reflects the contribution of additive genetic effects to phenotypic variance (Bijma *et al*., 2007a; Bijma 2011).

Figures 7 A & B show a comparison of h_y^2 and T_P^2 for a population without genetic variation in infectivity, with genetic variances in the logarithm of R_0 ranging from 0.1^2 through 0.5^2. In Figure 7A, the maximum value of h_y^2 equals 0.0625, for P = 0.5 and $\sigma^2_{A_{lR_0}} = 0.5^2$. Given that genetic variances greater than $\sigma^2_{A_{lR_0}} = 0.5^2$ are very large (as argued above), observed scale heritabilities of binary disease status greater than ~0.06 are unlikely for endemic infectious diseases. The heritabilities in Figure 7A agree with the findings of Hulst *et al*. (2021), who used stochastic simulation of actual endemics and analysis of the resulting binary disease status data with a linear animal model. Figure 7B shows that T_P^2 increases strongly when prevalence goes down. The difference between T_P^2 and h_y^2 shows that the additive genetic variance in prevalence is (much) greater than the additive genetic variance in individual diseases status, and may even exceed phenotypic variance at low prevalence.

In conclusion, in this section we have presented expressions for the breeding value for prevalence and for individual diseases status, and for the corresponding genetic variances. With realistic levels of heterogeneity, the breeding value for prevalence is a factor $1/P$ greater than the breeding value for individual disease status. This result suggests that response to selection should be considerably greater than expected based on ordinary heritability of individual disease status. We will test this hypothesis in the next section.

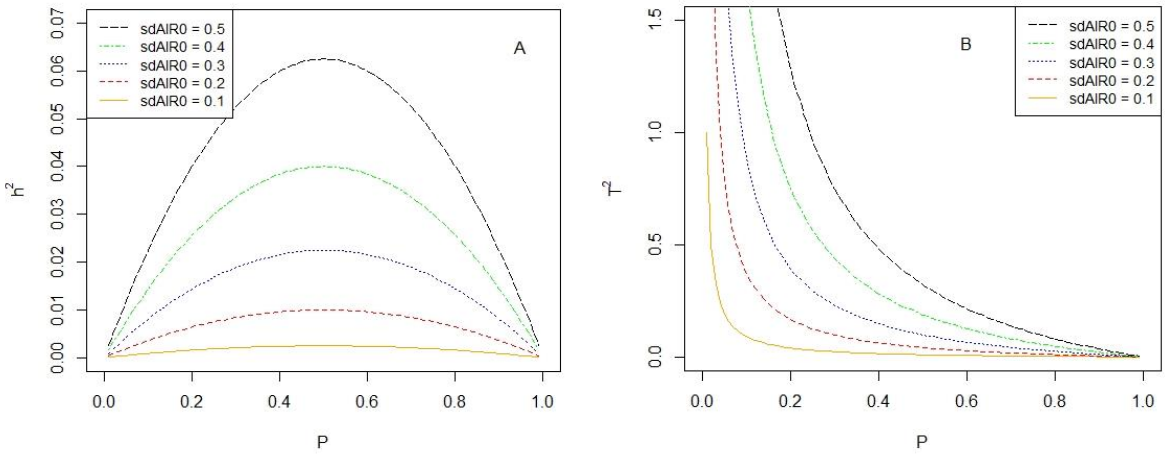


Figure 7. Panel A: Observed-scale heritability (h_y^2) of individual binary disease status (y = 0,1) as a function of the endemic prevalence, for different additive genetic standard deviations in the logarithm of R_0 (SDAlR0). From Equation 36. Panel B: Ratio of additive genetic variance in prevalence and phenotypic variance in disease status (T_P^2), as a function of the endemic prevalence. From Equations 37 and 32a. In both panels, there is no genetic variation in infectivity.

7. Response to selection

The increase in the genetic variance in prevalence when the prevalence decreases (Equations 30 & 32, Figure 5b & 6) suggests that response to selection should also increase when the prevalence decreases. To validate and illustrate this hypothesis, we stochastically simulated an endemic infectious disease in a large population undergoing mass selection based on individual disease status. Simulations were based on standard methods in epidemiology, without making use of the above theory (Appendix 6). Figure 8A shows the observed prevalence (*i.e.*, the mean binary disease status in each generation), the mean breeding value for prevalence and the mean breeding value for binary disease status, for ~70 generations of selection. Response in prevalence increases strongly when prevalence decreases, and the infection disappears in the final generation. There is excellent agreement between the observed prevalence and the breeding value for prevalence, showing that the change in $\bar{A}_P$ indeed predicts the change in prevalence. In contrast, the response in prevalence deviates substantially from the response in the breeding value for individual disease status ($\bar{A}_y$), particularly at lower values of the prevalence. Hence, while the breeding value for disease status correctly predicts individual disease status within a generation (Figure 4), the change in $\bar{A}_y$ considerably underestimates the response to selection. Furthermore, given the low value of the observe-scale heritability of binary diseases status, which did not exceed 0.022 in Figure 8A, response to selection in prevalence is quite large, unless prevalence is high.

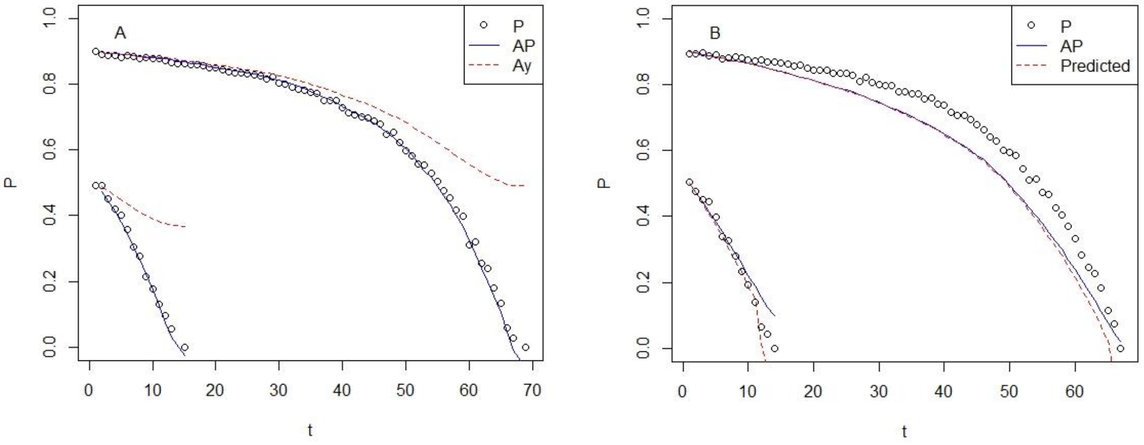


Figure 8 – Response to selection in prevalence for 70 generations of mass selection of the host population. For two populations, one starting at a prevalence of 90% (c = 10), the other starting at a prevalence of 50% (c = 2). Each generation, the 50% individuals with the lowest average disease status were selected as parents of the next generation. With genetic variation in susceptibility only, and $\sigma^2_{A_\gamma} = 0.3^2$. For a population of N = 4,000 individuals, a total of 15,000 events (sum of infections and recoveries) per generation, consisting of a burn-in of 10,000 events and 5,000 recorded events. Hence, selection is based on 1.25 events per individual on average, indicating a limited amount of phenotypic data. Observed scale heritability for binary disease status in any generation can be read from Figure 7A using an x-axis value corresponding to the prevalence in that generation. *Panel A:* Observed prevalence (circles), breeding value for prevalence ($\bar{A}_P$, solid blue line) and breeding value for individual disease status ($\bar{A}_y$, dashed red line). Results for breeding values are the cumulative change in breeding value plus the initial prevalence. *Panel B*: Predicted (lines) versus observed (circles) prevalence. Prevalence was predicted from Equation 38 (blue solid line) or Equation 39 (red dashed line).

We also briefly investigated prediction of response to mass selection with a very simple expression assuming linearity and with genetic variation in susceptibility only. We based our predictions on estimated breeding values for binary disease status, because these are typically available to breeders. Because the breeding value for prevalence and the breeding value for binary disease status differ by a factor 1/*P* when there is no genetic variation in infectivity (Equation 33), we simply upscaled the response to selection in binary disease status predicted from the breeder's equation (Walsh and Lynch, 2018) by this factor, giving

$$R_P = \iota \rho_{A_y,\bar{y}} \sigma_{A_y} \frac{1}{P} \tag{38}$$

where ι is the intensity of selection, defined as the standardized selection differential in mean individual disease status, $\iota = (\bar{y}_{\text{selected}} - \bar{\bar{y}})/\sigma_{\bar{y}}$, $\rho_{A_y,\bar{y}}$ is the accuracy of selection, which is the correlation between the selection criterion ($\bar{y}_i$) and the true breeding value for individual disease status, and *P* is the prevalence in the generation of the selection candidates. Hence, the numerator of Equation 38 represents the predicted response using parameters for binary disease status, which is multiplied by a factor 1/*P* to find response in prevalence. To implement Equation 38, we calculated the $\rho_{A_y,\bar{y}}$ as the correlation between the true breeding values for binary disease status (A_y) and the selection criterion ($\bar{y}$) in the candidates for selection. Hence, we did not attempt to predict the accuracy of selection.

Figure 8B shows a comparison of observed and predicted prevalence. Above a prevalence of ~0.5, response predicted from Equation 38 is somewhat larger than observed response, while the reverse is true below a prevalence of ~0.5. Nevertheless, agreement between observed and predicted response is remarkably good given the very unrealistic assumption of linearity in Equation 38 (*i.e.*, bivariate normality of A_y and $\bar{y}$). Because selection was based on mean individual disease status recorded over a period lasting on average only 1.25 events per individual (see legend Figure 8), many values were either 0 or 1, implying strong deviations from normality.

When prevalence was smaller than 0.5, response to selection was quite large. Hence, there was a meaningful difference in prevalence between the parent and offspring generation. Because the *P* in Equation 38 refers to the prevalence in the parent generation, while response is realized in the offspring generation, Equation 38 resulted in underprediction of response to selection when response was large. This underprediction disappeared when using prevalence in the offspring generation in the 1/*P* term in Equation 38. However, because prevalence in the offspring generation is initially unknown, as it depends on the response to selection, this prediction required solving the expression $R = i\rho_{A_y,\bar{y}}\sigma_{A_y} /(P + R)$, yielding

$$R_P = \frac{1}{2}\left(-P + \sqrt{P^2 + 4i\rho_{A_y,\bar{y}}\sigma_{A_y}}\right) \tag{39}$$

For a prevalence smaller than ~0.5, predictions from Equation 39 were very close to the observed response in prevalence (Figure 8B).

In conclusion, results in this section show that response to selection in the prevalence of endemic infectious diseases is a factor $1/P$ greater than suggested by the ordinary breeding values for individual binary disease status (Figure 8A). Thus breeders can predict response to selection by upscaling the selection differential in the usual estimated breeding values for binary disease status by a factor $1/P$ (Figure 8B).

8. Direct and indirect genetic variance in prevalence

In this section, we partition the total genetic variance in prevalence into direct and indirect genetic components. This partitioning is relevant, because IGE respond fundamentally different to selection than DGE (Griffing 1967; Griffing 1977; Moore *et al*., 1997; Muir 2005; Bijma 2010, 2011; see Discussion). We can partition the total breeding value for prevalence into a direct and an indirect component,

$$A_P = A_{P_D} + A_{P_I} \tag{40}$$

Analogously, we can partition the full additive genetic variance in prevalence into components due to direct genetic variance, indirect genetic variance and a covariance,

$$\sigma^2_{A_P} = \sigma^2_{A_{P_D}} + 2\sigma_{A_{P_D}A_{P_I}} + \sigma^2_{A_{P_I}} \tag{41}$$

In the absence of genetic variation in infectivity, the breeding value for own disease status is a fraction P of the breeding value for prevalence (Equation 33). Hence, a fraction P of the genetic effects of susceptibility and recovery on prevalence affects the disease status of the individual itself and is thus due to direct effects, while the remaining fraction $(1 - P)$ is due to indirect effects. For infectivity, the entire genetic effect is indirect, because an individual's infectivity does not affect its own disease status. Assuming limited heterogeneity, it follows from Equations 31b that

$$A_{P_D} = \frac{1}{R_0} P \left(A_{l\gamma} - A_{l\alpha}\right) \tag{42a}$$

$$A_{P_I} = \frac{1}{R_0} \left((1-P)A_{l\gamma} + A_{l\varphi} - (1-P)A_{l\alpha}\right) \tag{42b}$$

Note that Equation 42a is identical to the breeding value for individual disease status (A_y, Equation 34, using the substitution $1/R_0 = 1 - P$), but the current expression emphasizes the partitioning of A_P into direct and indirect effects. The direct and indirect genetic (co)variances are given by

$$\sigma^2_{A_{P_D}} = \frac{P^2}{R_0^2}\left(\sigma^2_{A_{l\gamma}} - 2\sigma_{A_{l\gamma}A_{l\alpha}} + \sigma^2_{A_{l\alpha}}\right) \tag{43a}$$

$$\sigma_{A_{P_D}A_{P_I}} = \frac{P}{R_0^2}\left\{(1-P)\left(\sigma^2_{A_{l\gamma}} - 2\sigma_{A_{l\gamma}A_{l\alpha}} + \sigma^2_{A_{l\alpha}}\right) + \sigma_{A_{l\gamma}A_{l\varphi}} - \sigma_{A_{l\gamma}A_{l\alpha}}\right\} \tag{43b}$$

$$\sigma^2_{A_{P_I}} = \frac{1}{R_0^2}\left\{(1-P)^2\left(\sigma^2_{A_{l\gamma}} - 2\sigma_{A_{l\gamma}A_{l\alpha}} + \sigma^2_{A_{l\alpha}}\right) + 2(1-P)\left(\sigma_{A_{l\gamma}A_{l\varphi}} - \sigma_{A_{l\varphi}A_{l\alpha}}\right) + \sigma^2_{A_{l\varphi}}\right\} \tag{43c}$$

Figure 9 shows the total additive genetic variance in the endemic prevalence and the fractions due to DGE, IGE and their covariance, for a scenario with equal genetic variances in susceptibility, infectivity and recovery and covariances equal to zero. For an endemic prevalence smaller than 0.5, IGE contribute the majority of the genetic variance. For example, for an endemic prevalence of 0.3, the full additive genetic variance consists of 6% direct genetic variance, 66% indirect genetic variance and 28%

direct-indirect genetic covariance. These results imply that IGE dominate the heritable variation and response to selection in the endemic prevalence of infectious diseases.

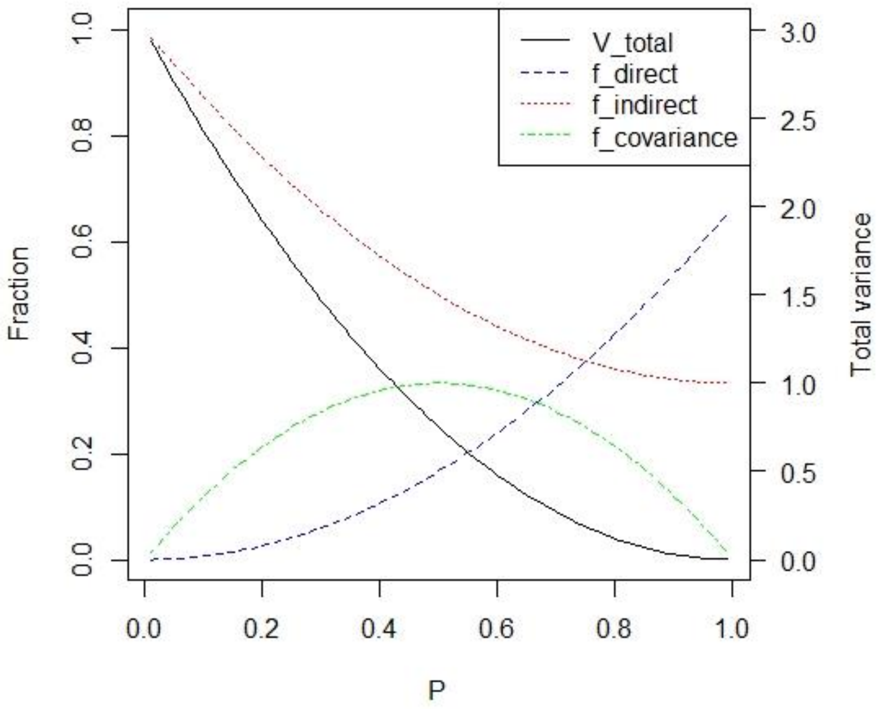


Figure 9 – Total additive genetic variance in endemic prevalence (V_{total}; secondary y-axis) and the relative contribution of DGE, IGE and their covariance (f_{direct}, f_{indirect}, and $f_{\text{covariance}}$; primary y-axis). For $\sigma^2_{A_\gamma}$ = $\sigma^2_{A_\varphi}$ = $\sigma^2_{A_\alpha}$ = 1 and covariances equal to zero. Results are obtained from Equations 32b, 3, and 43a-c.

Discussion

We presented quantitative genetic theory for endemic infectious diseases, with a focus on the genetic factors that determine the endemic prevalence. We defined an additive model for the logarithm of individual susceptibility, infectivity and rate of recovery, which results in normally distributed breeding values for the logarithm of R_0. Next we investigated the impact of genetic heterogeneity on the population level. Results showed that genetic heterogeneity has limited impact on R_0 and on the endemic prevalence. Subsequently, we considered genetic effects of individuals on their own disease status and on the endemic prevalence. Building on the breeding value for the logarithm of R_0, we showed that genotypic values and genetic parameters for the prevalence follow from the known properties of the log-normal distribution. In the absence of genetic variation in infectivity, genetic effects on the endemic prevalence are a factor 1/prevalence greater than the ordinary breeding values for individual binary disease status. Hence, even though prevalence is the simple average of individual binary disease status, breeding values for prevalence show much more variation than those for individual disease status. These results imply that the genetic variance that determines the potential response of endemic prevalence to selection is largely due to IGE, and thus hidden to classical genetic analysis and selection. For susceptibility and recovery, a fraction 1-*P* of the full genetic effect is due to IGE, whereas the effect of infectivity is entirely due to IGE. Hence, the genetic variance that determines the potential response of prevalence to selection must be much greater than currently believed, particularly at low levels of the prevalence (Figure 7). We evaluated this implication using stochastic simulation of endemics following standard methods in epidemiology, where parents of the next generation were selected based on their own disease status (mass selection). The results of these simulations showed that response to selection in the observed prevalence and in the breeding value for prevalence increased strongly when prevalence decreased, which supports our theoretical findings.

Model assumptions: Following Anacleto *et al.* (2015, 2019), Biemans *et al*. (2019) and Pooley *et al.* (2020), we assumed a linear additive model with normally distributed effects for the logarithm of susceptibility, infectivity and recovery, leading to a normal distribution of the additive genetic values for the logarithm of R_0. For complex traits, it is common to assume normally distributed genetic affects, based on the central limit theorem (Fisher 1918). Because R_0 is non-negative, we specified a normal distribution for its logarithm, to translate the $[0, \infty)$ domain of R_0 to the $(-\infty, +\infty)$ domain of the normal distribution. The same approach has been used to model variation in the genetic variance, which is also restricted to non-negative values (SanCristobal-Gaudy *et al*. 1998; Hill and Mulder 2010). The log-normal distribution of genotypic values for R_0 results in a decrease of the genetic standard deviation in R_0 with decreasing R_0 (Figure 2), which seems reasonable given the presence of a lower bound for R_0. Moreover, the log-normal distribution for R_0 is convenient, because it results in relatively simple expressions for the breeding value and the genetic variance for prevalence.

The assumption of a normal distribution for the logarithm of genotypic values for R_0 also agrees with the standard implementation of generalized linear (mixed) models (GLMM; Nelder and Wedderburn, 1972). R_0 refers to an expected number of infected individuals; In other words, R_0 is the expected value of count data. In GLMM, the default link function for count data is the log-link (McCullagh, 2019). Hence, our linear model for the logarithm of R_0 also agrees with common statistical practise.

Note that the strong increase of the genetic variance in prevalence with decreasing R_0 (Figure 5A) is not due to the assumption of lognormality of R_0. On the contrary, the log-normal distribution results in a decrease of the genetic standard deviation in R_0 with decreasing R_0 (Figure 2). The strong increase in the genetic variance in prevalence results from the relationship between R_0 and the endemic prevalence (Figure 1; Equation 3), which becomes steeper when R_0 is closer to one. This relationship is very well established in epidemiology (Weiss and Dishon, 1971; refs).

Positive feedback: The increasing difference between the breeding value for prevalence and the breeding value for individual disease status at lower prevalence (Equation 33) is a result of the increasing slope of the relationship between R_0 and the endemic prevalence (Equation 3, Figure 1). Equation 3 follows directly from a simple equilibrium condition (see text above Equation 3). However, the focus on the equilibrium partly obscures the underlying mechanism. Figures 10A and B illustrate that the difference between A_P and A_y originates from positive feedback effects in the transmission dynamics. (Figure 10 shows results for selection against susceptibility, selection for faster recovery would yield identical results). With lower susceptibility fewer individuals become infected, which subsequently translates into a reduced transmission rate, followed by a further reduction in the number of infected individuals, *etc*. This creates a positive feedback loop over cycles of the transmission-recovery loop

(Figure 10A). The initial change in prevalence before feedback effects manifest is equal to the selection differential in breeding value for individual disease status ($\Delta\bar{A}_y$; horizontal lines in Figure 9). This change represents the direct response due to reduced susceptibility, and does not include any change in exposure of susceptible individuals to infected herd mates. Next, prevalence decreases further because the initial decrease in prevalence reduces the exposure of susceptible individuals to infected herd mates. This additional decrease represents the indirect response to selection via the "social" environment. Without genetic variation in infectivity, the direct response makes up a fraction P of the total response in prevalence, and the indirect response a fraction $1 - P$.

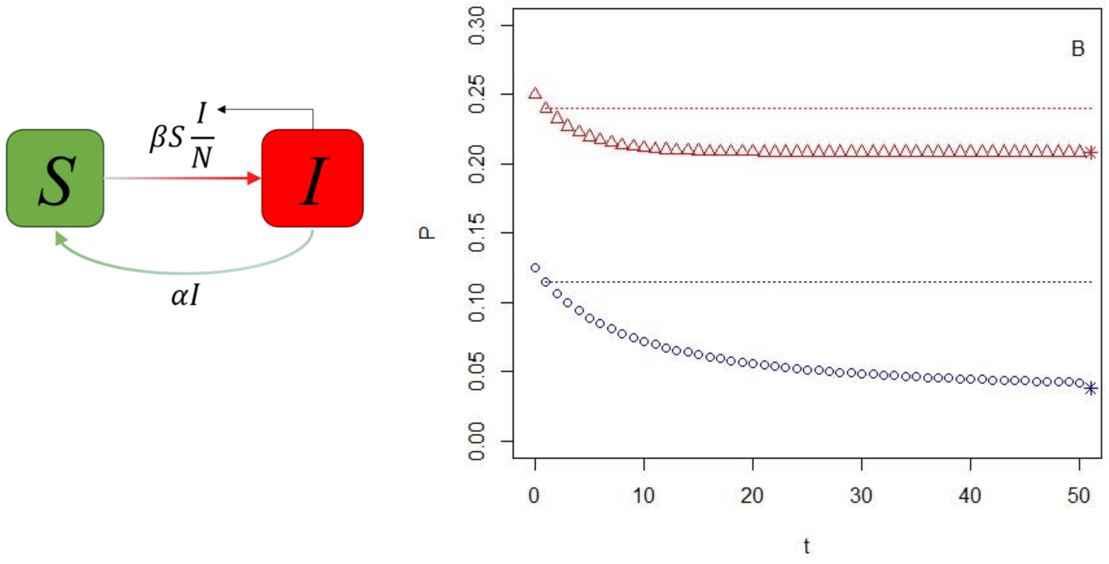


Figure 10 – Positive feedback after selection for lower susceptibility. Panel A: Diagram of the SIS compartmental model illustrating the feedback, with the number of susceptible (*S*) and infectious (*I*) individuals and the transmission and recovery rates (ignoring heterogeneity for simplicity). A reduction in the transmission rate parameter *β* reduces *I*, which in turn reduces the transmission rate, leading to a further reduction in *I*, *etc*. Panel B: Convergence of the prevalence to the new equilibrium after selection. For two populations; one starting at *P* = 0.25 (red triangles; *c* = 1.333), the other at *P* = 0.125 (blue circles; *c* = 1.143). The *x*-axis represents cycles of the transmission loop. The horizontal dotted lines show the prevalence predicted by the breeding value for binary disease status (the direct effect). The Asterix shows the equilibrium prevalence after convergence, which occurs a little later than *t* = 50 for the lower line. The genetic selection differential for binary disease status equals $\Delta\bar{A}_y = -0.01$ for both populations. The initial response to selection (the *y*-axis difference between *t* = 0 and *t* = 1) is equal to the $\Delta\bar{A}_y$ of -0.01 for both scenarios. Total response is -0.04 for the scenario with *P* = 0.25, and -0.08 for the scenario with *P* = 0.125, corresponding to -0.01/0.25 and -0.01/0.125. Results in panel B follow from iterating on Equation 20a, using a single value for $\mathcal{R}_{0,i}$, with $\alpha_i = \bar{\varphi}_{\text{inf}} = 1$ and choosing γ so that the selection differential $\Delta\bar{A}_y = -0.01$ (using $\gamma = \frac{G_y}{P(1-G_y)}/c$ from Equation 20a). In each iteration, the *P* in the righthand side of Equation 20a is replaced by the P_i calculated from Equation 20a in the previous iteration. This iteration converges to the prevalence given by Equation 3 (assuming negligible heterogeneity).

Herd immunity: In Figure 8A, the infection ultimately goes extinct due to mass selection for individual disease status. This happens due to a phenomenon known as herd immunity (Fine, 1993). In the final generation, the infection disappears because R_0 falls below a value of one; not because all the individuals have become fully resistant to infection. This result is similar to eradication of an infection by means of vaccination, which does not require full immunity of individuals and can also be achieved when only part of a population is vaccinated (Anderson and May, 1985). As can be seen in Figure 10 and in simulation results of Hulst *et al*. (2021), herd immunity develops over cycles of the transmission-recovery loop. Thus the full benefits of genetic selection or vaccination do not manifest immediately, as it takes some time for a population to converge to the new endemic prevalence.

The relevance of herd immunity can be illustrated using the data underlying Figure 8A. In the initial generation of the population starting at a prevalence of 0.5, the mean breeding value for log-susceptibility is equal to zero, and the contact rate is equal to two (c = 2, $\bar{A}_{l\gamma} = 0$, so that $R_0 \approx ce^{\bar{A}_{l\gamma}}$ = 2). In the final generation, the mean breeding value for log-susceptibility has dropped to -0.73, so that $R_0 \approx 2e^{-0.73}$ = 0.96. Hence, $R_0 < 1$, explaining extinction. However, if the average individual of the final generation would have been exposed to the infection pressure of the first generation, then the expected prevalence for this individual would have been 0.32 (From Equation 20a, with $R_{0,i} = 0.96$ and P = 0.5). Hence, the individual would have been infected 32% of the time. Nevertheless, in a population consisting entirely of this type of individual, as is the case in the final generation, the infection will no longer be present in the long term. This example illustrates the relevance of indirect effects for herd immunity and response to selection of infectious diseases.

Utilization of hidden genetic variation for genetic improvement: In this work, we have shown that a fraction $1 - P$ of the full individual genetic effect on the endemic prevalence represents an IGE, because only a fraction P of the full effect surfaces in the disease status of the individual itself (assuming no genetic variation in infectivity; Equation 33 and Appendix 5). In other words, a fraction $1 - P$ of the individual genetic effects of susceptibility and recovery on the prevalence are hidden to direct selection and classical genetic analysis. Nevertheless, results in Figure 8 show that prevalence responds rapidly to selection, particularly when prevalence is small. Hence, prevalence responds faster to selection when a greater proportion of its heritable variation is hidden, and when heritability is low (Figure 7A), which seems a paradox.

However, the IGEs due to susceptibility and recovery are a special kind, because they are fully correlated to the corresponding DGE. For each of the two traits, there is only a single genetic effect ($A_{l\gamma}$ and $A_{l\alpha}$, respectively), which has both a direct effect and an indirect effect on the prevalence. Hence, when selection changes the mean DGE, the mean IGE changes correspondingly. This can be seen from Equation 38, where the term σ_{A_y}/P represents the full additive genetic standard deviation in prevalence (as is clear from Equation 33), while the accuracy ($\rho_{A_y,\bar{y}}$) refers to selection for the direct effect only. Hence, without genetic variation in infectivity, the total response to selection based on individual disease status can be interpreted as the sum of a direct response in DGE and a correlated response in IGE,

$$R_{P,direct} = \iota\rho_{A_y,\bar{y}}\sigma_{A_y}$$

$$R_{P,correlated} = \iota\rho_{A_y,\bar{y}}\sigma_{A_y}\frac{1-P}{P}$$

and the sum of $R_{P,direct}$ and $R_{P,correlated}$ is equal to Equation 38. The $R_{P,direct}$ is the response to selection expected based on ordinary genetic analysis of individual disease status. The $R_{P,correlated}$ represents the additional response due to IGE. The direct response occurs immediately in the first cycle of the transmission loop (Figure 10B), while the indirect response manifests gradually over several cycles of the transmission loop, particularly when prevalence is small (see also result in Hulst *et al.* 2021).

The response due to the IGE of susceptibility and recovery arises naturally when selecting for lower individual disease status (*i.e.*, for the direct effect); it does not require any specific measures of the breeder. Thus, on the one hand, our results imply that response to genetic selection against infectious diseases should be considerably greater than currently believed, even when no changes are made to the selection strategy.

On the other hand, however, classical selection for direct effects is not the optimal way to reduce prevalence, for the following two reasons. First, classical selection does not target genetic effects on infectivity, because an individual's infectivity does not affect its own disease status (Lipschutz-Powell et al., 2012). Hence, infectivity changes merely due to a potential genetic correlation with susceptibility and/or recovery. When this correlation is unfavourable, infectivity will increase and response in prevalence will be smaller than expected based on the genetic selection differentials for susceptibility and recovery. (And thus smaller than the result of Equation 38). In theory, this could even lead to a negative net response (Griffing 1967). This is similar to the case with social behaviour-related IGEs on survival in laying hens and Japanese quail, where selection for individual survival may increase mortality (Craig and Muir, 1996; Muir 2005). This scenario seems unlikely for infectious diseases, but at present we lack knowledge of the multivariate genetic parameters of susceptibility, infectivity and recovery to make well-founded statements.

Second, even in the absence of genetic variation in infectivity, individual selection for susceptibility and recovery is non-optimal because the accuracy of selection is limited due to limited heritability, particularly at low prevalence (Figure 7A). The response to selection in traits affected by IGE can be increased by using kin selection and/or group selection (Griffing 1976; Muir 1996; Bijma 2011), or by including IGE in the genetic analysis (Muir 2005, Bijma *et al.* 2007b; Muir *et al.* 2013; Biemans *et al*. 2019, Anacleto *et al.* 2015, Pooley *et al.* 2020). Kin selection occurs when transmission takes place between related individuals, for example within groups of relatives (Anche *et al.* 2014). Group selection refers to the selection of parents for the next generation based on the prevalence in the group in which transmission takes place, rather than on individual disease status (Griffing 1976). Both theoretical and empirical work shows that kin and group selection lead to utilization of the full genetic variation, including both DGE and IGE (Griffing 1976, Muir 1996, 2005; Bijma and Wade, 2008; Bijma 2010, 2011). For infectious diseases, the work of Anche *et al.* (2014) illustrates the effect of kin selection, where favourable alleles for susceptibility increase much faster in frequency when disease transmission is between related individuals.

The mechanism of both kin and group selection relies on feed-back of an individual's IGE on its own value for the selection criterion. With kin selection, individuals with poor IGEs are on average exposed to the poor IGEs of their genetically related social partners, which reduces their trait value and thus their change of being selected as parent of the next generation. Irrespective of genetic relatedness, individuals with poor IGEs depress the performance of their group, which reduces their change of being selected as parent of the next generation in group selection scenarios.

Other compartmental models: In this work, we focused on endemic infectious diseases following a SIS-model, where individuals can either be susceptible (S, *i.e.*, non-infected) or infected (I). Hence, we assumed the infection does not confer any long-lasting immunity, and we ignored the potential existence of infection classes ("compartments") other than S and I, for example latently infected individuals that are not yet infectious. Moreover, we ignored the influx of new individuals into the population due to births, and the removal of individuals due to deaths, because the dynamics of transmission are often much faster than those of birth and death.

A key condition for validity of our results is that the pathogen can replicate only in the host individual, meaning that a reduction in, *e.g.*, susceptibility fully translates into reduced exposure of the host population to the pathogen (The mere survival of the pathogen in the environment does not violate our assumptions; see Hulst *et al*., 2021 for a discussion). This condition is not limited to the SIS-model, but is, for example, also met in the SEIS compartmental model. In the SEIS model, there is an incubation period after infection, where individuals have been exposed (E) but are not yet infectious. However, when birth and death can be ignored, R_0 and the equilibrium prevalence of the SEIS model are identical to those of the SIS model (Equations 3 and 4). Hence, our results also apply to cases with an incubation period after infection.

Infections that confer long-lasting immunity typically show epidemic, rather than endemic, behavior. Measles in the human population before the introduction of vaccination are a well-known example. For such infections, the same mechanisms as discussed above will play a role, but their quantitative effect is different. For epidemic infections, a fraction of the initially susceptible individuals typically escapes from becoming infected, and the scale of the epidemic is measured by the fraction of the population that has been infected when the epidemic is over (the so-called final size; Kermack & McKendrick, 1927). The Susceptible-Infected-Recovered (SIR) model is the simplest and best known compartmental model for epidemic infections. R_0 is the same for the SIR and the SIS model (Equation 4), and the final size follows from R_0 (Kermack & McKendrick, 1927). However, the relationship between the final size and R_0 in the SIR model, differs from the relationship between the endemic prevalence and R_0 in the SIS-model. At the same R_0, the final size is greater than the endemic prevalence because of an overshoot (*e.g.*, Diekmann *et al*. 2013; see figure in Appendix 7). Below (above) $R_0 \approx 2.15$, the slope of final size plotted as a function of R_0 is steeper (flatter) than the slope of endemic prevalence as a function of R_0 (Figure 1). Therefore, when $R_0 \lesssim 2.15$, the increase of the additive genetic variance for the final size of an epidemic infection when R_0 decreases is considerably stronger than for the prevalence of an endemic infection (the latter is in Figure 5A). Hence, we expect that the final size of epidemic infectious diseases with $R_0 \lesssim 2.15$ will respond very rapidly to selection, even more so than the equilibrium prevalence of endemic infections (assuming presence of genetic variance of course).

Exposure to infectious pathogens is a major driver of the evolution of host populations by natural selection, both in animals and plants (reviewed in Karlsson *et al*. 2014 and Ebert and Fields

2020). In the human species, for example, a study of genetic variation in 50 worldwide populations reveals that selection on infectious pathogens is the primary driver of local adaptation and the strongest selective force that shapes the human genome (Barreiro and Quintana-Murci 2010; Fumagalli *et al*. 2011). The key role of infectious pathogens in natural selection, together with the large contribution of IGE to the genetic variation in prevalence in the host population, indicates that IGE must have been an important fitness component. This, in turn, suggests that associating with kin may have evolved as an adaptive behaviour. In other words, natural selection might lead to social structures where individuals associate preferably with kin, because such behaviour has indirect fitness benefits. This is because interactions among kin lead to utilisation of the full heritable variation in fitness, including both DGE and IGE (Bijma, 2010), and thus accelerate response of fitness to selection. At low to moderate levels of the endemic prevalence, the genetic variation in prevalence might be sufficiently large for such behaviour to evolve even in the absence of direct fitness benefits, such as preferential behaviour towards kin. While this is a complex issue requiring careful quantitative modelling, including migration and emergence of selfish mutants, the size of IGE together with the key role of pathogens in natural selection strongly suggest the importance of kin selection in the history of life.

In agriculture, the implementation of kin selection may be feasible when animals can be kept in kin groups or plants can be grown in plots of a single genotype. In many cases, however, this will not be feasible, and other methods are required to optimally capture the IGE underlying the prevalence of infectious diseases. Current developments in sensing technology and artificial intelligence enable the development of tools for large scale automated collection of longitudinal data on individual disease status, and also on the contact structure between individuals (relevant mainly in animals). These advances, together with recently developed statistical methods for the estimation of the direct and indirect genetic effects underlying disease transmission (Pooley *et al*. 2020), could represent a much-needed breakthrough in artificial selection against infectious diseases in agriculture. Our results on genetic variation and response to selection suggest that such selection is way more promising than currently believed.

References

Anacleto, O., Garcia-Cortés, L. A., Lipschutz-Powell, D., Woolliams, J. A., & Doeschl-Wilson, A. B. (2015). A novel statistical model to estimate host genetic effects affecting disease transmission. *Genetics*, *201*(3), 871-884.

Anacleto, O., Cabaleiro, S., Villanueva, B., Saura, M., Houston, R. D., Woolliams, J. A., & Doeschl-Wilson, A. B. (2019). Genetic differences in host infectivity affect disease spread and survival in epidemics. Scientific reports, 9(1), 1-12.

Anche, M. T., De Jong, M. C. M., & Bijma, P. (2014). On the definition and utilization of heritable variation among hosts in reproduction ratio R 0 for infectious diseases. *Heredity*, *113*(4), 364-374.

Anderson, R., May, R. Population biology of infectious diseases: Part I. *Nature* **280,** 361–367 (1979). https://doi.org/10.1038/280361a0

Anderson, R. M., & May, R. M. (1985). Vaccination and herd immunity to infectious diseases. Nature, 318(6044), 323-329.

Aznar, I., Frankena, K., More, S. J., O'Keeffe, J., McGrath, G., & De Jong, M. C. M. (2018). Quantification of Mycobacterium bovis transmission in a badger vaccine field trial. *Preventive veterinary medicine*, *149*, 29-37.

Barreiro LB, Quintana-Murci L (2010) From evolutionary genetics to human immunology: How selection shapes host defence genes. Nat Rev Genet 11(1): 17–30.

Biemans, F., de Jong, M. C., & Bijma, P. (2017). A model to estimate effects of SNPs on host susceptibility and infectivity for an endemic infectious disease. *Genetics Selection Evolution*, *49*(1), 1-13.

Biemans, F., De Jong, M. C., & Bijma, P. (2019). Genetic parameters and genomic breeding values for digital dermatitis in Holstein Friesian dairy cattle: host susceptibility, infectivity and the basic reproduction ratio. *Genetics Selection Evolution*, *51*(1), 1-13.

Bijma, P. (2010). Fisher's fundamental theorem of inclusive fitness and the change in fitness due to natural selection when conspecifics interact. Journal of evolutionary biology, 23(1), 194-206.

Bijma, P. (2011). A general definition of the heritable variation that determines the potential of a population to respond to selection. *Genetics*, *189*(4), 1347-1359.

Bijma, P., & Wade, M. J. (2008). The joint effects of kin, multilevel selection and indirect genetic effects on response to genetic selection. Journal of evolutionary biology, 21(5), 1175-1188.

Bishop SC, Woolliams JA (2010) On the Genetic Interpretation of Disease Data. PLoS ONE 5(1): e8940. https://doi.org/10.1371/journal.pone.0008940.

Bishop, S., Doeschl-Wilson, A. B., & Woolliams, J. A. (2012). Uses and implications of field disease data for livestock genomic and genetics studies. Frontiers in genetics, 3, 114

Bolker, B. M., & Grenfell, B. T. (1996). Impact of vaccination on the spatial correlation and persistence of measles dynamics. *Proceedings of the National Academy of Sciences*, 93(22), 12648-12653.

Craig, J. V., & Muir, W. M. (1996). Group selection for adaptation to multiple-hen cages: beak-related mortality, feathering, and body weight responses. *Poultry Science*, *75*(3), 294-302.

De Villemereuil, P., Schielzeth, H., Nakagawa, S., & Morrissey, M. (2016). General methods for evolutionary quantitative genetic inference from generalized mixed models. *Genetics*, *204*(3), 1281-1294.

O Diekmann; J.A.P. Heesterbeek; J.A.J. Metz (1990). "On the definition and the computation of the basic reproduction ratio in models for infectious diseases in heterogeneous populations". Journal of Mathematical Biology. 28 (4): 356–382. doi:10.1007/BF00178324.

Diekmann, O., Heesterbeek, H., & Britton, T. (2012). Mathematical tools for understanding infectious disease dynamics (Vol. 7). Princeton University Press.

Ebert, D., & Fields, P. D. (2020). Host–parasite co-evolution and its genomic signature. Nature Reviews Genetics, 21(12), 754-768.

EFSA Panel on Animal Health and Welfare (AHAW). (2012). Scientific Opinion on Review of the European Union Summary Report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2009 and 2010 specifically for the data related to bovine tuberculosis, Echinococcus, Q fever, brucellosis and non-food borne diseases. EFSA Journal, 10(6), 2765.

Fine, P. E. (1993). Herd immunity: history, theory, practice. Epidemiologic reviews, 15(2), 265-302.

Frank, S. A. (2002). Immunology and evolution of infectious disease.

Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admettla, A., Pattini, L., & Nielsen, R. (2011). Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. PLoS Genet, 7(11), e1002355.

Gianola, D. (1982). Theory and analysis of threshold characters. *Journal of animal Science*, *54*(5), 1079-1096.

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(1), 127-140.

Griffing, B. (1976). Selection in reference to biological groups. V. Analysis of full-sib groups. Genetics, 82(4), 703-722.

Griffing, B. (1977) Selection for populations of interacting genotypes, in Proceedings of the International Congress on Quantitative Genetics, August 16-21, 1976 (Pollack, E., Kempthorne, O. and Bailey, T.B., eds), pp. 413–434, Iowa State University Press

Hethcote, H. W., 1989 Three basic epidemiological models, pp. 119-144 in *Applied mathematical ecology*. Springer.

Heringstad, B., G. Klemetsdal and T. Steine, 2007 Selection responses for disease resistance in two selection experiments with Norwegian red cows. Journal of dairy science 90**:** 2419-2426.

Hill, W. G., & Mulder, H. A. (2010). Genetic analysis of environmental variation. Genetics Research, 92(5-6), 381-395.

Karlsson, E. K., Kwiatkowski, D. P., & Sabeti, P. C. (2014). Natural selection and infectious disease in human populations. Nature Reviews Genetics, 15(6), 379-393. : Host genetics influences susceptibility to infectious disease.

Keeling, M. J., & Rohani, P. (2011). *Modeling infectious diseases in humans and animals*. Princeton university press.

Kermack, W. O. and McKendrick, A. G. 1927. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. Ser. A*, 115: 700–721.

Knap, P. W., & Doeschl-Wilson, A. (2020). Why breed disease-resilient livestock, and how?. *Genetics Selection Evolution*, *52*(1), 1-18.

Lipschutz-Powell, D., Woolliams, J. A., Bijma, P., & Doeschl-Wilson, A. B. (2012). Indirect genetic effects and the spread of infectious disease: are we capturing the full heritable variation underlying disease prevalence?. *PloS one*, *7*(6), e39551.

Lipschutz-Powell, D., Woolliams, J. A., & Doeschl-Wilson, A. B. (2014). A unifying theory for genetic epidemiological analysis of binary disease data. Genetics Selection Evolution, 46(1), 1-12.

Martin, P., H. Barkema, L. Brito, S. Narayana and F. Miglior, 2018 Symposium review: Novel strategies to genetically improve mastitis resistance in dairy cattle. Journal of dairy science 101**:** 2724-2736.

McCullagh, P. and Nelder, J. (1989) Generalized Linear Models, 2nd edn. London: Chapman and Hall.

Moore, A. J., Brodie III, E. D., & Wolf, J. B. (1997). Interacting phenotypes and the evolutionary process: I. Direct and indirect genetic effects of social interactions. Evolution, 51(5), 1352-1362.

Muir, W. M. (1996). Group selection for adaptation to multiple-hen cages: selection program and direct responses. *Poultry Science*, *75*(4), 447-458.

Muir, W. M. (2005). Incorporation of competitive effects in forest tree or animal breeding programs. *Genetics*, *170*(3), 1247-1259.

Muir, W. M., Bijma, P., & Schinckel, A. (2013). Multilevel selection with kin and non-kin groups, experimental results with Japanese quail (Coturnix japonica). *Evolution*, *67*(6), 1598-1606.

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. Journal of the Royal Statistical Society: Series A (General), 135(3), 370-384.

Nieuwhof, G. J., J. Conington and S. C. Bishop, 2009 A genetic epidemiological model to describe resistance to an endemic bacterial disease in livestock: application to footrot in sheep. Genetics Selection Evolution 41**:** 19.

Pooley, C. M., Marion, G., Bishop, S. C., Bailey, R. I., & Doeschl-Wilson, A. B. (2020). Estimating individuals' genetic and non-genetic effects underlying infectious disease transmission from temporal epidemic data. *PLOS Computational Biology*, *16*(12), e1008447.

Rushton J. The economics of animal health and production. Wallingford: CABI; 2009.

Russell, G. E. (2013). Plant breeding for pest and disease resistance: studies in the agricultural and food sciences.

SanCristobal-Gaudy, M., Elsen, J. M., Bodin, L., & Chevalet, C. (1998). Prediction of the response to a selection for canalisation of a continuous trait in animal breeding. Genetics Selection Evolution, 30(5), 423-451.

Schrag, S. J., & Wiener, P. (1995). Emerging infectious disease: what are the relative roles of ecology and evolution?. *Trends in ecology & evolution*, *10*(8), 319-324.

A J Springbett, K MacKenzie, J A Woolliams, S C Bishop, The Contribution of Genetic Diversity to the Spread of Infectious Diseases in Livestock Populations, *Genetics*, Volume 165, Issue 3, 1 November 2003, Pages 1465–1474, https://doi.org/10.1093/genetics/165.3.1465

Thanner, S., Drissner, D., & Walsh, F. (2016). Antimicrobial resistance in agriculture. *MBio*, *7*(2).

Tsairidou, S., Anacleto, O., Woolliams, J. A., & Doeschl-Wilson, A. (2019). Enhancing genetic disease control by selecting for lower host infectivity and susceptibility. *Heredity*, *122*(6), 742-758.

Walsh, B., & Lynch, M. (2018). Evolution and selection of quantitative traits. Oxford University Press.

Weiss, G. H., & Dishon, M. (1971). On the asymptotic behavior of the stochastic and deterministic models of an epidemic. *Mathematical Biosciences*, 11(3-4), 261-265.

Wolf, J. B., Brodie III, E. D., Cheverud, J. M., Moore, A. J., & Wade, M. J. (1998). Evolutionary consequences of indirect genetic effects. *Trends in ecology & evolution*, *13*(2), 64-69.

List of appendices

Appendix 1

R_0 with heterogeneity and log-normally distributed susceptibility, infectivity en recovery

We assume that the transmission rate from infected individual *j* to susceptible individual *i* is proportional to the product of the infectivity of *j* and the susceptibility of *i* (Equation 5),

$$\beta_{ij} = c\gamma_i\varphi_j.$$

So there is no interaction between i and j. (This property is known as separable mixing in the epidemiological literature; Diekmann *et al*. 1990; 2013). Moreover, we assume that susceptibility, infectivity and recovery follow a log-normal distribution (Equations 6 and 7). We also assume that the population is not very small, so that in the early phase of an endemic where only few individuals are infected, the composition of the remaining susceptible individuals is not affected.

Because R_0 refers to the "total number of individuals that become infected by a typical infected individual over its entire infectious lifetime", we define an individual lifetime infectivity, which is the product of an individual's infectivity per unit of time and the average duration of its infectious lifetime,

$$\phi_i = \varphi_i/\alpha_i,$$

which follows a log-normal distribution with parameters following from those of φ and α. Hence, we have condensed our three genetic effects into two.

We can find R_0 from

$$R_0 = c\,\bar{\gamma}\,\bar{\phi}_{typ}$$

where $\bar{\phi}_{typ}$ is the lifetime infectivity of the typical infected individual, and $\bar{\gamma}$ is the simple average susceptibility in the population,

$$\bar{\gamma} = \int_0^\infty \gamma\, g(\gamma)\, d\gamma.$$

where $g(\gamma)$ is the pdf of γ. We can use the simple average of susceptibility in this expression because we assume the population is large.

With separable mixing, the typical infected individual is created immediately in the first generation of disease transmission. This is the case because there is no interaction between γ and φ, so that the properties of the typical infected individual are determined entirely by susceptibility. Hence, the pdf of γ for the typical infected individual follows from weighing $g(\gamma)$ by γ,

$$g_{typ}(\gamma) = \frac{1}{\bar{\gamma}}\gamma\, g(\gamma)$$

Since the properties of the typical infected individual depend on susceptibility only, we can find $\bar{\phi}_{typ}$ by averaging ϕ over its distribution conditional on γ, and subsequently averaging over the distribution of γ,

$$\bar{\phi}_{typ} = \int_0^\infty \left(\int_0^\infty \phi\, f(\phi|\gamma)\, d\phi \right) g_{typ}(\gamma)\, d\gamma$$

Hence, we now have the elements of R_0, but still need to solve the integral expression.

Because conditional Normal distributions are also Normal and the logarithm is a bijective function, $\phi|\gamma$ follows a log-normal distribution with parameters being the conditional mean and variance of the Normal distribution,

$$\phi|\gamma \sim Lnorm\left(\mu = b_{\phi,\gamma} A_{l\gamma}\ ;\ \sigma^2 = \left(1-\rho_{\gamma,\phi}^2\right)\sigma_{A_{l\phi}}^2\right)$$

with $b_{\phi,\gamma} = cov(A_{l\gamma}, A_{l\phi})/var(A_{l\gamma})$ denoting the regression coefficient of $A_{l\phi}$ on $A_{l\gamma}$, and $\rho_{\gamma,\phi}^2 = cov^2(A_{l\gamma}, A_{l\phi})/[var(A_{l\gamma})var(A_{l\phi})]$ the squared correlation, where $A_{l\phi}$ denotes the breeding value for logarithm of lifetime infectivity.

Hence, the inner integral is the mean of a log-normal variate, which is of the form exp $(\mu + \frac{\sigma^2}{2})$,

$$\int_0^\infty \phi\, f(\phi|\gamma)\, d\phi = E[\phi|\gamma] = \exp\left(b_{\phi,\gamma} A_{l\gamma} + \tfrac{1}{2}\left(1-\rho_{\gamma,\phi}^2\right)\sigma_{A_{l\phi}}^2\right)$$

$$= e^{\frac{1}{2}\left(1-\rho_{\gamma,\phi}^2\right)\sigma_{A_{l\phi}}^2}\ e^{b_{\phi,\gamma} A_{l\gamma}}\ .$$

Since the first term of this expression is a constant,

$$\bar{\phi}_{typ} = e^{\frac{1}{2}\left(1-\rho_{\gamma,\phi}^2\right)\sigma_{A_{l\phi}}^2}\ \int_0^\infty e^{b_{\phi,\gamma} A_{l\gamma}}\, g_{typ}(\gamma)\, d\gamma.$$

Substituting $g_{typ}(\gamma) = \frac{1}{\bar{\gamma}}\gamma\, g(\gamma)$, and replacing $g(\gamma)$ by the corresponding log-normal pdf yields

$$\int_0^\infty e^{b_{\phi,\gamma} A_{l\gamma}}\ \ g_{typ}(\gamma)\, d\gamma \ = \ \frac{1}{\bar{\gamma}\sigma\sqrt{2\pi}} \int_0^\infty e^{-\frac{A^2}{2\sigma^2}+bA}\ d\gamma$$

where we simplified the notation for brevity, using $\sigma^2 = \sigma_{A_{l\gamma}}^2$, $b = b_{\phi,\gamma}$, and $A = A_{l\gamma}$.

Next, we change variable, using $d\gamma = e^A dA$, and adjust the bounds accordingly,

$$\frac{1}{\bar{\gamma}\sigma\sqrt{2\pi}} \int_{-\infty}^\infty e^{-\frac{A^2}{2\sigma^2}+bA}\ e^A dA \ = \ \frac{1}{\bar{\gamma}\sigma\sqrt{2\pi}} \int_{-\infty}^\infty e^{-\frac{A^2}{2\sigma^2}\ +(1+b)A}\ dA$$

Solving the integral term in Mathematica-online yields

$$\frac{1}{\bar{\gamma}\sigma\sqrt{2\pi}} \int_{-\infty}^\infty e^{-\frac{A^2}{2\sigma^2}\ +(1+b)A}\ dA = \frac{1}{\bar{\gamma}}\ e^{\frac{1}{2}\sigma_{A_{l\gamma}}^2 (1+b_{\phi,\gamma})^2}$$

$$\bar{\phi}_{typ} = \; e^{\frac{1}{2}(1-\rho^2_{\gamma,\phi})\sigma^2_{A_{l\phi}}} \; \frac{1}{\bar{\gamma}} \; e^{\frac{1}{2}\sigma^2_{A_{l\gamma}}(1+b_{\phi,\gamma})^2}$$

$$\bar{\phi}_{typ} = \frac{1}{\bar{\gamma}} \; e^{\frac{1}{2}\left[(1-\rho^2_{\gamma,\phi})\sigma^2_{A_{l\phi}}+\sigma^2_{A_{l\gamma}}(1+b_{\phi,\gamma})^2\right]}$$

$$R_0 = c \; \bar{\gamma} \; \bar{\phi}_{typ} = c \; e^{\frac{1}{2}\left[(1-\rho^2_{\gamma,\phi})\sigma^2_{A_{l\phi}}+\sigma^2_{A_{l\gamma}}(1+b_{\phi,\gamma})^2\right]}$$

Using $\bar{\gamma} = e^{\frac{1}{2}\sigma^2_{A_{l\gamma}}}$ and $\bar{\phi} = e^{\frac{1}{2}\sigma^2_{A_{l\phi}}}$ this simplifies to Equations 14 and 16 of the main text,

$$\bar{\phi}_{typ} = \bar{\phi} e^{\sigma_{A_{l\gamma}A_{l\phi}}}$$

$$R_0 = c \; \bar{\gamma} \; \bar{\phi} \; e^{\sigma_{A_{l\gamma}A_{l\phi}}}$$

Further simplification follows from expressing $\bar{\gamma}$, $\bar{\phi}$ and $e^{\sigma_{A_{l\gamma}A_{l\phi}}}$ in terms of variances and covariances of γ, φ and α.

$$\bar{\gamma} = e^{\frac{1}{2}\sigma^2_{A_{l\gamma}}}$$

$$\phi_i = \frac{\varphi_i}{\alpha_i} = e^{(A_{l\varphi,i}-A_{l\alpha,i})} = e^{A_{l\phi,i}}$$

where $A_{l\phi,i} = A_{l\varphi,i} - A_{l\alpha,i}$, which is the breeding value for the logarithm of lifetime infectivity, with

$$var(A_{l\phi,i}) = \sigma^2_{A_{l\varphi}} - 2\sigma_{A_{l\varphi}A_{l\alpha}} + \sigma^2_{A_{l\alpha}}$$

From the properties of the log-normal distribution,

$$\bar{\phi} = e^{\frac{1}{2}\sigma^2_{A_{l\phi}}} = e^{\frac{1}{2}(\sigma^2_{A_{l\varphi}}-2\sigma_{A_{l\varphi}A_{l\alpha}}+\sigma^2_{A_{l\alpha}})}$$

Furthermore,

$$e^{\sigma_{A_{l\gamma}A_{l\phi}}} = e^{(\sigma_{A_{l\gamma}A_{l\varphi}}-\sigma_{A_{l\gamma}A_{l\alpha}})}$$

Substitution of the expressions for $\bar{\gamma}$, $\bar{\phi}$ and $e^{\sigma_{A_{l\gamma}A_{l\phi}}}$ into $R_0 = c \; \bar{\gamma} \; \bar{\phi} \; e^{\sigma_{A_{l\gamma}A_{l\phi}}}$ yields

$$R_0 = c \;\; e^{\frac{1}{2}\sigma^2_{A_{l\gamma}}} \;\; e^{\frac{1}{2}(\sigma^2_{A_{l\varphi}}-2\sigma_{A_{l\varphi}A_{l\alpha}}+\sigma^2_{A_{l\alpha}})} \;\; e^{(\sigma_{A_{l\gamma}A_{l\varphi}}-\sigma_{A_{l\gamma}A_{l\alpha}})}$$

$$R_0 = c \; e^{\frac{1}{2}(\sigma^2_{A_{l\gamma}} + \sigma^2_{A_{l\varphi}} + \sigma^2_{A_{l\alpha}} + 2\sigma_{A_{l\gamma}A_{l\varphi}} - 2\sigma_{A_{l\varphi}A_{l\alpha}} - 2\sigma_{A_{l\gamma}A_{l\alpha}})}$$

$$R_0 = c \; e^{\frac{1}{2}\sigma^2_{A_{lR_0}}}$$

The right-hand side of this expression is identical to the mean genotypic value for R_0 (Equation 12).

Appendix 2

Numerical solution to find the endemic equilibrium prevalence with heterogeneity

To find the endemic prevalence, *P*, we partition the population into types, *i*, and numerically solve the expressions

$$P_i = \frac{\mathcal{R}_{0,i}P}{\mathcal{R}_{0,i}P+1}$$

and

$$\mathcal{R}_{0,i} = \frac{c\gamma_i\bar{\varphi}_{\text{inf}}}{\alpha_i}$$

for *P*. Here we develop this numerical solution for the case where susceptibility, infectivity and recovery follow a log-normal distribution, assuming separable mixing (see Appendix 1).

As can be seen from the expression for $\mathcal{R}_{0,i}$, the equilibrium prevalence for a type depends on both its susceptibility (γ_i) and its recovery rate (α_i). Individuals with above-average susceptibility are over-represented among the infecteds in the equilibrium, whereas individuals with above-average recovery are under-represented. Hence, as can be seen from the expression for $\mathcal{R}_{0,i}$, the partitioning into types should be based on γ_i/α_i. Therefore we define

$$\theta_i = \frac{\gamma_i}{\alpha_i} = e^{A_{\theta,i}}$$

$$A_{\theta_i} = A_{\gamma,i} - A_{\alpha,i}$$

$$\theta \sim Lnorm(\mu_{A_\theta} = 0, \sigma^2_{A_\theta} = \sigma^2_{A_\gamma} - 2\sigma_{A_\gamma A_\alpha} + \sigma^2_{A_\alpha})$$

To numerically solve the two equations given above, we also need $\bar{\varphi}_{\text{inf}}$. The $\bar{\varphi}_{\text{inf}}$ will depends on the θ_i of the infecteds when infectivity is correlated to susceptibility and/or recovery. Hence, we need the distribution of $\varphi|\theta$, which follows from

$$\sigma_{A_\theta A_\varphi} = \sigma_{A_\gamma A_\varphi} - \sigma_{A_\alpha A_\varphi}$$

$$b_{\varphi\theta} = \frac{\sigma_{A_\theta A_\varphi}}{\sigma^2_{A_\theta}}$$

$$\rho_{\varphi\theta} = \sigma_{A_\theta A_\varphi}/\sigma_{A_\theta}\sigma_{A_\varphi}$$

$$E\left(A_\varphi|A_\theta\right) = b_{\varphi\theta}A_\theta = \mu_{\varphi|\theta}$$

$$\text{var}\left(A_\varphi|A_\theta\right) = \left(1-\rho^2_{\varphi\theta}\right)\sigma^2_{A_\varphi} = \sigma^2_{\varphi|\theta}$$

so that

$$\varphi|\theta \sim Lnorm(\mu_{\varphi|\theta} = b_{\varphi\theta}A_\theta\ ,\ \sigma^2_{\varphi|\theta} = \left(1-\rho^2_{\varphi\theta}\right)\sigma^2_{A_\varphi})$$

From the log-normal distribution:

$$E(\varphi|\theta) = \mathrm{e}^{\mu_{\varphi|\theta} + \frac{1}{2}\sigma^2_{\varphi|\theta}}$$

Hence, we can partitioning θ into classes i, with

$$\mathcal{R}_{0,i} = c\ \theta_i\ \bar{\varphi}_{\text{inf}}$$

$$\bar{\varphi}_{\text{inf}} = \frac{1}{P}\sum_i f_i\ P_i\ E(\varphi_i|\theta_i)$$

$$P_i = \frac{\mathcal{R}_{0,i}P}{\mathcal{R}_{0,i}P + 1}$$

$$P = \sum_i f_i P_i$$

where f_i is the fraction of individuals of type i, $f_i = N_i/N$, and P_i is the prevalence in type i, $P_i = I_i/N_i$. The numerical solution follows from iterating on these four equations. An R-code is in Supplementary Material 1.

Appendix 3 - Methods for simulation of epidemics and validation of prevalence and genotypic value for individual disease status

We simulated endemics according to standard epidemiological theory to validate the numerical solution of the endemic prevalence (Equations 20a,b) and the genotypic values for binary disease status (Equation 24). We considered two compartments of individuals, susceptible individuals (S) and infected (I) individuals, and a so-called stochastic SIS-model where susceptible individuals can become infected, and infected individuals can recover and then immediately become susceptible again (Weiss and Dishon, 1971). For simplicity, we simulated genetic variation in susceptibility only, with $\gamma_i \sim \text{Lognormal}(0, \sigma^2_{A_{l\gamma}})$.

To limit Monte-Carlo error, we simulated a relatively large population of N = 2,000 genetically unrelated individuals for a total of 300,000 events (infection or recovery). We used a burn-in of 100,000 events before recording data on individual binary disease status. Hence, in the recorded data, the average individual experienced 100 events (50 infections and 50 recoveries).

The endemic was started by infecting a proportion P_0 = 1-1/c of the individuals, chosen at random. Subsequently, we sampled events (infection or recovery) and the individual involved using Gillespie's algorithm (Gillespie, 1977). For each infected individual, the probability of recovery was proportional to the recovery rate, α. For susceptible individual i the probability of infection was proportional to $c\gamma_i I/N$, I/N denoting the fraction of the population that is infected. Probabilities were accumulated over all individuals and scaled to a sum of 1 by dividing them by their sum. Finally, the specific event was sampled by drawing a random number, say x, from a standard uniform distribution and finding the event and the corresponding individual belonging to the probability interval $[x_l, x_h]$, where $x_l < x < x_h$. The disease status of that individual and I were updated before sampling the next event. The time of each event was not simulated. After 300,000 events, prevalence was calculated as the disease status averaged over the entire population, and also by individual, discarding the burn-in period. The regression coefficient of average individual disease status on G_γ was also estimated.

Additive genetic variance in log-susceptibility was $\sigma^2_{A_{l\gamma}} = 0.3^3$. Three scenarios were considered, differing in contact rate: c = 1.22 giving P = 0.2, c = 2 giving P = 5 and c = 5.15 giving P = 0.8. Those combinations of $\sigma^2_{A_{l\gamma}}$, c and P were found by numerically solving Equations 20a&b. The actual prevalences observed in the simulations were equal to these numerical solutions.

References

Gillespie, Daniel T. (1977). "Exact Stochastic Simulation of Coupled Chemical Reactions". The Journal of Physical Chemistry. 81 (25): 2340–2361.

Weiss, G. H., & Dishon, M. (1971). On the asymptotic behavior of the stochastic and deterministic models of an epidemic. Mathematical Biosciences, 11(3-4), 261-265.

Appendix 4. Additive genetic variance in log-normal traits.

We assumed log-normally distributed genotypic values for susceptibility, infectivity and recovery, also resulting in a log-normal distribution for G_{R_0} and for $1 - G_P$. Hence, genetic effects are additive on the log-scale, but taking the exponent introduces some non-additive genetic variance on the actual scale. Here we derive the fraction of the variance that is additive on the actual scale.

Because all genetic effects had a mean of zero on the log-scale, the problem is equivalent to finding the fraction of additive variance in $y = e^x$, where $x \sim N(\mu = 0, \sigma^2)$. From the properties of the long-normal distribution, $E(y) = e^{\mu}\, e^{\sigma^2/2}$. With a small change $d\mu$, the mean of y becomes $e^{d\mu}\, e^{\sigma^2/2}$. Hence, the mean of y changes by an amount $e^{\frac{\sigma^2}{2}}(e^{d\mu} - 1)$. Since $\lim_{d\mu \to 0} e^{d\mu} = 1 + d\mu$, this change corresponds to $e^{\frac{\sigma^2}{2}} d\mu$. Hence, the linear regression coefficient of y on x equals

$$b_{y,x} = e^{\sigma^2/2}.$$

Thus the additive effect for y equals

$$A_y = e^{\sigma^2/2}\, x,$$

and additive variance in y equals

$$\sigma^2_{A_y} = \sigma^2 e^{\sigma^2}.$$

The total variance in y follows from the properties of the log-normal distribution,

$$\sigma^2_y = (e^{\sigma^2} - 1)e^{\sigma^2}.$$

The non-additive variance in y, therefore, equals

$$\sigma^2_{NA_y} = \sigma^2_y - \sigma^2_{A_y} = \left(e^{\sigma^2} - 1\right)e^{\sigma^2} - \sigma^2 e^{\sigma^2} = e^{\sigma^2}(e^{\sigma^2} - 1 - \sigma^2)$$

Figure A3.1 illustrates that the additive fraction of $\sigma^2_{A_y}$ approaches 1 when σ^2 goes to zero. For $\sigma^2 = 0.5^2$, ~88% of the variance in y is additive. Variances on the log scale larger than 0.5^2 are unrealistic (see main text). This indicates that at least 88% of the genetic variance in susceptibility, infectivity, recovery, R_0 and prevalence is additive when they follow a log-normal distribution.

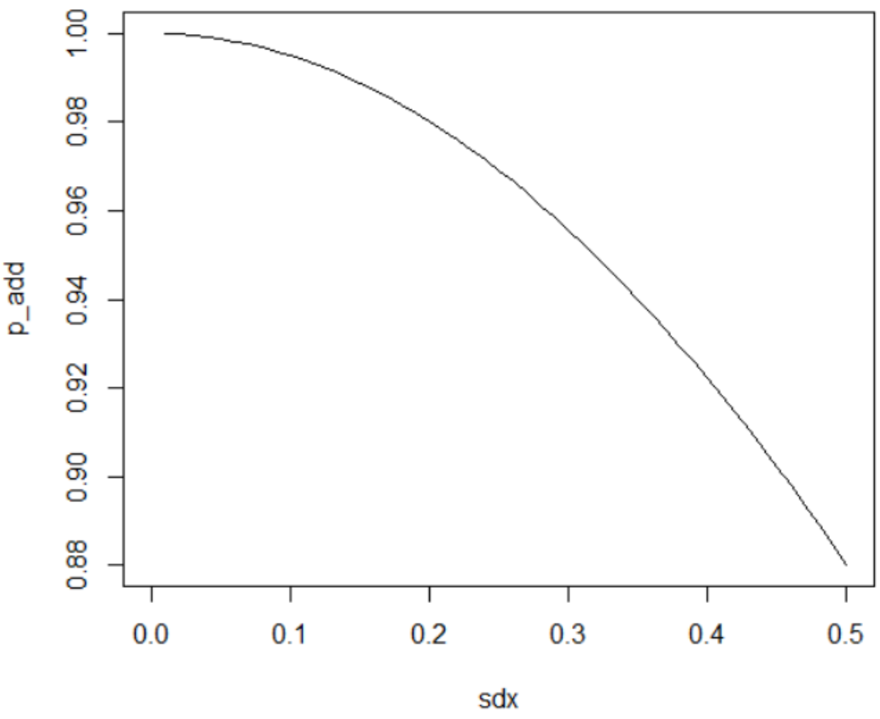


Figure A4.1 – The additive fraction of the variance in traits following a log-normal distribution.

Appendix 5. Genotypic value for individual disease status *vs.* genotypic value for prevalence, without genetic variation in infectivity

Without genetic variation in infectivity we have $\varphi_i = \varphi = 1$, because the scale is included in the effective contact rate c. From Equation 24, the genotypic value for individual binary disease status is,

$$G_{y,i} = \frac{\mathcal{R}_{0,i}P}{\mathcal{R}_{0,i}P + 1}$$

where, from Equation 20b,

$$\mathcal{R}_{0,i} = c\gamma_i/\alpha_i$$

From Equation 26a, the genotypic value for prevalence is,

$$G_{P,i} = 1 - \frac{1}{G_{R_0,i}}$$

where, from Equation 8,

$$G_{R_0,i} = c\gamma_i/\alpha_i$$

Hence, without genetic variation in infectivity, $\mathcal{R}_{0,i}$ and $G_{R_0,i}$ are identical, and we will use the symbol $G_{R_0,i}$ in the following.

The linear approximation of the relationship between G_y and G_P follows from a comparison of their first derivatives with respect to G_{R_0},

$$\frac{dG_p}{d\mathcal{R}_0} = \frac{1}{G_{R_0}^2}$$

$$\frac{dG_y}{d\mathcal{R}_0} = \frac{P\left(G_{R_0}P\ + 1\right) - G_{R_0}P^2}{(G_{R_0}P + 1)^2}$$

$$= \frac{P}{(G_{R_0}P + 1)^2}$$

Substituting Equation 3, assuming limited heterogeneity, yields

$$\frac{dG_y}{d\mathcal{R}_0} = \frac{G_{R_0} - 1}{G_{R_0}^3}$$

Hence,

$$\frac{dG_y}{dG_{R_0}} \bigg/ \frac{dG_p}{dG_{R_0}} = \frac{G_{R_0}}{G_{R_0} - 1} = P$$

Therefore, for a small change in an individual's genotypic value for R_0, the change in its genotypic value for binary disease status is only a fraction P of the change in its genotypic value for prevalence,

$$dG_y \,/\, dG_p = P$$

Hence, when expressed relative to their mean, G_y and G_p differ approximately by a factor P (see also Figure 4 in Bijma 2020). This result is approximate, because the true relationship is non-linear and the expression $P_{eq} = 1 - 1/R_0$ is approximate with variation among individuals. For realistic magnitudes of the genetic variance, however, the non-linearity is limited. Note that the above derivation does not require the assumption of a log-normal distribution of susceptibility and recovery.

Appendix 6 Methods for observed response to selection

First a base population was generated of N = 4000 unrelated individuals, with genetic variation in susceptibility only. No distinction was made between males and females. For each individual, breeding values for the logarithm of susceptibility were sampled from $A_{l\gamma} \sim N\left(\bar{\gamma} = 0, \sigma_{A_\gamma}^2 = 0.3^2\right)$, and individual susceptibility was calculated as $\gamma_i = e^{A_{\gamma,i}}$. The expected prevalence for the base generation was calculated as $P_0 = 1 - 1/c$, with a c of either 2 or 10, and the initial disease status of base generation individuals was sampled at random from $Bin(4000, P_0)$.

Next, an endemic was simulated as described in Appendix 3, for a total of 15,000 events (sum of infections and recoveries), consisting of a burn-in of 10,000 events and 5,000 recorded events. The 4,000 individuals were ordered based on their mean individual disease status over the 5,000 recorded events (so based on 1.25 events on average per individual), and the 2000 individuals with the lowest values were selected as parents of the next generation (corresponding to a selected proportion of 0.5).

Selected parents were mated at random. Each pair of parents produced two offspring, resulting in N = 4,000 offspring. Offspring inherited the breeding value for the logarithm of susceptibility in a Mendelian fashion; $A_{l\gamma,\text{offspring}} = \frac{1}{2} A_{l\gamma,parent1} + \frac{1}{2} A_{l\gamma,parent2} + N\left(0, \frac{1}{2}\sigma_{A_\gamma}^2\right)$. The initial disease status of offspring (*i.e.*, at the start of the burn-in period of their generation) was sampled at random from $Bin(4000, P_{\text{offspring}})$, where $P_{\text{offspring}}$ denotes the expected prevalence in the offspring generation, calculated as $P_{\text{offspring}} = \max\,[1 - \frac{1}{c\, e^{\bar{A}_\gamma}}; 0.02]$. The 0.02 guaranteed an average of at least 80 infected individuals at the start of the endemic in any generation, also when the expected prevalence was zero (*i.e*, when $1 - \frac{1}{c\, e^{\bar{A}_\gamma}} \leq 0$). Then an endemic was started, as described above for the base generation, *etc*. This process was repeated until the number of infected individuals dropped to zero, implying extinction of the infection.

Appendix 7 - Final size of epidemics vs. equilibrium prevalence of endemics

For endemic infections, the "size" is measured by the equilibrium prevalence, P, as given in Equation 3 (assuming limited heterogeneity), and illustrated in Figure 1 of the main text. For epidemic infections described by a Susceptible – Infectious – Recovered (SIR) model, the size is measured be the fraction of the population that has been infected when the epidemic has ended, $R_\infty/N = 1 - s_\infty$ (Kermack & McKendrick 1927; Diekmann *et al.*, 2013). Hence, s_∞ denotes the fraction of the population still susceptible after the epidemic has ended, *i.e.*, that has escaped from infection. The final size is determined by R_0, and follows from numerically solving the implicit equation

$$\log(s_\infty) = R_0(s_\infty - 1)$$

Figure A7.1 shows a comparison of the final size of an epidemic infection and the equilibrium prevalence of an endemic infection, as a function of R_0.

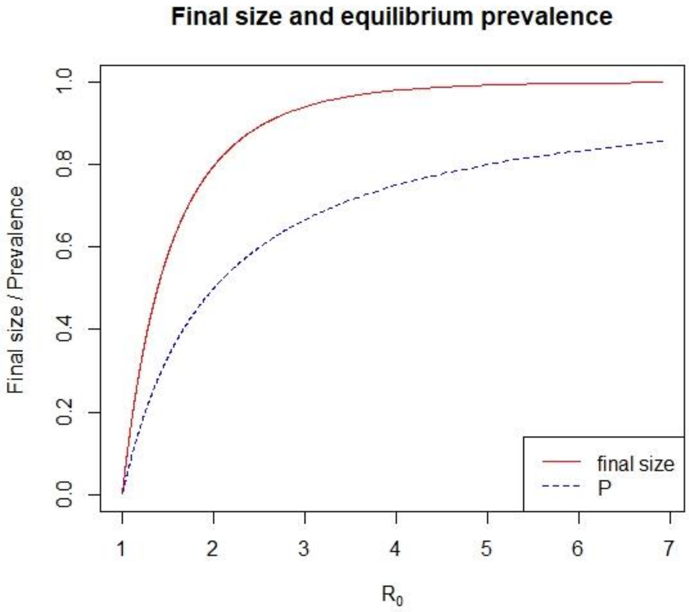


Figure A7.1 – Final size of an epidemic infection from a SIR model, and equilibrium prevalence of an endemic infection from a SIS model, as a function of R_0.

At low values of R_0, the line for final size is steeper, while at high values of R_0 the line for prevalence is steeper. The point where both slopes are equal follows from equating the derivatives of both lines with respect to the logarithm of R_0,

$$\frac{d(1-s_\infty)}{dlR_0} = -\frac{s_\infty(s_\infty - 1)\log(s_\infty)}{s_\infty - s_\infty\log(s_\infty) - 1}$$

$$\frac{dP}{dlR_0} = \frac{1}{R_0}$$

A numerical solution, also using the above expression for final size, yields $R_0 \approx 2.15$, $P \approx 0.54$ and $1 - s_\infty \approx 0.83$. Hence, below $R_0 \approx 2.15$ the change in final size due to a marginal change in R_0 is greater than the change in endemic prevalence due to that same change in R_0. Hence, at the same response to selection in R_0, the final size of an epidemic will respond faster to selection than the equilibrium prevalence of an endemic infection when $R_0 < \sim 2.15$.

Supplementary material

Supplementary Material 1: R-code to numerically find the endemic equilibrium prevalence.

See file "Supplementary Material 1 - Numerical Solution Prevalence Heterogeneity.R"

Supplementary Material 2: Additional results for the endemic equilibrium prevalence.

See file "Supplementary Material 2 - Effect of heterogeneity on endemic prevalence.xlsx"

Supplementary Material 3: Validation of Equation 34

Equation 34 states that, in the absence of genetic variation in infectivity, the breeding value for individual disease status is given by

$$A_y \approx P(1-P)A_{lR_0}$$

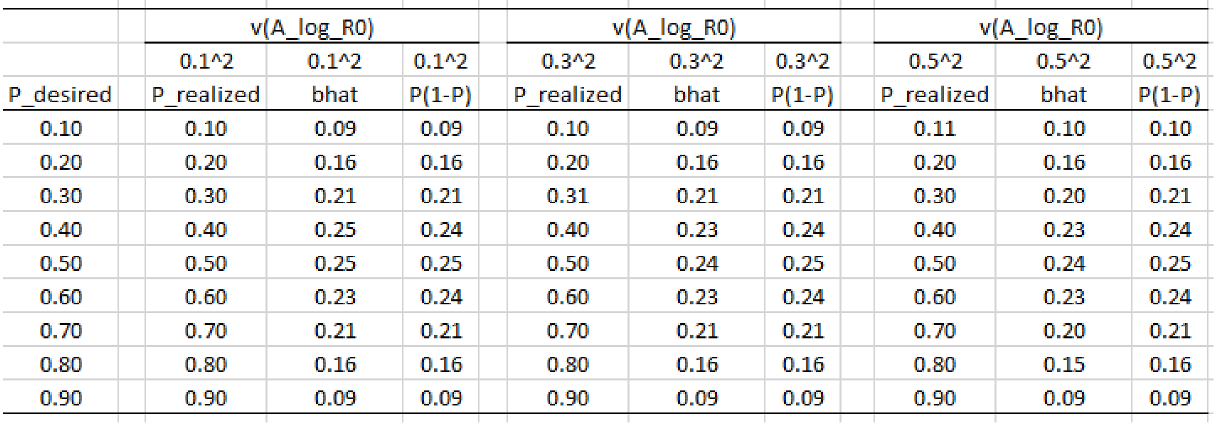
where P denotes prevalence and A_{lR_0} the breeding value for the logarithm om R_0. We validated this expression using stochastic simulation of endemics following standard epidemiological theory (as outlined in Appendix 4). We simulated a population of N = 10,000 individuals, for a total of 500,000 events (sum of infections and recoveries), using the first 100,000 events as burn in. We considered input values of $\sigma^2_{A_{l\gamma}}$ = 0.1^2, 0.3^2 and 0.5^2 and a prevalence ranging from 0.1 to 0.9, with steps of 0.1. We found the values for the contact rate that correspond to these prevalences numerically from solving Equations 20a and b (Table S3.1; R-code in Supplementary Material 1).

Table S3.1 - Contact rates required to find a certain prevalence, for 3 values of $\sigma^2_{A_{l\gamma}}$.

P	0.1^2	0.3^2	0.5^2
0.1	1.108	1.073	1.010
0.2	1.247	1.218	1.163
0.3	1.424	1.404	1.362
0.4	1.665	1.652	1.628
0.5	2.000	2.000	2.000
0.6	2.502	2.523	2.560
0.7	3.340	3.393	3.495
0.8	5.015	5.135	5.374
0.9	10.040	10.364	11.025

By definition, the individual breeding value equals the regression of individual trait value on additive genetic effects. Validation, therefore, focussed on the comparison of the $P(1-P)$ term in the above expression for A_y to the regression coefficient of individual disease status on A_{lR_0} ($b_{y,A_{lR_0}}$) estimated from the simulated data. Table S3.2 shows close agreement between these two parameters.

Table S3.2 - Comparison of estimated regression coefficient (bhat) of individual disease status on individual breeding value for the logarithm of R_0 with its expected value of $P(1\text{-}P)$. For three levels of genetic variance in log(R_0). P_desired denotes the desired prevalence; P_realized denotes the realized prevalence using contact rates given in Table S3.1.



	v(A_log_R0)			v(A_log_R0)			v(A_log_R0)		
	0.1^2	0.1^2	0.1^2	0.3^2	0.3^2	0.3^2	0.5^2	0.5^2	0.5^2
P_desired	P_realized	bhat	P(1-P)	P_realized	bhat	P(1-P)	P_realized	bhat	P(1-P)
0.10	0.10	0.09	0.09	0.10	0.09	0.09	0.11	0.10	0.10
0.20	0.20	0.16	0.16	0.20	0.16	0.16	0.20	0.16	0.16
0.30	0.30	0.21	0.21	0.31	0.21	0.21	0.30	0.20	0.21
0.40	0.40	0.25	0.24	0.40	0.23	0.24	0.40	0.23	0.24
0.50	0.50	0.25	0.25	0.50	0.24	0.25	0.50	0.24	0.25
0.60	0.60	0.23	0.24	0.60	0.23	0.24	0.60	0.23	0.24
0.70	0.70	0.21	0.21	0.70	0.21	0.21	0.70	0.20	0.21
0.80	0.80	0.16	0.16	0.80	0.16	0.16	0.80	0.15	0.16
0.90	0.90	0.09	0.09	0.90	0.09	0.09	0.90	0.09	0.09