Benefits of Dominance over Additive Models for the Estimation of Average Effects in the Presence of Dominance

Pascal Duenk,1 Mario P. L. Calus, Yvonne C. J. Wientjes, and Piter Bijma
Animal Breeding and Genomics Group, Wageningen University & Research, 6700 AH Wageningen, The Netherlands
ORCID IDs: 0000-0001-8403-968X (P.D.); 0000-0002-3213-704X (M.P. L.C.)

ABSTRACT In quantitative genetics, the average effect at a single locus can be estimated by an additive (A) model, or an additive plus dominance (AD) model. In the presence of dominance, the AD-model is expected to be more accurate, because the A-model falsely assumes that residuals are independent and identically distributed. Our objective was to investigate the accuracy of an estimated average effect ($\hat{a}$) in the presence of dominance, using either a single locus A-model or AD-model. Estimation was based on a finite sample from a large population in Hardy-Weinberg equilibrium (HWE), and the root mean squared error of $\hat{a}$ was calculated for several broad-sense heritabilities, sample sizes, and sizes of the dominance effect. Results show that with the A-model, both sampling deviations of genotype frequencies from HWE frequencies and sampling deviations of allele frequencies contributed to the error. With the AD-model, only sampling deviations of allele frequencies contributed to the error, provided that all three genotype classes were sampled. In the presence of dominance, the root mean squared error of $\hat{a}$ with the AD-model was always smaller than with the A-model, even when the heritability was less than one. Remarkably, in the absence of dominance, there was no disadvantage of fitting dominance. In conclusion, the AD-model yields more accurate estimates of average effects from a finite sample, because it is more robust against sampling deviations from HWE frequencies than the A-model. Genetic models that include dominance, therefore, yield higher accuracies of estimated average effects than purely additive models when dominance is present.

In quantitative genetics, the average effect at a single locus can be estimated by an additive (A) model, or an additive plus dominance (AD) model. In the presence of dominance, the AD-model is expected to be more accurate, because the A-model falsely assumes that residuals are independent and identically distributed. Our objective was to investigate the accuracy of an estimated average effect ($\hat{a}$) in the presence of dominance, using either a single locus A-model or AD-model. Estimation was based on a finite sample from a large population in Hardy-Weinberg equilibrium (HWE), and the root mean squared error of $\hat{a}$ was calculated for several broad-sense heritabilities, sample sizes, and sizes of the dominance effect. Results show that with the A-model, both sampling deviations of genotype frequencies from HWE frequencies and sampling deviations of allele frequencies contributed to the error. With the AD-model, only sampling deviations of allele frequencies contributed to the error, provided that all three genotype classes were sampled. In the presence of dominance, the root mean squared error of $\hat{a}$ with the AD-model was always smaller than with the A-model, even when the heritability was less than one. Remarkably, in the absence of dominance, there was no disadvantage of fitting dominance. In conclusion, the AD-model yields more accurate estimates of average effects from a finite sample, because it is more robust against sampling deviations from HWE frequencies than the A-model. Genetic models that include dominance, therefore, yield higher accuracies of estimated average effects than purely additive models when dominance is present.

$\alpha = a + (1 - 2p)d,$

where $a$ is half the difference in genotypic value between both homozygotes, and $d$ is the difference between the genotypic value of the heterozygote and the average genotypic value of both homozygotes. With genomic data, additive (A) models estimate $\alpha$ by linear regression of phenotypes on allele counts (i.e., genotypes). Additive plus dominance (AD) models estimate the additive and dominant gene effects separately, after which $\alpha$ can be obtained from Equation 1 (Lynch and Walsh 1998). For both models, the part of dominance
that is not captured by the average effect is called the dominance deviation (Falconer and Mackay 1996).

When the A-model is used, dominance deviations are not modeled and thus become part of the residual. As a consequence, the residuals are not independent and identically distributed (IID), because dominance deviations are different across genotypes (Ott and Longnecker 2010). The A-model may therefore give inaccurate estimates of \(\alpha\), because it falsely assumes that the residuals are IID. When the AD-model is used, dominance deviations are explicitly modeled, and the residuals will more likely be IID. In the presence of dominance, the AD-model may therefore yield more accurate estimates of \(\alpha\) than the A-model. In contrast to the A-model, however, the AD-model requires the estimation of two effects instead of one (for a single locus), which may reduce the accuracy with which these effects are estimated. Additionally, dominance effects are generally smaller and therefore harder to estimate than additive effects (Lynch and Walsh 1998). For these reasons, the AD-model may require more individuals to be sampled for an accurate estimation of \(\alpha\).

### Theory

#### Effect of dominance

In quantitative genetics, heritabilities, dominance effects, sample sizes, and allele frequencies are of interest. We start with some theory of a single locus model, then derive the expected heritabilities, dominance effects, sample sizes, and allele frequencies. Additionally, dominance effects are generally smaller and therefore harder to estimate than additive effects. In the presence of dominance, the AD-model may yield more accurate estimates of \(\alpha\) than the A-model.

#### Expectation of \(\alpha\)

If we take a random sample of \(N\) individuals from a large population in HWE that has allele frequency \(p\), the expectation of \(\alpha\) can be computed using probabilities and estimates of each possible sample composition. We define \(c\) as a set of variables \(\{n_0, n_1, n_2\}\) that describe unique sample compositions, where \(n_0\) is the number of individuals with genotype 0, \(n_1\) is the number of individuals with genotype 1, and \(n_2\) is the number of individuals with genotype 2. The probability of sampling \(c\) is calculated from the multinomial probability function

\[
P(c|N, p) = \frac{N!}{n_0! n_1! n_2!} p^{n_0} (1-p)^{n_1} (1-2p)^{n_2}.
\]

#### Estimation of \(\alpha\)

To investigate the effects of \(\alpha\) given \(c\), we use the AD-model to estimate the average effect at a single locus in the presence of dominance, using either an A-model or an AD-model. We start with some theory of a single locus model, then derive the expected heritabilities, dominance effects, sample sizes, and allele frequencies. We then calculate the mean RMSE for several degrees of dominance over the distribution of allele frequency, and identify mechanisms that underlie the differences between the A-model and AD-model.

#### THE MODEL

Our interest is to estimate the average effect \((\alpha)\) at a single locus in a large population that is in HWE, from data collected as a finite sample of that population. The average effect will be treated as a fixed effect (as in quantitative genetics), and not a random variable (e.g., as in genomic prediction) (de los Campos et al. 2015). In quantitative genetics, \(\alpha\) at a single locus can be estimated from the sample by linear regression using an A-model or an AD-model. The A-model estimates \(\alpha\) directly through linear regression of phenotypic values on allele counts, while the AD-model estimates \(\alpha\) and \(d\) gene effects by multiple linear regression.

#### Equation 8

\[
y = \mathbf{x} \alpha + \mathbf{m} d + \mathbf{e},
\]

where \(\mathbf{m}\) is a dominance indicator vector with \((0 - 2p_i (1 - p_i))\) for homozygous individuals, and \((1 - 2p_i (1 - p_i))\) for heterozygous individuals. Vectors \(y\) and \(\mathbf{x}\) are the same as in the A-model, and \(\mathbf{e}\) is a vector of residuals. Note that this is the genotypic parameterization as described by Vitezica et al. (2013). With the AD-model, the LSE of \(\alpha\) and \(d\) are

\[
\hat{\alpha}_{AD} = \hat{\alpha} + (1 - 2p) \hat{d}.
\]

By definition, \(\hat{\alpha}\) from both models gives an estimate of the average effect in the sample (Falconer and Mackay 1996). Because the size of the sample is finite, genotype and allele frequencies in the sample might deviate from the frequencies in the total population. These deviations might introduce error in the estimation of \(\alpha\).

### Conclusion

The AD-model estimates the additive \((\alpha)\) and dominant \((d)\) gene effects by multiple linear regression.

\[
\hat{\alpha}_{AD} = \begin{bmatrix} \mathbf{x} \mathbf{x} & \mathbf{x} \mathbf{m} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{x} \mathbf{y} \end{bmatrix}.
\]

The AD-model requires more individuals to be sampled for an accurate estimation of \(\alpha\). In contrast to the A-model, however, the AD-model requires the estimation of two effects instead of one (for a single locus), which may reduce the accuracy with which these effects are estimated. Additionally, dominance effects are generally smaller and therefore harder to estimate than additive effects. In the presence of dominance, the AD-model may yield more accurate estimates of \(\alpha\) than the A-model.
Root mean squared error

The RMSE is defined as the root of the expected squared difference between the ț estimated from the sample, and the true value of a

$$\text{RMSE}(\hat{a}) = \sqrt{E[(\hat{a} - a)^2]} = \sqrt{\frac{\sum_{i=0}^{N-1} \sum_{n_i=0}^{N-i} I(c)\delta(c)}}{\sum_{n_i=0}^{N-i} I(c)\delta(c)}, \quad (10)$$

where $\delta(c)$ is the contribution of finite sampling deviation to the RMSE

$$\delta(c) = P(c)(\hat{a}(c) - a)^2. \quad (11)$$

We define $\delta(c)$ here because we will later on focus on the contribution of a single finite sample c to the RMSE. The above expressions will be used to investigate the effect of N, $H^2$, p, and d on the RMSE of $\hat{a}$ with the A-model and the AD-model.

METHODS

We aim to illustrate the effect of sample size (N), broad-sense heritability ($H^2$), allele frequency p, and dominance effect d, on RMSE of estimated average effects ($\hat{a}$). As a base scenario, we chose one for both the additive and dominance effect of the gene (e.g., full dominance). The expected value of $\hat{a}$ was calculated for $N \in \{300, 500, 1000\}$, $H^2 \in \{0.01, 0.05, 1\}$, and $p = [0.001 – 0.999]$ (increments of 0.001), with the A-model (Equation 3) and AD-model (Equation 5). The variation in broad-sense heritability was achieved by adding random residuals to the phenotypes (y). In addition, we varied the dominance effect ($d \in \{0.0, 0.1, 0.2, 0.5\}$) for the scenario where $N = 500$ and $H^2 = 0.05$.

The $\hat{a}(c)$ from the AD-model were computed using the sample allele frequency ($\hat{p}$) in Equation 6 instead of the population allele frequency (p), because the latter is usually unknown. For samples where one of the genotypes was missing, $\hat{a}(c)$ with the AD-model was computed in the same way as with the A-model, because in those cases the vector of genotypes x was completely confounded with dominance vector m.

Additionally, to quantify the average accuracy of $\hat{a}$, we computed the mean RMSE of $\hat{a}$, assuming a distribution for the allele frequency. For this purpose, we used the RMSE as a function of p and numerically integrated over p using its expected distribution under a drift model,

$$\text{RMSE} = \int_{\hat{p}}^{1-\hat{p}} f(p)\text{RMSE}(p)dp. \quad (12)$$

Here, $N_i$ is the effective population size, $f(p)$ is the distribution of allele frequencies when mutation is ignored, p ranges from 1/2$N_i$ to $1 - 1/2N_i$ (Wright 1931; Goddard 2009), and

$$f(p) = \frac{k}{2p(1-p)} \quad (13)$$

To ensure that $\int f(p)dp = 1$, $k$ was given a value of 1/log(2$N_i$ – 1). The resulting distribution of allele frequencies is U-shaped, and a low $N_i$ yields a more uniform distribution than a high $N_i$. We computed the mean RMSE for several $N_i$ (50, 100, and 200), (N 200–600), and sizes of dominance effect d (0.5, 1, and 1.5). We considered sample sizes up to 600 instead of 1000 to reduce computation time. In these scenarios, both $H^2$ and the additive gene effect (a) were equal to one.

The data used can be regenerated exactly following the descriptions in this paper.

RESULTS

Root mean squared error

Figure 1 shows the RMSE of $\hat{a}$ with the A- or AD-model, for $a = 1$ and $d = 1$. For all scenarios, the RMSE of $\hat{a}$ was smaller with the AD-model than with the A-model.

In scenarios where $H^2 = 1$, RMSE was symmetrical around $p = 0.5$ with both the A- and AD-model. For brevity, we will therefore only describe the pattern for $p < 0.5$. For both models and all N, the RMSE was smallest when p was close to 0, and increased when allele frequency increased. With the A-model, RMSE was largest around $p = 0.04$ and then decreased when p moved toward 0.5. With the AD-model, RMSE was also largest around $p = 0.04$, then decreased when p moved toward 0.1, after which RMSE slightly increased again until $p = 0.5$.

With $H^2 < 1$, RMSE showed a similar pattern, but was not symmetrical around $p = 0.5$. Compared with $H^2 = 1$, the RMSE was larger for all p, but this contrast decreased when p increased. This asymmetry was a result of fixing $H^2$ in the simulations, which caused the ratio of the dominance variance and residual variance to increase with p. For all scenarios, RMSE decreased when N increased.

Figure 2 shows the RMSE of $\hat{a}$ with the A- or AD-model, for $a = 1$, $N = 500$, $H^2 = 0.05$, and different dominance effects (d). For d = 0 and $d = 1$, there was almost no difference in RMSE between the A- and AD-model. This indicates that in the absence of dominance, there was no disadvantage of using the AD-model in terms of RMSE. For $d = 0.1$, there was no apparent benefit from using the AD-model. For $d = 0.2$ and $d = 0.5$, however, the AD-model had lower RMSE than the A-model.

Contribution of finite sampling deviation to the root mean squared error

When there is no environmental variance ($H^2 = 1$) and the model is correct, the RMSE of $\hat{a}$ is expected to be zero. The results, however, show that the RMSE is larger than zero with both the A- and AD-model. To gain more insight into the sources of this error, we investigated the contribution of single samples to the RMSE, for one scenario where $H^2 = 1, N = 300, p = 0.10$, and $a = d = 1$, so that $\alpha = 1.8$. For this purpose, we studied the squared difference between $\hat{a}(c)$ and a (i.e., squared error), as a function of the realized number of individuals with genotype 2 (n2). The samples have different probabilities of occurring, so that some samples may contribute more to the total RMSE than others. We therefore investigated the contribution of finite sampling deviation to the RMSE ($\delta(c)$), by weighting the squared errors of $\hat{a}(c)$ with their probabilities (see Equation 11).

Additive model: Figure 3A shows the squared error as a function of the realized number of individuals with genotype 2, for the A-model. The realized number of individuals with genotype 2 in the sample is expressed as a departure from its expectation (i.e., $\Delta n_2$), where the expectation is $E(n_2) = \hat{p}N = 3$. The squared error was smallest when $\Delta n_2$ was zero and increased as $\Delta n_2$ moved away from zero. The remaining variance in squared error for a given value of $\Delta n_2$ (as shown by the boxplots) was due to variation in the difference between $\hat{p}$ and $p$ (i.e., $\Delta p$). For example, when $\Delta n_2 = 3$, the allele frequency in the sample can vary, because the number of sampled heterozygotes can vary. This variation in $\Delta p$ affects $\hat{a}(c)$, except when $\Delta n_2 = -3$. In that case, the number of individuals with genotype 2 was zero (in this example) and $\hat{a}(c)$ was always the slope of a line between two data points.
Figure 3B shows the effect of $D_n^2$ and $D_p$ on $\delta(c)$ for the A-model. The sample where $D_n^2 = 0$ and $D_p = 0$ did not contribute to the RMSE ($\delta(c) = 0$). Samples where $D_n^2 < 0$ had the largest contributions to the RMSE, and samples where $D_n^2 > 0$ had somewhat smaller contributions. Figure 3B also shows that $D_p$ contributed less to the RMSE than $D_n^2$; because there were samples where $D_p = 0$; but $\delta(c)$ was relatively large.

Additive plus dominance model: Figure 3C shows the squared error as a function of $D_n^2$, for the AD-model. The squared error was small and about equal for all $D_n^2$, except for $D_n^2 = -3$, where the squared error was largest and exactly the same as with the A-model (see Figure 3A), because there were no individuals with genotype 2 in the sample. Similar to the A-model, the remaining variance (as shown by the box-plots) was due to variation in the difference between $p$ and $\bar{p}$ ($\Delta p$).

Figure 3D shows the effect of $D_n^2$ and $\Delta p$ on $\delta(c)$ for the AD-model. Samples where both $D_n^2 \neq -3$ and $\Delta p = 0$, did not contribute to the RMSE ($\delta(c) = 0$). Samples where $D_n^2 = -3$ showed the largest contribution, while all other samples showed small $\delta(c)$. Similar to the A-model, Figure 3D shows that $\Delta p$ was not an important source of error.

A- vs. AD-model: In conclusion, even when the locus explains all variance (i.e., $H^2 = 1$), $\hat{a}$ shows error with both the A- and AD-model when it is based on a finite random sample from a population in HWE and dominance is present. With the A-model, the error originated mainly from sampling deviations of genotype frequencies from expected HWE frequencies ($\Delta HWE$), and to a lesser extent from sampling deviations of allele frequencies ($\Delta p$) (Figure 3B). With the AD-model, the error originated from $\Delta p$ only, provided that all three genotype classes were sampled (Figure 3D). These results partly explain the patterns of RMSE in Figure 1 (see Appendix A for more detail).

Mean RMSE across allele frequency distribution
Above, we illustrated the RMSE of $\hat{a}$ as a function of $p$. Now, we present the mean RMSE averaged over the distribution of $p$, for $a = 1$ and $H^2 = 1$, assuming a U-shaped distribution of $p$ as a function of $N_e$, and for different values for $N$ and $d$ (Figure 4). For all scenarios, the mean RMSE with the A-model was about twice as large as the mean RMSE with the AD-model.

With both models, the mean RMSE was zero when $d$ was zero (data not shown) and increased as $d$ increased. The mean RMSE decreased when $N$ increased. The mean RMSE decreased a little when $N_e$ increased, which was caused by differences in the U-shaped distribution of allele frequencies. For example, when $N_e = 50$, the percentage of loci with an allele frequency outside the 0.05–0.95 range was 36%, whereas when $N_e = 200$, this percentage was 51%. Loci in this range have a low RMSE (see Figure 1 and Appendix A), and therefore a higher $N_e$ results
in a lower mean RMSE. The effect of N, on the mean RMSE decreased as N increased. Results were identical when a was changed, because the mean RMSE scales linearly with the absolute dominance effect d, and not with the dominance coefficient d/a.

**DISCUSSION**

We investigated the accuracy (in terms of RMSE) of estimated average effects (\(\hat{\alpha}\)) in the presence of dominance, using a single locus model including only an additive or an additive plus dominance effect. In the presence of dominance, the A-model falsely assumes that residuals are IID. The AD-model was therefore expected to better fit the data and give more accurate estimates of \(\alpha\), but only when dominance is present and sample size sufficient for the dominance effect to be accurately estimated. Our results, however, show that the AD-model was always equally or more accurate than the A-model, even with small sample sizes (i.e., \(N = 300\)), a heritability lower than one (i.e., \(H^2 < 1\)), or in the absence of dominance.

With the A-model, both sampling deviations of genotype frequencies from HWE frequencies (\(\Delta HWE\)) and sampling deviations of allele frequencies (\(\Delta p\)) contributed to the error. With the AD-model, only sampling deviations of allele frequencies contributed to the error, provided that all three genotype classes were sampled. The contribution of \(\Delta p\) to the error was much smaller than the contribution of \(\Delta HWE\). The AD-model was therefore more accurate than the A-model. Thus, even when the locus explained all variance (i.e., \(H^2 = 1\)), the mean RMSE decreased as sample size increased, because with larger sample sizes, deviations from HWE that considerably affect \(\hat{\alpha}\) had a lower probability of occurring. Additionally, with larger sample sizes, the chance of missing one of the genotype classes was smaller, which further reduced the RMSE. The (mean) RMSE of \(\hat{\alpha}\) was always smaller with the AD-model than with the A-model. The RMSE of \(\hat{\alpha}\) scaled linearly with d; if d doubled, the RMSE also doubled. Remarkably, in the absence of dominance, there was no disadvantage of using the AD-model. Hence, the AD-model yielded equally or more accurate estimates of average effects than the A-model for all scenarios considered.

With the A-model, \(\hat{\alpha}\) is computed as the linear regression coefficient of genotypic values on allele counts (Fisher 1941), which yields the average effect in the sample (\(\alpha_s\)), rather than the average effect in the whole population (\(\alpha\)). Hence, the expectation of \(\hat{\alpha}\) is equal to

\[
E(\hat{\alpha}_A) = \alpha + (1 - 2p_0) d \left(\frac{1 - F_1}{1 + F_1}\right),
\]

where \(F_1\) measures the deviation from HWE in the sample (\(\Delta HWE\)) (Haldane 1954; Falconer 1985). Here, \(F_1\) is defined as one minus the ratio between the observed number of heterozygotes and the expected number of heterozygotes based on the sample allele frequency (Haldane 1954; Wright 1969). With the AD-model, \(\hat{\alpha}\) is computed from \(\alpha\) and \(d\), which are simultaneously estimated from the data. Unlike the A-model (where \(E(\hat{\alpha}) = \alpha\)), the expectation of \(\hat{\alpha}_{AD}\) is equal to

\[
E(\hat{\alpha}_{AD}) = \alpha + (1 - 2p_0) d,
\]

when all three genotype classes are sampled. Comparison of Equation 14 and Equation 15 shows that the error in \(\hat{\alpha}_A\) originates from both \(\Delta HWE\) and \(\Delta p\), while the error in \(\hat{\alpha}_{AD}\) originates from \(\Delta p\) only, except when one of the genotypes is missing in the sample. When only two genotype classes are sampled, the AD-model reduces to the A-model. With the AD-model, the contrast between the mean genotypic value of the homozygotes and the genotypic value of the heterozygotes (\(d\)) does not depend on the number of individuals in these two groups. This is why the AD-model is more robust against deviations from HWE than the A-model.

These results were confirmed by mathematical derivations of the error with the two models (Appendix B). In theory, the error from the A-model can be quantified when \(p_0\), \(F_1\), and \(d\) are known, and from the AD-model when \(p_0\), \(p\), and \(d\) are known. In real data, however, \(p\) (and also \(d\) with the A-model) is not known, and therefore the error cannot be quantified. As a result, the error cannot be removed from either of the two models. In conclusion, the AD-model is preferred for the estimation of average effects when dominance is present, because it yields more accurate estimates than the A-model, particularly when sample sizes are small.

In this study, we used the so-called genotypic parameterization of the AD-model, as opposed to the breeding parameterization (Vitezica et al. 2013). The results, however, were identical to the breeding parameterization (results not shown), because the two parameterizations are equivalent.

Additional to the contribution of dominance to additive variance, evidence for the contribution of epistasis is increasing (Mackay 2015; Monnahan and Kelly 2015). Our results show that modeling dominance improves estimated average effects, and it may therefore be tempting to hypothesize that modeling epistasis may also improve estimates. However, investigating the benefit of modeling epistasis for the accuracy of \(\hat{\alpha}\) is not straightforward, because it requires extension to multiple loci.

Taking a finite sample from a large population, which was done in this study, closely resembles a sharp reduction to a small population size, known as a bottleneck. In a small population, genotype frequencies deviate from HWE even under random mating. The expected genotypic frequency for heterozygotes is equal to \(2p_0(1 - p_0)(1 - F_1)\) (Haldane 1954). In turn, the expectation of \(F_1\) depends on the size of the bottleneck (or sample size, \(N\)), and is equal to \(-1/(2N - 1)\) with random
mating (Kimura and Crow 1963). This indicates that the expected heterozygosity in the sample is larger than the HWE frequency calculated from the sample allele frequency. The effect of $D_{\text{HWE}}$ on estimated average effects was studied by Wang et al. (1998), who focused on the consequences for the additive genetic variance. In agreement with our results, they showed that the average effect was not influenced by $D_{\text{HWE}}$ when $d = 0$, or when $p \approx 0.5$. Furthermore, the effect of $D_{\text{HWE}}$ on estimated average effects depended on the size of the bottleneck (or sample size, $N$) and the size of dominance effect ($d$) (Wang et al. 1998). Because the effects of a bottleneck are very similar to the effects of taking a small sample from a large population, the results of our study also apply to populations in a bottleneck.

We quantified the error in estimates of $\alpha$ that originated from $\Delta_{\text{HWE}}$ in random finite samples from a population of unrelated individuals. We purposefully used relatively small sample sizes to illustrate the effect. Although sample sizes taken in empirical studies may be larger, effective sample size may be much smaller, because actual populations often have small effective population size ($N_e$) (Hall 2016). This low $N_e$ is related to the family structure in the population, where many individuals are bred from a limited number of parents, so that
N_r \ll N. Hence, the effective sample size may be much smaller than N, because the sample will partly consist of related individuals. Because of this relatedness, sampling deviations in allele and genotype frequencies can be larger than expected based on sample size. The sample sizes chosen in this study may therefore be similar to effective sample sizes in empirical studies. As an example, we investigated the SD of $F_s$ across allele frequencies in a dataset of $\sim 3500$ pigs (Cleveland et al. 2012). The resulting value was comparable to the expected SD of $F_s$ for samples of 500–1000 animals (see Appendix C), which supports our expectation that effective number of sampled individuals may be smaller than the actual number of sampled individuals. Furthermore, in many studies that use genotype data, markers are removed if they show a significant deviation from HWE. The significance threshold that is used for HWE filtering, however, is often very liberal (Gondro et al. 2013). Consequently, there are still many markers left in the data that deviate from HWE and may give inaccurate estimates of average effects. As a result, we expect that the magnitude of $\Delta HWE$ simulated in this study may be similar to $\Delta HWE$ in empirical studies.

The estimation of average effects at single loci, as presented in this study, may be relevant for genome-wide association studies (GWAS). In GWAS, a large number of markers spread across the genome are each tested and $d$ (e.g., Lopes et al. 2014; Aliloo et al. 2015; Huang et al. 2015; Bennewitz et al. 2017) and, to our knowledge, none have investigated differences in accuracy of estimated average effects between the A-model and AD-model. The effects of sampling genotypes on $\alpha$ shown in this study apply to $\alpha_m$ in GWAS, because $\alpha_m$ are usually estimated by ordinary least squares. Using the AD-model in GWAS will therefore yield more accurate estimates of average effects and explained variance of markers.

The results presented in this study may also be relevant for genomic prediction. In genomic prediction, genomic estimated breeding values (GEBVs) are calculated as the sum of many estimated average effects multiplied by their marker genotypes (Meuwissen et al. 2001). Differences in accuracy of GEBVs may therefore be related to differences in accuracy of the estimated average effects. Our results, however, cannot be extrapolated directly to accuracy of GEBVs for several reasons. In this study, we considered a single locus, estimated $\alpha$ as a fixed effect, and assumed known genotypes of the quantitative trait locus (QTL). By contrast, GEBVs are based on many marker loci, for which all $\alpha$’s are estimated simultaneously as random effects (Meuwissen et al. 2001). In genomic prediction, the effect of a single QTL is likely to be explained by multiple markers, and errors of individual marker effects may cancel out to some extent when accumulated within individuals to compute their GEBVs. Additionally, random effect models shrink average effects toward zero (Whittaker et al. 2000), which may shrink the sampling error as well. In conclusion, to translate our results to accuracy of...
GEBVs, this research should be extended to the estimation of multiple random effects based on marker genotypes.

Neither GWAS nor genomic prediction are based on the genotypes at QTL directly, but rely on linkage disequilibrium (LD, measured by \( r \)) between observed markers and unknown QTL (Lewontin and Kojima 1960). For the additive effect at the QTL, the fraction captured by the marker is proportional to \( r \), whereas for the dominance effect, the fraction captured by the marker is proportional to \( r^2 \) (Weir 2008; Zhu et al. 2015). The proportion of the signal of the dominance part of \( \sigma_\alpha^2 \), that is captured is therefore expected to be smaller than of the additive part, because \( r^2 \leq r \). For this reason, a marker should be very close to a QTL to pick up its dominance effect (Wellmann and Bennetts 2012). As a result, the benefit of dominance models over additive models may be smaller with lower marker densities. We therefore argue that, when dominance is present and markers are able to capture dominance, the dominance model yields more accurate estimates of \( \alpha \) than the additive model.

CONCLUSIONS

When a single locus average effect is estimated in a random finite sample from a large population in HWE, both A-models and AD-models yield error in their estimates, even when the locus explains all variance (i.e., \( H^2 = 1 \)). Estimates from the AD-model, however, are more robust against chance deviations from HWE frequencies than estimates from the A-model. Genetic models that include dominance, therefore, yield higher accuracies of estimated average effects at single loci than purely additive models when dominance is present. In the absence of dominance, there was no penalty for fitting dominance. These results are important for GWAS, and potentially also for genomic prediction.

ACKNOWLEDGMENTS

The authors thank Chris Maliepaard for his suggestions on an earlier draft of the paper. This research is supported by the Netherlands Organization of Scientific Research (NWO) and the Breed4Food consortium partners Cobb Europe, CRV, Hendrix Genetics, and Topigs Norsvin. The authors declare that they have no competing interests.

Author contributions: P.D. performed the analysis and wrote the manuscript. M.P.L.C., Y.C.J.W., and P.B. supervised the study and assisted with the interpretation of results and writing of the manuscript. All authors read and approved the final manuscript.

LITERATURE CITED

Alilioo, H., J. E. Pryce, O. González-Recio, B. G. Cocks, and B. J. Hayes, 2015 Validation of markers with non-additive effects on milk yield and fertility in Holstein and Jersey cows. BMC Genet. 16: 89.

Bennewitz, J., C. Edel, R. Fries, T. H. E. Meuwissen, and R. Wellmann, 2017 Application of a Bayesian dominance model improves power in quantitative trait genome-wide association analysis. Genet. Sel. Evol. 49: 7.

Cleveland, M. A., J. M. Hickey, and S. Forni, 2012 A common dataset for genomic analysis of livestock populations. G3 2: 429–435.

de los Campos, G., D. Sorensen, and D. Gianola, 2015 Genomic heritability: what is it? PLoS Genet. 11: e1005048.

Falconer, D., and T. Mackay, 1996 Introduction to Quantitative Genetics. Pearson Education Limited, Harlow.

Falconer, D. S., 1985 A note on Fisher’s ‘average effect’ and ‘average excess’. Genet. Res. 46: 337–347.

Fisher, R. A., 1941 Average excess and average effect of a gene substitution. Ann. Eugen. 11: 53–63.

Goddard, M., 2009 Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136: 245–257.

Gondro, C., J. van der Werf, and B. Hayes, 2013 Genome-Wide Association Studies and Genomic Prediction. Humana Press, New York.

Haldane, J. B. S., 1954 An exact test for randomness of mating. J. Genet. 52: 631–635.

Hall, S. J. G., 2016 Effective population sizes in cattle, sheep, horses, pigs and goats estimated from census and herdbook data. Animal 10: 1778–1785.

Hayes, B., 2013 Overview of statistical methods for genome-wide association studies (GWAS), pp. 149–169 in Genome-Wide Association Studies and Genomic Prediction, edited by Gondro, C., J. van der Werf, and B. Hayes. Humana Press, New York.

Huang, X., S. Yang, J. Gong, Y. Zhao, Q. Feng et al., 2015 Genomic analysis of hybrid rice varieties reveals numerous superior alleles that contribute to heterosis. Nat. Commun. 6: 6258.

Huang, X., S. Yang, J. Gong, Q. Zhao, Q. Feng et al., 2016 Genomic architecture of heterosis for yield traits in rice. Nature 537: 629–633.

Kimura, M., and J. F. Crow, 1963 The measurement of effective population number. Evolution 17: 279–288.

Lewontin, R. C., and K. Kojima, 1960 The evolutionary dynamics of complex polymorphisms. Evolution 14: 458–472.

Lopes, M. S., J. W. M. Bastiaanssen, B. Harlizius, E. F. Knol, and H. Bovenhuis, 2014 A genome-wide association study reveals dominance effects on number of teats in pigs. PLoS One 9: e105867.

Lynch, M., and B. Walsh, 1998 Genetics and Analysis of Quantitative Traits. Sinauer, Sunderland, MA.

Mackay, T. F. C., 2015 Epistasis for quantitative traits in Drosophila, pp. 47–70 in Epistasis: Methods and Protocols, edited by Moore, J. H., and S. M. Williams. Springer, New York.

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

Mottahed, P. J., and J. K. Kelly, 2015 Epistasis is a major determinant of the additive genetic variance in Musimulus gattatus. PLoS Genet. 11: e1005201.

Morris, J. A., and F. E. Binet, 1966 Heterosis and selection in the domestic fowl. Aust. J. Biol. Sci. 19: 575–585.

Ott, L., and M. Longnecker, 2010 An Introduction to Statistical Methods and Data Analysis. Brooks/Cole, Cengage Learning, Boston.

Sellier, P., 1976 The basis of crossbreeding in pigs; a review. Livest. Prod. Sci. 3: 203–226.

Stuber, C. W., 2010 Heterosis in plant breeding, pp. 227–251 in Plant Breeding Reviews, edited by Janick, J.. John Wiley & Sons, Inc, New York.

Visscher, P., R. Pong-Wong, C. Whittemore, and C. Haley, 2000 Impact of biotechnology on (cross)breeding programmes in pigs. Livest. Prod. Sci. 65: 57–70.

Vitezica, Z. G., L. Varona, and A. Legarra, 2013 On the additive and dominant variance and covariance of individuals within the genomic selection scope. Genetics 195: 1223–1230.

Wang, L., A. Caballero, and W. G. Hill, 1998 The effect of linkage disequilibrium and deviation from Hardy-Weinberg proportions on the changes in genetic variance with bottlenecking. Heredity 81: 174–186.

Weir, B. S., 2008 Linkage disequilibrium and association mapping. Annu. Rev. Genomics Hum. Genet. 9: 129–142.

Wellmann, R., and J. Bennetts, 2012 Bayesian models with dominance effects for genomic evaluation of quantitative traits. Genet. Res. 94: 21–37.

Whittaker, J. C., R. Thompson, and M. C. Denham, 2000 Marker-assisted selection using ridge regression. Genet. Res. 75: 249–252.

Wright, S., 1931 Evolution in Mendelian populations. Genetics 16: 97–159.

Wright, S., 1969 The Theory of Gene Frequencies. The University of Chicago Press, Chicago.

Xiao, J., J. Li, L. Yuan, and S. D. Tanksley, 1995 Dominance is the major genetic basis of heterosis in rice as revealed by QTL analysis using molecular markers. Genetics 140: 745–754.

Zhu, Z., A. Bakshi, A. E. Vinkhuizen, G. Hemani, S. S. H. Lee et al., 2015 Dominance genetic variation contributes little to the missing heritability for human complex traits. Am. J. Hum. Genet. 96: 377–385.

Communicating editor: D.-J. de Koning
APPENDIX A

With the A-model, error originated mainly from sampling deviations of genotype frequencies from expected HWE frequencies (ΔHWE), and to a lesser extent from sampling deviations of allele frequencies (Δp). For all N, the RMSE was smallest at intermediate allele frequencies, and increased when allele frequency moved away from 0.5 (Figure 1). This increase was due to the increasing probability of sampling fewer individuals from the rare genotype class than expected based on HWE. The RMSE decreased again at extreme allele frequencies. This decrease was due to (1) the increasing probability of not sampling the rare homozygous genotype, and (2) the change in α. When one of the homozygous genotypes is not sampled, α is always equal to the slope of the regression line between the opposing homozygotes and the heterozygotes (a + d when n₂ = 0, or a − d when n_0 = 0). As the allele frequency approaches fixation, the probability of missing a homozygous genotype increases, while at the same time, the true α becomes close to the slope of the resulting regression line (see Equation 1) (a + d when p → 0, or a − d when p → 1). Hence, the RMSE as a result of missing one genotype class was small at extreme allele frequency and increased as the allele frequency increased. It is good to note that with only one genotype class sampled, the sample was disregarded because aα could not be estimated.

With the AD-model, error originated from Δp only, provided that all three genotype classes were sampled. For all N, the RMSE with the AD-model was always smaller than with the A-model, and showed a different pattern across allele frequencies (Figure 1). The RMSE decreased slightly when allele frequency moved away from 0.5, because the probability of drawing a sample with a very high Δp decreased. RMSE increased again for allele frequencies ~0.1 or 0.9. This increase was due to the increased probability of sampling no individuals from the rare genotype class, in which case the AD-model reduced to the A-model. The RMSE decreased again at even more extreme allele frequencies, for the same reasons as explained for the A-model.

APPENDIX B

Our aim is to estimate α in the entire population, but with the A-model, the estimate we get from the sample is equal to α_s. So, we can predict what the error will be by deriving the difference between α_s and α.

\[
\text{error} = E(\hat{\alpha}_s) - \alpha = \alpha_s - \alpha = \left[ a + (1 - 2p_s)d \left( \frac{1 - F_s}{1 + F_s} \right) \right] - [a + (1 - 2p)d] = (1 - 2p_s)d \left( \frac{1 - F_s}{1 + F_s} \right) - (1 - 2p)d = \left( \frac{1 - F_s}{1 + F_s} \right) - \left( \frac{1 - F_s}{1 + F_s} \right) (1 - 2p_s)d = (1 - 2p_s)d.
\]

Then, if there is no deviation from HWE (F = 0), the error due to deviations of p_s from p is

\[
\left[ (1 - 2p_s) - (1 - 2p) \right]d = 2(p - p_s)d.
\]

Similarly, if p_s = p, the error due to deviations from HWE is equal to

\[
\left( \frac{1 - F_s}{1 + F_s} - 1 \right) (1 - 2p_s)d.
\]

When we correct for deviations from HWE, assuming that we know d, we get

\[
\hat{\alpha}_c = \alpha_s - \text{error} = \left[ a + (1 - 2p_s)d \left( \frac{1 - F_s}{1 + F_s} \right) \right] - \left[ \left( \frac{1 - F_s}{1 + F_s} \right) - 1 \right] (1 - 2p_s)d = a + (1 - 2p_s)d - (1 - 2p_s)d = a + (1 - 2p_s)d.
\]

If we subsequently correct for deviations of p, assuming that we know p, we get

\[
\left[ a + (1 - 2p_s)d \right] - 2(p - p_s)d = \left[ a + (1 - 2p_s)d \right] - [(1 - 2p_s) - (1 - 2p)]d = a + (1 - 2p)d,
\]

which is the average effect in the population.

For the AD-model, we can predict error for samples that have all three genotype classes by

\[
\text{error} = E(\hat{\alpha}_{AD}) - \alpha = \left[ a + (1 - 2p_s)d \right] - [a + (1 - 2p)d] = \left( 1 - 2p_s \right) - (1 - 2p) = 2(p - p_s)d,
\]

which is equal to the error due to deviations of p_s from p with the A-model.
Appendix C Expected SD of $F_s$ for several sample sizes, compared with the SD of $F_s$ measured in an empirical dataset of pigs.