RESEARCH ARTICLE

Average semivariance yields accurate estimates of the fraction of marker-associated genetic variance and heritability in complex trait analyses

Mitchell J. Feldmann1, Hans-Peter Piepho2, William C. Bridges3, Steven J. Knapp1*

1 Department of Plant Sciences, University of California, Davis, California, United States of America, 2 Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Stuttgart, Germany, 3 Department of Mathematical Sciences, Clemson University, Clemson, South Carolina, United States of America

* sjknapp@ucdavis.edu

Abstract

The development of genome-informed methods for identifying quantitative trait loci (QTL) and studying the genetic basis of quantitative variation in natural and experimental populations has been driven by advances in high-throughput genotyping. For many complex traits, the underlying genetic variation is caused by the segregation of one or more ‘large-effect’ loci, in addition to an unknown number of loci with effects below the threshold of statistical detection. The large-effect loci segregating in populations are often necessary but not sufficient for predicting quantitative phenotypes. They are, nevertheless, important enough to warrant deeper study and direct modelling in genomic prediction problems. We explored the accuracy of statistical methods for estimating the fraction of marker-associated genetic variance ($p$) and heritability ($H^2_M$) for large-effect loci underlying complex phenotypes. We found that commonly used statistical methods overestimate $p$ and $H^2_M$. The source of the upward bias was traced to inequalities between the expected values of variance components in the numerators and denominators of these parameters. Algebraic solutions for bias-correcting estimates of $p$ and $H^2_M$ were found that only depend on the degrees of freedom and are constant for a given study design. We discovered that average semivariance methods, which have heretofore not been used in complex trait analyses, yielded unbiased estimates of $p$ and $H^2_M$. The cryptic bias problem described here is unrelated to selection bias, although both cause the overestimation of $p$ and $H^2_M$. The solutions we described are predicted to more accurately describe the contributions of large-effect loci to the genetic variation underlying complex traits of medical, biological, and agricultural importance.

Author summary

The contributions of individual genes to the phenotypic variation observed for genetically complex traits has been an ongoing and important challenge in biology, medicine, and...
agriculture. While many genes have statistically undetectable effects, those with large effects often warrant in-depth study and can be important predictors of complex phenotypes such as disease risk in humans or disease resistance in domesticated plants and animals. The genes identified through associations with genetic markers in complex trait analyses typically account for a fraction of the heritable variation, a genetic parameter we called ‘marker heritability’. We discovered that textbook statistical methods systematically overestimate marker heritability and thus overestimate the contributions of specific genes to the phenotypic variation observed for complex traits in natural and experimental populations. We describe the source of the upward bias, validate our findings through computer simulation, describe methods for bias-correcting estimates of marker heritability, and illustrate their application through empirical examples. The statistical methods we describe supply investigators with more accurate estimates of the contributions of specific genes or networks of interacting genes to the heritable variation observed in complex trait studies.

Introduction

The genetic variation observed in nature is frequently caused by genes with quantitative effects [1–7]. Their discovery and characterization has been a dominant feature of quantitative genetic studies in biology, evolution, agriculture, and medicine since the introduction of methods for genotyping DNA variants genome-wide [8–11], and the parallel development of statistical methods for finding associations between DNA variants and the underlying genes or quantitative trait loci (QTL) [2, 4, 5, 7, 12–16]. A significant breakthrough was achieved when Lander and Botstein [12] introduced ‘interval mapping’ and showed that genomes could be systematically searched to identify QTL in populations genotyped with a genome-wide framework of genetically mapped DNA markers. As genotyping technologies advanced and marker densities increased, genome-wide association study (GWAS) methods emerged to search genomes for genotype-to-phenotype associations by exploiting the historical recombination in populations [14, 15, 17–19]. The concept of genomic prediction emerged as a counterpart to GWAS, initially for estimating genomic-estimated breeding values (GEBVs) in domesticated plants and animals and later for estimating polygenic risk scores (PRSs) in humans and model organisms [20–23]. These technical advances precipitated a consequential shift in the study of quantitative traits from analyses of phenotypic variation limited and informed by pedigree or family data to genome-wide analyses of genotype-to-phenotype associations and genomic prediction informed by genotypic data [6, 7, 13, 16, 20, 24–31].

The phenotypic variation observed in a population is customarily partitioned into genetic and non-genetic components to estimate heritability, repeatability, and reliability of the quantitative traits under study [24, 25, 30, 32]. The genetic component can be caused by any number of genes with quantitative effects, even a single gene, but more often by multiple genes with a range of effects [31, 33–43]. For most quantitative traits, that number is unknown but presumed to be large and undiscoverable [3, 6, 7, 22, 32, 34]. Because genes with small effects are challenging to identify and validate, the ‘many genes with small effects’ hypothesis has been difficult to conclusively falsify [21, 22, 32]. Despite the uncertainty surrounding the identity, number, effects, and interactions of genes in the undiscovered fraction [6], three decades of complex trait analyses in humans, domesticated plants and animals, Drosophila,
Arabidopsis, yeast, mice, zebrafish, and other organisms have shown that the ‘discovered’ genes are typically small in number, large in effect, and collectively only explain a fraction of the genetic variance ($\sigma^2_G$) [13, 16, 28, 32, 34–36, 44, 45]. The unexplained fraction has been called ‘missing heritability’ [46–48].

The discovered genes in polygenic systems of genes are often necessary but not sufficient for predicting quantitative phenotypes, e.g., disease risk in humans or yield in domesticated plants and animals [3, 21, 34, 42, 44, 49]. There is a large body of evidence that the QTL effects for many quantitative traits are gamma family distributed, where the discovered genes are found in the upper or thin tail of the distribution above the threshold of statistical significance [34]. The presumption is that the lower or heavy tail of the gamma family distribution is caused by many genes with small effects, the chief tenet of the infinitesimal model of quantitative genetics [6, 26, 32, 50]. Genes with large effects often dominate the ‘non-missing heritability’, mask or obscure the effects of other quantitatively acting genes, and pleiotropically affect multiple quantitative phenotypes [16, 35, 39, 51], e.g., mutations in the BRCA2 gene can have large effects, are incompletely penetrant, interact with other genes, and are necessary but not sufficient for predicting breast, ovarian, and other cancer risks in women [52].

The large-effect QTL BTA19 pleiotropically affects milk yield, protein yield, and productive life in Guernsey cattle (Bos taurus) [43], and branching and pigment genes (BR, PHY, and HYP) have large effects, interact, and pleiotropically affect several genetically correlated seed biomass traits in sunflower (Helianthus annuus) [53]. Despite decades of directional selection, loci with large effects often segment (have not been fixed) in domesticated plant and animal populations [33, 34, 37, 38, 40, 54, 55]. The fractions of the genetic variances explained by BRCA2, BTA19, BR, PHY, and HYP were not reported in those studies. What fraction of the heritability for breast cancer risk, for example, can be explained by the known mutations in BRCA2? Our study explored the accuracy of methods for estimating that parameter.

Our surveys and others substantiate that the missing and non-missing fractions of the genetic variance are commonly either not estimated or inaccurately estimated in GWAS and other gene finding studies, e.g., the statistical significance of individual marker loci from sequential regression analyses are typically reported without correcting for the effects of other discovered marker loci through multilocus partial regression analyses or Type III ANOVA [17, 19, 22, 34, 56]. Such analyses are necessary for accurately assessing the statistical importance of the underlying gene and gene-gene interaction effects in a multilocus system, e.g., when multiple loci are identified by GWAS (sequential analyses of individual loci), their effects are more accurately estimated by simultaneous analysis using partial regression analysis approaches and even then can be upwardly biased [51]. The estimation problem we studied is intertwined with the broader problem of accurately describing multilocus systems of genes with large effects. We show that the discovered fraction of the genetic variance can be grossly overestimated and that the cause of the problem is a mathematical artifact in the expected values of variance components and their ratios. We revisited the problem of estimating the non-missing and missing fractions of heritability in candidate gene and other complex trait analyses, in part because of the systematic upward bias we discovered, in addition to inconsistencies in the methods commonly applied to the problem. The solutions to the problem presented here are straightforward and primarily applicable to the study of genes with large effects, especially those affecting the accuracy of genomic predictions for disease risk or breeding value [21, 43]. The optimum approaches for weighting or correcting for loci with large effects in genomic prediction are not completely clear; however, in artificial selection settings where the favorable alleles for discovered loci are unequivocally known, those alleles can be directly
selected via marker-assisted selection (MAS) with genomic selection exerting pressure on unknown loci underlying the additive genetic variance not explained by the segregation of known large effect loci [54, 57–61].

Lande and Thompson [62] proposed the parameter $p = \frac{\sigma^2_M}{\sigma^2_G}$ to estimate the discovered or non-missing fraction of the genetic variance, where $\sigma^2_M$ is the fraction of the genetic variance associated with statistically significant markers in linkage disequilibrium (LD) with genes or QTL affecting the trait under study (here QTL refers to a chromosome segment predicted to harbor a gene or genes affecting a quantitative trait). Similarly, marker heritability ($H^2_M = \frac{\sigma^2_M}{\sigma^2_P}$) estimates the non-missing fraction of the phenotypic variance ($\sigma^2_P$) associated with statistically significant markers in LD with causal genes or QTL. Here a distinction needs to be made between $H^2_M$ and genomic heritability, a parameter estimated by summing the effects of a dense genome-wide sample of markers, only some of which are predicted to be in LD with the underlying causal genes or QTL [27, 30, 63]. We are not proposing marker heritability as a replacement or substitute for genomic heritability but as a parameter for parsing out the non-missing fraction of heritability associated with discovered loci, especially loci like BRCA2 and BTA19 [43, 52]. The genetic variance component ($\sigma^2_G$) in these ratios can be estimated from pedigree or family information (as shown in our examples) or genomic information (as reviewed by [30] and [63]). For either, $\sigma^2_M$ is simply the variance explained by marker loci with effects large enough to be statistically detected and important enough to be specifically studied and modeled, perhaps as fixed effects [22, 39, 40, 51, 61]. Despite a direct and logical connection to heritability, estimates of $p$ and $H^2_M$ are seldom reported in complex trait studies, whereas genomic heritability estimates are commonly reported in genomic prediction studies [30, 34, 62].

Here we show that $p$ and $H^2_M$ are often overestimated in complex trait analyses. The problem we discovered is unrelated to selection bias, the phenomena where the effects of discovered QTL are inflated by biased sampling from truncated distributions with small sample sizes [64–69], and unrelated to the upward biases known to arise in GWAS [70]. While selection bias is a well known and widely cited problem in complex trait analyses, we describe a previously unreported and cryptic source of bias in estimates of $p$ and $H^2_M$. To identify the source of the bias and explore the problem in greater depth, we compared the accuracy of average marginal variance (AMV) [71, 72] and average semivariance (ASV) [73] methods for estimating $p$ and $H^2_M$. AMV is the acronym applied throughout this paper for the ANOVA and REML variance component estimation methods commonly described in textbooks and implemented in statistical software for the analysis of generalized linear mixed models (GLMMs), e.g., the ‘lme4’ R package and the SAS packages ‘GLM’ and ‘GLIMMIX’ [24, 25, 72, 74–79]. We introduced the average marginal variance terminology here to facilitate comparisons of the differences between AMV and ASV methods for estimating variance component ratios. The ASV methods we applied to the problem are extensions of those described by Piepho [73] for estimating the total variance and coefficient of determination ($R^2$) in GLMM analyses. For the AMV and ASV analyses shown throughout this paper, REML was used to estimate the variance components [56, 72, 75, 79]. The source of the bias was discovered, however, through algebraic analyses of the expected mean squares (EMSs) from ANOVA. We describe that source and approaches for bias-correcting ANOVA or REML estimates of $p$ and $H^2_M$ from the commonly applied AMV methods. We show that ASV methods directly yield unbiased estimates of $p$ and $H^2_M$ that are identical to bias-corrected AMV estimates. Finally, we discuss the connection of these random effects methods to the fixed effect methods commonly applied in QTL mapping and genome-wide association studies [51, 80, 81].
Results and discussion

Overestimation of the genetic variance explained by markers in linkage disequilibrium with causative genes or QTL

The overestimation problem described here was originally discovered in a reanalysis of data from genetic studies in plants where REML estimates of $H^2_M$ exceeded REML estimates of broad-sense heritability ($H^2$) and REML estimates of $p$ and $H^2_M$ exceeded 1.0, the theoretical upper limit for these parameters (Table 1). We initially suspected that selection bias might be the culprit [68, 69, 82–84] but concluded that selection bias alone could not explain $\hat{p} > 1.0$ or $H^2_M > 1.0$. Although proof was lacking and the bias was non-obvious, we hypothesized that many estimates in the theoretical range ($0.0 \leq p \leq 1.0$) must also be upwardly biased. The proof was found through algebraic analyses of the ANOVA estimators of $\sigma^2_M$, $\sigma^2_G$, and $\sigma^2_p$ for balanced and unbalanced data (S1, S2 and S4 Texts). Although variance components are commonly estimated using REML, as was done in the analyses shown throughout this paper, algebraic analyses of ANOVA expected mean squares (EMSs) identified the source of the bias and yielded explicit algebraic solutions for bias correcting ANOVA and REML estimates of $p$ and $H^2_M$.

The source of the bias was identified by expressing the estimator of $\sigma^2_M$ as a function of the ANOVA estimators of $\sigma^2_M$, $\sigma^2_G$, $\sigma^2_p$, and $H^2_M$. Consider a study where $n_G$ entries are phenotyped for a normally distributed quantitative trait using a balanced completely randomized study design with $r_G$ replications/entry, $n_M$ marker genotypes/locus, and $r_M$ replications/marker genotype. The LMM needed for estimating $\sigma^2_M$ (the between entry variance component) is:

$$y_{ijk} = \mu + G_i + \epsilon_{ijk}$$

where $y_{ijk}$ is the $ijk$th phenotypic observation, $\mu$ is the population mean, $G_i$ is the random effect of the $i$th entry, $\epsilon_{ijk}$ is the random effect of the $ijk$th residual, $G_i \sim N(0, \sigma^2_G)$, $\epsilon_{ijk} \sim N(0, \sigma^2)$, $i = 1, 2, \ldots, n_G$, and $k = 1, 2, \ldots, r_G$. Suppose entries are genotyped for a single marker locus $(M)$ in linkage disequilibrium with a gene or QTL affecting the quantitative phenotype ($y_{ijk}$). The between entry source of variation from LMM (1) can be partitioned into marker ($M$) and entry nested in marker ($G : M$) sources of variation (this is the residual genetic variation among entries not explained by markers in the model). The LMM for estimating $\sigma^2_M$ and $\sigma^2_{G,M}$ is:

$$y_{ijk} = \mu + M_i + G : M_{ij} + \epsilon_{ijk}$$

where $y_{ijk}$ is the $ijk$th phenotypic observation, $M_i$ is the random effect of the $i$th marker genotype at locus $M$, $G : M_{ij}$ is the random effect of the $i$th entry nested in the $j$th marker genotype, $\epsilon_{ijk}$ is the random effect of the $ijk$th residual, $i = 1, 2, 3, M_i \sim N(0, \sigma^2_M)$, $G : M_{ij} \sim N(0, \sigma^2_{G,M})$, and $\epsilon_{ijk} \sim N(0, \sigma^2)$.

The ANOVA estimator of the between-entry variance component ($\hat{\sigma}^2_G$) from LMM (1) with balanced data is:

$$\hat{\sigma}^2_G = \frac{MS_G - MS_s}{r_G} = \frac{SS_G/df_G - SS_s/df_s}{r_G} = \frac{1}{df_G r_G} SS_G - \frac{1}{df_s r_G} SS_s$$

(3)
Table 1. REML estimates of marker-associated variance ($\sigma^2_m$), the fraction of the genetic variance explained by markers ($p = \sigma^2_m/\sigma^2_g$), and marker heritability ($H^2_m = \sigma^2_m/\sigma^2_g$) from random marker effects analyses and coefficients of determination ($R^2$) from Type II and Type III fixed marker effects analyses for large effect loci identified in cattle, sunflower, and strawberry studies.

<table>
<thead>
<tr>
<th>Study</th>
<th>Source</th>
<th>df</th>
<th>Variance Component</th>
<th>Uncorrected $\sigma^2$</th>
<th>Bias Corrected $\sigma^2$</th>
<th>Type II $R^2$</th>
<th>Type III $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cattle White Spots$^a$</td>
<td>M</td>
<td>25</td>
<td>$\sigma^2_{G,M} + \ldots + \sigma^2_{G,M+20}$</td>
<td>7.92</td>
<td>0.76</td>
<td>3.88</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>rs10</td>
<td>2</td>
<td>$\sigma^2_{G,rs10}$</td>
<td>0.62</td>
<td>0.06</td>
<td>0.21</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>rs45</td>
<td>2</td>
<td>$\sigma^2_{G,rs45}$</td>
<td>2.91</td>
<td>0.28</td>
<td>1.20</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>rs20</td>
<td>2</td>
<td>$\sigma^2_{G,rs20}$</td>
<td>3.81</td>
<td>0.37</td>
<td>2.04</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>rs10×rs45</td>
<td>4</td>
<td>$\sigma^2_{G,rs10,rs45}$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>rs10×rs20</td>
<td>4</td>
<td>$\sigma^2_{G,rs10,rs20}$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>rs45×rs20</td>
<td>4</td>
<td>$\sigma^2_{G,rs45,rs20}$</td>
<td>0.37</td>
<td>0.04</td>
<td>0.26</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>rs10×rs45×rs20</td>
<td>7</td>
<td>$\sigma^2_{G,rs10,rs45,rs20}$</td>
<td>0.22</td>
<td>0.02</td>
<td>0.17</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>G : rs10×rs45×rs20</td>
<td>2.935</td>
<td>$\sigma^2_{G,rs10,rs45,rs20}$</td>
<td>5.26</td>
<td>—</td>
<td>5.26</td>
<td>—</td>
</tr>
<tr>
<td>Sunflower Oil Content$^b$</td>
<td>Entry (G)</td>
<td>145</td>
<td>$\sigma^2_G$</td>
<td>21.61</td>
<td>0.95</td>
<td>21.61</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>M + G : M</td>
<td>145</td>
<td>$\sigma^2_G$</td>
<td>30.76</td>
<td>1.42</td>
<td>30.76</td>
<td>1.42</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>7</td>
<td>$\sigma^2_G$</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>BR</td>
<td>1</td>
<td>$\sigma^2_{PB}$</td>
<td>1.57</td>
<td>0.06</td>
<td>1.57</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>PHY</td>
<td>1</td>
<td>$\sigma^2_{PB}$</td>
<td>2.9</td>
<td>0.13</td>
<td>2.9</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>HYP</td>
<td>1</td>
<td>$\sigma^2_{PB}$</td>
<td>0.21</td>
<td>0.01</td>
<td>0.21</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>BR × PHY</td>
<td>1</td>
<td>$\sigma^2_{PB}$</td>
<td>1.89</td>
<td>0.09</td>
<td>1.89</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>PHY × HYP</td>
<td>1</td>
<td>$\sigma^2_{PB}$</td>
<td>0.21</td>
<td>0.01</td>
<td>0.21</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>BR × PHY × HYP</td>
<td>1</td>
<td>$\sigma^2_{PB}$</td>
<td>0.21</td>
<td>0.01</td>
<td>0.21</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>G : BR × PHY × HYP</td>
<td>138</td>
<td>$\sigma^2_{PB}$</td>
<td>12.91</td>
<td>0.60</td>
<td>12.91</td>
<td>0.60</td>
</tr>
<tr>
<td>Strawberry Fusarium Wilt$^c$</td>
<td>Entry (G)</td>
<td>557</td>
<td>$\sigma^2_G$</td>
<td>2.07</td>
<td>—</td>
<td>2.07</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>M + G : M</td>
<td>557</td>
<td>$\sigma^2_G$</td>
<td>4.77</td>
<td>1.46</td>
<td>4.77</td>
<td>1.46</td>
</tr>
<tr>
<td></td>
<td>AX396 (M)</td>
<td>2</td>
<td>$\sigma^2_G$</td>
<td>4.84</td>
<td>1.37</td>
<td>4.84</td>
<td>1.37</td>
</tr>
<tr>
<td></td>
<td>G : AX396</td>
<td>555</td>
<td>$\sigma^2_G$</td>
<td>0.30</td>
<td>0.09</td>
<td>0.30</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>Residual (e)</td>
<td>1,631</td>
<td>$\sigma^2_e$</td>
<td>0.23</td>
<td>—</td>
<td>0.23</td>
<td>—</td>
</tr>
<tr>
<td>Strawberry Fusarium Wilt$^c$</td>
<td>Entry (G)</td>
<td>540</td>
<td>$\sigma^2_G$</td>
<td>3.30</td>
<td>0.98</td>
<td>3.30</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>M + G : M</td>
<td>540</td>
<td>$\sigma^2_G$</td>
<td>4.01</td>
<td>1.21</td>
<td>4.01</td>
<td>1.21</td>
</tr>
<tr>
<td></td>
<td>AX393 (M)</td>
<td>2</td>
<td>$\sigma^2_G$</td>
<td>1.48</td>
<td>0.45</td>
<td>1.48</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>G : AX393</td>
<td>538</td>
<td>$\sigma^2_G$</td>
<td>2.53</td>
<td>0.77</td>
<td>2.53</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>Residual (e)</td>
<td>1,584</td>
<td>$\sigma^2_e$</td>
<td>0.23</td>
<td>—</td>
<td>0.23</td>
<td>—</td>
</tr>
</tbody>
</table>

$^a$Statistics are shown for three marker loci (rs10, rs45, and rs20) associated with genetic variation for white spotting (%) in a cattle population ($n_G=2,973$) with a single phenotypic observation per individual and highly unbalanced marker data [85]. The marker loci were identified by GWAS. The linear mixed model for the cattle analysis was identical to that for the sunflower analysis without replications ($n_G=1$). $\theta_M$ coefficient equations for three loci with unbalanced data are shown in S3 Text.

$^b$Statistics are shown for three marker loci (BR, PHY, and HYP) associated with genetic variation for seed oil content (%) in a sunflower recombinant inbred line (RIL) population ($n_G=146$) with nearly balanced marker data and multiple phenotypic observations (replications) per RIL [53]. The marker loci were identified by QTL mapping. Variance components were estimated from LMM (27) for the AMV method and LMM (S13) for the ASV method.

$^c$Statistics are shown for two SNP markers (AX396 and AX493) associated with genetic variation for resistance to Fusarium wilt in a strawberry population ($n_G=565$) with unbalanced SNP marker data and multiple phenotypic observations per individual [86]. AX396 and AX493 are tightly linked and both were in LD with a dominant gene (Fw1) conferring resistance to Fusarium wilt but had significantly different genotypic ratios among individuals in the population. Variance components were estimated from LMM (2) for the AMV method and LMM (15) for the ASV method. The $k_{G,rs10}$ coefficient a single locus with unbalanced data are shown in S1 Text.

$^d$Type II $R^2$ is the coefficient of partial determination estimated from a Type II ANOVA, where the main and interactions effects of markers are fixed. For the cattle example, the reduction in sums of squares for main effects were estimated with the other main effects in the genetic model without interactions, e.g., the reduction in SS for rs10 was $R(rs10)$.

$^e$Type III $R^2$ is the coefficient of partial determination estimated from a Type III ANOVA, where the main and interactions effects of markers are fixed, e.g., the reduction in sums of squares for rs10 in the cattle example was estimated by fitting rs10 with all other factors in the model: $R(rs10)$.

https://doi.org/10.1371/journal.pgen.1009762.t001
where $MS_G = SS_G / df_G$ is the between entry mean square, $SS_G$ is the between entry sum of squares, $df_G = n_G - 1$ is the between entry degrees of freedom, $MS_i = SS_i / df_i = \sigma_i^2$ is the residual mean square, $SS_i$ is the residual sum of squares, $df_i = n_i(df_G - 1) - 1$ is the residual degrees of freedom, $\sigma_i^2$ is the residual variance component, and $r_G$ is the number of replications per entry [74]. The between-entry variance component has a theoretical genetic interpretation when entries are progeny with genetic relationships known from pedigrees, e.g., monozygotic twins, full-sib families, or recombinant inbred lines [24, 25, 30]. ANOVA estimators of the marker locus $G$ and entry nested in $M$ variance components from LMM (2) with balanced data are:

$$\hat{\sigma}_M^2 = \frac{MS_M - MS_{GM}}{r_G n_{GM}}$$

(4)

and

$$\hat{\sigma}_{GM}^2 = \frac{MS_{GM} - MS}{r_G}$$

(5)

respectively, where $n_{GM}$ is the number of entries nested in each marker genotype, $E(\hat{\sigma}_M^2) = \sigma_M^2$, $E(\hat{\sigma}_{GM}^2) = \sigma_{GM}^2$, $MS_{GM}$ is the entry nested in $M$ mean square, and $MS_M$ is the mean square for marker locus $M$. The residuals in LMMs (1) and (2) are identical when the data are balanced ($\hat{\sigma}_G^2 = MS_G$). Hence, for a single marker locus with balanced data, the ANOVA estimator of $p$ is:

$$\hat{p} = \frac{\hat{\sigma}_G^2}{\hat{\sigma}_G^2 + \hat{\sigma}_M^2}$$

(6)

and the ANOVA estimator of broad-sense marker heritability on an entry-mean basis is:

$$H_M^2 = \frac{\hat{\sigma}_G^2}{\hat{\sigma}_G^2 + \hat{\sigma}_G^2}$$

(7)

where $\hat{p} = \hat{\sigma}_G^2 / \hat{\sigma}_G^2 + \hat{\sigma}_M^2$ is the phenotypic variance on an entry-mean basis [25, 76].

The overestimation of $\hat{p}$ and $H_M^2$ was not obvious from inspection of ANOVA estimators (6) and (7). The source of the bias was discovered by substituting $SS_M + SS_{GM}$ for $SS_G$ in the ANOVA estimator of $\sigma_G^2$ from (3) and simplifying:

$$\hat{\sigma}_G^2 = \frac{1}{df_G r_G} (SS_M + SS_{GM}) - \frac{1}{df_G r_G} SS_G$$

$$= \frac{1}{df_G r_G} [df_M (\hat{\sigma}_G^2 + r_M \hat{\sigma}_{GM}^2 + r_M \hat{\sigma}_M^2) + df_{GM} (\hat{\sigma}_G^2 + r_G \hat{\sigma}_{GM,M}^2)] - \frac{1}{df_G r_G} df_G \hat{\sigma}_G^2$$

$$= \frac{df_M r_M \hat{\sigma}_G^2 + \hat{\sigma}_{GM}^2}{df_G r_G} = \frac{df_M n_{GM}}{df_G} \frac{\hat{\sigma}_M^2 + \hat{\sigma}_{GM}^2}{df_G} = k_M \hat{\sigma}_G^2 + \hat{\sigma}_{GM}^2$$

(8)

where the fraction $k_M$ is source of the bias, $0 < k_M < 1$, $r_M$ is the number of replications per marker genotype, $n_{GM}$ is the number of entries nested in marker loci, $SS_M$ is the marker sum of squares, $df_M$ is the marker degrees of freedom, $r_M$ is the number of replicates of each marker genotype, $SS_{GM}$ is the entry nested in marker sum of squares, and $df_{GM}$ is the entry nested in marker degrees of freedom. The term $k_M$ in (8) depends on degrees of freedom and $n_{GM}$ and is hereafter referred to as the $k_M$ bias coefficient, where the subscript $M$ indexes the intralocus and interlocus effects of marker loci.
We started with the special case of balanced data, which seldom arises in practice, but develop shown above for a single locus with balanced data, we found results here for the general case of unbalanced data. Following the same approach as that typic data, we found:

for a single marker locus with balanced genotypic data (S1, S2 and S3 Texts). For a single marker locus with balanced genotypic data.

Eq (8) shows that the sum of ANOVA estimates of $\sigma_M^2$ and $\sigma_{G,M}^2$ from LMM (1) are greater than the ANOVA estimate of $\sigma_G^2$ from LMM (2):

$$\hat{\sigma}_G^2 = \frac{\hat{\sigma}_M^2 + \hat{\sigma}_{G,M}^2}{k_M \hat{\sigma}_M^2 + \hat{\sigma}_{G,M}^2}$$

(9)

Although the SS for sources of variation in these LMMs are additive (SSM + SS_{G,M} = SS_G), the mean squares are not (MSM + MS_{G,M} $\neq$ MS_G). Because $\hat{\sigma}_G^2 = k_M \hat{\sigma}_M^2 + \hat{\sigma}_{G,M}^2$, the sum $\hat{\sigma}_M^2 + \hat{\sigma}_{G,M}^2$ from LMM (2) overestimates $\hat{\sigma}_G^2$ by a factor of $(1 - k_M) \hat{\sigma}_M^2$. The ANOVA estimators of $p$ and $H^2_M$ from analyses of LMMs (1) and (2) are upwardly biased because $\hat{\sigma}_M^2$ is multiplied by the fraction $k_M$ in their denominators, and not the numerators:

$$\hat{p} = \frac{\hat{\sigma}_M^2}{\hat{\sigma}_G^2} = \frac{\hat{\sigma}_M^2}{k_M \hat{\sigma}_M^2 + \hat{\sigma}_{G,M}^2}$$

(10)

and

$$\hat{H}_M^2 = \frac{\hat{\sigma}_M^2}{\hat{\sigma}_G^2} = \frac{\hat{\sigma}_M^2}{k_M \hat{\sigma}_M^2 + \hat{\sigma}_{G,M}^2 + \hat{\sigma}_G^2}$$

(11)

Substituting $\hat{\sigma}_M^2 + \hat{\sigma}_{G,M}^2$ for $\hat{\sigma}_G^2$ in the denominators of $p$ and $H^2_M$ decreases but does not eliminate the bias because $\hat{\sigma}_M^2$ is multiplied by $k_M$ in the denominator (S1 Fig). For a single marker with balanced data, we found that:

$$k_M = \frac{df_{SM}}{df_G} = \frac{df_M n_{G,M}}{df_G}$$

(12)

and

$$\left(\frac{\hat{\sigma}_M^2}{\hat{\sigma}_G^2} = \frac{\hat{\sigma}_M^2}{k_M \hat{\sigma}_M^2 + \hat{\sigma}_{G,M}^2} \right) > \left(\frac{\hat{\sigma}_M^2}{\hat{\sigma}_G^2} = \frac{\hat{\sigma}_M^2}{\hat{\sigma}_M^2 + \hat{\sigma}_{G,M}^2} \right)$$

(13)

where $0 < k_M < 1$. Hence, the bias is caused by the $k_M$ multiplier in the expected values of the ANOVA estimators of $p$ and $H^2_M$. As shown later, simulation analyses confirmed that (9) and (13) accurately predict the upward bias caused by $k_M$. Moreover, we concluded that the bias could be corrected by multiplying ANOVA or REML estimates of $\hat{\sigma}_G^2$ by $k_M$ in the numerators of $p$ and $H^2_M$ estimates.

**Genetic models with unbalanced genotypic data**

We started with the special case of balanced data, which seldom arises in practice, but develop results here for the general case of unbalanced data. Following the same approach as that shown above for a single locus with balanced data, we found $k_M$ coefficients for bias-correcting ANOVA and REML estimates of $p$ and $H^2_M$ for analyses of one to three marker loci with unbalanced genotypic data (S1, S2 and S3 Texts). For a single marker locus with unbalanced genotypic data, we found:

$$k_M = \frac{n_G - n_{G,M,k} \sum n_{G,M,k}}{df_G}$$

(14)

where $n_G$ is the number of entries, $df_G$ are the degrees of freedom for entries, and $n_{G,M,k}$ is the number of entries nested in the $k^{th}$ marker genotype (S1 Text). This simplifies to (12) for a single marker locus with balanced genotypic data.

The $k_M$ coefficients become slightly more complicated as the number of marker loci increases but nevertheless follow a predictable algebraic pattern, e.g., for a two locus genetic
model, see equations (S19)-(S25) in S3 Text. Similarly, for a three locus genetic model, see equations (S19)-(S25) in S3 Text. $k_M$ is greater ($k_M$ bias is proportionally smaller) for interaction than main effects, e.g., for two marker loci, $k_{M1} < k_{M1\times M2} < 1$ and $k_{M2} < k_{M1\times M2} < 1$, where $k_{M1}$ is the coefficient for $M_1$, $k_{M2}$ is the coefficient for $M_2$, and $k_{M1\times M2}$ is the coefficient for the $M_1 \times M_2$ epistatic interaction (S2 Text). $k_M$ for the two-locus interaction ($k_{M1\times M2}$) is larger than $k_M$ for the individual marker loci ($k_{M1}$ and $k_{M2}$) because the denominator ($d_{G1\times G2}$) is constant, whereas the numerators increase and approach the denominator as the degrees of freedom for marker effects increase. Therefore, the upward bias is proportionally smaller for the $M1 \times M2$ variance component than the $M1$ or $M2$ variance components for a two locus genetic model. Similarly, for a three locus genetic model, the upward bias is proportionally smaller for the $M1 \times M2 \times M3$ interaction variance component than the two-way interaction variance components ($M1 \times M2$, $M1 \times M3$, and $M2 \times M3$). These results naturally extend to genetic models with more than three loci. Algebraic results are only shown for three marker loci because we found that the the $k_M$ bias problem can be directly solved using average semivariance estimation methods when analyzing more complex genetic models (see below).

Although certainly not limited to three marker loci, the methods described herein are primarily designed to study the effects of one to a few genes with large effects, e.g., $BRCA2$ [52], BTA19 [43], and the examples shown in Tables 1 and 2, and not to replace GWAS or QTL mapping.

**Study designs without replications or repeated measures of individuals or families**

LMMs (1) and (2) arise in study designs where entries, e.g., individuals, families, or strains, are replicated, e.g., in studies with domesticated plants, biological replicates of half-sib or full-sib families, doubled haploid or recombinant inbred lines, or testcross hybrids are commonly phenotyped [24, 25, 31, 76, 87, 88] (see the sunflower example in Table 1). These same LMMs apply to study designs for monozygotic twins in humans and other mammals and clonally replicated individuals in asexually propagated plants, e.g., cassava ($Manihot esculenta$), strawberry ($Fragaria \times ananassa$), and apple ($Malus \times domestica$) (see the strawberry examples in Table 1). The extension of the proposed $k_M$ bias correction solutions to LMMs with repeated measures is straightforward and should have applications in studies where large effect loci are important determinants of the genetic variation underlying quantitative traits in both replicable and unreplicable organisms or populations [88–94].

When entries are unreplicated, the random error or residual source of variation in LMM (2) disappears ($\sigma_{G,M}^2$ becomes the residual) and $\sigma_V^2$, $\sigma_G^2$, and $p$ cannot be estimated; however, the marker heritability can be estimated using the phenotypic variance among unreplicated individuals ($k_M\sigma_V^2 + \sigma_G^2$). As before, this variance component ratio is upwardly biased by the factor $k_M$ (see the cattle example in Table 1). Without the insights gained from the algebra shown in equations (10), (S3), (S9), and (S18), and S1, S2 and S3 Texts, the bias would not be obvious unless one or more estimates of marker heritability exceeded 1, which only happens when the loci under study have very large effects. That was exactly how we originally discovered the bias problem in the first place (Table 1). The bias is systematic and ubiquitous but not immediately obvious when estimates fall within the expected range ($0 < H^2_G < 1$). The same bias correction solutions we proposed for study designs with replications of entries can be applied in study designs where entries are unreplicated. When unreplicated entries are genotyped with a dense genome-wide of markers, $\sigma_V^2$ be estimated using a genomic or pedigree relationship matrix [92, 95–97], which yields an estimate of $p$. 
Table 2. Type I, II, and III sums of squares for fixed effect analyses of markers associated with QTL identified in GWAS and QTL mapping experiments in cattle and sunflower.

<table>
<thead>
<tr>
<th>Study</th>
<th>Source</th>
<th>Type I SS</th>
<th>Type II SS</th>
<th>Type III SS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ABC</td>
<td>ACB</td>
<td>BAC</td>
</tr>
<tr>
<td>Cattle White Spots</td>
<td>rs10</td>
<td>3,552.3</td>
<td>3,552.3</td>
<td>1,707.2</td>
</tr>
<tr>
<td></td>
<td>rs45</td>
<td>6,339.7</td>
<td>4,259.6</td>
<td>8,384.8</td>
</tr>
<tr>
<td></td>
<td>rs20</td>
<td>4,880.5</td>
<td>7,160.7</td>
<td>4,880.5</td>
</tr>
<tr>
<td></td>
<td>rs10 × rs45</td>
<td>12.7</td>
<td>12.7</td>
<td>12.7</td>
</tr>
<tr>
<td></td>
<td>rs10 × rs20</td>
<td>132.7</td>
<td>132.7</td>
<td>132.7</td>
</tr>
<tr>
<td></td>
<td>rs45 × rs20</td>
<td>193.1</td>
<td>193.1</td>
<td>193.1</td>
</tr>
<tr>
<td></td>
<td>rs10 × rs45 × rs20</td>
<td>143.5</td>
<td>143.5</td>
<td>143.5</td>
</tr>
<tr>
<td></td>
<td>G : rs10 × rs45 × rs20</td>
<td>15,512.9</td>
<td>15,512.9</td>
<td>15,512.9</td>
</tr>
<tr>
<td>Sunflower Oil Content</td>
<td>BR</td>
<td>1,624.0</td>
<td>1,624.0</td>
<td>1,708.8</td>
</tr>
<tr>
<td></td>
<td>PHY</td>
<td>298.2</td>
<td>254.2</td>
<td>213.4</td>
</tr>
<tr>
<td></td>
<td>HYP</td>
<td>537.1</td>
<td>581.0</td>
<td>537.1</td>
</tr>
<tr>
<td></td>
<td>BR × PHY</td>
<td>57.9</td>
<td>57.9</td>
<td>57.9</td>
</tr>
<tr>
<td></td>
<td>BR × HYP</td>
<td>168.0</td>
<td>168.0</td>
<td>168.0</td>
</tr>
<tr>
<td></td>
<td>PHY × HYP</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
</tr>
<tr>
<td></td>
<td>BR × PHY × HYP</td>
<td>36.6</td>
<td>36.6</td>
<td>36.6</td>
</tr>
<tr>
<td></td>
<td>G : BR × PHY × HYP</td>
<td>4,113.4</td>
<td>4,113.4</td>
<td>4,113.4</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>553.8</td>
<td>553.8</td>
<td>553.8</td>
</tr>
</tbody>
</table>

*For each Type I ANOVA, the six possible orders of the three main effects (marker loci A, B, and C) were tested in the genetic model, where A = rs10, B = rs45, and C = rs20 for the cattle example and A = BR, B = PHY, and C = HYP for the sunflower example. The interactions were added to the genetic model in a single sequence: A × B, A × C, B × C, and A × B × C. The three sequences indicate the sequence with which markers loci entered the genetic model, e.g., for the ABC order, the sums of squares for the three main effects were SS(μ), SS(B)A, μ, and SS(C)A, B, μ, where μ is the population mean and factors to the right of the vertical bar were included in the model. Similarly, for the CBA order, the sums of squares for the three main effects were SS(C)μ, SS(B)C, μ, and SS(A)B, C, μ. The sequences with which interactions were added to the genetic model were identical in the six Type I analyses, e.g., the sums of squares for the A × B interaction was SS(A × B)A, B, C, μ and for the three-way interaction was SS(A × B × C)A, B, C, A × B, A × C, B × C, μ.

Statistics are shown for three marker loci (rs10, rs45, and rs20) associated with genetic variation for white spotting (%) in a cattle population (nC = 2, 973) with a single phenotype observation per individual and highly unbalanced marker data [85]. The markers were identified by GWAS. The linear model for the cattle analysis was identical to the linear model for the sunflower analysis without replications (rc = 1); hence, the residual in the cattle analysis was the entry nested in marker source of variation. k_M coefficients for three loci with unbalanced data are shown in S3 Text.

Statistics are shown for three marker loci (BR, PHY, and HYP) associated with genetic variation for seed oil content (%) in a sunflower recombinant inbred line (RIL) population (nC = 146) with nearly balanced marker data and multiple phenotype observations (replications) per RIL [85]. k_M coefficients for three loci with unbalanced data are shown in S3 Text.

https://doi.org/10.1371/journal.pgen.1009762.t002

**Average semivariance estimation directly solves the bias problem**

The AMV methods proposed above for bias correcting ANOVA or REML estimates of p and H_μ^2 are straightforward to apply in practice because they are the methods widely described in textbooks and implemented in popular statistical software packages, e.g., the R package ‘lme4’ and SAS package ‘GLIMMIX’ [78, 98]. Here we show that the bias problem can be directly solved by applying average semivariance (ASV) estimation methods [73]. As before, we start by showing results for a single marker locus with balanced genotypic data. AMV notation and estimators are reformulated in matrix notation here to build the foundation for describing ASV notation and estimators. The input for both are the adjusted entry-level means (μ̃_j) from LMM (1) stored in an n_C-element vector. These are the best linear unbiased estimates (BLUEs) for entries [73, 99]. The LMM equivalent to (2) for the entry-level means analysis of the effect
of a single marker locus ($M$) is:

$$\tilde{y}_{im} = \mu + M_i + G : M_{ij0} + \epsilon_{im}$$

(15)

where $\tilde{y}_{im}$ is the phenotypic mean for the $im$th entry, $\mu$ is the population mean, $M_i$ is the random effect of the $i$th marker genotype, $\var(M_i) = \sigma_M^2$, $G : M_{ij0}$ is the random effect of entries nested in $M$, $\var(G : M_{ij0}) = \sigma_{GM,M}^2$, $\epsilon_{im}$ is the residual error, and $\var(\epsilon_{im}) = \tau_G^2 \sigma_e^2$. The residual variance-covariance matrix ($R$) is estimated in the first stage of a two-stage analysis [99–101]. The between-entry variance can be partitioned into $\sigma_M^2$ and $\sigma_{GM,M}^2$ with individual variance-covariance matrices $G$, defined by the genetic model, e.g., different main and interaction effects among marker loci.

The AMV estimator of the phenotypic (total) variance among observations for LMM (15) is:

$$\hat{\theta}_p^{AMV} = \frac{1}{n_G} tr(V) = \sum_i \hat{\theta}_i^{AMV}$$

(16)

where $V$ is the variance-covariance matrix of the phenotypic observations, $n_G$ is the number of entries, $tr(V)$ is the trace of $V$, $\theta_i^{AMV} = \frac{1}{n_G} tr(Z_i G_i Z_i^T)$ is the marginal variance explained by the $i$th genetic factor in the model (e.g., $M$ and $G : M$), $Z_i$ are design matrices for the $c$ genetic factors, $\theta_i^{AMV} = \frac{1}{n_G} tr(R_i)$ is the AMV estimator of the residual variance, and $R_i$ is the residual variance-covariance matrix. The AMV estimator of the genetic variance among entries ($G$) is:

$$\hat{\theta}_G^{AMV} = \frac{1}{n_G} \tilde{\sigma}_G^2 = \frac{1}{n_G} tr(Z_{cg} G_{cg} Z_{cg}^T)$$

(17)

where $Z_{cg}$ is a $n_G$ identity matrix. From LMM (15), the AMV estimator of the variance associated with a single marker locus with balanced data is:

$$\hat{\theta}_M^{AMV} = \frac{1}{n_G} \tilde{\sigma}_M^2 = \frac{1}{n_G} tr(Z_{cm} G_{cm} Z_{cm}^T)$$

(18)

where $E(\tilde{\sigma}_M^2) = \sigma_M^2$, $Z_{cm} = I_{n_M} \otimes 1_{n_G,M}$, $1_{n_G,M}$ is an $n_G,M$-element unit vector, and $u_M$ is a vector of random effects for $M$. The AMV estimator of the variance associated with the residual genetic variation among entries nested in $M$ is:

$$\hat{\theta}_G^{AMV} = \frac{1}{n_G} \tilde{\sigma}_G^2 = \frac{1}{n_G} tr(Z_{cg} G_{cg} Z_{cg}^T)$$

(19)

where $n_G,M$ is a vector of random entry nested in $M$ effects and $Z_{cg}$ is a $n_G$ identity matrix.

Hence, the AMV estimators of $\sigma_M^2$ and $\sigma_G^2$ are identical to ANOVA estimators (4) and (5), respectively, with entry means as input for the former and original observations as input for the latter.

ASV, or the average variance of differences among observations, leads to a definition of the total variance that provides a natural way to account for the heterogeneity of variance and covariance among observations [73, 102]. ASV can be defined for any variance-covariance structure in a generalized LMM and allows for missing and unbalanced data [73]. The ASV estimator of total variance is half the average variance of pairwise differences among entries and can be partitioned into independent sources of variance, e.g., genetic and non-genetic or residual:

$$\hat{\theta}_p^{ASV} = \frac{1}{n_G-1} tr(V D_{nc}) = \sum_i \hat{\theta}_i^{ASV} + \hat{\theta}_G^{ASV}$$

(20)
where $D_{ng} = I_{ng} - n_g^{-1}J_{ng}$ is the idempotent matrix used for column-wise mean-centering, $I_{ng}$ is an $n_g \times n_g$ identity matrix, and $J_{ng}$ is an $n_g \times n_g$ unit matrix [73]. $\theta_{i,y}^{AV}$ accounts for the variance and covariance of the phenotypic observations. From (20), $\theta_i^{AV} = (n_g - 1)^{-1} tr(Z_n G^T D_{ng})$ is the variance explained by the $i^{th}$ genetic factor ($u_i$), where $c$ indexes genetic factors, the genetic factors are marker locus effects and entries nested in marker locus effects, and $\theta_i^{AV} = (n_g - 1)^{-1} tr(RD_{ng})$ is the residual variance. The variance explained by the $i^{th}$ genetic factor is $\theta_i^{AV} = E(s_i^2)$, e.g., for a single marker locus $M$, $\theta_i^{AV} = E(s_i^2) = E(s_i^2) = E(s_i^2)$, and the biases of these ASV estimators are defined in SI Text.

The ASV estimator of the genetic variance among entries ($G$) is:

$$\hat{\theta}_G^{AV} = (n_g - 1)^{-1} \hat{\sigma}_G^2 tr(Z_n G^T D_{ng}) = \hat{\sigma}_G^2$$

where $Z_n$ is a $n_g$ identity matrix. Hence, from Eqs (8), (17) and (21), AMV and ASV estimators of the between-entry variance component ($\sigma_y^2$) are equivalent ($\hat{\sigma}_G^2 = \hat{\sigma}_G^{AMV} = \hat{\sigma}_G^{AV}$). The ASV estimator of the variance associated with $M$ is:

$$\hat{\theta}_M^{AV} = (n_g - 1)^{-1} \hat{\sigma}_M^2 tr(Z_n M^T D_{ng}) = \frac{(n_g - 1)n_M}{n_g - 1} \hat{\sigma}_M^2$$

where $k_M = df_M n_g dM/df_g$ is the bias correction coefficient, $Z_n M^T = I_{ng} \otimes 1_{n_M}$, $df_g = n_g - 1$, $df_M = n_M - 1$, and $df_{GM} = df_M - df_g$. This definition of the $k_M$-bias coefficient is identical to the earlier definition with $r_{1g}$ factored out (see Eq 12). Eq (22) shows that the ASV estimator of $\sigma_M^2$ is corrected by the fraction $k_M$, which correctly scales the estimate of $\sigma_M^2$ to the genetic variance and yields unbiased estimates of $p$ and $H_M^2$. From Eqs (9) and (22), we found that $\hat{\theta}_M^{AV} < \hat{\theta}_M^{AMV}$ by the factor $k_M$. The ASV estimator of the variance associated with $G$ is:

$$\hat{\theta}_G^{AV} = (n_g - 1)^{-1} \hat{\sigma}_G^2 tr(Z_n G^T D_{ng}) = \frac{n_g - 1}{n_g - 1} \hat{\sigma}_G^2 = \hat{\sigma}_G^2$$

The ASV estimator of $p$ for a single marker locus ($M$) is:

$$\hat{p} = \frac{\hat{\theta}_M^{AV}}{\hat{\theta}_G^{AV}} = \frac{k_M \hat{\sigma}_M^2}{\hat{\sigma}_G^2}$$

Similarly, the ASV estimator of $H_M^2$ for a single marker locus is:

$$\hat{H}_M^2 = \frac{\hat{\theta}_M^{AV}}{\hat{\theta}_G^{AV}} = \frac{k_M \hat{\sigma}_M^2}{\hat{\sigma}_G^2 + r_M^{-1} \hat{\sigma}_G^2}$$

where $\hat{\sigma}_G^2 + r_M^{-1} \hat{\sigma}_G^2$ is the phenotypic variance on an entry-mean basis [25]. From these results, we found that:

$$\hat{\theta}_M^{AV} + \hat{\theta}_G^{AV} = \hat{\theta}_G^{AMV} = \hat{\sigma}_G^2 < \hat{\theta}_M^{AMV} + \hat{\theta}_{GM}^{AMV}$$

and showed that ASV estimators of $p$ and $H_M^2$ are unbiased (automatically corrected for $k_M$).

**Computer simulations confirmed that ASV-REML estimates of $p$ and $H_M^2$ are unbiased**

Computer simulations confirmed that AMV-REML estimates of $p$ (6) and $H_M^2$ (7) are upwardly biased by the factor $k_M$ and that ASV-REML estimates of these parameters form (24) and (25).
The mean of AMV-REML estimates of $p$ and $H^2_M$ from 21 different simulation study designs (S1 Table) were identical to those predicted by the $k_M$ coefficients shown in S1, S2 and S3 Texts. Several insights arose from the simulation analyses. First, the bias caused by $k_M$ increased as $H^2_M$ increased but was proportionally constant for different $H^2_M$ (Fig 1). These results show that the overestimation of $p$ and $H^2_M$ is greatest for genes and gene-gene interactions with large effects (Fig 1). Their effects could be inflated by selection bias over and above $k_M$ bias [67, 68, 82, 83, 103]; hence, we concluded that $k_M$ bias and selection bias could operate in combination to inflate estimates of the contribution of a locus to the heritable variation in a population (S1, S2 and S3 Texts). Moreover, because the bias increases as the effect of the locus increases, we concluded that the overestimation problem is worst for large-effect QTL (Fig 1). Second, $k_M$ bias was greater for unbalanced than balanced data (Fig 1D and 1E). The effect of unbalanced data was more extreme for the F2 simulation (Fig 1D) where the expected genotypic ratio was 1 AA: 2 Aa: 1 aa than for simulations where 10 or 33% of the observations were randomly missing for markers with roughly equal numbers of replicates.
marker genotype (Fig 1E and 1F). Third, the F2 and missing data simulations further showed that the precision of estimates of these parameters decreased as the genotypic data imbalance increased. Even though bias-corrected AMV and ASV estimates of these parameters are unbiased, the sampling variances among the simulated F2 samples were larger than observed for...
the 10 and 33% missing data samples and yielded a small percentage of $H_M^2$ estimates slightly greater than 1.0 (Fig 1D). For the other simulation study designs (Fig 1), none of the ASV estimates exceeded 1.0. The sample variances of $p$ and $H_M^2$ can be estimated using data resampling methods, e.g., bootstrapping [104], or the estimators we developed using the Delta method (S5 Text) [25, 105, 106]. Equations (S44) and (S45) in S5 Text show that ASV estimates are more precise than AMV estimates by a factor of $k_{M}^2$. These predictions perfectly aligned with the empirical bootstrap estimates. Fourth, the relative biases were not affected by the number of replications of entries or the number of entries, although the precision of $\sigma_{M}^2$ estimates increased as $n_G$ and $H_M^2$ increased (Fig 2). Predictably, the number of entries ($n_G$) dramatically affected the precision of estimates of $H_M^2$ (Fig 2C and 2D). The relative biases were not affected by $r_C$ or $H_M^2$; however, the sampling variances were strongly affected by $H_M^2$ and decreased as $H_M^2$ increased (Fig 2E and 2F and S2 Fig).

GWAS example: A single marker locus with highly unbalanced genotypic data

The bias-correction methods described above are illustrated here for highly unbalanced genotypic data from a GWAS experiment. Variance components were estimated for two SNP markers (AX493 and AX396) in LD with a gene (FW1) conferring resistance to Fusarium wilt in a strawberry (Fragaria × ananassa) GWAS population ($n_G = 564$) genotyped with a genome-wide framework of SNP markers [86]. Both SNP markers had highly significant GWAS effects with $-\log_{10}(p) = 6.61 \times 10^{-31}$ for AX493 and $2.95 \times 10^{-22}$ for AX396. Genotype frequencies were highly unbalanced for both markers with a scarcity of AA homozygotes (2.8%) for AX396 (16AA : 177Aa : 371aa) and a 1 : 2 : 1 ratio for AX493 (141AA : 282Aa : 141aa). For both loci, the minor allele frequency was $>0.05$. The $k_M$ for these data ($k_{AX493} = 0.62$ and $k_{AX396} = 0.47$) were calculated as shown in S1 Text. The AMV-REML estimate of $H_M^2$ for AX396 exceeded 1.0, a telltale sign of $k_M$-bias (Table 1). AMV-REML estimates of $\sigma_{M}^2$ and $H_M^2$ for both SNP markers were double or nearly double their bias-corrected ASV-REML estimates (Table 1). The bias-corrected estimate of marker heritability for AX396 was 0.62, versus 1.33 for the uncorrected estimate. Even with bias-correction, the sum of ASV-REML estimates of $\sigma_{M}^2$ and $\sigma_{G|M}^2$ for AX493 was slightly greater than the ASV-REML estimate of $\sigma_{G}^2$. This result was consistent with finding for highly unbalanced marker genotypic data in our simulation studies where a certain fraction of bias-corrected estimates exceeded the theoretical limit for heritability because of decreased precision (Fig 1). The $k_M$-bias problem would not necessarily have been detected in the analysis of AX396 because the $p$ and $H_M^2$ estimates fell within the expected range, e.g., $\hat{\theta}_{G|M|\text{AX396}} + \hat{\theta}_{G|M|\text{AX396}}/\hat{p} = 0.71$ (Table 1). Although both SNP markers were closely associated with FW1, they accounted for dramatically different fractions of genetic variance because of historic recombination and because neither are causal DNA variants or in complete LD with causal DNA variants [17, 19, 86, 107].

QTL mapping example: Three marker loci with slightly unbalanced genotypic data

Statistics are shown here for an analysis of three marker loci (BR, PHY, and HYP) affecting seed oil content in a sunflower (Helianthus annuus) RIL population using LMM (27) [53]. The genotypic data were only slightly unbalanced and the three marker loci were identified by QTL mapping. The $k_M$ needed for bias-correcting AMV-REML estimates of $p$ and $H_M^2$ are shown in S3 Text (Table 1). The AMV-REML estimates of $p$ and $H_M^2$ were nearly double the bias-corrected ASV-REML estimates, e.g., the AMV-REML estimate of $H_M^2$ for the three-locus genetic
model (0.79) was nearly two-fold greater than the ASV-REML estimate (0.41) (Table 1). Similarly, the AMV-REML estimate of \( p \) for the \( BR \) locus (0.54) was slightly more than double the bias-corrected (ASV-REML) estimate (0.26). Hence, the uncorrected REML estimates of \( p \) and \( H^2_M \) grossly inflated the predicted contributions of the three marker loci to genetic variation for seed oil content (Table 1).

**GWAS example: Three marker loci with unbalanced genotypic data and unreplicated entries**

The application of bias-correction is illustrated here for a genetic model with three marker loci, highly unbalanced genotypic data, and a single phenotypic observation per individual—\( \sigma_g^2 \) and \( p \) could not be estimated for this example because individuals were unreplicated. Variance components were estimated for three SNP markers (\( rs_{10}, rs_{45}, \) and \( rs_{20} \)) on chromosomes 2, 6, and 22, respectively, affecting white spotting (%) in a Holstein–Friesian cattle (\( Bos taurus \)) population (\( n_G = 2,973 \)) [85]. These SNP markers had the largest effects among those predicted to be in LD with genes affecting white spotting. The genotypic frequencies were 50AA : 586Aa : 2,337aa for \( rs_{10} \), 78AA : 736Aa : 2,159aa for \( rs_{45} \), and 237AA : 976Aa : 1,760aa for \( rs_{20} \). The \( k_M \) for these data (\( k_{rs_{10}} = 0.35, k_{rs_{45}} = 0.41, \) and \( k_{rs_{20}} = 0.54 \)) were calculated as shown in S3 Text. The uncorrected AMV-REML estimate of \( H^2_M \) for the three-locus genetic model (0.76) was substantially greater than the bias-corrected ASV-REML estimate (0.37) (Table 1). Similar differences were observed for the three marker loci.

**Candidate gene analysis: Fixed or random, BLUE or BLUP?**

Our study was partly motivated by inconsistencies in the statistical approaches applied in candidate gene and other complex trait analyses when testing hypotheses and fitting genetic models for multiple large-effect loci. With the high densities of genome-wide markers commonly assayed in gene finding studies, investigators often identify markers tightly linked to candidate or known causal genes, as exemplified by diverse real world examples [17, 19, 33, 34, 37, 38, 40, 42, 43, 52, 54, 108]. The candidate marker loci are nearly always initially identified by genome-wide searches using sequential (marker-by-marker) approaches [56, 72, 75, 79, 109, 110]. Complicated and often misunderstood problems arise in the estimation and interpretation of statistics from sequential fixed effect analyses when the data are unbalanced [79, 111, 112]. Most importantly, there are multiple model fitting and analysis options (Type I, II, and III ANOVA) and the reduction in error sums of squares (SSE), test statistics, and parameter estimates differ among them, a problem that disappears when the data are balanced or when single large effect loci are discovered [79, 111–113]. Our review of the literature uncovered substantial variation and inconsistencies in the statistical approaches applied to the problem of fitting multilocus genetic models, testing multilocus genetic hypotheses, and estimating best linear unbiased estimates (BLUEs), e.g., estimated marginal means from a fixed effects analysis of marker loci.

The problems that arise in fixed effect analyses of unbalanced data profoundly affect parameter estimates and statistical inferences but have not been universally recognized or addressed in complex trait analyses [79, 112]. We reanalyzed the cattle and sunflower examples with markers as fixed effects (Table 2) to show this, illustrate the challenges and nuances of fixed effects analyses of unbalanced data, and facilitate comparisons between random and fixed effects analyses of marker loci [56, 56, 75, 79, 109–112]. Following the discovery of statistically significant marker-trait associations from a marker-by-marker genome-wide scan, the natural progression would be to analyze multilocus genetic models where the effects of the discovered loci are simultaneously corrected for the effects of other discovered loci [79, 112], as shown in
our multilocus analysis examples (Tables 1 and 2). This is straightforward when the genotypic data are balanced or nearly balanced (as in the sunflower example) but more complicated and convoluted when the genotypic data are unbalanced (as in the cattle example) [75, 79, 111, 112]. Although methods for fixed effect analyses of factorial treatment designs (multilocus genetic models) with unbalanced data are well known [56, 79, 109, 110, 112], there are several model fitting and parameter estimation variations that can lead to dramatically different parameter estimates and statistical inferences. This is perfectly illustrated by the cattle example where the coefficients of determination (analogous but not identical to $H^2_M$) from Type I, II, and III analyses were substantially different from each other and from $H^2_M$ estimates from the random effects analysis (Tables 1 and 2). The differences and ambiguities among the different fixed effects approaches disappear when the random effects approach is applied to the problem.

The analysis of markers as random effects in a multilocus analyses of known or candidate genes with large effects with ASV, although historically uncommon, simultaneously yields unbiased estimates of the variance component ratios investigated in the present study ($\nu$ and $H^2_M$) and best linear unbiased predictors (BLUPs) of the additive and dominance effects of the causative loci identified by marker associations, in addition to solving the often ambiguous problems that arise in fixed effects analyses of unbalanced data [32, 75, 77, 79, 112, 113]. As discussed in depth below and illustrated through a reanalysis of the cattle and sunflower examples (Table 2), the random effects approach we described (ASV with REML estimation of the variance components) yields accurate estimates of the underlying genetic parameters (variance component ratios and BLUPs of marker effects) from a single unambiguous generalized linear mixed model analysis, whereas wildly different parameter estimates can arise among the multitude of fixed effects analyses that investigators might elect to apply in practice when the underlying genotypic and phenotypic data are unbalanced (Tables 1 and 2).

As substantiated by our simulation analyses (Figs 1 and 2), ASV with REML estimation of the underlying variance components yields accurate estimates of $\nu$ and $H^2_M$ for marker loci and interactions between marker loci, both individually and collectively, and BLUPs of the additive and dominance effects of marker loci [76, 113–115]. When the genotypic data are unbalanced, the order with which marker and marker $\times$ marker effects enter the genetic model profoundly affects parameter estimates and statistical inferences in fixed effect analyses [56, 72, 74, 116]. To illustrate this, the main effects of marker loci A, B, and C were estimated for the six possible Type I ANOVA orders of the three loci (ABC, ACB, BAC, BCA, CAB, and CBA) (Table 2). Predictably, the reduction in the error sums of squares for a particular locus differed for each Type I order in the cattle example: the Type I SS ranged from 591.4 to 3,552.3 for rs10, 4,880.5 to 9,504.4 for rs20, and 4,259.6 to 8,384.8 for rs45. The $R^2$, or PVE, estimates for marker loci were radically different among the six Type I ANOVA and Type II and III analyses. The Type I SS were, in addition, significantly greater than the Type III SS for nearly every factor. Although Type III statistics are commonly estimated and reported in analyses of factorial treatment designs with unbalanced data, there are compelling arguments for estimating Type II statistics [109, 110]; nevertheless, as we have argued, the fixed effects approach is unnecessary.

Broadly speaking, the large effect loci segregating in a population are typically necessary but not sufficient for predicting genetic merit or disease risks but are often important enough to warrant deeper study and, in animal and plant breeding, direct selection via MAS or direct modelling in genome selection applications [21, 32, 57]. The BLUP (random marker effects) approach we applied was designed to align the study of loci with large and highly predictive effects with the BLUP approaches commonly applied to genomic prediction problems that are
agnostic or indifferent to the effects of individual loci, the so-called “black box” of genomic prediction [6, 7, 20, 21, 88, 117–121]. The predictive markers associated with large effect marker loci can be integrated into the genome-wide framework of marker loci applied in genomic prediction or incorporated as fixed effects when estimating GEBVs or PRSs [21, 54, 57–61]. One of the greatest strengths of the random effects (BLUP) approach is that the genetic parameters can be estimated from a single REML analysis free of the challenges and uncertainty associated with the fixed effects model building process [79, 109, 110, 112]. Finally, if our conclusions are correct, the complex trait analysis literature is riddled with overestimates of the genotypic and phenotypic variances explained by specific genes or QTL.

Materials and methods

Simulation studies

We used computer simulation to estimate the bias and assess the accuracy of uncorrected and bias-corrected REML estimates of $p$ and $H^2_M$ for 21 study designs (S1 Table and S4 Text). Phenotypic observations ($y_{ijk}$) for LMMs (1) and (2) were simulated for $n_M = 3$ genotypes/marker locus and 21 combinations of study design variables ($n_G$, $r_G$, $r_M$, and $H^2$) with balanced or unbalanced data (S1 Table). Simulations were performed to assess the accuracy of REML estimates of $p$ and $H^2_M$ for 21 study designs with 1,000 replicates per study design (S1 Table). The phenotypic observations for each sample were obtained by generating random normal variables for entries, markers, and residuals using the R function `rnorm()` with known means and variances [122] as described by [123, 124]. The simulated random effects of entries, markers, and replications in LMMs (1) and (2) were summed to obtain $n = n_G r_G$ phenotypic observations for each study design. Variance components for the random effects in LMMs (1) and (2) were estimated using the REML function implemented in and assess the accuracy of AMV and ASV estimators of $p$ and $H^2_M$. For study designs 1–6, the true marker heritability randomly varied from 0 to 1. Study designs 1–6 demonstrate how different numbers of marker loci ($m$) and unbalanced data affect estimates of $p$ and $H^2_M$ (Fig 1; S1 Table). For study designs 5 and 6, we randomly deleted 10 and 33% of the phenotypic observations, respectively, to create unbalanced data. For study designs 7–21, the true variances of the independent variables were fixed for all samples, which allowed us to estimate the bias and relative bias associated with the different estimators (the biases are shown in S4 Text). Study designs 7–21 illustrate how $r_G$, $n_G$, and $H^2_M$ affected the biases and relative biases of $p$ and $H^2_M$ (Fig 2; S1 Table). The variance components were estimated using REML in the `lme4::lmer()` v1.1–21 [78] package in R v4.0.2 [122]. We estimated the sample variances of AMV and ASV estimates of $p$ for each study design (S1 Table). Finally, we developed estimators of the sampling variances of $p$ and $H^2_M$ using the delta method [25, 106], as shown in S5 Text.

Estimation examples

To illustrate the application of bias-correction methods and the differences between AMV and bias-corrected AMV estimates of $p$ and $H^2_M$, we reanalyzed data from a GWAS study in cattle (Bos taurus), a QTL mapping study in oilseed sunflower (Helianthus annuus L.) [53], and a GWAS study of Fusarium wilt resistance in strawberry (Fragaria × ananassa Duchesne ex Rozier) [86]. For the sunflower study, two replications ($r_G = 2$) of $n_G = 146$ recombinant inbred lines (RILs) were phenotyped for seed oil concentration (g/kg) and genotyped for three marker loci (BR, PHY, and HYP) with two homozygous marker genotypes/locus [53]. For the cattle study, unreplicated entries ($r_G = 1; n_G = 2, 973$) were phenotyped for white spotting (%) and genotyped for three marker loci (rs10, rs45, rs20) with three marker genotypes per locus [85].
LMM (2) expanded to three marker loci with all possible interactions among marker loci is:

\[ y_{ijkl} = \mu + BR_i + PHY_j + HYP_k + BR \times PHY_{ij} + BR \times HYP_{ik} + PHY \times HYP_{jk} + BR \times PHY \times HYP_{ijk} + G : (BR \times PHY \times HYP)_{ijkl} + e_{ijkl} \]  

(27)

where \( BR_i \) is the \( i^{th} \) effect of the BR locus, \( PHY_j \) is the \( j^{th} \) effect of the PHY locus, \( HYP_k \) is the \( k^{th} \) effect of the HYP locus, \( G : (BR \times PHY \times HYP)_{ijkl} \) is the \( ijk^{th} \) effect of entries nested in the \( hij^{th} \) \( BR \times PHY \times HYP \) interaction, and \( e_{ijkl} \) is the \( hijk^{th} \) residual effect. The data for RILs were balanced, whereas the data for marker genotypes were slightly unbalanced. Each of the eight \( BR \times PHY \times HYP \) homozygotes were observed in the RIL population; however, the number of entries nested in each marker genotype (\( n_{GM} \)) varied from \( n_{GABR} = 81 : 65 \), \( n_{GPHY} = 60 : 86 \), and \( n_{GPHY} = 70 : 76 \). Variance components for LMMs (1) and (27) were estimated using the REML method in \textit{nlme::lmer()} [78]. The marker-associated genetic variances for individual marker loci and two- and three-way interactions among marker loci were bias-corrected using the formula described in S1, S2 and S3 Texts.

For the strawberry study, four replications (\( r_G = 4 \)) of 565 entries (\( n_G = 565 \)) from a genome-wide association study (GWAS) were phenotyped for resistance to Fusarium wilt and genotyped for single nucleotide polymorphism (SNP) markers in LD with FW1, a dominant gene conferring resistance to \textit{Fusarium oxysporum f. sp. fragariae}, the causal pathogen [86]. The replications were asexually propagated clones of individuals; hence, the expected causal variance among individuals was equal to the total genetic variation in the population, analogous to monozygotic twins [25]. Genetic parameters were estimated for two SNP markers (AX493 and AX396) that were tightly linked to FW1 [86]. The genotypic data for both markers were highly unbalanced. Genotype numbers were 141 AA : 282 Aa : 141 aa for AX493 and 16 AA : 177 Aa : 371 aa for AX396, where A and a are alternate SNP alleles. The variance components were estimated for LMMs (1) and (2) using REML method implemented in the R package \textit{nlme::lmer()} [78]. REML estimates of the marker-associated genetic variances for both marker loci were bias-corrected using the approach described in S1 Text.

For the cattle study, we used a model similar to (27) for the analysis. However, because entries are unreplicated in this experiment, we cannot include the entries nested in the three-way marker interaction (\( G : M \)) term because it has the same levels as the residual. The LMM for this case study is:

\[ y_{ijkl} = \mu + rs10_{ih} + rs45_{ij} + rs20_{jk} + rs10 \times rs45_{ijk} + rs10 \times rs20_{ijk} + rs45 \times rs20_{ijk} + rs10 \times rs45 \times rs20_{ijkl} + G : (rs10 \times rs45 \times rs20)_{ijkl} \]  

(28)

where \( rs10_{ih} \) is the \( i^{th} \) effect of the peak SNP (rs109979909) on chromosome 2, \( rs45 \), is the \( j^{th} \) effect of the peak SNP on chromosome 6 (rs451683615), \( rs20 \), is the \( k^{th} \) effect of the peak SNP on chromosome 22 (rs209784468), and \( G : (rs10 \times rs45 \times rs20)_{ijkl} \) is the \( ijk^{th} \) residual effect comprising residual genetic effects \( G : M \) and residual error. In this experiment, there were \( k \) entries and \( k \) observations, and because of this we cannot fit LMM (1) without incorporating pedigree or genomic relatedness. In this single case, we estimate \( s_i^2 = 10.42 \) from the log-transformed data to use in the denominator of \( H_i^2 \).

We used the SAS package PROC GLM [77] for Type I and III analyses of the sunflower and cattle data with marker loci as fixed effects. Type I analyses were done for the six possible orders of main effects (ABC, ACB, BAC, BCA, CAB, and CBA) and a single order for marker \( \times \) marker interactions (\( A \times B, A \times C, B \times C, A \times B \times C \)), where \( A, B, \) and \( C \) are the three marker loci (factors). For the ABC order, the reduction SS for the main effects were \( R(A | \mu) \), \( R(B | \mu, A) \), and \( R(C | \mu, A, B) \), where \( \mu \) is the population mean. Similarly, for the ACB
order, the reduction SS for the main effects were \( R(A \mid \mu), R(C \mid \mu, A), \) and \( R(B \mid \mu, A, C), \) and so on for the other four orders (BAC, BCA, CAB, and CBA). C, A × B, A × C, B × C, A × B × C, the reduction SS for the main effect of B was \( R(B \mid B, C, A \times B, A \times C, B \times C, A \times B \times C). \) For comparison, the Type III reduction SS for the main effects were \( R(A \mid \mu, B, C, A \times B, A \times C, B \times C, A \times B \times C), \) \( R(B \mid \mu, A, C, A \times B, A \times C, B \times C, A \times B \times C), \) and \( R(C \mid \mu, A, B, A \times B, A \times C, B \times C, A \times B \times C). \)

**Supporting information**

**S1 Table. Simulation study designs and variables.** Normally distributed phenotypic observations were simulated for 21 study designs and associated linear mixed models by varying the number of observations \((n)\), the number of entries \((n_{G})\), the number of replications/entry \((r_{G})\), the number of marker loci \((m)\), \(n_{M} = 3\) genotypes/marker locus, the number of entries/marker genotype \((n_{G,M})\), and marker heritability \((H_{G}^{2})\). One thousand samples of size \(n\) were simulated for each study design. The segregation of a single marker locus in an F\(_{2}\) population was simulated in study design 4. The number of entries nested in marker genotypes for study design 4 was equivalent to the expected number for the segregation of a co-dominant DNA marker in a population segregating 1 AA : 2 Aa : 1 aa for a single marker locus. In this example, there are 135 entries nested in AA, 270 entries nested in Aa, and 135 entries nested in aa and each are replicated 5 times. Simulates the segregation of a single locus in an F\(_{2}\) population. The number of entries/genotype for study design 4.

**S1 Fig. Accuracy of AMV and ASV estimators of marker heritability when the phenotypic variance is estimated by pooling marker and residual genetic sources of variation \((\sigma_{s}^{2} + \sigma_{G,M}^{2})\).** AMV and ASV estimates of \(H_{G}^{2}\), when \(\sigma_{s}^{2}\) from LMM (1) is replaced with \(\hat{\sigma}_{s}^{2} + \hat{\sigma}_{G,M}^{2}\) for AMV from LMM (2) or \(k_{M}\hat{\sigma}_{s}^{2} + \hat{\sigma}_{G,M}^{2}\) for ASV. Estimates are shown for 1,000 segregating populations simulated for different numbers of entries \((n_{G}\) individuals, families, or strains), five replications/entry \((r_{G} = 5)\), true marker heritability \((H_{G}^{2})\) ranging from 0 to 1, and one to three marker loci with three genotypes/marker locus \((n_{M1} = 3)\). The AMV estimates (shown in red) equal \(\hat{\sigma}_{s}^{2}/(\hat{\sigma}_{s}^{2} + \hat{\sigma}_{G,M}^{2} + r_{G}^{-1}\hat{\sigma}_{G}^{2})\), whereas the ASV estimates (shown in blue) equal \(k_{M}\hat{\sigma}_{s}^{2}/(k_{M}\hat{\sigma}_{s}^{2} + \hat{\sigma}_{G,M}^{2} + r_{G}^{-1}\hat{\sigma}_{G}^{2})\). AMV estimates of marker heritability \((H_{G}^{2};\) red highlighted observations\) and ASV estimates of marker heritability \((H_{G}^{2};\) blue highlighted observations\) are shown for: (A) one locus with balanced data for \(n_{G} = 540\) entries (study design 1); (B) two marker loci with interaction \((M1, M2,\) and \(M1 \times M2)\) and balanced data for \(n_{G} = 540\) (study design 2); (C) three marker loci with interactions \((M1, M2, M3, M1 \times M2, M1 \times M3, M2 \times M3,\) and \(M1 \times M2 \times M3)\) and balanced data for \(n_{G} = 540\) (study design 3); (D) a population segregating 1:2:1 for a single marker locus with \(r_{G,M} = 135\) entries for both homozygotes and \(r_{G,M} = 270\) heterozygous entries, and \(n_{G} = 540\) (study design 4); (E) one locus with 10% randomly missing data among 540 entries (study design 5); and (F) one locus with 33% randomly missing data among 540 entries (study design 6). Study design details are shown in S1 Table.

**S2 Fig. Relative bias of AMV and ASV estimators of marker heritability.** Relative biases of AMV and ASV estimates of \(H_{G}^{2}\) are shown for 1,000 segregating populations simulated for different numbers of entries \((n_{G}\) individuals, families, or strains), five replications/entry \((r_{G} = 5)\), true marker heritability \((H_{G}^{2})\) ranging from 0 to 1, and one to three marker loci with three genotypes/marker locus \((n_{M1} = 3)\). AMV estimates of marker heritability \((H_{G}^{2};\) red highlighted observations\) and ASV estimates of marker heritability \((H_{G}^{2};\) blue highlighted observations\)
are shown for: (A) one locus with balanced data for \( n_G = 540 \) entries (study design 1); (B) two marker loci with interaction \((M_1, M_2, \text{and } M_1 \times M_2)\) and balanced data for \( n_G = 540 \) (study design 2); (C) three marker loci with interactions \((M_1, M_2, M_3, M_1 \times M_2, M_1 \times M_3, M_2 \times M_3, \text{and } M_1 \times M_2 \times M_3)\) and balanced data for \( n_G = 540 \) (study design 3); (D) an population segregating 1:2:1 for one marker locus with \( r_{GM} = 135 \) entries for both homozygotes and \( r_{GM} = 270 \) heterozygous entries, and \( n_G = 540 \) (study design 4); (E) one locus with 10% randomly missing data among 540 entries (study design 5); and (F) one locus with 33% randomly missing data among 540 entries (study design 6). Study design details are shown in S1 Table.

S1 Text. ASV estimator of the fraction of the genetic variance associated with a single marker locus for unbalanced data. (PDF)

S2 Text. ASV estimator of the fraction of the genetic variance associated with two marker loci for unbalanced data. (PDF)

S3 Text. ASV estimator of the fraction of the genetic variance associated with three marker loci for unbalanced data. (PDF)

S4 Text. Biases of AMV and ASV estimators of marker-associated variance. (PDF)

S5 Text. Sample variances for AMV and ASV estimators of \( p \) and \( H^2_M \). (PDF)

Author Contributions


Data curation: Mitchell J. Feldmann.


Funding acquisition: Hans-Peter Piepho, Steven J. Knapp.


Project administration: Steven J. Knapp.


Writing – original draft: Mitchell J. Feldmann, Hans-Peter Piepho, Steven J. Knapp.
Writing – review & editing: Mitchell J. Feldmann, Hans-Peter Piepho, William C. Bridges, Steven J. Knapp.

References


66. Utzf H, Melchinger AE, Schön CC. Bias and sampling error of the estimated proportion of genotypic variance explained by quantitative trait loci determined from experimental data in maize using cross


73. Piepho HP. A coefficient of determination (R²) for generalized linear mixed models. Biomet J. 2019; 61:860–872. PMID: 30957911


88. Mrode RA. Linear models for the prediction of animal breeding values. Cabi; 2014.


115. Molenaar H, Boehm R, Piepho HP. Phenotypic selection in ornamental breeding: it’s better to have the BLUPs than to have the BLUEs. Frontiers in plant science. 2018; 9:1511. https://doi.org/10.3389/fpls.2018.01511 PMID: 30455707


