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

Short Title: Average semivariance estimation methods

Mitchell J. Feldmann[1], Hans-Peter Piepho[2], William C. Bridges[3], Steven J. Knapp[1*]

1 Department of Plant Sciences, University of California, Davis, One Shields Avenue, Davis, California, United States of America, **2** Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Fruwirthstrasse 23, 70599 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, in addition to best linear unbiased predictors of the additive and dominance effects of the underlying loci. 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 (σ^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 segregate (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 $p = \sigma^2_M/\sigma^2_G$ to estimate the discovered or non-missing fraction of the genetic variance, where σ^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 ($H^2_M = \sigma^2_M/\sigma^2_P$) estimates the non-missing fraction of the phenotypic variance (σ^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 (σ^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, σ^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 σ^2_M, σ^2_G, and σ^2_P for balanced and unbalanced data (Appendices 2-4). 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 p as a function of the ANOVA estimators of σ^2_M and σ^2_G for balanced data and algebraically simplifying the equations. The linear mixed models (LMMs) and ANOVA estimators of the variance components needed to show this are described here. We start with the analysis of a single marker locus in an experiment where entries (e.g., individuals, families, or strains) are replicated, σ^2_G can be estimated, and the data for entries and markers are balanced. Extensions for one to three marker loci with unbalanced data are shown in Appendices 1-3. Two LMMs are needed for estimating σ^2_M, σ^2_G, 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 σ^2_G (the between entry variance component) is:

$$y_{jk} = \mu + G_j + \varepsilon_{jk} \tag{1}$$

where y_{jk} is the jk^{th} phenotypic observation, μ is the population mean, G_j is the random effect of the jth entry, ε_{jk} is the random effect of the jk^{th} residual, $G_j \sim N(0, \sigma^2_G)$, $\varepsilon_{jk} \sim N(0, \sigma^2_\varepsilon)$, j = 1, 2, ..., n_G, and k = 1, 2, ..., 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_{jk}). 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 σ^2_M and $\sigma^2_{G:M}$ is:

$$y_{ijk} = \mu + M_i + G:M_{i(j)} + \varepsilon_{ijk} \tag{2}$$

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

The ANOVA estimator of the between-entry variance component (σ^2_G) from LMM (1) with balanced data is:

$$\hat{\sigma}^2_G = \frac{MS_G - MS_\varepsilon}{r_G} = \frac{SS_G/df_G - SS_\varepsilon/df_\varepsilon}{r_G} = \frac{1}{df_G r_G} SS_G - \frac{1}{df_\varepsilon r_G} SS_\varepsilon \tag{3}$$

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, $MS_\varepsilon = SS_\varepsilon/df_\varepsilon = \sigma^2_\varepsilon$ is the residual mean square, SS_ε is the residual sum of squares, $df_\varepsilon = n_G(r_G - 1) - 1$ is the residual degrees of freedom, σ^2_ε 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 M and entry nested in M variance components from LMM (2) with balanced data are:

$$\hat{\sigma}^2_M = \frac{MS_M - MS_{G:M}}{r_G n_{G:M}} \tag{4}$$

Table 1. REML estimates of marker-associated variance (σ^2_M), the fraction of the genetic variance explained by markers ($p = \sigma^2_M/\sigma^2_G$), and marker heritability ($H^2_M = \sigma^2_M/\sigma^2_P$) 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.

					Uncorrected			Bias Corrected			Type II	Type III
Study	Source	df	k_M	Variance Component	$\hat{\sigma}^2$	$\hat{p}$	$\hat{H}^2$	$\hat{\sigma}^2_*$	$\hat{p}_*$	$\hat{H}^2_*$	$\hat{R}^{2d}$	$\hat{R}^{2e}$
Cattle	M	25	—	σ^2_{rs10} + ... + $\sigma^2_{rs10\times rs45\times rs20}$	7.92	—	0.76	3.88	—	0.37	—	—
White Spotting[a]	$rs10$	2	0.35	σ^2_{rs10}	0.62	—	0.06	0.21	—	0.02	0.04	0.00
	$rs45$	2	0.41	σ^2_{rs45}	2.91	—	0.28	1.20	—	0.11	0.21	0.08
	$rs20$	2	0.54	σ^2_{rs20}	3.81	—	0.37	2.04	—	0.20	0.23	0.10
	$rs10\times rs45$	4	0.58	$\sigma^2_{rs10\times rs45}$	0.00	—	0.00	0.00	—	0.00	0.00	0.00
	$rs10\times rs20$	4	0.67	$\sigma^2_{rs10\times rs20}$	0.00	—	0.00	0.00	—	0.00	0.01	0.01
	$rs45\times rs20$	4	0.70	$\sigma^2_{rs45\times rs20}$	0.37	—	0.04	0.26	—	0.02	0.01	0.01
	$rs10\times rs45\times rs20$	7	0.77	$\sigma^2_{rs10\times rs45\times rs20}$	0.22	—	0.02	0.17	—	0.02	0.01	0.01
	$G:rs10\times rs45\times rs20$	2,935	—	$\sigma^2_{G:rs10\times rs45\times rs20}$	5.26	—	—	5.26	—	—	—	—
Sunflower	Entry (G)	145	—	σ^2_G	21.61	—	0.95	21.61	—	0.95	—	—
Oil Content[b]	$M+G:M$	145	—	σ^2_B + ... + $\sigma^2_{G:M}$	30.76	1.42	1.35	22.15	1.02	0.98	—	—
	M	7	—	σ^2_B + ... + $\sigma^2_{B\times P\times HYP}$	17.85	0.83	0.79	9.24	0.43	0.41	—	—
	BR	1	0.48	σ^2_B	11.57	0.54	0.51	5.59	0.26	0.25	0.21	0.26
	PHY	1	0.47	σ^2_{PHY}	1.26	0.06	0.06	0.60	0.03	0.03	0.02	0.04
	HYP	1	0.49	σ^2_{HYP}	2.9	0.13	0.13	1.41	0.07	0.06	0.05	0.10
	$BR\times PHY$	1	0.77	$\sigma^2_{B\times P}$	0.21	0.01	0.01	0.17	0.01	0.01	0.01	0.01
	$BR\times HYP$	1	0.78	$\sigma^2_{B\times HYP}$	1.89	0.09	0.08	1.46	0.07	0.06	0.03	0.04
	$PHY\times HYP$	1	0.77	$\sigma^2_{PHY\times HYP}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$BR\times PHY\times HYP$	1	0.88	$\sigma^2_{B\times PHY\times HYP}$	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01
	$G:BR\times PHY\times HYP$	138	—	$\sigma^2_{G:B\times PHY\times HYP}$	12.91	0.60	0.57	12.91	0.60	0.57	—	—
	Residual (ε)	144	—	σ^2_ε	2.07	—	—	2.07	—	—	—	—
Strawberry	Entry (G)	557	—	σ^2_G	3.26	—	0.98	3.26	—	0.98	—	—
Fusarium Wilt[c]	$M+G:M$	557	—	σ^2_{AX396} + $\sigma^2_{G:AX396}$	4.77	1.46	1.44	2.39	0.73	0.72	—	—
	$AX396$ (M)	2	0.47	σ^2_{AX396}	4.48	1.37	1.35	2.09	0.64	0.63	0.84	0.84
	$G:AX396$	555	—	$\sigma^2_{G:AX396}$	0.30	0.09	0.09	0.30	0.09	0.09	—	—
	Residual (ε)	1,631	—	σ^2_ε	0.23	—	—	0.23	—	—	—	—
Strawberry	Entry (G)	540	—	σ^2_G	3.30		0.98	3.30	—	0.98	—	—
Fusarium Wilt[c]	$M+G:M$	540	—	σ^2_{AX493} + $\sigma^2_{G:AX493}$	4.01	1.21	1.20	3.45	1.05	1.03	—	—
	$AX493$ (M)	2	0.62	σ^2_{AX493}	1.48	0.45	0.44	0.93	0.28	0.28	0.22	0.22
	$G:AX493$	538	—	$\sigma^2_{G:AX493}$	2.53	0.77	0.75	2.53	0.77	0.75	—	—
	Residual (ε)	1,584	—	σ^2_ε	0.23	—	—	0.23	—	—	—	—

[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 ($r_G = 1$). k_M coefficient equations for three loci with unbalanced data are shown in Appendix 4.

[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 (A13) 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_M coefficient a single locus with unbalanced data are shown in Appendix 2.

[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 rs_{10} was $\mathrm{R}(rs_{10}|rs_{45}, rs_{20})$. Similarly, the reduction in SS for each two-locus interaction was estimated without main or three-way interaction effects in the genetic model, e.g., the Type II reduction in sum of squares for the $rs_{10}\times rs_{45}$ interaction was $\mathrm{R}(rs_{10}\times rs_{45}|rs_{45}, rs_{20}, rs_{10}\times rs_{20}, rs_{45}\times rs_{20})$ and so on for the other two-locus interactions. Finally, the reduction in SS for the three-locus interaction was $\mathrm{R}(rs_{10}\times rs_{45}\times rs_{20}|rs_{10}, rs_{45}, rs_{20}, rs_{10}\times rs_{45}, rs_{10}\times rs_{20}, rs_{45}\times rs_{20})$.

[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 rs_{10} in the cattle example was estimated by fitting rs_{10} with all other factors in the model: $\mathrm{R}(rs_{10}|rs_{45}, rs_{20}, rs_{10}\times rs_{45}, rs_{10}\times rs_{20}, rs_{45}\times rs_{20}, rs_{10}\times rs_{45}\times rs_{20})$.

and

$$\hat{\sigma}^2_{G:M} = \frac{MS_{G:M} - MS_\varepsilon}{r_G} \tag{5}$$

respectively, where $n_{G:M}$ is the number of entries nested in each marker genotype, $E(\hat{\sigma}^2_M) = \sigma^2_M$, $E(\hat{\sigma}^2_{G:M}) = \sigma^2_{G:M}$, $MS_{G:M}$ 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 ($\hat{\sigma}^2_\varepsilon = MS_\varepsilon$). Hence, for a single marker locus with balanced data,

the ANOVA estimator of p is:

$$\hat{p} = \frac{\hat{\sigma}^2_M}{\hat{\sigma}^2_G} \tag{6}$$

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

$$\hat{H}^2_M = \frac{\hat{\sigma}^2_M}{\hat{\sigma}^2_G + \hat{\sigma}^2_\varepsilon / r_G} \tag{7}$$

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

The overestimation of p and H^2_M was not obvious from inspection of ANOVA estimators (6) and (7). The source of the bias was discovered by substituting $SS_M + SS_{G:M}$ for SS_G in the ANOVA estimator of σ^2_G from (3) and simplifying:

$$\begin{aligned}
\hat{\sigma}^2_G =& \frac{1}{df_G r_G}(SS_M + SS_{G:M}) - \frac{1}{df_\varepsilon r_G} SS_\varepsilon \\
=& \frac{1}{df_G r_G}[df_M(\hat{\sigma}^2_\varepsilon + r_{G:M}\hat{\sigma}^2_{G:M} + r_M\hat{\sigma}^2_M) + df_{G:M}(\hat{\sigma}^2_\varepsilon + r_{G:M}\hat{\sigma}^2_{G:M})] - \frac{1}{df_\varepsilon r_G} df_\varepsilon \hat{\sigma}^2_\varepsilon \\
=& \frac{df_M r_M}{df_G r_G}\hat{\sigma}^2_M + \hat{\sigma}^2_{G:M} = \frac{df_M n_{G:M}}{df_G}\hat{\sigma}^2_M + \hat{\sigma}^2_{G:M} = k_M\hat{\sigma}^2_M + \hat{\sigma}^2_{G:M}
\end{aligned} \tag{8}$$

where the fraction k_M is source of the bias, $0 < k_M < 1$, r_M is the number of replications per marker genotype, $n_{G:M}$ 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, $SS_{G:M}$ is the entry nested in marker sum of squares, and $df_{G:M}$ is the entry nested in marker degrees of freedom. The term k_M in (8) depends on degrees of freedom and $n_{G:M}$ and is hereafter referred to as the k_M bias coefficient, where the subscript M indexes the intralocus and interlocus effects of marker loci.

Equation (8) shows that the sum of ANOVA estimates of σ^2_M and $\sigma^2_{G:M}$ from LMM (1) are greater than the ANOVA estimate of σ^2_G from LMM (2):

$$\hat{\sigma}^2_M + \hat{\sigma}^2_{G:M} > \hat{\sigma}^2_G = k_M\hat{\sigma}^2_M + \hat{\sigma}^2_{G:M} \tag{9}$$

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

$$\hat{p} = \frac{\hat{\sigma}^2_M}{\hat{\sigma}^2_G} = \frac{\hat{\sigma}^2_M}{k_M\hat{\sigma}^2_M + \hat{\sigma}^2_{G:M}} \tag{10}$$

and

$$\hat{H}^2_M = \frac{\hat{\sigma}^2_M}{\hat{\sigma}^2_{\bar{P}}} = \frac{\hat{\sigma}^2_M}{k_M\hat{\sigma}^2_M + \hat{\sigma}^2_{G:M} + \hat{\sigma}^2_\varepsilon / r_G} \tag{11}$$

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

$$k_M = \frac{df_M r_M}{df_G r_G} = \frac{df_M n_{G:M}}{df_G} \tag{12}$$

and

$$\left(\frac{\hat{\sigma}^2_M}{\hat{\sigma}^2_G} = \frac{\hat{\sigma}^2_M}{k_M\hat{\sigma}^2_M + \hat{\sigma}^2_{G:M}}\right) > \frac{\hat{\sigma}^2_M}{\hat{\sigma}^2_M + \hat{\sigma}^2_{G:M}} > \left(\frac{k_M\hat{\sigma}^2_M}{k_M\hat{\sigma}^2_M + \hat{\sigma}^2_{G:M}} = \frac{k_M\hat{\sigma}^2_M}{\hat{\sigma}^2_G}\right) \tag{13}$$

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 σ^2_M 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 (Appendices 1-3). For a single marker locus with unbalanced genotypic data, we found:

$$k_M = \frac{n_G - n_G^{-1}\sum_h n^2_{G:M_h}}{df_G} \tag{14}$$

where n_G is the number of entries, df_G are the degrees of freedom for entries, and $n_{G:M_h}$ is the number of entries nested in the hth marker genotype (Appendix 1). 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 (A10)-(A12) in Appendix 2. Similarly, for a three locus genetic model, see equations (A19)-(A25) in Appendix 3. 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 $k_{M_1\times M_2}$ is the coefficient for the $M_1 \times M_2$ epistatic interaction (Appendix 2). 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 ($df_G r_G$) 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 M_1 or M_2 variance components for a two locus genetic model. Similarly, for a three locus genetic model, the upward bias is proportionally smaller for the $M_1 \times M_2 \times M_3$ interaction variance component than the two-way interaction variance components ($M_1 \times M_2$, $M_1 \times M_3$, and $M_2 \times M_3$). 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 Table (1-2), 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* × *ananassa*), and apple (*Malus* × *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 ($\sigma^2_{G:M}$ becomes the residual) and σ^2_G, σ^2_ε, and p cannot be estimated; however, the marker heritability can be estimated using the phenotypic variance among unreplicated individuals ($k_M\sigma^2_M + \sigma^2_{G:M}$). 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), (A3), (A9), and (A18), and Appendices 1-3, 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 ($0 < \hat{H}^2_M < 1$). 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, σ^2_G be estimated using a genomic or pedigree relationship matrix [92, 95–97], which yields an estimate of p.

Average Semivariance Estimation Directly Solves the Bias Problem

The AMV methods proposed above for bias correcting ANOVA or REML estimates of p and H^2_M 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 ($\bar{y}_{ij\bullet}$) from LMM (1) stored in an n_G-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:

$$\bar{y}_{ij\bullet} = \mu + M_i + G : M_{i(j)} + \bar{\varepsilon}_{ij\bullet} \tag{15}$$

where $\bar{y}_{ij\bullet}$ is the phenotypic mean for the ij^{th} entry, μ is the population mean, M_i is the random effect of the ith marker genotype, $var(M_i) = \sigma^2_M$, $G : M_{i(j)}$ is the random effect of entries nested in M, $var(G : M_{i(j)}) = \sigma^2_{G:M}$, $\bar{\varepsilon}_{ij\bullet}$ is the residual error, and $var(\bar{\varepsilon}_{ij\bullet}) = r_G^{-1}\sigma^2_\varepsilon$. 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 σ^2_M and $\sigma^2_{G:M}$ with individual variance-covariance matrices G_c 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:

$$\hat{\theta}^{AMV}_P = {n_G}^{-1} tr(V) = \sum_c \hat{\theta}^{AMV}_{g_c} + \hat{\theta}^{AMV}_{\bar{\varepsilon}} \tag{16}$$

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, $\theta^{AMV}_{g_c} = {n_G}^{-1} tr(Z_c G_c Z_c^T)$ is the marginal variance explained by the cth genetic factor in the model (e.g., M and $G : M$), Z_c are design matrices for the c genetic factors, $\theta^{AMV}_{\bar{\varepsilon}} = {n_G}^{-1} tr(R)$ is the AMV estimator of the residual variance, and R is the residual variance-covariance matrix. The AMV estimator of the genetic variance among entries (G) is:

$$\hat{\theta}^{AMV}_G = (n_G)^{-1} \hat{\sigma}^2_G tr(Z_{u_G} Z^T_{u_G}) = \hat{\sigma}^2_G \tag{17}$$

where Z_{u_G} 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:

$$\hat{\theta}^{AMV}_M = (n_G)^{-1} \hat{\sigma}^2_M tr(Z_{u_M} Z^T_{u_M}) = \frac{(n_M) n_{G:M}}{n_G} \hat{\sigma}^2_M = \hat{\sigma}^2_M \tag{18}$$

where $E(\hat{\sigma}^2_M) = \sigma^2_M = \theta^{AMV}_M$, $Z_{u_M} = I_{n_M} \otimes 1_{n_{G:M}}$, I_{n_M} is a n_M identity matrix, $1_{n_{G:M}}$ 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:

$$\hat{\theta}^{AMV}_{G:M} = (n_G)^{-1} \hat{\sigma}^2_{G:M} tr(Z_{u_{G:M}} Z^T_{u_{G:M}}) = \frac{n_G}{n_G} \hat{\sigma}^2_{G:M} = \hat{\sigma}^2_{G:M} \tag{19}$$

where $u_{G:M}$ is a vector of random entry nested in M effects and $Z_{u_{G:M}}$ is a n_G identity matrix. Hence, the AMV estimators of σ^2_M and σ^2_G 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:

$$\hat{\theta}^{ASV}_{\bar{P}} = (n_G - 1)^{-1} tr(VD_{n_G}) = \sum_c \hat{\theta}^{ASV}_{g_c} + \hat{\theta}^{ASV}_{\bar{\varepsilon}} \tag{20}$$

where $D_{n_G} = I_{n_G} - {n_G}^{-1} J_{n_G}$ is the idempotent matrix used for column-wise mean-centering, I_{n_G} is an $n_G \times n_G$ identity matrix, and J_{n_G} is an $n_G \times n_G$ unit matrix [73]. $\theta^{ASV}_{\bar{P}}$ accounts for the variance and covariance of the phenotypic observations. From (20), $\theta^{ASV}_{g_c} = (n_G - 1)^{-1} tr(Z_c G_c Z_c^T D_{n_G})$ is the variance explained by the cth genetic factor (u_c), where c indexes genetic factors, the genetic factors are marker locus effects and entries nested in marker locus effects, and $\theta^{ASV}_{\bar{\varepsilon}} = (n_G - 1)^{-1} tr(RD_{n_G})$ is the residual variance. The variance explained by the cth genetic factor is $\theta^{ASV}_{g_c} = E(s^2_{g_c})$, e.g., for a single marker locus M, $\theta^{ASV}_M = E(s^2_M)$. $E(s^2_{g_c})$, $E(s^2_M)$, and the biases of these ASV estimators are defined in Appendix 4.

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

$$\hat{\theta}^{ASV}_G = (n_G - 1)^{-1} \hat{\sigma}^2_G tr(Z_{u_G} Z^T_{u_G} D_{n_G}) = \hat{\sigma}^2_G \tag{21}$$

where Z_{u_G} is a n_G identity matrix. Hence, from equations (8), (17), and (21), AMV and ASV estimators of the between-entry variance component (σ^2_G) are equivalent ($\hat{\sigma}^2_G = \hat{\theta}^{AMV}_G = \hat{\theta}^{ASV}_G$). The ASV estimator of the variance associated with M is:

$$\hat{\theta}^{ASV}_M = (n_G - 1)^{-1} \hat{\sigma}^2_M tr(Z_{u_M} Z^T_{u_M} D_{n_G}) = \frac{(n_M - 1) n_{G:M}}{n_G - 1} \hat{\sigma}^2_M = \frac{df_M n_{G:M}}{df_G} \hat{\sigma}^2_M = k_M \hat{\sigma}^2_M \tag{22}$$

where $k_M = df_M n_{G:M} / df_G$ is the bias correction coefficient, $Z_{u_M} = I_{n_M} \otimes 1_{n_{G:M}}$, $df_G = n_G - 1$, $df_M = n_M - 1$, and $df_{G:M} = df_G - df_M$. This definition of the k_M-bias coefficient is identical to the earlier definition with r_G factored out (see equation 12). Equation (22) shows that the ASV estimator of σ^2_M is corrected by the fraction k_M, which correctly scales the estimate of σ^2_M to the genetic variance and yields unbiased estimates of p and H^2_M. From equations (9) and (22), we found that $\hat{\theta}^{ASV}_M < \hat{\theta}^{AMV}_M$ by the factor k_M. The ASV estimator of the variance associated with $G : M$ is:

$$\hat{\theta}^{ASV}_{G:M} = (n_G - 1)^{-1} \hat{\sigma}^2_{G:M} tr(Z_{u_{G:M}} Z^T_{u_{G:M}} D_{n_G}) = \frac{n_G - 1}{n_G - 1} \hat{\sigma}^2_{G:M} = \hat{\sigma}^2_{G:M} \tag{23}$$

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

$$\hat{p}_* = \frac{\hat{\theta}^{ASV}_M}{\hat{\theta}^{ASV}_G} = \frac{k_M \hat{\sigma}^2_M}{\hat{\sigma}^2_G} = \frac{k_M \hat{\sigma}^2_M}{k_M \hat{\sigma}^2_M + \hat{\sigma}^2_{G:M}} \tag{24}$$

Similarly, the ASV estimator of H^2_M for a single marker locus is:

$$\hat{H}^2_{M*} = \frac{\hat{\theta}^{ASV}_M}{\hat{\theta}^{ASV}_{\bar{P}}} = \frac{k_M \hat{\sigma}^2_M}{\hat{\sigma}^2_G + r_G^{-1} \hat{\sigma}^2_{\bar{\varepsilon}}} = \frac{k_M \hat{\sigma}^2_M}{k_M \hat{\sigma}^2_M + \hat{\sigma}^2_{G:M} + r_G^{-1} \hat{\sigma}^2_{\bar{\varepsilon}}} \tag{25}$$

where $\hat{\sigma}^2_G + \hat{\sigma}^2_{\bar{\varepsilon}} / r_G = \hat{\sigma}^2_{\bar{P}}$ is the phenotypic variance on an entry-mean basis [25]. From these results, we found that:

$$\hat{\theta}^{ASV}_M + \hat{\theta}^{ASV}_{G:M} = \hat{\theta}^{ASV}_G = \hat{\theta}^{AMV}_G = \hat{\sigma}^2_G < \hat{\theta}^{AMV}_M + \hat{\theta}^{AMV}_{G:M} \tag{26}$$

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

Computer Simulations Confirmed That ASV-REML Estimates of p and H^2_M are Unbiased

Computer simulations confirmed that AMV-REML estimates of p (equation 6) and H^2_M (equation 7) are upwardly biased by the factor k_M and that ASV-REML estimates of these parameters (equations 24) and 25) are unbiased (Fig 1 and 2). The mean of AMV-REML estimates of p and H^2_M from 21 different simulation study designs (Table S1 were identical to those predicted by the k_M coefficients shown in Appendices 1-3. 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 1A-C). Their effects could be inflated by selection bias over and above k_M bias [67, 68, 82, 83]; 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 (Appendix 2-4). 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-E). The effect of unbalanced data was more extreme for the F_2 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-F). Third, the F_2 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 F_2 samples were larger than observed for the 10 and 33% missing data samples and yielded a small percentage of H^2_M estimates slightly greater than 1.0 (Fig 1D). For the other simulation study designs (Fig 1A-C and E-F), none of the ASV estimates exceeded 1.0. The sample variances of p and H^2_M can be estimated using data resampling methods, e.g., bootstrapping [103], or the estimators we developed using the Delta method (Appendix 5). Equations (A44) and (A45) in Appendix 5 show that ASV estimates are more precise than AMV estimates by a factor of k^2_M. 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 σ^2_M estimates increased as n_G and H^2_M increased (Fig 2). Predictably, the number of entries (n_G) dramatically affected the precision of estimates of H^2_M (Fig 2C-D). The relative biases were not affected by r_G or H^2_M; however, the sampling variances were strongly affected by H^2_M and decreased as H^2_M increased (Fig 2E-F and S2) .

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 2.95×10^{-222} 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 Appendix 1. The AMV-REML estimate of H^2_M for $AX396$ exceeded 1.0, a telltale sign of k_M-bias (Table 1). AMV-REML estimates of σ^2_M and H^2_M 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 σ^2_M and $\sigma^2_{G:M}$ for $AX493$ was slightly greater than the ASV-REML estimate of σ^2_G. 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^2_M estimates fell within the expected range, e.g., $\hat{\theta}^{ASV}_{AX396} + \hat{\theta}^{ASV}_{G:AX396} / \hat{\theta}^{ASV}_{\bar{P}} = 0.71$ (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, 104].

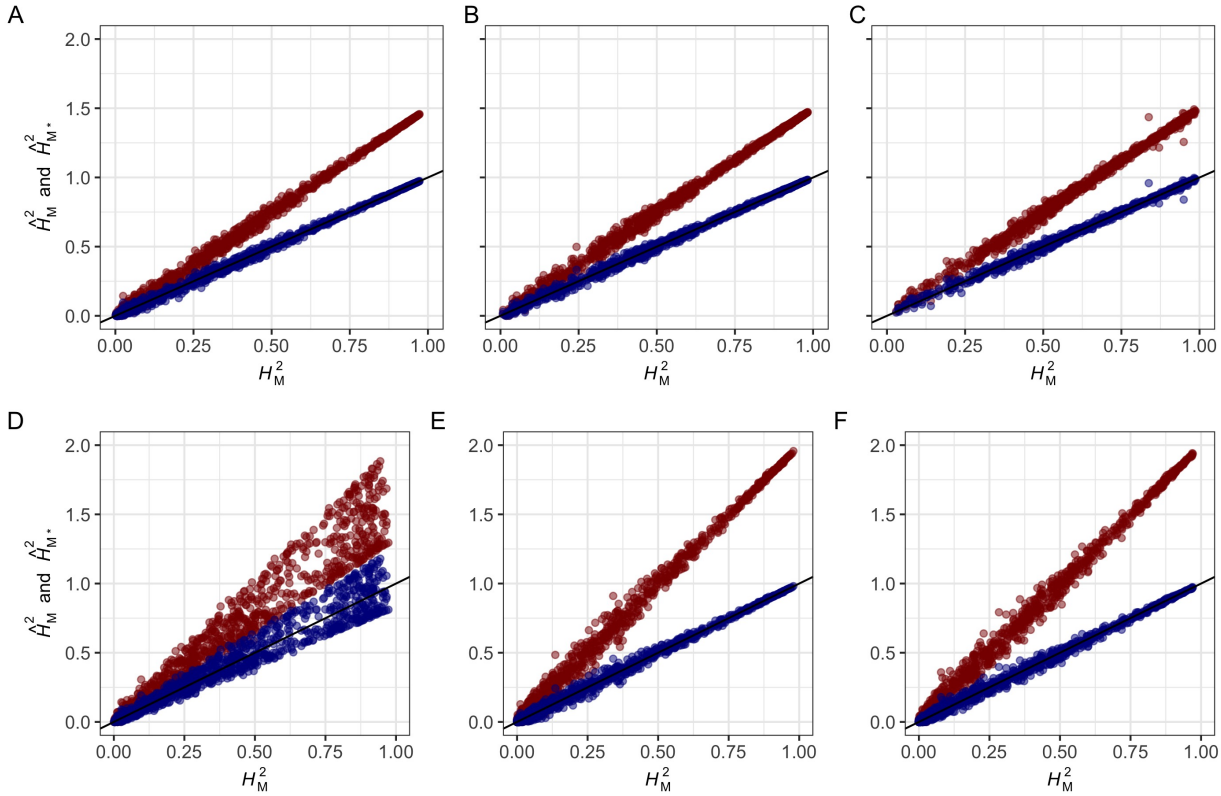

Fig 1. Accuracy of AMV and ASV Estimators of Marker Heritability. AMV and ASV estimates of H^2_M are shown for 1,000 segregating populations simulated for different numbers of entries (n_G individuals, families, or strains), five replications/entry (r_G = 5), true marker heritability (H^2_M) ranging from 0 to 1, and one to three marker loci with three genotypes/marker locus (n_{M1} = 3). AMV estimates of marker heritability ($\hat{H}^2_M$; red highlighted observations) and ASV estimates of marker heritability ($\hat{H}^2_{M*}$; 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) an population segregating 1:2:1 for one 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 Table S1.

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^2_M are shown in Appendix 3 (Table 1). The AMV-REML estimates of p and H^2_M were nearly double the bias-corrected ASV-REML estimates, e.g., the AMV-REML estimate of H^2_M 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).

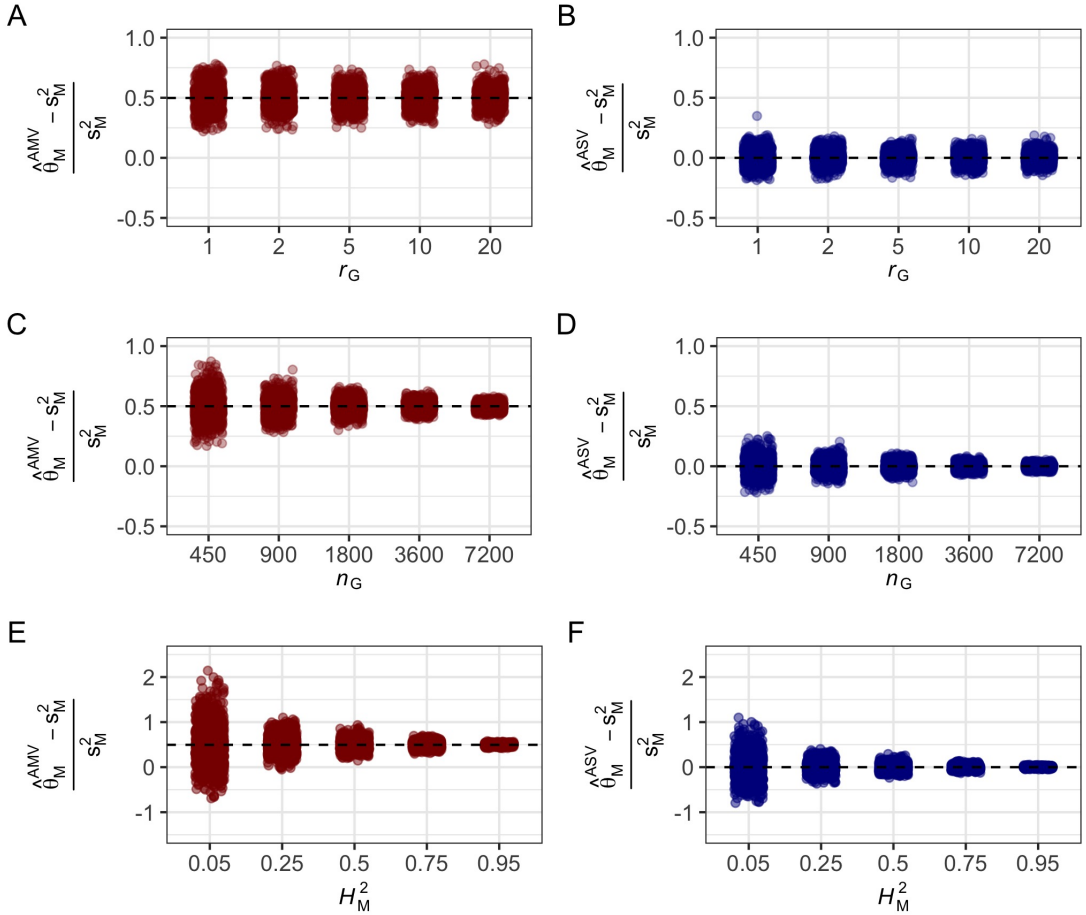

Fig 2. Effect of r_G, n_G, and H^2_M on the Relative Bias of AMV and ASV Estimators of σ^2_M. (A-B) Phenotypic observations were simulated for 1,000 populations segregating for a single marker locus with three genotypes ($n_M = 3$), $n_G = 900$ progeny, and r_G = 1, 2, 5, 10, or 20 (study designs 7-11). The marker locus was assumed to be in complete linkage disequilibrium with a single QTL that explains 50% of the phenotypic variance ($H^2_M = 0.50$). (A) Distribution of the relative biases of AMV estimates of σ^2_M for different r_G. The relative bias $RB[\theta^{AMV}_M] = 0.498$ was identical for different r_G. (B) Distribution of the relative biases of ASV estimates of σ^2_M for different r_G. The relative bias $RB[\theta^{ASV}_M] = 0.00$ was identical for different r_G. (C-D) Phenotypic observations were simulated for 1,000 populations segregating for a single marker locus with three genotypes ($n_M = 3$), five replications/entry ($r_G = 5$), and n_G = 450, 900, 1,800, 3,600, or 7,200 entries/population (study designs 12-16). The marker locus was assumed to be in complete linkage disequilibrium with a single QTL that explains 50% of the phenotypic variance ($H^2_M = 0.50$). (C) Distribution of the relative biases of AMV estimates of σ^2_M for different n_G. The relative bias $RB[\theta^{AMV}_M] = 0.499$ was identical across the variables tested. (D) Distribution of the relative biases of ASV estimates of σ^2_M for different n_G. The relative bias ($RB[\theta^{ASV}_M] = 0.00$) was identical across the variables tested. (E-F) Phenotypic observations were simulated for 1,000 populations segregating for a single marker locus with three genotypes ($n_M = 3$), five replications/entry ($r_G = 5$), and $n_G = 450$ entries/population. The marker locus was assumed to be in complete linkage disequilibrium with a single QTL that explains 5-95% of the phenotypic variance ($H^2_M = 0.05$ to 0.95 (study designs 17-21). (E) Distribution of the relative biases of AMV estimates of σ^2_M for different H^2_M. The relative bias $RB[\theta^{AMV}_M] = 0.496$ was identical across the variables tested. (F) Distribution of the relative biases of ASV estimates of σ^2_M for different H^2_M. The relative bias $RB[\theta^{ASV}_M] = 0.0$ was identical across the variables tested.

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—σ_G^2 and p could not be estimated for this example because individuals were unreplicated. Variance components were estimated for three SNP markers ($rs10$, $rs45$, and $rs20$) 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 $rs10$, 78AA: 736Aa: $2,159aa$ for $rs45$, and 237AA: 976Aa: $1,760aa$ for $rs20$. The k_M for these data ($k_{rs10} = 0.35$, $k_{rs45} = 0.41$, and $k_{20} = 0.54$) were calculated as shown in Appendix 3. The uncorrected AMV-REML estimate of H_M^2 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, 105]. The candidate marker loci are nearly always initially identified by genome-wide searches using sequential (marker-by-marker) approaches [56, 72, 75, 79, 106, 107]. Complicated and often misunderstood problems arise in the estimation and interpretation of statistics from sequential fixed effect analyses when the data are unbalanced [79, 108, 109]. 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, 108–110]. 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, 109]. 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, 106–109]. 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, 109], as shown in our multilocus analysis examples (Table 1-2). 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, 108, 109]. Although methods for fixed effect analyses of factorial treatment designs (multilocus genetic models) with unbalanced data are well known [56, 79, 106, 107, 109], 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_M^2) from Type I, II, and III analyses were substantially different from each other and from H_M^2 estimates from the random effects analysis (Table 1-2). 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 (p and H_M^2) 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, 109, 110]. 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 1-2).

As substantiated by our simulation analyses (Fig 1-2), ASV with REML estimation of the underlying variance components yields accurate estimates of p and H^2_M for marker loci and interactions between marker loci, both individually and collectively, and BLUPs of the the additive and dominance effects of marker loci [76, 110–112]. 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, 113]. 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 *rs*10, 4,880.5 to 9,504.4 for *rs*20, and 4,259.6 to 8,384.8 for *rs*45. 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 [106, 107]; 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, 114–118]. 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, 106, 107, 109]. 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 (Table S1; Appendix 4). 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 (Table S1). 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 (Table S1). 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 [119] as described by [120, 121]. 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; Table S1). 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 Appendix 1). 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; Table S1; Appendix 4). The variance components were estimated using REML in the *lme4::lmer()* v1.1-21 [78] package in R v4.0.2 [119]. We

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.

		Type I SS[a]							
Study	Source	ABC	ACB	BAC	BCA	CAB	CBA	Type II SS	Type III SS
Cattle	$rs10$	3,552.3	3,552.3	1,707.2	591.4	1,208.7	591.4	542.5	22.1
White Spotting[b]	$rs45$	6,539.7	4,259.6	8,384.8	8,384.8	4,259.6	4,876.8	4,282.7	1,394.9
	$rs20$	4,880.5	7,160.7	4,880.5	5,996.4	9,504.4	9,504.4	4,834.3	1,788.4
	$rs10 \times rs45$	12.7	12.7	12.7	12.7	12.7	12.7	14.3	47.4
	$rs10 \times rs20$	132.7	132.7	132.7	132.7	132.7	132.7	107.4	234.0
	$rs45 \times rs20$	193.1	193.1	193.1	193.1	193.1	193.1	193.1	91.5
	$rs10 \times rs45 \times rs20$	143.5	143.5	143.5	143.5	143.5	143.5	143.5	143.5
	$G : rs10 \times rs45 \times rs20$	15,512.9	15,512.9	15,512.9	15,512.9	15,512.9	15,512.9	15,512.9	15,512.9
Sunflower	BR	1,624.0	1,624.0	1,708.8	1,904.0	1,829.7	1,904.0	1,881.4	1,711.2
Oil Content[c]	PHY	298.2	254.2	213.4	213.4	254.3	180.0	220.2	208.3
	HYP	537.1	581.0	537.1	342.0	375.4	375.4	507.0	511.6
	$BR \times PHY$	57.9	57.9	57.9	57.9	57.9	57.9	49.7	50.0
	$BR \times HYP$	168.0	168.0	168.0	168.0	168.0	168.0	172.1	195.5
	$PHY \times HYP$	11.1	11.1	11.1	11.1	11.1	11.1	11.1	7.6
	$BR \times PHY \times HYP$	36.6	36.6	36.6	36.6	36.6	36.6	36.6	36.6
	$G : BR \times PHY \times HYP$	4,113.4	4,113.4	4,113.4	4,113.4	4,113.4	4,113.4	4113.4	4,113.4
	Residual	553.8	553.8	553.8	553.8	553.8	553.8	553.8	553.8

[a]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 $A = rs_{10}$, $B = rs_{45}$, and $C = rs_{20}$ 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: $A \times B$, $A \times C$, $B \times C$, and $A \times B \times C$. The three letters 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 $\mathrm{SS}(A|\mu)$, $\mathrm{SS}(B|A,\mu)$, and $\mathrm{SS}(C|A,B,\mu)$, 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 $\mathrm{SS}(C|\mu)$, $\mathrm{SS}(B|C,\mu)$, and $\mathrm{SS}(A|B,C,\mu)$. 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 $A \times B$ interaction was $\mathrm{SS}(A \times B|A,B,C,\mu)$ and for the three-way interaction was $\mathrm{SS}(A \times B \times C|A,B,C,A \times B,A \times C,B \times C,\mu)$.
[b]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 markers were identified by GWAS. The linear model for the cattle analysis was identical to the linear model for the sunflower analysis without replications ($r_G = 1$); 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 Appendix 4.
[c]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]. k_M coefficients for three loci with unbalanced data are shown in Appendix 4.

estimated the sample variances of AMV and ASV estimates of p for each study design (Table S1). Finally, we developed estimators of the sampling variances of p and H^2_M using the delta method [25, 122], as shown in Appendix 5.

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:

$$\begin{aligned} y_{hijkl} = &\mu + BR_h + PHY_i + HYP_j + BR \times PHY_{hi} + BR \times HYP_{hj} + PHY \times HYP_{ij} \\ &+ BR \times PHY \times HYP_{hij} + G : (BR \times PHY \times HYP)_{hij(k)} + \varepsilon_{hijkl} \end{aligned} \tag{27}$$

where BR_h is the hth effect of the BR locus, PHY_i is the ith effect of the PHY locus, HYP_j is the jth effect of the HYP locus, $G:(BR \times PHY \times HYP)_{hij(k)}$ is the kth effect of entries nested in the hij^{th} $BR \times PHY \times HYP$ interaction, and ε_{hijkl} is the $hijkl^{th}$ 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 ($n_{G:M}$) varied from $n_{G:BR} = 81:65$, $n_{G:PHY} = 60:86$, and $n_{G:HYP} = 70:76$. Variance components for LMMs (1) and (27) were estimated using the REML method in *lme4::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 Appendix 4.

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 *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 *lme4::lmer()* [78]. REML estimates of the marker-associated genetic variances for both marker loci were bias-corrected using the approach described in Appendix 2.

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:

$$\begin{aligned} y_{hijk} = & \mu + rs10_h + rs45_i + rs20_j + rs10 \times rs45_{hi} + rs10 \times rs20_{hj} + rs45 \times rs20_{ij} \\ & + rs10 \times rs45 \times rs20_{hij} + G:(rs10 \times rs45 \times rs20)_{hij(k)} \end{aligned} \tag{28}$$

where $rs10_h$ is the hth effect of the peak SNP ($rs109979909$) on chromosome 2, $rs45_i$ is the ith effect of the peak SNP on chromosome 6 ($rs451683615$), $rs20_j$ is the jth effect of the peak SNP on chromosome 22 ($rs209784468$), and $G:(rs10 \times rs45 \times rs20)_{hij(k)}$ is the $hij(k)^{th}$ 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 $\hat{s}_P^2 = 10.42$ from the log transformed data to use in the denominator of $\hat{H}_M^2$.

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, and 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 $\mid \mu$), R(B $\mid \mu$, A), and R(C $\mid \mu$, A, B), where μ 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 $\times$ B, A $\times$ C, B $\times$ C, A $\times$ B $\times$ 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, and A $\times$ B $\times$ C), R(B $\mid \mu$, A, C, A $\times$ B, A $\times$ C, B $\times$ C, and A $\times$ B $\times$ C), and R(C $\mid \mu$, A, B, A $\times$ B, A $\times$ C, B $\times$ C, and A $\times$ B $\times$ C).

Data Availability

The custom R scripts for reproducing our simulations have been deposited in a public GitHub repository [123]. The simulated data shown in Figs 1 and 2 have been deposited in a public Zenodo repository [124].

Abbreviations

AMV: average marginal variance; ANOVA: analysis of variance; ASV: average semivariance; GEBV: genome estimated breeding value; GWAS: genome-wide association study; LD: linkage disequilibrium; LSR: least squares

regression; MAS: marker-assisted selection; QTL: quantitative trait locus; REML: restricted maximum likelihood; SNP: single nucleotide polymorphism; WGR: whole-genome regression.

Acknowledgements

The authors thank the anonymous reviewers for comments and suggestions that improved the manuscript.

Conflict of interest

The authors declare no conflict of interest.

Funding

This research was supported by grants to SJK from the United Stated Department of Agriculture (`http://dx.doi.org/10.13039/100000199`) National Institute of Food and Agriculture (NIFA) Specialty Crops Research Initiative (# 2017-51181-26833) and California Strawberry Commission (`http://dx.doi.org/10.13039/100006760`). (`http://dx.doi.org/10.13039/100007707`). HPP was supported by the German Research Foundation (DFG) grant PI 377/18-1. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supporting information

Table S1 Simulation study designs and variables.

Figure S1 Accuracy of AMV and ASV Estimators of Marker Heritability When the Phenotypic Variance is Estimated by Pooling Marker and Residual Genetic Sources of Variation ($\sigma^2_M + \sigma^2_{G:M}$).

Figure S2 Relative Bias of AMV and ASV Estimators of Marker Heritability.

Author Contributions

Conceptualization: Steven J. Knapp, William C. Bridges, Mitchell J. Feldmann, Hans-Peter Piepho
Data curation: Mitchell J. Feldmann
Formal Analysis: Mitchell J. Feldmann, Hans-Peter Piepho, Steven J. Knapp, William C. Bridges
Funding Acquisition: Steven J. Knapp and Hans-Peter Piepho
Investigation: Mitchell J. Feldmann, Hans-Peter Piepho, Steven J. Knapp
Methodology: Mitchell J. Feldmann, Hans-Peter Piepho, Steven J. Knapp, William C. Bridges
Project administration: Steven J. Knapp
Resources: Steven J. Knapp and Mitchell J. Feldmann
Software: Mitchell J. Feldmann
Supervision: Steven J. Knapp and Mitchell J. Feldmann
Validation: William C. Bridges, Hans-Peter Piepho, Mitchell J. Feldmann, Steven J. Knapp
Visualization: Mitchell J. Feldmann, Steven J. Knapp, Hans-Peter Piepho
Writing – original draft preparation: Mitchell J. Feldmann, Steven J. Knapp, Hans-Peter Piepho
Writing – review & editing: Mitchell J. Feldmann, Hans-Peter Piepho, Steven J. Knapp, William C. Bridges

References

1. Lander ES, Schork NJ. Genetic dissection of complex traits. Science. 1994;265(5181):2037–2048.
2. Risch NJ. Searching for genetic determinants in the new millennium. Nature. 2000;405(6788):847–856.
3. Glazier AM, Nadeau JH, Aitman TJ. Finding genes that underlie complex traits. science. 2002;298(5602):2345–2349.
4. Consortium CT, et al. The nature and identification of quantitative trait loci: a community's view. Nature reviews Genetics. 2003;4(11):911.
5. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nature reviews genetics. 2005;6(2):95–108.
6. Hill WG. Understanding and using quantitative genetic variation. Philos Trans R Soc London, Ser B. 2010;365(1537):73–85.
7. Hill WG. Quantitative genetics in the genomics era. Curr Genomics. 2012;13(3):196–206.
8. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. American journal of human genetics. 1980;32(3):314.
9. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001;409(6822):928–934.
10. Huang X, Feng Q, Qian Q, Zhao Q, Wang L, Wang A, et al. High-throughput genotyping by whole-genome resequencing. Genome research. 2009;19(6):1068–1076.
11. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, et al. SNP detection for massively parallel whole-genome resequencing. Genome research. 2009;19(6):1124–1132.
12. Lander ES, Botstein D. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics. 1989;121(1):185–199.
13. Mackay TFC. The genetic architecture of quantitative traits. Annu Rev Genet. 2001;35(1):303–339.
14. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, et al. Complement factor H polymorphism in age-related macular degeneration. Science. 2005;308(5720):385–389.
15. Consortium WTCC, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661.
16. Mackay TFC, Stone EA, Ayroles JF. The genetics of quantitative traits: challenges and prospects. Nat Rev Genet. 2009;10(8):565.
17. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012;90(1):7–24.
18. Pasaniuc B, Price AL. Dissecting the genetics of complex traits using summary association statistics. Nat Rev Genet. 2017;18(2):117.
19. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet. 2017;101(1):5–22.
20. Meuwissen T, Hayes B, Goddard M. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157(4):1819–1829.

21. Wray NR, Kemper KE, Hayes BJ, Goddard ME, Visscher PM. Complex trait prediction from genome data: contrasting EBV in livestock to PRS in humans: genomic prediction. Genetics. 2019;211(4):1131–1141.

22. Crouch DJ, Bodmer WF. Polygenic inheritance, GWAS, polygenic risk scores, and the search for functional variants. Proceedings of the National Academy of Sciences. 2020;117(32):18924–18933.

23. Wray NR, Lin T, Austin J, McGrath JJ, Hickie IB, Murray GK, et al. From basic science to clinical application of polygenic risk scores: a primer. JAMA psychiatry. 2021;78(1):101–109.

24. Falconer D, Mackay T. Introduction to Quantitative Genetics. Harlow, Essex, UK. Longmans Green; 1996.

25. Lynch M, Walsh B. Genetics and analysis of quantitative traits. vol. 1. Sinauer Sunderland, MA; 1998.

26. Walsh B. Quantitative genetics in the age of genomics. Theoretical Population Biology. 2001;59(3):175–184.

27. Visscher PM, Medland SE, Ferreira MA, Morley KI, Zhu G, Cornes BK, et al. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genet. 2006;2(3).

28. Roff DA. A centennial celebration for quantitative genetics. Evolution. 2007;61(5):1017–1032.

29. Hill WG, Goddard ME, Visscher PM. Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet. 2008;4(2):e1000008.

30. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era—concepts and misconceptions. Nat Rev Genet. 2008;9(4):255–266.

31. Roff DA. Evolutionary quantitative genetics. Springer Science & Business Media; 2012.

32. Bernardo R. Reinventing quantitative genetics for plant breeding: something old, something new, something borrowed, something BLUE. Heredity. 2020;125(6):375–385.

33. Andersson L. Genetic dissection of phenotypic diversity in farm animals. Nature Reviews Genetics. 2001;2(2):130–138.

34. Hayes B, Goddard ME. The distribution of the effects of genes affecting quantitative traits in livestock. Genetics Selection Evolution. 2001;33(3):1–21.

35. Mackay TF. Quantitative trait loci in Drosophila. Nature reviews genetics. 2001;2(1):11–20.

36. Andersson L, Georges M. Domestic-animal genomics: deciphering the genetics of complex traits. Nature Reviews Genetics. 2004;5(3):202–212.

37. Anderson JA, Chao S, Liu S. Molecular breeding using a major QTL for Fusarium head blight resistance in wheat. Crop Science. 2007;47:S–112.

38. Septiningsih EM, Pamplona AM, Sanchez DL, Neeraja CN, Vergara GV, Heuer S, et al. Development of submergence-tolerant rice cultivars: the Sub1 locus and beyond. Annals of Botany. 2009;103(2):151–160.

39. Lorenz K, Cohen BA. Small-and large-effect quantitative trait locus interactions underlie variation in yeast sporulation efficiency. Genetics. 2012;192(3):1123–1132.

40. Saatchi M, Schnabel RD, Taylor JF, Garrick DJ. Large-effect pleiotropic or closely linked QTL segregate within and across ten US cattle breeds. BMC Genomics. 2014;15(1):1–17.

41. Huang H, Cao J, Hanif Q, Wang Y, Yu Y, Zhang S, et al. Genome-wide association study identifies energy metabolism genes for resistance to ketosis in Chinese Holstein cattle. Animal genetics. 2019;50(4):376–380.

42. Freebern E, Santos DJ, Fang L, Jiang J, Gaddis KLP, Liu GE, et al. GWAS and fine-mapping of livability and six disease traits in Holstein cattle. BMC genomics. 2020;21(1):1–11.

43. Li B, VanRaden P, Null D, O'Connell J, Cole J. Major quantitative trait loci influencing milk production and conformation traits in Guernsey dairy cattle detected on Bos taurus autosome 19. Journal of Dairy Science. 2021;104(1):550–560.

44. Bernardo R. Molecular markers and selection for complex traits in plants: learning from the last 20 years. Crop Sci. 2008;48(5):1649–1664.

45. Bernardo R. Bandwagons I, too, have known. Theor Appl Genet. 2016;129(12):2323–2332.

46. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747.

47. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010;11(6):446–450.

48. Young AI. Solving the missing heritability problem. PLoS Genet. 2019;15(6):e1008222.

49. Bernardo R. What if we knew all the genes for a quantitative trait in hybrid crops? Crop Sci. 2001;41(1):1–4.

50. Hill WG. Applications of population genetics to animal breeding, from Wright, Fisher and Lush to genomic prediction. Genetics. 2014;196(1):1–16.

51. De Villemereuil P, Morrissey MB, Nakagawa S, Schielzeth H. Fixed-effect variance and the estimation of repeatabilities and heritabilities: issues and solutions. Journal of Evolutionary Biology. 2018;31(4):621–632.

52. Gaudet MM, Kirchhoff T, Green T, Vijai J, Korn JM, Guiducci C, et al. Common genetic variants and modification of penetrance of BRCA2-associated breast cancer. PLoS Genet. 2010;6(10):e1001183.

53. Tang S, Leon A, Bridges WC, Knapp SJ. Quantitative trait loci for genetically correlated seed traits are tightly linked to branching and pericarp pigment loci in sunflower. Crop Sci. 2006;46(2):721–734.

54. Hayes BJ, Pryce J, Chamberlain AJ, Bowman PJ, Goddard ME. Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. PLoS genetics. 2010;6(9):e1001139.

55. Seabury CM, Oldeschulte DL, Saatchi M, Beever JE, Decker JE, Halley YA, et al. Genome-wide association study for feed efficiency and growth traits in US beef cattle. BMC genomics. 2017;18(1):1–25.

56. Littell RC. Analysis of unbalanced mixed model data: a case study comparison of ANOVA versus REML/GLS. Journal of Agricultural, Biological, and Environmental Statistics. 2002;7(4):472.

57. Daetwyler HD, Calus MP, Pong-Wong R, de Los Campos G, Hickey JM. Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics. 2013;193(2):347–365.

58. Gianola D. Priors in whole-genome regression: the Bayesian alphabet returns. Genetics. 2013;194(3):573–596.

59. Moore JK, Manmathan HK, Anderson VA, Poland JA, Morris CF, Haley SD. Improving Genomic Prediction for Pre-Harvest Sprouting Tolerance in Wheat by Weighting Large-Effect Quantitative Trait Loci. Crop Science. 2017;57(3):1315–1324.

60. Rice B, Lipka AE. Evaluation of RR-BLUP Genomic Selection Models that Incorporate Peak Genome-Wide Association Study Signals in Maize and Sorghum. The Plant Genome. 2019;12(1).

61. Spindel J, Begum H, Akdemir D, Collard B, Redoña E, Jannink J, et al. Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. Heredity. 2016;116(4):395–408.

62. Lande R, Thompson R. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics. 1990;124(3):743–756.

63. de los Campos G, Sorensen D, Gianola D. Genomic heritability: what is it? PLoS Genet. 2015;11(5):e1005048.

64. Beavis WD. QTL analyses: power, precision, and accuracy. In AH Patterson (ed) Molecular Dissection of Complex Traits. 1998; p. 145–162.

65. Melchinger AE, Utz HF, Schön CC. Quantitative trait locus (QTL) mapping using different testers and independent population samples in maize reveals low power of QTL detection and large bias in estimates of QTL effects. Genetics. 1998;149(1):383–403.

66. Utz HF, 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 validation and validation with independent samples. Genetics. 2000;154(4):1839–1849.

67. Allison DB, Fernandez JR, Heo M, Zhu S, Etzel C, Beasley TM, et al. Bias in estimates of quantitative-trait–locus effect in genome scans: demonstration of the phenomenon and a method-of-moments procedure for reducing bias. Am J Hum Genet. 2002;70(3):575–585.

68. Xu S. Theoretical basis of the Beavis effect. Genetics. 2003;165(4):2259–2268.

69. Bernardo R. What proportion of declared QTL in plants are false? Theor Appl Genet. 2004;109(2):419–424.

70. Göring HH, Terwilliger JD, Blangero J. Large upward bias in estimation of locus-specific effects from genomewide scans. Am J Hum Genet. 2001;69(6):1357–1369.

71. Henderson CR. Estimation of variance and covariance components. Biometrics. 1953;9(2):226–252.

72. Searle SR, Gruber MH. Linear models. Wiley Online Library; 1971.

73. Piepho HP. A coefficient of determination (R^2) for generalized linear mixed models. Biom J. 2019;61:860–872.

74. Searle SR. Linear models for unbalanced data. Wiley; 1987.

75. Littell RC, Milliken GA, Stroup WW, Wolfinger RD, Schabenberger O. SAS system for mixed models. vol. 633. SAS institute Cary, NC; 1996.

76. Bernardo R. Breeding for quantitative traits in plants. vol. 1. Stemma press Woodbury, MN; 2002.

77. Inc SI. SAS/STAT 13.1 User's Guide: Chapter 43 - The GLIMMIX Procedure. Author Cary, NC; 2013. Available from: `https://support.sas.com/documentation/onlinedoc/stat/131/glimmix.pdf`.

78. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. J Stat Softw. 2015;67(1):1–48. doi:10.18637/jss.v067.i01.

79. Gbur EE, Stroup WW, McCarter KS, Durham S, Young LJ, Christman M, et al. Analysis of generalized linear mixed models in the agricultural and natural resources sciences. vol. 156. John Wiley & Sons; 2020.

80. Broman KW, Sen S. A Guide to QTL Mapping with R/QTL. vol. 46. Springer; 2009.

81. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. The American Journal of Human Genetics. 2010;86(1):6–22.

82. Beavis W, Smith O, Grant D, Fincher R. Identification of quantitative trait loci using a small sample of topcrossed and F4 progeny from maize. Crop Sci. 1994;34(4):882–896.

83. Luo L, Mao Y, Xu S. Correcting the bias in estimation of genetic variances contributed by individual QTL. Genetica. 2003;119(2):107–114.

84. Zhang J, Yue C, Zhang Y. Bias correction for estimated QTL effects using the penalized maximum likelihood method. Heredity. 2012;108(4):396–402.

85. Jivanji S, Worth G, Lopdell TJ, Yeates A, Couldrey C, Reynolds E, et al. Genome-wide association analysis reveals QTL and candidate mutations involved in white spotting in cattle. Genet Sel Evol. 2019;51(1):62.

86. Pincot DD, Poorten TJ, Hardigan MA, Harshman JM, Acharya CB, Cole GS, et al. Genome-wide association mapping uncovers Fw1, a dominant gene conferring resistance to Fusarium wilt in strawberry. G3: Genes, Genomes, Genet. 2018;8(5):1817–1828.

87. Conner JK, Hartl DL, et al. A primer of ecological genetics. Sinauer Associates Incorporated; 2004.

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

89. Choy Y, Brinks J, Bourdon R. Repeated-measure animal models to estimate genetic components of mature weight, hip height, and body condition score. Journal of animal science. 2002;80(8):2071–2077.

90. Dekkers JC. Commercial application of marker-and gene-assisted selection in livestock: strategies and lessons. J Anim Sci. 2004;82(suppl_13):E313–E328.

91. Dekkers J. Prediction of response to marker-assisted and genomic selection using selection index theory. J Anim Breed Genet. 2007;124(6):331–341.

92. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91(11):4414–4423.

93. Sun X, Habier D, Fernando RL, Garrick DJ, Dekkers JC. Genomic breeding value prediction and QTL mapping of QTLMAS2010 data using Bayesian methods. In: BMC proceedings. BioMed Central; 2011. p. 1–8.

94. Isik F, Holland J, Maltecca C. Genetic data analysis for plant and animal breeding. Springer; 2017.

95. Astle W, Balding DJ. Population structure and cryptic relatedness in genetic association studies. Statistical Science. 2009;24(4):451–471.

96. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42(7):565.

97. Endelman JB, Jannink JL. Shrinkage estimation of the realized relationship matrix. G3: Genes, Genomes, Genet. 2012;2(11):1405–1413.

98. Cary N. SAS/STAT 13.1 User's Guide; 2013.

99. Piepho HP, Moehring J, Schulz-Streeck T, Ogutu JO. A stage-wise approach for the analysis of multi-environment trials. Biom J. 2012;54(6):844–860.

100. Damesa TM, Möhring J, Worku M, Piepho HP. One step at a time: stage-wise analysis of a series of experiments. Agron J. 2017;109(3):845–857.

101. Damesa TM, Hartung J, Gowda M, Beyene Y, Das B, Semagn K, et al. Comparison of Weighted and Unweighted Stage-Wise Analysis for Genome-Wide Association Studies and Genomic Selection. Crop Sci. 2019;59:2572–2584.

102. Schmidt P, Hartung J, Bennewitz J, Piepho HP. Heritability in Plant Breeding on a Genotype-Difference Basis. Genetics. 2019;212(4):991–1008.

103. Efron B, Tibshirani R. The bootstrap method for assessing statistical accuracy. Behaviormetrika. 1985;12(17):1–35.

104. Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. Plant Methods. 2013;9(1):29.

105. Jensen J, Su G, Madsen P. Partitioning additive genetic variance into genomic and remaining polygenic components for complex traits in Dairy cattle. BMC Genet. 2012;13(1):44.

106. Herr DG. On the history of ANOVA in unbalanced, factorial designs: The first 30 years. The American Statistician. 1986;40(4):265–270.

107. Langsrud Ø. ANOVA for unbalanced data: Use Type II instead of Type III sums of squares. Statistics and Computing. 2003;13(2):163–167.

108. Stroup W, Littelll R. Impact of variance component estimates on fixed effect inference in unbalanced linear mixed models. Proceedings of the Kansas State University Conference on Applied Statistics in Agriculture. 2002;14:32–48.

109. Stroup WW, Milliken GA, Claassen EA, Wolfinger RD. SAS for mixed models: introduction and basic applications. SAS Institute; 2018.

110. Piepho H, Möhring J, Melchinger A, Büchse A. BLUP for phenotypic selection in plant breeding and variety testing. Euphytica. 2008;161(1):209–228.

111. Searle SR, Casella G, McCulloch C. Variance components. John Wiley & Sons; 1992.

112. 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.

113. Hector A, Von Felten S, Schmid B. Analysis of variance with unbalanced data: an update for ecology & evolution. Journal of animal ecology. 2010;79(2):308–316.

114. Gianola D, Perez-Enciso M, Toro MA. On marker-assisted prediction of genetic value: beyond the ridge. Genetics. 2003;163(1):347–365.

115. Goddard M, Hayes B. Genomic selection. J Anim Breed Genet. 2007;124(6):323–330.

116. Habier D, Fernando RL, Garrick DJ. Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics. 2013;194(3):597–607.

117. Meuwissen T, Hayes B, Goddard M. Genomic selection: A paradigm shift in animal breeding. Animal Frontiers. 2016;6(1):6–14.

118. Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de los Campos G, et al. Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. 2017;22(11):961–975.

119. R Core Team. R: A Language and Environment for Statistical Computing; 2020. Available from: `https://www.R-project.org/`.

120. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Stat Med. 2006;25(24):4279–4292.

121. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38(11):2074–2102.

122. Johnson NL, Kemp AW, Kotz S. Univariate discrete distributions. vol. 444. John Wiley & Sons; 2005.

123. Feldmann MJ. VarCompSim; 2020. `https://github.com/mjfeldmann/VarCompSim`.

124. Feldmann MJ, Piepho HP, Bridges WC, Knapp SJ. Data for "Accurate Estimation of Marker-Associated Genetic Variance and Heritability in Complex Trait Analyses". Zenodo. 2020;doi:10.5281/zenodo.3742421.

125. Estaghvirou SBO, Ogutu JO, Schulz-Streeck T, Knaak C, Ouzunova M, Gordillo A, et al. Evaluation of approaches for estimating the accuracy of genomic prediction in plant breeding. BMC Genomics. 2013;14(1):860.

126. Oehlert GW. A note on the delta method. The American Statistician. 1992;46(1):27–29.

Appendices

A1 ASV Estimator of the Fraction of the Genetic Variance Associated With a Single Marker Locus for Unbalanced Data

We developed AMV and ASV estimators of p and H^2_M for analyses of balanced data because of their algebraic simplicity; however, phenotypic and genotypic data are nearly always unbalanced in practice. Here we extend the bias-correction solutions developed for balanced data to unbalanced data. ASV estimators of σ^2_M and $\sigma^2_{G:M}$ are developed here for a single marker locus (M) with unbalanced data. The phenotypic observations are entry-means ($\bar{y}_{hi\bullet}$).

The ASV estimator of the genetic variance associated with marker locus M for LMM (2) in its general form is:

$$\hat{\theta}^{ASV}_M = (n_G - 1)^{-1}\hat{\sigma}^2_M tr(Z_{u_M} Z^T_{u_M} D_{n_G}) = \frac{n_G - n_G^{-1}\sum_h n^2_{G:M_h}}{df_G}\hat{\sigma}^2_M = k_M\hat{\sigma}^2_M \tag{A1}$$

where $\hat{\sigma}^2_M$ is the genetic variance associated with the marker locus effect, df_G is the degrees of freedom for entries, and $Z_{u_M} = \oplus_h 1_{n_{G:M_h}}$. $\hat{\theta}^{ASV}_M$ is a k_M-bias corrected estimate of the marker-associated genetic variance. This form of k_M can be thought of as the general form of the correction term relative to (12) which is specific to balanced study designs.

The ASV estimate of the residual genetic variance among entries nested in the marker locus M is:

$$\hat{\theta}^{ASV}_{G:M} = (n_G - 1)^{-1}\hat{\sigma}^2_{G:M} tr(Z_{u_{G:M}} Z^T_{u_{G:M}} D_{n_G}) = \frac{n_G - 1}{n_G - 1}\hat{\sigma}^2_{G:M} = \hat{\sigma}^2_{G:M} \tag{A2}$$

where $\sigma^2_{G:M}$ is the residual genetic variance among entries nested in the marker locus effect and $Z_{u_{G:M}}$ is a n_G identity matrix.

From equation (A1), the ASV estimator of p for a single marker locus (M) with unbalanced data is:

$$\hat{p}_* = \frac{\hat{\theta}^{ASV}_M}{\hat{\theta}^{ASV}_G} = \frac{k_M\hat{\theta}^{AMV}_M}{\hat{\theta}^{AMV}_G} = \frac{k_M\hat{\sigma}^2_M}{\hat{\sigma}^2_G} = \frac{n_G - n_G^{-1}\sum_h n^2_{G:M_h}}{df_G} \times \frac{\hat{\sigma}^2_M}{\hat{\sigma}^2_G} \tag{A3}$$

Hence, ASV yields unbiased estimates of the fraction of the genetic variance associated with the M locus (p) for unbalanced data. AMV estimates of p can be bias-corrected using the k_M coefficient shown above for unbalanced data. We review this here because most of the currently available linear mixed model software solutions produce AMV estimates, e.g., REML estimates of σ^2_M from LMM (2) are AMV estimates.

A2 ASV Estimator of the Fraction of the Genetic Variance Associated with Two Marker Loci for Unbalanced Data

ASV estimators of σ^2_M and $\sigma^2_{G:M}$ are developed here for two marker loci ($M1$ and $M2$). The phenotypic observations are entry-means ($\bar{y}_{hij\bullet}$), the underlying data are unbalanced, and the LMM for the entry-mean analysis is:

$$\bar{y}_{hij\bullet} = \mu + M1_h + M2_i + M1 \times M2_{hi} + G : M_{hi(j)} + \bar{\varepsilon}_{hij\bullet} \tag{A4}$$

where $\bar{y}_{hij\bullet}$ is the entry-mean, μ is the population mean, h = 1, 2, or 3, i = 1, 2, or 3, j = 1, 2, ..., n_G, k = 1, 2, ..., r_G, $M1_h$ is the random effect of marker locus 1 with $var(M1_h) = \sigma^2_{M1}$, $M2_i$ is the random effect of marker locus 2 with $var(M2_i) = \sigma^2_{M2}$, $(M1 \times M2)_{hi}$ is the random effect of the interaction between marker loci 1 and 2 with $var(M1 \times M2_{hi}) = \sigma^2_{M1\times M2}$, $G : M_{hi(j)}$ is the random effect of entries nested in marker loci with $var(G : M_{hi(j)}) = \sigma^2_{G:M}$, M refers to the $M1 \times M2$ interaction, and $\bar{\varepsilon}_{hij\bullet}$ is the residual with $var(\bar{\varepsilon}_{hij\bullet}) = r_G^{-1}\sigma^2_\varepsilon$.

The ASV estimator of the genetic variance associated with marker locus $M1$ from LMM (A4) is:

$$\hat{\theta}_{M1}^{ASV} = (n_G - 1)^{-1}\hat{\sigma}_{M1}^2 tr(Z_{u_{M1}} Z_{u_{M1}}^T D_{n_G}) = \frac{n_G - n_G^{-1}\sum_h n_{G:M1_h}^2}{df_G}\hat{\sigma}_{M1}^2 = k_{M1}\hat{\sigma}_{M1}^2 \tag{A5}$$

where $Z_{u_{M1}} = \oplus_h 1_{n_{G:M1_h}}$ as is the incidence matrix for $M1$, σ_{M1}^2 is the genetic variance associated with marker locus $M1$, k_{M1} is the bias-correction factor for $M1$. Similarly, the ASV estimator of the genetic variance associated with marker locus $M2$ from LMM (A4) is:

$$\hat{\theta}_{M2}^{ASV} = (n_G - 1)^{-1}\hat{\sigma}_{M2}^2 tr(Z_{u_{M2}} Z_{u_{M2}}^T D_{n_G}) = \frac{n_G - n_G^{-1}\sum_i n_{G:M2_i}^2}{df_G}\hat{\sigma}_{M2}^2 = k_{M2}\hat{\sigma}_{M2}^2 \tag{A6}$$

where $Z_{u_{M2}} = \oplus_i 1_{n_{G:M2_i}}$ is the incidence matrix for $M2$.

The ASV estimator of the genetic variance associated with the interaction between marker loci $M1$ and $M2$ is:

$$\hat{\theta}_{M1\times M2}^{ASV} = (n_G - 1)^{-1}\hat{\sigma}_{M1\times M2}^2 tr(Z_{u_{M1\times M2}} Z_{u_{M1\times M2}}^T D_{n_G}) = \frac{n_G - n_G^{-1}\sum_{hi} n_{G:M_{hi}}^2}{df_G}\hat{\sigma}_{M1\times M2}^2 = k_{M1\times M2}\hat{\sigma}_{M1\times M2}^2 \tag{A7}$$

where $Z_{u_{M1\times M2}} = \oplus_{hi} 1_{n_{M1\times M2_{hi}}}$ is the incidence matrix for the $M1$ by $M2$ interaction, $I_{n_{M1\times M2}}$ is a $n_{M1\times M2}$ identity matrix, $1_{n_{G:M}}$ is a , $k_{M1\times M2}$ is the bias-correction factor for the interaction, $\hat{\sigma}_{M1\times M2}^2$ is the genetic variance associated with the interaction, and $\theta_{M1\times M2}^{ASV} \neq \sigma_{M1\times M2}^2$.

The ASV estimator of the variance associated with the residual genetic variance among entries nested in marker loci $M1$ and $M2$ is:

$$\hat{\theta}_{G:M}^{ASV} = (n_G - 1)^{-1}\hat{\sigma}_{G:M}^2 tr(Z_{u_{G:M}} Z_{u_{G:M}}^T D_{n_G}) = \frac{n_G - 1}{n_G - 1}\hat{\sigma}_{G:M}^2 = \hat{\sigma}_{G:M}^2 \tag{A8}$$

where $Z_{u_{G:M}} = I_{n_G}$ is a n_G incidence matrix.

Hence, as for the two-marker analysis, ASV yields k_M-bias corrected estimates of the marker-associated genetic variance for two loci ($M1$ and $M2$) for unbalanced data:

$$\hat{p}_* = \frac{\hat{\theta}_{M1}^{ASV} + \hat{\theta}_{M2}^{ASV} + \hat{\theta}_{M1\times M2}^{ASV}}{\hat{\theta}_G^{ASV}} = \frac{k_{M1}\hat{\sigma}_{M1}^2 + k_{M2}\hat{\sigma}_{M2}^2 + k_{M1\times M2}\hat{\sigma}_{M1\times M2}^2}{\hat{\sigma}_G^2} \tag{A9}$$

The genetic variance associated with marker loci can be partitioned into intra-locus (additive and dominance) and inter-locus (additive $\times$ additive, additive $\times$ dominance, and dominance $\times$ dominance) components [24, 25]. The example shown here includes all components: two degrees of freedom each for $M1$ and $M2$ and four degrees of freedom for $M1 \times M2$.

From (A5) and (A6), the coefficients for bias-correcting AMV estimates of σ_{M1}^2 (the genetic variance explained by $M1$) and σ_{M2}^2 (the genetic variance explained by $M2$) are:

$$k_{M1} = \frac{n_G - n_G^{-1}\sum_h n_{G:M1_h}^2}{df_G} \tag{A10}$$

$$k_{M2} = \frac{n_G - n_G^{-1}\sum_i n_{G:M2_i}^2}{df_G} \tag{A11}$$

Similarly, from (A7), the coefficient for bias-correcting AMV estimates of $\sigma_{M1\times M2}^2$ (the genetic variance explained by the interaction between $M1$ and $M2$) is:

$$k_{M1\times M2} = \frac{n_G - n_G^{-1}\sum_{hi} n_{G:M1\times M2_{hi}}^2}{df_G} \tag{A12}$$

These k_M-coefficients (A10-A12) can be substituted in (A9) to obtain bias-corrected estimates of p.

A3 ASV Estimator of the Fraction of the Genetic Variance Associated with Three Marker Loci for Unbalanced Data

ASV estimators of the genetic variance associated with intra-locus and inter-locus effects of three marker loci are developed here for unbalanced data. As before, the phenotypic observations are entry-means ($\bar{y}_{hijk\bullet}$) and the LMM for the entry-mean analysis is:

$$\begin{aligned}\bar{y}_{hijk\bullet} =& \mu + M1_h + M2_i + M2_j + M1\times M2_{hi} + M1\times M3_{hj} + M2\times M3_{ij} \\ &+ M1\times M2\times M3_{hij} + G:M_{hij(k)} + \bar{\varepsilon}_{hijk\bullet}\end{aligned} \tag{A13}$$

where $\bar{y}_{hijk\bullet}$ is the entry-mean, μ is the population mean, h = 1, 2, or 3, i = 1, 2, or 3, j = 1, 2, ..., n_G, k = 1, 2, ..., r_G, $M1_h$ is the random effect of marker locus 1 with $var(M1_h) = \sigma^2_{M1}$, $M2_i$ is the random effect of marker locus 2 with $var(M2_i) = \sigma^2_{M2}$, $M3_j$ is the random effect of marker locus 2 with $var(M3_j) = \sigma^2_{M3}$, $M1\times M2_{hi}$ is the random effect of the interaction between marker loci 1 and 2 with $var(M1\times M2_{hi}) = \sigma^2_{M1\times M2}$, $M1\times M3_{hj}$ is the random effect of the interaction between marker loci 1 and 3 with $var(M1\times M3_{hj}) = \sigma^2_{M1\times M3}$, $M2\times M3_{ij}$ is the random effect of the interaction between marker loci 2 and 3 with $var(M2\times M3_{ij}) = \sigma^2_{M2\times M3}$, $M1\times M2\times M3_{hij}$ is the random effect of the interaction between marker loci 1, 2, and 3 with $var(M1\times M2\times M3_{hij}) = \sigma^2_{M1\times M2\times M3}$, $G:M_{hij(k)}$ is the random effect of entries nested in marker loci with $var(G:M_{hij(k)}) = \sigma^2_{G:M)}$, M refers to the $M1\times M2\times M3$ interaction, and $\bar{\varepsilon}_{hijk\bullet}$ is the residual with $var(\bar{\varepsilon}_{hijk\bullet}) = r_G^{-1}\sigma^2_{\varepsilon}$.

The ASV estimator of the genetic variance associated with marker locus $M1$ from LMM (A13) is:

$$\hat{\theta}^{ASV}_{M1} = (n_G-1)^{-1}\hat{\sigma}^2_{M1}tr(Z_{u_{M1}}Z^T_{u_{M1}}D_{n_G}) = \frac{n_G - n_G^{-1}\sum_h n^2_{G:M1_h}}{df_G}\hat{\sigma}^2_{M1} \tag{A14}$$

where $P = I_{n_G} - n_G^{-1}J_{n_G}$, J_{n_G} is a n_G unit matrix, $Z_{u_{M1}} = \oplus_h 1_{n_{G:M1_h}}$ is the incidence matrix for $M1$. Substituting the appropriate values for marker loci $M2$ or $M3$ into (A14) yields the ASV estimators for these loci: $\hat{\theta}^{ASV}_{M2}$ and $\hat{\theta}^{ASV}_{M3}$.

The ASV estimator of the genetic variance associated with the two-locus interaction between marker loci $M1$ and $M2$ is:

$$\hat{\theta}^{ASV}_{M1\times M2} = (n_G-1)^{-1}\hat{\sigma}^2_{M1\times M2}tr(Z_{u_{M1\times M2}}Z^T_{u_{M1\times M2}}D_{n_G}) = \frac{n_G - n_G^{-1}\sum_{hi} n^2_{G:M1\times M2_{hi}}}{df_G}\hat{\sigma}^2_{M1\times M2} \tag{A15}$$

where $Z_{u_{M1:M2}} = \oplus_{hi} 1_{n_{M1:M2_{hi}}}$ is the incidence matrix for $M1\times M2$ interaction. ASV estimates of the other two-locus interactions are similarly defined. The ASV estimator of the genetic variance associated with the three-locus interaction ($M1\times M2\times M3$) is:

$$\hat{\theta}^{ASV}_{M1\times M2\times M3} = (n_G-1)^{-1}\hat{\sigma}^2_{M1\times M2\times M3}tr(Z_{u_{M1\times M2\times M3}}Z^T_{u_{M1\times M2\times M3}}D_{n_G}) = \frac{n_G - n_G^{-1}\sum_{hij} n^2_{G:M1\times M2\times M3_{hij}}}{df_G}\hat{\sigma}^2_{M1\times M2\times M3} \tag{A16}$$

where $Z_{u_{M1\times M2\times M3}} = \oplus_{hij} 1_{n_{M1\times M2\times M3_{hij}}}$ is the incidence matrix for the three-locus interaction. Finally, the ASV estimator of the residual genetic variance among entries nested in marker loci is:

$$\hat{\theta}^{ASV}_{G:M} = (n_G-1)^{-1}\hat{\sigma}^2_{G:M}tr(Z_{u_{G:M}}Z^T_{u_{G:M}}D_{n_G}) = \frac{n_G-1}{n_G-1}\hat{\sigma}^2_{G:M} = \hat{\sigma}^2_{G:M} \tag{A17}$$

where $Z_{u_{G:M}} = I_{n_G}$ is a n_G identity matrix.

Hence, as for the three-marker analysis, ASV yields k_M-bias corrected estimates of the marker-associated genetic variance for three loci ($M1$, $M2$, and $M3$) for unbalanced data:

$$\hat{p}_* = \frac{\sum_{i=1}^{3} \hat{\theta}_{Mi}^{ASV} + \sum_{j=1}^{2} \sum_{k=j+1}^{3} \hat{\theta}_{Mj\times Mk}^{ASV} + \hat{\theta}_{M1\times M2\times M3}^{ASV}}{\hat{\theta}_G^{ASV}} \tag{A18}$$
$$= \frac{\sum_{i=1}^{3} k_{Mi}\hat{\sigma}_{Mi}^2 + \sum_{j=1}^{2} \sum_{k=j+1}^{3} k_{Mj\times Mk}\hat{\sigma}_{Mj\times Mk}^2 + k_{M1\times M2\times M2}\hat{\sigma}_{M1\times M2\times M3}^2}{\hat{\sigma}_G^2}$$

where $i = 1,2,3,\ j = 1,2$, and $k = 2,3$ and $j \neq k$ are used to specify $M1$, $M2$, $M3$ and all two-way locus-locus interactions.

From (A14), the k_M-coefficient for bias-correcting AMV estimates of σ_{M1}^2, σ_{M1}^2, and σ_{M1}^2 are:

$$k_{M1} = \frac{n_G - n_G^{-1}\sum_h n_{G:M1_h}^2}{df_G} \tag{A19}$$
$$k_{M2} = \frac{n_G - n_G^{-1}\sum_i n_{G:M2_i}^2}{df_G} \tag{A20}$$
$$k_{M3} = \frac{n_G - n_G^{-1}\sum_j n_{G:M3_j}^2}{df_G} \tag{A21}$$

Similarly, from (A15), the k_M-coefficient for bias-correcting AMV estimates of $\sigma_{M1\times M2}^2$, $\sigma_{M1\times M2}^2$, and $\sigma_{M2\times M3}^2$ are:

$$k_{M1\times M2} = \frac{n_G - n_G^{-1}\sum_{hi} n_{G:M1\times M2_{hi}}^2}{df_G} \tag{A22}$$
$$k_{M1\times M3} = \frac{n_G - n_G^{-1}\sum_{hj} n_{G:M1\times M3_{hj}}^2}{df_G} \tag{A23}$$
$$k_{M2\times M3} = \frac{n_G - n_G^{-1}\sum_{ij} n_{G:M2\times M3_{ij}}^2}{df_G} \tag{A24}$$

From (A16), the k_M-coefficient for bias-correcting AMV estimates of $\sigma_{M1\times M2\times M3}^2$ (the genetic variance explained by the three-way marker loci interaction) is:

$$k_{M1\times M2\times M3} = \frac{n_G - n_G^{-1}\sum_{hij} n_{G:M_{hij}}^2}{df_G} \tag{A25}$$

These k_M-coefficients (A19-A25) are substituted in (A18) to obtain bias-corrected REML estimates of p.

A4 Biases of AMV and ASV Estimators of Marker-Associated Variance

Following [68] and [125], the true value of the variance for the between-entry effect in LMM (1) is:

$$s_G^2 = (n_G - 1)^{-1} \sum_{j=1}^{n_G} (g_j - \bar{g}_\bullet)^2 \tag{A26}$$

where $\bar{g}$ are the entry means, $\bar{g}_j$ is the jth element of $\bar{g}$, $j = 1, 2, 3, ..., n_G$, and $\bar{g}_\bullet$ is the mean of $\bar{g}$. Note that $E(s_G^2) = E(\hat{\sigma}_G^2) = \sigma_G^2$. The variance of the residual effect in LMM (1) and (2) is:

$$s_{\bar{\varepsilon}}^2 = (n_G - 1)^{-1} \sum_{j=1}^{n_G} (\bar{\varepsilon}_j - \bar{\varepsilon}_\bullet)^2 \tag{A27}$$

where $\bar{\varepsilon}$ are the mean residuals across entries, $\bar{\varepsilon}_j$ is the jth element of $\bar{\varepsilon}$, $j = 1, 2, 3, ..., n_G$, and $\bar{\varepsilon}_\bullet$ is the mean of $\bar{\varepsilon}$. We note that $E(s^2_{\bar{\varepsilon}}) = E(\hat{\sigma}^2_{\bar{\varepsilon}}) = \sigma^2_{\bar{\varepsilon}}$. The variance for the entries nested in M effect in LMM (2) is:

$$s^2_{G:M} = (n_G - 1)^{-1} \sum_{j=1}^{n_G} (\bar{q}_j - \bar{q}_\bullet)^2 \tag{A28}$$

where $\bar{q}_j$ is the effect of the jth entry nested in marker locus 1, $j = 1, 2, 3, ..., n_G$, and $\bar{q}_\bullet$ is the mean of $\bar{q}$ across entries. Note that $E(s^2_{G:M}) = E(\hat{\sigma}^2_{G:M}) = \sigma^2_{G:M}$. Finally, the variance for the effect of a single marker locus (M) in LMM (2) is:

$$s^2_M = (n_G - 1)^{-1} \sum_{j=1}^{n_G} (m_j - \bar{m}_\bullet)^2 \tag{A29}$$

where m_j is the effect of the jth genotype for a single marker locus, $j = 1, 2, 3, ..., n_G$, and $\bar{m}_\bullet$ is the mean of m across entries. Note that the summation in (A29) is over entries and not marker alleles. We demonstrate in the main body of this text that $E(s^2_M) \leq E(\hat{\sigma}^2_M) = \sigma^2_M$ which causes p and H^2_M to be systematically overestimated. The expected value of s^2_M is:

$$E(s^2_M) = (n_G - 1)^{-1} \sigma^2_M tr(Z_{n_G} Z^T_{n_G} D_{n_G}) = \theta^{ASV}_M \tag{A30}$$

Note that the ASV estimator θ^{ASV}_M defines the expected value of s^2_M which, as we demonstrate in the main text, is a fraction of the AMV estimator $\theta^{AMV}_M = E(\sigma^2_M)$. Our bias argument, supported by overwhelming simulation evidence, applies both in relation to the samples s^2_M and $E(s^2_M)$, but we focus on the former as it is more tangible.

The bias of the AMV estimator of the marker-associated genetic variance is:

$$bias(\hat{\theta}^{AMV}_M) = E[\hat{\theta}^{AMV}_M] - s^2_M \tag{A31}$$

where $\hat{\theta}^{AMV}_M$ is the AMV estimate of the marker associated genetic variance and $E[\hat{\theta}^{AMV}_M]$ is the expected value of $\hat{\theta}^{AMV}_M$. Similarly, the bias of the ASV estimator is:

$$bias(\hat{\theta}^{ASV}_M) = E[\hat{\theta}^{ASV}_M] - s^2_M \tag{A32}$$

where $\hat{\theta}^{ASV}_M$ is the ASV estimator of the marker associated genetic variance and $E[\hat{\theta}^{ASV}_M]$ is the expected value of $\hat{\theta}^{ASV}_M$.

The relative bias of the AMV estimator is:

$$RB(\hat{\theta}^{AMV}_M) = \frac{E[\hat{\theta}^{AMV}_M] - s^2_M}{s^2_M} \tag{A33}$$

and the relative bias of the ASV estimator is:

$$RB(\hat{\theta}^{ASV}_M) = \frac{E[\hat{\theta}^{ASV}_M] - s^2_M}{s^2_M} \tag{A34}$$

The biases and relative biases were empirically estimated through computer simulations.

A5 Sample Variances for AMV and ASV Estimators of p and H^2_M

Sample variances for AMV and ASV estimators of p and H^2_M were developed using the delta method and basic covariance algebra [25, 122, 126]. The sample variance of the AMV estimator of p is:

$$var(\hat{p}) \approx \left(\frac{\hat{\sigma}^2_M}{\hat{\sigma}^2_G}\right)^2 \left(\frac{var(\hat{\sigma}^2_M)}{(\hat{\sigma}^2_M)^2} + \frac{var(\hat{\sigma}^2_G)}{(\hat{\sigma}^2_G)^2} - \frac{2cov(\hat{\sigma}^2_M, \hat{\sigma}^2_G)}{\hat{\sigma}^2_M \times \hat{\sigma}^2_G}\right) \tag{A35}$$

where $var(\hat{\sigma}^2_G)$ is the variance of σ^2_G, $var(\hat{\sigma}^2_M)$ is the variance of $\hat{\sigma}^2_M$, and $cov(\hat{\sigma}^2_M, \hat{\sigma}^2_G)$ is the covariance between $\hat{\sigma}^2_M$ and $\hat{\sigma}^2_G$. Similarly, the sample variance of the AMV estimator of H^2_M is:

$$var(\hat{H}^2_M) \approx \left(\frac{\hat{\sigma}^2_M}{\hat{\sigma}^2_{\bar{P}}}\right)^2 \left(\frac{var(\hat{\sigma}^2_M)}{(\hat{\sigma}^2_M)^2} + \frac{var(\hat{\sigma}^2_{\bar{P}})}{(\hat{\sigma}^2_{\bar{P}})^2} - \frac{2cov(\hat{\sigma}^2_M, \hat{\sigma}^2_{\bar{P}})}{\hat{\sigma}^2_M \times \hat{\sigma}^2_{\bar{P}}}\right) \tag{A36}$$

where $cov(\hat{\sigma}^2_M, \hat{\sigma}^2_G)$ is the covariance between $\hat{\sigma}^2_M$ and $\hat{\sigma}^2_G$, $var(\hat{\sigma}^2_{\bar{P}})$ is the variance for the phenotypic variance on an entry-mean basis ($\hat{\sigma}^2_{\bar{P}}$), and $cov(\hat{\sigma}^2_M, \hat{\sigma}^2_{\bar{P}})$ is the covariance between $\hat{\sigma}^2_M$ and $\hat{\sigma}^2_{\bar{P}}$ from equation (7). The variance of $\hat{\sigma}^2_{\bar{P}}$ is:

$$var(\hat{\sigma}^2_{\bar{P}}) = var(\hat{\sigma}^2_G + \hat{\sigma}^2_\varepsilon / r_G) = var(\hat{\sigma}^2_G) + \frac{1}{r^2_G} var(\hat{\sigma}^2_\varepsilon) + \frac{2}{r_G} cov(\hat{\sigma}^2_G, \hat{\sigma}^2_\varepsilon) \tag{A37}$$

where $cov(\hat{\sigma}^2_M, \hat{\sigma}^2_\varepsilon)$ is the covariance between $\hat{\sigma}^2_M$ and $\hat{\sigma}^2_\varepsilon$. The covariance between $\hat{\sigma}^2_M$ and $\hat{\sigma}^2_G)$ is:

$$cov(\hat{\sigma}^2_M, \hat{\sigma}^2_G) = cov(\hat{\sigma}^2_M, \hat{\sigma}^2_G) = cov(\hat{\sigma}^2_M, \hat{\sigma}^2_M + \hat{\sigma}^2_{G:M}) = var(\hat{\sigma}^2_M) + cov(\hat{\sigma}^2_M, \hat{\sigma}^2_{G:M}) \tag{A38}$$

where $cov(\hat{\sigma}^2_M, \hat{\sigma}^2_{G:M})$ is the covariance between $\hat{\sigma}^2_M$ and $\hat{\sigma}^2_{G:M}$. The covariance between $\hat{\sigma}^2_M$ and $\hat{\sigma}^2_{\bar{P}}$ from equation (7) is:

$$cov(\hat{\sigma}^2_M, \hat{\sigma}^2_{\bar{P}}) = cov(\hat{\sigma}^2_M, \hat{\sigma}^2_G + \hat{\sigma}^2_\varepsilon / r_G) = cov(\hat{\sigma}^2_M, \hat{\sigma}^2_G) + \frac{1}{r_G} cov(\hat{\sigma}^2_M, \hat{\sigma}^2_\varepsilon) \tag{A39}$$

These sample variances can be extracted from the asymptotic variance-covariance matrices estimated with widely used software for linear mixed model analyses, e.g., *lme4* [78].

Using the delta method [25, 126] and recalling from equation (8) that $k_M \hat{\sigma}^2_M + \hat{\sigma}^2_{G:M} = \hat{\sigma}^2_G$, the sample variance of the ASV estimates of p is:

$$var(\hat{p}_{M*}) \approx \left(\frac{k_M \hat{\sigma}^2_M}{\hat{\sigma}^2_G}\right)^2 \left(\frac{var(\hat{\sigma}^2_M)}{(\hat{\sigma}^2_M)^2} + \frac{var(\hat{\sigma}^2_G)}{(\hat{\sigma}^2_G)^2} - \frac{2cov(\hat{\sigma}^2_M, \hat{\sigma}^2_G)}{\hat{\sigma}^2_M \times \hat{\sigma}^2_G}\right) \tag{A40}$$

where $var(\hat{\sigma}^2_G)$ is the variance of σ^2_G, $var(k_M \hat{\sigma}^2_M)$ is the variance of $k_M \hat{\sigma}^2_M$, and $cov(k_M hat \sigma^2_M, \hat{\sigma}^2_G)$ is the covariance between $k_M \hat{\sigma}^2_M$ and $\hat{\sigma}^2_G$.

Similarly, the sample variance for the ASV estimator of H^2_M is:

$$var(\hat{H}^2_{M*}) \approx \left(\frac{k_M \hat{\sigma}^2_M}{\hat{\sigma}^2_{\bar{P}}}\right)^2 \left(\frac{var(\hat{\sigma}^2_M)}{(\hat{\sigma}^2_M)^2} + \frac{var(\hat{\sigma}^2_{\bar{P}})}{(\hat{\sigma}^2_{\bar{P}})^2} - \frac{2cov(\hat{\sigma}^2_M, \hat{\sigma}^2_{\bar{P}})}{\hat{\sigma}^2_M \times \hat{\sigma}^2_{\bar{P}}}\right) \tag{A41}$$

where $cov(k_M \hat{\sigma}^2_M, \hat{\sigma}^2_G)$ is the covariance between $k_M \hat{\sigma}^2_M$ and $\hat{\sigma}^2_G$, $var(\hat{\sigma}^2_{\bar{P}})$ is the variance for the phenotypic variance on an entry-mean basis, and $cov(k_M \hat{\sigma}^2_M, \hat{\sigma}^2_{\bar{P}})$ is the covariance between $k_M \hat{\sigma}^2_M$ and $\hat{\sigma}^2_{\bar{P}}$ from equation (7). The covariance between $k_M \hat{\sigma}^2_M$ and $\hat{\sigma}^2_{\bar{P}}$ from equation (2) is:

$$\begin{aligned} cov(k_M \hat{\sigma}^2_M, \hat{\sigma}^2_{\bar{P}}) &= cov(k_M \hat{\sigma}^2_M, \hat{\sigma}^2_G + \hat{\sigma}^2_\varepsilon / r_G) = k_M cov(\hat{\sigma}^2_M, \hat{\sigma}^2_G) + \frac{k_M}{r_G} cov(\hat{\sigma}^2_M, \hat{\sigma}^2_\varepsilon) \\ &= k^2_M var(\hat{\sigma}^2_M) + k_M cov(\hat{\sigma}^2_M, \hat{\sigma}^2_{G:M}) + \frac{k_M}{r_G} cov(\hat{\sigma}^2_M, \hat{\sigma}^2_\varepsilon) \end{aligned} \tag{A42}$$

where $cov(\hat{\sigma}^2_M, \hat{\sigma}^2_{G:M})$ is the covariance between $\hat{\sigma}^2_M$ and $\hat{\sigma}^2_{G:M}$ from equation (2). The covariance between $k_M \hat{\sigma}^2_M$ and $\hat{\sigma}^2_G$ is:

$$cov(k_M \hat{\sigma}^2_M, \hat{\sigma}^2_G) = k_M cov(\hat{\sigma}^2_M, \hat{\sigma}^2_G) = k_M cov(\hat{\sigma}^2_M, k_M \hat{\sigma}^2_M + \hat{\sigma}^2_{G:M}) = k^2_M var(\hat{\sigma}^2_M) + k_M cov(\hat{\sigma}^2_M, \hat{\sigma}^2_{G:M}) \tag{A43}$$

We found that the sample variances of the ASV estimators are consistently smaller than AMV estimators of p and H^2_M by a factor of $1 - k_M$, as shown here:

$$\frac{var(\hat{p}_*)}{var(\hat{p})} \approx \frac{\left(\frac{k_M\hat{\sigma}_M^2}{\hat{\sigma}_G^2}\right)^2\left(\frac{var(\hat{\sigma}_M^2)}{(\hat{\sigma}_M^2)^2}+\frac{var(\hat{\sigma}_G^2)}{(\hat{\sigma}_G^2)^2}-\frac{2cov(\hat{\sigma}_M^2,\hat{\sigma}_G^2)}{\hat{\sigma}_M^2\times\hat{\sigma}_G^2}\right)}{\left(\frac{\hat{\sigma}_M^2}{\hat{\sigma}_G^2}\right)^2\left(\frac{var(\hat{\sigma}_M^2)}{(\hat{\sigma}_M^2)^2}+\frac{var(\hat{\sigma}_G^2)}{(\hat{\sigma}_G^2)^2}-\frac{2cov(\hat{\sigma}_M^2,\hat{\sigma}_G^2)}{\hat{\sigma}_M^2\times\hat{\sigma}_G^2}\right)} = \frac{\left(\frac{k_M\hat{\sigma}_M^2}{\hat{\sigma}_G^2}\right)^2}{\left(\frac{\hat{\sigma}_M^2}{\hat{\sigma}_G^2}\right)^2} = \frac{k_M^2\left(\frac{\hat{\sigma}_M^2}{\hat{\sigma}_G^2}\right)^2}{\left(\frac{\hat{\sigma}_M^2}{\hat{\sigma}_G^2}\right)^2} = \frac{k_M^2}{1} < 1 \tag{A44}$$

and

$$\frac{var(\hat{H}_{M*}^2)}{var(\hat{H}_M^2)} \approx \frac{\left(\frac{k_M\hat{\sigma}_M^2}{\hat{\sigma}_{\bar{P}}^2}\right)^2\left(\frac{var(\hat{\sigma}_M^2)}{(\hat{\sigma}_M^2)^2}+\frac{var(\hat{\sigma}_{\bar{P}}^2)}{(\hat{\sigma}_{\bar{P}}^2)^2}-\frac{2cov(\hat{\sigma}_M^2,\hat{\sigma}_{\bar{P}}^2)}{\hat{\sigma}_M^2\times\hat{\sigma}_{\bar{P}}^2}\right)}{\left(\frac{\hat{\sigma}_M^2}{\hat{\sigma}_{\bar{P}}^2}\right)^2\left(\frac{var(\hat{\sigma}_M^2)}{(\hat{\sigma}_M^2)^2}+\frac{var(\hat{\sigma}_{\bar{P}}^2)}{(\hat{\sigma}_{\bar{P}}^2)^2}-\frac{2cov(\hat{\sigma}_M^2,\hat{\sigma}_{\bar{P}}^2)}{\hat{\sigma}_M^2\times\hat{\sigma}_{\bar{P}}^2}\right)} = \frac{\left(\frac{k_M\hat{\sigma}_M^2}{\hat{\sigma}_{\bar{P}}^2}\right)^2}{\left(\frac{\hat{\sigma}_M^2}{\hat{\sigma}_{\bar{P}}^2}\right)^2} = \frac{k_M^2\left(\frac{\hat{\sigma}_M^2}{\hat{\sigma}_{\bar{P}}^2}\right)^2}{\left(\frac{\hat{\sigma}_M^2}{\hat{\sigma}_{\bar{P}}^2}\right)^2} = \frac{k_M^2}{1} < 1 \tag{A45}$$

Importantly, k_M^2 is the square of a value less than 1 indicating that the sampling variance of ASV estimates of p_* and H_{M*}^2 are less than those of p and H_M^2.

Table S1. 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 ($n = n_G \times r_G$), 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^2_M). 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

Design	m	$n_{G:M}$	n_G	r_G	n	H^2_M
1	1	180	540	5	2,700	$(0.0, 1.0)$
2	2	180	540	5	2,700	$(0.0, 1.0)$
3	3	180	540	5	2,700	$(0.0, 1.0)$
4[a]	1	135 : 270 : 135	540	5	2,700	$(0.0, 1.0)$
5[b]	1	$(126, 180)$	$(486, 540)$	$(0, 5)$	2,430	$(0.0, 1.0)$
6[c]	1	$(0, 180)$	$(360, 540)$	$(0, 5)$	1,800	$(0.0, 1.0)$
7	1	300	900	1	900	0.50
8	1	300	900	2	1,800	0.50
9	1	300	900	5	4,500	0.50
10	1	300	900	10	9,000	0.50
11	1	300	900	20	18,000	0.50
12	1	150	450	5	2,250	0.50
13	1	300	900	5	4,500	0.50
14	1	600	1,800	5	9,000	0.50
15	1	1200	3,600	5	18,000	0.50
16	1	2,400	7,200	5	36,000	0.50
17	1	150	450	5	2,250	0.05
18	1	150	450	5	2,250	0.25
19	1	150	450	5	2,250	0.50
20	1	150	450	5	2,250	0.75
21	1	150	450	5	2,250	0.95

[a]The number of entries nested in marker genotypes for study design 4 was
[b]The numerical ranges for r_M, n_G, and r_G varied among simulated samples
[c]The numerical ranges for r_M, n_G, and r_G varied among simulated samples

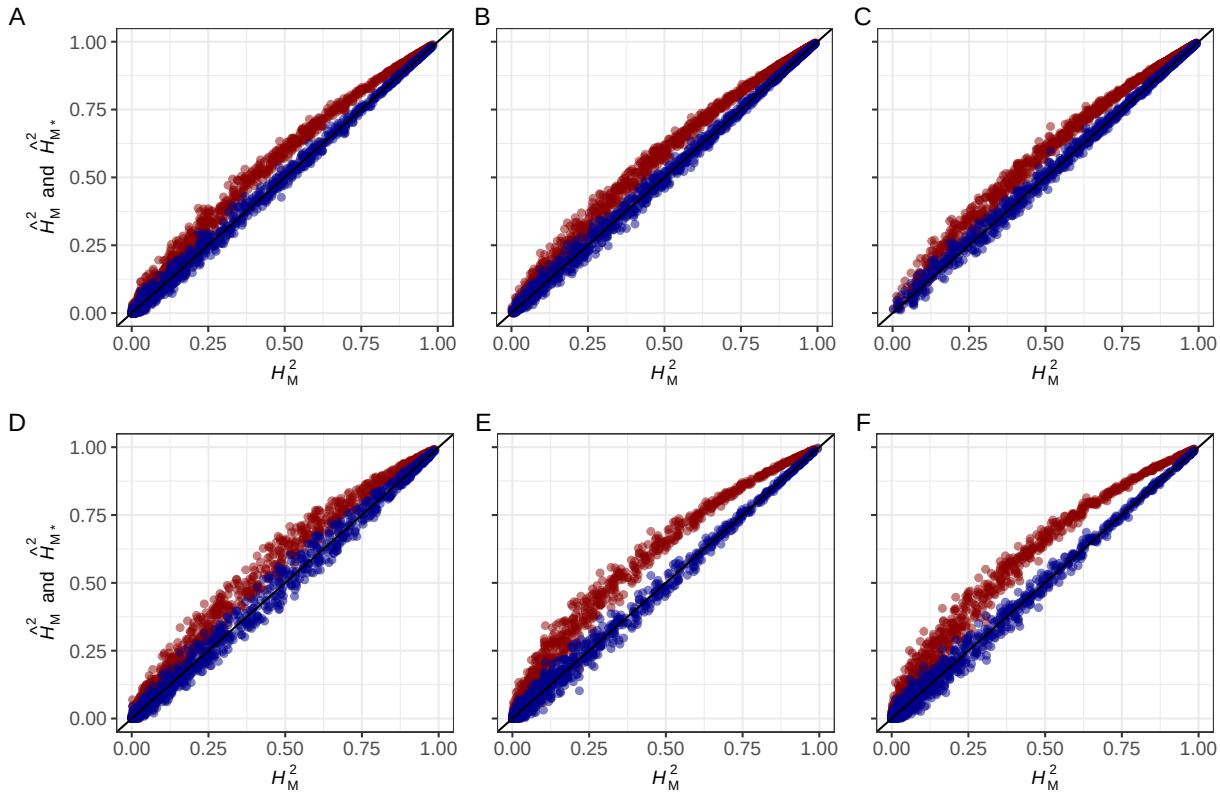

Fig S1. Accuracy of AMV and ASV Estimators of Marker Heritability When the Phenotypic Variance is Estimated by Pooling Marker and Residual Genetic Sources of Variation ($\sigma^2_M + \sigma^2_{G:M}$). AMV and ASV estimates of H^2_M when σ^2_G from LMM (1) is replaced with $\hat{\sigma}^2_M + \hat{\sigma}^2_{G:M}$ for AMV from LMM (2) or $k_M\hat{\sigma}^2_M + \hat{\sigma}^2_{G:M}$ 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^2_M) 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 $\hat{\sigma}^2_M/(\hat{\sigma}^2_M + \hat{\sigma}^2_{G:M} + r_G^{-1}\hat{\sigma}^2_\varepsilon)$, whereas the ASV estimates (shown in blue) equal $k_M\hat{\sigma}^2_M/(k_M\hat{\sigma}^2_M + \hat{\sigma}^2_{G:M} + r_G^{-1}\hat{\sigma}^2_\varepsilon)$. AMV estimates of marker heritability ($\hat{H}^2_M$; red highlighted observations) and ASV estimates of marker heritability ($\hat{H}^2_{M*}$; 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 Table S1.

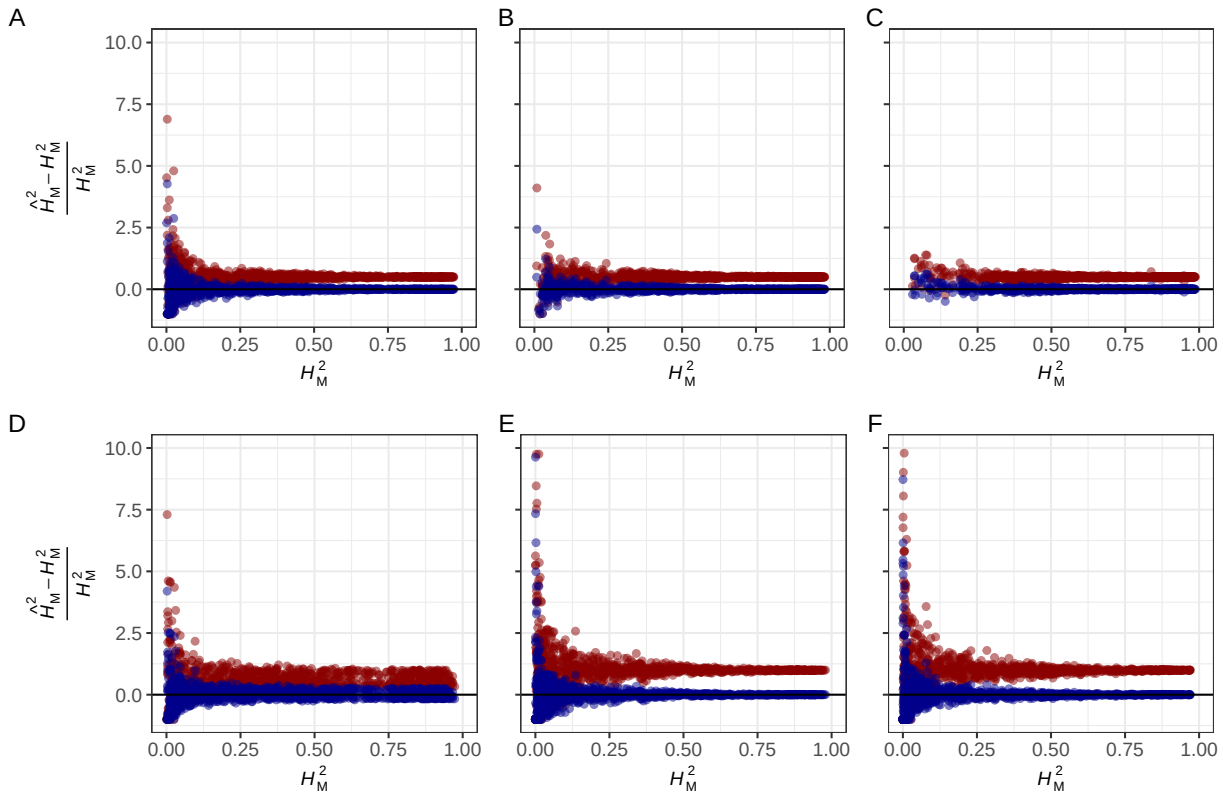

Fig S2. Relative Bias of AMV and ASV Estimators of Marker Heritability. Relative biases of AMV and ASV estimates of H^2_M are shown for 1,000 segregating populations simulated for different numbers of entries (n_G individuals, families, or strains), five replications/entry (r_G = 5), true marker heritability (H^2_M) ranging from 0 to 1, and one to three marker loci with three genotypes/marker locus (n_{M1} = 3). AMV estimates of marker heritability ($\hat{H}^2_M$; red highlighted observations) and ASV estimates of marker heritability ($\hat{H}^2_{M*}$; 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) an population segregating 1:2:1 for one 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 Table S1.