



# Effect of genetic architecture on the prediction accuracy of quantitative traits in samples of unrelated individuals

Fabio Morgante<sup>1,2,3</sup> · Wen Huang<sup>1,2,3,4,6</sup> · Christian Maltecca<sup>1,5</sup> · Trudy F.C. Mackay<sup>1,2,3,4</sup>

Received: 2 August 2017 / Revised: 16 November 2017 / Accepted: 22 November 2017 / Published online: 10 February 2018  
© The Author(s) 2018. This article is published with open access

## Abstract

Predicting complex phenotypes from genomic data is a fundamental aim of animal and plant breeding, where we wish to predict genetic merits of selection candidates; and of human genetics, where we wish to predict disease risk. While genomic prediction models work well with populations of related individuals and high linkage disequilibrium (LD) (e.g., livestock), comparable models perform poorly for populations of unrelated individuals and low LD (e.g., humans). We hypothesized that low prediction accuracies in the latter situation may occur when the genetics architecture of the trait departs from the infinitesimal and additive architecture assumed by most prediction models. We used simulated data for 10,000 lines based on sequence data from a population of unrelated, inbred *Drosophila melanogaster* lines to evaluate this hypothesis. We show that, even in very simplified scenarios meant as a stress test of the commonly used Genomic Best Linear Unbiased Predictor (G-BLUP) method, using all common variants yields low prediction accuracy regardless of the trait genetic architecture. However, prediction accuracy increases when predictions are informed by the genetic architecture inferred from mapping the top variants affecting main effects and interactions in the training data, provided there is sufficient power for mapping. When the true genetic architecture is largely or partially due to epistatic interactions, the additive model may not perform well, while models that account explicitly for interactions generally increase prediction accuracy. Our results indicate that accounting for genetic architecture can improve prediction accuracy for quantitative traits.

---

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1038/s41437-017-0043-0>) contains supplementary material, which is available to authorized users.

---

✉ Trudy F.C. Mackay  
trudy\_mackay@ncsu.edu

- <sup>1</sup> Program in Genetics, North Carolina State University, Raleigh, NC 27695-7614, USA
- <sup>2</sup> Department of Biological Sciences, North Carolina State University, Raleigh, NC 27695-7614, USA
- <sup>3</sup> W. M. Keck Center for Behavioral Biology, North Carolina State University, Raleigh, NC 27695-7614, USA
- <sup>4</sup> Initiative in Biological Complexity, North Carolina State University, Raleigh, NC 27695-7614, USA
- <sup>5</sup> Department of Animal Science, North Carolina State University, Raleigh, NC 27695-7621, USA
- <sup>6</sup> Present address: Department of Animal Science, Michigan State University, East Lansing, MI 48824, USA

## Introduction

Most phenotypic variation in natural populations is continuous. Fisher (1918) reconciled the opposing ideas of Mendelian segregation and continuous phenotypic variation for quantitative traits by formulating the classical ‘infinitesimal’ model of inheritance, whereby alleles at many (effectively infinite) loci affecting quantitative traits exhibit Mendelian properties of homozygous (additive) effects, dominance effects, and non-additive interlocus interactions (epistasis), but the effects are small and sensitive to variation in the environment (Falconer and Mackay 1996). The effects of these quantitative trait loci (QTLs) are too small to be assessed individually in pedigrees. However, QTL effects in aggregate can be described in terms of additive ( $V_A$ ), dominance ( $V_D$ ), and epistatic ( $V_I$ ) variance components, the summation of which comprises the total genetic variance ( $V_G$ ) of a quantitative trait; the remainder of the observed phenotypic variance ( $V_P$ ) is due to variation in environmental influences ( $V_E$ ). Importantly, Fisher showed that the ratio of additive genetic variance to the total phenotypic variance,  $V_A/V_P$ , the narrow sense heritability ( $h^2$ ),

determines the phenotypic correlations among relatives.  $h^2$  is the fraction of the variance of the trait that is transmissible from parents to offspring, and is therefore critical for predicting responses to natural and artificial selection, as well as disease risk, in the absence of knowledge of the underlying genetic details (Fisher 1918; Falconer and Mackay 1996).

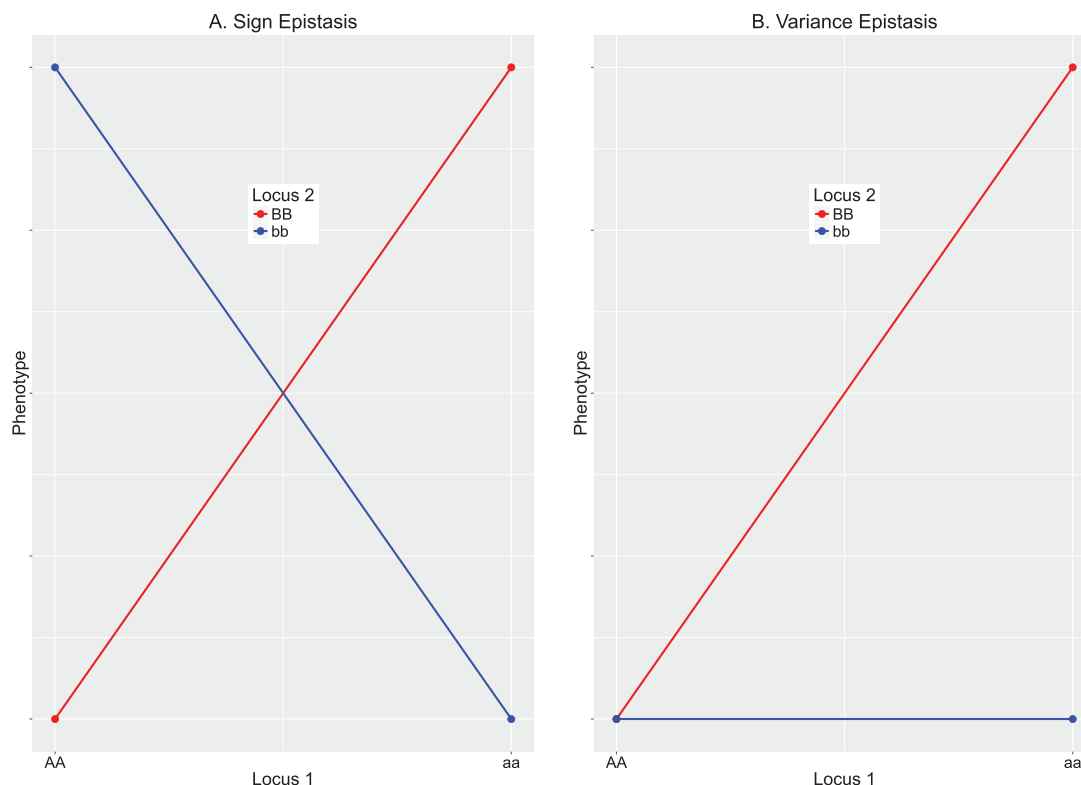
The current availability of abundant polymorphic molecular markers, and in some cases population scale genome sequences, have enabled high resolution and well powered genome wide association studies (GWAS) to map individual QTLs, particularly in human populations. GWAS assessing additive (marginal) effects of QTLs have identified many loci affecting complex traits, contributing to our knowledge of the biology of these traits. However, these loci cumulatively generally explain a small proportion of the narrow sense heritability, limiting our ability to predict quantitative trait phenotypes from high resolution genetic polymorphism data. The most cited example of this phenomenon is human height, for which  $h^2$  from pedigree studies is  $\sim 0.8$  and the loci identified by GWAS collectively explain  $\sim 10\%$  of the total phenotypic variation (Manolio et al. 2009; Lango Allen et al. 2010).

There are many possible and not mutually exclusive causes of this “missing heritability” phenomenon, including the stringent significance threshold imposed by the multiple testing correction, limiting sample sizes if most QTLs have small effects, rare alleles with large effects that are not assessed by single locus GWAS, and failure to account for the possibility of non-additive effects (Manolio et al. 2009). Recognizing the difficulty of mapping causal QTLs, Meuwissen et al. (2001) proposed using whole genome regressions of additive effects of many molecular markers simultaneously to predict individual genetic values. Such genomic prediction methods have shown high prediction accuracies in domesticated animals and crops and have transformed animal and plant breeding (Goddard 2009). When Yang et al. (2010) applied the whole genome regression method to a large sample of unrelated humans with dense common single nucleotide polymorphism (SNP) genotype data and height phenotypes, the proportion of phenotypic variance explained by SNPs (the ‘genomic heritability’,  $h^2_g$ ) was  $\sim 0.45$ ; about half of the narrow sense heritability. However, the ability to explain a substantial fraction of genetic variation does not necessarily translate into high predictive ability; the same model used by Yang et al. (2010) gave poor predictive ability for human height (de los Campos et al. 2013). The discrepancy between high prediction accuracies with domesticated animals and crops and low accuracies in humans is likely due to different patterns of overall relatedness and consequently different patterns of linkage disequilibrium (LD). High relatedness and consequently high LD in selected crops and animals

mean that any marker is likely to be in strong LD with causal QTL(s); but largely unrelated humans and relatively low LD mean that markers will not necessarily capture effects of causal QTLs (de los Campos et al. 2013). Variation in LD is also thought to be the explanation for poor performance of genomic prediction across breeds (De Roos et al. 2009).

The most commonly used genomic prediction model, Genomic Best Linear Unbiased Predictor (G-BLUP, e.g., Habier et al. 2007) assumes Fisher’s additive, infinitesimal genetic architecture. Given a single population in which genotype data and measures of multiple phenotypes have been obtained, the same G-BLUP model would be applied to all phenotypes, which is biologically unrealistic and neither gives nor uses any biological insight regarding the genetic basis of variation in the traits. Thus, an additional reason for poor prediction accuracy can be departure of the real genetic architecture of quantitative traits from the infinitesimal model. Information on relatedness from large panels of molecular markers can dilute the true signal of genetic relatedness at causal loci when many fewer loci affect the trait than markers, even if the true genetic architecture is additive. Further, epistatic interactions have been shown to be an important component of genetic architecture of quantitative traits in model organisms (Mackay 2014; Taylor and Ehrenreich 2015), and if they occur, accounting for non-linear interactions between loci may improve prediction (Jiang and Reif 2015; Ober et al. 2015; Martini et al. 2016).

If molecular quantitative genetics is to fulfill the dual goal of biological insight and genotype-phenotype prediction in the context of precision medicine (de los Campos et al. 2010) and precision agriculture (e.g., Goddard and Hayes 2009; Heffner et al. 2009), we need to develop accurate prediction methods that work well for populations of unrelated individuals and that utilize trait architecture information to simultaneously give insight into the numbers, identity and gene action of causal loci. A method combining GWAS and G-BLUP has recently been proposed to achieve this goal (Ober et al. 2015). Briefly, phenotypic records of time to recover from a chill-induced coma were obtained for multiple individuals from the sequenced, inbred lines from the *Drosophila* Genetic Reference Panel (DGRP) (Mackay et al. 2012; Huang et al. 2014). The broad sense heritability ( $H^2$ , the proportion of phenotypic variance due to all sources of genetic variation) of line means, on which GWAS was based, was very high,  $\sim 0.9$ . GWAS revealed sex-specific genetic architecture including major effect loci and evidence of epistasis. The prediction accuracy estimated using G-BLUP and cross-validation was zero in both sexes. However, incorporating genetic architecture by performing GWAS for additive effects and epistatic interactions in the training data and using the top variants to



**Fig. 1** Types of epistasis. **a** Sign epistasis. The effect of Locus 1 is of the same magnitude but opposite direction depending on the genotype of Locus 2. **b** Variance epistasis. The effect of Locus 1 changes magnitude depending on the genotype of Locus 2

build a genomic relationship matrix (GRM) to predict phenotypes in the test data greatly improved prediction accuracy, despite the small sample size, low LD between closely linked loci, and minimal relatedness among the DGRP lines.

Motivated by the results of Ober et al. (2015), we further investigated the utility of combining mapping and prediction to simultaneously infer genetic architecture and develop a robust prediction model under the unfavorable scenario of low relatedness and LD. To overcome the limited sample size of the DGRP, we simulated large numbers of genotypes based on the DGRP polymorphisms and allele frequency spectrum, as well as whole genomes with similar LD decay and pairwise relatedness to the DGRP, for a range of simplified genetic architectures. (We use the term ‘genetic architecture’ throughout to denote the number of causal loci, their allele frequencies and gene actions, although we recognize that this term may embrace other factors). The simulated genetic architectures are more simplified and extreme than in reality, but serve as a ‘stress test’ of the G-BLUP model. We found that the G-BLUP model performed poorly when all common polymorphisms are used for prediction, and very well when the genetic architecture was estimated by GWAS in the training data, and then incorporated in the prediction model.

## Materials and methods

### Simulation of causal genotypes and phenotypes

We simulated true QTL genotypes for sample sizes of 205 (the size of the DGRP), 1000, 2500, and 5000 by randomly sampling 0s and 2s, i.e., the number of copies of the reference allele, with probability equal to the genotype/allele frequency spectrum of a reduced version of the DGRP genotype data, including 8665 variants meeting the following criteria. The genotypes were pruned for LD such that in every 1000-variant window, there were no pairs of variants whose  $r^2$  was greater than 0.05; the genotype call rate was greater than 0.8; and the minor allele frequency (MAF) was greater than 0.25 for all variants. We chose the high MAF threshold because prediction will fail for any model that includes a causal allele in the training set that is not present in the test set, which will happen for low MAF when the population size is small (the case of 205 lines). The stringent LD threshold was chosen to represent the simplified scenario where causal loci are not in LD.

We simulated true QTLs for several different genetic architectures. QTL effects were randomly drawn from a *Gamma* distribution with shape and scale parameters of 0.4

and 1.66, respectively. We randomly assigned the sign of the effects such that some were positive and some were negative (Meuwissen et al. 2001). The true genetic values were calculated by multiplying the genotypes by the QTL effects. Environmental effects were randomly drawn from a *Normal* distribution with a mean of 0 and variance =  $V_g \frac{(1-H^2)}{H^2}$ , where  $V_g$  is the variance of the genetic values and  $H^2$  is the broad sense heritability of the trait. Phenotypes for the individuals were then obtained by summing the genetic and environmental effects. We simulated four additive genetic architectures in which 1, 20, 100, and 1000 QTLs explained all the genetic variation using this approach.

We also simulated genetic architectures for 50 and 500 interacting pairs of variants exhibiting sign epistasis and variance epistasis (Fig. 1) using a slightly modified approach from the one we used to simulate the additive architectures. With sign epistasis, the effect on the phenotype of one locus is of the same magnitude but in opposite directions depending on the genotype at the interacting locus, so negligible additive variance is produced when the allele frequencies at both loci are intermediate. With variance epistasis, where the effect on the phenotype of one locus is in the same direction but of different magnitude depending on the genotype at the interacting locus, additive variance is produced when both loci are polymorphic (Mackay 2014).

Once the two variants defining the interaction were randomly sampled, an interaction matrix was created by assigning a coefficient of  $-1$  if the two genotypes were 0 0 or 2 2, and a coefficient of  $+1$  if the two genotypes were 2 0 or 0 2 for sign epistasis. For variance epistasis, the coefficient was  $-1$  if the two genotypes were 0 0, 2 2, or 2 0, and  $+1$  if the two genotypes were 0 2. Then, the simulation of phenotypes followed the procedure described above using the interaction matrix instead of individual QTL genotypes. This procedure thus assigns two-locus haplotypes (as appropriate for inbred lines) rather than single-locus effects.

Further, we simulated two additional genetic architectures composed of a mixture of additive and epistatic effects with 100 additive QTLs and 50 pairwise sign or variance epistatic interactions following the procedure described above. For each of these more complex genetic architectures, the proportion of total genetic variance explained by the additive component and the epistatic component, respectively, was varied to be 0.75:0.25, 0.50:0.50, and 0.25:0.75 by adjusting the epistatic effects with an appropriate constant. Importantly, these proportions do not necessarily lead to  $V_A$  and  $V_I$  partitions of the same numerical values since epistatic effects can also produce  $V_A$ . These two genetic architectures were simulated for sample sizes of 205, 1000, and 2500.

We performed simulations for all genetic architectures for  $H^2 = 0.4$  and  $H^2 = 0.9$ , and 30 replicates per genetic architecture/heritability value were produced.

### Simulation of whole genome sequences and phenotypes

Whole haplotype sequences for 10,000 lines were simulated using the Markovian Coalescent Simulator (MaCS). Then, each haplotype was duplicated to create 10,000 diploid, inbred genomes (one for each line). MaCS takes two main parameters as input:  $\rho = 4N_e r$  and  $\theta = 4N_e \mu$ ; where  $N_e$  is the effective population size,  $r$  is the recombination rate and  $\mu$  is the mutation rate (Chen et al. 2009). We chose values for these parameters that produce LD decay and the distribution of pairwise relatedness among the lines similar to those observed in the DGRP (Huang et al. 2014) and that are plausible for a natural population of *D. melanogaster*. The values chosen were  $N_e = 1,000,000$ ;  $r = 1 \times 10^{-8}$  and  $\mu = 1 \times 10^{-8}$  (Charlesworth 2009; Comeron et al. 2012; Huang et al. 2016). The simulated whole genomes were composed of 3 chromosome arms of 5 Mb each. The small simulated genome compared to the actual *Drosophila* genome was used so the number of polymorphic sites was similar to the DGRP. The resulting sequences had 5,871,537 polymorphic sites, of which 1,761,219 were common (MAF > 0.05).

The four additive and four pure epistatic architectures described in the previous section were simulated by randomly sampling variants to be QTLs or interactions, and then using the same procedure. The simulated genotype data were pruned for LD and MAF using the same thresholds as above and consisted of 18,795 variants. All the additive and epistatic architectures were simulated for  $H^2 = 0.4$  and  $H^2 = 0.9$ , and 30 replicates per architecture/heritability value were produced.

### Statistical analysis: estimation and prediction

The data were analyzed using alternative genomic models:

1. Additive model:  $\mathbf{y} = 1\mu + \mathbf{g}_A + \mathbf{e}$ , where  $\mathbf{y}$  is an  $n$ -vector of phenotypes,  $\mathbf{1}$  is an  $n$ -vector of ones,  $\mu$  is the population mean,  $\mathbf{g}_A$  is an  $n$ -vector of random additive line effects [ $\mathbf{g}_A \sim N(\mathbf{0}, \mathbf{G}\sigma_{gA}^2)$ ] and  $\mathbf{e}$  is an  $n$ -vector of random residual effects [ $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$ ].  $\mathbf{G}$  is the additive GRM built using all common variants (MAF > 0.05) according to the formula  $\frac{\mathbf{W}\mathbf{W}'}{p}$  where  $\mathbf{W}$  is the matrix of centered and standardized genotypes for all the lines and  $p$  is the number of variants;  $\mathbf{I}$  is the identity matrix.
2. Epistatic model:  $\mathbf{y} = 1\mu + \mathbf{g}_E + \mathbf{e}$ , where  $\mathbf{y}$  is an  $n$ -vector of phenotypes,  $\mathbf{1}$  is an  $n$ -vector of ones,  $\mu$  is the population mean,  $\mathbf{g}_E$  is an  $n$ -vector of random epistatic

line effects [ $\mathbf{g}_E \sim N(\mathbf{0}, \text{EpiG}\sigma_{gE}^2)$ ] and  $\mathbf{e}$  is an  $n$ -vector of random residual effects [ $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$ ]. EpiG is the additive  $\times$  additive epistatic GRM and was built by  $\mathbf{G}\#\mathbf{G}$  as in Su et al. (2012);  $\mathbf{I}$  is the identity matrix.

3. Combined additive and epistatic model:  $\mathbf{y} = \mathbf{1}\mu + \mathbf{g}_A + \mathbf{g}_E + \mathbf{e}$ , where  $\mathbf{y}$  is an  $n$ -vector of phenotypes,  $\mathbf{1}$  is an  $n$ -vector of ones,  $\mu$  is the population mean,  $\mathbf{g}_A$  is an  $n$ -vector of random additive line effects [ $\mathbf{g}_A \sim N(\mathbf{0}, \mathbf{G}\sigma_{gA}^2)$ ],  $\mathbf{g}_E$  is an  $n$ -vector of random epistatic line effects [ $\mathbf{g}_E \sim N(\mathbf{0}, \text{EpiG}\sigma_{gE}^2)$ ] and  $\mathbf{e}$  is an  $n$ -vector of random residual effects [ $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$ ].  $\mathbf{G}$  and EpiG are the additive and epistatic GRMs, respectively, as described above;  $\mathbf{I}$  is the identity matrix.

When the three models were used with the true QTL data, the additive and epistatic GRMs were built using only the true QTLs/interactions that generated the phenotypes. In the additive scenarios, only the additive model was fitted. For scenarios with sign epistasis and variance epistasis, all three models were fitted; the GRM in the additive model was built using all the 100 or 1000 variants that defined the 50 or 500 pairwise interactions. In the epistatic and combined additive and epistatic models, we implemented a modified version of the epistatic GRM proposed by Ober et al. (2015), which accounts only for the true interactions that generated the phenotypes.

The true QTL data were analyzed using 10 replicates of 5-fold cross-validation (CV) for each of the 30 replicates of simulated phenotypes, as this is the most appropriate technique when analyzing non-simulated data, where only one dataset is usually available, and/or sample size is small (Daetwyler et al. 2013). For simulations with whole genome sequences, where the sample size was 10,000, we randomly drew training (80% of the data) and test (20% of the data) samples to reduce the computational demand.

The prediction accuracy was calculated as the squared correlation coefficient between the true and predicted phenotypes ( $r^2$ ), averaged over the 30 replicates of each genetic architecture; when CV was used,  $r^2$  was also averaged over folds and CV replicates. This statistic, which measures the proportion of true phenotypic variability accounted for by the predicted phenotypes, is useful for evaluating prediction models because the asymptotic upper bound of  $r^2$  is the heritability of the trait when fitting a model with only genomic data (de Los Campos et al. 2013). For a number of representative architectures and methods, we also calculated the bias of prediction as the slope of the regression of true phenotypes onto predicted phenotypes. The genomic heritability was calculated in the training set as  $h_g^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$  using the additive model, averaged over the 30 replicates of each architecture (de los Campos et al. 2015).

These analyses were performed using R (R Core Team 2015). Variance components for the datasets with sample size of 10,000 were estimated using GCTA (Yang et al. 2011).

### Statistical analysis: variable selection

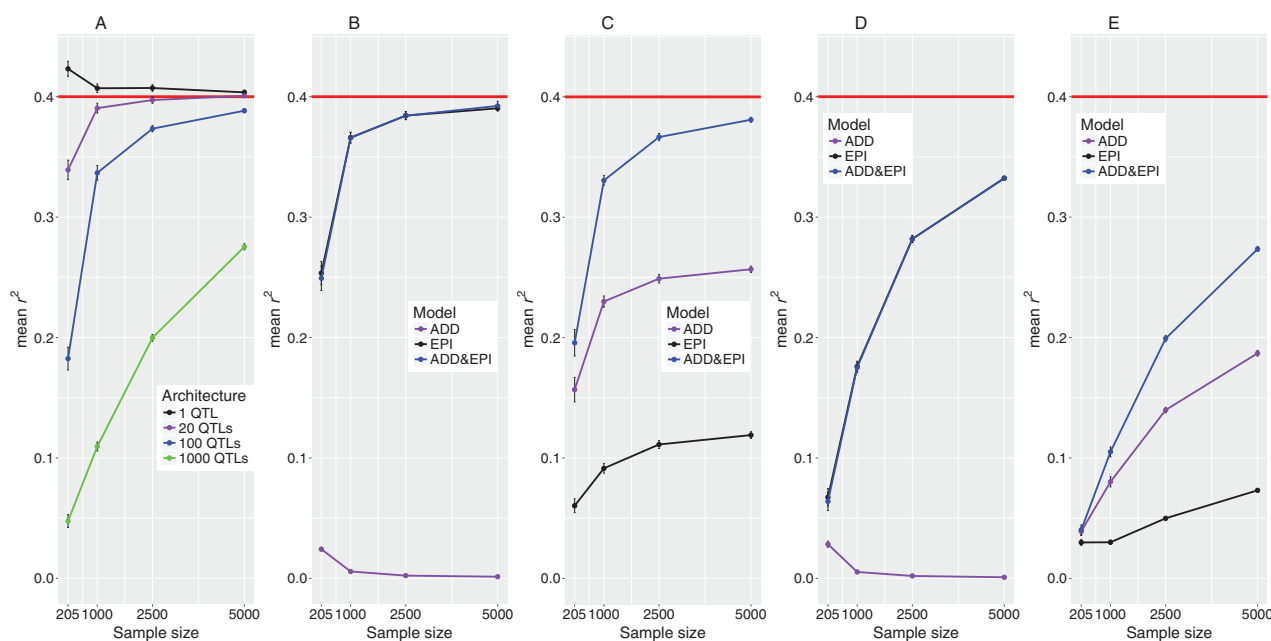
To combine mapping and prediction and select informative features for the prediction models when whole genome sequences were used, we performed marker selection based on genotype-phenotype associations in the training population; we then used only the selected markers to derive the model in the training population and predict phenotypes in the test population. Specifically, we performed a single marker regression without any adjustment to rank variants whose additive effects are associated with the phenotype using PLINK (Purcell et al. 2007) using all ~1,800,000 common variants (i.e., not only the ~18,000 variants used to sample the true causal variants and interactions). We selected (1) the top  $t$  variants (for  $t = 100, 1000, 10,000$  or  $100,000$ ) with the smallest  $P$ -values and (2) variants with  $P$ -values for association smaller than  $10^{-3}, 10^{-5}, 10^{-7}$ , and  $10^{-9}$  to build the additive GRM to use in the prediction analysis.

We also performed a pairwise interaction GWAS (EpiGWAS) using FastEpistasis (Schüpbach et al. 2010), testing all possible  $\binom{18,795}{2}$  variant-variant combinations on the pruned genotype data described in the previous section. We used the pruned genotype data here because it was not computationally feasible to test all possible variant-variant combinations among all common variants. The top  $q$  pairwise interactions ( $q = 50, 100, \text{ or } 500$ ) with the smallest  $P$ -value were then selected and used to build the modified version of the epistatic GRM accounting only for these interactions (Ober et al. 2015) to use in the prediction analysis.

### Relaxing the MAF and LD assumptions

To achieve a more realistic test on the different models, we repeated the simulations with whole genome sequences described above using less stringent thresholds for MAF (MAF > 0.05) and LD ( $r^2 < 0.2$ ) of causal variants/interactions. This resulted in 144,537 pruned genotypes from which causal variants and interactions were drawn. These simulated phenotypes were analyzed using the same methods described above, except that we did not use a  $P$ -value threshold in the additive GWAS pre-selection of variants. Moreover, EpiGWAS was performed by testing all possible  $\binom{144,537}{2}$  variant-variant combinations.





**Fig. 2** G-BLUP analyses of purely additive or epistatic simulated traits ( $H^2 = 0.4$ ) using only the true QTLs and/or interactions. Error bars are standard errors of the mean. The y-axes are mean  $r^2$  values and the x-axes show the sample size. **a** Four additive scenarios for different numbers of QTLs. **b** 50 sign epistasis interactions. **c** 50 variance

epistasis interactions. **d** 500 sign epistasis interactions. **e** 500 variance epistasis interactions. Panels (**b–e**) show results from three models: additive (ADD), epistatic (EPI) and, combined additive and epistatic (ADD&EPI)

## Results

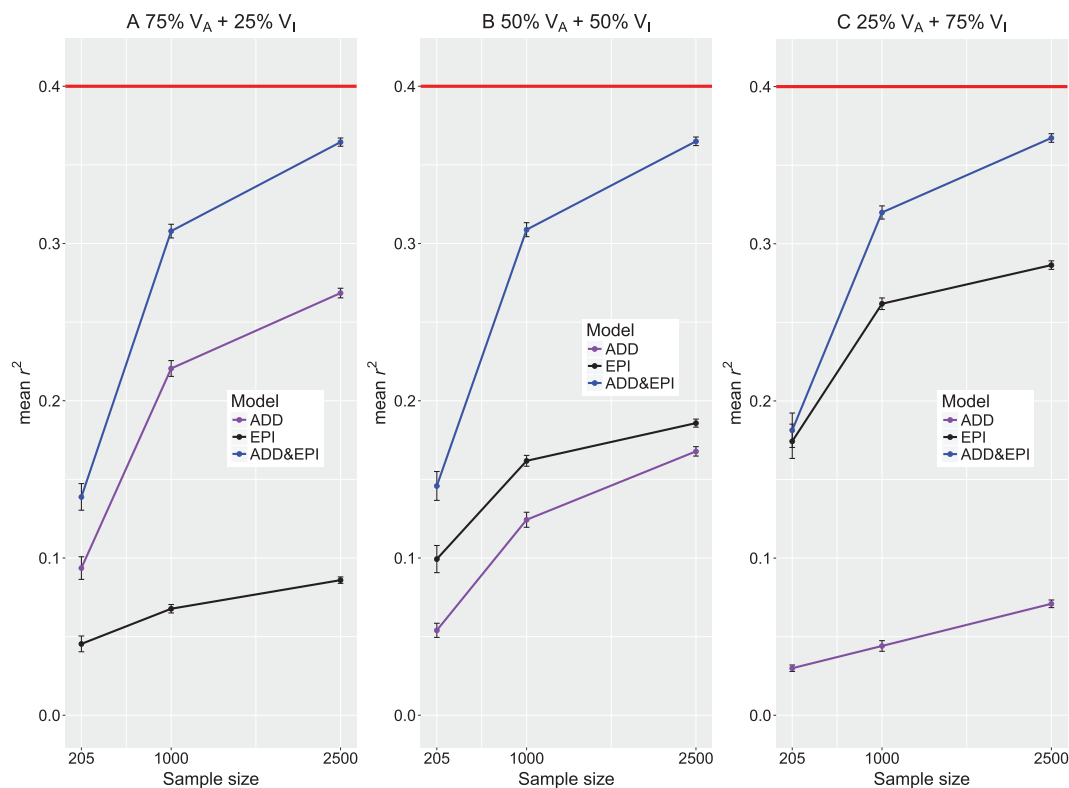
We assessed the effects of sample size, genetic architecture and different genomic prediction models on predictive ability for a population of unrelated individuals in which LD decays rapidly with physical distance. We performed all analyses for heritabilities of  $H^2 = 0.4$  and  $H^2 = 0.9$ . The former represents a heritability typical of many quantitative traits assessed on an individual basis in natural outbreeding populations. The latter represents a very high heritability, which can occasionally occur in nature (e.g., human height (Manolio et al. 2009)) but are more common in replicated mapping populations such as recombinant inbred lines, the DGRP, and naturally inbreeding or clonally reproducing organisms, where many individuals of the same genotype can be measured. We present the results for  $H^2 = 0.4$  in the main text and those for  $H^2 = 0.9$  in the supplementary material since the conclusions from both analyses are qualitatively similar.

### Prediction accuracy when true QTLs/interactions are known

We first assessed the accuracy of genomic prediction using G-BLUP when the true QTLs/interactions are known and used to build the GRMs, for different genetic architectures (purely additive, purely epistatic, and three different

proportions of additive and epistatic architectures), for sample sizes ranging from 205 (the size of the DGRP) to 5000. These simulations show how genomic prediction models perform in an idealized situation where all true QTLs are known. As expected, the additive model performed well when the true genetic architecture was additive, but required larger sample sizes to approach the upper bound for prediction accuracy as the number of QTLs increased (Fig. 2a). For example, with  $H^2 = 0.4$  and a single QTL explaining all the genetic variance, a sample size of 205 was sufficient to achieve the upper bound of prediction accuracy. However, with 1000 QTLs, a sample size of 205 yielded very poor prediction accuracy. Prediction accuracy increased significantly for the 1000 QTL scenario with increasing sample size, but even with a sample size of 5000 the average prediction accuracy was only 0.27, ~68% of the theoretical maximum (Fig. 2a).

In contrast, when the true genetic architecture was determined by 50 pairs of loci with sign epistasis (Fig. 1a), an additive model with a GRM built from the causal variants but ignoring the genetic architecture produced a prediction accuracy of  $r^2 = 0$  regardless of sample size (Fig. 2b). The epistatic GRM, however, gave good prediction accuracy, even with smaller sample sizes. The prediction accuracies from the combined additive and epistatic model were in this case indistinguishable from those of the epistatic model alone. These results are expected for sign



**Fig. 3** G-BLUP analyses on the simulated traits ( $H^2 = 0.4$ ) with both additive and sign epistasis components, using only the true QTLs and/or interactions. The y-axes are mean  $r^2$  values and the x-axes show the sample size. Error bars are standard errors of the mean. **a** Results for a trait for which 75 and 25% of the total genetic variance is additive and epistatic, respectively. **b** Results for a trait for which 50% of the total

genetic variance is additive and 50% is epistatic. **c** Results for a trait for which 25 and 75% of the total genetic variance is additive and epistatic, respectively. Three models are fitted for each scenario: additive (ADD), epistatic (EPI) and combined additive and epistatic (ADD&EPI)

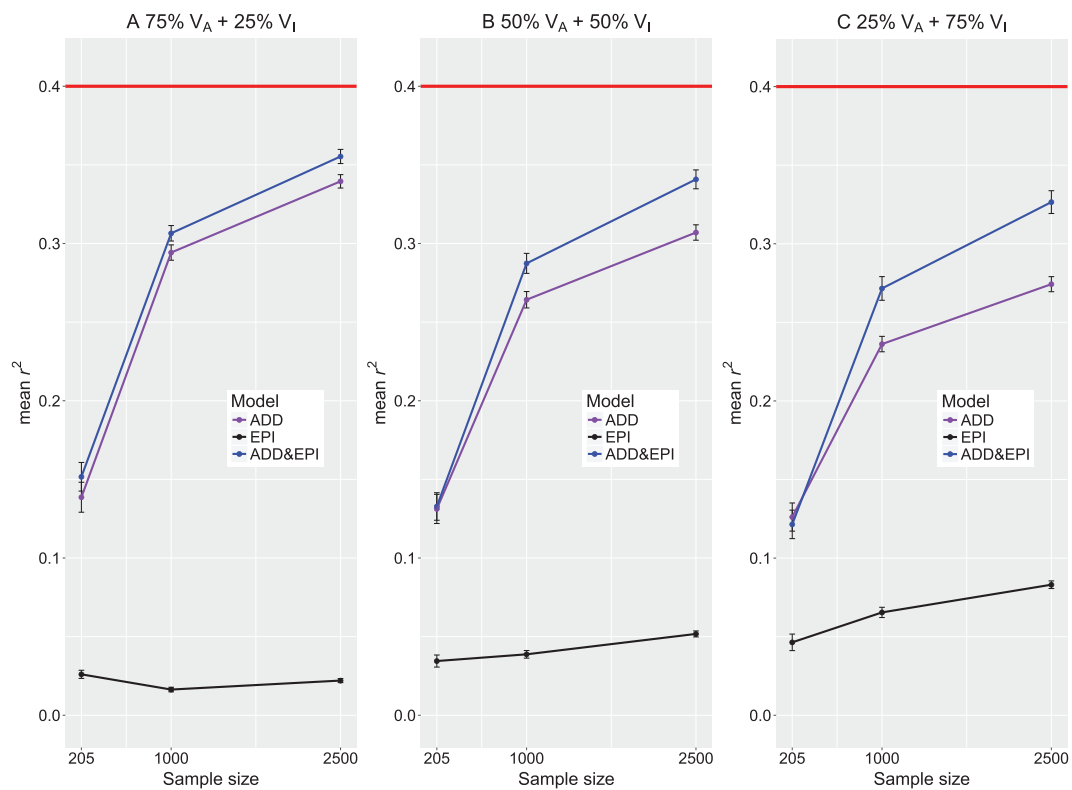
epistasis and intermediate allele frequencies of interacting loci, as these conditions produce negligible additive genetic variance.

When the true genetic architecture was defined by 50 pairs of loci with variance epistasis (Fig. 1b), an additive model with a GRM built from the causal interacting variants produced significant prediction accuracy (Fig. 2c). However, in this case the prediction accuracy of the additive model appeared to quickly reach an asymptote whereby increasing the number of lines did not correspond to an increase in accuracy; this asymptote fell well below the upper bound of accuracy. The same trend was observed when using the epistatic GRM, although with even lower accuracy. However, the combined additive and epistatic model achieved much higher prediction accuracy than the summation of the accuracies of the separate additive and epistatic models. These results are not surprising because variance epistasis generates both additive and epistatic variance with intermediate frequencies of interacting loci.

Increasing the numbers of epistatically interacting loci to 500 yielded qualitatively similar patterns (Fig. 2d–e). However, larger sample sizes were required to get closer to

the upper bound of prediction accuracy compared to the 50 pairs of loci scenario, similar to the results of increasing the number of causal loci in the additive scenarios.

We next explored the effect of mixed additive and epistatic genetic architectures for both sign and variance epistasis. We varied the sample size from 205 to 2500 and kept the genetic architecture constant at 100 additive QTLs and 50 interacting pairs of QTLs, and assessed different contributions to additive genetic and epistatic variance (additive:epistatic variance ratios of 75:25, 50:50, and 25:75) (Figs. 3 and 4). In all cases, the combined additive and epistatic model gave consistently better prediction accuracies than the pure additive or epistatic models. However, the single component models achieved different performances depending on whether sign or variance epistasis was modeled, as well as the percentage of total genetic variance explained by additive QTLs and interactions. For example, the additive model gave low prediction accuracies even for large sample sizes when the proportion of genetic variance explained by additive QTLs and interactions was 0.25:0.75 and sign epistasis was modeled (Fig. 3c). In this case, the epistatic model gave much higher prediction



**Fig. 4** G-BLUP analyses on the simulated traits ( $H^2 = 0.4$ ) with both additive and variance epistasis components, using only the true QTLs and/or interactions. The y-axes are the mean  $r^2$  values and the x-axes show the sample size. Error bars are standard errors of the mean. **a** Results for a trait for which 75 and 25% of the total genetic variance is additive and epistatic, respectively. **b** Results for a trait for which 50%

of the total genetic variance is additive and 50% is epistatic. **c** Results for a trait for which 25% and 75% of the total genetic variance is additive and epistatic, respectively. Three models are fitted for each scenario: additive (ADD), epistatic (EPI) and combined additive and epistatic (ADD&EPI)

accuracies. However, the additive model gave high prediction accuracies when the proportion of genetic variance explained by additive QTLs and interactions was 0.75:0.25 and variance epistasis was modeled (Fig. 4a).

Similar patterns were observed for the pure and mixed genetic architectures when  $H^2 = 0.9$  (Figs. S1-S3), but smaller sample sizes were needed to achieve proportionally higher accuracy. For the case of truly additive QTLs and an additive GRM, these observations agree with predictions from deterministic formulae and simulation studies (Godard 2009; Daetwyler et al. 2010). For example, with sample size of 5000 and 1000 QTLs, prediction accuracy averaged 0.87, 97% of the maximum (Fig. S1).

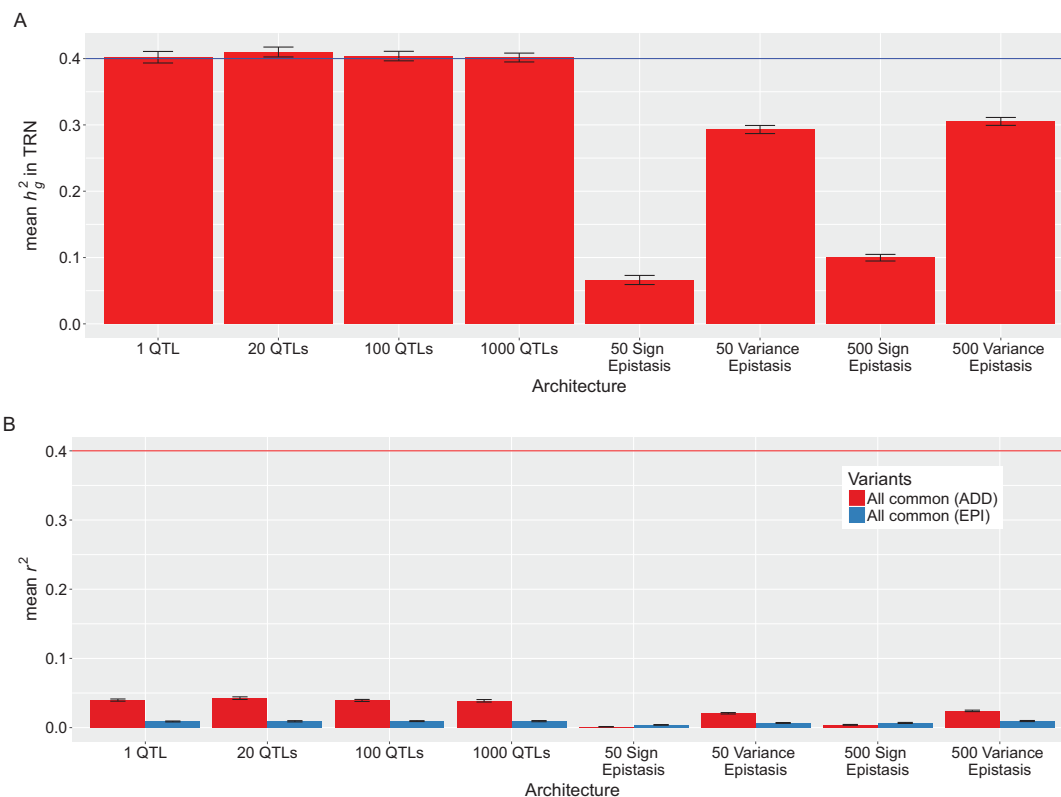
In summary, we show that when true QTLs and genetic architectures are known, which represents a situation where standard GWAS are perfect and could identify all and only causal variants, taking account of the genetic architecture always improves prediction accuracy for all sample sizes, and the combined additive and epistatic model gives highest prediction accuracies when there is epistasis. Further, applying the additive model when there is considerable sign

epistasis gives poor predictive ability no matter how large the sample size is. In the presence of variance epistasis, the predictive ability of the additive model improves, but the asymptotic predictive ability with increasing sample size is much less than the theoretical maximum given by the heritability.

### Prediction accuracy using all common variants when true QTLs/interactions are unknown

In reality, we do not know the true QTLs, and therefore we must use markers as proxies in genomic prediction analyses. To investigate this more realistic scenario, we simulated whole genomes for 10,000 individuals using estimates of effective population size and recombination and mutation rates to generate a distribution of pairwise genomic relationships and LD decay similar to that observed in the DGRP. LD decayed rapidly with physical distance and asymptoted to a low background level (Fig. S4), and the distribution of the off-diagonal elements of the additive GRM shows that the individuals were unrelated (Fig. S5).





**Fig. 5** G-BLUP analyses of purely additive or epistatic simulated traits ( $H^2 = 0.4$ ) using all common variants (MAF > 0.05). Error bars are standard errors of the mean. **a** The genomic heritability,  $h_g^2$ , computed in the training set (TRN). The y-axis is the mean  $h_g^2$  and the x-axis gives the different simulated architectures. **b** The predictive ability on

the test set. The y-axis is the mean  $r^2$  value and the x-axis gives the different simulated architectures. Red bars show the predictive ability calculated using an additive model (ADD) and blue bars show the predictive ability calculated using an epistatic model (EPI)

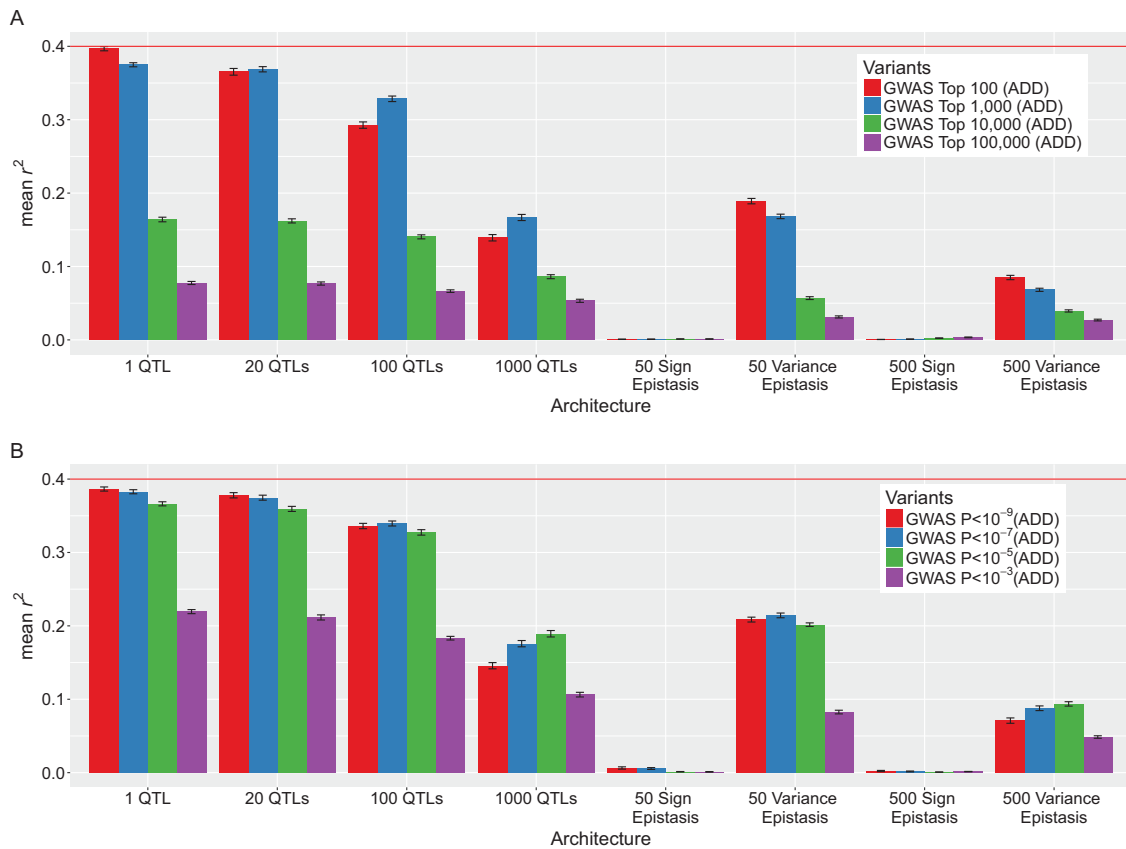
We again simulated different levels of genetic complexity (1, 20, 100, and 1000 additive QTLs); 50 and 500 interacting pairs of QTLs with sign epistasis; and 50 and 500 interacting pairs of QTLs with variance epistasis. We did not simulate the mixed additive and epistatic architectures as the previous analyses indicated the results from these analyses would fall within the boundaries set by the pure models.

First, we asked to what extent the additive GRM using all common markers explained the heritability in the training data. For the additive genetic architectures, the additive model could recover all the theoretical genetic variation (Fig. 5a, Fig. S6A), as expected for a true additive architecture and sequence data (de los Campos et al. 2015). However, applying the additive GRM to training data in which the true genetic architecture was 50 pairs of loci with sign epistasis only explained ~16% of the total heritability, which increased to ~73% when the true genetic architecture was 50 pairs of loci with variance epistasis. Similar results were obtained for 500 interactions (Fig. 5a, Fig. S6A). Next, we assessed the prediction accuracy of the same additive GRM using all common markers for the different

genetic architectures, and found it was universally very low and 0 for the sign epistasis genetic architectures (Fig. 5b, Fig. S6B). In agreement with previous studies, additive architecture with drastically different numbers of true QTLs had similar prediction accuracies using G-BLUP (Daetwyler et al. 2010). We observed a similar result for the epistatic architectures with different number of interactions. Finally, the pure epistasis GRM using all common markers gave even lower prediction accuracies than the additive model in all scenarios except the sign epistasis architecture (Fig. 5b, Fig. S6B). The pattern of results is similar regardless of the heritability; except that proportionately higher accuracies are achieved with the higher heritability.

### Prediction accuracy accounting for trait genetic architecture using GWAS when the true QTLs/interactions are unknown

One plausible reason for the low predictive ability when using all common markers when the genetic architectures are truly additive is the low signal to noise ratio—there are many more markers than QTLs, and the individuals are



**Fig. 6** G-BLUP analyses of purely additive or epistatic simulated traits ( $H^2 = 0.4$ ) using different subsets of variants prioritized by GWAS and the additive model (ADD). The y-axes are mean  $r^2$  values and the x-axes give the different simulated architectures. Error bars are standard errors of the mean. **a** The top  $t$  variants ( $t = 100; 1000; 10,000; \text{ or } 100,000$ ) from GWAS were used for prediction. **b** Variants with  $P < 10^{-x}$  ( $x = 3, 5, 7, \text{ or } 9$ ) from GWAS were used for prediction. In the 50 (500) sign epistasis scenario, only 13 (6) and 8 (2) replicates had at least 1 variant with  $P < 10^{-7}$  and  $P < 10^{-9}$ , respectively; the average is based only on those replicates

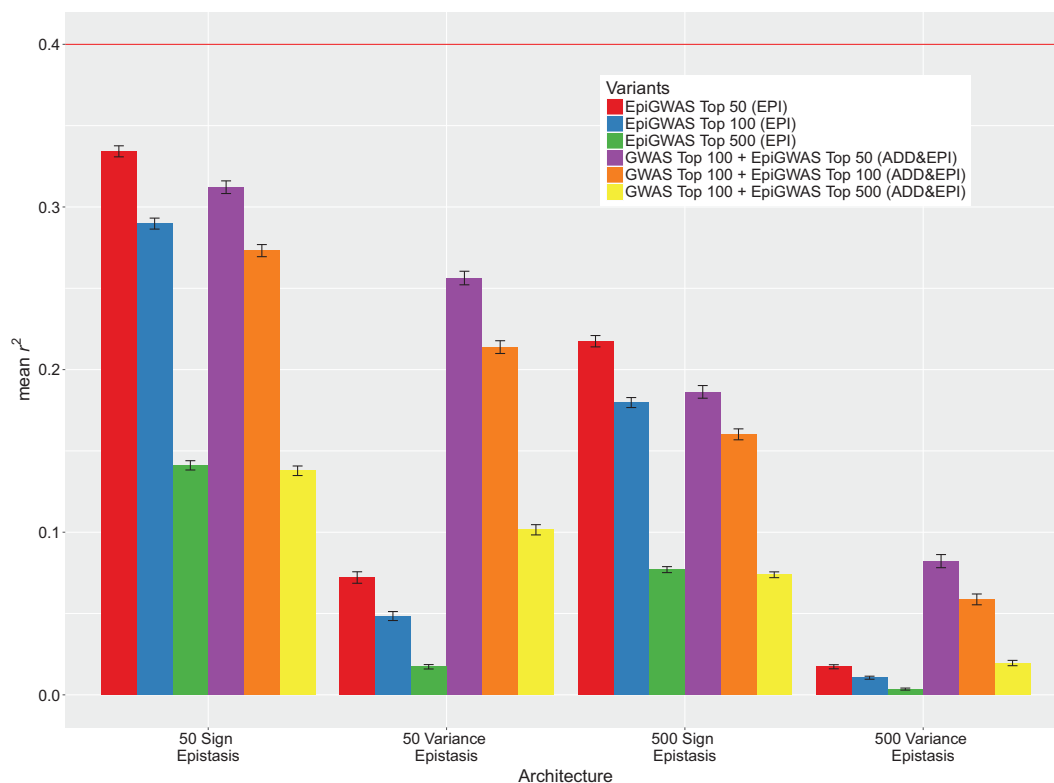
unrelated with low LD, so the true signal is diluted by the non-associated markers. Combining mapping with prediction could alleviate this problem (Zhang et al. 2015; Ober et al. 2015; Tiezzi and Maltecca 2015). Therefore, we performed a GWAS for additive effects on the training data for each of the different genetic architectures, and then used only the top variants associated with the trait to predict phenotypes in the test population. We used two different procedures to select the top variants—the top  $t$  variants or the top variants with  $P < 10^{-x}$ .

Combining mapping and prediction using an additive model worked well when the genetic architecture was additive (Fig. 6, Fig. S7). The prediction accuracies increased substantially over the additive GRM using all common variants for all the additive architectures regardless of the procedure used to select the top variants to include in the prediction models. The prediction accuracy deteriorated as more markers were added, consistent with increasing dilution of the signal. Using the top  $t$  variants or all top variants with  $P$ -values  $< 10^{-x}$  yielded similar results within

each architecture, although results were somewhat more sensitive to the number of variant threshold than to the  $P$ -value threshold (Fig. 6, Fig. S7). The average number of variants selected by the different  $P$ -value threshold is shown in Table S1.

This procedure, however, did not improve prediction accuracy when the genetic architecture was pure sign epistasis—prediction accuracy remained at 0 for all thresholds used (Fig. 6, Fig. S7). However, there was a conspicuous improvement in prediction accuracy over using all common variants when the genetic architecture was of variance epistasis, although the improvement was less pronounced than for the corresponding additive architecture (100 or 1000 QTLs) and for larger number of interactions (Fig. 6, Fig. S7). Similar to the additive scenarios, the accuracy of prediction for variance epistasis declined as the number of selected variants increased.

The pattern of results for all genetic architectures is again similar regardless of the heritability; except that proportionately higher accuracies are achieved with the higher



**Fig. 7** G-BLUP analyses of the purely epistatic simulated traits ( $H^2 = 0.4$ ) using subsets of interactions prioritized by a pairwise GWAS (EpiGWAS), or subsets of interactions prioritized by both EpiGWAS and GWAS. The top  $q$  pairwise interactions ( $q = 50, 100, \text{ or } 500$ ) from EpiGWAS, or the top  $t$  variants ( $t = 100$ ) from GWAS and the top  $q$

pairwise interactions ( $q = 50, 100, \text{ or } 500$ ) from EpiGWAS were used for prediction. The epistatic model (EPI) or the combined additive and epistatic model (ADD&EPI) were used. The y-axis is the mean  $r^2$  value and the x-axis shows the two types of simulated epistatic architectures. Error bars are standard errors of the mean

heritability, as expected from deterministic analyses (Goddard 2009; Daetwyler et al. 2010).

### Prediction accuracy using EpiGWAS and GWAS +EpiGWAS when the true QTLs/interactions are unknown

Using GWAS to select variants marginally associated with traits increased prediction accuracy for completely additive traits, but did not perform as well for epistatic traits. While there was some improvement in prediction accuracy with variance epistasis, it was not as great as that for the scenario with 100 or 1000 additive QTLs; and for sign epistasis the strategy failed completely. Because the epistatic GRM and combined additive and epistatic GRMs prediction models greatly improved prediction accuracy for epistatic scenarios when the true QTLs were known, we asked whether we could enrich for true interactions using an epistatic GWAS to select the top  $q$  interactions associated with the trait to build the epistatic GRM to use for prediction. We also tried a combined strategy performing an additive GWAS plus epistatic GWAS in the training data to select the top  $t$

variants to build the additive GRM and the top  $q$  interactions to build the epistatic GRM.

When all QTLs exhibited sign epistasis, using the top  $q$  interactions from the epistatic GWAS greatly improved prediction accuracy over all other models, for which prediction accuracy was 0 or close to 0 (Fig. 7). For  $H^2 = 0.4$ , the prediction accuracy was greatest when the number of interactions included in the epistatic GRM was 50, declined slightly as the number of interactions increased to 100, and dropped when the number further increased to 500, recapitulating the dilution effect seen previously for the additive models and additive gene action (Fig. 7). This pattern was observed when the true architecture was made of either 50 or 500 interactions but did not occur for  $H^2 = 0.9$ , where the three epistatic GWAS models yielded more similar accuracies (Fig. S8). Incorporating the additive GRM built using the top 100 additive GWAS variants to the two epistatic GWAS GRMs also gave improved prediction accuracies, but the mean  $r^2$  values were less than the corresponding pure epistatic GWAS models for  $H^2 = 0.4$  (but not for the higher heritability, where all four models had comparable accuracies) (Fig. 7, Fig. S8).

When all interactions exhibited variance epistasis, a GRM built from the top epistatic GWAS variants gave a significant prediction accuracy, but it was much lower than the best additive model for either heritability (Figs. 6 and 7, Figs. S7, S8). However, the combined additive and epistatic GRMs performed much better than either the additive or the epistatic GWAS models when the true architecture consisted of 50 interactions for both heritabilities, and 500 interactions for  $H^2 = 0.9$ . However, when the true architecture consisted of 500 pairwise interactions with  $H^2 = 0.4$ , the combined additive and epistatic GRM model did not give any improvement over the additive GWAS model for traits (Fig. 7, Fig. S8). These results are consistent with the theoretical expectation that variance epistasis produces both additive and epistatic variance when the causal variants are common (Mackay 2014).

### Prediction accuracy with less stringent MAF and LD thresholds

In the additive scenarios, the results obtained for architectures with causal  $MAF > 0.05$  were very similar to those with causal  $MAF > 0.25$  for all analyses (Figs. S9, S10, S12, S13). The additive model with all common variants explained all the genetic variance in the training set, but prediction accuracy in the test set remained very low (Figs. S9, S12), and increased when only the top  $t$  variants from GWAS were used (Figs. S10, S13). However, the additive model with all common variants explained a large amount of genetic variance for both sign and variance epistasis when the MAF threshold was lowered. This was expected since epistatic interactions generate increasing amounts of additive variance as allele frequencies decrease (Hill et al. 2008; Mackay 2014) (Figs. S9, S12). This model still gave low prediction accuracies (Figs. S9, S12). Using only the top  $t$  variants from additive GWAS provided a great improvement in prediction accuracy in all epistatic scenarios, including sign epistasis, for the same reason (Figs. S10, S13).

For cases of sign and variance epistasis, using only the top  $q$  interactions from epistatic GWAS provided an improvement in prediction accuracy over the additive model with all common variants, but accuracy was generally lower than using the additive GWAS GRM (except for sign epistasis with 50 interactions) (Figs. S11, S14). However, the combined additive and epistatic GWAS GRMs performed better than either the additive GWAS GRM or the epistatic GWAS GRM for all scenarios except for variance epistasis with 500 interactions and  $H^2 = 0.4$ , where the additive GWAS model performs best (Figs. S11, S14).

As shown in Table S2, when using all common variants predictions are unbiased (i.e., the slope of the regression is not significantly different from 1) regardless the heritability,

genetic architecture and causal MAF. However, when using preselected variants/interactions, predictions are generally biased. While this observation is consistent with the literature (e.g., Veerkamp et al. 2016), the problem of biased predictions can be solved using the method of Kim et al. (2017).

## Discussion

Here, we have shown how genetic architecture affects prediction accuracy of complex traits. In a population of unrelated individuals and low LD, the additive GRM built using all common variants (~1,800,000) did not yield good prediction accuracy regardless of the genetic architecture of the trait, even a purely additive architecture. This is likely due to the infinitesimal assumption made by the G-BLUP model, where each marker in the GRM is assumed to contribute the same amount of information to the calculation of relationships among individuals (de Los Campos et al. 2013; Ober et al. 2015). This assumption was not met in the simulated traits and is generally not realistic for many complex traits. Thus, the true signal of causal loci is diluted and masked by 'noise' from markers either not associated nor in LD with the causal loci affecting the trait—it is the relationships among individuals at causal loci for the trait under analysis that matters for prediction.

It is important to note that the same additive GRM explained all the additive genetic variance in the training population for the additive genetic architectures. This is expected from theory when the true causal variants are included in the marker panel, as when genotypes are inferred from genome sequence data, and when markers and causal variants have the same MAF and LD properties (de los Campos et al. 2015). A similar inference has been drawn from actual data on human height in samples of unrelated individuals, where an additive model gave  $h^2_g \approx 0.4$  but a prediction  $r^2 \approx 0.03$  (de Los Campos et al. 2013) using 400,000 markers. Therefore, while sequencing data may increase the genomic heritability for completely additive traits, an equivalent increase in prediction accuracy (the ultimate objective of precision medicine and agriculture) may not occur. The ability of a statistical model built using a set of predictors to explain variation in the response variable (i.e., inference) and its ability to predict yet-to-be observed responses (i.e., prediction) are two different properties that should not be confused. In fact, a model that can explain substantial variation in the response might not be the best for predicting future observations, and vice versa (Shmueli 2010). Recently, much higher prediction accuracies have been obtained for human height using extremely large datasets (Kim et al. 2017). However, these higher accuracies were obtained using both explicit and implicit

variable selection methods, therefore agreeing with our hypothesis.

In addition to having no or extremely low predictive ability, the additive model with all common variants could recover only a fraction of the total heritability for completely epistatic traits, especially the sign type. This observation is in accord with theoretical expectation, and may be related to the missing heritability (the difference between pedigree- $h^2$  and  $h^2_g$ ) observed for many complex traits because epistasis can inflate estimates of pedigree-based heritability (Manolio et al. 2009; Zuk et al. 2012; Mackay 2014).

Since different traits likely have different genetic architectures and this difference is not accounted for by the G-BLUP model, we sought to enrich the model for true causal variants/interactions. We did this by combining mapping (additive and/or epistatic GWAS on the training population) and prediction such that the results from mapping (the top variants/interactions) serve as prior information for G-BLUP through architecture-specific additive and epistatic GRMs (Ober et al. 2015). In the additive scenarios, using the results of an additive GWAS considerably improved prediction accuracy for any number of QTLs. The architectures with fewer QTLs benefited the most from this procedure due to greater power to map true causal variants.

With completely epistatic architectures, using only an additive GWAS was often not sufficient to achieve a reasonable prediction accuracy, especially with sign epistasis and higher causal MAF. Although this is an extreme case that is unlikely to occur in nature, it clearly demonstrates the possibility of such failure when models are misspecified. Prediction accuracies using an additive GWAS were increased for variance epistasis relative to the sign epistasis case, but were less than the corresponding additive architecture. While an epistatic GWAS GRM alone provided high prediction accuracy for sign epistasis when the causal MAF was greater than 0.25, this was not true for sign epistasis with causal MAF greater than 0.05, and variance epistasis with either causal MAF threshold. However, fitting a model with both an additive GWAS GRM and an epistatic GWAS GRM generally performed better than either the additive GWAS or epistatic GWAS GRM alone.

These results clearly show how accounting for the genetic architecture of complex traits may help predict future observations. The fact that the additive GRM with all common variants has worked well, for example, within dairy cattle breeds for most traits is likely attributable to closely related individuals and high LD whereby each marker is potentially in LD with at least one causal variant, simulating the infinitesimal model assumption. This condition, however, is not satisfied in samples of unrelated individuals (such as human populations or *Drosophila* lines) and leads to low predictive ability. Previous studies

(de los Campos et al. 2013; Ober et al. 2015) showed that prioritizing variants/interactions via GWAS to use in prediction could improve prediction accuracy. Here, we investigated a wide variety of simulated scenarios with different genetic architectures and showed that this approach works. These scenarios were for very simplified and extreme conditions that are unlikely to be present in nature and may have overemphasized the importance of epistasis in prediction analysis. Nonetheless, they served our purpose as a 'stress test' to demonstrate that the additive infinitesimal model may not always work, and should stimulate further investigations to evaluate more realistic scenarios.

Some of the limitations of the method used here are the need for a reasonable sample size providing mapping power, and an LD structure that allows mapping resolution and minimizes confounding of founder effects. This is particularly true for mapping interactions, as highlighted when the true architecture consisted of 500 interactions. In fact, higher accuracy was obtained when selecting only the top 50 mapped interactions and decreased when adding more interactions, suggesting that the additionally mapped interactions contained spurious associations. This point is also illustrated by the higher accuracy of the additive GWAS GRM only over the combined additive and epistatic GWAS GRMs for 500 variance epistatic interactions with  $H^2 = 0.4$ . This, however, did not occur with  $H^2 = 0.9$ , where each interaction explains a higher proportion of the total variance. Clearly, larger sample sizes are needed to map interactions precisely. Other methods which use prior biological information to account for genetic architecture in prediction do not suffer for these limitations, but require a well annotated genome (Edwards et al. 2016).

The role of epistasis in complex trait genetics is controversial. While there is substantial evidence, especially from model organisms, that epistatic interactions are an integral part of the genetic architecture of complex traits (Mackay 2014), they are expected to translate mostly into additive genetic variation if the variants making up the interactions have low frequencies (Hill et al. 2008; Mäki-Tanila and Hill 2014). However, evidence is accumulating that causal variants may be more common than previously thought (Fuchsberger et al. 2016). Thus, we should not neglect epistasis a priori in complex trait analysis, even though it is hard to map in outbred populations. With respect to prediction, we have shown that even with variance epistasis, where most genetic variation is indeed additive, prediction accuracy obtained with additive models (with or without variable selection) was generally low. Remarkably, when we used the two variants that made up the true causal interactions to build the additive GRM, an asymptote well below the upper bound of accuracy was reached, and increasing the sample size did not improve



prediction accuracy. On the other hand, using the combined additive and epistatic model immediately increased prediction accuracy, which approached the upper bound when using the true causal interactions. Thus, whenever it is present, accounting for epistasis has the potential to improve prediction accuracy of complex traits. The magnitude of the improvement will largely depend on causal MAF, with higher values benefiting the most.

Our analysis also highlights the relative importance of sample size and statistical model. When the statistical model used for the analysis closely matches the true model that generated genetic variation in the trait, larger sample sizes improve prediction accuracy of the more complicated architectures. However, larger sample size does not translate to improved prediction accuracy when the statistical model deviates substantially from the true biological model. Although the cost of genotyping and sequencing has decreased tremendously and allows studies with large sample sizes, equivalent efforts in modeling directed towards identifying the causal variants and their mode of action are needed.

Variance components and prediction analysis are not suitable to infer the genetic architecture of complex traits (Falconer and Mackay, 1996; Huang and Mackay 2016). An additive model could capture most genetic variation and yield significant prediction accuracy for a trait that was in fact purely epistatic (for the case of variance epistasis with any causal MAF or sign epistasis with lower causal MAF). In addition, different parameterizations of the genetic effects (i.e. additive, dominant and epistatic) can lead to different partitions of the total genetic variance into its components for the same data (Huang and Mackay 2016). This makes it very difficult to obtain a reliable estimate of the relative importance of additive and epistatic gene action underlying quantitative trait variation. This motivated our choice to vary the relative proportion of additive and epistatic variance from one extreme to the other.

This study has some limitations. First, we limited our simulation to pairwise interactions because it is extremely computationally challenging to map higher order interactions. However, there is no reason to assume that only pairwise interactions affect complex traits, and high order interactions have been empirically validated (Taylor and Ehrenreich 2015). Possibly, pairwise epistasis is an emergent property of higher order epistasis, similar to additivity being an emergent property of pairwise epistasis (Mackay 2014). Hence, performing a pairwise scan, which can be performed in a semi-exhaustive way in a reasonable amount of time, may be able to capture higher order interactions. Second, to make the scenarios as simple as possible, our simulated genotypes were based on inbred lines, for which only four genotypes segregate with pairwise epistasis, and all interactions are of the additive by additive type. (In

outbred populations there are nine possible genotypes for pairwise interactions, and four possible types of epistatic interactions involving additive and dominance effects at both loci).

## Data Archiving

The code used to simulate all the data can be found in the Supplementary material.

**Acknowledgements** This work was supported by National Institutes of Health Grants R01 AA016560 and R01 AG043490 to TFCM and the Center for Genomic Selection in Animals and Plants (GenSAP) funded by The Danish Council for Strategic Research to FM and TFCM. FM would like to thank Dr. Peter Sørensen for help with some code, and Dr. Jeremy Howard for writing the C++ code to compute the genomic relationship matrix.

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, and provide a link to the Creative Commons license. You do not have permission under this license to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

## References

- Charlesworth B (2009) Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* 10:195–205
- Chen GK, Marjoram P, Wall JD (2009) Fast and flexible simulation of DNA sequence data. *Genome Res* 19:136–142
- Comeron JM, Ratnappan R, Bailin S (2012) The many landscapes of recombination in *Drosophila melanogaster*. *PLoS Genet* 8:33–35
- Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA (2010) The impact of genetic architecture on genome-wide evaluation methods. *Genetics* 185:1021–1031
- Daetwyler HD, Calus MPL, Pong-Wong R, de los Campos G, Hickey JM (2013) Genomic prediction in animals and plants: Simulation of data, validation, reporting, and benchmarking. *Genetics* 193:347–365
- de los Campos G, Gianola D, Allison DB (2010) Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nat Rev Genet* 11:880–886
- de los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D (2013) Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genet* 9:e1003608
- de los Campos G, Sorensen D, Gianola D (2015) Genomic heritability: what is it? *PLoS Genet* 11:e1005048

- De Roos APW, Hayes BJ, Goddard ME (2009) Reliability of genomic predictions across multiple populations. *Genetics* 183:1545–1553
- Edwards SM, Sørensen IF, Sarup P, Mackay TFC, Sørensen P (2016) Genomic prediction for quantitative traits is improved by mapping variants to gene ontology categories in *Drosophila melanogaster*. *Genetics* 203:1871–1883
- Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics. Longman, Harlow, Essex
- Fisher RA (1918) The correlation between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edinb* 52:399–433
- Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V, Gaulton KJ et al. (2016) The genetic architecture of type 2 diabetes. *Nature* 536:41–47
- Goddard ME, Hayes BJ (2009) Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat Rev Genet* 10:381–391
- Goddard M (2009) Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* 136:245–257
- Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177:2389–2397
- Heffner EL, Sorrells ME, Jannink J (2009) Genomic selection for crop improvement. *Crop Sci* 49:1–12
- Hill WG, Goddard ME, Visscher PM (2008) Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet* 4:e1000008
- Huang W, Massouras A, Inoue Y, Peiffer J, Ramia M, Tarone AM et al. (2014) Natural variation in genome architecture among 205 *Drosophila melanogaster* Genetic Reference Panel lines. *Genome Res* 24:1193–1208
- Huang W, Lyman RF, Lyman RA, Carbone MA, Harbison ST, Magwire MM et al. (2016) Spontaneous mutations and the origin and maintenance of quantitative genetic variation. *Elife* 5:e14625
- Huang W, Mackay TFC (2016) The genetic architecture of quantitative traits cannot be inferred from variance component analysis. *PLoS Genet* 12:e1006421
- Jiang Y, Reif JC (2015) Modeling epistasis in genomic selection. *Genetics* 201:759–768
- Kim H, Grueneberg A, Vazquez AI, Hsu S, de los Campos G (2017) Will big data close the missing heritability gap? *Genetics* 207:1135–1145
- Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F et al (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467:832–838
- Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D et al (2012) The *Drosophila melanogaster* Genetic Reference Panel. *Nature* 482:173–178
- Mackay TFC (2014) Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nat Rev Genet* 15:22–33
- Mäki-Tanila A, Hill WG (2014) Influence of gene interaction on complex trait variation with multilocus models. *Genetics* 198:355–367
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorf LA et al. (2009) Finding the missing heritability of complex diseases. *Nature* 461:747–753
- Martini JWR, Wimmer V, Erbe M, Simianer H (2016) Epistasis and covariance: how gene interaction translates into genomic relationship. *Theor Appl Genet* 129:963–976
- Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829
- Ober U, Huang W, Magwire M, Schlather M, Simianer H, Mackay TFC (2015) Accounting for genetic architecture improves sequence based genomic prediction for a *Drosophila* fitness trait. *PLoS One* 10:1–17
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575
- R Core Team (2015) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- Schüpbach T, Xenarios I, Bergmann S, Kapur K (2010) FastEpistasis: a high performance computing solution for quantitative trait epistasis. *Bioinformatics* 26:1468–1469
- Shmueli G (2010) To explain or to predict? *Stat Sci* 25:289–310
- Su G, Christensen OF, Ostersen T, Henryon M, Lund MS (2012) Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. *PLoS One* 7:1–7
- Taylor MB, Ehrenreich IM (2015) Higher-order genetic interactions and their contribution to complex traits. *Trends Genet* 31:34–40
- Tiezzi F, Maltecca C (2015) Accounting for trait architecture in genomic predictions of US Holstein cattle using a weighted realized relationship matrix. *Genet Sel Evol* 47:24
- Veerkamp RF, Bouwman AC, Schrooten C, Calus MP (2016) Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein–Friesian cattle. *Genet Sel Evol* 48:95
- Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR et al. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 42:565–569
- Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: A tool for genome-wide complex trait analysis. *Am J Hum Genet* 88:76–82
- Zhang Z, Erbe M, He J, Ober U, Gao N, Hao Z et al. (2015) Accuracy of whole genome prediction using a genetic architecture enhanced variance-covariance matrix. *G3: Genes Genom Genet* 5: 615–627
- Zuk O, Hechter E, Sunyaev SR, Lander ES (2012) The mystery of missing heritability: genetic interactions create phantom heritability. *Proc Natl Acad Sci* 109:1193–1198