OPEN ACCESS Freely available online

PLOS ONE

Enhancing Genome-Enabled Prediction by Bagging Genomic BLUP

Daniel Gianola[1,2,3]*, Kent A. Weigel[2], Nicole Krämer[4], Alessandra Stella[5], Chris-Carolin Schön[4]

1 Department of Animal Sciences, University of Wisconsin-Madison, Madison, Wisconsin, United States of America, **2** Department of Dairy Science, University of Wisconsin-Madison, Madison, Wisconsin, United States of America, **3** Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin, United States of America, **4** Department of Plant Breeding, Technical University of Munich, Weihenstephan Center, Munich, Germany, **5** Bioinformatics and Statistics Unit, Parco Tecnologico Padano, Lodi, Italy

Abstract

We examined whether or not the predictive ability of genomic best linear unbiased prediction (GBLUP) could be improved via a resampling method used in machine learning: bootstrap aggregating sampling (''bagging''). In theory, bagging can be useful when the predictor has large variance or when the number of markers is much larger than sample size, preventing effective regularization. After presenting a brief review of GBLUP, bagging was adapted to the context of GBLUP, both at the level of the genetic signal and of marker effects. The performance of bagging was evaluated with four simulated case studies including known or unknown quantitative trait loci, and an application was made to real data on grain yield in wheat planted in four environments. A metric aimed to quantify candidate-specific cross-validation uncertainty was proposed and assessed; as expected, model derived theoretical reliabilities bore no relationship with cross-validation accuracy. It was found that bagging can ameliorate predictive performance of GBLUP and make it more robust against overfitting. Seemingly, 25–50 bootstrap samples was enough to attain reasonable predictions as well as stable measures of individual predictive mean squared errors.

Citation: Gianola D, Weigel KA, Krämer N, Stella A, Schön C-C (2014) Enhancing Genome-Enabled Prediction by Bagging Genomic BLUP. PLoS ONE 9(4): e91693. doi:10.1371/journal.pone.0091693

Editor: Qin Zhang, China Agricultrual University, China

Received December 18, 2013; **Accepted** February 13, 2014; **Published** April 10, 2014

Funding: Research was partially supported by the Federal Ministry of Education and Research (BMBF, Germany) within the AgroClustEr 15 Synbreed-''Synergistic plan and animal breeding'' (FKZ 03115528A), by a U.S. Department of Agriculture Hatch Grant (142- PRJ63CV) to DG, and by the Wisconsin Agriculture Experiment Station. The funders had no role on study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: gianola@ansci.wisc.edu

Introduction

A method for whole-genome enabled prediction of quantitative traits known as GBLUP, standing for ''genomic best linear unbiased prediction'', was seemingly suggested first by Van Raden [1–2]. In GBLUP, a pedigree-based relationship matrix among individuals [3] is replaced by a matrix valued measure of genomic similarities constructed using molecular markers, such as single nucleotide polymorphisms (SNPs) [4]. This ''genomic relationship'' matrix or variants thereof [5–6], referred to as $\mathbf{G}$ hereinafter, defines a covariance structure among individuals (even if genetically unrelated in the standard sense), stemming from ''molecular similarity'' in state at additive marker loci among members of a sample. Given $\mathbf{G}$ and values of some needed variance components, one can use the theory of best linear unbiased prediction (BLUP) to obtain point and interval (e.g., prediction error variances) estimates of the genetic values of a set of individuals as marked by the battery of SNPs. While $\mathbf{G}$ can be constructed in different manners we do not address this issue here. However, we note that it is possible to separate ''genomic'' from ''residual'' variance components statistically even in the absence of genetic relatedness. Hence, care must be exercised when relating the ''genomic'' to the ''genetic'' variance; this is discussed in [7].

Using theory developed by Henderson [8–10] it can be shown that GBLUP is equivalent to a linear regression model on additive genotypic codes of markers, with the allelic substitution effects at marker loci treated as drawn independently from a distribution possessing a constant variance over markers [11–13]. There is also an equivalence between pedigree-based BLUP or G-BLUP (or of models using both pedigree and marker relationships) and non-parametric regression [14]. For instance, if the $n \times p$ marker matrix $\mathbf{X}$ (n = number of observations, p = number of markers) is used to construct a kernel matrix $\mathbf{XX}'$, it can be established that GBLUP is a reproducing kernel Hilbert spaces regression procedure [15–16]. Also, BLUP and GBLUP can be represented as linear neural networks where inputs are entries of the pedigree-based or $\mathbf{G}$ matrices, respectively [17]. Hence, GBLUP can be motivated from several different perspectives.

There are many competing procedures for genome-enabled prediction, such as the members of the growing Bayesian alphabet [14],[18–19], but most of these require a Bayesian Markov chain Monte Carlo (MCMC) implementation. On the other hand, GBLUP is simple, easy to understand, explain and compute, and there is software available for likelihood-based variance component estimation and for prediction of random effects. Also, GBLUP handles cross-sectional and longitudinal data flexibly and extends to multiple-trait settings in a direct manner. Further, GBLUP delivers a competitive predictive ability since members of the Bayesian alphabet typically differ by little in predictive performance and differences among methods are typically masked by cross-validation noise [19–23]. Last but not least, some members of this alphabet produce predictions that are sensitive

to hyper-parameter specification [24]. Given these considerations, GBLUP or extensions thereof [25] are good candidates for routine whole-genome prediction in animal and plant breeding applications and possibly for prediction of complex traits in humans as well [7],[26].

The closeness between predictions and realized values depends mainly on three factors: prediction bias, variance of prediction error and amount of noise associated with future observations. The latter cannot be reduced by any prediction machine based on training data, so it is impossible to construct predictors attaining a perfect predictive correlation, even if the model holds. Theoretical and empirical results [1], [27–30] indicate that the proportion of (cross-validation) variance explained by a linear predictor increases up to a point with training sample size, then reaching a plateau. However, when a small number of individuals is available, any prediction machine is bound to produce predictions with a large variance. In this context, it seems worthwhile to explore avenues for enhancing accuracy (i.e., reduce bias) and reliability (i.e., decrease variance) of predictions when training size is small.

The question examined here is whether or not the predictive ability of GBLUP can be improved by recourse to resampling methods used in machine learning. These include bootstrap aggregating sampling ("bagging") and iterated bagging or "debiasing" [31–34]. Bagging uses bootstrap sampling to reduce variance and can enhance reliability and reduce mean squared error [31]. The conditional bias of GBLUP [19] cannot be removed by bagging, but iterated bagging has the potential of reducing variance while removing bias simultaneously. This study investigates bagging of GBLUP, with consideration of debiasing deferred to a future investigation. The second section of this paper gives a review of GBLUP and of its inherent inaccuracy (bias). The third section describes bagging in the context of GBLUP at the level of the genetic signal and of marker effects. The fourth section examines the performance of bagging in four simulated case studies, and the fifth section presents an application to real data on grain yield in wheat planted in four environments. The paper concludes with a discussion and with a proposal of a metric aimed to quantify candidate-specific cross-validation uncertainty.

Materials and Methods

GBLUP

Idealized conditions. Assume that effects of nuisance factors (e.g., year to year variation) have been removed in a pre-processing stage (this can also be done in a single-stage, but we ignore this for simplicity). GBLUP can be motivated by positing the linear regression model on markers

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{e}, \tag{1}$$

where $\mathbf{y}$ is an $n\times 1$ vector of observations or pre-processed data measured on a set of individuals or lines; $\mathbf{X}$ is an $n\times p$ matrix of marker genotypes, with its typical element x_{ij} being the genotype code at locus j observed in individual i, and with $rank(\mathbf{X})\leq \min(n,p)$; $\boldsymbol{\beta}$ is a $p\times 1$ vector of unknown allelic substitution effects when marker genotypes are coded, e.g., as $-1,0$ and 1 for *aa*, *Aa* and *AA* at locus *A*, say, or when these coded values are deviated from the corresponding column means or standardized. Above, $\mathbf{e}\sim N\left(\mathbf{0},\mathbf{N}\sigma_e^2\right)$ is a vector of residuals where σ_e^2 is the residual variance and $\mathbf{N}$ is an $n\times n$ diagonal matrix with typical element N_{ii}; if $\mathbf{y}$ consists of single measurements on individuals, $N_{ii}=1$ for all $i=1,2,...,n$, and if $\mathbf{y}$ is a vector of means, N_{ii} would be the number of observations entering into the mean. In BLUP, a distribution is assigned to $\boldsymbol{\beta}$ and the simplest one is $\boldsymbol{\beta}|\sigma_\beta^2\sim N\left(\mathbf{0},\mathbf{I}\sigma_\beta^2\right)$, where σ_β^2 is the variance of marker allele substitution effects. Using this assumption together with model (1) gives as marginal distribution of the data (after assuming that $\mathbf{X}$ and $\mathbf{y}$ have been centered) $\mathbf{y}|\mathbf{X}\sim N\left(\mathbf{0},\mathbf{X}\mathbf{X}'\sigma_\beta^2+\mathbf{N}\sigma_e^2\right)$. In BLUP σ_β^2 and σ_e^2 are treated as known but these parameters are typically estimated from data at hand [35]. With markers, most often $n<p$ so it is convenient to form the best linear unbiased predictor of $\boldsymbol{\beta}$ as $\hat{\boldsymbol{\beta}}=\sigma_\beta^2\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$, where $\mathbf{V}=\mathbf{X}\mathbf{X}'\sigma_\beta^2+\mathbf{N}\sigma_e^2$. If model (1) holds, it can be shown that $\hat{\boldsymbol{\beta}}$ is unbiased in the sense that $E\left(\hat{\boldsymbol{\beta}}|\mathbf{X}\right)=E(\boldsymbol{\beta})=\mathbf{0}$. Its covariance matrix (given $\mathbf{X}$) is

$$\mathbf{V}_{\hat{\boldsymbol{\beta}}}=\sigma_\beta^4\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}, \tag{2}$$

and its prediction error covariance matrix (also covariance matrix of the conditional distribution of $\boldsymbol{\beta}$ given $\mathbf{y}$ and $\mathbf{X}$ under normality assumptions) is

$$\mathbf{V}_{\boldsymbol{\beta}|\mathbf{y}}=\mathbf{V}_{\boldsymbol{\beta}}-\mathbf{V}_{\hat{\boldsymbol{\beta}}}=\sigma_\beta^2\left(\mathbf{I}-\sigma_\beta^2\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right), \tag{3}$$

where $\mathbf{V}_{\boldsymbol{\beta}}=\mathbf{I}\sigma_\beta^2$. The diagonal elements of $\mathbf{V}_{\boldsymbol{\beta}|\mathbf{y}}$ (v_{jj}) lead to a measure of reliability of prediction of marker effect j: $rel_j=1-\frac{v_{jj}}{\sigma_\beta^2}$. A matrix of reliabilities and co-reliabilities is

$$\mathbf{R}=\mathbf{I}-\mathbf{V}_{\boldsymbol{\beta}|\mathbf{y}}\mathbf{V}_{\boldsymbol{\beta}}^{-1}=\mathbf{V}_{\hat{\boldsymbol{\beta}}}\mathbf{V}_{\boldsymbol{\beta}}^{-1}. \tag{4}$$

If one wishes to predict a future vector $\mathbf{y}_f=\mathbf{X}_f\boldsymbol{\beta}+\mathbf{e}_f$, with future residuals independent of past ones and provided that future and past residuals stem from the same stochastic process, under normality assumptions the predictive distribution [19] is

$$\mathbf{y}_f|\mathbf{y},\mathbf{X},\mathbf{X}_f\sim N\left(\mathbf{X}_f\hat{\boldsymbol{\beta}},\mathbf{X}_f\mathbf{V}_{\boldsymbol{\beta}|\mathbf{y}}\mathbf{X}_f'+\mathbf{I}_f\sigma_e^2\right). \tag{5}$$

Further, if $\mathbf{y}_f$ is the predictand and $BLUP\left(\mathbf{X}_f\boldsymbol{\beta}\right)=\mathbf{X}_f\hat{\boldsymbol{\beta}}=\hat{\mathbf{y}}_f$ is the predictor, BLUP theory yields

$$Var\left(\mathbf{y}_f|\mathbf{y}\right)=Var\left(\mathbf{y}_f\right)-Var\left(\hat{\mathbf{y}}_f\right), \tag{6}$$

so that $Var\left(\hat{\mathbf{y}}_f\right)=\mathbf{X}_f\mathbf{V}_{\hat{\boldsymbol{\beta}}}\mathbf{X}_f'$. If the marker effects could be estimated such that $\mathbf{V}_{\boldsymbol{\beta}|\mathbf{y}}\to 0$, then $Var\left(\hat{\mathbf{y}}_f\right)\to\mathbf{X}_f\mathbf{V}_{\boldsymbol{\beta}}\mathbf{X}_f'$ and $Var\left(\mathbf{y}_f|\mathbf{y}\right)\to\mathbf{I}_f\sigma_e^2$. Hence, the distribution in (5) would still have variance, indicating that it is not possible to attain a predictive correlation equal to 1 even if a training set has an infinite size (more formally, if the training process conveys an infinite amount of information about markers); [7] gives a discussion of related matters.

Blup is conditionally inaccurate

While BLUP theory is well established, quantitative geneticists tend to interpret the unbiasedness property of BLUP as if it pertained to the true unknown $\boldsymbol{\beta}$, when in fact it applies to the average of the distribution of $\boldsymbol{\beta}$, that is, 0. If $\boldsymbol{\beta}$ in (1) were viewed as a model on unknown effects of known quantitative trait loci (QTL), it is obvious that one should think in terms of a fixed effects model, as per the standard finite number of loci model of

quantitative genetics [36]. Accordingly, if effects of markers are sought because these "flag" some genomic region of interest, the random sampling assumption made in BLUP is not relevant, although it might lead to a more stable estimator. In the fixed effects case both $\boldsymbol{\beta}$ and the marked genetic signal $\mathbf{X}\boldsymbol{\beta}$ are estimated with bias by BLUP even if the model holds [19].

Markers are not QTL and the latter are "causes" of generating a signal to phenotype. Suppose that the "true model" is linear on effects ($\mathbf{q}$) of QTL relating to phenotypes via incidence matrix $\mathbf{Q}$, that is

$$\mathbf{y}=\mathbf{Q}\mathbf{q}+\mathbf{e}. \tag{7}$$

If the QTL effects $\mathbf{q}$ are viewed as fixed entities (arguably geneticists have this in mind in their quest of finding genes), $E(\mathbf{y}|\mathbf{Q}\mathbf{q})=\mathbf{Q}\mathbf{q}$. In this situation BLUP produces the following average outcomes

$$E\left(\hat{\boldsymbol{\beta}}|\mathbf{q},\mathbf{Q},\mathbf{X}\right)=E\left[\sigma^2_{\beta}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}|\mathbf{q}\right]=\sigma^2_{\beta}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Q}\mathbf{q},$$

and

$$E\left(\mathbf{X}\hat{\boldsymbol{\beta}}|\mathbf{q},\mathbf{Q},\mathbf{X}\right)=E\left[\sigma^2_{\beta}\mathbf{X}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}|\mathbf{q}\right]=\sigma^2_{\beta}\mathbf{X}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Q}\mathbf{q},$$

so the bias of the estimated signal is $\mathbf{Q}\mathbf{q}-\sigma^2_{\beta}\mathbf{X}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Q}\mathbf{q}=\left[\mathbf{I}-\sigma^2_{\beta}\mathbf{X}\mathbf{X}'\mathbf{V}^{-1}\right]\mathbf{Q}\mathbf{q}$.

On the other hand, if $\mathbf{q}$ is assigned a distribution, say, $N\left(\mathbf{0},\mathbf{I}\sigma^2_q\right)$ and the p markers are treated as random as well, e.g., $\boldsymbol{\beta}|\sigma^2_{\beta}\sim N\left(\mathbf{0},\mathbf{I}\sigma^2_{\beta}\right)$, under normality assumptions one has

$$E\left(\mathbf{q}|\hat{\boldsymbol{\beta}},\mathbf{Q},\mathbf{X}\right)=E(\mathbf{q})+Cov(\mathbf{q},\hat{\boldsymbol{\beta}}')\mathbf{V}^{-1}_{\hat{\boldsymbol{\beta}}}\hat{\boldsymbol{\beta}}=\hat{\mathbf{q}}_{\text{markers}} \tag{8}$$

where

$$\begin{aligned}Cov(\mathbf{q},\hat{\boldsymbol{\beta}}')&=Cov(\mathbf{q},\sigma^2_{\beta}\mathbf{y}'\mathbf{V}^{-1}\mathbf{X})\\&=\sigma^2_{\beta}Cov\left(\mathbf{q},\mathbf{q}'\right)\mathbf{Q}'\mathbf{V}^{-1}\mathbf{X}=\sigma^2_{\beta}\sigma^2_q\mathbf{Q}'\mathbf{V}^{-1}\mathbf{X}.\end{aligned}$$

Using this in (8)

$$\hat{\mathbf{q}}_{\text{markers}}=\sigma^2_{\beta}\sigma^2_q\mathbf{Q}'\mathbf{V}^{-1}\mathbf{X}\mathbf{V}^{-1}_{\hat{\boldsymbol{\beta}}}\hat{\boldsymbol{\beta}}$$

where $\mathbf{Q}'\mathbf{V}^{-1}\mathbf{X}$ conveys unknown linkage disequilibrium relationships between QTL and marker genotypes; similarly, the best statement about the signal is $E\left(\mathbf{Q}\mathbf{q}|\hat{\boldsymbol{\beta}},\mathbf{Q},\mathbf{X}\right)=\mathbf{Q}\hat{\mathbf{q}}_{\text{markers}}$. Unfortunately, neither σ^2_q nor $\mathbf{Q}$ are known, so statements made about QTL from markers are based on untestable assumptions, including the view that the QTL effects and the genotypes are linearly related, as in (7).

Predictive correlation when markers are the QTL

Imagine a best case scenario where the markers are the QTL, and consider predicting $\mathbf{y}_f=\mathbf{Q}_f\mathbf{q}+\mathbf{e}_f$. Here, $\mathbf{Q}_f$ is the incidence matrix relating QTL effects to yet to be realized phenotypes $\mathbf{y}_f$, and $\mathbf{e}_f$ is a future vector of residuals. BLUP theory, using $\mathbf{Q}_f\hat{\mathbf{q}}$ (here, $\hat{\mathbf{q}}$ is the $BLUP(\mathbf{q})$ under the true model) as predictor, gives the following squared correlation between the ith elements of $\mathbf{y}_f$ and $\mathbf{Q}_f\hat{\mathbf{q}}$ (below $\mathbf{Q}'_{f,i}$ is the ith row of $\mathbf{Q}_f$)

$$r^2_{ii}=\frac{Var\left(\mathbf{Q}'_{f,i}\hat{\mathbf{q}}\right)}{Var\left(y_{f,i}\right)}=\frac{\mathbf{Q}'_{f,i}Var(\hat{\mathbf{q}})\mathbf{Q}_{f,i}}{\sigma^2_q\mathbf{Q}'_{f,i}\mathbf{Q}_{f,i}+\sigma^2_e}.$$

Let now $n\rightarrow\infty$ in which case BLUP theory yields $\mathbf{Q}'_{f,i}Var(\hat{\mathbf{q}})\mathbf{Q}_{f,i}\rightarrow\mathbf{Q}'_{f,i}\mathbf{Q}_{f,i}\sigma^2_q$, so that

$$r^2_{ii}\rightarrow 1/\left[1+\sigma^2_e/\left(\sigma^2_q\mathbf{Q}'_{f,i}\mathbf{Q}_{f,i}\right)\right].$$

This shows that it is impossible to attain a perfect predictive correlation even when the markers are the QTL. Further, $\mathbf{Q}'_{f,i}\mathbf{Q}_{f,i}=\sum_{j=1}^{nq}Q^2_{f,ij}$, where $Q_{f,ij}$ is the genotype at QTL locus j ($j=1,2...,n_q$) of individual i in the testing set. If QTL genotypes are centered and assumed to be in Hardy-Weinberg equilibrium $E\left(Q^2_{f,ij}\right)=2p_j(1-p_j)$, so approximately

$$r^2_{ii}\approx\left[1+\frac{\sigma^2_e}{\sum_{j=1}^{nq}2p_j(1-p_j)\sigma^2_q}\right]^{-1}=h^2, \tag{9}$$

where p_j is the frequency of a reference allele, $\sum_{j=1}^{nq}2p_j(1-p_j)\sigma^2_q$ is additive genetic variance and

$$h^2=\frac{\sum_{j=1}^{nq}2p_j(1-p_j)\sigma^2_q}{\sum_{j=1}^{nq}2p_j(1-p_j)\sigma^2_q+\sigma^2_e},$$

is trait heritability. Hence, the predictive correlation has an upper bound at h.

If instead of predicting individual phenotypes the problem is that of predicting an average, the upper bound for the predictive correlation is higher. The corresponding formula is easy to arrive at and it just requires replacing σ^2_e in (9) by σ^2_e/n_{Ave}, the number of observations intervening in the average. Then

$$r^2_{ii}\approx\left[1+\frac{\sigma^2_e/n_{Ave}}{\sum_{j=1}^{nq}2p_j(1-p_j)\sigma^2_q}\right]^{-1}=\frac{n_{Ave}h^2}{1+(n_{Ave}-1)h^2},$$

which is heritability as used in plant breeding, or "heritability of a mean" [36]. The predictive correlation never reaches 1 but can get close to 1 if n_{Ave} is very large. Values of squared predictive correlations in the range of 0.5–0.75 have been attained in dairy cattle progeny tests where predictands are processed averages of production records of cows sired by a bull [2].

Bagging GBLUP (BGBLUP) and marker effects

Bagging GBLUP. "Bagging" exploits the idea that predictors can be rendered more stable by repeated bootstrapping and averaging over bootstrap samples, thus reducing variance [31].

Bagging has been found to have advantages in cases where predictors are unstable, i.e., when small perturbations of the training set produce marked changes in model training [31], [34], [37]; for example, with ordinary least-squares under severe multicolinearity. An important application of bagging is in prediction using random forests [38].

Prediction methods that use regularization, such as those applied in genome-enabled selection, are often stable because penalties on model complexity reduce the effective number of parameters drastically, thus lowering variance. However, this is attained at the expense of bias with respect to marker effects and to the unknown function to be predicted (marked genetic value). A priori it would seem that bagging would not bring advantages in the context of a regularized method such as GBLUP. However, this issue has not been examined so far and there may be cases, e.g., in "small" populations, where random variation in training sets of small sizes has a marked impact on predictive ability.

To motivate bagging, we recall that GBLUP is a regression of phenotypes on genomic relationships between individuals. Let $\mathbf{g}=\mathbf{X}\boldsymbol{\beta}$ be a vector of "genomic values", $\mathbf{G}\propto\mathbf{X}\mathbf{X}'$ and assume that $\mathbf{G}^{-1}$ exists; then (1) can be written in equivalent form as

$$\mathbf{y}=\mathbf{G}\mathbf{G}^{-1}\mathbf{g}+\mathbf{e}=\mathbf{G}\mathbf{g}^{*}+\mathbf{e} \tag{10}$$

where $\mathbf{g}\sim\mathbf{N}\left(\mathbf{0},\mathbf{G}\sigma^2_{\beta}\right)$ and $\mathbf{g}^{*}=\mathbf{G}^{-1}\mathbf{g}\sim\mathbf{N}\left(\mathbf{0},\mathbf{G}^{-1}\sigma^2_{\beta}\right)$. In scalar form, the datum for individual i is expressible under this parameterization as the regression

$$\begin{aligned} y_i&=\sum_{j=1}^{nq}G_{ij}g^{*}_{j}+e_i\\ &=\mathbf{G}_i'\mathbf{g}^{*}+e_i\ , \end{aligned} \tag{11}$$

where G_{ij} is an element of the $n\times n$ matrix $\mathbf{G}$ and $\mathbf{G}_i'$ is its i^{th} row. From basic principles,

$$BLUP(\mathbf{g}^{*})=Cov(\mathbf{g}^{*},\mathbf{y}')\mathbf{V}^{-1}\mathbf{y}=\sigma^2_{\beta}\mathbf{V}^{-1}\mathbf{y}=\hat{\mathbf{g}}^{*},$$

and since $\mathbf{G}\mathbf{g}^{*}=\mathbf{g}$, use of invariance gives

$$BLUP(\mathbf{g})=\mathbf{G}BLUP(\mathbf{g}^{*})=\sigma^2_{\beta}\mathbf{G}\mathbf{V}^{-1}\mathbf{y}=\hat{\mathbf{g}} \tag{12}$$

Note that $\sigma^2_{\beta}\mathbf{G}\mathbf{V}^{-1}$ is an $n\times n$ matrix of "heritabilities" and "coheritabilities"; the diagonal elements of $\mathbf{G}$ may be distinct from each other, so each individual may have a different heritability ascribed to it, as noted by de los Campos et al. (2013).

The intuitive idea behind bagging was outlined in [1]. Suppose there is a large number of training samples from the same population; by averaging over the predictions made from these samples we would end up with a reduction of variance but with the bias properties remaining the same as those from the predictor derived from a single training set. This large supply of training sets can be emulated by bootstrap sampling: by averaging over samples, one gets "closer" (in the mean squared error sense) to the true value, on average. Technical details of why this works are given in Appendix S1.

Let the predictor formed from a single training set of size N_{Train} be $\hat{\boldsymbol{\phi}}\left(\mathbf{G}_i'\right)=\mathbf{G}_i'\ \hat{\mathbf{g}}^{*}(i=1,2,...,N_{\text{Train}})$. Its variance can be lowered by taking B bootstrap copies of size N_{Train} (i.e., sampling with replacement from the training set) and then averaging over copies. A given $\left(y_i,\mathbf{G}_i'\right)$ may not appear at all or may be repeated several times over the B bootstrap samples. The bagging algorithm is:

- For each copy b ($b=1,2,...,B$) run GBLUP using estimates of variance components obtained from the entire data set (to simplify computations), find the regressions $\hat{\mathbf{g}}^{*}_{b}$ and form a bootstrap draw for $BLUP(\mathbf{g})$ as

$$\hat{\mathbf{g}}_b=\mathbf{G}_b\hat{\mathbf{g}}^{*}_{b};\ b=1,2,...,B. \tag{13}$$

- After running the B GBLUP implementations take the following averages

$$\hat{\mathbf{g}}^{*}_{\text{Bagged}}=\frac{\sum_{b=1}^{B}\hat{\mathbf{g}}^{*}_{b}}{B},$$

and

$$\hat{\boldsymbol{\phi}}_{\text{bagged}}=\frac{\sum_{b=1}^{B}\hat{\mathbf{g}}_{b}}{B}. \tag{14}$$

- Predict vector $\mathbf{y}_{\text{Test}}$ in the testing set as $\hat{\boldsymbol{\phi}}_{\text{Test}}=\mathbf{G}_{\text{Test,Train}}\hat{\mathbf{g}}^{*}_{\text{Bagged}}$ where $\mathbf{G}_{\text{Test,Train}}$ is a matrix of genomic relationships between individuals in testing and training sets. If $\mathbf{y}_{\text{Test}}$ pertains to records on the same individuals the predictor is $\hat{\boldsymbol{\phi}}_{\text{Test}}=\hat{\boldsymbol{\phi}}_{\text{bagged}}$.

To see how improvement arises consider the following argument. We seek to learn signal $\mathbf{g}$ from the GBLUP predictor $\hat{\mathbf{g}}$. Under the setting of BLUP (where $\mathbf{g}$ and $\mathbf{y}$ both vary at random over conceptual repeated sampling), the best predictor of $\mathbf{g}$ in the mean squared error sense is $\hat{\mathbf{g}}$ because this is the conditional expectation under normality [39–40]. However, if $\mathbf{g}$ is the signal of a fixed target set of candidates, $\hat{\mathbf{g}}$ is biased as shown earlier. Thus, $E(\hat{\mathbf{g}}|\mathbf{g})=\mathbf{g}+\boldsymbol{\delta}$, where $\boldsymbol{\delta}$ is a vector of biases, and the mean squared error matrix of $\hat{\boldsymbol{\phi}}_{\text{bagged}}$ is

$$\begin{aligned} MSE\left(\hat{\boldsymbol{\phi}}_{\text{bagged}}|\mathbf{g}\right)&=\left[E\left(\hat{\boldsymbol{\phi}}_{\text{bagged}}-\mathbf{g}\right)E\left(\hat{\boldsymbol{\phi}}_{\text{bagged}}-\mathbf{g}\right)'\right]\\ &\quad+Var\left(\hat{\boldsymbol{\phi}}_{\text{bagged}}\right). \end{aligned}$$

Now, if the B bootstrap copies of GBLUP are drawn from the same distribution, $\hat{\boldsymbol{\phi}}_{\text{bagged}}$ has the same expectation as $\hat{\mathbf{g}}$ and, therefore, the same bias: $E\left(\hat{\boldsymbol{\phi}}_{\text{bagged}}-\mathbf{g}|\mathbf{g}\right)=\boldsymbol{\delta}$. Further, if the B copies are viewed as drawn independently, the variance of $\hat{\boldsymbol{\phi}}_{\text{bagged}}$ should be B times smaller than that of any $\hat{\mathbf{g}}_b$. Thus

$$MSE\left(\hat{\boldsymbol{\phi}}_{\text{bagged}}\right)=\boldsymbol{\delta}\boldsymbol{\delta}'+\frac{Var(\hat{\mathbf{g}})}{B};$$

a formula for taking the correlation between samples into account is not available but the reduction in MSE would be obviously smaller than in the idealized situation where samples are independent. Hence, $\hat{\boldsymbol{\phi}}_{\text{bagged}}$ has the same bias of GBLUP but at best is B times less variable. If a prediction machine has little

variance, as it is the case for shrinkage methods with a large amount of regularization, predictive performance might be degraded [31], but whether this occurs in a particular problem or not can be assessed empirically only.

Bagging marker effects

Alternatively, one can work directly with the linear model (1), but this is more involved computationally because the problem becomes a p-dimensional one (in GBLUP there are $n<p$ unknowns). Assuming centered data, the ridge regression estimator (BLUP of marker effects) is

$$\hat{\boldsymbol{\beta}}=\left(\mathbf{X}'\mathbf{X}+\mathbf{I}\lambda\right)^{-1}\mathbf{X}'\mathbf{y},$$

where some $\lambda=\frac{\sigma_e^2}{\sigma_\beta^2}$ (in the BLUP framework) is available, estimated from data at hand or chosen over a cross-validation grid. Given the data matrix $\mathbf{D}=[\mathbf{y},\mathbf{X}]$, of order $n\times(1+p)$, draw B bootstrap copies by randomly sampling n rows from $\mathbf{D}$ with replacement, such that a particular bootstrap sample is $\mathbf{D}_b=[\mathbf{y}_b,\mathbf{X}_b]$. The bagged ridge regression estimator is

$$\hat{\boldsymbol{\beta}}_{\text{Bagged}}=\frac{1}{B}\sum_{b=1}^{B}\left(\mathbf{X}_b'\mathbf{X}_b+\mathbf{I}\lambda\right)^{-1}\mathbf{X}_b'\mathbf{y}_b.$$

Then, given an out of sample case with marker matrix $\mathbf{X}_{\text{Test}}$, the yet-to be realized phenotypes are predicted as $\hat{\mathbf{y}}_{\text{Bagged}}=\mathbf{X}_{\text{Test}}\hat{\boldsymbol{\beta}}_{\text{Bagged}}$.

Using the relationship $\mathbf{g}=\mathbf{X}\boldsymbol{\beta}$, one can obtain "indirect" samples of the bootstrap distribution of $\hat{\mathbf{g}}$ as $\mathbf{X}_b\left(\mathbf{X}_b'\mathbf{X}_b+\mathbf{I}\lambda\right)^{-1}\mathbf{X}_b'\mathbf{y}_b=\mathbf{H}_b\mathbf{y}_b$ (where $\mathbf{H}_b=\mathbf{X}_b\left(\mathbf{X}_b'\mathbf{X}_b+\mathbf{I}\lambda\right)^{-1}\mathbf{X}_b'$ is the "hat" matrix for sample b) and form the "indirect" bagged GBLUP as

$$\hat{\mathbf{g}}_{\text{Bagged,indirect}}=\frac{1}{B}\sum_{b=1}^{B}\mathbf{H}_b\mathbf{y}_b. \quad (15)$$

Results

Simulated case studies

Case Study 1: model training with 200 known QTL and $N_{\text{Train}}=500$. To evaluate the impact of bagging on model training, we simulated 200 QTL with known binary genotypes (as in inbred lines, where genotypes at a given locus are either *aa* or *AA*). Genotypes were sampled with a frequency of 0.05 at all loci and their effects were drawn independently from a $N(0,\frac{1}{2})$ distribution; sample size was $N_{\text{Train}}=500$ so all QTL effects were likelihood identified (estimable) in the regression model described below. Any resulting disequilibrium was due to finite sample size only. Phenotypes were formed by summing the product of QTL genotype codes (0,1) times their corresponding effects over the 200 QTL and then adding a random residual drawn from $N(0,10)$; the "effective" heritability (variance among simulated realized genetic values as a fraction of the total variance) attained in the simulation was 0.34. The "true" model was employed in the training process using the "true" variance ratio $\lambda=20$ and the effect of the number of bootstrap samples (B), each of size $N_{\text{Train}}=500$, on the bagged predictor was examined by taking $B=50$, 200 and 500. While $B=25$ or 50 is often adequate [31], a larger number of bootstrap samples is not harmful.

Regressions of the 200 elements of $\mathbf{q}$ on either their ordinary least-squares estimates (OLS), BLUP-ridge regression ($\lambda=20$) and on a bagged mean obtained by averaging over bootstrap sample estimates of the BLUPs of $\mathbf{q}$ were calculated. This was also done for the regression of the 500 simulated genetic signals ($\mathbf{g}=\mathbf{Qq}$) on either GBLUP and on the average of the GBLUP bootstrap samples $\left(\hat{\mathbf{g}}_{\text{Bagged}}\right)$. Table 1 presents results. As expected from BLUP theory [10] the regressions of either $\mathbf{q}$ or $\mathbf{g}$ on their corresponding BLUPs were near their expected value: 1. By construction, $Var[BLUP(\mathbf{q})]=Cov(\mathbf{q},BLUP(\mathbf{q}))$, so the expected regression is necessarily 1. For QTL effects, OLS, even though being an unbiased estimator of $\mathbf{q}$, produced a regression of about 0.42. The bagging procedure produced regressions that exceeded 1, both at the level of the QTL effects and of the genetic signal $\mathbf{g}$. From the point of view of goodness of fit to the data, ridge regression and bagging produced models that accounted for more variation of the training data than OLS: for OLS R^2 was about 0.49 whereas it ranged between 0.51 and 0.53 for bagging and ridge regression. Increasing the number of bootstrap copies in the bag increased R^2 mildly with no sizable gain resulting from increasing B from 200 to 500. The regressions of $\mathbf{g}$ on $\hat{\mathbf{g}}_{\text{Bagged}}$ were larger than 1 but the bagged means accounted for the same proportion of variation in true $\mathbf{g}$ values as $\hat{\mathbf{g}}_{\text{GBLUP}}$ did, with $R^2=0.63$ for the latter and for bagged GBLUP with $B=200$ or 500.

Figure 1 (left panel) gives the distribution of estimates of the 200 QTL effects within each of 4 bootstrap samples for the run with $B=500$ as well as within the vector of bagged means. As expected bagging produced less variability among individual effect estimates, as "extreme" values are tempered by the averaging procedure. The impact of bagging is seen in the right panel of Figure 1: the variance among bagged means of individual effects was much less than the variance within any of the 500 bootstrap copies. Figure 2 (left panel) shows the variance reduction when bagging was applied to the GBLUPs of individuals: the horizontal line at the bottom is the variance among the bagged GBLUPs of the 500 individuals in the training sample. The right panel of Figure 2 shows that bagged GBLUP (BGBLUP) produces an understatement of genetic values relative to what is predicted by GBLUP for individuals ranked as "high" by the latter, but an overstatement otherwise; however, the two procedures give aligned predictions of genetic values. Bagging regresses extreme GBLUP estimates towards their average, which might attenuate influence from idiosyncratic samples on which GBLUP is trained. Hence, bagging enhances the shrinkage towards 0 inherent to GBLUP.

Case Study 2: model training with 1000 known QTL and $N_{\text{Train}}=500$. As in case 1, sample size was $N_{\text{Train}}=500$ but 1000 QTL with known binary genotypes and unknown effects were simulated. Here $n<rank(Q)$ so least-squares cannot be used due to lack of estimability. The setting was as in case 1, but "effective heritability" was 0.68, that is, twice as when 200 QTL were simulated. The "true" model was employed in the training process using the "true" variance ratio $\lambda=20$, and the number of bootstrap samples for bagging was $B=25$,50 or 100.

Results are presented in Table 2. The regressions of $\mathbf{q}$ on $\hat{\mathbf{q}}_{\text{Bagged}}$ were much larger than 1 and the slope seemingly stabilized at about 1.32 with a bag consisting of 50 bootstrap copies. As expected, the regression of $\mathbf{q}$ on $\hat{\mathbf{q}}_{\text{Ridge}}$ was near 1. However, the proportion of variation in true $\mathbf{q}$ accounted for by the variation in

Table 1. Regression coefficients of true QTL effects ($\mathbf{q}$) on their ordinary least-squares ($\hat{\mathbf{q}}_{\mathrm{OLS}}$), ridge regression BLUP ($\hat{\mathbf{q}}_{\mathrm{Ridge}}$) and bagged ridge regression BLUP estimates ($\hat{\mathbf{q}}_{\mathrm{Bagged}}$), and regressions of true genetic signal ($\mathbf{g}$) on genomic BLUP ($\hat{\mathbf{g}}_{\mathrm{GBLUP}}$) and bagged genomic BLUP ($\hat{\mathbf{g}}_{\mathrm{Bagged}}$) at varying number of bootstrap samples (B).

	No. Bootstrap samples→	$B=20$	$B=200$	$B=500$
Regressions of q on ↓				
$\hat{\mathbf{q}}_{\mathrm{OLS}}$	$0.42(R^2=0.49)$	–	–	–
$\hat{\mathbf{q}}_{\mathrm{Ridge}}$	$0.98\ (R^2=0.53)$	–	–	–
$\hat{\mathbf{q}}_{\mathrm{Bagged}}$	–	$1.05\ (R^2=0.51)$	$1.07(R^2=0.53)$	$1.08\ (R^2=0.53)$
Regressions of g on ↓				
$\hat{\mathbf{g}}_{\mathrm{GBLUP}}$	$0.99(R^2=0.63)$	–	–	–
$\hat{\mathbf{g}}_{\mathrm{Bagged}}$	-	$1.05(R^2=0.61)$	$1.08(R^2=0.63)$	$1.08(R^2=0.63)$

R^2 is the coefficient of determination of the regression fitted. The simulation involved a training set of 500 individuals with 200 true additive QTL fitted in the model.
doi:10.1371/journal.pone.0091693.t001

bagged or ridge regression estimates was smaller than in case 1, with $R^2 \sim 0.29-0.30$ for bagging and ridge regression versus $R^2 \sim 0.51-0.53$ in case 1. On the other hand, bagged GBLUP accounted for about 73% of the variation of the true genetic values $\mathbf{g}$, providing a "fit to signal" similar to that of GBLUP. The regression of $\mathbf{g}$ on $\hat{\mathbf{g}}_{\mathrm{GBLUP}}$ was also near its expected value of 1, whereas true differences in genetic values among individuals were exaggerated by a factor of about 1.25–1.27 by bagging. The left panel in Figure 3 shows the alignment between 4 randomly chosen bootstrap samples and the 500 GBLUP estimates obtained in N_{Train}. The right panel illustrates clearly that bagging ($B=100$) "pulls down" individuals with large GBLUP values and "pulls up" those placed at the left tail of the distribution of GBLUPs.

Case Study 3: model training with 20 unknown QTLs, 200 markers and $N_{\mathrm{Train}}=500$. The setting was as in case 1 ($N_{\mathrm{Train}}=500$) but here the genetic signal was generated from $q=20$ QTL with unknown binary genotypes that were in linkage disequilibrium (LD) with $p=200$ binary markers as specified below. The true model was

$$\mathbf{y}=\mathbf{Qq}+\mathbf{e},$$

where $\mathbf{q}$ is a 20×1 vector of allelic substitution effects. QTL genotypes were equally frequent at all loci and their effects were

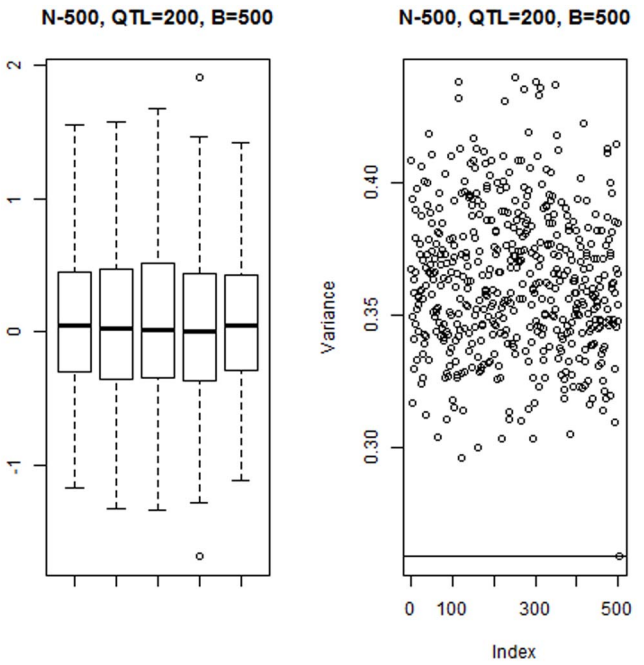

Figure 1. Simulation with 200 known QTL, 500 individuals in the training sample and 500 bootstrap copies for bagging. Left panel: distribution of 200 effects within each of 4 bootstrap samples (1–4), and within the average (bag) of 500 samples (5). Right panel: distribution of variance among 200 effects within each of 500 bootstrap samples and within their average (item 501, flagged with arrow).
doi:10.1371/journal.pone.0091693.g001

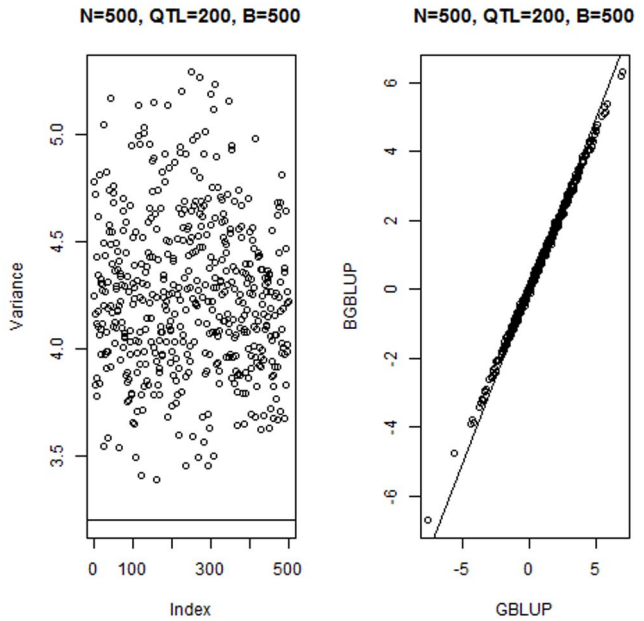

Figure 2. Simulation with 200 known QTL, 500 individuals in the training sample and 500 bootstrap copies for bagging. Left panel: variance (VAR) among 500 GBLUPs in each of 500 bootstrap samples. The horizontal line gives the variance among bootstrap (bagged) means for the 500 individuals. Right panel: scatter plot of bagged GBLUP (BGBLUP) versus exact GBLUP for 500 individuals.
doi:10.1371/journal.pone.0091693.g002

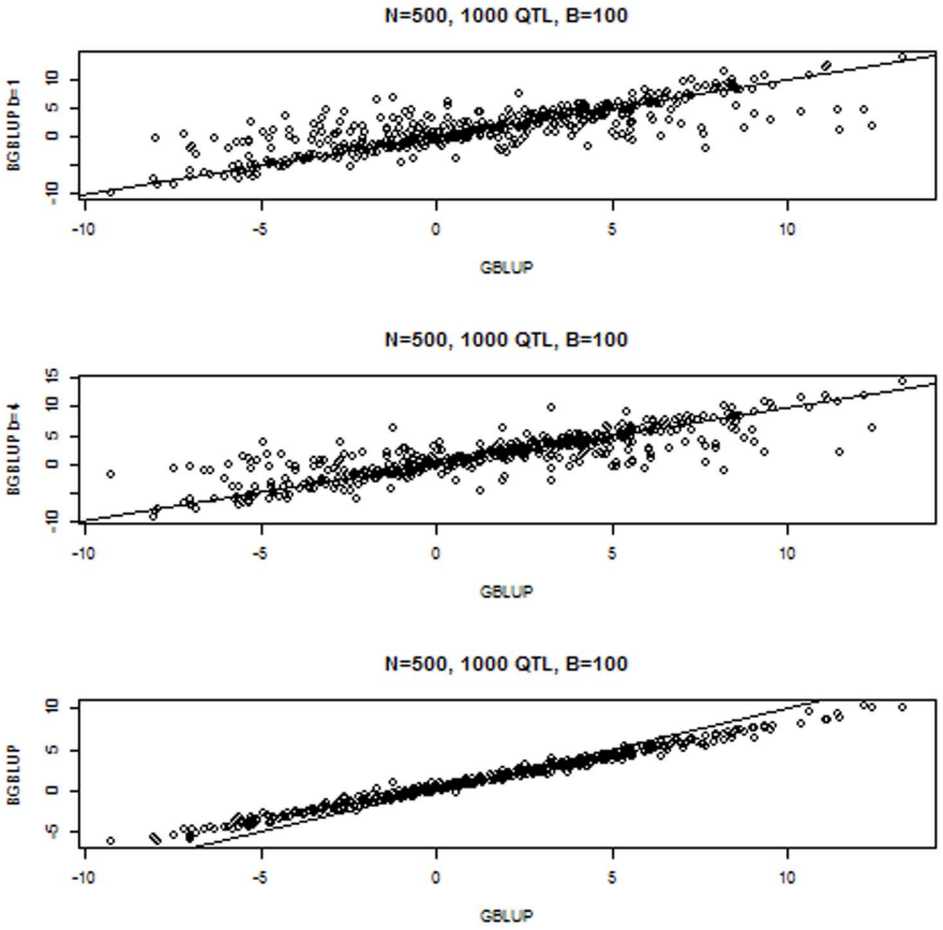

Figure 3. Simulation with 1000 known QTL, 500 individuals in the training sample and 100 bootstrap copies for bagging. Left panel: relationship between GBLUP with the entire sample and GBLUPs in 4 bootstrap samples of size 500. Right panel: relationship between bagged GBLUP (BGBLUP) and GBLUP of the 500 individuals.
doi:10.1371/journal.pone.0091693.g003

drawn independently from a $N(0,\frac{1}{2})$ distribution. Phenotypes were formed by summing the product of QTL genotype codes (0,1) times their corresponding effects over the 20 QTL, and adding a random residual drawn from $N(0,10)$. The method employed for training was ridge regression (BLUP) on 200 markers using the "true" variance ratio $\lambda=20$, and the number of bootstrap samples for bagging was $B=100$. The simulation generated an effective heritability of about 0.19.

LD was simulated statistically by introducing correlations among columns of the $\mathbf{Q}$ (QTL genotypes) and $\mathbf{X}$ (marker genotypes) matrices, respectively. This was done by drawing N_{Train}

Table 2. Regression coefficients of true QTL effects ($\mathbf{q}$) on their ridge regression BLUP $\left(\hat{\mathbf{q}}_{\text{Ridge}}\right)$ and bagged ridge regression BLUP estimates $\left(\hat{\mathbf{q}}_{\text{Bagged}}\right)$, and regressions of true genetic signal ($\mathbf{g}$) on genomic BLUP ($\hat{\mathbf{g}}_{\text{GBLUP}}$) and bagged genomic BLUP $\left(\hat{\mathbf{g}}_{\text{Bagged}}\right)$ at varying number of bootstrap samples (B).

	No. Bootstrap samples→	$B=25$	$B=50$	$B=100$
Regressions of q on ↓				
$\hat{\mathbf{q}}_{\text{Ridge}}$	$0.97(R^2=0.30)$	–	–	–
$\hat{\mathbf{q}}_{\text{Bagged}}$	–	$1.30(R^2=0.29)$	$1.32(R^2=0.29)$	$1.32(R^2=0.29)$
Regressions of g on ↓				
$\hat{\mathbf{g}}_{\text{GBLUP}}$	$0.99(R^2=0.73)$	–	–	–
$\hat{\mathbf{g}}_{\text{Bagged}}$	-	$1.25\ (R^2=0.73)$	$1.27(R^2=0.73)$	$1.27(R^2=0.73)$

R^2 is coefficient of determination of the regression fitted.
The simulation involved a training set of 500 individuals with 1000 true additive QTL fitted in the model.
doi:10.1371/journal.pone.0091693.t002

independent *Beta*(*a*,*b*) random variables corresponding to the rows of these matrices and then sampling a Bernoulli random variable (i.e., 0 for *aa* and 1 to *AA*, say) with probability of success given by the draw from the *Beta* distribution. Thus, columns of matrices **Q** or **X** (with entries 0 or 1) had a beta-binomial distribution (e.g., Casella and George, 1992) and an expected correlation equal to $1/(a+b+1)$; employing $a=0.30$ and $b=0.70$ a correlation equal to $\frac{1}{2}$ would be expected. Using this approach the "first" 10 QTL were in LD among themselves as well as in LD with the "first" 100 markers; these markers were in mutual LD themselves with a correlation of $\frac{1}{2}$ also. On the other hand, QTL 11–20 were in mutual linkage equilibrium as well as with all other markers. To illustrate, in the sample simulated realized genotypes at QTL 1 and 2 had a correlation of 0.48, whereas QTL 1 and 15 had a correlation of -0.02. Likewise, the correlation between markers 1 and 2 was 0.50, that between markers 1 and 200 was -0.04 and QTL 2 was correlated with marker 105 at 0.03. QTL genotypes at each locus were multiple-regressed on the 200 marker genotypes: for QTL 1–10 (in LD with markers) the R^2 of the regression of QTL genotypes on the 200 marker genotypes was about 0.70 or larger. For the 10 QTL in LE with markers R^2 fluctuated around 0.40. This last result illustrates that even a null association accounts for some variation, merely because the likelihood increases monotonically with model complexity (in this case there are 200 partial regressions of each QTL genotype on markers).

The regression of phenotypes on QTL genotypes or on markers had an R^2 at 0.24 and 0.43, respectively, with the latter being larger simply because more parameters are fitted in the model. On the other hand, the squared correlation between true signal and fitted values was 0.87 for the QTL model versus 0.15 for the marker-based model: even though markers captured more variation (because of higher model complexity) than a regression on true genotypes, their ability of capturing signal was much less.

We measured similarity among individuals by constructing "genomic correlation matrices". This was done by centering both **Q** and **X**, calculating $\mathbf{Q}_{\text{Centered}}\mathbf{Q}'_{\text{Centered}}$ and $\mathbf{X}_{\text{Centered}}\mathbf{X}'_{\text{Centered}}$, and converting these into correlation matrices $\mathbf{R}_Q$ and $\mathbf{R}_M$, respectively. A plot of the off-diagonal elements of $\mathbf{R}_M$ versus those of $\mathbf{R}_Q$ is shown in Figure 4 for the corresponding pairs of individuals. Although there is an association between genomic correlations at the QTL and marker levels, the latter ones were smaller in absolute values. This association was not perfect: at a given level of correlation at the QTL level, there was much variation in relationships when measured by markers. Implications of this on accuracy of genome-enabled prediction are discussed in [7].

GBLUP and BGBLUP were calculated at each of $\frac{\lambda}{2}, \lambda$ and $\frac{3}{2}\lambda$ (with $\lambda=20$) as variance ratio, to investigate the impact of regularization on the ability of capturing "true" signal, **Qq**. The regression of signal on GBLUP was 0.48, 0.54 and 0.60 for the three values of the regularization parameter, respectively, whereas that for BGBLUP was 0.50, 0.59 and 0.65, respectively. This illustrates that the regression of "true values" on GBLUP predictions is not 1 when the model is incorrect, as it is the case when markers are not QTL, as simulated here. The corresponding R^2 values were 0.25, 0.27 and 0.28 for GBLUP, and 0.24, 0.28 and 0.30 for BGBLUP. While $\lambda=20$ is the "correct" regularization to be exerted at the QTL level, this is not so for the model based on markers, where a stronger degree of regularization is needed (the number of markers was larger than the number of QTL). An approximation is that the "correct" variance ratio for the marker based model should be $p/n_q=10$ times larger than λ, where $n_q=20$ is the number of QTL. We found (results not shown) that the regressions of signal on GBLUP and BGBLUP increased as larger values of the regularization parameter were applied to the marker based model, with the "optimum" being near 10λ, as expected.

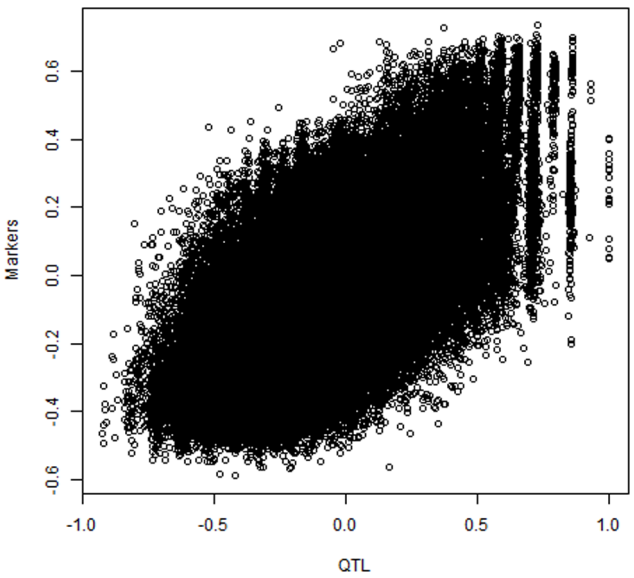

Figure 4. Simulation with 20 unknown QTL, 200 markers and 500 individuals: "genomic correlations" among 500 individuals at QTL or marker loci.
doi:10.1371/journal.pone.0091693.g004

Case Study 4: predictive cross-validation with 100 unknown QTLs and 500 markers. The simulation posed 100 unknown QTL whose additive effects were drawn from a $N(0,1.25)$ distribution and 500 individuals genotyped for 500 markers. The LD structure was similar to that of case 3: the first 50 QTL were in mutual linkage disequilibrium as well as in LD with the first 250 markers. QTL 51–100 were in LE among themselves as well with all markers; the first 250 markers were in mutual LD and markers 251–500 were in LE. Residuals were drawn from $N(0,10)$ and the "effective" heritability attained was about 0.74.

The 500 individuals were distributed at random into two non-overlapping training and testing sets with 250 members in each. This was done 100 times at random, producing 100 training-testing pairs enabling estimation of the cross-validation distribution. GBLUP and BGBLUP were fitted to the training data ($N_{\text{Train}}=250$) for each of 13 values of the regularization parameter λ in the sequence

$$\lambda_{\text{Seq}}=\left(\frac{1}{16},\frac{1}{8},\frac{1}{4},\frac{1}{2},1,2,4,8,16,32,64,128,256\right)\times\lambda_{\text{''True''}}, \qquad (16)$$

where $\lambda_{\text{''True''}}=\frac{\sigma_e^2}{\sigma_\beta^2}$ with $\sigma_\beta^2\approx\sigma_q^2\,\frac{n_q}{p}$. In each training instance, BGBLUP was implemented with $B=25$ bootstrap samples of size N_{Train} drawn from the training set of size 250. Three metrics were used to evaluate the two prediction methods: goodness of fit in the training set (correlation between fitted and observed values), predictive correlation (predicted phenotypes in the testing set and realized values) and predictive mean-squared error, that is, average squared difference between predicted and realized values over the $N_{\text{Test}}=250$ cases in the testing set.

The preceding involved 13 BGBLUP and GBLUP implementations in each of the 100 random cross-validations, for a total of 1300 comparisons. Overall (Figure 5), BGBLUP attained a better predictive performance than GBLUP because predictive correlations were typically larger (in some cases more than twice as large as GBLUP) and mean squared errors of prediction were lower as well. Thus, BGBLUP was more reliable (larger correlation) and more accurate (smaller mean squared error) than GBLUP. The superiority of BGBLUP over GBLUP became smaller when regularization was stronger over the grid defined by λ_{Seq}. However, it was not until $\log_{10}\lambda_{\text{Seq}}$ became 5 or more times larger than ($\log_{10}\lambda_{\text{"True"}}$) that GBLUP "caught up" with bagging but was never better. Actually, it was not until the λ value used for model training was 32 times larger than λ_{True} that the two predictors delivered the same performance, suggesting a "robustness" property of BGBLUP that GBLUP seems to lack. Model complexity is reduced as λ increases (the effective number of parameters decreases) so the variance of GBLUP decreases as well, in which case bagging offers little help. Contrary to what was stated by [7] we did not find evidence that bagging damaged predictive performance under any of the regularization regimes entertained. Results for selected settings are discussed in the following paragraph.

Figure 6 shows training and predictive correlations and mean squared errors for BGBLUP (y-axis) and GBLUP (x-axis) at 2 levels of "under-regularization": $\lambda=\lambda_{\text{True}}/16$ and $\lambda=\lambda_{\text{True}}/2$. Here, where shrinkage is less than it should be, given the model, BGBLUP reduced overfitting (smaller training set correlations), increased predictive correlations and reduced mean squared errors, relative to BLUP. Differences were marked: in the upper (lower) middle panels of Figure 6 it is seen that the largest predictive correlation attained by GBLUP was smaller than 0.30 (0.39) and that bagging increased the corresponding correlation to about 0.38 (0.41). The effect of bagging on reducing predictive mean squared was also clear.

Figure 7 presents results when the "true" value of λ (400) was used for training. GBLUP was again close to overfitting and BGBLUP reduced training correlations, thus tempering the problem. Predictively, BGBLUP was better than GBLUP in all 100 comparisons, both in the correlation and MSE senses. Over the 100 cross-validation runs, the predictive correlation ranged between 0.195 and 0.391 (median 0.282) for GBLUP and between 0.243 and 0.423 (median 0.330) for BGBLUP. MSE of prediction ranged between 33.64 and 46.66 (median 38.47) for GBLUP and between 30.24 and 43.41 (median 35.01) for BGBLUP. Comparing the medians of the distributions, BGBLUP enhanced the predictive correlation by 17% and MSE was 91% of that of GBLUP. Figure 8 depicts what was found when "excessive" shrinkage was applied in training: the regularization parameter values were $8\lambda_{\text{True}}$ (upper panel) and $16\lambda_{\text{True}}$ (lower panel). BGBLUP was only marginally better than GBLUP. Strong shrinkage rendered the training model exceedingly simple so both methods delivered similar same predictive ability. Differences between methods vanished when $256\lambda_{\text{True}}$ was used as shrinkage parameter.

We repeated the experiment with the setting of case study 4 but increasing training and testing sample sizes to 2500 each. Here, training sample size was 5 times larger than the number of markers: no differences between BGBLUP and GBLUP were found at any level of regularization. The reason is that n now exceeds p, making GBLUP fairly stable, in which case the variance reduction property of BGBLUP does not help much.

In summary, in our simulations BGBLUP was typically better than GBLUP at most points of the regularization grid considered. Its performance was better than that of GBLUP at values close to "optimal" regularization, and differences were large when shrinkage was small, because bootstrap sampling with averaging reduced variance. If λ is small, GBLUP tends to overfit and to be variable but BGBLUP alleviates these problems. In addition to helping with overfitting, BGBLUP was robust with respect to departures from optimal regularization, e.g., to errors in the variance ratio. The experiment with a much larger training sample size than the number of markers indicated that the performance of bagging depends on p/N_{Train}: we conjecture that as this ratio increases (as it will surely be the case with DNA sequence data) use of BGBLUP may enhance predictive ability in some real data situations, simply because overfitting and colinearity will be exacerbated by introducing a massive number of variates in the model.

Analysis of wheat data

The data set is at http://cran.r-project.org/web/packages/BLR/index.html, in the BLR package in *R*, and has been used by, e.g., [17],[42–43]. The data represents 599 wheat inbred lines each genotyped with 1279 DArT (Diversity Array Technology) markers, and planted in 4 distinct environments. DArT markers may take on one of two values, denoting presence or absence of an allele. Records came from several international trials conducted at the International Maize and Wheat Improvement Center (CIMMYT), Mexico. The trait considered was average grain yield for each line in each of the four environments. This response variable is an average over a balanced set of plots and replicates and, within environment, the residual variance is expected to be constant over lines.

To provide "proof of concept", we used a ridge-regression BLUP model with 600 randomly chosen markers. All markers were not employed in order to facilitate calculations, since matrix inversion was used for every bootstrap sample and cross-validation. With a large number of markers or individuals or for routine application, GBLUP can be computed in using iterative algorithms, but this is a numerical, as opposed to conceptual, issue. Further, $N_{\text{Train}}=449$ and $N_{\text{Test}}=599-449=150$; the model was trained using a grid with 10 values of λ: 5, 10, 50, 100, 150, 200, 250, 300, 350 and 400. All lines were present in each partition of the data into the two disjoint training and testing sets, with the process repeated at random 100 times, to estimate the cross-validation distribution. Each implementation of BGBLUP used 25 bootstrap samples and performance was evaluated via predictive correlation, predictive mean squared error and mean absolute deviation between predicted and realized phenotypes. We found (results not shown) that the "optimum" λ was around 50–100 in environments 1 and 2 with stronger regularization needed in environments 3 and 4. BGBLUP was slightly better than GBLUP in environment 1 near optimum regularization, and clearly better at the lower values of λ because GBLUP nearly overfitted; a similar picture emerged in environment 2. Overall, BGBLUP was better than GBLUP when λ was below the optimum, sometimes slightly better at the optimum, with no difference if regularization was excessive, due to the fact that the two models were rendered effectively simpler as λ grew.

The upper and lower panels of Figure 9 give results for environments 3 and 4, respectively; results for environments 1 and 2 were similar. As found with the simulated data, over the 100 repetitions $\times$ 10 values of $\lambda=1000$ comparisons, BGBLUP tended to produce larger predictive correlations and smaller mean squared errors than GBLUP. However, this superiority was not uniform and depended on whether or not the model was "under" or "over" regularized, with BGBLUP having slightly better

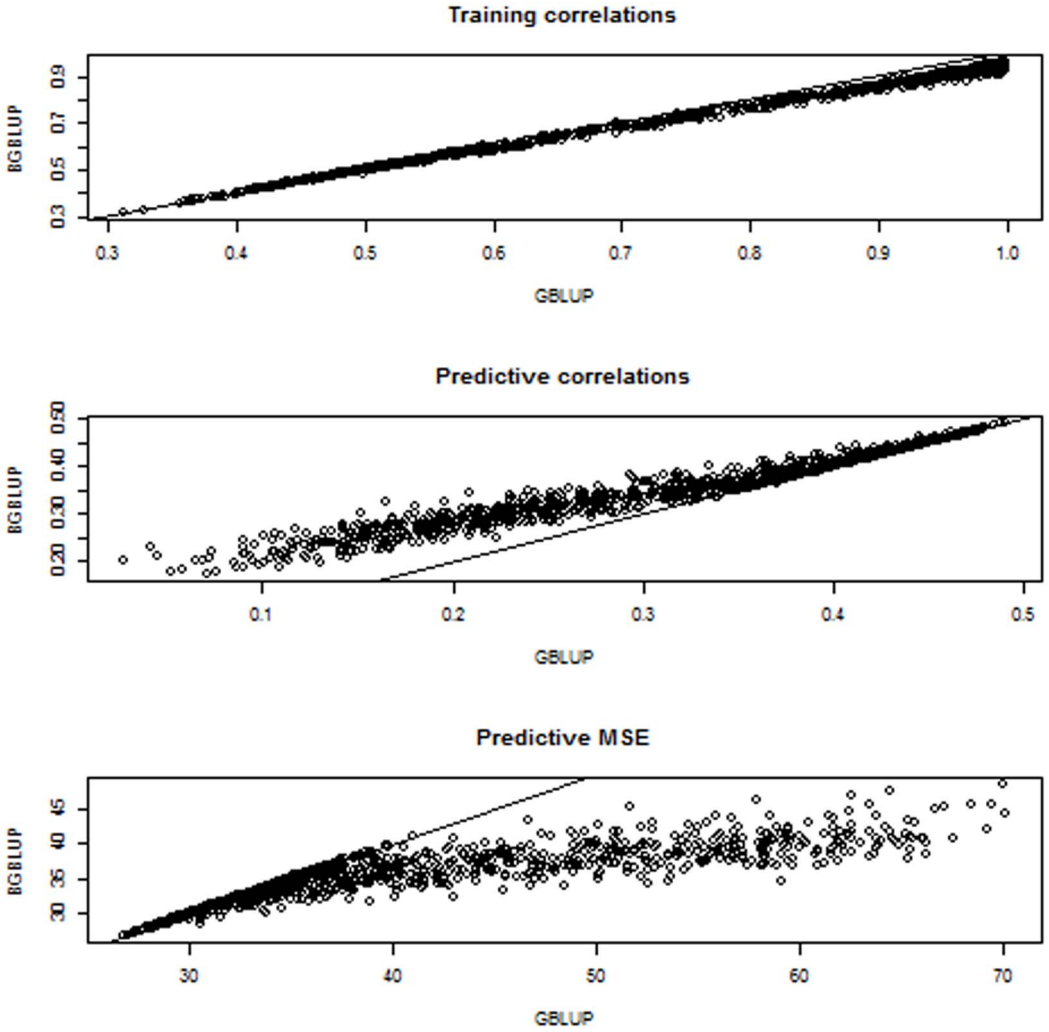

Figure 5. Simulation with 100 unknown QTL, 500 markers and 250 individuals in each training and testing set: training correlations, predictive correlations and predictive mean-squared error (MSE) for 1300 comparisons between bagged GBLUP (BGBLUP, 25 bootstrap samples) and GBLUP.
doi:10.1371/journal.pone.0091693.g005

performance in the former situation but slightly worse in the latter. This is illustrated for environment 4 in the upper left panel of Figure 10, where average differences between BGBLUP and GBLUP over the grid of lambda values are shown for predictive correlations, predictive mean square error and average absolute difference between predicted and realized values. The other panels of Figure 10 indicate that, over the 100 cross-validations, BGBLUP was better at low values of λ, slightly better at near optimum values of λ and mildly worse when regularization was extreme ($\lambda = 400$). Hence, it would seem that BGBLUP performs at least as well as GBLUP unless a gross error is made in assessing the value of the regularization parameter in model training. Such a large error is unlikely if the variances are estimated from training data (unless the sample is small) or evaluated over a grid of suitable candidate values. One should be cautious about elicitations of the regularization parameter based on simple theoretical arguments that may not hold.

Discussion

We examined whether or not bootstrap sampling in the context of GBLUP can enhance predictive ability in cross-validation. Simulation (with known or unknown QTL) and a wheat data set with grain yield information were used for this purpose. In the simulations, it was found that bagging BLUP estimates of marker effects or of genomic signal increased the slope of the regression of true marker or marked breeding value on predictor relative to what is expected under BLUP theory. When an individual was evaluated as "extreme" by GBLUP, bagging made the estimate less extreme. If the linear model entertained holds, the regression of true signal on GBLUP is expected to be 1, but the regression on BGBLUP is steeper because the latter has smaller variance. This is easy to see: if T is a predictand and $\hat{T}$ is its BLUP, then $Cov(T,\hat{T}) = Var(\hat{T})$ so the slope of the regression of T on $\hat{T}$ is one. Now, if $\hat{T}_{\text{Bagged}}$ is the average of B bootstrap copies of $\hat{T}$, the variance of $\hat{T}_{\text{Bagged}}$ is about B times (assuming samples are mildly correlated) smaller than that that of $\hat{T}$, but $Cov(T,\hat{T}) = Cov(T,\hat{T}_{\text{Bagged}})$. Hence, the regression must be larger than 1.

It was also found that bagging conferred robustness to GBLUP because it is less prone to over-fitting and often delivered better predictions in terms of correlation and mean squared error even when regularization was "optimal". At least in simulated data,

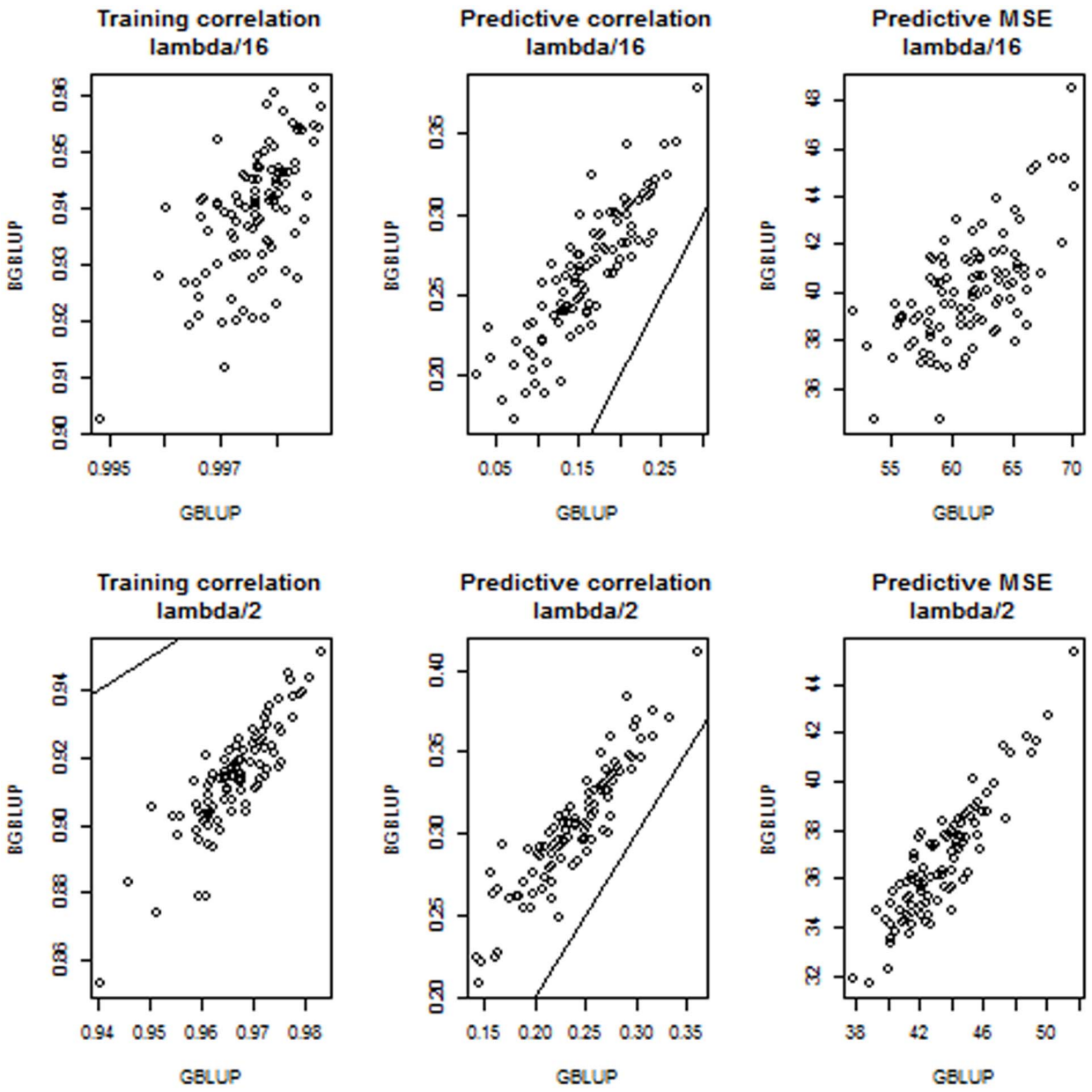

Figure 6. Simulation with 100 unknown QTL, 500 markers and 250 individuals in each training and testing set: training correlations, predictive correlations and predictive mean-squared error (MSE) for 100 comparisons between bagged GBLUP (BGBLUP, 25 bootstrap samples) and GBLUP at two levels of ''under-regularization''.
doi:10.1371/journal.pone.0091693.g006

BGBLUP was not inferior to GBLUP when shrinkage was beyond what it should be.

Bagging allows estimating a cross-validation prediction error mean squared error for each subject tested. In theory, given training data, GBLUP in a testing set can be computed as

$$\hat{\mathbf{g}}_{\text{Train}} = \mathbf{G}_{\text{Train}}\left(\mathbf{G}_{\text{Train}} + \mathbf{I}_{\text{Ntrain}}\frac{\sigma_e^2}{p\sigma_\beta^2}\right)^{-1}$$

$$\mathbf{y} = \left(\mathbf{I}_{\text{Ntrain}} + \mathbf{G}_{\text{Train}}^{-1}\lambda_{\text{Optimum}}\right)^{-1}\mathbf{y},$$

with λ_{Optimum} being some ''optimum'' value of the regularization parameter. If the problem is that of predicting a future set of records $\mathbf{y}_{\text{Test}}$ of the same individuals, the variance-covariance matrix of prediction errors (under normality assumptions) is

$$Var(\mathbf{y}_{\text{Test}}|\mathbf{y}_{\text{Train}}) = Var(\mathbf{g}|\mathbf{y}_{\text{Train}}) + \mathbf{I}\sigma_e^2.$$

Given that the assumptions hold, there are two sources of uncertainty here. The first is uncertainty about signal (breeding value) given training data, and the second is noise variability associated with the yet to be realized observations. The model derived (expected) reliabilities of the predicted genomic value values from the training data are given by the diagonals of matrix

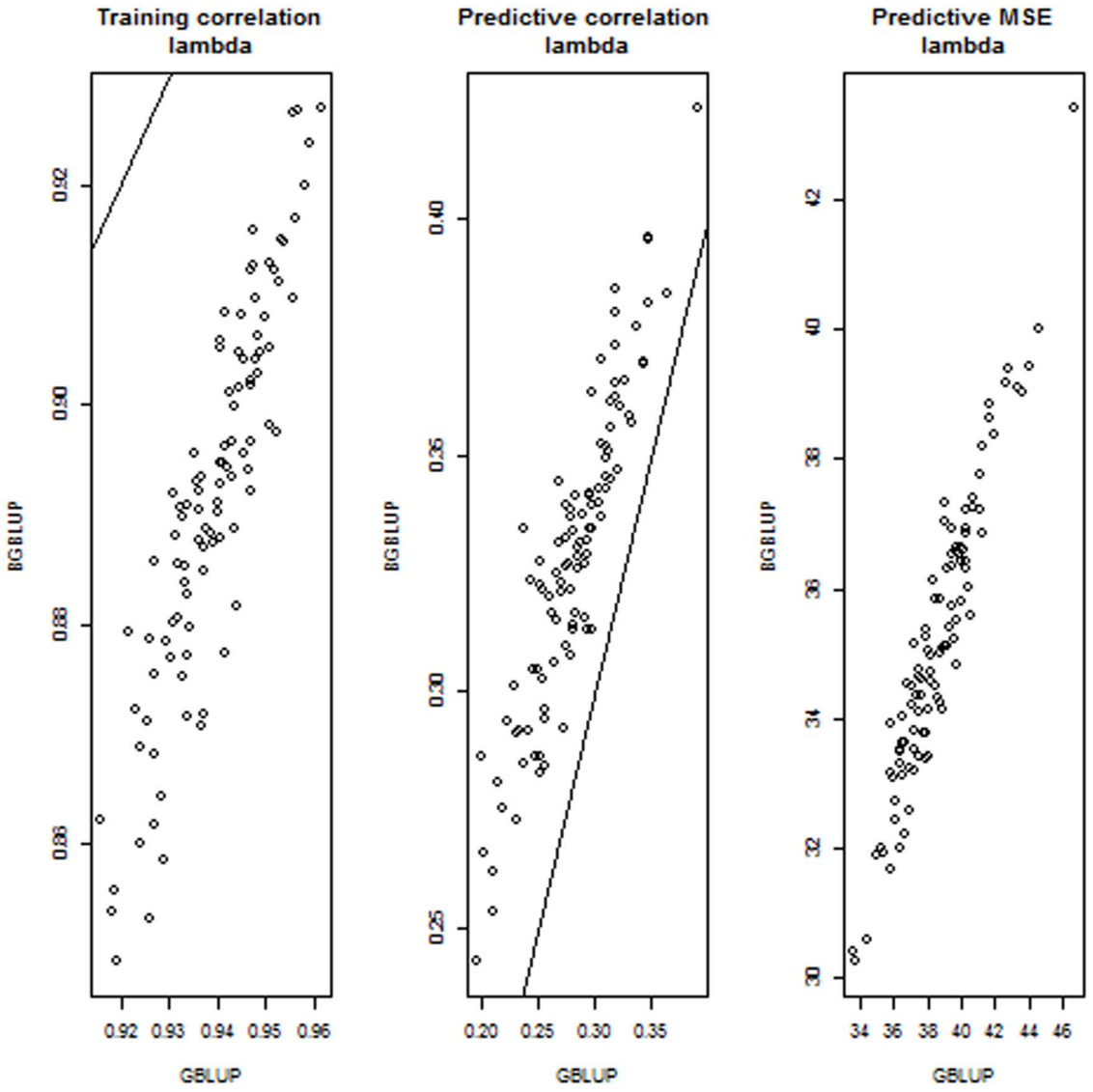

Figure 7. Simulation with 100 unknown QTL, 500 markers and 250 individuals in each training and testing set: training correlations, predictive correlations and predictive mean-squared error (MSE) for 100 comparisons between bagged GBLUP (BGBLUP, 25 bootstrap samples) and GBLUP at the ''correct'' level of regularization.
doi:10.1371/journal.pone.0091693.g007

$$\boldsymbol{\Omega}_{\mathrm{G}} = \mathbf{I} - \sigma_e^2\left(\mathbf{I}_{\mathrm{Ntrain}} + \mathbf{G}_{\mathrm{Train}}^{-1}\lambda_{\mathrm{Optimum}}\right)^{-1}$$
$$Var^{-1}\left(\mathbf{X}_{\mathrm{Train}}\mathbf{X}_{\mathrm{Train}}'\sigma_\beta^2\right), \quad (17)$$

where $\mathbf{X}_{\mathrm{Train}}$ is the marker matrix in the training set. Note that (17) does not make reference at all to realized outcomes (testing set phenotypes) or to realized predictions, so using the term ''accuracy'' in lieu of ''reliability'' is misleading. While the diagonal elements of $\boldsymbol{\Omega}_{\mathrm{G}}$ may be close to 1, this does not give assurance that predictions will be any good. The reliability matrix uses only information on marker genotypes (via the $\mathbf{X}$ matrix) and variance components, but do not exploit information on phenotypes. Importantly, the two measures ignore model uncertainty, thus exaggerating prediction reliability relative to what would be observed in a cross-validation distribution. Model goodness of fit statistics in training data lead to expectations that seldom translate into what is observed in cross-validation (e.g., [44]) and examples of this are in a study of human height with molecular markers by [7] and [45]. The problem of developing credible individual-specific measures of reliability in cross-validation has not been solved yet [30] but a practical solution can be arrived at by use of bagging.

Let the fixed, observed, outcome (e.g., the mean of an inbred line of wheat, a daughter yield deviation of an artificial insemination bull or the phenotype of a subject) in a testing set be $\tilde{y}_i$, $i=1,2,...,N_{\mathrm{Test}}$, and let the prediction from GBLUP be $\hat{\boldsymbol{\phi}}_i$, so the realized prediction error is $\tilde{y}_i - \hat{\boldsymbol{\phi}}_i$. In BLUP theory, predictor and predictand (the latter with eventual realized value $\tilde{y}_i$) vary at random over conceptual repeated sampling, given some linear model, but here $\tilde{y}_i$ is an observed realization from an unknown process. Using bagging, B bootstrap samples of the distribution of $\hat{\boldsymbol{\phi}}_i$ are available, so one can form the set of prediction errors $E_i = \left\{\tilde{y}_i - \hat{\boldsymbol{\phi}}_i^{(1)}, \tilde{y}_i - \hat{\boldsymbol{\phi}}_i^{(2)}, ..., \tilde{y}_i - \hat{\boldsymbol{\phi}}_i^{(B)}\right\}$ for

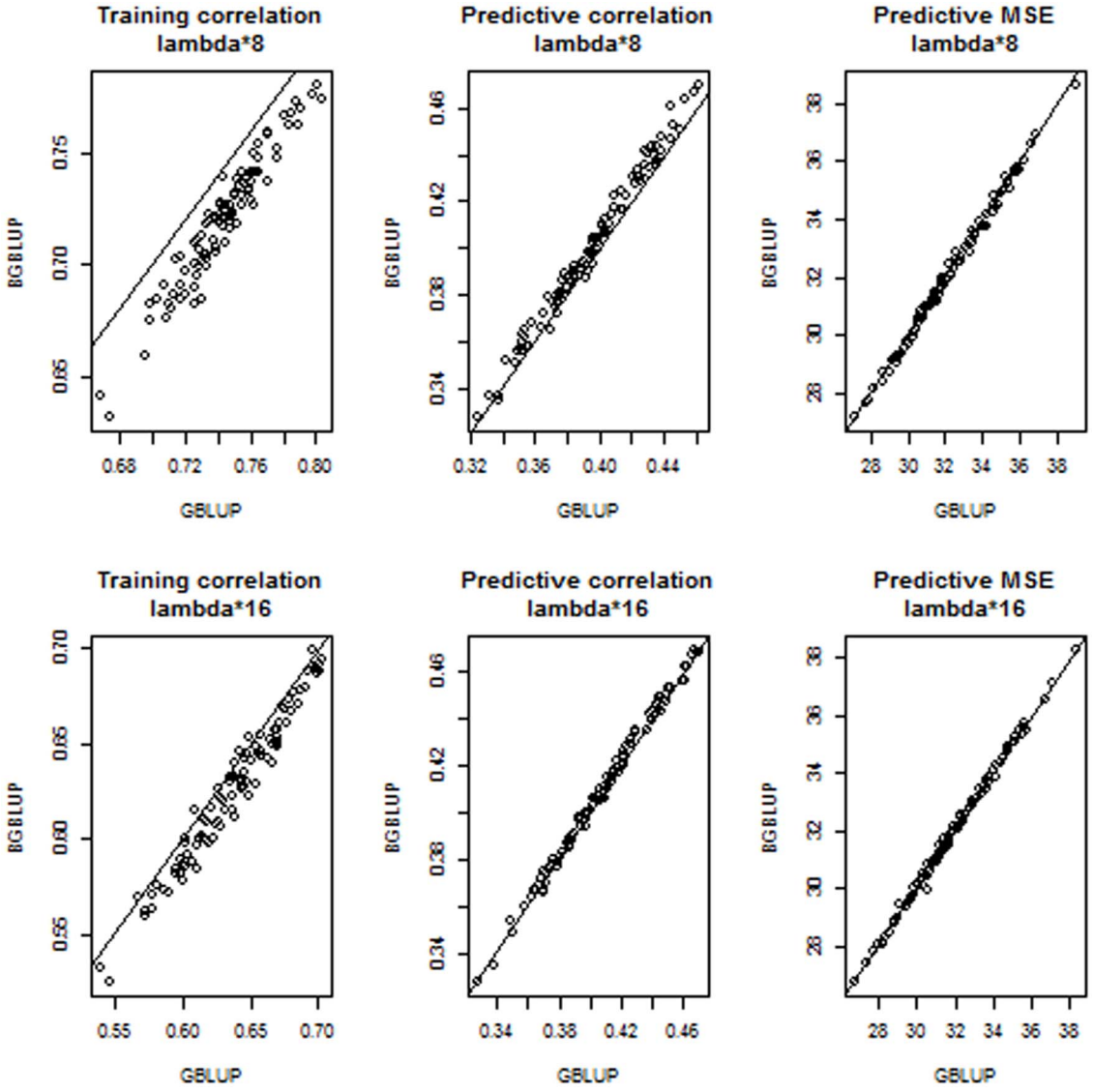

Figure 8. Simulation with 100 unknown QTL, 500 markers and 250 individuals in each training and testing set: training correlations, predictive correlations and predictive mean-squared error (MSE) for 100 comparisons between bagged GBLUP (BGBLUP, 25 bootstrap samples) and GBLUP at two levels of ''over-regularization''.
doi:10.1371/journal.pone.0091693.g008

$i=1,2,...,N_{\text{Test}}$. For each i, the bootstrap average squared prediction error associated with GBLUP (given $\tilde{y}_i$, $\mathbf{X}_{\text{Train}}$ and $\mathbf{X}_{\text{Test}}$) is assessed as

$$BPE_i=\frac{\sum_{b=1}^{B}\left[\tilde{y}_i-\hat{\phi}_i^{(b)}\right]^2}{B};i=1,...,N_{\text{Test}}, \tag{18}$$

noting that this squared cross-validation prediction error reflects both squared bias (unknown) and variance. Similarly, a cross-validation reliability measure can be constructed as

$$BPREL_i=1-\frac{BPE_i}{v_{\text{Test}}};i=1,...,N_{\text{Test}}, \tag{19}$$

where

$$v_{\text{Test}}=\frac{\sum_{i=1}^{N_{\text{Test}}}\left[\tilde{y}_i-\bar{\tilde{y}}_i\right]^2}{N_{\text{Test}}-1}.$$

BPREL takes values between 0 and 1 provided that $BPE_i\leq v_{\text{Test}}$, which cannot be assured unless one replaces v_{Test} by, say, $\max_{j\in\text{Testing set}}\left(BPE_j\right)$. A disadvantage of $BPREL_j$ is that it does not take into account the fact that, given $\mathbf{X}$, all observations are expected to have a different phenotypic variance, depending on how a genomic relationship matrix is constructed in GBLUP. Recall that GBLUP poses $\mathbf{g}|\sigma_g^2\sim N\left(0,\mathbf{G}\sigma_g^2\right)$, leading to the testing set expected variance-covariance structure

$$\mathbf{V}_{\text{Test}}=\mathbf{G}_{\text{Test}}\sigma_g^2+\mathbf{I}_{\text{Test}}\sigma_e^2.$$

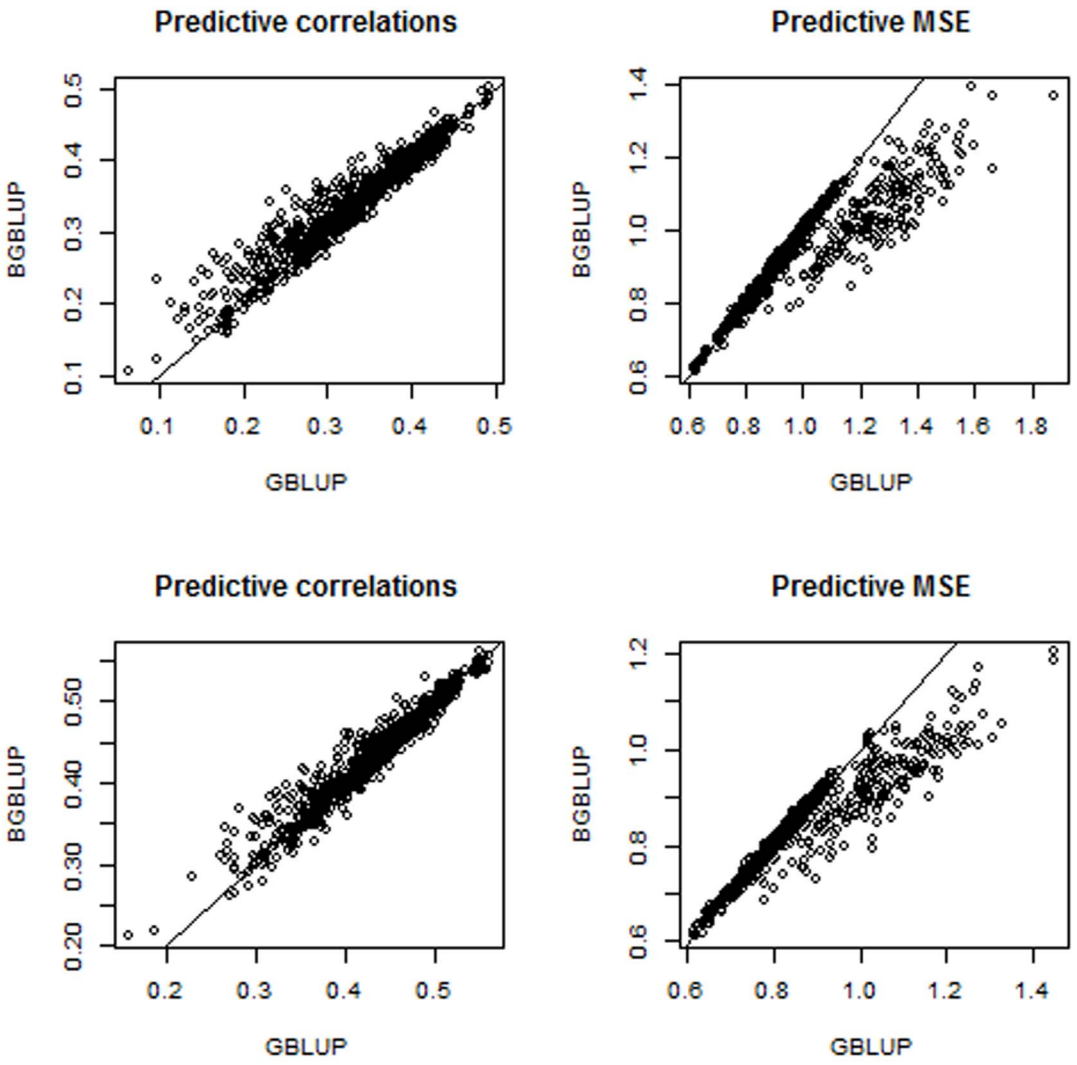

Figure 9. Wheat data in environments 3 and 4: predictive correlations and mean-squared errors in 1000 cross-validations (100 random partitions intro training-testing sets and 10 levels of the regularization parameter).
doi:10.1371/journal.pone.0091693.g009

In the absence of some scaling (with the latter having consequences on the definition of σ_g^2) the diagonal elements of $\mathbf{G}$ vary over individuals, so the diagonals of $\mathbf{V}_{\text{Test}}$ vary as well; this does not occur in a pedigree-based model if all individuals have the same level of inbreeding. One way of taking this into account is to modify the "reliability" measure (19) into

$$BPREL_i' = 1 - \frac{BPE_i}{diag_i(\mathbf{V}_{\text{Test}})}, \tag{20}$$

where $diag_i(\mathbf{V}_{\text{Test}})$ is the *ith* diagonal element of $\mathbf{V}_{\text{Test}}$.

We examined this proposal under the setting of case 4, with 100 unknown QTL, 500 binary markers, $N_{\text{Train}} = N_{\text{Test}} = 250$. In the simulation (results not shown), we found in cross-validation that the "optimum" λ in terms of predictive correlation and mean-squared error sense was 3200. We trained the model using $\lambda = 1600$, 3200 and 6400 and $B = 25$,50 and 100. Differences in BPE_i obtained with the three values of B were very small and the three levels of regularization produced the same qualitative picture, with prediction mean-squared error increasing with stronger shrinkage. Figure 11 illustrates the disconnect between prediction error variances derived in the training process and bootstrap average squared prediction errors, which make use of both training phenotypes, via $\hat{\boldsymbol{\phi}}_i^{(b)}$ and realized values $\tilde{y}_i$. Likewise, as shown in Figure 12 the empirical *BPREL* (top panel) and the adjusted reliabilities (bottom panel) are unrelated to model derived reliabilities. The adjusted reliabilities were calculated as

$$BPREL_i'' = 1 - \frac{BPE_i}{\max\left[Var(y_{\text{Test}}), \max_{i \in \text{Testing set}}(BPE_j)\right]} \tag{21}$$

with $Var(y_{\text{Test}})$ constant over the training set. When using *BPREL* median reliability (using $B = 25$) was estimated at 0.702, 0.725 and 0.722 at the three level of regularization but a few ones were negative. After the adjustment in (21) all reliabilities varied mostly between about 0.40 and slightly less than 1 and these values were unrelated to theoretical reliabilities. Predictions were quite accurate, in general (recall that the same stochastic process was used to create training and training sets).

There does not seem to be a theoretical reason leading to expect an agreement between model based reliabilities and measures of

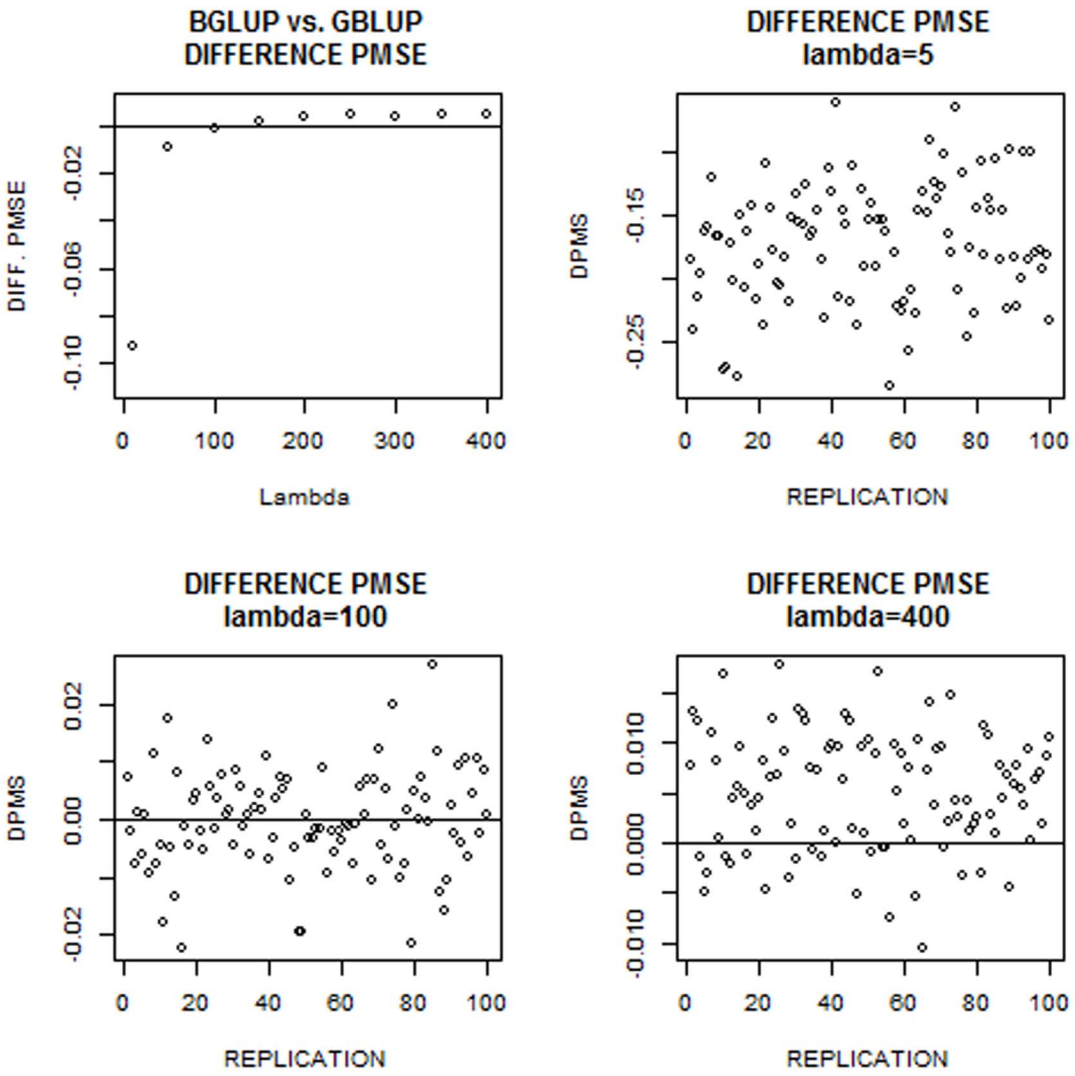

Figure 10. Wheat data in environment 4. Upper left panel: average differences (over 100 cross-validations) between bagged GBLUP and GBLUP for predictive correlations (PCOR), mean-squared error (PMSE) and absolute value of differences (PABS) between prediction and realization at 10 values of the regularization parameter. Upper right, lower left and lower right panels give the three metrics for lambda values of 5, 100 and 400, respectively.
doi:10.1371/journal.pone.0091693.g010

cross-validation performance, because the latter gauge different things. The theoretical reliabilities, based on a model deduced quantity, are just indicators of the amount of information in the training data set without making reference to the "goodness" of any prediction. On the other hand, *BPREL* or variants thereof take into account "closeness" between prediction and realized value, with bagging enhancing the stability of the prediction. Hence, we argue that bagging is sensible because it reduces the variance of GBLUP, seemingly without hampering predictive ability, and provides a means for ascertaining the (conditional) prediction bias in a strict sense. If BPE_i is close to 0 the squared prediction error is small, so that the prediction has a small variance, a small bias, or both. Irrespective of the cause, the cross-validation measure of reliability would be close to 1.

To discuss influences that theoretical reliability may have on predicted values in a testing set ($\hat{\mathbf{y}}_{\text{Test}}$), note that, when using ridge regression BLUP,

$$\hat{\mathbf{y}}_{\text{Test}} = \mathbf{X}_{\text{Test}}\left(\mathbf{X}'_{\text{Train}}\mathbf{X}_{\text{Train}} + \mathbf{I}_{\text{Ntrain}}\lambda\right)^{-1}\mathbf{X}'_{\text{Train}}\mathbf{y}_{\text{Train}}.$$

The influence training data have on predictions via GBLUP can be measured by the derivative or "hat matrix"

$$\frac{\partial\hat{\mathbf{y}}_{\text{Test}}}{\partial\mathbf{y}'_{\text{Train}}} = \mathbf{X}_{\text{Test}}\left(\mathbf{X}'_{\text{Train}}\mathbf{X}_{\text{Train}} + \mathbf{I}_{\text{Ntrain}}\lambda\right)^{-1}\mathbf{X}'_{\text{Train}}$$

Now, the matrix of "reliabilities of marker effects" is

$$\mathbf{R}_{\beta} = \mathbf{I}_p - \lambda\left(\mathbf{X}'_{\text{Train}}\mathbf{X}_{\text{Train}} + \mathbf{I}_{\text{Ntrain}}\lambda\right)^{-1},$$

so that

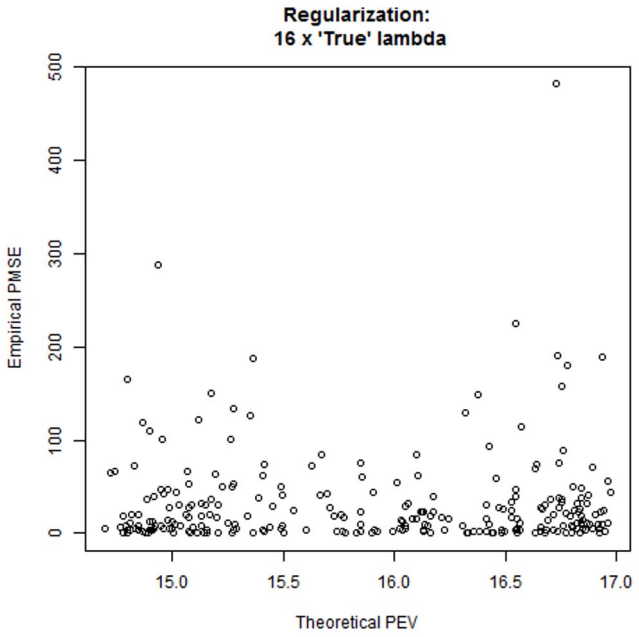

Figure 11. Disconnect between expected prediction error variances (theoretical PEV) and empirical bootstrap average squared prediction errors (empirical PEV) for a simulation under the settings of case study 4. True lambda = 20.
doi:10.1371/journal.pone.0091693.g011

$$\frac{\partial \hat{\mathbf{y}}_{\text{Test}}}{\partial \mathbf{y}'_{\text{Train}}} = \mathbf{X}_{\text{Test}} \left(\frac{\mathbf{I}_p - \mathbf{R}_\beta}{\lambda} \right) \mathbf{X}'_{\text{Train}}.$$

Hence, the predicted values can be seen to be less sensitive with respect to variation in training data when reliabilities (in this case of marker effects, but the same logic carries for GBLUP) increase and when λ gets larger; when $\lambda \to \infty$ the model becomes essentially null as the effective number of parameters goes to 0. Informally, $\mathbf{R}_\beta$ has an upper bound at $\mathbf{I}_p$ so, when reliabilities are perfect, predictions are insensitive with respect to variation in training data. However, even in this perfect case and assuming the model is correct, there is no clear connection between reliability and predictive outcome.

Bagging did reduce the variability of GBLUP predictions and, as observed in our case studies, it enhanced predictive performance when the model was "under-regularized". When, regularization was near optimum, bagging did not improve predictive performance, but it provided a means for assessing predictive mean squared error for any individual or candidate item in a testing set. This is because bagging can emulate variation in training data sets of a given size, allowing calculation of conditional (given $\mathbf{X}_{\text{Train}}$, $\mathbf{X}_{\text{Test}}$ and $\mathbf{y}_{\text{Test}}$) mean squared errors and of a measure of "reliability" connecting directly to predictive outcomes. These measures reflect variation in the predictor (rendered small by bagging), prediction bias and, of course, noise inherent to the fact that prediction can never be perfect. We did not find that bagging deteriorated predictive performance in any

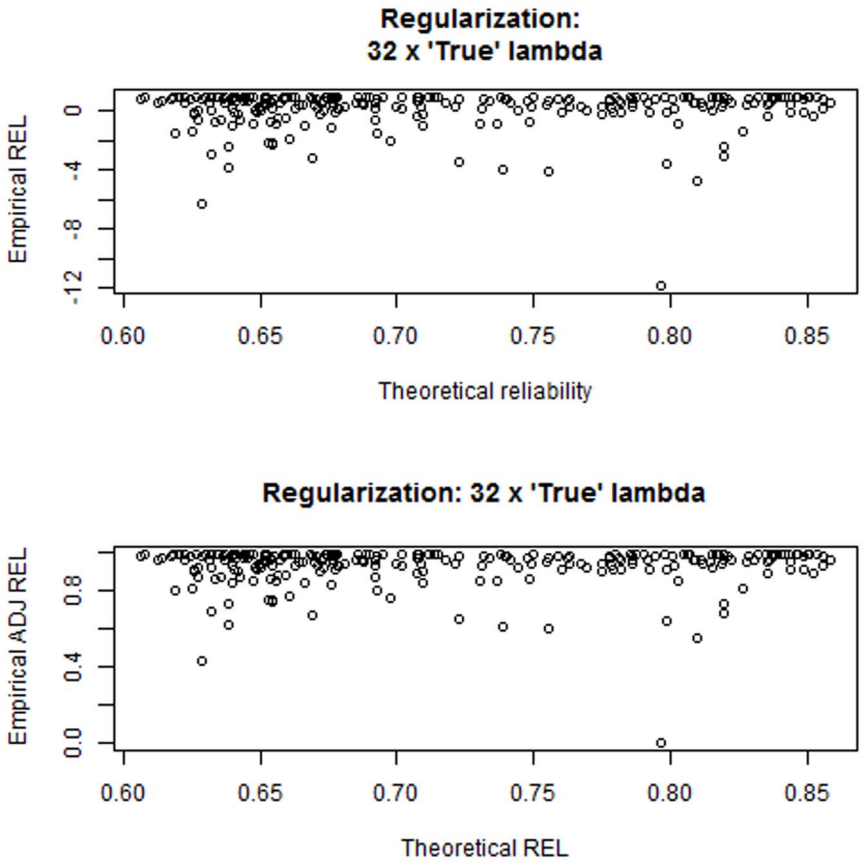

Figure 12. Relationship between expected reliabilities and empirical reliabilities (see text) in the top panel, and empirical adjusted reliabilities (see text) in the bottom panel.
doi:10.1371/journal.pone.0091693.g012

of the settings simulated, with only a slight hint in the wheat data set when regularization was excessive. As anticipated by [31] bagging helped when the predictor was more variable, due to small shrinkage. Coupled with the finding that predictive performance was not degraded otherwise, it seems that bagging confers robustness to the GBLUP prediction machine.

Conclusions

In short, bagging ameliorated the predictive performance of GBLUP, providing a means for developing candidate-specific measures of cross-validation reliability. It is computationally intensive when one searches for an optimum value of λ because of the simultaneous bootstrapping. In our study it seemed that 25–50 bootstrap samples were enough to attain reasonable predictions as well as stable measures of individual predictive mean squared errors. In practice, λ can be assessed by estimating the variance components in some data set and this may need to be done only once; the optimum λ in cross-validation is often close to what one obtains from estimating λ in the training set (de los Campos, personal communication), but regularization depends on the p/n ratio, so studies from other studies with different sample sizes (even from the same population) may not provide a good guide to attain optimum regularization in a given problem.

Bagging may not be feasible for immense data sets, but a question is whether or not such huge data sets are really needed for attaining an optimal predictive performance. Typically, predictive ability increases with training set size [27], [30] but it plateaus at some point. A smaller training data set with less "molecular redundancy" than a huge data set that spans less genomic variation may deliver a better predictive performance. Perhaps the design of training data sets needs to be studied in more depth. An unresolved problem is that of reducing empirical prediction bias, which is a research area of potential interest.

Supporting Information

Appendix S1 Bagging may decrease predictive mean-squared error.
(PDF)

Author Contributions

Conceived and designed the experiments: DG. Performed the experiments: DG. Analyzed the data: DG. Wrote the paper: DG. Contributed ideas and made critical comments: KAW NK AS CCS.

References

1. Van Raden PM (2008) Efficient methods to compute genomic predictions. Journal of Dairy Science 91:4414–4423.
2. VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, et al. (2009) Reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science 92:16–24.
3. Henderson CR (1976) A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding value. Biometrics 32:69–83.
4. Nejati-Javaremi A, Smith C, Gibson J (1997) Effect of total allelic relationship on accuracy of evaluation and response to selection. Journal of Animal Science 75:1738–1745.
5. Astle W, Balding DJ (2009) Population structure and cryptic relatedness in genetic association studies. Statistical Science 4:451–471.
6. Forni S, Aguilar I, Misztal I (2011) Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genetics, Selection, Evolution 43: 1.
7. 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 Genetics 9 e1003608.
8. Henderson CR (1973) Sire Evaluation and Genetic Trends. *Proceedings of the Animal Breeding and Genetics Symposium in Honor of Dr. Jay L. Lush*, 10–41, Champaign: American Society of Animal Science and the American Dairy Science Association.
9. Henderson CR (1977) Best linear unbiased prediction of breeding values not in the model for records. Journal of Dairy Science 60:783–787.
10. Henderson CR (1984) *Application of Linear Models in Animal Breeding*, Guelph: University of Guelph.
11. Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship Information on genome-assisted breeding values. Genetics 177:2389–2397.
12. Goddard ME (2008) Genomic selection: prediction of accuracy and maximisation of long term response. Genetica Epub PMID: 18704696.
13. Janss L, de los Campos G, Sheehan N, Sorensen D (2012) Inferences from genomic models in stratified populations. Genetics 92: 693–704.
14. de los Campos G, Gianola D, Rosa GJM (2009) Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. Journal of Animal Science 87:1883–1887.
15. Gianola D, Fernando RL, Stella A (2006) Genomic assisted prediction of genetic value with semi-parametric procedures. Genetics 173:1761–1776.
16. Gianola D, van Kaam JBCHM (2008) Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits 178(4): 2289–2303.
17. Gianola D, Okut H, Weigel KA, Rosa GJM (2011) Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. BMC Genetics 2011 12:87 doi:10.1186/1471-2156-12-87.
18. Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.
19. Gianola D (2013) Priors in whole-genome regression: the Bayesian alphabet returns. Genetics 194:573–596.
20. Ober U, Ayroles JF, Stone EA, Richards S, Zhu D, et al. (2012) Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster. PLos Genetics 8 e1002685.
21. Heslot N, Yang HP, Sorrells ME, Jannink JL (2012) Genomic selection in plant breeding: a comparison of models. Crop Science 52:146–160.
22. Ogutu JO, Schulz-Streeck T, Piepho HP (2012) Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. BMC Proceedings 6 (Suppl 2): S10. http://www.biomedcentral.com/1753-6561/6/S2/S10
23. Wimmer V, Lehermeier C, Albrecht T, Auinger HJ, Wang Y, et al. (2013) Genome-wide prediction of traits with different genetic architecture through efficient variable selection. Genetics 195:573–587.
24. Lehermeier C, Wimmer V, Albrecht T, Auinger H, Gianola D, et al. (2012) Sensitivity to prior specification in Bayesian genome-based prediction models. Statistical Applications in Genetics and Molecular Biology 12: 375–391. doi: 10.1515/sagmb-2012-0042.
25. Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, et al. (2010) A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. Journal of Dairy Science 93:743–752.
26. de los Campos G, Gianola D, Allison DAB (2010) Predicting genetic predisposition in humans: the promise of whole-genome markers. Nature Reviews Genetics 11:880–886.
27. Daetwyler HD, Villanueva B, Woolliams JA (2008) Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS One 3: e3395.
28. Daetwyler HD, Hickey JM, Henshall JM, Dominik S, Gredler B, et al. (2010) Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. Animal Production Science 50:1004–1010.
29. Erbe M, Pimentel ECG, Sharifi AR, Simianer H (2010) Assessment of Cross-validation Strategies for Genomic Prediction in Cattle. Book of Abstracts of the 9th WCGALP, S. 129. Leipzig, Germany.
30. Erbe M (2013) Accuracy of genomic prediction in dairy cattle. PhD Thesis, George-August University, Göttingen, Germany.
31. Breiman L (1996) Bagging predictors. Machine Learning 24:123–140.
32. Breiman L (2001a) Using iterated bagging to debias regressions. Machine Learning 45:261–277.
33. Suen Y-L, Melville P, Mooney RJ (2005) Combining bias and variance reduction techniques for regression trees. Proceedings of the 16th European Conference on Machine Learning. pp. 741–749. Porto, Portugal.
34. Valle C, Nanculef R, Allende H, Moraga C (2007) Two bagging algorithms with coupled learners to encourage diversity. In: Berthold MR, Shawe-Taylor J, Lavrac N, editors. Advances in Intelligent Data Analysis VII. LNCS 4723, pp. 130–139.
35. Gianola D, Foulley JL (1999) Variance estimation from integrated likelihoods (VEIL). Genetics, Selection, Evolution 22:403–417.
36. Falconer DS, Mackay TFC (1996) *Introduction to Quantitative Genetics*. Ed. 4. Longmans Green, Harlow, Essex, UK.
37. Inoue A, Kilian L (2008) How useful is bagging in forecasting economic time series? A case study of U.S. Consumer Price Inflation. Journal of the American Statistical Association 103:511–522.

38. Breiman L (2001b) Random Forests. Machine Learning 45:5–32.
39. Bulmer MG (1980) *The Mathematical Theory of Quantitative Genetics*, Oxford: Oxford University Press.
40. Fernando RL, Gianola D (1986) Optimal properties of the conditional mean as a selection criterion. Theoretical and Applied Genetics 72: 822–825.
41. Casella G, George EI (1992) Explaining the Gibbs Sampler. The American Statistician 46:167–174.
42. Crossa J, de Los Campos G, Pérez P, Gianola D, Burgueño J, et al. (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724.
43. Long N, Gianola D, Rosa GJM, Weigel KA (2011) Marker-assisted prediction of non-additive genetic values. Genetica 139:843–854.
44. Takezawa K (2006) *Introduction to Non-parametric Regression*. Wiley-Interscience, Hoboken.
45. Makowsky R, Pajewski NM, Klimentidis YC, Vázquez AI, Duarte CW, et al. (2011) Beyond missing heritability: prediction of complex traits. PLoS Genetics doi:10.1371/journal.pgen.1002051.