On the accuracy of genomic selection

Charles-Elie Rabier, Philippe Barre, Torben Asp, Gilles Charmet, Brigitte Mangin

PLOS ONE

RESEARCH ARTICLE

On the Accuracy of Genomic Selection

Charles-Elie Rabier[1]*, Philippe Barre[2], Torben Asp[3], Gilles Charmet[4], Brigitte Mangin[5]*

1 MIAT, Université de Toulouse, INRA, Castanet-Tolosan, France, **2** UR4, INRA, Unité de Recherche Pluridisciplinaire, Prairies et Plantes Fourragères, Lusignan, France, **3** Department of Molecular Biology and Genetics, Aarhus University, Slagelse, Denmark, **4** GDEC, UMR INRA-UBP, Clermont-Ferrand, France, **5** LIPM, Université de Toulouse, INRA, CNRS, Castanet-Tolosan, France

* cerabier@insa-toulouse.fr; ce.rabier@gmail.com (CER); Brigitte.Mangin@toulouse.inra.fr (BM)

Abstract

Genomic selection is focused on prediction of breeding values of selection candidates by means of high density of markers. It relies on the assumption that all quantitative trait loci (QTLs) tend to be in strong linkage disequilibrium (LD) with at least one marker. In this context, we present theoretical results regarding the accuracy of genomic selection, i.e., the correlation between predicted and true breeding values. Typically, for individuals (so-called test individuals), breeding values are predicted by means of markers, using marker effects estimated by fitting a ridge regression model to a set of training individuals. We present a theoretical expression for the accuracy; this expression is suitable for any configurations of LD between QTLs and markers. We also introduce a new accuracy proxy that is free of the QTL parameters and easily computable; it outperforms the proxies suggested in the literature, in particular, those based on an estimated effective number of independent loci (M_e). The theoretical formula, the new proxy, and existing proxies were compared for simulated data, and the results point to the validity of our approach. The calculations were also illustrated on a new perennial ryegrass set (367 individuals) genotyped for 24,957 single nucleotide polymorphisms (SNPs). In this case, most of the proxies studied yielded similar results because of the lack of markers for coverage of the entire genome (2.7 Gb).

Citation: Rabier C-E, Barre P, Asp T, Charmet G, Mangin B (2016) On the Accuracy of Genomic Selection. PLoS ONE 11(6): e0156086. doi:10.1371/journal.pone.0156086

Editor: Tongming Yin, Nanjing Forestry University, CHINA

Received: September 10, 2015

Accepted: May 9, 2016

Published: June 20, 2016

Data Availability Statement: Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.jb17n.

Funding: This work was supported by the CROPDL project, which is a part of the INRA Meta-program SELGEN (http://www.selgen.inra.fr/). The funders had no role in 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.

Introduction

During the last decades, investigators have mainly concentrated on linkage analysis to detect the regions of DNA, so-called quantitative trait loci (QTLs), responsible for quantitative variation. The linkage analysis (LA) is specific because it relies on family data and on pedigrees: segregation of a QTL is studied within a family by means of information related to the family.

In this context, the most popular statistical method for QTL mapping is interval mapping [1]. It involves scanning the genome by means of genetic markers and testing for the presence/absence of a QTL at every location in the genome. The mathematical theory behind this concept has been extensively studied for many years and is now well established [2–4]. According to [5], thousands of QTLs have been detected in plants, animals, and humans by means of interval mapping as a statistical tool. For instance, [6] detected QTLs responsible for a

reduction in grain shattering in cultivated rice, and [7] uncovered a QTL responsible for tomato fruit size.

More recently, researchers moved on to genome-wide association studies (GWAS) that are based on unrelated individuals, in contrast to LA. A GWAS allows researchers to analyze individuals without knowing their pedigree. One of the most popular methods relies on the model proposed by [8]. Hundreds of SNP-trait associations have been discovered in humans [9, 10] by means of GWAS: 30 loci are now known to be linked to Crohn's disease [11], and approximately 40 loci are associated with human height [12, 13].

Nonetheless, both approaches have a drawback: QTLs with very small effects are difficult to detect. Note that most traits of interest can be characterized as complex traits: they are presumably governed by a large number of small-effect QTLs [14–16]. In a large number of studies, the detected QTLs could not explain all genetic variation [17, 18]. It should be noted that this phenomenon explains a part of the so-called missing heritability. Typically, predictions based on selected SNPs have not been reliable.

At present, genomic selection (GS) is focused on prediction of breeding values of selection candidates by means of high density of markers. In contrast to LA and GWAS, the main goal of GS is not to detect QTLs anymore but to predict the future phenotype of young candidates as soon as their DNA has been collected. GS relies on the expectation that some QTLs will be in strong linkage disequilibrium (LD) with at least one marker [19]. From a theoretical point of view, GS differs from LA and GWAS because GS can be viewed as a whole-genome regression analysis [20, 21]: all the marker effects are estimated simultaneously. This way, it accounts for the correlation among SNPs, which is not the case when each SNP is analyzed separately. GS was first applied to animal breeding, especially dairy cattle (see [22]), where this new method has been found to be particularly promising. It was later tested on plants [23], with recent studies on apple [24], sugar beet [25], pea [26], and on inbred lines of rice [27].

A large number of methods can be chosen to make predictions in GS: penalized regression methods (see [28] for a review), Bayesian methods (see for instance [29]), and reproducing kernel Hilbert spaces methods [30, 31] are the most popular tools. Quality of the prediction is usually evaluated by accuracy criteria, such as the correlation between predicted and true breeding values. A large number of formulas for accuracy are now available in the literature. Most of them are inspired by the work of [32] who derived the formulas while relying only on the causal model with fixed effects and assuming independence of causal loci. Later, this work was extended in [33], in order to allow for the presence of a large number of loci (in the genome) that can not be considered independent due to linkage and a fixed genome size. The authors proposed, in particular, to substitute the effective number of independent loci M_e into the original formula of [32]. Subsequently, a large number of research groups built on this concept and proposed different ways of estimating M_e. Those methods are either based on the effective population size (e.g., [34, 35]), or on the number of independent tests in association mapping [36]. A comparison among the methods relying on the effective population size is presented in [37].

There are many questions about GS. The choice of the training (TRN) population with respect to the test (TST) population is a hot area of research. This procedure seems to have a strong influence on predictions [38, 39]. In the mixed model framework, [40] proposed an optimization method based on the coefficient of determination. Note that by choosing the most informative individuals to phenotype (i.e., TRN individuals), researchers can use GS as a tool for reducing phenotyping costs. Another area of active research is the long-term behavior of GS [35, 41], for example, the influence of selection as a function of time, or the reliability of the predicted model as a function of time when only the first generation is phenotyped. With the increase in the number of genomic markers because of next-generation sequencing technologies, the question of selecting genomic regions, prior to the learning step has been

addressed in simulation studies [42] as well as studies on real-life data [43]. It has been shown that additional biological information can increase GS accuracy.

In the present study, we propose to focus on mathematical properties of the accuracy based on the regression model called random regression best linear unbiased predictor (RRBLUP) or genomic best linear unbiased predictor (GBLUP). This model, initially proposed by [44], is one of the most popular methods for prediction of breeding values. We present here a closed-form expression for the accuracy; this formula is suitable for any configurations of LD between QTLs and markers. Theoretical developments are made possible by analyzing the causal model and prediction model differently; this is generally not the case for investigators working on the mixed model [34, 40], except [45]. Our theoretical formula enables identification of the terms affecting the accuracy in GS, e.g., LD and the link between TRN and TST sets. Besides, with the help of our formula, we can obtain the key result of [32] regarding the accuracy, when we use the same assumptions that those authors used. Another interesting result in our paper is introduction of a new proxy for the accuracy; this proxy is free of the QTL parameters and is easily computable. We show that substituting an estimated effective number of independent loci (M_e) into [32]'s formula is not the appropriate way to work with the high dimensional framework. Another quantity is suggested here. This way, our study can be viewed as an answer to the article [37], where the authors expressed doubt about the existing proxies after comparing 145 accuracy values collected from 13 articles on GS.

In the text below, after a description of the mathematical theory, our theoretical results and existing formulas are compared on simulated data. At the end, an illustration of real-life data is presented. We analyzed GS in plant height in perennial ryegrass, using 24,957 SNPs obtained via genotyping by sequencing (GBS) from 367 genotypes.

Materials and Methods

The theory

In this section, we assume, without a loss of generality, that coded genotypes at the markers and at QTLs are centered, as well as the phenotypic observations.

The causal linear model. The quantitative trait is observed in n_{TRN} TRN individuals, and we denote the observations as $Y_1, \ldots, Y_{n\mathrm{TRN}}$. C QTLs are present in the genome and have an effect on the quantitative trait. In the text below, θ_j refers to the fixed QTL effect of the j-th QTL and $Q_{i,j}$ denotes the corresponding coded genotype for individual i. We assume the following causal linear model for the quantitative trait:

$$Y_i = \sum_{j=1}^{C} Q_{i,j}\,\theta_j + e_i \quad (i = 1, \cdots, n_{\mathrm{TRN}})$$

where $e_i \sim N(0, \sigma_e^2)$ and σ_e^2 denotes the environmental variance.

With matrix and vector notation, this model can be rewritten as

$$\boldsymbol{Y} = \boldsymbol{Q\theta} + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{Q}$ is a $n_{\mathrm{TRN}} \times C$ matrix, $\boldsymbol{Y} = (Y_1, \ldots, Y_{n\mathrm{TRN}})'$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_C)'$, $\mathbf{e} \sim N(0, \sigma_e^2 \boldsymbol{I}_{n_{\mathrm{TRN}}})$ and $\boldsymbol{I}_{n\mathrm{TRN}}$ is the identity matrix of size n_{TRN}.

In the text that follows, $\boldsymbol{q_i}$ denotes a vector of size $C \times 1$ that refers to the "causal genome" of individual i.

Introducing a TST individual. A supplementary individual, a so-called TST individual (denoted as $n_{\mathrm{TRN}+1}$) is genotyped but not phenotyped. With the same notation as in the TRN population, $\boldsymbol{q}_{n\mathrm{TRN}+1}$ denotes the genome at QTL locations of individual $n_{\mathrm{TRN}+1}$. As a result,

the quantitative trait $Y_{n_{\mathrm{TRN}}+1}$ can be expressed as

$$Y_{n_{\mathrm{TRN}}+1} = \boldsymbol{q}'_{\boldsymbol{n}_{\mathrm{TRN}}+1}\boldsymbol{\theta} + e_{n_{\mathrm{TRN}}+1}$$

where $e_{n_{\mathrm{TRN}}+1} \sim N(0, \sigma_e^2)$. Next, $\boldsymbol{q}_{\boldsymbol{n}\mathrm{TRN}+1}$ will be considered random. Recall that $\boldsymbol{\theta}$ is fixed.

Accuracy. In GS, we are interested in predicting either the genotypic value $\boldsymbol{q}'_{\boldsymbol{n}_{\mathrm{TRN}}+1}\boldsymbol{\theta}$, or the phenotypic value $Y_{n\mathrm{TRN}+1}$. In both cases, a predictor $\hat{Y}_{n_{\mathrm{TRN}}+1}$ is constructed by means of a prediction model developed through learning on n_{TRN} TRN individuals. Then, the quality of the prediction is evaluated according to some accuracy criteria. In particular, the phenotypic accuracy, ρ, and the genotypic accuracy, $\tilde{\rho}$, are defined as follows:

$$\rho = \frac{\mathrm{Cov}\left(\hat{Y}_{n_{\mathrm{TRN}}+1},\ Y_{n_{\mathrm{TRN}}+1}\right)}{\sqrt{\mathrm{Var}\left(\hat{Y}_{n_{\mathrm{TRN}}+1}\right)\mathrm{Var}\left(Y_{n_{\mathrm{TRN}}+1}\right)}}\ ,\ \tilde{\rho} = \frac{\mathrm{Cov}\left(\hat{Y}_{n_{\mathrm{TRN}}+1},\ \boldsymbol{q}'_{\boldsymbol{n}_{\mathrm{TRN}}+1}\boldsymbol{\theta}\right)}{\sqrt{\mathrm{Var}\left(\hat{Y}_{n_{\mathrm{TRN}}+1}\right)\mathrm{Var}\left(\boldsymbol{q}'_{\boldsymbol{n}_{\mathrm{TRN}}+1}\boldsymbol{\theta}\right)}}\ . \tag{2}$$

These two types of accuracy are linked by the relation $\rho/\tilde{\rho} = h$, where h is the square root of the heritability of the trait:

$$h^2 = \frac{\boldsymbol{\theta}'\mathrm{Var}\left(\boldsymbol{q}_{\boldsymbol{n}_{\mathrm{TRN}}+1}\right)\boldsymbol{\theta}}{\boldsymbol{\theta}'\mathrm{Var}\left(\boldsymbol{q}_{\boldsymbol{n}_{\mathrm{TRN}}+1}\right)\boldsymbol{\theta} + \mathrm{Var}\left(e_{n_{\mathrm{TRN}}+1}\right)}\ . \tag{3}$$

Next, we set $\sigma_G^2 = \boldsymbol{\theta}'\mathrm{Var}(\boldsymbol{q}_{\boldsymbol{n}\mathrm{TRN}+1})\boldsymbol{\theta}$, and as a consequence, we have the relationship $h^2 = \sigma_G^2/(\sigma_G^2 + \sigma_e^2)$. Depending on a research group, investigators focus either on phenotypic accuracy ρ (e.g., [46]), or on genotypic accuracy $\tilde{\rho}$ (e.g., [32, 33]).

The oracle situation. Suppose this situation denotes the settings where the QTL locations and their effects are known. Then, the natural predictor, $\hat{Y}_{n_{\mathrm{TRN}}+1}$, of the quantity $Y_{n\mathrm{TRN}+1}$ is

$$\hat{Y}_{n_{\mathrm{TRN}}+1} = \boldsymbol{q}'_{\boldsymbol{n}_{\mathrm{TRN}}+1}\boldsymbol{\theta}\ .$$

As a result, the oracle accuracies, so-called ρ_{oracle} and $\tilde{\rho}_{\mathrm{oracle}}$ for phenotypic and genotypic accuracy, satisfy the following equations:

$$\rho_{\mathrm{oracle}} = \frac{\mathrm{Cov}\left(\boldsymbol{q}'_{\boldsymbol{n}_{\mathrm{TRN}}+1}\boldsymbol{\theta},\ Y_{n_{\mathrm{TRN}}+1}\right)}{\sqrt{\mathrm{Var}\left(\boldsymbol{q}'_{\boldsymbol{n}_{\mathrm{TRN}}+1}\boldsymbol{\theta}\right)\mathrm{Var}\left(Y_{n_{\mathrm{TRN}}+1}\right)}} = h \quad \text{and} \quad \tilde{\rho}_{\mathrm{oracle}} = 1\ .$$

A marker-based model. In practice, the QTL effects and their locations are typically unknown. As a consequence, the prediction is based on information from p genetic markers located in the genome. Suppose $\boldsymbol{X}$ denotes the TRN incidence matrix of size $n_{\mathrm{TRN}} \times p$. The TRN marker-based model is typically the following random effects model:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_{n\mathrm{TRN}})', \boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)' \sim N(0, \sigma_\beta^2\boldsymbol{I}_{\boldsymbol{p}}), \boldsymbol{\varepsilon} \sim N(0, \sigma_\varepsilon^2\boldsymbol{I}_{\boldsymbol{n}_{\mathrm{TRN}}})$. This setting allows to work with the high dimensional framework, i.e. the situation where $p > n_{\mathrm{TRN}}$. σ_β^2 and σ_ε^2 denote respectively the variance of the marker effects and the residual variance.

By the same token, $\boldsymbol{x_i}$ is a vector of size $p \times 1$ that refers to genomic markers of individual i. Recall that $\boldsymbol{q_i}$ represents the "causal genome" of individual i. Besides, in the text that follows, we use the notation $X_{i,j}$ for the coded genotype of individual i at the j-th marker.

This model was initially proposed by [44]. In the literature, it is known as GBLUP or RRBLUP. As a consequence, the estimated effects of SNPs are

$$\hat{\boldsymbol{\beta}} = \mathbb{E}(\boldsymbol{\beta} \mid \boldsymbol{Y}, \boldsymbol{X}) = \left(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1} \boldsymbol{X}'\boldsymbol{Y} \quad \text{where} \quad \lambda = \sigma_\varepsilon^2/\sigma_\beta^2.$$

Suppose $\boldsymbol{x}_{n_{\mathrm{TRN}}+1}$ denotes the random variable corresponding to the genomic markers of the TST individual. Then, the prediction is

$$\begin{aligned} \hat{Y}_{n_{\mathrm{TRN}}+1} = \boldsymbol{x}'_{n_{\mathrm{TRN}}+1}\hat{\boldsymbol{\beta}} &= \boldsymbol{x}'_{n_{\mathrm{TRN}}+1}(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}'\boldsymbol{Y} \\ &= \boldsymbol{x}'_{n_{\mathrm{TRN}}+1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y} \text{ where } \boldsymbol{V} = \boldsymbol{X}\boldsymbol{X}' + \lambda \boldsymbol{I}_{n_{\mathrm{TRN}}}. \end{aligned} \tag{4}$$

The RRBLUP model is also called ridge regression and the parameter λ is viewed as a regularization parameter (see for instance [47]).

The main result of this paper is the following. Recall that $\boldsymbol{\theta}$ is fixed, and that $\boldsymbol{q}_{n_{\mathrm{TRN}}+1}$ and $\boldsymbol{x}_{n_{\mathrm{TRN}}+1}$ are random. Conditionally on both the TRN incidence matrix $\boldsymbol{X}$, and the TRN causal matrix $\boldsymbol{Q}$, the phenotypic accuracy is

$$\rho_{\mathrm{RR}} = \frac{\boldsymbol{\theta}' \, \mathbb{E}\left(\boldsymbol{q}_{n_{\mathrm{TRN}}+1}\boldsymbol{x}'_{n_{\mathrm{TRN}}+1}\right)\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Q}\boldsymbol{\theta}}{\left(\sigma_e^2 \, \mathbb{E}\left(\| \, \boldsymbol{x}'_{n_{\mathrm{TRN}}+1}\boldsymbol{X}'\boldsymbol{V}^{-1} \, \|^2\right) \; + \; \boldsymbol{\theta}'\boldsymbol{Q}'\boldsymbol{V}^{-1}\boldsymbol{X}\mathrm{Var}\left(\boldsymbol{x}_{n_{\mathrm{TRN}}+1}\right)\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Q}\boldsymbol{\theta}\right)^{1/2}\left(\sigma_G^2 + \sigma_e^2\right)^{1/2}} \tag{5}$$

where $\|.\|$ is the L^2 norm, and $\mathrm{Var}(\boldsymbol{x}_{n_{\mathrm{TRN}}+1})$ is the covariance matrix of size $p \times p$. The proof is provided in S1 Text.

Finally, we want to emphasize that this closed-form expression for the accuracy was derived without any assumptions on QTL locations and marker locations. In other words, the formula deals with the configuration where QTLs match a few genetic markers as well as the configuration where QTLs are not located on markers.

A new proxy (QTLs in perfect LD with some markers). Let us assume now that the C QTLs are located exactly on genetic markers, but at the same time, let us allow the number of genetic markers to be much larger than the number of QTLs (i.e., $p >> C$). Suppose that

$$\boldsymbol{x}'_{n_{\mathrm{TRN}}+1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Q}\boldsymbol{\theta} = \boldsymbol{q}'_{n_{\mathrm{TRN}}+1}\boldsymbol{\theta}.$$

This is an ideal situation where each QTL is in perfect LD with its associated marker, with respect to the $\boldsymbol{V}^{-1}$ matrix. Besides, each QTL is in linkage equilibrium with other markers, with respect to the $\boldsymbol{V}^{-1}$ matrix. This assumption is appropriate in a random mating population (with a large number of individuals), evolving during a large number of generations because the LD decreases at an exponential rate (e.g., [48, 49]).

Then, according to formula (5) and the proof provided in S1 Text, the accuracy becomes

$$\rho_{\mathrm{pLD}} = h \sqrt{\frac{h^2/(1-h^2)}{\mathbb{E}\left(\| \, \boldsymbol{x}'_{n_{\mathrm{TRN}}+1}\boldsymbol{X}'\boldsymbol{V}^{-1} \, \|^2\right) + \frac{h^2}{1-h^2}}}. \tag{6}$$

Contrary to the general accuracy presented in formula (5), this accuracy can be computed easily because it depends on known or estimable quantities: the heritability of the trait is usually known, and the expectation on the denominator can be estimated using the empirical mean in the TST sample. Finally, the tuning parameter λ that is present in the expression for $\boldsymbol{V}$, can be estimated by several statistical methods. In this paper, we used Restricted Maximum Likelihood (REML) [50] or deduced λ from the heritability.

The link with the work of [32]. In [32], a seminal formula for the accuracy is presented. We would like to show here that with the help of our general formula (5), we can obtain this previously published result, if we use the same assumptions that those authors used. In particular, [32] assumed that the QTL locations are known and that each QTL is in perfect LD with its associated marker. The formula was obtained by performing regression analysis of the trait on each QTL separately; this approach is equivalent to assuming that $\boldsymbol{Q}'\boldsymbol{Q}$ is diagonal and thus invertible. Even the case $C >> n_{\mathrm{TRN}}$ can be analyzed because of this above assumption. Then, the estimated QTL effects and the prediction are

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{Q}'\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}'\boldsymbol{Y} \quad , \quad \hat{Y}_{n_{\mathrm{TRN}}+}1 = \boldsymbol{q}'_{n_{\mathrm{TRN}}+1}\left(\boldsymbol{Q}'\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}'\boldsymbol{Y} \quad .$$

Note that $\hat{\boldsymbol{\beta}}$ is obtained by assuming that $\lambda = 0$ in formula (4). In this context, according to our general formula (5) and calculations shown in S1 Text, the accuracies are the following:

$$\tilde{\rho} = \sqrt{\frac{h^2/(1-h^2)}{\frac{C}{n_{\mathrm{TRN}}} + \frac{h^2}{1-h^2}}} \quad , \quad \rho = h\sqrt{\frac{h^2/(1-h^2)}{\frac{C}{n_{\mathrm{TRN}}} + \frac{h^2}{1-h^2}}} \, . \tag{7}$$

This expression for $\tilde{\rho}$ is suitable for any values of σ^2_G and σ^2_e. This is not the case in [32] because those authors analyzed the case $\sigma^2_G + \sigma^2_e = 1$ and used the approximation $\sigma^2_e = 1$. We refer readers to S1 Text for more details.

Simulation study

In order to verify the validity of our theoretical results, a simulation study was performed.

Simulated data. Genomic data were generated by means of the hypred R package [51]. Populations were simulated by random mating between haploid individuals, during (a) 30, (b) 50, or (c) 70 generations. Recombination was modeled according to Haldane [52]. Mutations were not taken into consideration. In generation zero, two haploid founder lines were crossed. These two lines were completely different genetically. Generation 1 consisted of (a) 400 or (b) 800 haploid offspring of these two founders. After that, the population kept evolving by random mating with a constant size at each generation and no overlapping generation. This type of simulation mimics recombinant inbred line (RIL) or double haploid (DH) evolving populations. In the final generation, 2 individuals were randomly selected, and 100 (resp. 200) full sibs were generated under the 400 (resp. 800) offspring scenario, in order to get some closely related individuals.

The focus was on one chromosome of length 1 Morgan. We considered 4 different densities of genetic markers equally spaced on the chromosome: (a) 100, (b) 1,000, (c) 5,000, or (d) 10,000 SNPs. We considered two configurations for the phenotypic model: (a) 2 QTLs located at 3cM and 80cM with effects +1 and −2, respectively, or (b) 100 QTLs located every centimorgan, with the same effect +0.15. The environmental variance σ^2_e was set to 1. Table 1 shows the estimated heritabilities corresponding to the different scenarios studied. These heritabilities are based on the overall population (TRN+TST). Indeed, although a few closely related individuals were present in the TRN set, we did not observe a noticeable change in terms of heritability between the TST set and the TRN set.

Genetic markers without polymorphism were filtered out. Besides, identical SNPs along the chromosome were also filtered out; we kept only the first occurrence of that SNP on the chromosome.

A set of either (a) 500 or (b) 1,000 TRN individuals was used for the learning step. Note that in both cases, the TRN set included the full sibs: among the 500 (resp. 1,000) TRN, 100 (resp. 200) were full sibs. The prediction model was evaluated on 100 TST (in all cases), that were produced in the last generation.

Table 1. Estimated heritability (h^2) as a function of the simulation setup.

Nb QTLs	Nb Generations	n_{TRN}	h^2
2	30	500	0.54
		1,000	0.53
	50	500	0.53
		1,000	0.53
	70	500	0.51
		1,000	0.49
100	30	500	0.75
		1,000	0.77
	50	500	0.65
		1,000	0.69
	70	500	0.57
		1,000	0.61

The average of 100 replicates. In each sample, heritability estimated using the estimator $\widehat{\mathrm{Var}}(q_i'\boldsymbol{\theta})/(\widehat{\mathrm{Var}}(q_i'\boldsymbol{\theta})+1)$ based on the overall population (n_{TRN} TRN + 100 TST).$\widehat{\mathrm{Var}}$ denotes empirical variance.

doi:10.1371/journal.pone.0156086.t001

In the text that follows, an "architecture" is a fixed number of: (a) SNPs; (b) generations; (c) QTL numbers, effects, and locations; (d) TRN individuals. A total of 100 replicates were generated according to a given architecture. Table 2 shows a summary of the different configurations studied, and Table 3 provides the number of remaining markers after filtering. As in [53], the QTL locations did not vary across replicates. Nonetheless, contrary to that article, the QTL effects always had the same values here.

Regularization parameter λ. For each replicate, predicted phenotypes of the TST dataset were obtained by RRBLUP. The regularization parameter λ was estimated in two ways. The first method relies on variance components estimated by REML. Then, the corresponding regularization parameter called λ_{REML} is

$$\lambda_{REML} = \hat{\sigma}^2_{\varepsilon}/\hat{\sigma}^2_{\beta}$$

where $\hat{\sigma}^2_{\varepsilon}$ and $\hat{\sigma}^2_{\beta}$ denote respectively the estimates of the environmental variance σ^2_{ε} and the variance σ^2_{β} of each SNP effect. The rrBLUP R package and in particular its function kin.blup were used to compute these variance components.

The second method relies on the heritability of the quantitative trait assuming that the genetic variance is spread out uniformly across all the genetic markers. Then, the tuning parameter called λ_{h^2} is defined as follows:

$$\lambda_{h^2} = \frac{1-h^2}{n_{\mathrm{TRN}}\ h^2}\ \sum_{j=1}^{\tilde{p}}\sum_{i=1}^{n_{\mathrm{TRN}}} X^2_{i,j}$$

Table 2. The different configurations studied.

Nb markers	100 / 1,000 / 5,000 / 10,000
Nb Generations	30 / 50 / 70
Nb QTLs	2 / 100
n_{TRN}	500 / 1,000

doi:10.1371/journal.pone.0156086.t002

Table 3. The average number of markers after filtering (based on 100 replicates).

Nb generations	Nb Markers	n_{TRN}	Nb Markers after filtering
30	100	500	100
		1,000	100
	1,000	500	766.51
		1,000	929.22
	5,000	500	1,262
		1,000	2,066.17
	10,000	500	1,353
		1,000	2,345.53
50	100	500	100
		1,000	100
	1,000	500	801.49
		1,000	950.08
	5,000	500	1,392.73
		1,000	2,267.54
	10,000	500	1,518
		1,000	2,609.32
70	100	500	99.98
		1,000	100
	1,000	500	812.37
		1,000	950.68
	5,000	500	1,451.6
		1,000	2,380.81
	10,000	500	1,591
		1,000	2,781.81

doi:10.1371/journal.pone.0156086.t003

where $\tilde{p}$ is the number of markers after filtering, and h^2 is the estimated heritability given in Table 1.

Empirical accuracy and Theoretical accuracy. In the text below, n_{TST} denotes the number of TST individuals, and $\boldsymbol{x}_{\boldsymbol{n}\mathrm{TRN}} + \boldsymbol{i}$ means the genomic markers of the i-th TST individual. In order to compute the so-called Theoretical accuracy, introduced in formula (5), we used the following estimators:

$$\frac{1}{n_{\mathrm{TST}}}\sum_{i=1}^{n_{\mathrm{TST}}}\left\|\boldsymbol{x}'_{\boldsymbol{n}_{\mathrm{TRN}}+i}\boldsymbol{X}'\boldsymbol{V}^{-1}\right\|^2,\ \boldsymbol{\theta}'\left(\frac{1}{n_{\mathrm{TST}}}\sum_{i=1}^{n_{\mathrm{TST}}}\boldsymbol{q}_{\boldsymbol{n}_{\mathrm{TRN}}+\boldsymbol{i}}\boldsymbol{x}'_{\boldsymbol{n}_{\mathrm{TRN}}+\boldsymbol{i}}\right)\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Q}\boldsymbol{\theta},$$

$$\frac{1}{n_{\mathrm{TST}}}\left[\boldsymbol{x}_{\boldsymbol{n}_{\mathrm{TRN}}+1},\ldots,\boldsymbol{x}_{\boldsymbol{n}_{\mathrm{TRN}}+\boldsymbol{n}_{\mathrm{TST}}}\right]\times\left[\boldsymbol{x}_{\boldsymbol{n}_{\mathrm{TRN}}+1},\ldots,\boldsymbol{x}_{\boldsymbol{n}_{\mathrm{TRN}}+\boldsymbol{n}_{\mathrm{TST}}}\right]',$$

$$\frac{1}{n_{\mathrm{TST}}}\sum_{i=1}^{n_{\mathrm{TST}}}\left(\boldsymbol{q}'_{\boldsymbol{n}_{\mathrm{TRN}}+\boldsymbol{i}}\boldsymbol{\theta}-\frac{1}{n_{\mathrm{TST}}}\sum_{i=1}^{n_{\mathrm{TST}}}\boldsymbol{q}'_{\boldsymbol{n}_{\mathrm{TRN}}+\boldsymbol{i}}\boldsymbol{\theta}\right)^2,$$

to estimate the quantities $\mathbb{E}(\|\boldsymbol{x}'_{\boldsymbol{n}\mathrm{TRN}+1}\boldsymbol{X}'\boldsymbol{V}^{-1}\|^2)$, $\boldsymbol{\theta}'\ \mathbb{E}(\boldsymbol{q}_{\boldsymbol{n}_{\mathrm{TRN}}+1}\boldsymbol{x}'_{\boldsymbol{n}_{\mathrm{TRN}}+1})\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Q}\boldsymbol{\theta}$, $\mathrm{Var}(\boldsymbol{x}_{\boldsymbol{n}\mathrm{TRN}+1})$, and σ^2_G, respectively. Besides, the true value was used for the environmental variance, i.e., $\sigma^2_e = 1$. The empirical accuracy was computed in the R software, with the empirical correlation between the predicted values and the true values.

Mean accuracy on replicates and the TRN incidence matrix. Recall that our theoretical result in formula (5) was obtained conditionally on the TRN incidence matrix $\boldsymbol{X}$ and

conditionally on the TRN causal matrix $\boldsymbol{Q}$. In most of the simulation results presented in this paper, $\boldsymbol{X}$ and $\boldsymbol{Q}$ are different across replicates. Indeed, for each replicate, a new population was generated by random mating according to the given architecture. The accuracy was computed according to formula (5) on each replicate, and finally, the mean accuracy was calculated on the 100 replicates. As a consequence, the focus was on a mean accuracy corresponding to a given architecture.

On the other hand, we also analyzed the case where $\boldsymbol{X}$ and $\boldsymbol{Q}$ do not vary across replicates. In this case, $\boldsymbol{X}$ and $\boldsymbol{Q}$ were obtained by generating only one TRN population associated with a given architecture. Only the TST incidence matrix was allowed to change across replicates. In particular, TST individuals were regenerated by random mating between individuals from the penultimate generation. New phenotypes (TRN+TST) were regenerated for every replicate, and as previously, the mean accuracy on the 100 replicates was computed.

The effective number of segments M_e. Most of the published proxies for the accuracy use the so-called effective number of independent loci (M_e). We focused on the three following representations of M_e:

$$M_{e1} = \frac{2N_eL}{\log(4N_el)}\ ,\ M_{e2} = \frac{2N_eL}{\log(2N_el)}\ ,\ M_{e3} = \frac{2N_eL}{\log(N_el)}$$

where L, l, and N_e denote the genome length, average chromosome length, and effective population size respectively. M_{e1} was proposed by [35], whereas M_{e2} and M_{e3} are from [34]. In our simulation study, N_e was estimated using theoretical results of [54]. In particular, the LD was computed between all SNPs (in the TRN incidence matrix), and the estimated N_e was the least-squares estimate of the fitted nonlinear model (cf. S2 Text). Another approach developped to handle the issue of multiple testing in association mapping studies [36] was also considered. Later, this method was applied to GS (e.g. [55]). The drawback of this method is that it requires that $p < n_{\mathrm{TRN}} + n_{\mathrm{TST}}$. To overcome this problem, the chromosome was split into 2 parts for the 1,000 SNPs scenario, and into 3 parts when 5,000 SNPs or 10,000 SNPs were analyzed. After that, the overall M_e was obtained by summing up the numbers of independent tests obtained separately for each part.

Real data study on perennial ryegrass

Our plant material belongs to the perennial ryegrass species (*Lolium perenne* L.), a diploid species (2n = 14) with a haploid genome size of 2.7 Gb [56]. This genome size was estimated by flow cytometry in picograms and transformed in to the number of bases assuming 978 Mb/pg [57]. *Lolium perenne* L. is a highly heterozygous species with strong inbreeding depression and a self-incompatibility system. Besides, the varieties are synthetics. The population was provided by the private breeding company Gie GRASS. The dataset consisted of 367 genotypes obtained after the multiplication by intercrossing of 12 genotypes during three generations. Moreover, the 12 genotypes were obtained from pair-crosses involving 8 different genotypes. Seeds were sown in individual pots in the first week of August 2013. They were cut regularly to promote tillering and were cloned. On April 16th, 2014, four clones per genotype were planted in the field at INRA Lusignan France (43°36'55.59"N; 4° 0'36.59"E) in randomized block design. On August 21st, 2014, the plants were cut at approximately 5 cm and plant height was measured immediately with a ruler. Plant height was remeasured on August 28th, 2014. The plant growth rate was calculated as the difference in plant height between the two dates divided by the number of growing degree days, with the base temperature of zero (138.5°C.days).

Molecular data were obtained by GBS following the same protocol as in [58]. DNA was extracted from 50 mg of dried leaves by the protocol described in [59]. PstI was used for

complexity reduction. The sequencing was performed by means of the Hiseq 2500 (Illumina; pair-end 2 × 150, but only one pair was used for analysis). Scythe (https://github.com/vsbuffalo/scythe) was used to demultiplex the GBS raw reads and to trim adaptor contamination, using the prior contamination rate set to 0.40. Sickle (https://github.com/najoshi/sickle) was used to quality-trim the demultiplexed reads using the parameters -q 20 -l 40. The demultiplexed and quality trimmed reads were aligned against a draft assembly of the perennial ryegrass genome (48,415 scaffolds) in the BWA aln software [60]. The resulting BAM files were further processed using the Genome Analysis Toolkit (GATK) version 2.7-4 [61]. Variant calling was performed by means of GATK's UnifiedGenotyper, and high-quality SNPs were extracted. The resulting SNPs were quality-filtered according to several criteria, for example, only variants with a quality score higher than 30 were retained. A total of 24,957 SNPs with the minimum allele frequency of 5% were scored. Plink [62] was used to calculate the Pearson coefficient of correlation between pairs of SNPs belonging to the same scaffold. Phenotypic and SNP data are available at doi:10.5061/dryad.jb17n.

Results

Simulated data

This section starts by considering QTLs in perfect LD with some markers. Later, the case of imperfect LD is also reviewed. We studied successively, with the help of simulated data, (a) reliability of the Theoretical accuracy from formula (5), (b) sensitivity to the regularization parameter λ, (c) the effects of a fixed TRN incidence matrix, (d) the pertinence of the proxy suggested by formula (6), and (e) a substitute for the effective number of segments. Note that in the following text, the TRN incidence matrix varies across replicates, unless stated otherwise.

Empirical accuracy versus Theoretical accuracy. Fig 1 shows a comparison between the Empirical accuracy and Theoretical accuracy. The tuning parameter λ was estimated by REML in both cases. Each point on the graph corresponds to mean accuracy (based on 100 replicates) associated with a given architecture.

According to the figure, the Theoretical accuracy matched the Empirical accuracy regardless of the architecture being considered. Besides, readers can see that the accuracy increased with the number of QTLs because the heritability increased. As expected, for given numbers of QTLs, generations, and markers, the greater the number of TRN individuals, the higher the accuracy was. For instance, when we considered 50 generations, 2 QTLs and 10K SNPs, the Theoretical Accuracy was estimated to be 0.65 for n_{TRN} = 500 and 0.68 for n_{TRN} = 1,000, although the heritability was the same.

To complete our simulation study, it is worth to consider the case of a mixture between major genes and multiple small QTLs which mimics probably better the common architecture for a lot of traits. This type of architecture was also used to investigate a larger range of heritability. So, we generated two large QTLs located at 3cM and 80cM, and 98 small QTLs located every centimorgan (except at 3cM and 80cM). We considered three scenarios: (a) large QTLs with effects +0.5 and −0.6, small QTLs with the same effect +0.07, (b) large QTLs with effects +1 and −0.7, small QTLs with the same effect +0.1, (c) large QTLs with effects +2 and −2, small QTLs with the same effect +0.1. In all cases, we focused on the configuration 1,000 SNPs, 500 TRN individuals, and 50 generations. The heritabilities associated to the different scenarios were: (a) h^2 = 0.34, (b) h^2 = 0.54, (c) h^2 = 0.71. According to our simulated data, the Theoretical accuracy matched exactly the Empirical accuracy for scenario (a) (0.52), and scenario (b) (0.68). A very good agreement was also observed for scenario (c): the Empirical accuracy was found to be equal to 0.80, whereas the Theoretical accuracy took the value 0.79.

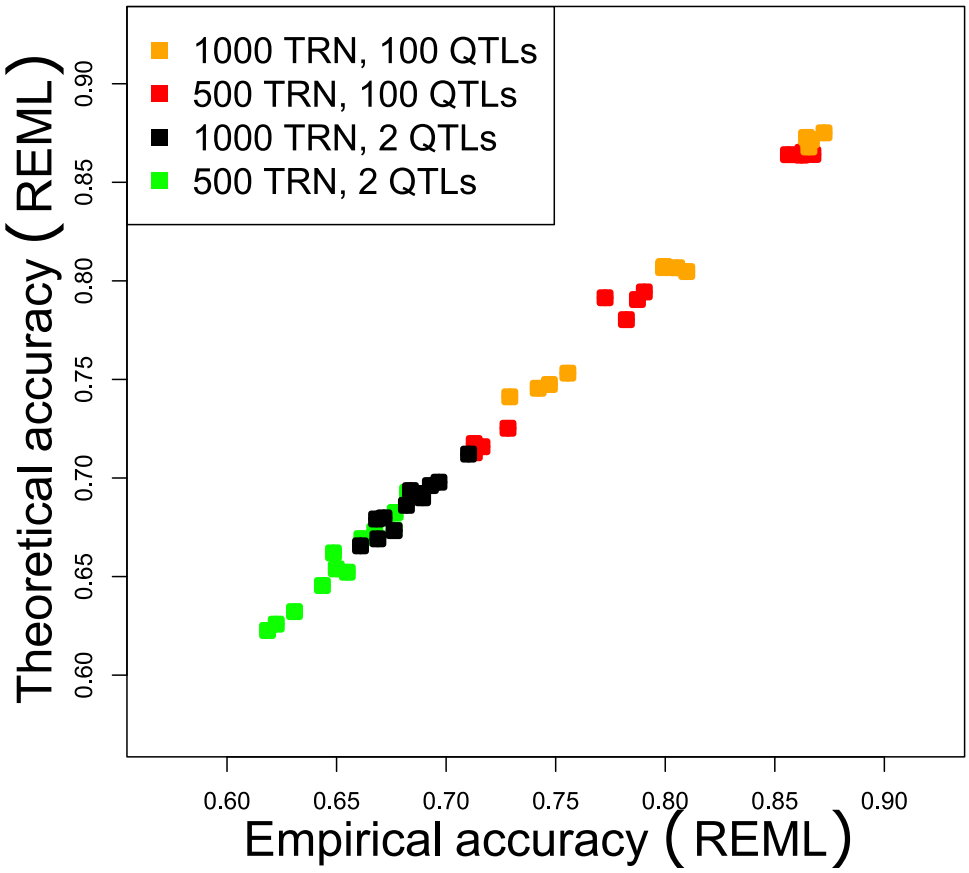

Fig 1. Comparison between the Theoretical accuracy and the Empirical accuracy, as a function of the number of TRN individuals and the number of QTLs. Tuning parameter λ was estimated by REML in both cases.

doi:10.1371/journal.pone.0156086.g001

Tuning parameter λ. Fig 2 shows analysis of sensitivity of the Theoretical and Empirical accuracies to the regularization parameter λ. We focused on two ways of estimating λ: one used REML, whereas the second one relied on heritability of the trait. According to Fig 2A, the Theoretical accuracy remained unchanged regardless of the method chosen for estimating λ. Because in practice, only approximated heritability is known to geneticists, we considered also the case where the tuning parameter was based on wrongly inferred heritability (90% of the true value). According to Fig 2B, the accuracy did not deteriorate: there was still good agreement between an Empirical accuracy based on a false heritability, and a Theoretical accuracy dependent on the true quantity.

QTLs in imperfect LD with some markers. Because our previous analysis implied that QTLs were in perfect LD with some markers, here, we analyze the case of imperfect LD. To mimic imperfect LD, the causal SNPs were unobserved in the TRN and TST populations. Fig 3A focuses on the 2 QTL scenario, and highlights the fact that our theoretical formula is also suitable under imperfect LD. Indeed, for simulated data without the causal SNPs in the marker-based model, the Theoretical accuracy matched the Empirical accuracy for all the different architectures. We also studied the effects of the presence/absence of the causal SNPs in the marker-based model as a function of marker density (Fig 3B). Readers can notice that the Theoretical accuracy was not affected by the absence of the causal SNPs, provided that the density of markers remained high (at least 1,000 markers). As explained in [63], on a dense map,

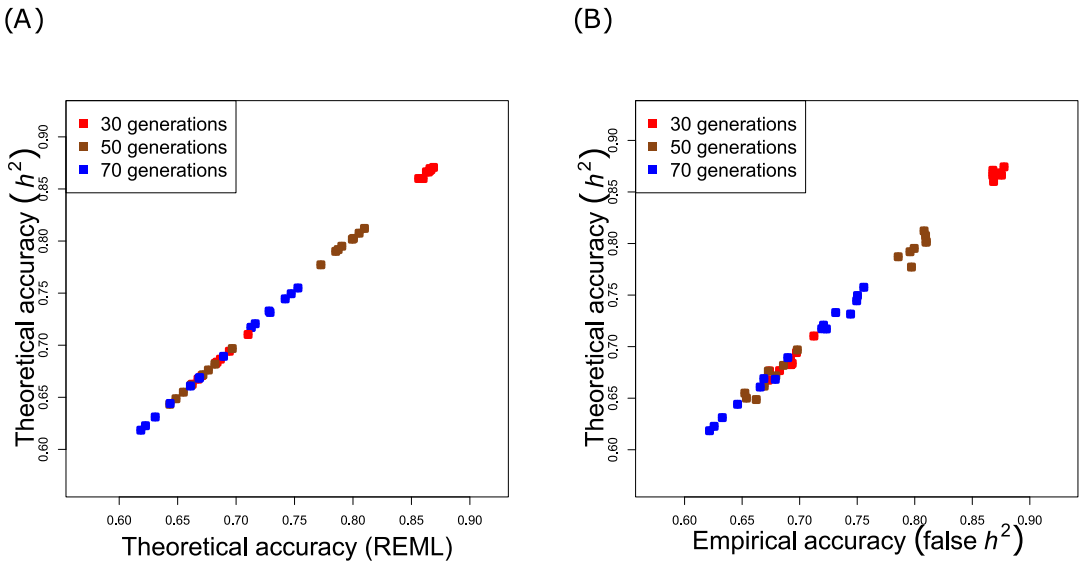

Fig 2. Comparison between the Theoretical accuracies and Empirical accuracies as a function of the tuning parameter λ, and as a function of the number of generations considered. λ was either estimated by REML, or based on either heritability h^2 or on false heritability.

doi:10.1371/journal.pone.0156086.g002

each QTL tends to be in perfect LD with at least one SNP. In contrast, when the density of markers is low, the Theoretical accuracy decreases due to the lack of markers. For instance, according to Fig 3B, when 100 SNPs and 500 TRN individuals were considered, the accuracy decreased respectively from 0.68 to 0.59, from 0.66 to 0.52, or from 0.64 to 0.46, when the population evolved during 30 generations, 50 generations and 70 generations respectively.

Only one TRN incidence matrix. Fig 4 shows analysis of the case where the TRN incidence matrix ***X*** and the TRN causal matrix ***Q*** did not vary across replicates (for a given

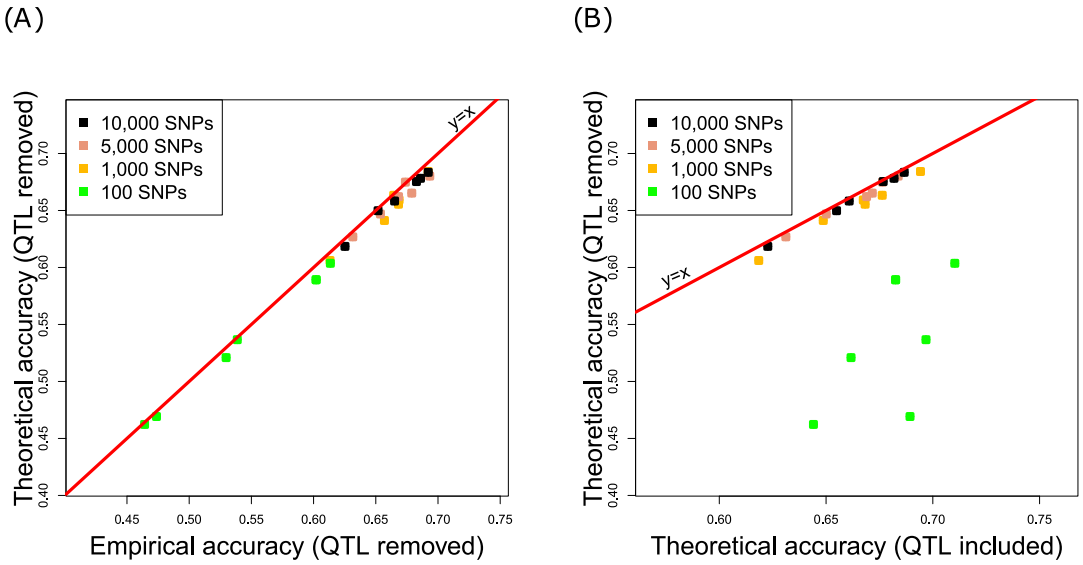

Fig 3. Theoretical accuracies and Empirical accuracies, as a function of the density of markers, and depending on whether the causal SNP was observed. The focus was on the 2 QTL scenario. "QTL removed" means the configuration where the causal SNP was not observed in the TRN and TST populations, whereas "QTL included" means opposite. In all cases, the tuning parameter λ was based on the heritability.

doi:10.1371/journal.pone.0156086.g003

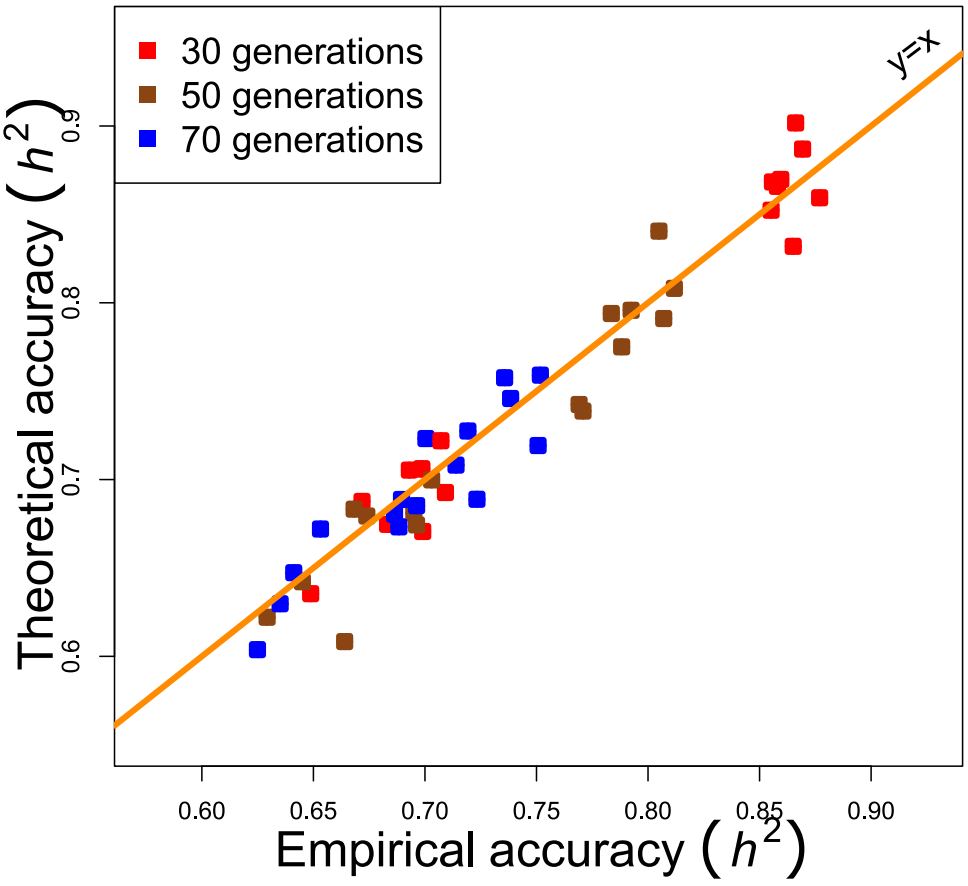

Fig 4. Comparison between the Theoretical accuracy and the Empirical accuracy, as a function of the number of generations. For a given architecture, the TRN incidence matrix did not vary across replicates. The tuning parameter λ was based on the heritability.

doi:10.1371/journal.pone.0156086.g004

architecture). Readers can see that the Theoretical accuracy and the Empirical accuracy were still a good match. It is noticeable, however, that there was more variability than when $\boldsymbol{X}$ and $\boldsymbol{Q}$ varied across replicates.

New proxy vs existing proxies. In order to predict the accuracy in genomic selection, most of the methods are based on the original formula of [32]. They consist of replacing the number of independent QTLs by the effective number of independent loci M_e. In most cases, M_e is computed according to assumptions of population genetics [34, 35]. Note that M_e can also be computed by inferring the number of independent tests in association mapping studies [36].

In this context, Fig 5 shows a comparison of performance of five different proxies in terms of the accuracy. Three of these proxies, the ones based on M_{e1}, M_{e2}, and M_{e3}, rely on the effective population size, whereas the fourth, an M_{LJ}-based proxy, comes from association studies. The fifth proxy is the one suggested in this paper (formula 6). In Fig 5A, the TRN incidence matrix $\boldsymbol{X}$ varies across replicates, whereas it is fixed in Fig 5B. In both cases, we can notice that the proxy based on M_{LJ} underestimated the Empirical accuracy. Table 4 shows the mean squared errors (MSE) corresponding to each method. As expected, the Theoretical accuracy yielded the best performances. Recall that it cannot be computed in practice because it depends on unknown quantities: the QTL effects and their locations. Furthermore, our new proxy outperformed the existing proxies. In comparison with the MSE corresponding to the best proxy

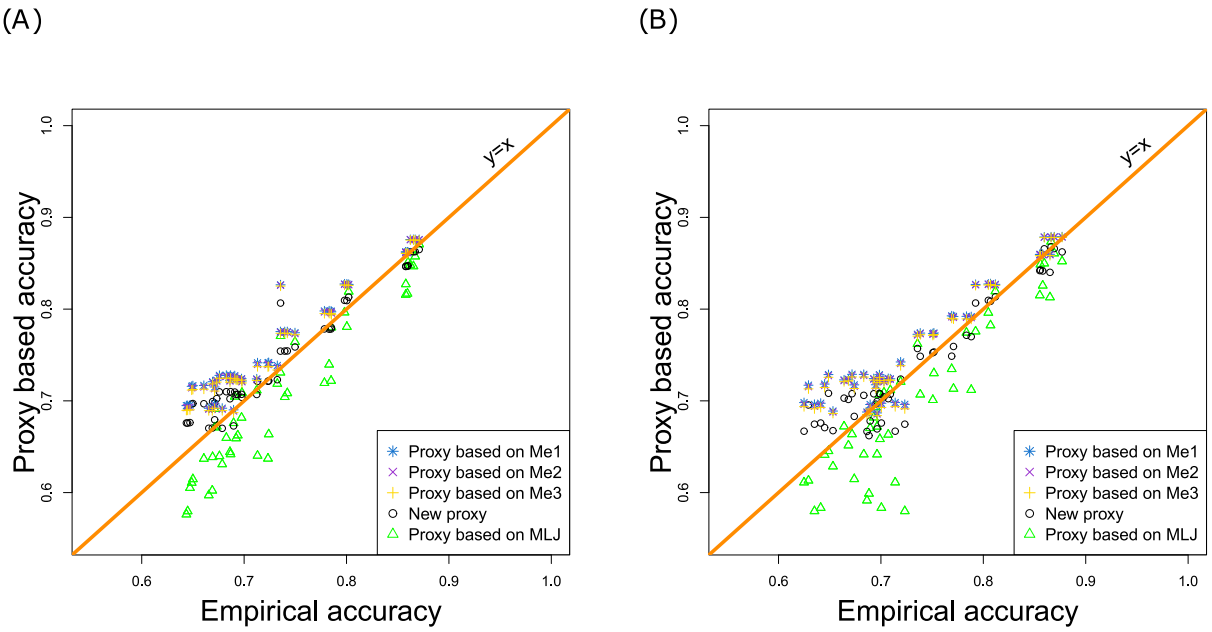

Fig 5. Performances of 5 proxies as a function of the Empirical accuracy. (A) For a given architecture, the TRN incidence matrix varied across replicates. (B) For a given architecture, the TRN incidence matrix did not vary.

doi:10.1371/journal.pone.0156086.g005

(the M_{e3}-based one), our proxy MSE were 2.3- and 1.8-fold smaller, respectively when $\boldsymbol{X}$ varied and when $\boldsymbol{X}$ did not vary across replicates. Note also that for each method, the MSE was smaller when $\boldsymbol{X}$ varied than when $\boldsymbol{X}$ was fixed (Fig 5).

Comparison between the effective number of segments and the quantity $n_{\mathrm{TRN}}\mathbb{E}(\| \boldsymbol{x}'_{\mathrm{TRN}+1}\boldsymbol{X}'\boldsymbol{V}^{-1} \|^2)$. According to our theoretical analysis, we should substitute the quantity $n_{\mathrm{TRN}}\mathbb{E}(\|\boldsymbol{x}'_{n_{\mathrm{TRN}}+1}\boldsymbol{X}'\boldsymbol{V}^{-1}\|^2)$ into [32]'s formula, instead of the number of independent loci, which is usually computed. Table 5 shows that $n_{\mathrm{TRN}}\mathbb{E}(\|\boldsymbol{x}'_{n_{\mathrm{TRN}}+1}\boldsymbol{X}'\boldsymbol{V}^{-1}\|^2)$ and M_{e1}, M_{e2}, M_{e3} and M_{LJ} are completely different quantities. Note that we focused on the configuration $n_{\mathrm{TRN}} = 500$ (the case $n_{\mathrm{TRN}} = 1{,}000$ is shown in S1 Table). In particular, we can see that $n_{\mathrm{TRN}}\mathbb{E}(\|\boldsymbol{x}'_{n_{\mathrm{TRN}}+1}\boldsymbol{X}'\boldsymbol{V}^{-1}\|^2)$ varied with the number of QTLs considered; this was not the case for other quantities. Table 6 shows analysis of the case where the TST population evolved during 30, 40, or 70 generations, where we kept a TRN population that evolved during 30 generations.

Table 4. Mean squared error (with respect to the Empirical accuracy) corresponding to 5 proxies. The MSE corresponding to the Theoretical accuracy is also shown (λ is based on the heritability). MSE $= \sum_{a=1}^{48}(\mathrm{AccP}_a - \mathrm{AccE}_a)^2/48$ where 48 is the number of studied architectures. AccE_a and AccP_a are averages on 100 replicates, and denote respectively, for architecture a, the Empirical Accuracy and the Accuracy based on the chosen proxy.

	Fixed TRN matrix	TRN matrix varied
Theoretical accuracy	3.6710×10^{-4}	4.204×10^{-5}
Our proxy	5.9858×10^{-4}	4.628×10^{-4}
Proxy based on M_{e1}	1.2643×10^{-3}	1.228×10^{-3}
Proxy based on M_{e2}	1.207×10^{-3}	1.157×10^{-3}
Proxy based on M_{e3}	1.1335×10^{-3}	1.0669×10^{-3}
Proxy based on M_{LJ}	2.1906×10^{-3}	1.474×10^{-3}

doi:10.1371/journal.pone.0156086.t004

Table 5. Comparison among different estimators (M_{e1}, M_{e2}, M_{e3} and M_{LJ}) of the number of effective loci and the quantity $n_{TRN}\mathbb{E}(\|x'_{n_{TRN}+1}X'V^{-1}\|^2)$. For a given architecture, a mean was computed on 100 replicates (variance is shown in brackets) and the TRN incidence matrix did not vary across replicates (n_{TRN} = 500, λ is based on the heritability).

Nb QTLs	Nb generations	Nb Markers	$n_{TRN}\mathbb{E}(\|x'_{n_{TRN}+1}X'V^{-1}\|^2)$	M_{LJ}	M_{e1}	M_{e2}	M_{e3}
2	30	100	47.16 (0.75)	50.13 (0.67)	11.96	14.02	16.92
		5,000	46.98 (0.58)	177.45 (4.72)	11.76	13.78	16.66
	50	100	53.63 (1.39)	59.73 (0.56)	17.71	20.43	24.12
		5,000	55.32 (0.61)	233.49 (7.36)	18.52	21.32	25.12
	70	100	51.26 (0.76)	64.39 (0.34)	22.68	25.93	30.27
		5,000	56.56 (0.67)	258.63 (7.89)	22.46	25.70	30.02
100	30	100	71.85 (2.40)	50.13 (0.70)	11.96	14.02	16.92
		5,000	74.83 (1.93)	177.45 (4.71)	11.76	13.78	16.66
	50	100	66.94 (2.31)	59.73 (0.50)	17.71	20.43	24.12
		5,000	70.66 (1.18)	233.49 (7.36)	18.52	21.32	25.12
	70	100	58.49 (1.06)	64.39 (0.34)	22.68	25.93	30.27
		5,000	66.68 (1.04)	258.63 (7.89)	22.46	25.70	30.02

doi:10.1371/journal.pone.0156086.t005

This scenario is particularly realistic in plants, where a large number of generations can be obtained easily, and typically, the prediction model is not refitted with time. According to the table, for a given number of markers, the quantity $n_{TRN}\mathbb{E}(\|x'_{n_{TRN}+1}X'V^{-1}\|^2)$ increased with the number of generations in the TST populations. In contrast, the usual quantities M_{e1}, M_{e2}, and M_{e3} could not capture the changes regarding the TST population because they depend only on the TRN population.

Real data

Accuracy. Fig 6A shows the Empirical accuracy estimated by means of the perennial ryegrass dataset, as a function of the TRN/TST samples under study. Readers will recall that 90% of the individuals were chosen randomly for the TRN set, and that the remaining 10% were considered TST individuals. According to the graph, there are large fluctuations between the different samples; this result points to the importance of a good match between TRN and TST sets (see [19, 38] regarding maize data, and [39] regarding simulated data).

Fig 6B illustrates results from 5-fold cross-validation. Readers can see that there is less variablity between the empirical accuracies than when the TRN incidence matrix is fixed for a given sample (see Fig 6A). This is in agreement with conclusions based on our simulation

Table 6. Comparison among different estimators (M_{e1}, M_{e2}, M_{e3} and M_{LJ}) of the number of effective loci and the quantity $n_{TRN}\mathbb{E}(\|x'_{n_{TRN}+1}X'V^{-1}\|^2)$ as a function of the number of generations during which the TST population evolved (TRN population is always based on 30 generations). For a given architecture, a mean was computed on 100 replicates (variance is shown in brackets) and the TRN incidence matrix did not vary across replicates (n_{TRN} = 500, and λ is based on the heritability).

Nb Markers	Nb generations for TST	$n_{TRN}\mathbb{E}(\|x'_{n_{TRN}+1}X'V^{-1}\|^2)$	M_{LJ}	M_{e1}	M_{e2}	M_{e3}
100	30	47.16 (0.75)	50.13 (0.67)	11.96	14.02	16.92
	40	52.17 (0.93)	52.55 (0.84)	11.96	14.02	16.92
	70	58.17 (0.89)	55.95 (0.63)	11.96	14.02	16.92
5,000	30	46.98 (0.58)	177.45 (4.72)	11.76	13.78	16.66
	40	51.46 (0.60)	197.62 (5.21)	11.76	13.78	16.66
	70	52.86 (0.45)	229.45 (9.56)	11.76	13.78	16.66

doi:10.1371/journal.pone.0156086.t006

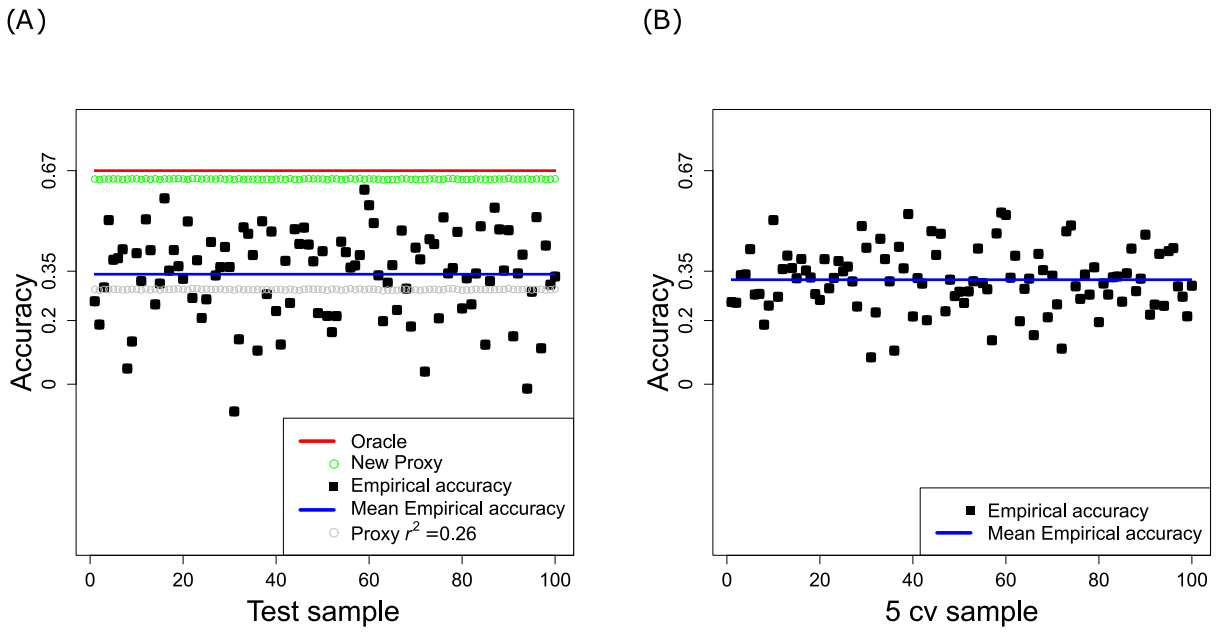

Fig 6. Empirical accuracy and proxies obtained for the perennial ryegrass dataset. (A) 90% TRN and 10% TST. (B) 5 fold cross-validation.

doi:10.1371/journal.pone.0156086.g006

study because the cross-validation process can be viewed as an "unfixed" TRN matrix case, as opposed to the 90% TRN/10% TST process, which is a case of a fixed TRN matrix. As expected, the mean accuracy on all those samples was fairly similar in both analyses: it was estimated to be 0.35 for the 90% TRN/10% TST configuration, and 0.33 for the 5-fold cross-validation.

We computed the oracle accuracy (h), that is to say, the accuracy that would be achieved if the QTL locations and the QTL effects were known. The square root of the heritability was estimated to be 0.67 ($\hat{h}^2 = 0.45$), because of the observed clones. This parameter was later evaluated for every TST sample, and only slight changes were observed (data not shown).

Next, we assessed perfomance of our new proxy. Although the proxy was reestimated for each TRN/TST configuration, we did not notice any fluctuations among the samples. The proxy should be regarded only as an upper bound: it is the optimal accuracy that can be achieved if the QTLs are in perfect LD with markers.

Because we suspected that the marker density was insufficient to cover the entire perennial ryegrass genome, our proxy was later adapted to the case of imperfect LD. In particular, we considered the work of [55], which is a generalization of [32]. Those authors assumed a constant LD r^2 between each QTL and its associated marker, and they assumed the independence of each marker-QTL pair. In this context, our proxy can be rewritten as follows:

$$\rho_{\mathrm{pLD}} = r^2\, h \sqrt{\frac{h^2/(1-h^2)}{\mathbb{E}\left(\parallel \boldsymbol{x}'_{n_{\mathrm{TRN}}+1} X' \boldsymbol{V}^{-1} \parallel^2\right) + r^2 \frac{h^2}{1-h^2}}}. \qquad (8)$$

The r^2 was estimated with the Pearson coefficient of correlation between pairs of SNPs belonging to the same scaffold. Due to the GBS method and the high level of polymorphism in perennial ryegrass (1 SNP/15 bp, see [64]), the distance between successive SNPs within a scaffold was not constant. Many SNPs were grouped within 150 bp because this value corresponded to the read length. The average distance between pairs of SNPs within the scaffolds separated by at

least 150 bp was estimated to be 25,000 bp. In other words, blocks of several SNPs were separated on average by 25,000 bp. Assuming that these blocks are randomly distributed in the genome, the maximal distance between a QTL and a marker (SNP) is approximately 12,500 bp (25,000/2). As a consequence, the average r^2 between pairs of SNPs separated by less than 12,500 bp, was computed: it was estimated to be 0.26 from the values of r^2 as a function of genomic distance in base pairs (see S1 Fig). This value was later substituted into formula (8). Note that the method we used for calculating the quantity r^2 is largely inspired by the work of [55].

According to Fig 6A, there is now a fairly good agreement between the proxy adapted for imperfect LD and the mean accuracy. This finding confirms our orginal supposition: the lack of markers must be responsible for the difference between our upper bound and the observed Empirical accuracy.

Discussion

General aspects

We present here a theoretical formula for the accuracy of GS. The theoretical advances were possible because we analyzed a causal model different from the prediction model (so-called marker-based model); this is usually not the case for investigators working on the mixed models (e.g. [34, 40]). Due to the recent progress in molecular biology, more and more genetic markers are becoming available, and it seems reasonable to assume that there are fewer QTLs than genetic markers in the genome. Recently, [53] incorporated this idea into the mixed-model framework. Nevertheless, those authors had to make approximations in order to obtain analytical formulas. In particular, those authors assumed that the TRN genomic relationships at causal loci were known, and then proposed to perform the regression of genomic relationships based on markers, on those based on causal loci. This idea was motivated by [65]. Although this concept seems interesting, it is not easy to implement and remains an empirical approach. In contrast, our proposed theoretical formula was derived rigorously, without approximations. The marker-based model chosen for our study, is the one corresponding to RRBLUP, also known as ridge regression. In other words, we considered the same high-dimensional prediction model and the same sparse causal model as those addressed in some recent statistical studies [66, 67].

With the help of our general formula (5), we are now able to quantify the influence of various parameters on the accuracy. The theoretical result depends on the QTL effects, QTL locations, TRN causal matrix, TRN incidence matrix, TST causal matrix and the TST incidence matrix. Although [68] highlighted the difficulty of decoding GBLUP, the final result is somewhat more complicated than what we expected and the results in the literature. For instance, according to our study, the average LD between markers is not proportional to the accuracy, contrarely to the results of [34]. In particular, we show that it is the quantity $\boldsymbol{X}' \boldsymbol{V}^{-1} \boldsymbol{Q}$ that has an impact (that is to say, the LD between markers and QTLs in the TRN population with respect to the metric $\boldsymbol{V}^{-1}$). This weighted LD can be viewed as an extension of the work of [69] where the authors introduced new LD measures corrected for population structure and relatedness. Moreover, according to our formula, the covariance between SNPs in the TST population affects the accuracy.

Our present study can be viewed as an answer to the analysis of [37], where the authors raised important questions regarding accuracy in GS. They compared 145 accuracy values extracted from 13 articles, either based on simulated data or real data. An analysis of variance model was fitted to the data, in order to test effects of 4 existing formulas and parameters, on the accuracy. The number of TRN individuals n_{TRN} and the effective number of segments M_e were found to have a strong influence on the accuracy. Besides, a "big formula effect" was

observed, and those authors were unable to demonstrate superiority of one method to the others. One criticism voiced by those authors was that the different formulas in the literature did not take into account the relation between TRN and TST populations. Our theoretical formula now involves explicitly the link between the two populations.

The Theoretical accuracy depends on unknown parameters such as the QTL effects and their locations. It may be useful to first perform an association study on the TRN population in order to identify the QTLs. After that, the detected QTLs could be plugged into our theoretical formula (5), to approximate the accuracy. In the simulation study, since QTLs were perfectly known, such analysis was not considered relevant. Nonetheless, for the perennial ryegrass set, we could have explored this topic.

Most of the existing methods for computation of the accuracy are inspired by the work of [32] and [33]. In [33], the authors proposed to substitute the effective number of independent loci M_e into the original formula of [32]. Then, a large number of research groups elaborated on this concept, and proposed different ways of estimating M_e, using either the effective population size (e.g. [34, 35]), or the number of independent tests [36].

Our theoretical analysis shows that plugging M_e into [32]'s formula is not the way to properly work with the high dimensional framework. We propose to use of another quantity, $n_{\text{TRN}}\mathbb{E}(\|\boldsymbol{x}'_{\boldsymbol{n}_{\text{TRN}}+1}\boldsymbol{X}'\boldsymbol{V}^{-1}\|^2)$. We were able to show on simulated data that our corresponding proxy for the accuracy outperforms existing proxies. In the ryegrass dataset, however, most of the proxies studied yielded similar results because of the lack of markers to cover the entire genome.

An important question in GS is the choice of the TRN population. Various studies have shown sensitivity of the accuracy to relatedness between individuals. [38] focused on a population of maize and demonstrated that predictions are much more reliable when they are performed within families than across different families. Nevertheless, if we are willing to predict breeding values of a somewhat general TST population (not necessarily linked to the TRN population), [39] showed with the help of simulated data, that it is more advantageous to keep large variability in the TRN set.

Similarly, [19] proposed to merge populations belonging to different groups. Similar conclusions are also present in studies on sugar beets [25] and oats [70]. If we assume a limited budget (and as a consequence, a fixed number of TRN individuals), then an interesting finding in our study is the following: by minimizing the quantity $\mathbb{E}(\|\boldsymbol{x}'_{\boldsymbol{n}_{\text{TRN}}+1}\boldsymbol{X}'\boldsymbol{V}^{-1}\|^2)$, we can choose the optimal TRN and TST sets, that maximize the accuracy. This protocol is an alternative to the "CDmean" method proposed by [40] that is based on the coefficient of determination (CD) to optimize the calibration sets. However, CDmean has a shortcoming: it does not differentiate the causal model from prediction model. In that sense, our approach to choosing the TRN and TST sets is expected to be more reliable than the one from [40]. This is a topic for future research. Concerning the computational burden, the Theoretical accuracy requires the inversion of a $n_{\text{TRN}} \times n_{\text{TRN}}$ matrix (complexity of $O(n^3_{\text{TRN}})$).

GS in perennial ryegrass

In forage grasses, traits related to leaf growth (e.g., leaf length, plant height, and leaf elongation rate measured on spaced plants) are correlated with forage productivity measured in a dense canopy (i.e., sward; regarding short-term intake by cows and plant survival in a sward, see [71–73]). In perennial ryegrass, these traits represent a complex genetic architecture: many genes are involved in the overall variability. The corresponding heritability can be either medium (0.4) or high (0.7) [59, 74]. In this case, marker-assisted selection can detect only a small part of the genetic variation and very quickly becomes inefficient after fixation of the largest QTLs.

In contrast, GS, which analyzes all the markers simultaneously, is potentially fruitful. Another interesting characteristic of GS is that it can reduce costs of phenotyping. In perennial ryegrass, the phenotyping of traits related to forage productivity is expensive and requires 2 or 3 years [75]. LD decreases rapidly with r, generally below 0.2 after less than 1kb [64].

In our study, we wanted to evaluate the accuracy of GS in a specific population created from elite material with a narrow genetic base (8 genotypes intercrossed during three generations). In this specific population, LD was expected to decrease slowly due to relatedness. In that case, we thought that a relatively small number of markers would be sufficient to cover the genome if the LD was large enough. According to our study, the mean accuracy is approximately 0.35. This value is similar to the results obtained on rice [27] and wheat [76]. The estimated accuracy was below that of our suggested proxy, which acted as an upper bound of the accuracy. Contrary to our original thoughts, the relatedness does not help to increase the accuracy and to compensate for the lack of markers. Indeed, our theoretical analysis shows that the accuracy depends on a weighted LD, $\boldsymbol{X}^{'}\boldsymbol{V}^{-1}\boldsymbol{Q}$, which can be viewed as LD corrected for the relatedness. On the other hand, GBS can also be considered as limiting factor for the accuracy. Although 20,000 markers were obtained by GBS, the number of independent markers is actually much smaller (approximately 4,500). Recall that a large number of markers is required to capture the genetic variability from many QTLs in the genome of perennial ryegrass. Finally, the accuracy was found to be strongly affected by the configuration of TRN and TST sets: the same phenomenon was observed in other studies (e.g., on oats [70]).

For all these reasons, in order to improve the accuracy of GS in perennial ryegrass, we propose to perform denser genotyping, and to select individuals to be phenotyped by minimizing the quantity $\mathbb{E}(\|\boldsymbol{x}^{'}_{\boldsymbol{n}_{\mathrm{TRN}}+1}\boldsymbol{X}^{'}\boldsymbol{V}^{-1}\|^{2})$, according to our theoretical analysis. Fig 6A) shows that our suggested proxy was not very sensitive to the choice of the TRN set; we expect that an increase in the marker density will introduce some variability and help to choose the optimal TRN set.

Supporting Information

S1 Text. It includes the mathematical proof of the various formulas introduced in the section Materials and Methods.
(PDF)

S2 Text. Explanation of how to estimate N_e using the Hill and Weirr formula [54], and the LD computed between all SNPs.
(PDF)

S1 Table. This table is similar to Table 5 dealing with the case n_{TRN} = 1,000.
(PDF)

S1 Fig. LD, measured with r^2, computed for the perennial ryegrass dataset, and associated with each pair of SNPs within scaffolds.
(PDF)

Acknowledgments

This work was supported by the CROPDL project, which is a part of the INRA Meta-program SELGEN.

The authors thank all the people who participated in the experiment on perennial ryegrass: i) from INRA URP3F: Sabrina Delaunay, Philippe Cormenier, Françoise Durand and Sébastien Blugeon for DNA extraction and GBS libraries; Magali Caillaud, Jean-François Bourcier, Rodrigue Véron, Marylin Vandier, Fabien Surault, Bernard Vignault, and Franck Rondard for

the field experiment; Franck Gelin and Pascal Vernou for plant management in a greenhouse; Nathalie Bonnet and Liliane Jean for administrative work; ii) from INRA EPGV: Aurélie Bérard, Elodie Marquand, and Marie-Christine LePaslier for GBS libraries and sequencing; iii) technicians from the CNS Evry for sequencing; and iv) from the University of Aahrus: Stephan Hentrup, Stephen Byrne and Istvan Nagy for bioinformatics. The authors are grateful to Jérôme Garon, Sandrine Flajoulot, and Vincent Béguier from Gie GRASS who provided the plant material.

We thank anonymous reviewers for their very constructive comments.

Author Contributions

Conceived and designed the experiments: BM GC PB. Performed the experiments: CER PB. Analyzed the data: CER BM PB. Wrote the paper: CER BM PB. Contributed to genotyping data analysis: TA.

References

1. Lander ES, Botstein D. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics. 1989; 121(1):185–199. PMID: 2563713
2. Cierco C. Asymptotic distribution of the maximum likelihood ratio test for gene detection. Statistics. 1998; 31(3):261–285. doi: 10.1080/02331889808802639
3. Chen Z, Chen H. On some statistical aspects of the interval mapping for QTL detection. Statistica Sinica. 2005; 15(4):909–925.
4. Azais JM, Delmas C, Rabier CE. Likelihood ratio test process for Quantitative Trait Locus detection. Statistics. 2014; 48(4):787–801. doi: 10.1080/02331888.2012.760093
5. Wu R, Ma C, Casella G. Statistical genetics of quantitative traits: linkage, maps and QTL. Springer Science & Business Media; 2007.
6. Li C, Zhou A, Sang T. Rice domestication by reducing shattering. Science. 2006; 311(5769):1936–1939. doi: 10.1126/science.1123604 PMID: 16527928
7. Frary A, Nesbitt TC, Frary A, Grandillo S, Van Der Knaap E, Cong B, et al. fw2. 2: a quantitative trait locus key to the evolution of tomato fruit size. Science. 2000; 289(5476):85–88. doi: 10.1126/science.289.5476.85 PMID: 10884229
8. Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature genetics. 2006; 38 (2):203–208. doi: 10.1038/ng1702 PMID: 16380716
9. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences. 2009; 106(23):9362–9367. doi: 10.1073/pnas.0903103106
10. Donnelly P. Progress and challenges in genome-wide association studies in humans. Nature. 2008; 456(7223):728–731. doi: 10.1038/nature07631 PMID: 19079049
11. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nature genetics. 2008; 40(8):955–962. doi: 10.1038/NG.175 PMID: 18587394
12. Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, Mangino M, et al. Genome-wide association analysis identifies 20 loci that influence adult height. Nature genetics. 2008; 40(5):575–583. doi: 10.1038/ng.121 PMID: 18391952
13. Lettre G, Jackson AU, Gieger C, Schumacher FR, Berndt SI, Sanna S, et al. Identification of ten loci associated with height highlights new biological pathways in human growth. Nature genetics. 2008; 40 (5):584–591. doi: 10.1038/ng.125 PMID: 18391950
14. Hayes B. QTL mapping, MAS, and genomic selection. Short course organized by Iowa State University. 2007;Available from: http://www.anslab.iastate.edu/class/ABG%20Short%20Course/QTL%20Mapping,%20MAS,%20and%20Genomic%20Selection%20Dr.%20Ben%20Hayes.pdf [cited 06/12/2015].
15. Goddard ME, Hayes BJ. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nature Reviews Genetics. 2009; 10(6):381–391. doi: 10.1038/nrg2575 PMID: 19448663

16. Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, et al. The genetic architecture of maize flowering time. Science. 2009; 325(5941):714–718. doi: 10.1126/science.1174276 PMID: 19661422
17. Maher B. Personal genomes: The case of the missing heritability. Nature News. 2008; 456(7218):18–21. doi: 10.1038/456018a
18. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009; 461(7265):747–753. doi: 10.1038/nature08494 PMID: 19812666
19. Schulz-Streeck T, Ogutu J, Karaman Z, Knaak C, Piepho H. Genomic selection using multiple populations. Crop Science. 2012; 52(6):2453–2461. doi: 10.2135/cropsci2012.03.0160
20. de Los Campos G, Gianola D, Allison DB. Predicting genetic predisposition in humans: the promise of whole-genome markers. Nature Reviews Genetics. 2010; 11(12):880–886. doi: 10.1038/nrg2898 PMID: 21045869
21. de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MP. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics. 2013; 193(2):327–345. doi: 10.1534/genetics.112.143313 PMID: 22745228
22. Hayes B, Bowman P, Chamberlain A, Goddard M. Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of dairy science. 2009; 92(2):433–443. doi: 10.3168/jds.2008-1646 PMID: 19164653
23. Jannink JL, Lorenz AJ, Iwata H. Genomic selection in plant breeding: from theory to practice. Briefings in functional genomics. 2010; 9(2):166–177. doi: 10.1093/bfgp/elq001 PMID: 20156985
24. Kumar S, Chagné D, Bink MC, Volz RK, Whitworth C, Carlisle C. Genomic selection for fruit quality traits in apple (Malus × domestica Borkh.). PloS One. 2012; 7(5):e36674. doi: 10.1371/journal.pone.0036674 PMID: 22574211
25. Würschum T, Reif JC, Kraft T, Janssen G, Zhao Y. Genomic selection in sugar beet breeding populations. BMC genetics. 2013; 14(1):85. doi: 10.1186/1471-2156-14-85 PMID: 24047500
26. Burstin J, Salloignon P, Martinello M, Magnin-Robert JB, Siol M, Jacquin F, et al. Genetic diversity and trait genomic prediction in a pea diversity panel. BMC genomics. 2015; 16(1):105. doi: 10.1186/s12864-015-1266-1 PMID: 25765216
27. Spindel J, Begum H, Akdemir D, Virk P, Collard B, Redoña E, et al. Genomic Selection and Association Mapping in rice (Oryza sativa): Effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLoS Genetics. 2015; 11(2):e1004982. doi: 10.1371/journal.pgen.1004982 PMID: 25689273
28. Li Z, Sillanpää MJ. Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. Theoretical and Applied Genetics. 2012; 125(3):419–435. doi: 10.1007/s00122-012-1892-9 PMID: 22622521
29. Kärkkäinen HP, Sillanpää MJ. Back to basics for Bayesian model building in genomic selection. Genetics. 2012; 191(3):969–987. doi: 10.1534/genetics.112.139014 PMID: 22554888
30. Gianola D, Fernando RL, Stella A. Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics. 2006; 173(3):1761–1776. doi: 10.1534/genetics.105.049510 PMID: 16648593
31. de los Campos G, Gianola D, Rosa GJ, Weigel KA, Crossa J. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genetics Research. 2010; 92(04):295–308. doi: 10.1017/S0016672310000285 PMID: 20943010
32. Daetwyler HD, Villanueva B, Woolliams JA. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS One. 2008; 3(10):e3395. doi: 10.1371/journal.pone.0003395 PMID: 18852893
33. Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA. The impact of genetic architecture on genome-wide evaluation methods. Genetics. 2010; 185(3):1021–1031. doi: 10.1534/genetics.110.116855 PMID: 20407128
34. Goddard M, Hayes B, Meuwissen T. Using the genomic relationship matrix to predict the accuracy of genomic selection. Journal of Animal Breeding and Genetics. 2011; 128(6):409–421. doi: 10.1111/j.1439-0388.2011.00964.x PMID: 22059574
35. Goddard M. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica. 2009; 136(2):245–257. doi: 10.1007/s10709-008-9308-0 PMID: 18704696
36. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity. 2005; 95(3):221–227. doi: 10.1038/sj.hdy.6800717 PMID: 16077740
37. Brard S, Ricard A. Is the use of formulae a reliable way to predict the accuracy of genomic selection? Journal of Animal Breeding and Genetics. 2015; 132(3):207–217. doi: 10.1111/jbg.12123 PMID: 25377121

38. Albrecht T, Wimmer V, Auinger HJ, Erbe M, Knaak C, Ouzunova M, et al. Genome-based prediction of testcross values in maize. Theoretical and Applied Genetics. 2011; 123(2):339–350. doi: 10.1007/s00122-011-1587-7 PMID: 21505832

39. Pszczola M, Strabel T, Mulder H, Calus M. Reliability of direct genomic values for animals with different relationships within and to the reference population. Journal of dairy science. 2012; 95(1):389–400. doi: 10.3168/jds.2011-4338 PMID: 22192218

40. Rincent R, Laloë D, Nicolas S, Altmann T, Brunel D, Revilla P, et al. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (Zea mays L.). Genetics. 2012; 192(2):715–728. doi: 10.1534/genetics.112.141473 PMID: 22865733

41. Habier D, Fernando R, Dekkers J. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 2007; 177(4):2389–2397. doi: 10.1534/genetics.107.081190 PMID: 18073436

42. Pérez-Enciso M, Rincón JC, Legarra A. Sequence-vs. chip-assisted genomic selection: accurate biological information is advised. Genetics Selection Evolution. 2015; 47(1):43. doi: 10.1186/s12711-015-0117-5

43. Boichard D, Guillaume F, Baur A, Croiseau P, Rossignol MN, Boscher MY, et al. Genomic selection in French dairy cattle. Animal Production Science. 2012; 52(3):115–120. doi: 10.1071/AN11119

44. Meuwissen T, Hayes B, Goddard M. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001; 157(4):1819–1829. PMID: 11290733

45. Gianola D, Weigel KA, Krämer N, Stella A, Schön CC. Enhancing genome-enabled prediction by bagging genomic BLUP. PLoS One. 2014; 9(4):e91693. doi: 10.1371/journal.pone.0091693 PMID: 24722227

46. Visscher PM, Yang J, Goddard ME. A commentary on common SNPs explain a large proportion of the heritability for human height by Yang et al. (2010). Twin Research and Human Genetics. 2010; 13 (06):517–524. doi: 10.1375/twin.13.6.517 PMID: 21142928

47. Endelman JB. Ridge regression and other kernels for genomic selection with R package rrBLUP. The Plant Genome. 2011; 4(3):250–255. doi: 10.3835/plantgenome2011.08.0024

48. Falconer DS. Introduction to quantitative genetics. DS Falconer; 1960.

49. Crow JF, Kimura M. An introduction to population genetics theory. New York, Evanston and London: Harper & Row, Publishers; 1970.

50. Corbeil RR, Searle SR. Restricted maximum likelihood (REML) estimation of variance components in the mixed model. Technometrics. 1976; 18(1):31–38. doi: 10.2307/1267913

51. Technow, F. R Package hypred: Simulation of Genomic Data in Applied Genetics. 2014;Available from: http://cran.r-project.org/src/contrib/Archive/hypred/ [cited 06/12/2015].

52. Haldane J. The combination of linkage values and the calculation of distances between the loci of linked factors. J Genet. 1919; 8(29):299–309.

53. de los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D. Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genetics. 2013; 9(7):1–15. doi: 10.1371/journal.pgen.1003608

54. Hill W, Weir B. Variances and covariances of squared linkage disequilibria in finite populations. Theoretical population biology. 1988; 33(1):54–78. doi: 10.1016/0040-5809(88)90004-4 PMID: 3376052

55. Lian L, Jacobson A, Zhong S, Bernardo R. Genomewide prediction accuracy within 969 maize biparental populations. Crop Science. 2014; 54(4):1514–1522. doi: 10.2135/cropsci2013.12.0856

56. Šmarda P, Bureš P, Horová L, Foggi B, Rossi G. Genome size and GC content evolution of Festuca: ancestral expansion and subsequent reduction. Annals of botany. 2008; 101(3):421–433. doi: 10.1093/aob/mcm307 PMID: 18158307

57. Dolezel J, Bartos J, Voglmayr H, Greilhuber J. Nuclear DNA content and genome size of trout and human. Cytometry Part A. 2003;(51):127–8.

58. Byrne S, Czaban A, Studer B, Panitz F, Bendixen C, Asp T. Genome wide allele frequency fingerprints (GWAFFs) of populations via genotyping by sequencing. PloS One. 2013; 8(3):e57438. doi: 10.1371/journal.pone.0057438 PMID: 23469194

59. Pauly L, Flajoulot S, Garon J, Julier B, Béguier V, Barre P. Detection of favorable alleles for plant height and crown rust tolerance in three connected populations of perennial ryegrass (Lolium perenne L.). Theoretical and Applied Genetics. 2012; 124(6):1139–1153. doi: 10.1007/s00122-011-1775-5 PMID: 22234605

60. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009; 25(14):1754–1760. doi: 10.1093/bioinformatics/btp324 PMID: 19451168

61. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 2010; 20(9):1297–1303. doi: 10.1101/gr.107524.110 PMID: 20644199

62. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics. 2007; 81(3):559–575. doi: 10.1086/519795 PMID: 17701901

63. Hayes B, Goddard M. Genome-wide association and genomic selection in animal breeding. Genome. 2010; 53(11):876–883. PMID: 21076503

64. Auzanneau J, Huyghe C, Julier B, Barre P. Linkage disequilibrium in synthetic varieties of perennial ryegrass. Theoretical and Applied Genetics. 2007; 115(6):837–847. doi: 10.1007/s00122-007-0612-3 PMID: 17701396

65. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nature genetics. 2010; 42(7):565–569. doi: 10.1038/ng.608 PMID: 20562875

66. Shao J, Deng X. Estimation in high-dimensional linear models with deterministic design matrices. The Annals of Statistics. 2012; 40(2):812–831. doi: 10.1214/12-AOS982

67. Bühlmann P. Statistical significance in high-dimensional linear models. Bernoulli. 2013; 19(4):1212–1242. doi: 10.3150/12-BEJSP11

68. Habier D, Fernando RL, Garrick DJ. Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics. 2013; 194(3):597–607. doi: 10.1534/genetics.113.152207 PMID: 23640517

69. Mangin B, Siberchicot A, Nicolas S, Doligez A, This P, Cierco-Ayrolles C. Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. Heredity. 2011; 108 (3):285–291. doi: 10.1038/hdy.2011.73 PMID: 21878986

70. Asoro FG, Newell MA, Beavis WD, Scott MP, Jannink JL. Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. The Plant Genome. 2011; 4 (2):132–144. doi: 10.3835/plantgenome2011.02.0007

71. Barre P, Emile JC, Betin M, Surault F, Ghesquière M, Hazard L. Morphological characteristics of perennial ryegrass leaves that influence short-term intake in dairy cows. Agronomy journal. 2006; 98(4):978–985. doi: 10.2134/agronj2005.0213

72. Hazard L, Ghesquière M. Evidence from the use of isozyme markers of competition in swards between short-leaved and long-leaved perennial ryegrass. Grass and Forage Science. 1995; 50(3):241–248. doi: 10.1111/j.1365-2494.1995.tb02319.x

73. Horst G, Nelson C, Asay K. Relationship of leaf elongation to forage yield of tall fescue genotype. Crop Science. 1978; 18(5):715–719. doi: 10.2135/cropsci1978.0011183X001800050005x

74. Shinozuka H, Cogan NO, Spangenberg GC, Forster JW. Quantitative Trait Locus (QTL) meta-analysis and comparative genomics for candidate gene prediction in perennial ryegrass (Lolium perenne L.). BMC genetics. 2012; 13(1):101. doi: 10.1186/1471-2156-13-101 PMID: 23137269

75. Hayes BJ, Cogan NO, Pembleton LW, Goddard ME, Wang J, Spangenberg GC, et al. Prospects for genomic selection in forage plant species. Plant Breeding. 2013; 132(2):133–143. doi: 10.1111/pbr.12037

76. Zhao Y, Mette M, Gowda M, Longin C, Reif J. Bridging the gap between marker-assisted and genomic selection of heading time and plant height in hybrid wheat. Heredity. 2014; 112(6):638–645. doi: 10.1038/hdy.2014.1 PMID: 24518889