RESEARCH ARTICLE

Genome-Assisted Prediction of Quantitative Traits Using the R Package *sommer*

Giovanny Covarrubias-Pazaran*

Department of Horticulture, University of Wisconsin, Madison, Wisconsin, Unites States of America

* covarrubiasp@wisc.edu

Citation: Covarrubias-Pazaran G (2016) Genome-Assisted Prediction of Quantitative Traits Using the R Package *sommer*. PLoS ONE 11(6): e0156744. doi:10.1371/journal.pone.0156744

Editor: Aimin Zhang, Institute of Genetics and Developmental Biology, CHINA

Received: March 20, 2016

Accepted: May 18, 2016

Published: June 6, 2016

Data Availability Statement: All genotypic and phenotypic information used in this research are freely accessible and can be found in the R package documentation. The maize data can be accessed as data (cornHybrid), and data (wheatLines). The script for all analysis can be found in S1 File.

Funding: CONACYT (Mexico) funded PhD studies of GCP, but the funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The author has declared that no competing interests exist.

Abstract

Most traits of agronomic importance are quantitative in nature, and genetic markers have been used for decades to dissect such traits. Recently, genomic selection has earned attention as next generation sequencing technologies became feasible for major and minor crops. Mixed models have become a key tool for fitting genomic selection models, but most current genomic selection software can only include a single variance component other than the error, making hybrid prediction using additive, dominance and epistatic effects unfeasible for species displaying heterotic effects. Moreover, Likelihood-based software for fitting mixed models with multiple random effects that allows the user to specify the variance-covariance structure of random effects has not been fully exploited. A new open-source R package called *sommer* is presented to facilitate the use of mixed models for genomic selection and hybrid prediction purposes using more than one variance component and allowing specification of covariance structures. The use of *sommer* for genomic prediction is demonstrated through several examples using maize and wheat genotypic and phenotypic data. At its core, the program contains three algorithms for estimating variance components: Average information (AI), Expectation-Maximization (EM) and Efficient Mixed Model Association (EMMA). Kernels for calculating the additive, dominance and epistatic relationship matrices are included, along with other useful functions for genomic analysis. Results from *sommer* were comparable to other software, but the analysis was faster than Bayesian counterparts in the magnitude of hours to days. In addition, ability to deal with missing data, combined with greater flexibility and speed than other REML-based software was achieved by putting together some of the most efficient algorithms to fit models in a gentle environment such as R.

Introduction

With next generation sequencing technologies (NGS) becoming cheaper and consequently more feasible for all crops, huge genomic data sets have become available to help during selection and decision making in plant breeding programs [1,2]. The idea of using genetic markers to accelerate and improve plant and animal breeding systems originated with Sax in 1923 who first reported an association of a simply inherited genetic marker with a quantitative trait in plants [3,4]. On the other hand, the development of statistical tools for breeding purposes,

Abbreviations: AI, average information algorithm; BLUE, best linear unbiased estimator; BLUP, best linear unbiased predictor; EM, expectation maximization; EMMA, efficient mixed model association; GCA, general combining ability; GEBV, genomic-estimated breeding value; GV, genetic value; MAS, marker assisted selection; ML, maximum likelihood; NR, Newton-Raphson algorithm; QTL, quantitative trait loci; RAM, random access memory; REML, restricted maximum likelihood; SCA, specific combining ability.

particularly, Henderson's mixed model equations in 1949 (not published until 1963 with the help of Searle), gave rise in animal breeding to kinship-based selection, breeding value estimation, and prediction of breeding materials [5–7]. More recently, genetic markers have been exploited in plant breeding to detect quantitative trait loci (QTL) for marker assisted selection (MAS). However, after decades of QTL studies, the real application and value of the QTL paradigm in plant breeding has been questioned [8,9].

With the advent of inexpensive and high-throughput genotyping technologies in the last decade, a new plant and animal breeding selection paradigm called genomic selection has emerged [4]. Genomic selection allows the prediction of the phenotypes of individuals based on known marker effects or genetic relationships (kinship-based), and in plants it has been used for predicting trait performance of hybrids and unrealized crosses. One of the first methods proposed for genomic selection was a statistical method called ridge regression (RR), where the ridge parameter (λ) can be observed in a mixed model framework as the σ^2_e / σ^2_u ratio between the residual and random effect variances. This can be applied in the genomic context where σ^2_u is the genetic variance and best linear unbiased predictor (BLUP) can be interpreted as the genomic estimated breeding values (GEBV), where the random effect refers to genotype effects and the variance-covariance structure is the additive or genomic relationship matrix (**A** or $\mathbf{A}_g$). The genetic variance can also be interpreted in terms of marker effects in the form of marker-based BLUPs [10–13].

The use of mixed models to estimate breeding and genetic values can be generalized to more complex scenarios. Mixed models can be used to address general and specific combining abilities in hybrid populations. In particular, they can be used to predict the performance of unrealized crosses, such as single cross hybrids in species which commonly display additive and dominance (heterosis) effects [7,14]. These effects, also called general and specific combining abilities (GCA and SCA, respectively), can be dissected in a mixed model as random effects with a particular variance-covariance structure (**G**) and with distribution:

$$\begin{bmatrix} y \\ u \\ \varepsilon \end{bmatrix} \sim \left(\begin{bmatrix} \boldsymbol{X}\beta \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{ZGZ}' + \boldsymbol{R} & \boldsymbol{ZG} & \boldsymbol{R} \\ \boldsymbol{GZ}' & \boldsymbol{G} & 0 \\ \boldsymbol{R} & 0 & \boldsymbol{R} \end{bmatrix} \right) \text{ being } \boldsymbol{G} = \begin{bmatrix} \boldsymbol{K}_1\sigma^2_{u1} & 0 & 0 & 0 \\ 0 & \boldsymbol{K}_2\sigma^2_{u2} & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \boldsymbol{K}_j\sigma^2_{uj} \end{bmatrix}$$

Here, **X** and **Z** are incidence matrices for fixed and random effects respectively, $\mathbf{R} = \mathbf{I}\sigma^2_e$, $\mathbf{K}_i$ is the variance covariance structure for the ith random effect and σ^2_{ui} is the variance component for the ith random effect. Such covariance structures in a general mixed model are usually unknown, but in genomic selection theory such covariance structures are expressed as relationships among individuals, estimated by an additive, genomic, or other type of relationship matrix [13]. Despite all the molecular and statistical advances that allow genomic selection, there is few open source genomic selection or mixed model software that allows the modeling of several variance components at a time and particularly the modeling of SCA effects by likelihood methods, such as some popular R packages; regress, and EMMREML [13,15–19]. The purpose of this paper is to describe the R package *sommer* (solving mixed model equations in R), an open-source REML-based package that can handle more than one variance component, and at the same time allows for flexible specification of variance-covariance structures of random effects and compare it to popular Bayesian and Likelihood-based software. *Sommer* is especially useful for hybrid prediction of species displaying strong heterotic or specific combining ability effects. The package relies on three algorithms based on maximum likelihood (ML) and restricted maximum likelihood

(REML); efficient mixed model association (EMMA) [20], direct average information (AI) [21,22], and expectation maximization (EM) [23,24]. In addition, *sommer* includes other kernels for calculating additive, dominance and epistatic relationship matrices [25] and perform genome wide association studies (GWAS) (the software can be found and downloaded at https://cran.rstudio.com/web/packages/sommer/ [verified 10 May 2016]). The key features of the package are demonstrated using wheat data (*Triticum aestivum* L.) for genomic prediction of species displaying small or null heterotic effects where only the additive kernel is required (a single random effect), and prediction of single-cross maize hybrids (*Zea mays* L.), which require the use of additive and dominance kernels (multiple random effects), and can be extended to any species displaying heterotic effects.

Materials and Methods

Algorithms

The *sommer* package solves the mixed model equations proposed by Henderson [6], and it has been implemented to work with incidence matrices and known variance covariance matrices for each random effect through the use of the *mmer* function and a ASReml-type version named *mmer2*. If an incidence or a variance-covariance matrix is omitted, the software assumes an identity matrix. Currently, three algorithms for variance component estimation are supported; efficient mixed model association (EMMA) [18], average information (AI) [21,22], and expectation maximization (EM) [23,24]. The EMMA method is useful when only one variance component other than the error variance component (σ^2_e) is estimated [25]. When more than one variance component needs to be estimated, the AI and EM algorithms should be used. The AI algorithm is the default, similar to other commercial software such as ASReml [21].

Genomic breeding value estimation in a wheat population

We performed genomic breeding value estimation (GEBV) and hybrid prediction with wheat data, and the results were compared to other genomic selection and mixed model software, including rrBLUP [13], ASReml [21], regress (used by synbreed as well) [17,18], EMMREML [19], MCMCglmm [15], and BGLR [16]. We used the wheat data contained in the R package BGLR consisting of 599 inbred lines genotyped with 1279 diversity array technology (DArT) markers [16]. Phenotypic data consisted of grain yield (GY) for the 599 lines from the historical CIMMYT's Global Wheat Program evaluated in four mega-environments.

From the 599 wheat lines, 179,101 distinct single crosses can be performed. Kinship-based BLUP prediction for the 599 lines were obtained using rrBLUP (ridge regression), ASReml (average information), regress (Newton-Raphson), EMMREML (modified EMMA), BGLR (using the Reproducing kernel Hilbert space [RKHS] kernel), MCMCglmm (Gibbs sampling) and the three algorithms implemented in *sommer* (AI, EM, and EMMA). Similarity among BLUPs using all software was performed in R and displayed in tables and figures [26]. The genomic estimated breeding values (GEBV) for each of the 599 inbred lines was used to predict the performance of possible crosses as the average among the breeding value of the parental lines. The mixed model fitted has the form:

$$y = \boldsymbol{X}\beta + \boldsymbol{Z}u + \varepsilon$$

with variance:

$$V(y) = V(\boldsymbol{Z}u + \varepsilon) = \mathbf{ZGZ}' + \boldsymbol{R}$$

and the mixed model equations for this model are:

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix}^{-1} \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix} = \begin{bmatrix} \beta \\ u \end{bmatrix}$$

Here, $\mathbf{G} = \mathbf{K}\sigma^2_u$, is the variance covariance matrix of the random effect u, from a multivariate normal distribution u ~ MVN(0, $\mathbf{K}\sigma^2_u$), K being, in the genomics context, the additive or genomic relationship matrix ($\mathbf{A}$ or $\mathbf{A}_g$). **X** and **Z** are incidence matrices for fixed and random effects respectively, and **R** is the matrix for residuals (here $\mathbf{I}\sigma^2_e$). A mixed model with a single variance component other than the error (σ^2_e) can be used to estimate the genetic variance (σ^2_u) along with genotype BLUPs to exploit the genetic relationships between individuals coded in **K** (**A**). The genomic relationship matrix was constructed according to VanRaden where $\mathbf{K} = \mathbf{ZZ}'/2\Sigma p_i(1\text{-}p_i)$ [27]. Genotype BLUPs were calculated and considered equal to the GEBV and these were used to predict the performance of the 179,101 possible crosses as the average of parental genomic breeding values. We fitted this model using the *sommer* package by specifying the incidence and variance-covariance matrices and using the three algorithms implemented (AI, EM, EMMA). In addition, a five-fold cross validation was performed to calculate the predictive correlation for grain yield in the four mega environments available for the wheat data using the *sommer* package. In addition, heritability was estimated as $h^2 = \sigma^2_u / \sigma^2_u + \sigma^2_e$.

Single cross hybrid prediction in corn

Genotypic data was simulated consisting of 511 SNP markers in 40 inbred lines belonging to two heterotic groups (20 in each). Phenotypic data was simulated consisting of grain yield (GY) and plant height (PH) for the 40 parents and 100 out the 400 possible hybrids produced from the single-cross of both heterotic groups allowing for heterosis. Genotypes of the 40 parents were used to estimate the genomic relationship matrices as $\mathbf{K} = \mathbf{ZZ}'/2\Sigma p_i(1\text{-}p_i)$ [27] for both heterotic groups ($\mathbf{K}_1$ and $\mathbf{K}_2$), and the genomic relationship matrix for the 400 possible hybrids was obtained as the Kronecker product of the parental genomic relationship matrices $\mathbf{K}_1 \otimes \mathbf{K}_2$ ($\mathbf{K}_3$). Given that the phenotypic data for the possible crosses was not masked, the hybrids were predicted by estimating the BLUPs for general combining abilities in males and females (GCA_{female}, GCA_{male}) and specific combining abilities (SCA) of crosses along with their variance components (σ^2_{GCA1}, σ^2_{GCA2}, σ^2_{SCA}). The model has the form:

$$y = X\beta + Z_1 u_{GCA1} + Z_2 u_{GCA2} + Z_3 u_{SCA} + \varepsilon$$

The mixed model equations for this model are:

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z_1 & X'R^{-1}Z_2 & X'R^{-1}Z_3 \\ Z_1'R^{-1}X & Z_1'R^{-1}Z_1 + G_1^{-1} & Z_1'R^{-1}Z_2 & Z_1'R^{-1}Z_3 \\ Z_2'R^{-1}X & Z_2'R^{-1}Z_1 & Z_2'R^{-1}Z_2 + G_2^{-1} & Z_2'R^{-1}Z_3 \\ Z_3'R^{-1}X & Z_3'R^{-1}Z_1 & Z_3'R^{-1}Z_2 & Z_3'R^{-1}Z_3 + G_3^{-1} \end{bmatrix}^{-1} \begin{bmatrix} X'R^{-1}y \\ Z_1'R^{-1}y \\ Z_2'R^{-1}y \\ Z_3'R^{-1}y \end{bmatrix} = \begin{bmatrix} \beta \\ u_{GCA1} \\ u_{GCA2} \\ u_{SCA} \end{bmatrix}$$

where β is the vector of fixed effects, u_{GCA1}, u_{GCA2}, u_{SCA} are the BLUPs for GCA_{female}, GCA_{male}, and SCA effects, **X** and **Z**s are incidence matrices for fixed and random effects respectively, **R** is the matrix for residuals (here $\mathbf{I}\sigma^2_e$), and $\mathbf{G}^{-1}_1$, $\mathbf{G}^{-1}_2$, $\mathbf{G}^{-1}_3$ are the inverse of the variance-covariance matrices for random effects. The BLUPs u_{GCA1}, u_{GCA2}, u_{SCA} were used to predict the rest of the single-crosses as the sum of their respective GCA and SCA effects.

We fitted this model using the *sommer* package by specifying the incidence and variance-covariance matrices and using the AI and EM algorithms, given that EMMA method cannot estimate more than one variance component. The model could not be implemented in rrBLUP which is also limited to a single variance component. In the BGLR package the Reproducing kernel Hilbert space [RKHS] kernel was used, in ASReml and MCMCglmm the 'ginverse' argument was used to specify the variance-covariance structures, and in the regress package the multiple random effects model using the **ZKZ**' kernel was fitted. EMMREML uses a similar syntax than *sommer*. Results from other software were compared with *sommer*. In addition, a five-fold cross validation was performed to calculate the prediction accuracy for plant height and grain yield in this population.

In order to show the advantage of fitting a model including dominance (SCA) compared to a pure additive models (GCA) with respect to the prediction ability for species displaying heterotic effects, two additional models were fitted including only GCA effects; 1) both parents having the same variance component and 2) each parent from a different heterotic group having a different variance component:

$$\boldsymbol{G} = \left[\boldsymbol{K}\sigma_u^2\right] \qquad \text{and} \qquad \boldsymbol{G} = \begin{bmatrix} \boldsymbol{K}_1\sigma_{u1}^2 & 0 \\ 0 & \boldsymbol{K}_2\sigma_{u2}^2 \end{bmatrix}$$

Models were compared with respect to their prediction ability after 500 runs of a five-fold cross validation for plant height and grain yield. Models were fitted using *sommer* with the default AI algorithm. In addition, heritability for both trait was estimated as; $h^2 = (\sigma^2_{GCA1} + \sigma^2_{GCA2}) / (\sigma^2_{GCA1} + \sigma^2_{GCA2} + \sigma^2_e)$.

Capabilities with big data sets and comparison with other software

In order to test the capabilities of *sommer* compared with other software, posterior analysis were performed with REML-based counterparts: rrBLUP, regress, ASReml, EMMREML, and Bayesian-based: BGLR (iterations = 13000, burn-in = 2000), and MCMCglmm (iterations = 13000, burn-in = 2000; default parameters). Such comparisons were performed using bigger data sets. We simulated phenotypic and genotypic data for 5000 individuals with 10000 markers for a single trait and single additive kernel, with heritability $h^2 = 0.5$ and GEBVs were estimated. Computing time as a function of the population size (N) for the different ML/REML algorithms found across software packages for a single variance component scenario was recorded. We recorded elapsed times for population sizes from 500 to 5000 in intervals of 500 increments and plotted using R.

The phenotypic and genotypic data available from Technow et al. [28] was used to predict the genetic value (GV) of 10578 possible single cross hybrids from the cross of the Flint by Dent heterotic groups, which included additive and dominance effects (three variance components). The same model was fitted with *sommer* counterparts when possible for time and flexibility comparison purposes. Flint and Dent lines were genotyped with 35,432 SNP markers. Computing time as a function of the population size (N) was recorded for population sizes from 1000 to 8000 in intervals of 1000 increments using *sommer* and other REML-based software with ability to fit multiple random effects.

All genotypic and phenotypic information used in this research is freely accessible and can be found in the R package documentation. The maize data can be accessed as data(cornHybrid), data(wheatLines), and data(Technow_data). The script for all analysis can be found in S1 File.

Results and Discussion

At the core of the *sommer* package is the function 'mmer' which solves the mixed model equations proposed by Henderson [6], and it has been implemented to work directly with incidence and variance covariance matrices for each random effect. The function returns the variance components, the maximized log-likelihood, best linear unbiased estimators (BLUEs) for fixed effects and the BLUP solutions for random effects, along with other information of interest such as residuals, Akaike information criterion (AIC), Bayesian information criterion (BIC), etc.

In addition to the mixed model solver able to fit genome wide association (GWA) models in diploid and polyploidy organisms based on Yu et al. [29] and Rosyara et al. [30], the *sommer* package has been equipped with kernels to estimate additive relationship matrices based on Endelman [13] and VanRaden [27], and dominance and epistatic relationship matrices based on Su et al. [31] and Muñoz et al. [25], calculated respectively as:

$$\boldsymbol{A} = \frac{\boldsymbol{ZZ'}}{2\sum_{j=1}^{m} p_j(1-p_j)}$$

$$\boldsymbol{D} = \frac{\boldsymbol{ZZ'}}{\sum_{j=1}^{m} 2p_jq_j(1-p_jq_j)}$$

$$\begin{aligned} \boldsymbol{E}_{aa} &= \boldsymbol{A}\#\boldsymbol{A} && \text{(additive by additive interactions)} \\ \boldsymbol{E}_{dd} &= \boldsymbol{D}\#\boldsymbol{D} && \text{(dominance by dominance interactions)} \\ \boldsymbol{E}_{ad} &= \boldsymbol{A}\#\boldsymbol{D} && \text{(additive by dominance interactions)} \end{aligned}$$

Denoting # the Hadamard product between matrices, **Z** being the scaled marker matrix, p and q the allelic frequencies for the j^{th} marker (j = 1. . .m). Markers are coded -1, 0, 1 with respect to a reference allele for the null homozygous, heterozygote, and positive homozygote respectively, for the additive relationship matrix **A**. On the other hand, markers are coded as 0 and 1's for homozygotes and heterozygotes respectively for the dominance relationship matrix **D**.

Additional functions to 1) draw genetic maps, 2) convert letter format to numeric data, and 3) design matrices for half diallel designs, have been included in the package as well and are documented within the software.

Genomic breeding value estimation in wheat

Given that the realized genomic relationship matrix enters as a special case of the covariance structure for a random effect in a mixed model, and its incidence matrix represents the genotypes of such relationship matrix, the BLUPs of random effects are equal to the GEBV of the genotypes. In order to show the capabilities of the *sommer* package to predict the progeny performance when crossing lines without heterotic effects, such as in wheat lines, we used the dataset available in the BGLR package consisting in 599 lines which can be crossed hypothetically to produce 179,101 possible hybrids. We estimated the GEBV for the 599 lines for grain yield. Wheat, a self-pollinated crop, is usually bred by developing inbred lines by continuous cycles of selfing. Crosses among such lines usually do not display heterosis. This feature makes the prediction of a cross straight forward as the average among breeding values of the parental lines. On the other hand, predicting a maize cross which is known by being an outcross specie, heterosis needs to be taken into account. Therefore, is not possible to predict crosses by using the average breeding value among parental lines. Instead, independent general combining abilities are assumed for each parent and a specific combining ability is required to predict a maize cross (see [7,14] for a review of prediction in self pollinated and outcross species). In Fig 1 the two methods of kinship-based genomic prediction used in this investigation are shown. The

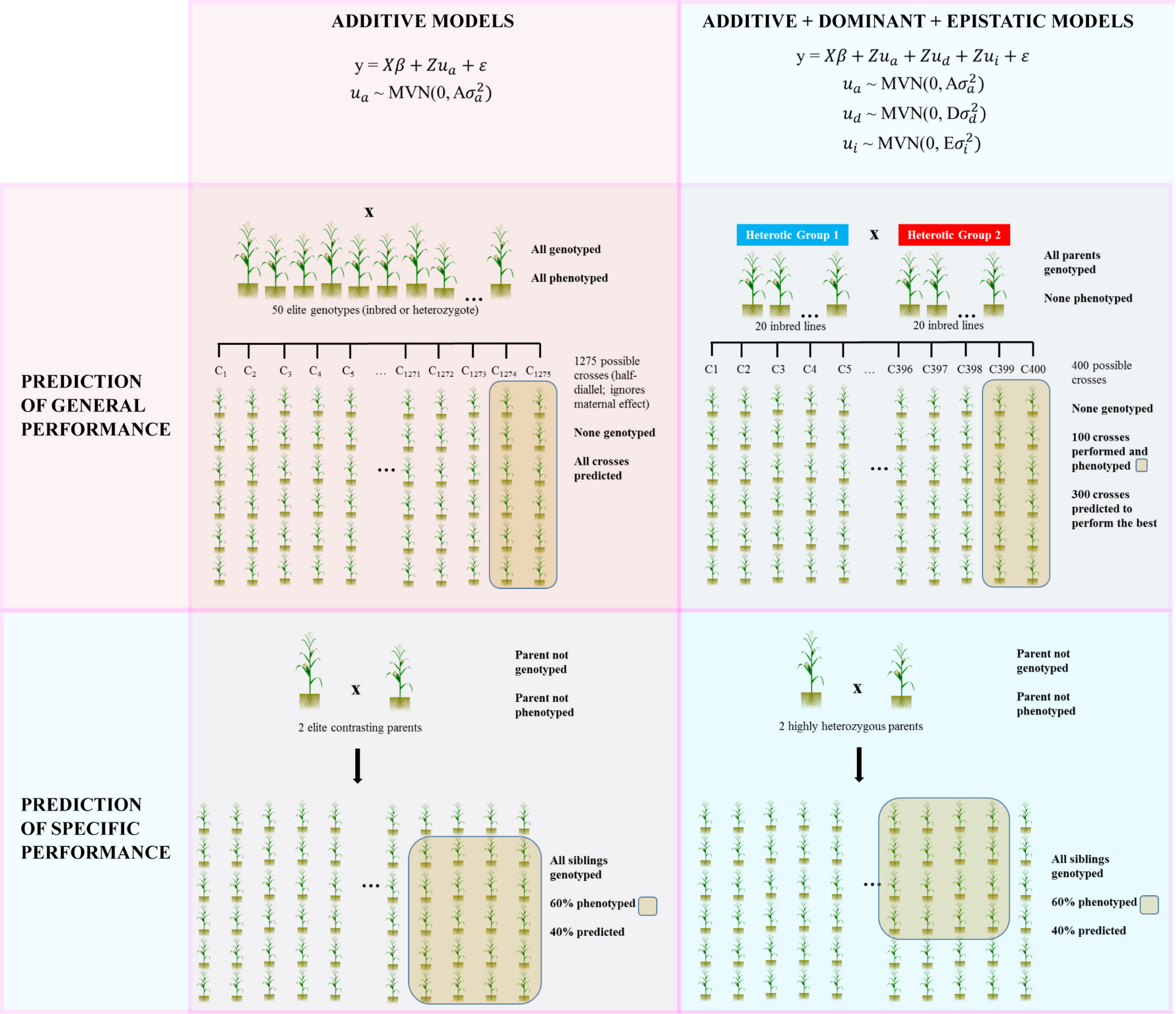
ADDITIVE MODELS
$y = X\beta + Zu_a + \varepsilon$
$u_a \sim \mathrm{MVN}(0, A\sigma_a^2)$
ADDITIVE + DOMINANT + EPISTATIC MODELS
$y = X\beta + Zu_a + Zu_d + Zu_i + \varepsilon$
$u_a \sim \mathrm{MVN}(0, A\sigma_a^2)$
$u_d \sim \mathrm{MVN}(0, D\sigma_d^2)$
$u_i \sim \mathrm{MVN}(0, E\sigma_i^2)$
PREDICTION OF GENERAL PERFORMANCE
x
All genotyped
All phenotyped
50 elite genotypes (inbred or heterozygote)
C_1 C_2 C_3 C_4 C_5 … C_{1271} C_{1272} C_{1273} C_{1274} C_{1275}
1275 possible crosses (half-diallel; ignores maternal effect)
None genotyped
All crosses predicted
Heterotic Group 1
x
Heterotic Group 2
All parents genotyped
None phenotyped
20 inbred lines
20 inbred lines
C1 C2 C3 C4 C5 … C396 C397 C398 C399 C400
400 possible crosses
None genotyped
100 crosses performed and phenotyped
300 crosses predicted to perform the best
PREDICTION OF SPECIFIC PERFORMANCE
x
Parent not genotyped
Parent not phenotyped
2 elite contrasting parents
All siblings genotyped
60% phenotyped
40% predicted
x
Parent not genotyped
Parent not phenotyped
2 highly heterozygous parents
All siblings genotyped
60% phenotyped
40% predicted

Fig 1. Examples of genome-assisted prediction performed using *sommer*. In the first row of the figure, genomic *prediction of general performance (cross performance prediction)* is summarized; in the 1st square, the model to predict crosses with a single additive kernel (wheat example) is depicted. In the 2nd square the prediction including additive, dominance and epistatic kernels is shown (single cross hybrid example). In the second row, genomic *prediction of specific performance (within cross performance prediction)* is shown; in the 1st square, prediction within a biparental cross is shown using a single additive kernel for species displaying small or null heterotic effects whereas in the 2nd square prediction within biparental populations is shown using additive, dominance and epistatic kernels (examples are included in the package).

doi:10.1371/journal.pone.0156744.g001

model is easily fitted by specifying the variance-covariance structure **G** = **K**σ^2_u, with **K** being the additive or genomic relationship matrix (**A** or **A**$_g$) among genotypes.

The BLUPs obtained by *sommer* were compared with rrBLUP [13], ASReml [21], regress [17], EMMREML [19], BGLR [16], and MCMCglmm [15]. We found all statistics such as BLUPs (u), BLUEs (βs), and variance components (σ^2_s) to be equal, showing that all software provided similar and sometimes identical results. Similar results were obtained using all the algorithms implemented in *sommer* (EMMA, EM, AI) to estimate variance components and

Table 1. Cross-validation of prediction accuracies using *sommer* (5-fold) for wheat and maize populations.

	Wheat†					Maize†			
	Env1	Env2	Env3	Env4	h^2	$A_{(1)}$	$A_{(2)}$	$A\text{-}D_{(3)}$	h^2
Accuracy grain yield	0.51±.09	0.48±.10	0.38±.10	0.46±.09	0.21	0.18±0.14	0.21±0.16	0.37±0.16	0.18
Accuracy plant height	‡	‡	‡	‡	‡	0.41±0.12	0.43±0.13	0.68±0.06	0.62

Prediction accuracies for grain yield were obtained for each of the 4 mega environments available for the 599 lines of wheat. Prediction accuracies for grain yield and plant height were obtained for a maize population consisting of 100 hybrids tested in 4 locations using only additive (GCA) effects with a single variance component for both parents [$A_{(1)}$], one variance component for each parent (GCA_1 and GCA_2; $A_{(2)}$), and additive (GCA) and dominance (SCA) effects [$A\text{-}D_{(3)}$].

† Accuracy values are averages over 500 runs of a 5-fold cross validation.

‡ Trait not evaluated in wheat.

doi:10.1371/journal.pone.0156744.t001

other parameters. However, the EM and AI algorithms converged slower than EMMA when only one variance component was estimated due to the iterative procedure used with EM and AI. We found prediction accuracies of 0.51, 0.48, 0.38 and 0.46 for grain yield in the 4 mega-environments, respectively, which was consistent with the expected predictability upper bound expected for the prediction accuracy, which is the square root of heritability, $h^2 = 0.21$ and $\sqrt{h^2} = 0.46$ (Table 1) in the kinship-based prediction.

Single cross hybrid prediction in corn

One of the strengths of the *sommer* package is the ability to specify more than one random effect and their variance-covariance structure, which is usually necessary when genomic selection needs to be performed in species displaying heterotic effects. The model requires the estimation of GCA effects for parents and SCA effects for specific crosses. We used the simulated data for crosses between different heterotic groups that displayed heterotic effects for grain yield. We estimated the genomic relationship matrix ($\mathbf{A}_g$) for both groups of lines ($\mathbf{K}_1$ and $\mathbf{K}_2$) and calculated the Kronecker product of both matrices to obtain the SCA relationship matrix ($\mathbf{K}_3$), in order to predict the other 300 hybrids. We fitted the following model specifying the variance-covariance matrices with *sommer* and compared with ASReml [21], regress [17], EMMREML [19], BGLR [16] and MCMCglmm [15]. In sommer the model can be fitted in the following manner:

```
library(sommer)
data(cornHybrid)
A = cornHybrid$K
y = cornHybrid$Yield
X1 = model.matrix(~ Location, data = cornHybrid)
Z1 = model.matrix(~ GCA1 - 1, data = cornHybrid)
Z2 = model.matrix(~ GCA2 - 1, data = cornHybrid)
Z3 = model.matrix(~ SCA - 1, data = cornHybrid)
K1 = A[levels(cornHybrid$GCA1), levels(cornHybrid$GCA1)]
K2 = A[levels(cornHybrid$GCA2), levels(cornHybrid$GCA2)]
K3 = kronecker(K1, K2)
ETA = list(GCA1 = list(Z = Z1, K = K1), GCA2 = list(Z = Z2, K = K2), SCA = list(Z = Z3, K = K3))
ans = mmer(y = y, Z = ETA, method = AI)
```

or using a data frame

$$ans = mmer2(y \sim 1, random = \sim GCA1 + GCA2 + SCA, G = list(GCA1 = K2, GCA2 = K2, SCA = K3), data = cornHybrid)$$

The BLUPs for u_{GCA1}, u_{GCA2}, and u_{SCA} effects obtained by *sommer* and ASReml were the same (Fig 2) since they use by default the AI algorithm. On the other hand, *sommer* performed faster due to the use of the direct inversion AI algorithm compared to the mixed model equation based AI algorithm found in ASReml, based on the mathematical properties found by Lee et al. [22] when multiple random effects with dense covariance structures are present, and the use of an eigen decomposition on the **G** structure when there is a simple random effect [32] (please refer to the cites for more details on the differences between the two AI algorithms). Slight differences were found when comparing BLUPs from *sommer* with BGLR and MCMCglmm. However, this was expected due to the fact that Bayesian methods are based on Gibbs sampling, and require a high number of iterations to achieve the same parameters than likelihood software. In addition, *sommer*, regress and EMMREML were the fastest taking ~5 seconds to estimate the variance components and perform all calculations, but regress overestimated the error variance and returned some negative variance components, indicating that some constraints have not been implemented yet. EMMREML handles multiple random effects but does not return the variance components other than the first one. On the other hand, BGLR (iterations = 13000, burn-in = 2000) took 2 minutes and 2 seconds (*sommer* was 25 times faster), and MCMCglmm (iterations = 13000, default) took 7 minutes and 13 seconds (*sommer* was 87 times faster), showing that speed is one of the strengths of using REML-based software in dense genetic models.

Given that phenotypic data was masked for the corn hybrids, cross validation was conducted in order to assess the prediction accuracy of hybrids for plant height and grain yield using the hybrid prediction method stated above. The estimated heritability (h^2) for grain yield was 0.18, and 0.62 for plant height in this population. According to selection theory, these values for grain yield set the upper bound for prediction to 0.43 for grain yield and 0.79 for plant height ($\sqrt{h^2}$). We found consistent results by using five-fold cross validations, which resulted in an average of 0.37±0.16 prediction accuracy for grain yield and 0.67±0.061 for plant height (Table 2; Fig 3). As expected, hybrids resulting from different heterotic groups were

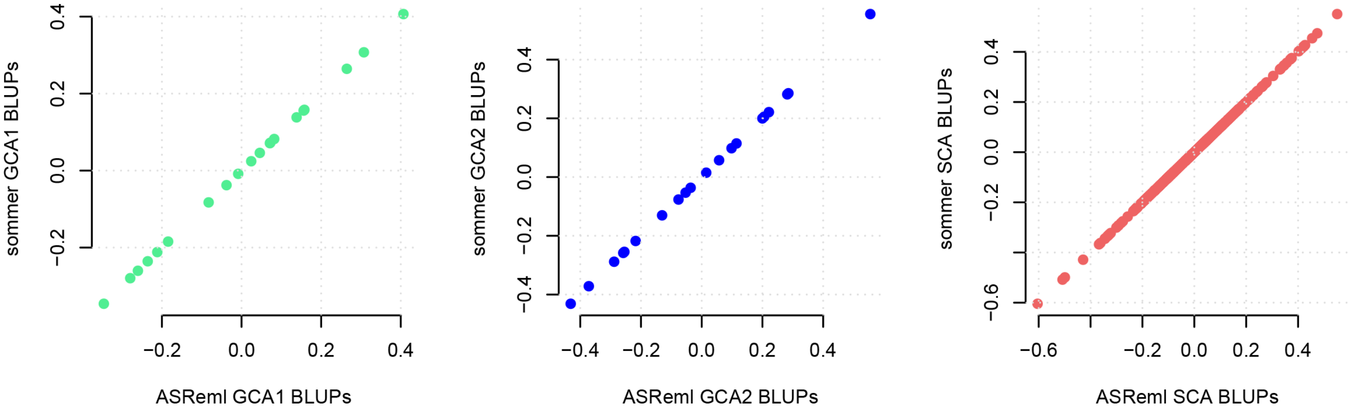

Fig 2. Best linear unbiased prediction (BLUP) comparisons for general and specific combining ability effects (GCA and SCA) using *sommer* versus ASReml. BLUPs for GCA and SCA related to grain yield were computed in a corn population with 400 individuals evaluated in 4 locations using *sommer* and ASReml. *Sommer* estimates are shown on the y axis and are similar to results from the ASReml estimates shown on the x axis.

doi:10.1371/journal.pone.0156744.g002

Table 2. Comparison of *sommer* versus the most common mixed model software available for genomic selection.

Feature	SAS	ASReml	lme4	rrBLUP	MCMC-glmm	BGLR	*sommer*	regress	EMM-REML
Open source			✓	✓	✓	✓	✓	✓	✓
Ability to specify var-cov structures for random effects	✓	✓	†	✓	✓	✓	✓	✓	✓
Estimation of more than one variance component	✓	✓	✓		✓	✓	✓	✓	✓
Basic expertise §	✓	✓	✓	✓	§	§	✓	✓	✓
Platform independent	✓	✓	✓	✓	✓	✓	✓	✓	✓
Ability to specify var-cov structures for residual structures	✓	✓			✓				
Use of sparse methods		✓	✓	‡	‡	‡	✓	‡	‡
Handles missing data	✓	✓	✓	✓	✓	✓	✓	✓	

Advantages and disadvantages of each software with check marks indicating whether they possess the stated feature.

† The *pedigreemm* package which is an extension of *lme4* allows the user to introduce pedigrees, but it does not allow the user to provide the variance-covariance matrices directly. Examples available in the *pedigreemm* package were run using *sommer* obtaining similar results, but *sommer* ran 4 times faster than *pedigreemm*. Examples are included in *sommer* documentation.

‡ Information not available.

§ These packages are based on Bayesian methods requiring the user to have a more advanced statistical background to decide the correct number of iterations, burn-in length and ability to analyze trace plots, and therefore the feature was defined as 'Basic expertise'.

doi:10.1371/journal.pone.0156744.t002

predicted taller and more productive than predicted hybrids resulting from a cross within the same heterotic group.

Additive vs additive-dominance models

We found an increase in prediction ability when models fitted included additive and dominance effects (GCA+SCA) relative to a pure additive model (GCA) under the two assumptions that both parents have the same additive variance or each parent from a different heterotic group has its own additive variance. For grain yield the prediction accuracy changed from ~0.18–0.20 for the purely additive models to 0.52 in a model including GCA and SCA effects, whereas for plant height the prediction accuracy changed from 0.41–0.43 for the additive models to 0.85 for the model including additive and dominance effects (GCA+SCA; Table 2). This highlights the importance of considering dominance effects in addition to the only-additive models. The package provides kernels to calculate additive, dominance, and epistatic

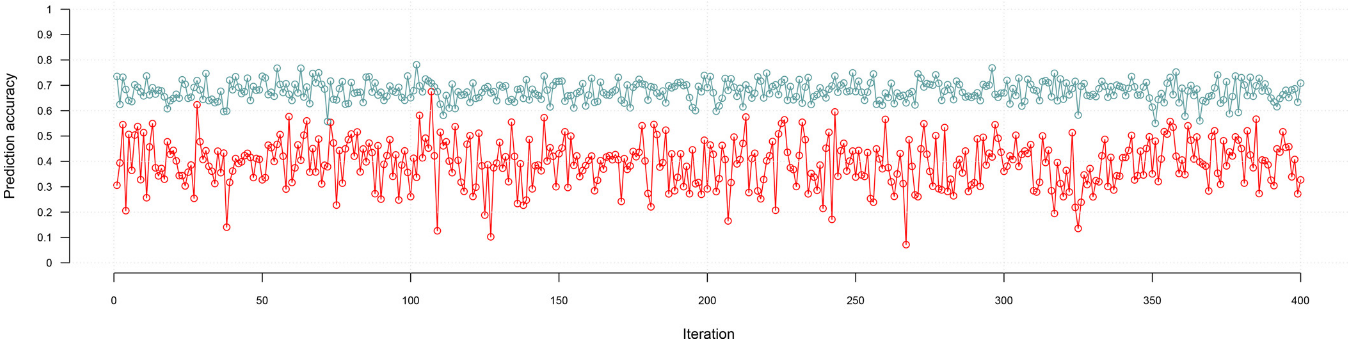

Fig 3. Prediction accuracy results using *sommer* in corn hybrids with a 5-fold cross validation. Cross validation results for plant height in a corn population of 400 individuals evaluated in 4 locations are shown in blue, whereas grain yield cross validation results for the same population is shown in red. Each dot represents the average of one run of a 5-fold cross validation.

doi:10.1371/journal.pone.0156744.g003

Table 3. Time comparison among different software for densest genomic models tested in the study.

No.Var.Comp.	Time	sommer	ASReml §	rrBLUP	regress	BGLR	MCMCglmm	EMMREML
One Var.	User	232.34	438.43	765.27	1679.88	1104.16	529527.09	610.85
Component	System	7.69	3.59	0.94	2.79	181.89	3715.65	0.18
N = 5000	Elapsed	240.04 #	442.73	766.43	1683.94	1291.71	533556.1	611.10
Three Var.	User	352.71	6860.85	‡	1858.99	11712.25	> 1058886 †	1130.92
Components	System	59.7	6.35	‡	3.53	5610.17	> 7431 †	4.89
N = 10585	Elapsed	425.96	6873.60	‡	1868.25	17364.27	> 1067112 †	1136.63 ¶

Time consumption for a GBLUP model with a single variance component (additive) with 5000 individuals and 10000 markers is shown in the first row. A GBLUP model with 3 variance components (additive, dominance, epistasis) with 10,585 hybrids to be predicted genotyped with 35,432 SNPs is displayed in the second row. The two models represent the biggest population sizes used in the study to highlight the differences when big data is encountered.

‡ No more than one variance component other that the error can be estimated.

† Work stopped after 12 days running the model.

§ In both models ASReml returned a warning message of abnormal termination.

¶ Although EMMREML can handle multiple random effects does not return the value for the variance components and cannot handle missing data.

# Using the average information with eigen decomposition proposed by Lee et al. [22].

doi:10.1371/journal.pone.0156744.t003

relationship matrices and examples of their use in the documentation, which have been shown to increase prediction accuracies in certain scenarios and family structures [31].

Mixed model software availability and big data sets

There is limited open-source and user-friendly mixed models software that allow flexible specification of variance-covariance structures for random effects. We compared *sommer* and some of the most popular mixed model software available to highlight the strengths of this new software (Table 2). This software represents a valuable resource for genomic selection and GWAS studies, but it can also be used as any other mixed model software for analysis of non-plant and animal breeding experiments.

Although we have shown genomic selection applications using a kinship-based estimation of GEBV, some models depend on knowing marker effects, such as marker × environment interactions [33–35], which can be implemented as well using the *sommer* package. Half diallel designs and genome wide association studies (GWAS), and general mixed model analysis, can

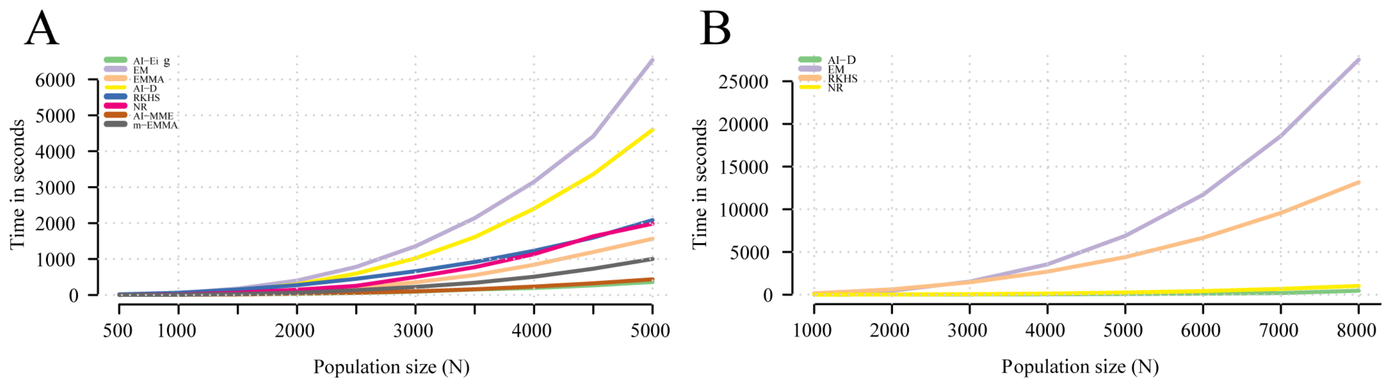

Fig 4. Time performance for different algorithms. In A) different color lines represent the different likelihood-based algorithms tested for populations sizes from 500 to 5000 in steps of 500 as a function of population size (N) for a single random effect. In B) the different color lines represent the algorithms able to deal with multiple random effects for a model with different population sizes (N), from 1000 to 8000 individuals for 3 random effects (GCA_1, GCA_2, and SCA).

doi:10.1371/journal.pone.0156744.g004

be performed in *sommer* as well, and examples are included in the documentation of the package.

A bigger data set with 5000 individuals genotyped with 10000 markers each, was simulated in order to test the performance of *sommer* against their counterparts under a bigger data set scenario with a single random effect. *Sommer* took 4 minutes to estimate all parameters for N = 5000 whereas, ASReml took 7.37 minutes, rrBLUP took 12.8 minutes, regress took 28.06 minutes, EMMREML 10.18 minutes, BGLR 21.5 minutes, and MCMCglmm 148 hours (using the default parameters specified in the M&M section. All calculations performed in a PC-Dell with processor 3.4 GHz Intel Core i7 and a 16 GB RAM memory). For a model with 3 variance components (GCA_1, GCA_2, SCA) with dense covariance structures consisting in 10578 hybrids *sommer* took 7.1 minutes to estimate all parameters, rrBLUP cannot estimate more than one variance component, regress took 31.13 minutes, ASReml took 114.56 minutes, EMMREML 15.61 minutes, BGLR took 289.4 minutes, and MCMCglmm took more than 12 days. This shows an important time reduction using *sommer* selected algorithms in comparison with other genomic prediction software in the magnitude of minutes for REML-based, and hours to days compared to Bayesian-based software (Table 3).

We use the simulated dataset with 5000 individuals and 10000 markers to show how the computing time behaves as a function of the population size (N) for the different ML/REML algorithms found across software packages. We recorded the elapsed time for different population sizes from 500 to 5000 in intervals of 500. We found all algorithms to have a similar computation time for small population sizes (Fig 4A). As the problem increased in complexity (i.e. above 2000 individuals) we found the EMMA and AI algorithm using the eigen decomposition (AI-eig) to perform better than other algorithms such as EM and NR and AI-D. The time increment followed a quadratic behavior which exemplifies the issue of dealing with big populations (Fig 4B). The use of sparse methods such as the eigen decomposition on covariance matrices proposed by Lee and van der Werf [32] and Zhou and Stephens [36] constitute an important alternative to confront the population size increment that the researcher should take advantage of. On the other hand, when dealing with multiple random effects and more complex structures, we found the AI-D and NR algorithms to perform better than EM and RKHS sampling method.

The efficiency of *sommer* compared to most software relies on the use of the direct average algorithm (AI) proposed by Lee et al. [22], which surpasses in performance most algorithms when multiple random effects and dense covariance structures are present, the use the EMMA algorithm [20] or the use of an eigen decomposition in the **G** component when a single random effect is present [32], and the capability to switch to the expectation maximization algorithm (EM) [23] when covariance structures are rather sparse.

In addition, packages such as BGLR and regress require to form the kernel product **ZKZ**' for each random effect as an input. This becomes an expensive operation as the model increase in size and complexity, although the Newton-Raphson algorithm performs quite fast, once this kernel has been formed. In the same way, packages such as MCMCglmm and ASReml-R require the inversion of the covariance matrix for the 'ginverse' argument, becoming quite computationally expensive as the covariance structures of the model increases in complexity. This was reflected in the computation time for the three random effects model, where inverting a 10,585 x 10,585 matrix (SCA matrix), corresponded to most of the computation time reported in Table 3 for those programs.

The most important strengths of the *sommer* package can be summarized as great flexibility to use different methods (algorithms), fast execution, and a friendly and intuitive interface that will help researchers to perform fast and easy genome wide association studies, and genomic selection strategies in research and breeding programs.

Conclusion

The purpose of this paper was to describe and make available a general and flexible mixed model solver with popular and efficient algorithms in order to fit genomic selection models, genome wide association studies (GWAS) in diploid and polyploidy organisms, and other non-genetic analyses. Efficient mixed model association (EMMA), expectation-maximization (EM) and average information (AI) algorithms along with kernel methods for estimating additive, dominance and epistatic relationship matrices were developed and presented for plant breeders and scientists through the new R package *sommer*. At the core of the package, the 'mmer' function allows specification of flexible variance-covariance structures and can be used to solve marker-based and kinship-based versions of the genomic prediction and selection models. Examples using maize data illustrated the strengths of *sommer* to increase prediction accuracy in species displaying heterotic effects, which require the estimation of GCA and SCA effects with covariance structures for such random effects. We also showed the functionality of *sommer* in species with null or small heterotic effects and mainly additive effects by analyzing a dataset of wheat lines and obtaining similar results at a smaller computation time, and greater flexibility compared to those obtained with other popular software for genomic selection, based on Bayesian statistics and other REML-based software. Future implementations in *sommer* will include the addition of residual structures (R structures) to allow users to model spatio-temporal trends or longitudinal data, and other popular algorithms to provide more flexibility.

Supporting Information

S1 File. Code for analysis and tables. R code to recreate the analysis.
(R)

Acknowledgments

The author thanks Natalia de Leon, Jeffrey Endelman, Juan Zalapa, Brandon Schlautman, Maria Kamenetsky, and Brett Burdo for their helpful comments on the manuscript, and two anonymous reviewers who helped to enhance the quality of the paper with valuable comments and ideas.

Author Contributions

Conceived and designed the experiments: GCP. Performed the experiments: GCP. Analyzed the data: GCP. Contributed reagents/materials/analysis tools: GCP. Wrote the paper: GCP.

References

1. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005, 437:376–380. PMID: 16056220
2. Goddard ME, and Hayes BJ. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat. Rev. Genet. 2009, 10:381–391. doi: 10.1038/nrg2575 PMID: 19448663
3. Gianola D, and Rosa GJM. One Hundred Years of Statistical Developments in Animal Breeding. Annu. Rev. Anim. Biosci. 2015, 3:19–56. doi: 10.1146/annurev-animal-022114-110733 PMID: 25387231
4. Sax K. The Relation between Chromosome Number, Morphological Characters and Rust Resistance in Segregates of Partially Sterile Wheat Hybrids. Genetics 1923, 8(4):301–321. PMID: 17246015
5. Henderson CR. Selection index and expected genetic advance. In Statistical Genetics and Plant Breeding, ed. Hanson WD, Robinson HF, Publ. 992. Washington, DC: Natl. Acad. Sci. Natl. Res. Counc. 1963 pp. 141–63.
6. Henderson CR. Best Linear Unbiased Estimation and Prediction under a Selection Model. Biometrics 1975, 31(2):423–447. PMID: 1174616

7. Bernardo R. Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Sci. 1994, 34:20–25.
8. Bertrand CY, and Mackill DJ. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Phil. Trans. R. Soc. B. 2008, 363:557–572. doi: 10.1098/rstb.2007.2170 PMID: 17715053
9. Xu Y, and Crouch JH. Marker-Assisted Selection in Plant Breeding: From Publications to Practice. Crop Sci. 2008, 48:391–407. doi: 10.2135/cropsci2007.04.0191
10. Whittaker JC, Thompson R, and Denham MC. Marker-assisted selection using ridge regression. Genet. Res. Camb. 2000, 75:249–252.
11. Meuwissen TH, Hayes BJ, and Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics 2001, 157:1819–1829. PMID: 11290733
12. Searle SR, Casella G, and McCulloch CE. Variance components. John Wiley & Sons, Hoboken, NJ. 2006
13. Endelman JB. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 2011, 4:250–255.
14. Zhao Y, Mette MF, and Reif JC. Genomic selection in hybrid breeding. Plant Breeding 2015, 134:1–10.
15. Hadfield JD. MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. Journal of Statistical Software 2010, 33(2), 1–22. Available: http://www.jstatsoft.org/v33/i02/.
16. Pérez P, de los Campos G. Genome-Wide Regression & Prediction with the BGLR Statistical Package. Genetics 2014; doi: 10.1534/genetics.114.164442
17. Clifford D and McCullagh P. The regress function. R News 2006, 6(2):6–10
18. Wimmer V, Albrecht T, Auinger HJ and Schoen CC (2012) synbreed: a framework for the analysis of genomic prediction data using R. Bioinformatics, 28: 2086–2087 doi: 10.1093/bioinformatics/bts335 PMID: 22689388
19. Deniz Akdemir and Okeke Uche Godfrey (2015). EMMREML: Fitting Mixed Models with Known Covariance Structures. R package version 3.1. Available: http://CRAN.R-project.org/package=EMMREML
20. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, et al. Efficient control of population structure in model organism association mapping. Genetics 2008, 178:1709–1723. doi: 10.1534/genetics.107.080101 PMID: 18385116
21. Gilmour AR, Thompson R, Cullis BR. Average Information REML: An efficient algorithm for variance parameter estimation in linear mixed models. Biometrics 1995, 51(4):1440–1450.
22. Lee SH, Werf JH van der. An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree. Genetics, Selection, Evolution 2006, 38(1):25–43. doi: 10.1186/1297-9686-38-1-25 PMID: 16451790
23. Searle SR. Applying the EM algorithm to calculating ML and REML estimates of variance components. Paper invited for the 1993 American Statistical Association Meeting, San Francisco.
24. Bernardo R. Breeding for quantitative traits in plants. Second edition. Stemma Press. Woodbury, MN; 2010.
25. Munoz PR, Marcio FR, Resende Jr., Gezan SA, Vilela Resende MD, de los Campos G, et al. Unraveling Additive from Nonadditive Effects Using Genomic Relationship Matrices. Genetics 2014, 198:1759–1768. doi: 10.1534/genetics.114.171322 PMID: 25324160
26. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria 2016. Available: http://www.R-project.org/.
27. VanRaden PM. Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science 2008, 91 (11):4414–23. doi: 10.3168/jds.2007-0980 PMID: 18946147
28. Technow F, Schrag TA, Schipprack W, Bauer E, Simianer H, and Melchinger AE. Genome Properties and Prospects of Genomic Prediction of Hybrid Performance in a Breeding Program of Maize. Genetics 2014, 197:1343–1355. doi: 10.1534/genetics.114.165860 PMID: 24850820
29. 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. Genetics 2006, 38:203–208. PMID: 16380716
30. Rosyara UR, De Jong WS, Douches DS, Endelman JB. Software for Genome-Wide Association Studies in Autopolyploids and Its Application to Potato. The Plant Genome 2016, doi: 10.3835/plantgenome2015.08.0073
31. Su G, Christensen OF, Ostersen T, Henryon M, Lund MS. Estimating Additive and Non-Additive Genetic Variances and Predicting Genetic Merits Using Genome-Wide Dense Single Nucleotide

Polymorphism Markers. PLoS ONE 2012, 7(9): e45293. doi: 10.1371/journal.pone.0045293 PMID: 23028912

32. Lee SH and van der Werf JHJ. MTG2: An efficient algorithm for multivariate linear mixed model analysis based on genomic information. Bioinformatics 2016, doi: 10.1093/bioinformatics /btw012
33. Crossa J, de los Campos G, Pérez P, Gianola D, Burgueño J, Araus JL, et al. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 2010, 186:713–724. doi: 10.1534/genetics.110.118521 PMID: 20813882
34. Bernardo R, Moreau L, and Charcosset A. Number and fitness of selected individuals in marker-assisted and phenotypic recurrent selection. Crop Sci. 2006, 46:1972–1980.
35. Zhong S, and Jannink JL. Using quantitative trait loci results to discriminate among crosses on the basis of their progeny mean and variance. Genetics 2007, 177:567–576. PMID: 17660556
36. Zhou X and Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat Meth 2014, 11(4):407–409.