bioRxiv preprint doi: https://doi.org/10.1101/014100; this version posted January 26, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

# Integrating crop growth models with whole genome prediction through approximate Bayesian computation

Frank Technow[1*], Carlos D. Messina[2], L. Radu Totir[1], Mark Cooper[2]
**1 Breeding Technologies, DuPont Pioneer, Johnston, IA, USA**
**2 Trait Characterization & Development, DuPont Pioneer, Johnston, IA, USA**
∗ **E-mail: Frank.Technow@pioneer.com**

## Abstract

Genomic selection, enabled by whole genome prediction (WGP) methods, is revolutionizing plant breeding. Existing WGP methods have been shown to deliver accurate predictions in the most common settings, such as prediction of across environment performance for traits with additive gene effects. However, prediction of traits with non-additive gene effects and prediction of genotype by environment interaction (G×E), continues to be challenging. Previous attempts to increase prediction accuracy for these particularly difficult tasks employed prediction methods that are purely statistical in nature. Augmenting the statistical methods with biological knowledge has been largely overlooked thus far. Crop growth models (CGMs) attempt to represent the functional relationships between plant physiology and the environment in the formation of yield and similar output traits of interest. Thus, they can explain the impact of G×E and certain types of non-additive gene effects on the expressed phenotype. Approximate Bayesian computation (ABC), a novel and powerful computational procedure, allows the incorporation of CGMs directly into the estimation of whole genome marker effects in WGP. Here we provide a proof of concept study for this novel approach and demonstrate its use with a simulated data set. We show that this novel approach can be considerably more accurate than the benchmark WGP method GBLUP in predicting performance in environments represented in the estimation set as well as in previously unobserved environments for traits determined by non-additive gene effects. We conclude that this proof of concept demonstrates that using ABC for incorporating biological knowledge in the form of CGMs into WGP is a very promising novel approach to improving prediction accuracy for some of the most challenging scenarios of interest to applied geneticists.

## Introduction

Genomic selection (Meuwissen et al. [1]), enabled by whole genome prediction (WGP) methods, is revolutionizing plant breeding (Cooper et al. [2]). Since its inception, attempts to improve prediction accuracy have focused on: developing improved and specialized statistical models (Yang and Tempelman [3], Heslot et al. [4], Kärkkäinen and Sillanpää [5], Technow and Melchinger [6]), increasing the marker density used (Meuwissen and Goddard [7], Erbe et al. [8], Ober et al. [9]), increasing the size and defining optimal designs of estimation sets (Rincent et al. [10], Windhausen et al. [11], Technow et al. [12], Hickey et al. [13]) and better understanding the genetic determinants driving prediction accuracy (Daetwyler et al. [14], Habier et al. [15]).

bioRxiv preprint doi: https://doi.org/10.1101/014100; this version posted January 26, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In-silico based phenotypic prediction, enabled by dynamic crop growth models (CGMs), dates back to the late 60's (van Ittersum et al. [16]) and it has constantly evolved since then through inclusion of scientific advances made in plant and crop physiology, soil science and micrometeorology (Keating et al. [17]; van Ittersum et al. [16]). CGMs used in plant breeding are structured around concepts of resource capture, utilization efficiency and allocation among plant organs (Cooper et al. [18]; Hammer et al. [19]; Passioura [20]; Yin et al. [21]) and are used to: characterize environments (Chapman et al. [22]; Löffler et al [23]), predict consequences of trait variation on yield within a genotype x environment x management context (Hammer et al. [24]), evaluate breeding strategies (Chapman et al. [25]; Messina et al. [26]; Messina et al. [27]), and assess hybrid performance (Cooper et al. [2]).

Early attempts to extend the use of CGMs to enable genetic prediction have focused on developing genetic models for parameters of main process equations within the CGM (Chenu et al. [28]; Messina et al. [29]; Yin et al. [21]). Linking quantitative trait locus (QTL) models and GCMs for complex traits motivated adapting CGMs to improve the connectivity between physiology and genetics of the adaptive traits (Hammer et al. [30]; Messina et al. [27]; Yin et al. [21]). However, despite a tremendous body of knowledge and experience, CGMs were largely ignored for the purpose of WGP.

There is ample evidence for the importance of epistasis in crops, including for economically important traits such as grain yield in maize (Wolf and Hallauer [31], Eta-Ndu and Openshaw, [32]; Holland [33]). Yield and other complex traits are the product of intricate interactions between component traits on lower hierarchical levels (Cooper et al. [34], Hammer et al. [19], Riedelsheimer et al. [35]). If the relationship between the underlying component traits is non-additive, epistatic effects can occur on the phenotypic level of complex traits even if the gene action is purely additive when characterized at the level of the component traits (Holland [33]). This phenomenon was first described for multiplicative relationships between traits by Richey [36] and later quantified by Melchinger et al. [37]. CGMs, which explicitly model these non-linear relationships among traits, have therefore the potential to open up novel avenues towards accounting for epistatic effects in WGP models by explicit incorporation of biological knowledge.

The target population of environments for plant breeding programs is subject to continuous change, expansion and increases in diversity (Cooper et al. [2]). To select for performance in specific environments, genotype by environment (G×E) interactions have to be predicted. Genomic prediction of G×E interactions is therefore of great interest to scientists and practitioners alike. Previous attempts incorporated G×E interactions in WGP models through environment specific marker effects (Schulz-Streeck et al. [38]) or genetic and environmental covariances (Burgeño et al. [39]). Later Jarquín et al. [40] and Heslot et al. [41] developed WGP models that accounted for G×E interactions by means of environmental covariates.

While these previous attempts are promising, they are purely statistical in nature and do not leverage the substantial biological insights into the functional mechanisms determining performance in specific environments. CGMs are an embodiment of this biological knowledge and might serve as a key component in novel WGP models for predicting G×E interactions. In fact, Heslot et al. [41] recognized this potential for CGMs. However, they employed them only for computing stress covariates from environmental data, which were subsequently used as covariates in purely statistical WGP models.

Given the potential merits of integrating CGMs in WGP, the question arises of how to combine the two in a unified predictive system. The ever increasing computational power of modern computing environments allows for efficient simulation from the most complex of models, such as CGMs (Messina et al. [27]). This computational power is

bioRxiv preprint doi: https://doi.org/10.1101/014100; this version posted January 26, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

leveraged by *approximate Bayesian computation* (ABC) methods, which replace the calculation of a likelihood function with a simulation step, and thereby facilitate analysis when calculation of a likelihood function is impossible or computationally prohibitive. ABC methods were developed in population genetics, where they helped solve otherwise intractable problems (Tavare et al. [42], Pritchard et al. [43], Csilléry et al. [44], Lopes and Beaumont [45]). However, ABC methods are rapidly adopted in other scientific fields, such as ecology (Lawson Handley et al. [46]), systems biology (Liepe et al. [47]) and hydrology (Sadegh and Vrugt [48]). Recently, Marjoram et al. [49] proposed using ABC methods for incorporating the biological knowledge represented in gene regulatory networks into genome-wide association studies, arguing that this might present a solution to the "missing heritability" problem.

Here we make the case that ABC may hold great promise for enabling novel approaches to WGP as well. Thus, the objective of this study is to provide a proof of concept, based on a simulated data set, for using ABC as a mechanism for incorporating the substantial biological knowledge embodied in CGMs into a novel WGP approach.

# Materials and Methods

## CGM and environmental data

We used a modified version of the maize CGM developed by Muchow et al. [50], which models maize grain yield development as a function of plant population (plants $m^{-2}$), daily temperature ($^{\circ}C$) and solar radiation ($MJ\ m^{-2}$) as well as several genotype dependent physiological traits. These traits were total leaf number (TLN), area of largest leaf (AM), solar radiation use efficiency (SRE) and thermal units to physiological maturity (MTU). Details on how the values for these traits were generated for the genotypes in the synthetic data set are provided later. However, the values used were within typical ranges reported in the literature. The simulated intervals for TLN, AM, SRE and MTU were $[6, 18]$ (Meghji et al. [51], Muchow et al. [50]), $[700, 800]$ (Muchow et al. [50], Elings [52]) , $[1.5, 1.7]$ (Muchow and Davis [53] and $[1050, 1250]$ (McGarrahan and Dale [54], Muchow [55], Nielsen et al. [56]), respectively, with average values at the midpoints of the intervals.

We chose Champaign/Illinois (40.08° N, 88.24° W) as a representative US corn belt location. Temperature and solar radiation data were obtained for the years 2012 and 2013 (Data provided by the Water and Atmospheric Resources Monitoring Program, a part of the Illinois State Water Survey (ISWS) located in Champaign and Peoria, Illinois, and on the web at `www.sws.uiuc.edu/warm`). In 2012, we used as sowing date April 1st and a plant population of 5 plants $m^{-2}$. In 2013, the sowing date used was May 1st and the plant population 10 plants $m^{-2}$. The 2012 and 2013 environments therefore differed not only in temperature and solar radiation but also in simulated management practices. The temperature and solar radiation from date of sowing is shown in Fig. 1 and typical growth curves obtained from the CGM for the two environments are shown in Fig. 2.

The CGM can be viewed as a function $F$ of the genotype specific inputs (the physiological traits) and the environment data

$$F(y_{TLN_i}, y_{SRE_i}, y_{AM_i}, y_{MTU_i}, \Omega_k) \tag{1}$$

where $y_{TLN_i}$ etc. are the values of the physiological traits observed for the $i^{th}$ genotype and the weather and management data of environment $k$ is represented as $\Omega_k$. To simplify notation, we will henceforth use $F(\cdot)_{ik}$ to represent the CGM and its inputs for genotype $i$ in environment $k$.

bioRxiv preprint doi: https://doi.org/10.1101/014100; this version posted January 26, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

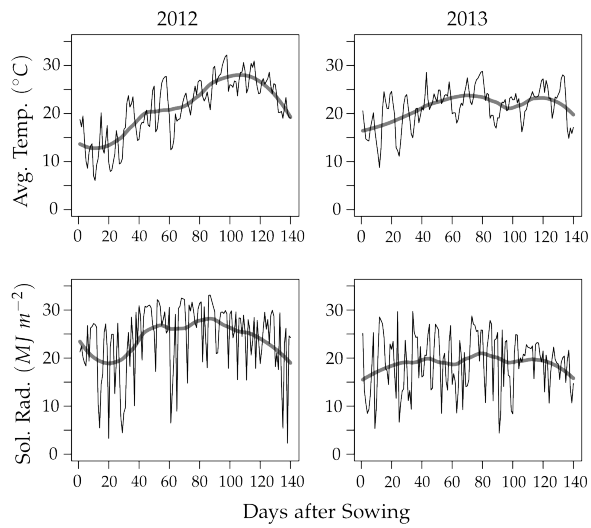


**Figure 1. Daily average temperature and solar radiation at Champaign, Illinois in 2012 and 2013.** The thick grey line shows a smoothed curve.

## Approximate Bayesian Computation (ABC)

ABC replaces likelihood computation with a simulation step (Tavare et al. [42]). An integral component of any ABC algorithm is therefore the simulation model operator Model($y^*_{ik} \mid \theta$) which generates simulated data $y^*_{ik}$ given parameters $\theta$. In our proof of concept study, the crop growth model $F(\cdot)_{ik}$ represents the deterministic component of Model($y^*_{ik} \mid \theta$), to which a Gaussian noise variable distributed as $\mathcal{N}(0, \sigma^2_e)$ is added as a stochastic component. If Model($y^*_{ik} \mid \theta$) is fully deterministic, the distribution sampled with the ABC algorithm will not converge to the true posterior distribution when the tolerance for the distance between the simulated and observed data goes to zero (Sadegh and Vrugt [48]).

The weather and management data $\Omega_k$ was assumed to be known, the physiological traits, however, were unknown and modeled as linear functions of the trait specific marker effects

$$\begin{aligned} y_{TLN_i} &= \mu_{TLN} + \mathbf{z}_i \mathbf{u}_{TLN} \\ y_{AM_i} &= \mu_{AM} + \mathbf{z}_i \mathbf{u}_{AM} \\ y_{SRE_i} &= \mu_{SRE} + \mathbf{z}_i \mathbf{u}_{SRE} \\ y_{MTU_i} &= \mu_{MTU} + \mathbf{z}_i \mathbf{u}_{MTU}, \end{aligned} \tag{2}$$

where $\mathbf{z}_i$ is the genotype vector of the observed biallelic single nucleotide polymorphism (SNP) markers of genotype $i$, $\mu_{TLN}$ etc. denote the intercepts and $\mathbf{u}_{TLN}$ etc. the marker effects. For brevity, we will use $\theta$ to denote the joint parameter vector $[\mu_{TLN}, \ldots, \mu_{MTU}, \mathbf{u}_{TLN}, \ldots, \mathbf{u}_{MTU}]$.

We used independent Normal distribution priors for all components of $\theta$. The prior for $\mu_{TLN}$ was $\mathcal{N}(m_{TLN}, \sigma^2_{\mu_{TLN}})$. To simulate imperfect prior information, we drew the prior mean $m_{TLN}$ from a Uniform distribution over the interval $[0.8 \cdot \overline{TLN}, 1.2 \cdot \overline{TLN}]$, where $\overline{TLN}$ is the observed population mean of TLN. The average difference between $m_{TLN}$ and $\overline{TLN}$ then is 10% of the latter value. The prior variance $\sigma^2_{\mu_{TLN}}$, which represents the prior uncertainty was equal to $2.25^2$. The prior means of AM, SRE and MTU were obtained accordingly and the prior variances $\sigma^2_{\mu_{AM}}$, $\sigma^2_{\mu_{SRE}}$ and $\sigma^2_{\mu_{MTU}}$ were $150^2$, $0.3^2$ and $225^2$, respectively.

The prior for the marker effects $\mathbf{u}_{TLN}$ was $\mathcal{N}(0, \sigma^2_{u_{TLN}})$, which corresponds to the *BayesC* prior (Habier et al. [57]). In BayesC, the prior variance of marker effects $\sigma^2_{u_{TLN}}$,

bioRxiv preprint doi: https://doi.org/10.1101/014100; this version posted January 26, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

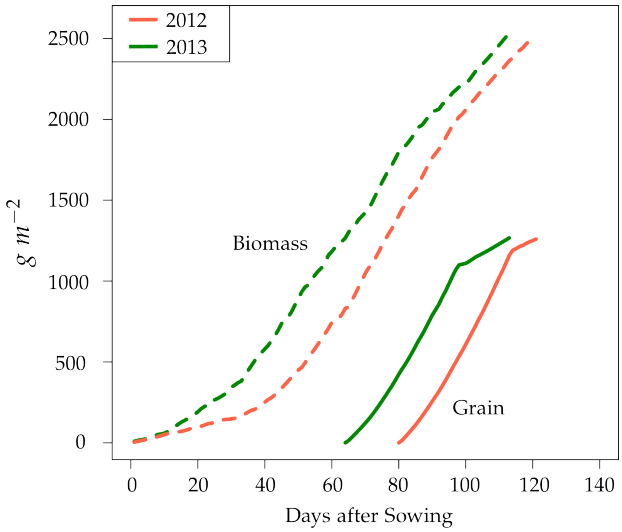


**Figure 2. Simulated grain and total biomass development from the crop growth model in 2012 and 2013 environments.** The growth curves correspond to a hypothetical genotype with average values for physiological traits TLN, AM, SRE and MTU.

which introduces shrinkage, is the same across markers. For simplicity, we set this variance to a constant value and did not attempt to estimate it. Also in this case we simulated imperfect information by drawing the value of $\sigma^2_{u_{TLN}}$ from a Uniform distribution over the interval $[0.8 \cdot var(TLN)/M, 1.2 \cdot var(TLN)/M]$, where $M$ is the number of markers and $var(TLN)$ the observed population variance of TLN. The prior variances of marker effects of the other traits were obtained accordingly.

The value of $\sigma^2_e$, the variance of the Gaussian noise variable that is part of the model operator Model$(y^*_{ik} \mid \theta)$, was drawn from a Uniform distribution over the interval $[0.8 \cdot v_e, 1.2 \cdot v_e]$, where $v_e$ is the residual variance component of the phenotypic grain yield values used to fit the model.

Algorithm 1 in Table 1 shows pseudocode for the ABC rejection sampling algorithm we used. We used the Euclidean distance to measure the distance between the simulated and observed data. The tolerance level $\epsilon$ for the distance between the simulated and observed data was tuned in a preliminary run of the algorithm to result in an acceptance rate of approximately $1 \cdot 10^{-6}$. The number of posterior samples drawn was 100. We will refer to this ABC based prediction method that incorporates the CGM as *ABC-CGM*. The ABC-CGM algorithm was implemented as a C routine integrated with the R software environment (R Core Team [58]).

## Synthetic data set

To test the performance of ABC-CGM, we created a biparental population of 1,550 doubled haploid (DH) inbred lines in silico. The genome consisted of a single chromosome of 1.5 Morgan length. The genotypes of the DH lines were generated by simulating meiosis events with the software package hypred (Technow [59]) according to the Haldane mapping function. On the chromosome we equidistantly placed 140 informative SNP markers. A random subset of 40 of these markers were assigned to be QTL with additive effects on either TLN, AM, SRE or MTU. Each physiological trait was controlled by 10 of the 40 QTL, which were later removed from the set of observed markers available for analysis.

bioRxiv preprint doi: https://doi.org/10.1101/014100; this version posted January 26, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

**Table 1. Pseudo code of ABC rejection sampling algorithm**

```
while x <= no. posterior samples do
    while d > ε do
        draw candidate θ* from prior(θ)
        for all i = 1, 2, ..., N do
            generate simulated data y*_ik from Model(y*_ik | θ*)
        end for
        compute d = sqrt(Σ_{i=1}^{N} (y_ik − y*_ik)^2)
    end while
    accept and store θ*
    increment x
end while
```

Basic ABC rejection sampling algorithm to sample from the approximate posterior distribution of $\theta$.

The additive substitution effects of the QTL were drawn from a Standard Normal distribution. Then raw genetic scores for each physiological trait were computed by summing the QTL effects according to the QTL genotypes of each DH line. These raw scores were subsequently re-scaled linearly to the aforementioned value ranges. Finally, phenotypic grain yield values were created as

$$y_{ik} = F(\cdot)_{ik} + e_{ik}, \tag{3}$$

where $e_{ik}$ is a Gaussian noise variable with mean zero and variance $v_e$. The value of $v_e$ was chosen such that the within-environment heritability of $y_{ik}$ was equal to 0.85. We generated 25 synthetic data sets by repeating the whole process.

## Estimation, prediction and testing procedure

The models were fitted using $N = 50$ randomly chosen DH lines as an estimation set. The remaining 1500 DH lines were used for testing model performance. Separate models were fitted using the 2012 and the 2013 grain yield data of the estimation set lines. The environment from which data for fitting the model was used will be referred to as *estimation environment*. Parameter estimates from each estimation environment were subsequently used to predict performance of the lines in the test set in both environments. Predictions for the same environment as the estimation environment will be referred to as *observed environment predictions* (e.g., predictions for 2012 with models fitted with 2012 data). Predictions for an environment from which no data were used in fitting the model will be referred to as *new environment predictions* (e.g., predictions for 2013 with models fitted with 2012 data).

As a point estimate for predicted grain yield performance in a specific environment we used the mean of the posterior predictive distribution for the DH line in question. The posterior predictive distribution was obtained by evaluating $F(\cdot)_{ik}$ over the accepted $\theta$ samples, using the weather and management data $\Omega_k$ pertaining to that environment.

Prediction accuracy was computed as the Pearson correlation between predicted and true performance in the environment for which the prediction was made. The true grain yield performance was obtained by computing $F(\cdot)_{ik}$ with the true values of the physiological traits.

As a performance benchmark we used genomic best linear unbiased prediction

bioRxiv preprint doi: https://doi.org/10.1101/014100; this version posted January 26, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(GBLUP, Meuwissen et al. [1]). The model is

$$y_{ik} = \beta_0 + \mathbf{z}_i\mathbf{u} + e_i \tag{4}$$

where $\beta_0$ is the intercept, $\mathbf{u}$ the vector of marker effects and $e_i$ a residual. As before, $\mathbf{z}_i$ denotes the marker genotype vector. The GBLUP model was fitted with the R package rrBLUP (Endelman [60]). GBLUP and BayesC are comparable in their shrinkage behavior because both use a constant variance across markers. For GBLUP, predicted values were computed according to eq. (4) as $\beta_0 + \mathbf{z}_i\mathbf{u}$. Note that because the conventional GBLUP model does not utilize information about the target environment, observed and new environment predictions are identical.

# Results and Discussion

## Predicting performance in observed environments

The accuracy for observed environment predictions achieved by ABC-CGM was larger than that of the benchmark method GBLUP by 0.15 points, on average (Table 2, Fig. 3). This superiority of ABC-CGM over GBLUP can be explained by the presence of non-additive gene effects which cannot be captured fully by the latter. In the example scenario we studied, the non-additive gene effects on grain yield are a result of non-linear functional relationships between the physiological traits and grain yield, which was particularly pronounced for TLN (Fig. 4). This thus presents a particular case of epistasis, which might be called *biological epistasis*, that can arise even if the gene effects on the physiological component traits underlying the final trait of interest (grain yield in our case) are purely additive (Holland [33]).

**Table 2. Accuracy of grain yield predictions of test DH lines**

| Estimation Env. | Prediction Env. | ABC-CGM | GBLUP |
|---|---|---|---|
| 2012 | 2012 | 0.81 | 0.68 |
| | 2013 | 0.58 | 0.38 |
| 2013 | 2012 | 0.60 | 0.46 |
| | 2013 | 0.78 | 0.64 |

Prediction accuracy for grain yield of DH lines in the test set, averaged over 25 replications. All differences within a row are statistically significant at a significance level of $< 0.005$.

We accounted for non-linear functional relationships among traits with the CGM. This enabled us to capture biological epistasis through simple linear models relating marker genotypes to the unobserved underlying physiological traits. Previously developed WGP models attempted to capture epistasis by directly fitting non-linear marker effects to the final trait of interest (e.g., Xu [61], Sun et al. [62], Howard et al. [63]). While these models showed some promise, they have not been adopted by practitioners on a larger scale. By combining statistics with biological insights captured by CGMs, ABC-CGM takes a fundamentally different approach and presents a potentially powerful alternative to purely statistical WGP models.

## Predicting performance in new environments

New environment prediction accuracy was considerably lower than observed environment prediction accuracy, for both prediction methods (Table 2, Fig. 3). The average prediction accuracy for performance in 2012 when using the 2013 estimation

bioRxiv preprint doi: https://doi.org/10.1101/014100; this version posted January 26, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

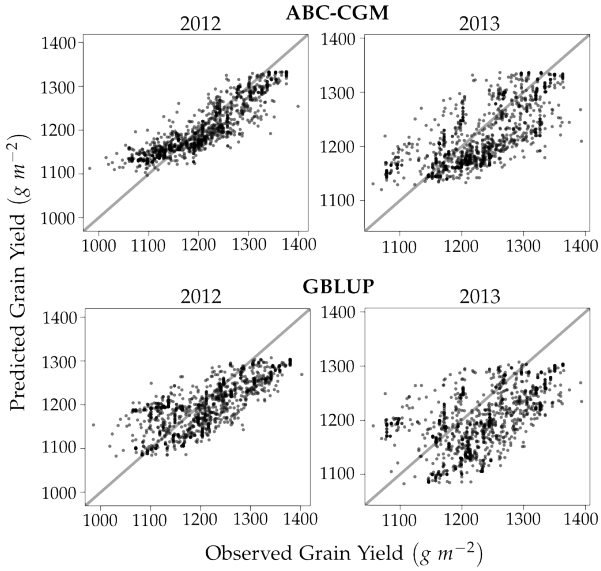


**Figure 3. Predicted vs. observed grain yield** of 1500 DH lines in testing set for prediction methods ABC-CGM (top row) and GBLUP (bottom row). The estimation environment was 2012. Results shown are from a representative example data set. In this example, the accuracy for same environment predictions was 0.82 (ABC-CGM) and 0.72 (GBLUP). For new environment predictions it was 0.64 (ABC-CGM) and 0.47 (GBLUP).

environment was 74% (ABC-CGM) and 68% (GBLUP) of the respective prediction accuracy achieved when using the 2012 estimation environment. The corresponding values for the accuracy of predicting performance in 2013 were 74% (ABC-CGM) and 59% (GBLUP).

The rank correlation between true performance in 2012 and 2013 was only 0.73 (averaged over 25 simulated data sets), which indicated the presence of considerable G×E interactions, including changes in rank (Fig. 5). That new environment prediction under the presence of G×E interactions is considerably less accurate than observed environment prediction was expected and already observed in other studies (Resende et al. [64], Windhausen et al. [11]). That the reduction in accuracy for ABC-CGM was less severe than for the conventional benchmark method GBLUP is encouraging because it indicates that the former method did succeed in predicting G×E interactions to some degree.

Predicting G×E interactions in new environments in which no yield testing was carried out previously, requires WGP models that link genetic effects (e.g., marker effects) with information that characterizes the environments. Jarquín et al. [40] accomplished this by fitting statistical interactions between markers and environmental covariates. A similar approach was taken by Heslot et al. [41], who in addition used a CGM to extract stress covariates from a large set of environmental variables. ABC-CGM takes this approach a step further by making the CGM an integral part of the estimation procedure.

However, while novel prediction methods might succeed in narrowing the gap between new and observed environment prediction, the former should always be expected to be less accurate than the latter. Field testing should therefore be carried out in environments of particular importance for a breeding program to achieve the maximum attainable prediction accuracy for these. The same applies for target environments in which G×E interaction effects are expected to be particularly strong.

bioRxiv preprint doi: https://doi.org/10.1101/014100; this version posted January 26, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

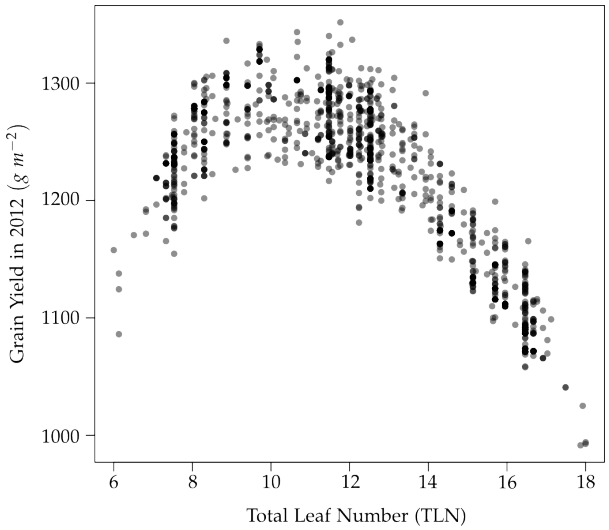


**Figure 4. Relationship between grain yield in 2012 and total leaf number (TLN)**. Results shown are from a representative example data set.

CGMs can help in identifying such environments, inform experimental design and utilization of managed environments (Messina et al. [27], Messina et al. [29]). However, the range of the target population of environments of modern plant breeding programs is much too large for facilitating yield testing in all of them (Cooper et al. [2]). Predicting performance in completely new environments will therefore always be required and novel methods like ABC-CGM are anticipated to be instrumental for enabling and enhancing success in this particularly daunting task.

## Further developments

**More sophisticated CGMs** For this first proof of concept study, we assumed that the crop growth model used in the estimation process fully represented the systematic component of the data generating process, besides the random noise component. This was clearly a "best case scenario". However, decades of crop growth modeling research have provided the know-how necessary to approximate real crop development to an almost arbitrary degree of accuracy (Keating et al. [17]), Renton [65], Hammer et al. [30]). Advanced CGMs such as *APSIM* (Keating et al. [17]), for example, model functional relationships between various crop parameters and external factors such as water and nutrient availability, soil properties as well as weed, insect and pathogen pressure. Thus, the tools are principally available for applying ABC-CGM in much more complex scenarios than the one studied by us.

**Stochastic CGMs** There are examples of the use of fully deterministic model operators in ABC (Toni et al. [66], Liepe et al. [67]). However, with fully deterministic model operators the sampled distribution would not converge to the true posterior when the tolerance level $\epsilon$ goes to zero (Sadegh and Vrugt [48]) and instead reduce to a point mass over those parameter values that can reproduce the data. The CGM we used was fully deterministic. We therefore followed the example of Sadegh and Vrugt [48], who constructed a stochastic model operator by adding a random noise variable, with the same probabilistic properties as assumed for the residual component of the phenotype, to the deterministic functional model. A more elegant and possibly superior solution, however, would be to integrate stochastic processes directly into the CGM. While the

bioRxiv preprint doi: https://doi.org/10.1101/014100; this version posted January 26, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

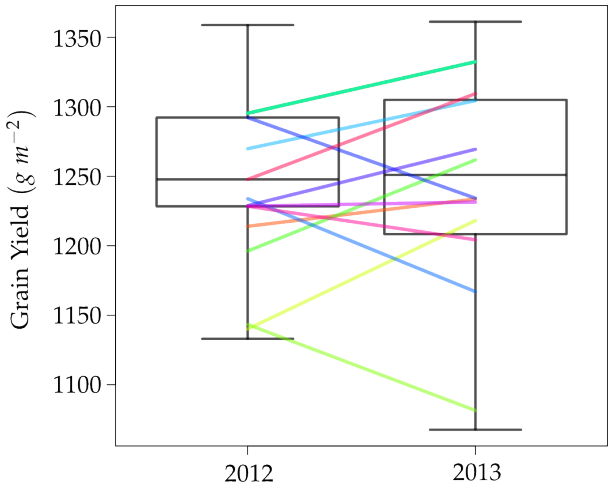


**Figure 5. Distribution of simulated grain yield in 2012 and 2013 environments.** The colored lines indicate the performance of specific genotypes in both environments. Data shown is from a representative example replication.

vast majority of CGMs are deterministic (Keating et al. [17], van Ittersum et al. [16]), there are examples of stochastic CGMs (Brun et al. [68]). In addition to incorporating inherently stochastic processes of development (Curry et al. [69]), stochastic CGMs could also serve to account for uncertainty in the parameters of the functional equations comprising the model (Wallach et al. [70]).

**Advanced ABC algorithms** For this proof of concept study we used the basic ABC rejection sampling algorithm (Tavare et al. [42], Pritchard et al. [43]). Considerable methodology related advances have been made, however, over the last decade that have led to algorithms with improved computational efficiency. Of particular interest here are population or sequential Monte Carlo algorithms, which are based on importance sampling (Sisson et al. [71], Toni et al. [66], Peters et al. [72]). These algorithms can dramatically increase acceptance rates without compromising on the tolerance levels. They achieve this by sampling from a sequence of intermediate proposal distributions of increasing similarity to the target distribution. Unfortunately, importance sampling fails when the number of parameters gets large, because then the importance weights tend to concentrate on very few samples, which leads to an extremely low effective sample size (Bengtsson et al. [73]). In the context of sequential Monte Carlo, this is known as particle depletion and was addressed by Peters et al. [72]. We implemented their approach, but were not able to overcome the problem of particle depletion. The number of parameters we estimated was 404 (100 marker effects per physiological trait plus an intercept), which seems well beyond the dimensionality range for importance sampling (Bengtsson et al. [73]).

Another interesting development is *MCMC-ABC*, which incorporates ABC with the Metropolis-Hastings algorithm (Marjoram et al. [74]). MCMC-ABC should result in high acceptance rates if the sampler moves into parameter regions of high posterior probability. However Metropolis-Hastings sampling too can be inefficient when the parameter space is of high dimension.

The greatest computational advantage of the original ABC rejection algorithm over Monte Carlo based ABC methods is that it generates independent samples and therefore readily lends itself to "embarrassingly" parallel computation (Marjoram et al. [74]). The computation time thus scales linearly to the number of processors available. Using the ABC rejection algorithm therefore allowed us to fully leverage the

bioRxiv preprint doi: https://doi.org/10.1101/014100; this version posted January 26, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

high performance computing cluster of DuPont Pioneer. In the era of cloud computing (Buyya et al. [75]), high performance computing environments are readily available to practitioners and scientists in both public and private sectors. Generality, scalability to parallel computations, and ease of implementation make the basic rejection sampler a viable alternative to more sophisticated approaches.

**Using prior information** We used mildly informative prior distributions, the parameters of which were derived from the population means and variances of the physiological traits. In practice, the required prior information must be obtained from extraneous sources, such as past experiments or from the literature (Brun et al. [68]). Such information is imperfect and only partially matches the true population parameters of the population in question. We determined the prior parameters from the population itself, but perturbed them considerably to simulate erroneous prior information. Specifically, the average relative discrepancy between the prior parameter used and the true population parameter was 10%. When we increased the relative discrepancy to 25%, prediction accuracy dropped, but only slightly (Table S1). Thus, ABC-CGM seems to be relatively insensitive to moderate prior miss specification, as long as the value range supported by the prior distribution is not completely out of scope. If prior information is unavailable, uninformative priors, such as Normal distributions with mean zero and a very large variance, have to be used. This, however, will require a reduced acceptance rate to retain a given tolerance threshold because most of the parameter vectors drawn from the prior in step 1 of the algorithm (1) will be far from the bulk of the posterior. As a consequence, computational efficiency could be reduced considerably.

In contrast to the complex trait of interest, component physiological traits may be realistically modeled based on a relatively simple genetic architecture and for such traits QTL explaining a sizable proportion of genetic variance can be mapped and characterized (Reymond et al. [76], Bogard et al. [77], Yin et al. [78], Welcker et al. [79], Tardieu et al. [80]). In fact, such component trait QTL were already successfully used to parametrize CGMs for studying genotype dependent response to environmental conditions (Chenu et al. [28], Messina et al. [29], Bogard et al. [77], Chenu et al. [81], Yin et al. [78], ). Knowledge about the location of such QTL or of transgenes (Dong et al. [82], Guo et al. [83], Habben et al. [84]), could be incorporated as an additional source of prior information. Then, instead of estimating marker effects for the whole genome, ABC-CGM could focus on the QTL regions, which reduces the dimensionality of the parameter space dramatically. The effects of the QTL estimated in previous experiments could inform prior distributions, too. Prior knowledge about QTL positions and effects should therefore be leveraged for improving prediction accuracy and computational efficiency of ABC-CGM.

**Other applications** The idea of incorporating biological insights into WGP models is not limited to CGMs. Plant metabolites are chemical compounds produced as intermediate or end products of biochemical pathways. They are seen as potential bridges between genotypes and phenotypes of plants (Keurentjes [85]) and are therefore of particular interest in plant breeding (Fernie and Schauer, [86]). Metabolic networks model the interrelationships between genes, intermediate metabolites and end products through biochemistry pathways (Schuster et al. [87]). Elaborate metabolic network models are available today that allow studying and simulating complex biochemical processes related to important crop properties, such as flowering time, seed growth, nitrogen use efficiency and biomass composition (Dong et al. [82], Pilalis et al. [88], Simons et al. [89], Saha et al. [90]). Liepe et al. [47] demonstrate how ABC can be used for parameter estimation with metabolic and other biochemical networks. Using the

bioRxiv preprint doi: https://doi.org/10.1101/014100; this version posted January 26, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

principles outlined here for ABC-CGM, metabolic networks might add valuable biological information for the purpose of WGP, too.

Despite ever increasing sample sizes and marker densities, most of the genetic variance of complex traits remained unaccounted for in genome-wide association studies (Maher [91]). Marjoram et al. [49] argued that signal detection power could be increased by augmenting the purely statistical association models used thus far with biological knowledge. They demonstrate their approach by using ABC for incorporating gene regulatory networks into their analysis. Here we showed that the same principle can be applied to WGP by using ABC for integrating a CGM in the estimation of whole genome marker effects. While this study is only a first step and many open questions remain, we conclude that ABC-CGM presents a promising novel path forward towards a new class of WGP models that leverage previously unused sources of knowledge and information and thereby increase prediction accuracy in settings that have proved challenging thus far for plant breeding.

# Supporting Information

## S1 Table

**Table S1. Accuracy of grain yield predictions of test DH lines with increased error in prior parameters**

# Acknowledgments

# References


1. Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.

2. Cooper M, Messina CD, Podlich D, Totir LR, Baumgarten A, et al. (2014) Predicting the future of plant breeding: complementing empirical evaluation with genetic prediction. Crop and Pasture Sci 64: 311-336.

3. Yang W, Tempelman RJ (2012) A Bayesian antedependence model for whole genome prediction. Genetics 190: 1491-1501.

4. Heslot N, Yang HP, Sorrells ME, Jannink JL (2012) Genomic selection in plant breeding: a comparison of models. Crop Sci 52: 146–160.

5. Kärkkäinen HP, Sillanpää MJ (2012) Back to basics for Bayesian model building in genomic selection. Genetics 191: 969–987.

6. Technow F, Melchinger AE (2013) Genomic prediction of dichotomous traits with Bayesian logistic models. Theor Appl Genet 126: 1133-1143.

7. Meuwissen T, Goddard M (2010) Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics 185: 623–631.

8. Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, et al. (2012) Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci 95: 4114–4129.

bioRxiv preprint doi: https://doi.org/10.1101/014100; this version posted January 26, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

9. Ober U, Ayroles JF, Stone EA, Richards S, Zhu D, et al. (2012) Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster. PLoS Genet 8: e1002685.

10. Rincent R, Laloe D, Nicolas S, Altmann T, Brunel D, et al. (2012) Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize. Genetics 192: 715–728.

11. Windhausen VS, Atlin GN, Hickey JM, Crossa J, Jannink JL, et al. (2012) Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3 2: 1427-1436.

12. Technow F, Bürger A, Melchinger AE (2013) Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. G3 3: 197–203.

13. Hickey JM, Dreisigacker S, Crossa J, Hearne S, Babu R, et al. (2014) Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. Crop Sci in press: doi:10.2135/cropsci2013.03.0195.

14. Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA (2010) The impact of genetic architecture on genome-wide evaluation methods. Genetics 185: 1021–1031.

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

16. van Ittersum MK, Leffelaar PA, Van Keulen H, Kropff MJ, Bastiaans L, et al. (2003) On approaches and applications of the Wageningen crop models. Eur J Agron 18: 201–234.

17. Keating BA, Carberry PS, Hammeer GL, Probert ME, Robertson MJ, et al. (2003) An overview of APSIM, a model designed for farming systems simulation. Eur J Agron 18: 267–288.

18. Cooper M, van Eeuwijk FA, Hammer GL, Podlich D, Messina C (2009) Modeling QTL for complex traits: detection and context for plant breeding. Curr Opin Plant Biol 12: 231–240.

19. Hammer G, Cooper M, Tardieu F, Welch S, Walsh B, et al. (2006) Models for navigating biological complexity in breeding improved crop plants. Trends Plant Sci 11: 587–593.

20. Passioura JB (1983) Roots and drought resistance. Agr Water Manage 7: 265–280.

21. Yin X, Struik PC, Kropff MJ (2004) Role of crop physiology in predicting gene-to-phenotype relationships. Trends Plant Sci 9: 426–432.

22. Chapman S, Cooper M, Hammer G, Butler D (2000) Genotype by environment interactions affecting grain sorghum. ii. frequencies of different seasonal patterns of drought stress are related to location effects on hybrid yields. Aust J Agric Res 51: 209–222.

bioRxiv preprint doi: https://doi.org/10.1101/014100; this version posted January 26, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

23. Löffler CM, Wei J, Fast T, Gogerty J, Langton S, et al. (2005) Classification of maize environments using crop simulation and geographic information systems. Crop Sci 45: 1708–1716.

24. Hammer G, Dong Z, McLean G, Doherty A, Messina C, et al. (2009) Can changes in canopy and/or root system architecture explain historical maize yield trends in the U.S. corn belt? Crop Sci 49: 299–312.

25. Chapman S, Cooper M, Podlich D, Hammer G (2003) Evaluating plant breeding strategies by simulating gene action and dryland environment effects. Agron J 95: 99–113.

26. Messina C, Hammer G, Dong Z, Podlich D, Cooper M (2009) Chapter 10 - Modelling crop improvement in a G×E×M framework via gene-trait-phenotype relationships. In: Sadras V, Calderini D, editors, Crop Physiology, San Diego: Academic Press. pp. 235 – 581.

27. Messina CD, Podlich D, Dong Z, Samples M, Cooper M (2011) Yield-trait performance landscapes: from theory to application in breeding maize for drought tolerance. J Exp Bot 62: 855–868.

28. Chenu K, Chapman SC, Hammer GL, McLean G, Salah HB, et al. (2008) Short-term responses of leaf growth rate to water deficit scale up to whole-plant and crop levels: an integrated modelling approach in maize. Plant Cell Environ 31: 378–391.

29. Messina CD, Jones JW, Boote KJ, Vallejos CE (2006) A gene-based model to simulate soybean development and yield responses to environment. Crop Sci 46: 456–466.

30. Hammer GL, van Oosterom E, McLean G, Chapman SC, Broad I, et al. (2010) Adapting APSIM to model the physiology and genetics of complex adaptive traits in field crops. J Exp Bot 61: 2185–2202.

31. Wolf DP, Hallauer AR (1997) Triple testcross analysis to detect epistasis in maize. Crop Sci 37: 736–770.

32. Eta-Ndu JT, Openshaw SJ (1999) Epistasis for grain yield in two $F_2$ populations of maize. Crop Sci 39: 346–352.

33. Holland JB (2001) Epistasis and plant breeding. In: Janick J, editor, Plant Breeding Reviews, Volume 21, Hoboken, NJ: John Wiley & Sons, Inc. pp. 27–92.

34. Cooper M, Chapman SC, Podlich D, Hammer G (2002) The GP problem: quantifying gene-to-phenotype relationships. In Silico Biol 2: 151–164.

35. Riedelsheimer C, Lisec J, Czedik-Eysenberg A, Sulpice R, Flis A, et al. (2012) Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. Proc Natl Acad Sci 109: 8872-8877.

36. Richey FD (1942) Mock-dominance and hybrid vigor. Science 96: 280–281.

37. Melchinger AE, Singh M, Link W, Utz H, von Kittlitz E (1994) Heterosis and gene effects of multiplicative characters: theoretical relationships and experimental results from Vicia faba L. Theor Appl Genet 88: 343-348.

38. Schulz-Streeck T, Ogutu JO, Gordillo A, Karaman Z, Knaak C, et al. (2013) Genomic selection allowing for marker-by-environment interaction. Plant Breeding 132: 532–538.

bioRxiv preprint doi: https://doi.org/10.1101/014100; this version posted January 26, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

39. Burgeño J, de los Campos G, Weigel K, Crossa J (2012) Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. Crop Sci 52: 702–719.

40. Jarquín D, Crossa J, Lacaze X, Du Cheyron P, Daucourt J, et al. (2014) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor Appl Genet 127: 595-607.

41. Heslot N, Akdemir D, Sorrells ME, Jannink JL (2014) Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theor Appl Genet 127: 463-480.

42. Tavare S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. Genetics 145: 505–518.

43. Pritchard J, Seielstad M, Perez-Lezaun A, Feldman M (1999) Population growth of human Y chromosomes: A study of Y chromosome microsatellites. Mol Biol Evol 16: 1791-1798.

44. Csilléry K, Blum MG, Gaggiotti OE, François O (2010) Approximate Bayesian Computation (ABC) in practice. Trends Ecol Evol 25: 410–418.

45. Lopes JS, Beaumont MA (2010) ABC: A useful Bayesian tool for the analysis of population data. Infect Genet and Evol 10: 825–832.

46. Lawson Handley LJ, Estoup A, Evans DM, Thomas CE, Lombaert E, et al. (2011) Ecological genetics of invasive alien species. BioControl 56: 409-428.

47. Liepe J, Kirk P, Filippi S, Toni T, Barnes CP, et al. (2014) A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. Nat Protoc 9: 439–456.

48. Sadegh M, Vrugt JA (2014) Approximate bayesian computation using Markov chain Monte Carlo simulation: DREAM(ABC). Water Resour Res 50: 6767–6787.

49. Marjoram P, Zubair A, Nuzhdin SV (2014) Post-GWAS: where next? More samples, more SNPs or more biology? Heredity 112: 79–88.

50. Muchow RC, R ST, Bennett JM (1990) Temperature and solar radiation effects on potential maize yield across locations. Agron J 82: 338–343.

51. Meghji MR, Dudley JW, Lambert RJ, Sprague GF (1984) Inbreeding depression, inbred and hybrid grain yields, and other traits of maize genotypes representing three eras. Crop Sci 24: 545–549.

52. Elings A (2000) Estimation of leaf area in tropical maize. Agron J 92: 436–444.

53. Muchow RC, Davis R (1988) Effect of nitrogen supply on the comparative productivity of maize and sorghum in a semi-arid tropical environment ii. radiation interception and biomass accumulation. Field Crop Res 18: 17–30.

54. McGarrahan JP, Dale RF (1984) A trend toward a longer grain-filling period for corn: a case study in Indiana. Agron J 76: 518–522.

55. Muchow R (1990) Effect of high temperature on grain-growth in field-grown maize. Field Crop Res 23: 145–158.

bioRxiv preprint doi: https://doi.org/10.1101/014100; this version posted January 26, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

56. Nielsen RL, Thomison PR, Brown GA, Halter AL, Wells J, et al. (2002) Delayed planting effects on flowering and grain maturation of dent corn. Agron J 94: 549–558.

57. Habier D, Fernando R, Kizilkaya K, Garrick D (2011) Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12: 186.

58. R Core Team (2014) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

59. Technow F (2013) hypred: Simulation of genomic data in applied genetics. R package version 0.4.

60. Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4: 250-255.

61. Xu S (2007) An empirical Bayes method for estimating epistatic effects of quantitative trait loci. Biometrics 63: 513–521.

62. Sun X, Ma P, Mumm RH (2012) Nonparametric method for genomics-based prediction of performance of quantitative traits involving epistasis in plant breeding. PLoS ONE 7: e50604.

63. Howard R, Carriquiry AL, Beavis WD (2014) Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. G3 4: 1027-1046.

64. Resende MF, Munoz P, Acosta JJ, Peter GF, Davis JM, et al. (2012) Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. New Phytol 193: 617–624.

65. Renton M (2011) How much detail and accuracy is required in plant growth sub-models to address questions about optimal management strategies in agricultural systems? AoB Plants 2011: plr006.

66. Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MP (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J R Soc Interface 6: 187–202.

67. Liepe J, Barnes C, Cule E, Erguler K, Kirk P, et al. (2010) ABC-SysBio approximate Bayesian computation in Python with GPU support. Bioinformatics 26: 1797–1799.

68. Brun F, Wallach D, Makowski D, Jones JW (2006) Working with dynamic crop models: Evaluation, analysis, parameterization, and applications. Amsterdam: Elsevier.

69. Curry GL, Feldman RM, Sharpe PJH (1978) Foundations of stochastic development. J Theor Biol 74: 397–410.

70. Wallach D, Keussayan N, Brun F, Lacroix B, Bergez JE (2012) Assessing the uncertainty when using a model to compare irrigation strategies. Agron J 104: 1274–1283.

71. Sisson SA, Fan Y, Tanaka MM (2007) Sequential Monte Carlo without likelihoods. Proc Nat Acad Sci 104: 1760-1765.

bioRxiv preprint doi: https://doi.org/10.1101/014100; this version posted January 26, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

72. Peters GW, Fan Y, Sisson SA (2012) On sequential Monte Carlo, partial rejection control and approximate Bayesian computation. Stat Comput 22: 1209–1222.

73. Bengtsson T, Bickel P, Li B (2008) Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems. In: IMS Collections Probability and Statistics: Essays in Honor of David A. Freedman, Institute of Mathematical Statistics, volume 2. pp. 316–334.

74. Marjoram P, Molitor J, Plagnol V, Tavaré S (2003) Markov chain Monte Carlo without likelihoods. Proc Natl Acad Sci 100: 15324-15328.

75. Buyya R, Yeo CS, Venugopal S (2008) Market-oriented cloud computing: vision, hype, and reality for delivering IT services as computing utilities. In: High Performance Computing and Communications, 2008. HPCC '08. 10th IEEE International Conference on. pp. 5-13.

76. Reymond M, Muller B, Leonardi A, Charcosset A, Tardieu F (2003) Combining quantitative trait Loci analysis and an ecophysiological model to analyze the genetic variability of the responses of maize leaf growth to temperature and water deficit. Plant Physiol 131: 664–675.

77. Bogard M, Ravel C, Paux E, Bordes J, Balfourier F, et al. (2014) Predictions of heading date in bread wheat (Triticum aestivum L.) using QTL-based parameters of an ecophysiological model. J Exp Bot 10.1093/jxb/eru328.

78. Yin X, Struik PC, van Eeuwijk FA, Stam P, Tang J (2005) QTL analysis and QTL-based prediction of flowering phenology in recombinant inbred lines of barley. J Exp Bot 56: 967–976.

79. Welcker C, Boussuge B, Bencivenni C, Ribaut JM, Tardieu F (2007) Are source and sink strengths genetically linked in maize plants subjected to water deficit? A QTL study of the responses of leaf growth and of Anthesis-Silking Interval to water deficit. J Exp Bot 58: 339–349.

80. Tardieu F, Reymond M, Muller B, Granier C, Simonneau T, et al. (2005) Linking physiological and genetic analyses of the control of leaf growth under changing environmental conditions. Crop and Pasture Sci 56: 937–946.

81. Chenu K, Chapman SC, Tardieu F, McLean G, Welcker C, et al. (2009) Simulating the yield impacts of organ-level quantitative trait loci associated with drought response in maize: a "gene-to-phenotype" modeling approach. Genetics 183: 1507–1523.

82. Dong Z, Danilevskaya O, Abadie T, Messina C, Coles N, et al. (2012) A gene regulatory network model for floral transition of the shoot apex in maize and its dynamic modeling. PLoS ONE 7: e43450.

83. Guo M, Rupe MA, Wei J, Winkler C, Goncalves-Butruille M, et al. (2014) Maize ARGOS1 (ZAR1) transgenic alleles increase hybrid maize yield. J Exp Bot 65: 249–260.

84. Habben JE, Bao X, Bate NJ, DeBruin JL, Dolan D, et al. (2014) Transgenic alteration of ethylene biosynthesis increases grain yield in maize under field drought-stress conditions. Plant Biotechnol J 12: 685–693.

85. Keurentjes JJB (2009) Genetical metabolomics: closing in on phenotypes. Curr Opin Plant Biol 12: 223–230.

bioRxiv preprint doi: https://doi.org/10.1101/014100; this version posted January 26, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

86. Fernie AR, Schauer N (2009) Metabolomics-assisted breeding: a viable option for crop improvement? Trends Genet 25: 39–48.

87. Schuster S, Fell DA, Dandekar T (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. Nat Biotechnol 18: 326–332.

88. Pilalis E, Chatziioannou A, Thomasset B, Kolisis F (2011) An in silico compartmentalized metabolic model of brassica napus enables the systemic study of regulatory aspects of plant central metabolism. Biotechnol and Bioeng 108: 1673–1682.

89. Simons M, Saha R, Guillard L, Clément G, Armengaud P, et al. (2014) Nitrogen-use efficiency in maize (Zea mays L.): from 'omics' studies to metabolic modelling. J Exp Bot 65: 5657–5671.

90. Saha R, Suthers PF, Maranas CD (2011) Zea mays RS1563: A comprehensive genome-scale metabolic reconstruction of maize metabolism. PLoS ONE 6: e21784.

91. Maher B (2008) Personal genomes: The case of the missing heritability. Nature 456: 18–21.