Integrating crop growth models with whole genome prediction through approximate Bayesian computation
Frank Technow1*, Carlos D. Messina2, L. Radu Totir1, Mark Cooper2
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 (GxE), 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 impact of 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 GxE 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 synthetic data sets. 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 and novel approach to improving prediction accuracy for some of the most challenging scenarios in plant breeding and applied genetics.

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]).
In-silico phenotypic prediction, enabled by dynamic crop growth models (CGMs), dates back to the late 1960’s (van Ittersum et al. [16]) and it has constantly evolved through inclusion of scientific advances made in plant 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 CGMs 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 among the underlying component traits is nonlinear, 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 among traits by Richey [36] and later quantified by Melchinger et al. [37]. CGMs, which explicitly model these nonlinear 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 re-evaluation (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 for practical applications of breeding theory. 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 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 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 were 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 synthetic data sets, 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 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 the calculation of trait values 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, 23] (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 ($40.08^\circ$ N, $88.24^\circ$ 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). The sowing date in 2012 was April 15th and in 2013 it was May 15th. We modified the original CGM of Muchow et al. [50] by enforcing a maximum length of the growing season, after which crop growth simulation was terminated, regardless of whether the genotype reached full physiological maturity or not. The length of the growing season in 2012 was 120 days from sowing and in 2013 it was 130 days from sowing. Both durations are within the range typically observed in the US Corn Belt (Neild and Newman [57]). In 2012 the plant population was 8 plants $m^{-2}$ and in 2013 the plant population was 10 plants $m^{-2}$. The 2012 and 2013 environments therefore differed not only in temperature and solar radiation but also in management practices. The temperature and solar radiation from date of sowing is shown in Fig. 1. Typical total biomass and grain yield development curves for early, intermediate and late maturing genotypes in the 2012 and 2013 environments are shown in Fig. 2 and corresponding curves for development of total and senescent leaf area in Fig. S1.

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

$$F(y_{TLN}, y_{SRE}, y_{AM}, y_{MTU}, \Omega_k)$$

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

Figure 2. Simulated development of total biomass and grain yield. The early, intermediate and late maturing genotypes had a total leaf number (TLN) of 6, 14.5 and 23, respectively. The values for the other three traits were 750 for AM, 1.6 for SRE and 1150 for MTU and in common for all genotypes. The full and dotted vertical lines indicate the end of the 2012 and 2013 growing season, respectively.

where $y_{TLN}$, etc. are the values of the physiological traits observed for the $i^{th}$ genotype and the weather and management data of environment $k$ are 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 $k$.

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 $\text{Model}(y^*_{ik} \mid \theta)$ 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 $\text{Model}(y^*_{ik} \mid \theta)$, to which a Gaussian noise variable distributed as $N(0, \sigma^2)$ is added as a stochastic component. If $\text{Model}(y^*_{ik} \mid \theta)$ 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

\[ y_{TLN} = \mu_{TLN} + z_i u_{TLN} \]
\[ y_{AM} = \mu_{AM} + z_i u_{AM} \]
\[ y_{SRE} = \mu_{SRE} + z_i u_{SRE} \]
\[ y_{MTU} = \mu_{MTU} + z_i u_{MTU}, \]

where $z_i$ is the genotype vector of the observed biallelic single nucleotide polymorphism (SNP) markers of genotype $i$, $\mu_{TLN}$ etc. denote the intercepts and $u_{TLN}$ etc. the marker effects. For brevity, we will use $\theta$ to denote the joint parameter vector $[\mu_{TLN}, \ldots, \mu_{MTU}, u_{TLN}, \ldots, u_{MTU}]$. We used independent Normal distribution priors for all components of $\theta$. The prior for $\mu_{TLN}$ was $N(m_{TLN}, \sigma^2_{\mu_{TLN}})$. To simulate imperfect prior information, we drew the prior mean $m_{TLN}$ from a Uniform distribution over the interval $[0.8 \cdot TLN, 1.2 \cdot TLN]$, where $TLN$ is the observed population mean of TLN. The average difference between $m_{TLN}$ and $TLN$ 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 $u_{TLN}$ was $N(0, \sigma^2_{u_{TLN}})$, which corresponds to the BayesC prior (Habier et al. [58]). In BayesC, the prior variance of marker effects $\sigma^2_{u_{TLN}}$, 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($y_{ik} | \theta$), 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. As distance measure between the simulated and observed data we used the Euclidean distance. 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 WGP method that incorporates the CGM as CGM-WGP. The CGM-WGP algorithm was implemented as a C routine integrated with the R software environment (R Core Team [59]).

**Synthetic data set**

To test the performance of CGM-WGP, 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 [60]) 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.
Table 1. Pseudocode of ABC rejection sampling algorithm

while \( x \leq \text{no. posterior samples} \) do
    while \( d > \epsilon \) do
        draw candidate \( \theta^* \) from prior(\( \theta \))
        for all \( i = 1, 2, \ldots, N \) do
            generate simulated data \( y_{ik}^* \) from Model(\( y_{ik}^* \mid \theta^* \))
        end for
        compute \( d = \sqrt{\sum_{i=1}^{N} (y_{ik} - y_{ik}^*)^2} \)
    end while
    accept and store \( \theta^* \)
    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. 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

\[
y_{ik} = F(\cdot)_{ik} + \epsilon_{ik},
\]

where \( \epsilon_{ik} \) is a Gaussian noise variable with mean zero and variance \( \nu_e \). The value of \( \nu_e \) was chosen such that the within-environment heritability of \( y_{ik} \) was equal to 0.85. We generated 50 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
(GBLUP, Meuwissen et al. [1]). The model is

\[ y_{ik} = \beta_0 + z_i u + e_i \]  

(4)

where \( \beta_0 \) is the intercept, \( u \) the vector of marker effects and \( e_i \) a residual. As before, \( z_i \) denotes the marker genotype vector. The GBLUP model was fitted with the R package rrBLUP (Endelman [61]). 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 \( \beta_0 + z_i u \). Note that because the conventional GBLUP model does not utilize information about the environment for which predictions are made, observed and new environment predictions are identical.

Results and Discussion

Predicting performance in observed environments

The accuracy of observed environment predictions achieved by CGM-WGP was considerably larger than that of the benchmark method GBLUP in both environments (Table 2, Fig. 3, Fig. S2). This superiority of CGM-WGP 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 nonlinear functional relationships between the physiological traits and grain yield, which was particularly pronounced for TLN (Fig. 4).

TLN is closely related with the maturity rating of genotypes (Muchow et al. [50]). The higher it is, the later the onset of the reproductive phase and the later the maturity. Late genotypes have a higher yield potential than earlier genotypes because of a greater leaf area (Fig. S1). However, if the growing season is too short, they cannot realize this yield potential because of their slower development and later onset of the generative phase (Fig. 2). Very early genotypes on the other hand, have a low leaf area and do not make use of the full growing season. As a consequence, their realized yield is low, too. The relationship between TLN and grain yield therefore follows an optimum curve (Fig. 4). This was particularly pronounced in 2012, which had the shorter growing season and therefore penalized the late maturing genotypes more. The more decidedly nonlinear relationship between grain yield and TLN in 2012 also explains why the difference in prediction accuracy between CGM-WGP and GBLUP was greater in this season than in 2013 (0.23 points in 2012 compared to 0.13 points in 2013, on average).

Table 2. Accuracy of grain yield predictions of DH lines in the test set

<table>
<thead>
<tr>
<th>Estimation Env.</th>
<th>Prediction Env.</th>
<th>CGM-WGP</th>
<th>GBLUP</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>2012</td>
<td>0.77</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>2013</td>
<td>0.48</td>
<td>0.10</td>
</tr>
<tr>
<td>2013</td>
<td>2012</td>
<td>0.42</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>2013</td>
<td>0.75</td>
<td>0.62</td>
</tr>
</tbody>
</table>

Prediction accuracy for grain yield of DH lines in the test set, averaged over 50 replications. All differences within a row are statistically significant at a significance level of < 0.005.

The scenario we studied is an example of 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]). We accounted for nonlinear functional relationships among traits with the CGM. This enabled us to capture biological epistasis through simple
Figure 3. Predicted vs. observed grain yield of 1500 DH lines in testing set for prediction methods CGM-WGP (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 observed environment predictions was 0.83 (CGM-WGP) and 0.69 (GBLUP). For new environment predictions it was 0.39 (CGM-WGP) and 0.11 (GBLUP).

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, Fig. S2). The average prediction accuracy for performance in 2012 when using the 2013 estimation environment was 54% (CGM-WGP) and 15% (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 64% (CGM-WGP) and 16% (GBLUP). Thus, CGM-WGP still delivered a decent accuracy for predicting performance in new environments, while GBLUP largely failed in this task. The prediction accuracy of GBLUP was in fact negative, and sometimes strongly so, for close to 50% of the synthetic data sets (Fig. S2). For CGM-WGP negative accuracies were observed in only 14% (2012) and 4% (2013) of the cases.

The rank correlation between true performance in 2012 and 2013 was only 0.54 (averaged over 50 synthetic data sets), which indicated the presence of considerable G×E interactions, including changes in rank (Fig. S3). The interaction between the environment and TLN again explains the occurrence of G×E to a large degree. In the
shorter 2012 season, the late maturing genotypes cannot realize their growth and yield potential and are outperformed by the genotypes with early and intermediate maturity (Fig. 2 and Fig. 4). In the 10 day longer growing season of 2013, however, the late maturing genotypes can realize their greater yield potential better and outperform the early maturing genotypes and have a similar performance as genotypes with intermediate maturity. This dynamic leads to cross over G\times E interactions between the 2012 and 2013 environments.

That new environment prediction under the presence of G\times E interaction is considerably less accurate than observed environment prediction was expected and already observed in other studies (Resende et al. [65], Windhausen et al. [11]). It is encouraging that the reduction in accuracy for CGM-WGP was considerably less severe than for the conventional benchmark method GBLUP because this indicates that the former method did succeed in predicting G\times E interactions to some degree.

Predicting G\times E interactions in new environments for which no yield data are available, 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. CGM-WGP takes this approach a step further by making the CGM and the environmental data that inform it, an integral part of the estimation procedure.

Nonetheless, 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 performed 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\times E interaction effects are expected to be particularly strong. CGMs can help to identify such environments and to 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 yield testing across the whole breadth (Cooper et al. [2]). Predicting performance in new environments will therefore always be required and novel methods like CGM-WGP 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 CGM used in the estimation process fully represented the systematic component of
the data generating process, besides the random noise. 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 any desired degree of accuracy (Keating et al. [17], Renton [66], 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, tools are principally available for applying CGM-WGP in more complex scenarios than the one addressed in this study.

Stochastic CGMs There are examples of the use of fully deterministic model operators in ABC (Toni et al. [67], Liepe et al. [68]). 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 vast majority of CGMs are deterministic (Keating et al. [17], van Ittersum et al. [16]), there are examples of stochastic CGMs (Brun et al. [69]). In addition to incorporating inherently stochastic processes of development (Curry et al. [70]), stochastic CGMs could also serve to account for uncertainty in the parameters of the functional equations comprising the model (Wallach et al. [71]).

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. [72], Toni et al. [67], Peters et al. [73]). 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. [74]). In the context of sequential Monte Carlo, this is known as particle depletion and was addressed by Peters et al. [73]. 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. [74]).

Another interesting development is MCMC-ABC, which incorporates ABC with the Metropolis-Hastings algorithm (Marjoram et al. [75]). 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. [75]). The computation time thus scales linearly to the number of processors.
available. Using the ABC rejection algorithm therefore allowed us to fully leverage the high performance computing cluster of DuPont Pioneer. In the era of cloud computing (Buyya et al. [76]), 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. [69]). 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 (bias) between the prior parameter used and the true population parameter was 10%. When we increased the relative discrepancy to 25% (i.e., a maximum discrepancy of 50%), prediction accuracy dropped somewhat (Table S1). The reduction was only slight for observed environment prediction but more pronounced for new environment prediction. However, CGM-WGP was still considerably more accurate than the benchmark GBLUP. Thus, CGM-WGP seems to be relatively robust to moderate prior miss specification, as long as the value range supported by the prior distribution is not out of scope. In the ideal case of no prior bias, on the other hand, new and observed environment prediction accuracy increased slightly as compared to a bias of 10%.

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. [77]; Bogard et al. [78]; Yin et al. [79]; Welcker et al. [80]; Tardieu et al. [81]). In fact, such component trait QTL have been successfully used to parametrize CGMs for studying genotype dependent response to environmental conditions (Chenu et al. [28]; Messina et al. [29] Bogard et al. [78]; Chenu et al. [82] Yin et al. [79]). Knowledge about the location and effect of such QTL, or of transgenes (Dong et al. [83]; Guo et al. [84]; Habben et al. [85]), could be incorporated as an additional source of prior information. Then, instead of estimating marker effects for the whole genome, CGM-WGP could focus on regions of particular importance, which reduces the dimensionality of the parameter space dramatically. Thus prior knowledge can be leveraged for improving prediction accuracy and computational efficiency of CGM-WGP.

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 [86]) and are therefore of particular interest in plant breeding (Fernie and Schauer [87]). Metabolic networks model the interrelationships between genes, intermediate metabolites and end products through biochemistry pathways (Schuster et al. [88]). Elaborate metabolic network models are available today that allow studying and simulating complex biochemical processes related to crop properties, such as flowering time, seed growth, nitrogen use efficiency and biomass composition (Dong et al. [83]; Pilalis et al. [89]; Simons et al. [90]; Saha et al. [91]). Liebe et al. [47] demonstrated how ABC can be used for parameter estimation with metabolic and other biochemical networks. Using the principles outlined here for CGM-WGP, 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 remains unaccounted for in genome-wide association studies (Maher [92]). 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 demonstrated 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. Yield is a product of plant genetics and physiology, the environment and crop management and integrating information pertaining to these components will ultimately enable us to better predict it [93]. While this study is only a first step and many questions remain, we conclude that CGM-WGP presents a promising and novel path forward towards a new class of WGP models that leverage previously unused sources of knowledge and thereby increase prediction accuracy in settings that have proved challenging for plant breeding and applied genetics.

Supporting Information

S1 Table

Table S1. Accuracy of grain yield predictions of test DH lines with increased bias in prior hyperparameters

S1 Figure

Figure S1. Simulated development of total and senescent leaf area. The early, intermediate and late maturing genotypes had a total leaf number (TLN) of 6, 14.5 and 23, respectively. The values for the other three traits were 750 for AM, 1.6 for SRE and 1150 for MTU and in common for all genotypes. The full and dotted vertical lines indicate the end of the 2012 and 2013 growing season, respectively.

S2 Figure

Figure S2. CGM-WGP vs. GBLUP prediction accuracy in 50 synthetic data sets.

S3 Figure

Figure S3. Distribution of simulated grain yield in 2012 and 2013 environments. The grey lines indicate the performance of specific genotypes in both environments. Data shown is from a representative example replication.

S4 Figure

Figure S4. Relationship between physiological traits and total grain yield. Data shown are a random sample of 1000 genotypes from a representative example replication.
Acknowledgments

References


