PLOS ONE

RESEARCH ARTICLE

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

* Frank.Technow@pioneer.com

Citation: Technow F, Messina CD, Totir LR, Cooper M (2015) Integrating Crop Growth Models with Whole Genome Prediction through Approximate Bayesian Computation. PLoS ONE 10(6): e0130855. doi:10.1371/journal.pone.0130855

Editor: Ive De Smet, University of Nottingham, UNITED KINGDOM

Received: February 3, 2015

Accepted: May 25, 2015

Published: June 29, 2015

Data Availability Statement: All relevant data and instructions to repeat the simulations are within the paper. The weather data can be obtained from www.sws.uiuc.edu/warm. A synthetic data set is also included as supplemental material.

Funding: The authors have no support or funding to report.

Competing Interests: The authors are employed by DuPont Pioneer, which owns a pending unpublished patent application covering concepts disclosed in the manuscript. The authors have no financial interest in the patent application. Part of the authors'

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 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 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 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 [1], enabled by whole genome prediction (WGP) methods, is revolutionizing plant breeding [2]. Since its inception, attempts to improve prediction accuracy have

employment responsibilities at DuPont Pioneer include developing systems and methods disclosed in the manuscript. The authors declare they are unaware of any competing interests. This does not alter the authors' adherence to PLOS ONE policies on sharing data and materials.

focused on: developing improved and specialized statistical models [3–6], increasing the marker density used [7–9], increasing the size and defining optimal designs of estimation sets [10–13] and better understanding the genetic determinants driving prediction accuracy [14, 15].

In-silico phenotypic prediction, enabled by dynamic crop growth models (CGMs), dates back to the late 1960's [16] and it has constantly evolved through inclusion of scientific advances made in plant physiology, soil science and micrometeorology [16, 17]. CGMs used in plant breeding are structured around concepts of resource capture, utilization efficiency and allocation among plant organs [18–21] and are used to: characterize environments [22, 23], predict consequences of trait variation on yield within a genotype × environment × management context [24], evaluate breeding strategies [25–27], and assess hybrid performance [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 [21, 28, 29]. 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 [21, 27, 30]. 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 [31–33]. Yield and other complex traits are the product of intricate interactions between component traits on lower hierarchical levels [19, 34–37]. 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 [33]. This phenomenon was first described for multiplicative relationships among traits by Richey [38] and later quantified by Melchinger et al. [39]. 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 [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 [40] or genetic and environmental covariances [41]. Later Jarquín et al. [42] and Heslot et al. [43] 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. [43] 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 [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 [44–47]. However, ABC methods were rapidly adopted in other scientific fields, such as ecology [48], systems biology [49] and hydrology [50]. Recently, Marjoram et al. [51] 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. [52], which models maize grain yield development as a function of plant population (PPOP, plants m^{-2}), daily temperature (°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] [52, 53], [700, 800] [52, 54], [1.5,1.7] [55] and [1050, 1250] [56–58], 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). The sowing date in 2012 was April 15th and in 2013 it was May 15th. We modified the original CGM of Muchow et al. [52] 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 [59]. In 2012 PPOP was 8 plants m^{-2} and in 2013 PPOP 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 S1 Fig.

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

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

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 are represented as Ω_k. To simplify notation, we will henceforth use $F(\cdot)_{ik}$ to represent the CGM and its inputs for genotype i in environment k.

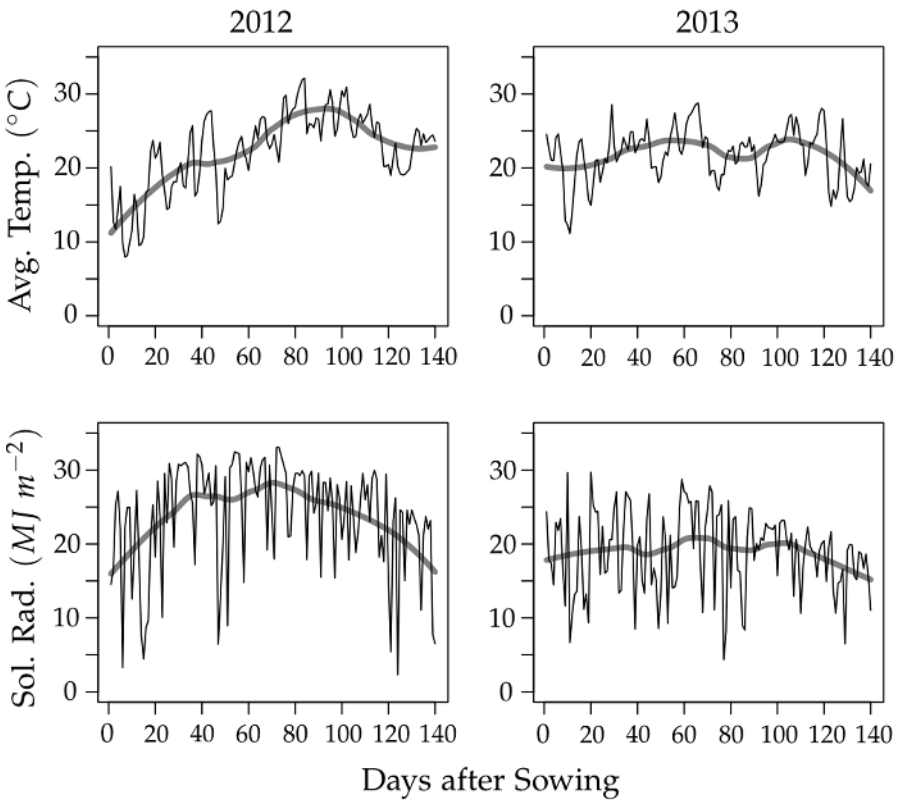

Fig 1. Daily average temperature and solar radiation at Champaign, Illinois in 2012 and 2013. The thick grey line shows a smoothed curve.

doi:10.1371/journal.pone.0130855.g001

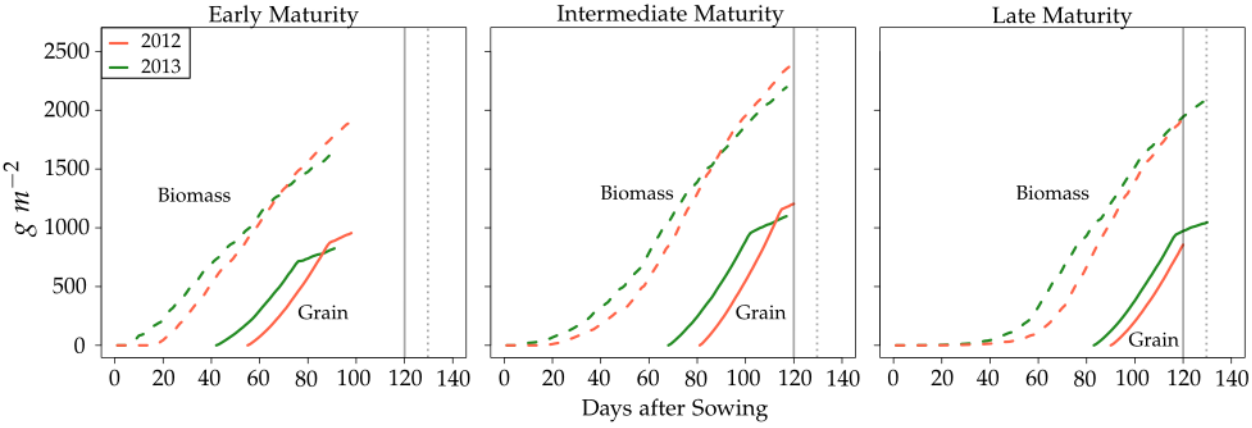

Fig 2. Simulated development of total biomass and grain yield. The early, intermediate and late maturing genotypes had a total leaf number (TLN) of 6, 14.5 and 23, respectively. The values for the other three traits were 750 for AM, 1.6 for SRE and 1150 for MTU and in common for all genotypes. The full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

doi:10.1371/journal.pone.0130855.g002

Approximate Bayesian Computation (ABC)

ABC replaces likelihood computation with a simulation step [44]. An integral component of any ABC algorithm is therefore the simulation model operator $\mathrm{Model}(y^*_{ik} \mid \theta)$ which generates simulated data y^*_{ik} given parameters θ. In our proof of concept study, the crop growth model $F(\cdot)_{ik}$ represents the deterministic component of $\mathrm{Model}(y^*_{ik} \mid \theta)$, to which a Gaussian noise variable distributed as $\mathcal{N}(0, \sigma^2_e)$ is added as a stochastic component. If $\mathrm{Model}(y^*_{ik} \mid \theta)$ 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 [50].

The weather and management data $\mathbf{\Omega}_k$ was assumed to be known, the physiological traits, however, were unknown and treated as latent or hidden variables, that were 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, μ_{TLN} etc. denote the intercepts and $\mathbf{u}_{TLN}$ etc. the marker effects. For brevity, we will use θ 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 θ. The prior for μ_{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 $\mathcal{N}(0, \sigma^2_{\mu_{TLN}})$, which corresponds to the *BayesC* prior [60]. In BayesC, the prior variance of marker effects $\sigma^2_{\mu_{TLN}}$, 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 $\sigma^2_{\mu_{TLN}}$ 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 σ^2_e, the variance of the Gaussian noise variable that is part of the model operator $\mathrm{Model}(y^*_{ik} \mid \theta)$, 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. As distance measure between the simulated and observed data we used the Euclidean distance. The tolerance level ϵ 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 [61].

Table 1. Pseudocode of ABC rejection sampling algorithm.

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

Basic ABC rejection sampling algorithm to sample from the approximate posterior distribution of θ.

doi:10.1371/journal.pone.0130855.t001

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 [62] 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.

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

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

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 50 synthetic data sets by repeating the whole process. An example synthetic data set is available as supplemental material (S1 Dataset).

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 θ samples, using the weather and management data $\mathbf{\Omega}_k$ 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 [1]). The model is

$$y_{ik} = \beta_0 + \mathbf{z}_i\mathbf{u} + e_i \quad (4)$$

where β_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 [63]. 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 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, S2 Fig). 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).

For any point in time (t) during the maize growth cycle, dry matter growth (DM_g) results from the interception of solar radiation (SR) and its conversion into mass with efficiency SRE. Light interception in turn depends on the size of the canopy, which is determined by the leaf area per plant (LAPP) and the plant population (PPOP), and the distribution of light within the canopy, which is modeled using a coefficient of light extinction (k). The relationship

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

Estimation Env.	Prediction Env.	CGM-WGP	GBLUP
2012	2012	0.77	0.54
	2013	0.48	0.10
2013	2012	0.42	0.08
	2013	0.75	0.62

Prediction accuracy for grain yield of DH lines in the test set, averaged over 50 replications.

doi:10.1371/journal.pone.0130855.t002

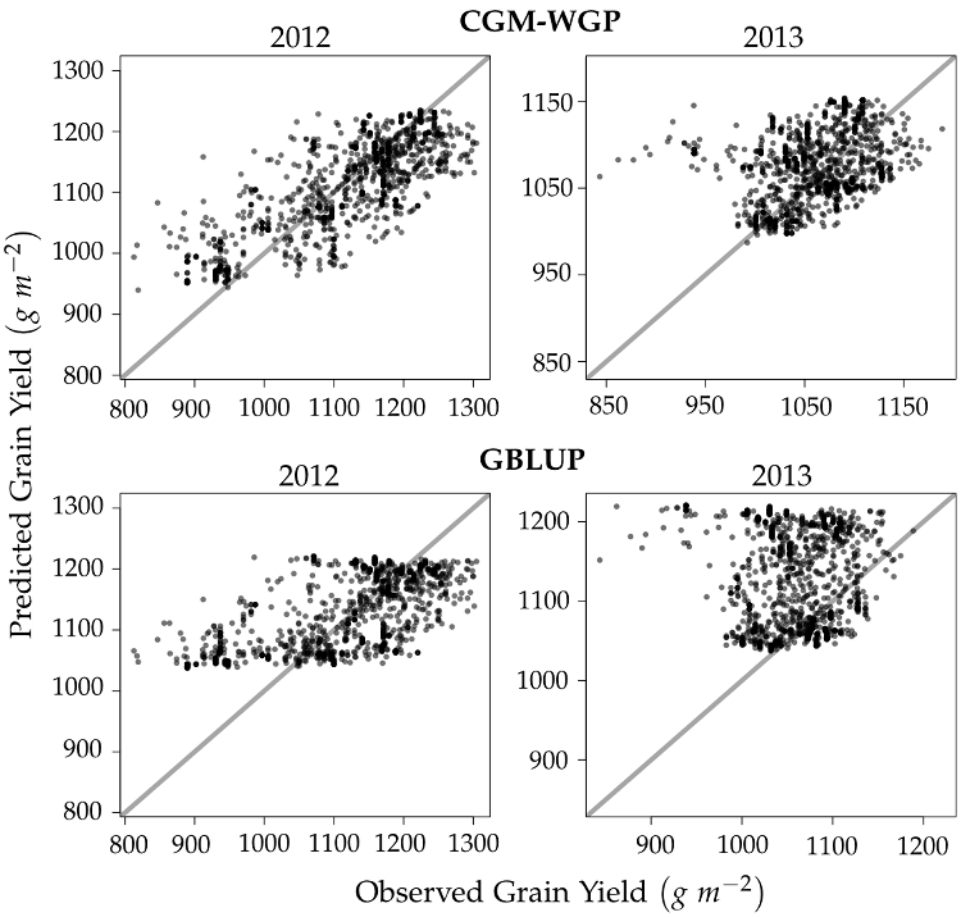

Fig 3. Predicted vs. observed grain yield of 1500 DH lines in testing set for prediction methods CGM-WGP (top row) and GBLUP (bottom row). The estimation environment was 2012. Results shown are from a representative example data set. In this example, the accuracy for observed environment predictions was 0.83 (CGM-WGP) and 0.69 (GBLUP). For new environment predictions it was 0.39 (CGM-WGP) and 0.11 (GBLUP).

doi:10.1371/journal.pone.0130855.g003

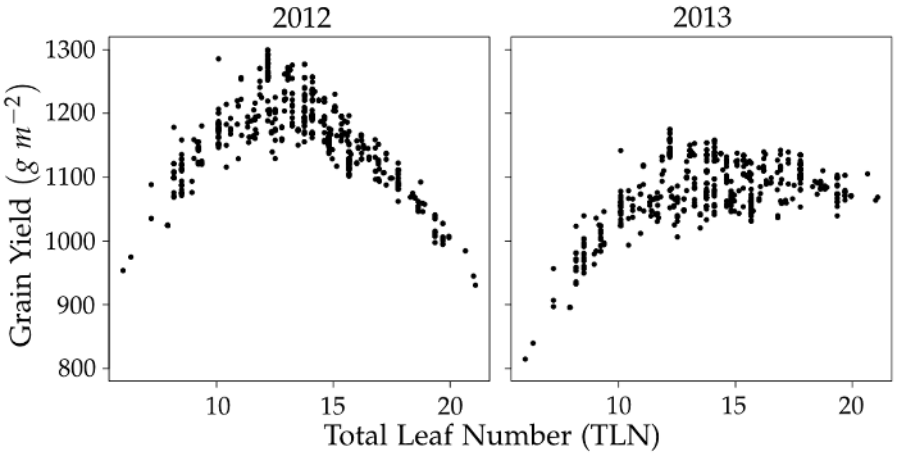

Fig 4. Relationship between total leaf number (TLN) and grain yield. Results shown are from a representative example data set.

doi:10.1371/journal.pone.0130855.g004

between DM_g, SR, SRE, LAPP and PPOP is non-linear [52]

$$DM_{g_t} = SR_t \times SRE \times (1 - e^{-k \times LAPP_t \times PPOP}). \quad (5)$$

Because LAPP is determined by TLN and AM [52], DM_g increases with increasing AM but only up to a point when canopy size maximizes light interception. Because grain yield is a fraction of the integral of DM_g over the growing season, there is a non-linear relationship between AM and grain yield, which can often be detected as a weak correlation (S3 Fig). From Eq (5) one can see that an increase in SRE is always beneficial for DM_g, which was reflected by the more or less linear relationship between SRE and grain yield (S3 Fig). The longer the length of the period between silking and physiological maturity, the more time the genotype has for grain filling. However, the end of the growing season can forestall exploitation of the longer grain filling periods of high MTU genotypes. Thus, increasing MTU beyond a point determined by the growing season length will have no further effect on grain yield. Such a saturation curve was indeed observed in 2013 (S3 Fig). The relationship between MTU and grain yield tended to be less clear in 2012, where many genotypes did not reach physiological maturity because of the shorter growing season. The relationships between grain yield and the physiological traits AM, SRE and MTU, were generally weak, however, and not obvious in all data sets. We will therefore focus on discussing the relationship between grain yield and TLN, which was very distinct and consistent.

TLN is closely related with the maturity rating of genotypes [52]. 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 (S1 Fig). 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).

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 [33]. We accounted for nonlinear 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 nonlinear marker effects to the final trait of interest [64–66]. 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, CGM-WGP 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, S2 Fig). 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 (S2 Fig). For CGM-WGP negative accuracies were observed in only 14% (2012) and 4% (2013) of the cases.

The genetic 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 (S4 Fig). A genetic correlation of 0.54 between environments is within the range typically observed in plant breeding data sets [67] and crossover interaction between environments is a common phenomenon in plant breeding [34, 68, 69].

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 (Figs 2 and 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 crossover G×E interactions between the 2012 and 2013 environments.

That new environment prediction under the presence of G×E interaction is considerably less accurate than observed environment prediction was expected and already observed in other studies [11, 70]. 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×E interactions to some degree.

Predicting G×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. [42] accomplished this by fitting statistical interactions between markers and environmental covariates. A similar approach was taken by Heslot et al. [43], 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×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 [27, 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 [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.

Areas of further research and development

Alternatives to CGM-WGP. With continual technology improvements for phenotyping traits it is becoming increasingly feasible to assay phenotypic variation for many of the physiological traits underlying the CGM [19]. There would then be no need to treat them as latent, hidden variables as in CGM-WGP. Such improved precision phenotyping capabilities thus open up possible alternatives and extensions to the CGM-WGP methodology introduced here.

One alternative that can be considered is a two-step procedure, in which (1) physiological traits are predicted based on QTL identified in dedicated mapping experiments and (2) the so obtained physiological trait values are used to parametrize CGMs and predict the expected yield performance of novel genotypes in the same or different environments [71–73]. Using WGP instead of QTL mapping in step 1, could further enhance that procedure.

One shortcoming of this approach is that all relevant physiological traits have to be measured for all genotypes in the estimation set. This may prove to be unfeasible in practice, particularly when done on an industrial scale (i.e., for many populations and repeated year after year). The situation is exacerbated when more sophisticated models like APSIM [74], which can model plant-soil interactions related to water and nutrient uptake, are used. The set of relevant physiological traits for these CGMs includes root traits for example [75], which are particularly difficult to measure in a routine, high-throughput fashion [76].

A key novelty of CGM-WGP is that it can accommodate partially or fully unobserved physiological traits by treating them as hidden variables. It could thus facilitate incorporating a CGM in WGP even when phenotyping all relevant physiological traits is not feasible.

However, CGM-WGP and the described two-step approach have a common objective, which is to apply a suitable CGM to capture non-linear relationships among traits and the environment to succeed in the crucial but challenging task of predicting yield in future environments. At this stage it is premature to suggest one approach ahead of these or other possibilities. However, the results of the present study indicate that there are opportunities to improve predictions for quantitative traits influenced by epistasis and G × E interaction, through effective integration of appropriate CGMs into the genetic prediction methodology.

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 a high degree of accuracy [17, 30, 77]. Advanced CGMs such as *APSIM*[74], 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.

With multiple possible CGMs to choose from, model selection becomes an issue. The ABC algorithm underlying CGM-WGP could in principle be used to perform model selection simultaneously with parameter estimation [49, 51, 78]. It could thus provide a statistically formal way of comparing the fit of several CGMs.

Stochastic CGMs. There are examples of the use of fully deterministic model operators in ABC [78, 79]. However, with fully deterministic model operators the sampled distribution would not converge to the true posterior when the tolerance level ϵ goes to zero [50] 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 [50], 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 [16, 17], there are examples of stochastic CGMs [80]. In addition to incorporating inherently stochastic processes of development [81], stochastic CGMs could also serve to account for uncertainty in the parameters of the functional equations comprising the model [82].

Advanced ABC algorithms. For this proof of concept study we used the basic ABC rejection sampling algorithm [44, 45]. 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 [78, 83, 84]. 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 [85]. In the context of sequential Monte Carlo, this is known as particle depletion and was addressed by Peters et al. [84]. 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 [85].

Another interesting development is *MCMC-ABC*, which incorporates ABC with the Metropolis-Hastings algorithm [86]. 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 [86]. 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 [87], 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 [80]. 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 (S1 Table). 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 the synthetic data sets we generated, the component traits were controlled by QTL with independent effects and thus uncorrelated. In practice, however, the traits might be correlated, because of pleiotropic QTL, for example. In this situation, marker effects are correlated too. We modeled the marker effects as *a priori* independent and note that this does not preclude posterior correlation if CGM-WGP in its current implementation is applied to a scenario with correlated component traits. However, it is possible to model whole genome marker effects as *a priori* correlated. This was explored for the purpose of fitting WGP models with breed specific

but correlated marker effects in animal breeding [88]. Modeling a correlation structure could allow information sharing across traits. It could also improve computational efficiency of CGM-WGP, because the prior distribution would be closer to the posterior and thereby result in fewer rejected samples.

Lastly, CGM-WGP can be modified to use 'BayesB' [1] or 'BayesCπ' [60] as priors of marker effects, which would allow marker effects to be exactly zero. By allowing the effects of markers to be zero, marker effect estimation and implicit SNP model selection are done simultaneously. This could present an interesting compromise between a continuous WGP approach (BayesC) and a discrete, QTL based approach to prediction.

Number of markers. We applied CGM-WGP to a biparental population, which is by far the most common population type in commercial plant breeding programs [89]. Previous WGP studies found that marker density is typically not the most limiting factor in these type of populations and that densities achievable with around 200 genomwide markers suffice for accurate predictions [13, 90–92]. This proof of concept showed that computations for CGM-WGP are feasible for 100 markers. Thus, while more challenging, we expect that computations can be facilitated for the numbers of markers required for biparental populations. However, applying CGM-WGP to data sets with tens of thousands of markers is likely not possible with current ABC algorithms, in particular if more sophisticated CGMs are used that require specification of more physiological traits. Technow and Melchinger [12] showed that using a more realistic but complex WGP model with a lower marker density can result in a higher prediction accuracy than using a less realistic WGP model with a higher marker density. Thus, the greater realism of CGM-WGP might compensate for the fact that it can currently be applied only at low to intermediate marker densities.

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 [71, 93–96]. In fact, such component trait QTL have been successfully used to parametrize CGMs for studying genotype dependent response to environmental conditions [28, 29, 73, 94, 95]. Knowledge about the location and effect of such QTL, or of transgenes [97–99], 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 genomic regions of particular importance. This reduces the dimensionality of the parameter space dramatically and enables CGM-WGP to be used in settings that traditionally required high marker densities, such as WGP in diverse germplasm [100].

Identifiability. It is possible that the CGM generates the same yield for two or more sets of component trait values. There are often several possible biological strategies with equivalent outcome, so this does not necessarily indicate model missspecification. In this situation, however, it is not possible to identify from the observed yield data alone which set of trait values is more appropriate. By extension, the same applies to the sets of marker effects of which the component trait values are linear functions. This is referred to as likelihood nonidentifiability (short 'nonidentifiability') and is a known problem in biological modeling with hidden variables [101]. When analyses are conducted under the Bayesian statistical paradigm, nonidentifiability does not necessarily preclude inference and estimation, because informative prior distributions can identify the parameters nonetheless [102, 103]. This is another argument in favor of using informative prior distributions. Nonidentifiability also does not preclude prediction, because the posterior predictive distribution is a function that is identified even if the parameters are not [104]. If prediction is the sole purpose, the nonidentified parameters can be viewed as nuisance variables [101], that are averaged over in the posterior predictive distribution. In fact, Gianola [104] argues that in a predictive setting, parameters are merely 'tools

enabling one to go from past to future observations'. However, nonidentifiability can be associated with computational problems [102] and is of course an issue if the latent variables are of interest themselves. A special case of nonidentifiability occurs when the parameters are not identifiable for the estimation data set at hand, out of sheer coincidence [101]. However, when applied to new observations, i.e., for prediction, the parameters might be identifiable, with one set of parameters being more appropriate than the others. If this is the case, nonidentifiability might lead to a reduction in prediction accuracy.

In addition to using informative prior distributions, increasing the informativeness of the data with respect to the parameters is a direct way to improve their identifiability. In our case, this can be achieved by using additional response variables next to final grain yield. One possible choice is to use grain yield measurements from multiple environments, because several sets of component trait values might generate the same grain yield in one environment, but not in the others. Other possible choices are intermediate traits generated by the CGM, such as early biomass development or leaf area index, which can be measured non-destructively and with high-throughput [105–107].

Actually measuring the underlying physiological traits obviates the need to treat them as latent, hidden variables. This would obviously guarantee identifiability. As mentioned before, it might be possible to measure at least some of the traits. If these are key traits in the development of grain yield, observing them would identify the unobserved traits, too. One way of exploiting the information from observed physiological traits in CGM-WGP is to treat them as constants in the estimation procedure. In this framework, physiological trait values of new genotypes have to be predicted from conventional QTL or WGP models, as described above. However, CGM-WGP could also be extended to estimate marker effects for observed and hidden physiological traits simultaneously.

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 [108] and are therefore of particular interest in plant breeding [109]. Metabolic networks model the interrelationships between genes, intermediate metabolites and end products through biochemistry pathways [110]. 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 [97, 111–113]. Liepe et al. [49] 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 [114]. Marjoram et al. [51] 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 [115]. While this study is only a first step and many questions remain, we conclude that CGM-WGP presents a promising novel path forward towards a new class of WGP models that integrate genomics, quantitative genetics, and systems biology and thereby increase prediction accuracy in settings that have proved challenging for plant breeding and applied genetics.

Supporting Information

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

S1 Dataset. SNP genotypes and trait phenotypes of synthetic maize doubled haploid lines. The data set is a representative example of the 50 synthetic data sets used in the study.
(CSV)

S1 Fig. 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.
(PDF)

S2 Fig. CGM-WGP vs. GBLUP prediction accuracy in 50 synthetic data sets.
(PDF)

S3 Fig. Relationship between physiological traits and total grain yield. Data shown are a random sample of 1000 genotypes from a representative example replication.
(PDF)

S4 Fig. 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.
(PDF)

Author Contributions

Conceived and designed the experiments: FT CDM LRT MC. Performed the experiments: FT. Analyzed the data: FT. Contributed reagents/materials/analysis tools: FT CDM LRT MC. Wrote the paper: FT CDM LRT MC.

References

1. Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829. PMID: 11290733
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 Pasture Sci 64: 311–336. doi: 10.1071/CP14007
3. Yang W, Tempelman RJ (2012) A Bayesian antedependence model for whole genome prediction. Genetics 190: 1491–1501. doi: 10.1534/genetics.111.131540 PMID: 22135352
4. Heslot N, Yang HP, Sorrells ME, Jannink JL (2012) Genomic selection in plant breeding: a comparison of models. Crop Sci 52: 146–160. doi: 10.2135/cropsci2011.09.0297
5. Kärkkäinen HP, Sillanpää MJ (2012) Back to basics for Bayesian model building in genomic selection. Genetics 191: 969–987. doi: 10.1534/genetics.112.139014 PMID: 22554888
6. Technow F, Melchinger AE (2013) Genomic prediction of dichotomous traits with Bayesian logistic models. Theor Appl Genet 126: 1133–1143. doi: 10.1007/s00122-013-2041-9 PMID: 23385660
7. Meuwissen T, Goddard M (2010) Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics 185: 623–631. doi: 10.1534/genetics.110.116590 PMID: 20308278
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. doi: 10.3168/jds.2011-5019 PMID: 22720968

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. doi: 10.1371/journal.pgen.1002685 PMID: 22570636
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. doi: 10.1534/genetics.112.141473 PMID: 22865733
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. doi: 10.1534/g3.112.003699 PMID: 23173094
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. doi: 10.1534/g3.112.004630 PMID: 23390596
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 54: 1476–1488. 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. doi: 10.1534/genetics.110.116855 PMID: 20407128
15. Habier D, Fernando RL, Garrick DJ (2013) Genomic-BLUP decoded: a look into the black box of genomic prediction. Genetics 194: 597–607. doi: 10.1534/genetics.113.152207 PMID: 23640517
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. doi: 10.1016/S1161-0301(02)00106-5
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. doi: 10.1016/S1161-0301(02)00108-9
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. doi: 10.1016/j.pbi.2009.01.006 PMID: 19282235
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. doi: 10.1016/j.tplants.2006.10.006 PMID: 17092764
20. Passioura JB (1983) Roots and drought resistance. Agr Water Manage 7: 265–280. doi: 10.1016/0378-3774(83)90089-6
21. Yin X, Struik PC, Kropff MJ (2004) Role of crop physiology in predicting gene-to-phenotype relationships. Trends Plant Sci 9: 426–432. doi: 10.1016/j.tplants.2004.07.007 PMID: 15337492
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. doi: 10.1071/AR99108
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. doi: 10.2135/cropsci2004.0370
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. doi: 10.2135/cropsci2008.03.0152
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. doi: 10.2134/agronj2003.0099
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. doi: 10.1093/jxb/erq329 PMID: 21041371
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. doi: 10.1111/j.1365-3040.2007.01772.x PMID: 18088328

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. doi: 10.2135/cropsci2005.04-0372

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. doi: 10.1093/jxb/erq095 PMID: 20400531

31. Wolf DP, Hallauer AR (1997) Triple testcross analysis to detect epistasis in maize. Crop Sci 37: 736–770. doi: 10.2135/cropsci1997.0011183X003700030012x

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. Allard RW, Bradshaw AD (1964) Implications of genotype-environmental interactions in applied plant breeding. Crop Sci 4: 503–508. doi: 10.2135/cropsci1964.0011183X000400050021x

35. Cooper M, Chapman SC, Podlich D, Hammer G (2002) The GP problem: quantifying gene-to-phenotype relationships. In Silico Biol 2: 151–164. PMID: 12066839

36. Slafer G (2003) Genetic basis of yield as viewed from a crop physiologist's perspective. Ann Appl Biol 142: 117–128. doi: 10.1111/j.1744-7348.2003.tb00237.x

37. 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. doi: 10.1073/pnas.1120813109 PMID: 22615396

38. Richey FD (1942) Mock-dominance and hybrid vigor. Science 96: 280–281. doi: 10.1126/science.96.2490.280 PMID: 17840481

39. 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. doi: 10.1007/BF00223643 PMID: 24186017

40. 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. doi: 10.1111/pbr.12105

41. 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.

42. 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. doi: 10.1007/s00122-013-2243-1 PMID: 24337101

43. 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. doi: 10.1007/s00122-013-2231-5 PMID: 24264761

44. Tavare S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. Genetics 145: 505–518. PMID: 9071603

45. 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. doi: 10.1093/oxfordjournals.molbev.a026091 PMID: 10605120

46. Csilléry K, Blum MG, Gaggiotti OE, François O (2010) Approximate Bayesian Computation (ABC) in practice. Trends Ecol Evol 25: 410–418. doi: 10.1016/j.tree.2010.04.001 PMID: 20488578

47. Lopes JS, Beaumont MA (2010) ABC: A useful Bayesian tool for the analysis of population data. Infect Genet Evol 10: 825–832. doi: 10.1016/j.meegid.2009.10.010

48. Lawson Handley LJ, Estoup A, Evans DM, Thomas CE, Lombaert E, et al. (2011) Ecological genetics of invasive alien species. BioControl 56: 409–428. doi: 10.1007/s10526-011-9386-2

49. 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. doi: 10.1038/nprot.2014.025 PMID: 24457334

50. Sadegh M, Vrugt JA (2014) Approximate Bayesian computation using Markov Chain Monte Carlo simulation: DREAM(ABC). Water Resour Res 50: 6767–6787. doi: 10.1002/2014WR015386

51. Marjoram P, Zubair A, Nuzhdin SV (2014) Post-GWAS: where next? More samples, more SNPs or more biology? Heredity 112: 79–88. doi: 10.1038/hdy.2013.52 PMID: 23759726

52. Muchow RC, Sinclair TR, Bennett JM (1990) Temperature and solar radiation effects on potential maize yield across locations. Agron J 82: 338–343. doi: 10.2134/agronj1990.00021962008200020033x

53. 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. doi: 10.2135/cropsci1984.0011183X002400030028x

54. Elings A (2000) Estimation of leaf area in tropical maize. Agron J 92: 436–444. doi: 10.2134/agronj2000.923436x

55. 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. doi: 10.1016/0378-4290(88)90057-3

56. 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. doi: 10.2134/agronj1984.00021962007600040004x

57. Muchow R (1990) Effect of high temperature on grain-growth in field-grown maize. Field Crop Res 23: 145–158. doi: 10.1016/0378-4290(90)90109-O

58. 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. doi: 10.2134/agronj2002.0549

59. Neild RE, Newman JE (1987) Growing season characteristics and requirements in the Corn Belt. Rep. NCH 40. Purdue Univ., West Lafayette, IN.

60. Habier D, Fernando R, Kizilkaya K, Garrick D (2011) Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12: 186. doi: 10.1186/1471-2105-12-186 PMID: 21605355

61. 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/.

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

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

64. Xu S (2007) An empirical Bayes method for estimating epistatic effects of quantitative trait loci. Biometrics 63: 513–521. doi: 10.1111/j.1541-0420.2006.00711.x PMID: 17688503

65. 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. doi: 10.1371/journal.pone.0050604 PMID: 23226325

66. 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. doi: 10.1534/g3.114.010298 PMID: 24727289

67. Cooper M, DeLacy IH (1994) Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. Theor Appl Genet 88: 561–572. doi: 10.1007/BF01240919 PMID: 24186111

68. Singh M, Ceccarelli S, Grando S (1999) Genotype × environment interaction of crossover type: detecting its presence and estimating the crossover point. Theor Appl Genet 99: 988–995. doi: 10.1007/s001220051406

69. Cooper M, Gho C, Leafgren R, Tang T, Messina C (2014) Breeding drought-tolerant maize hybrids for the us corn-belt: discovery to product. Journal of Experimental Botany 65: 6191–6204. doi: 10.1093/jxb/eru064 PMID: 24596174

70. 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. doi: 10.1111/j.1469-8137.2011.03895.x PMID: 21973055

71. 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. doi: 10.1093/jxb/erl227 PMID: 17130185

72. Yin X, Kropff MJ, Goudriaan J, Stam P (2000) A model analysis of yield differences among recombinant inbred lines in barley. Agron J 92: 114–120. doi: 10.2134/agronj2000.921114x

73. 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. doi: 10.1534/genetics.109.105429 PMID: 19786622

74. Holzworth DP, Huth NI, Zurcher EJ, Herrmann NI, McLean G, et al. (2014) APSIM-evolution towards a new generation of agricultural systems simulation. Environ Modell Softw 62: 327–350. doi: 10.1016/j.envsoft.2014.07.009

75. Wang E, Ridoutt BG, Luo Z, Probert ME (2013) Using systems modelling to explore the potential for root exudates to increase phosphorus use efficiency in cereal crops. Environ Modell Softw 46: 50–60. doi: 10.1016/j.envsoft.2013.02.009

76. de Dorlodot S, Forster B, Pagés L, Price A, Tuberosa R, et al. (2007) Root system architecture: opportunities and constraints for genetic improvement of crops. Trends Plant Sci 12: 474–481. doi: 10.1016/j.tplants.2007.08.012 PMID: 17822944

77. 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. doi: 10.1093/aobpla/plr006 PMID: 22476477

78. 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. doi: 10.1098/rsif.2008.0172 PMID: 19205079

79. 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. doi: 10.1093/bioinformatics/btq278 PMID: 20591907

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

81. Curry GL, Feldman RM, Sharpe PJH (1978) Foundations of stochastic development. J Theor Biol 74: 397–410. doi: 10.1016/0022-5193(78)90222-9 PMID: 723284

82. 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. doi: 10.2134/agronj2012.0038

83. Sisson SA, Fan Y, Tanaka MM (2007) Sequential Monte Carlo without likelihoods. Proc Nat Acad Sci 104: 1760–1765. doi: 10.1073/pnas.0607208104 PMID: 17264216

84. Peters GW, Fan Y, Sisson SA (2012) On sequential Monte Carlo, partial rejection control and approximate Bayesian computation. Stat Comput 22: 1209–1222. doi: 10.1007/s11222-012-9315-y

85. 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.

86. Marjoram P, Molitor J, Plagnol V, Tavaré S (2003) Markov chain Monte Carlo without likelihoods. Proc Natl Acad Sci 100: 15324–15328. doi: 10.1073/pnas.0306899100 PMID: 14663152

87. 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.

88. Lund MS, Su G, Janss L, Guldbrandtsen B, Brøndum RF (2014) Invited review: genomic evaluation of cattle in a multi-breed context. Livest Sci 166: 101–110. doi: 10.1016/j.livsci.2014.05.008

89. Mikel MA, Dudley JW (2006) Evolution of North American dent corn from public to proprietary germplasm. Crop Sci 46: 1193–1205. doi: 10.2135/cropsci2005.10-0371

90. Guo Z, Tucker D, Lu J, Kishore V, Gay G (2012) Evaluation of genome-wide selection efficiency in maize nested association mapping populations. Theor Appl Genet 124: 261–275. doi: 10.1007/s00122-011-1702-9 PMID: 21938474

91. Combs E, Bernardo R (2013) Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers. Plant Genome 6: 1–7. doi: 10.3835/plantgenome2012.11.0030

92. Zhang X, Perez-Rodriguez P, Semagn K, Beyene Y, Babu R, et al. (2015) Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and GBS SNPs. Heredity 114: 291–299. doi: 10.1038/hdy.2014.99 PMID: 25407079

93. 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. doi: 10.1104/pp.013839 PMID: 12586890

94. 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. doi: 10.1093/jxb/eru328 PMID: 25148833

95. 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. doi: 10.1093/jxb/eri089 PMID: 15710636

96. 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. doi: 10.1071/AR05156

97. 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. doi: 10.1371/journal.pone.0043450 PMID: 22912876

98. 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. doi: 10.1093/jxb/ert370 PMID: 24218327

99. 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. doi: 10.1111/pbi.12172 PMID: 24618117

100. Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisec J, Technow F, et al. (2012) Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat Genet 44: 217–220. doi: 10.1038/ng.1033 PMID: 22246502

101. Ponciano JM, Burleigh JG, Braun EL, Taper ML (2012) Assessing parameter identifiability in phylogenetic models using data cloning. Syst Biol 61: 955–972. doi: 10.1093/sysbio/sys055 PMID: 22649181

102. Gelfand AE, Sahu SK (1999) Identifiability, improper priors and Gibbs sampling for generalized linear models. J Am Stat Assoc 94: 247–253. doi: 10.1080/01621459.1999.10473840

103. Robert C (2007) The Bayesian choice: from decision theoretic foundations to computational implementation. New York: Springer.

104. Gianola D (2013) Priors in whole-genome regression: the Bayesian alphabet returns. Genetics 194: 573–596. doi: 10.1534/genetics.113.151753 PMID: 23636739

105. Montes JM, Technow F, Dhillon B, Mauch F, Melchinger AE (2011) High-throughput non-destructive biomass determination during early plant development in maize under field conditions. Field Crop Res 121: 268–273. doi: 10.1016/j.fcr.2010.12.017

106. Araus JL, Cairns JE (2014) Field high-throughput phenotyping: the new crop breeding frontier. Trends Plant Sci 19: 52–61. doi: 10.1016/j.tplants.2013.09.008 PMID: 24139902

107. Liebisch F, Kirchgessner N, Schneider D, Walter A, Hund A (2015) Remote, aerial phenotyping of maize traits with a mobile multi-sensor approach. Plant Methods 11: 9. doi: 10.1186/s13007-015-0048-8 PMID: 25793008

108. Keurentjes JJB (2009) Genetical metabolomics: closing in on phenotypes. Curr Opin Plant Biol 12: 223–230. doi: 10.1016/j.pbi.2008.12.003 PMID: 19162531

109. Fernie AR, Schauer N (2009) Metabolomics-assisted breeding: a viable option for crop improvement? Trends Genet 25: 39–48. doi: 10.1016/j.tig.2008.10.010 PMID: 19027981

110. 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. doi: 10.1038/73786 PMID: 10700151

111. 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. doi: 10.1002/bit.23107

112. 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. doi: 10.1093/jxb/eru227 PMID: 24863438

113. Saha R, Suthers PF, Maranas CD (2011) Zea mays RS1563: A comprehensive genome-scale metabolic reconstruction of maize metabolism. PLoS ONE 6: e21784. doi: 10.1371/journal.pone.0021784 PMID: 21755001

114. Maher B (2008) Personal genomes: The case of the missing heritability. Nature 456: 18–21. doi: 10.1038/456018a PMID: 18987709

115. Nature Genetics Editorial (2015) Growing access to phenotype data. Nat Genet 47: 99. doi: 10.1038/ng.3213 PMID: 25627896