Coupling Day Length Data and Genomic Prediction tools for Predicting Time-Related Traits under Complex Scenarios

Diego Jarquin[¥,*], Hiromi Kajiya-Kanegae[†], Chen Taishen[†], Shiori Yabe[‡], Reyna Persa[¥], Jianming Yu[**], Hiroshi Nakagawa[a], Masanori Yamasaki[§], Hiroyoshi Iwata[†,*]

[¥]Department of Agronomy and Horticulture, University of Nebraska – Lincoln, Lincoln NE, 68583, USA.

[†]Graduate School of Agricultural and Life Sciences, The University of Tokyo, Bunkyo, Tokyo 113-8657, Japan.

[‡]Institute of Crop Sciences, National Agriculture and Food Research Organization (NARO), Tsukuba, Ibaraki 305-8518, Japan

[**]Department of Agronomy, Iowa State University, Ames, USA.

[§]Food Resources Education and Research Center, Graduate School of Agricultural Science, Kobe University, Kasai, Hyogo 675-2103, Japan

[a]Institute for Agro-Environmental Sciences, National Agriculture and Food Research Organization (NARO), Tsukuba Ibaraki 305-8604

[*]Corresponding authors: University of Nebraska – Lincoln, Lincoln NE, 68583, USA. E-mail: diego.jarquin@gmail.com ORCID ID 0000-0002-5098-2060 (DJ). Graduate School of Agricultural and Life Sciences, The University of Tokyo, Bunkyo, Tokyo 113-8657, Japan. E-mail: iwata@ut-biomet.org, ORCID 0000-0002-6747-7036 (HI).

Abstract

Genomic selection (GS) has proven to be an efficient tool for predicting crop-rank performance of untested genotypes; however, when the traits have intermediate optima (phenology stages) this implementation might not be the most convenient. GS might deliver high rank correlations but incurring in serious bias. Days to heading (DTH) is a crucial development stage in rice for regional adaptability with significant impact in yield potential. We developed two methods that incorporate day length information (DL) in the prediction process for two relevant scenarios: CV0, predicting tested genotypes in unobserved environments (C method); and CV00, predicting untested genotypes in unobserved environments (CB method). The use of DL has advantages over weather data since it can be determined in advance knowing the location and planting date. Under CV00, CB method returned a square root of the mean square error (SMSE) of 7.3 days, a Pearson correlation of 0.93 and the environmental mean differences (EMD) ranged between -6.4 and 4.1 days while GS method respectively produced 18.1 days, 0.41 and (-31.5 to 28.7 days). For CV0, C method respectively returned 3.9 days, 0.98 and (-4.95, 4.67 days). The proposed methods showed that DL information aids to improve significantly the predictive ability of DTH.

Introduction

To satisfy the daily demand of crop consumption for a growing population, it is necessary to improve the production systems to allow a sustainable increase of the yield potential and nutritional value of the new and up-coming varieties[1]. In plant breeding there are different tools and methodologies to increase the rate of genetic gain[2]. Traditional plant breeding methods use phenotypic and/or pedigree information to perform selection of experimental lines for developing superior breeding lines. However, phenotypic selection may not be permissible in many situations due to the high phenotyping costs; also, pedigree selection may not be accurate since recombination effects are not accounted for under this method.

Genomic selection[3] is an emerging tool that allows screening genotypes from a very large population without having to observe them in fields[4,5]. This method only requires phenotypic and genomic information for calibrating models, then other genotyped candidates are selected based on the predicted values obtained via their marker profiles[6]. In this context, the implementation of genomic tools and resources translates our understanding of the relationship between genotype and phenotype into predicted genetic merits for direct selection. This is highly desirable dealing with those traits that are controlled by a large number of genes with small individual effects, also known as complex traits[5,7]. GS has shown to be an effective tool for plant and animal breeding applications, especially dealing with complex traits[8]. The successfulness of GS predicting phenotypes from genotypic information has been reported in various crop species[9].

In GS, the estimated breeding values (EBVs) of complex traits of untested genotypes are computed based on genome-wide marker information. The EBVs can be considered as deviations around a fixed mean (e.g., zero). Once the EBVs are obtained, the breeders may conduct selections based on the ranked values by selecting those individuals in the tails. Depending on the trait, the individuals are selected either on the upper or lower tails of the ranked EBVs. For example, the upper side is selected for increasing genetic gains of traits like grain yield, stress resistance, or the content of essential nutrients. On the other hand, genotypes in the lower tail are selected for reducing lodging, shattering, the content of anti-nutrients, etc. In these cases, the EBVs provide sufficient information when the correlation between the EBVs and the direct genetic values (or phenotypic values) is high. However, the rank-based selection criteria might not be the most accurate when the target trait has an intermediate optima. This is the case of those traits related to phenological stages of crop plants. For example, days to heading (DTH; typically defined as the number of days after planting when more than 50% of individuals show the heading stage) is one of the critical traits of ideotypes with an increased yield potential. Hence it is important to optimize DTH for high yield potential in different environments[10].

Usually multi environment trials (METs) are performed for developing stable cultivars across a wide range of environmental conditions, and the breeders may be interested in estimating DTH for a large set of environments instead of a single one. However, the EBVs may contain a large bias if phenotypic information from a targeted unobserved environment is not available. The EBVs of one environment may not be valid for another environment due to variations in the environmental *stimuli*. These differences remain also for same genotypes

observed in same locations but tested at different times (planting date, years). As consequence of the aforementioned issues, the problem of correctly estimating DTH for a set of untested genotypes in a wide range of unobserved environments becomes harder to solve than for non-time related traits. As pointed before, the advantage of GS techniques over conventional phenotypic methods is its ability for performing accurate rank predictions of untested genotypes based exclusively on their marker profiles without having to observe these in fields. Thus, a method capable of delivering highly accurate predictions of untested genotypes in unobserved environments without incurring in a large bias is desirable.

As was pointed before, DTH in rice is important for regional adaptability and has usually great impact on yield potential[10], genetic systems controlling it have been studied for a long time by integrating Mendelian genetics and genomics[11]. The potential of GS in DTH has been evaluated in rice[12,13] and sorghum[14]. Onogi *et al.*[15] focused on the genomic prediction of DTH in multiple environments and demonstrated the potential of a novel method that integrates GS with a crop growth model. Li *et al.*[14] also focused on the genomic prediction of DTH in multiple environments and demonstrated the potential of models combining GS with joint regression analysis on photothermal time. To our knowledge, no studies combining DL at the day when genotypes reach DTH with GS tools have been done before.

Three cross-validation schemes of interest for designing and planning experiments were considered. CV2, predicting tested genotypes in observed environments (incomplete field trials); CV0, predicting tested genotypes in unobserved environments (targeting environments); and CV00, predicting untested genotypes in unobserved environments (targeting genotypes and environments).

For assessing the proposed methods the square root of the mean square error (SMSE), the Pearson correlation and the differences of the environmental means (EMD) between predicted and observed values were computed.

In this study, we propose two novel methods to predict DTH in a precise and accurate way (i.e., small bias and high correlation between predicted and observed values) in target observed (C method) and unobserved (CB method) environments. Under the hardest cross-validation scheme (CV00), the results of the proposed method (CB) are compared with those obtained using the conventional GS model (E+G). For studying other realistic scenarios, the two remaining cross validation schemes were also considered (CV2 and CV0). Results from CV2 scheme using the conventional GS model (E+G) were used to contrast the performance of the proposed method when complete information on genotypes and environments are available. Thus, these results may provide a better idea of the model performance of the proposed method when no information of genotypes and environments are available. Results under the CV0 scheme correspond to the proposed C method. Thus, these results are used for assessing the levels of predictability that can be reached when phenotypic information of a genotype is available for other environments.

Here the environments are considered "target" in a strict sense because based on theoretical DL values the best location and planting date can be determined before conducting field experiments.

Results

Descriptive statistics. The Box-plot of DTH is depicted in Figure S1. The environments were ordered based on the environmental medians. In general, those environments with late planting date appeared first. The medians ranged between 59 ("Fukuyama 2010 Late") and 116 ("Akita 2015") days. The SDs of the environments varied between 5.07 and 16.83 days. Also, the smallest and the largest DTH values were 42 and 148 days and these were observed in "Tsukubamirai 2016 Late" and "Akita 2015" environments, respectively.

Evaluating the bias and accuracy of the methods. Table 1 contains the observed and the predicted environmental means (E-mean) for four different strategies (cross-validation scheme-method combination) as well as the corresponding SMSE and the Pearson correlation between the predicted and observed values. These four strategies were used for solving different prediction problems. CB and GS methods were implemented under the CV00 scheme (CV00_CB, CV00_GS) for comparing the model performance of the proposed model with the conventional implementation (GS) predicting untested genotypes in unobserved environments. C method was implemented under the CV0 scheme (CV0_C) for assessing the predictive ability of this method predicting tested genotypes in untested environments (also this is the intermediate step of the CB method used in the previous scheme). Lastly, the conventional method (GS) was implemented under the CV2 scheme (CV2_GS) to assess the levels of predictability that can be reached when complete information of genotypes and environments are available.

CV2 is the scenario that includes the most complete sets of information on genotypes and environments for calibrating models. Across environments, the CV2_GS strategy returned a SMSE of 3.8 days and a Pearson correlation of 0.98; within environments, the average SMSE and Pearson correlation values were of 3.7 days and 0.97, respectively. The difference of the environmental means (EMD) between predicted and observed DTH values ranged between -0.02 and 0.03 days. Since this scenario uses the most complete sets of information for calibrating models, these results can be used as reference for assessing the performance of the proposed methods under more complex prediction scenarios.

CV0_C results provide information of the ability for predicting tested genotypes in unobserved environment when considering only phenotypic and DL information of the same genotype via the C method. Across environments, this method returned a SMSE of 3.9 days and a Pearson correlation of 0.98; within environments, the average SMSE and Pearson correlation were of 3.7 days and 0.98. The EMD ranged between -4.95 and 4.67 days.

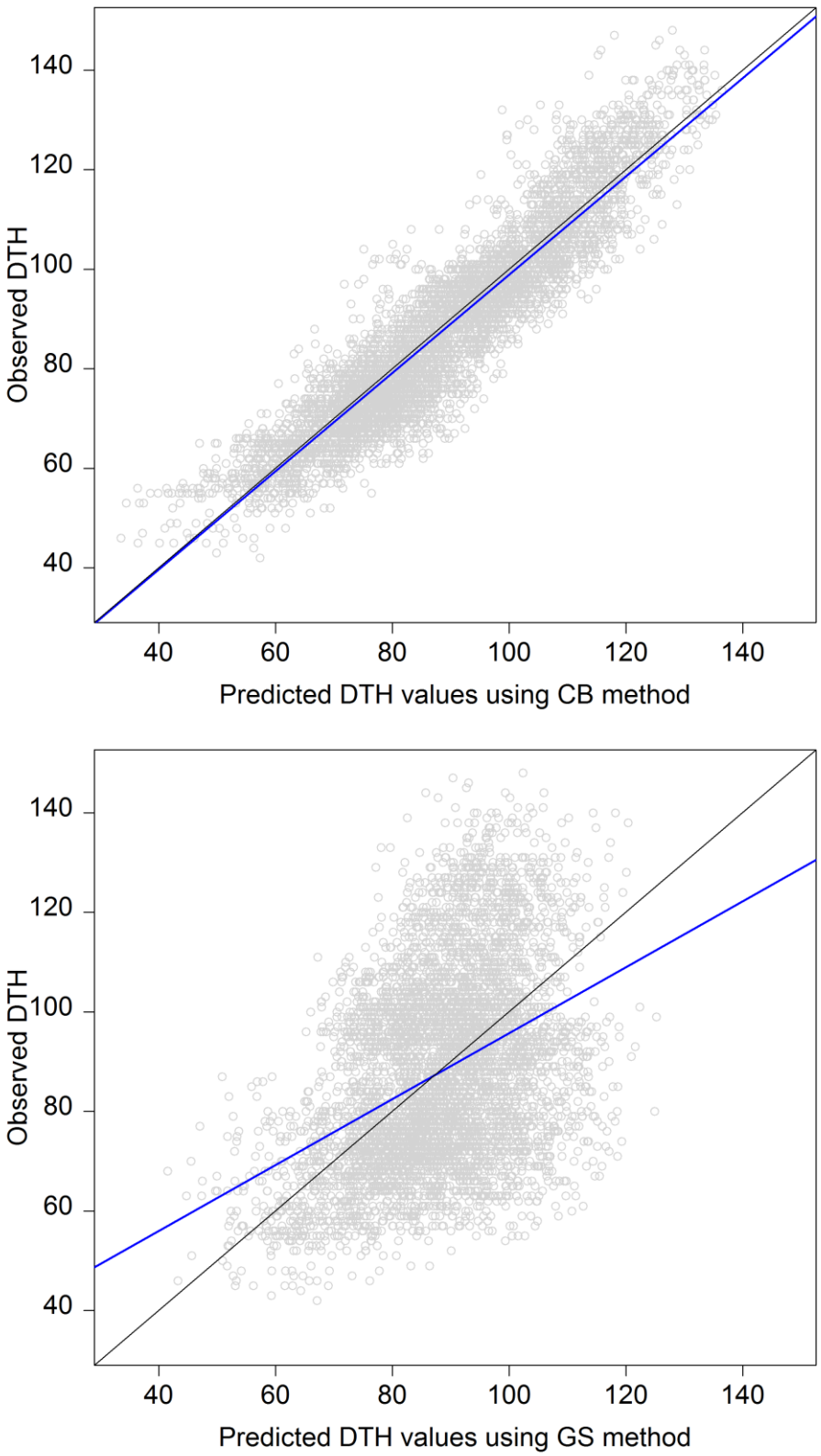
CV00 corresponds to the scenario of predicting untested genotypes in unobserved environments and it is considered the hardest and most interesting prediction problem because no phenotypic information is available for connecting untested genotypes in unobserved

environments with calibration sets. Across environments, the conventional GS method (CV00_GS) respectively returned a SMSE and a Pearson correlation of 18.1 days and 0.41; within environments, the average SMSE and Pearson correlation were of 17.5 days and 0.69. The EMD ranged between -31.5 and 28.7 days. For the same scenario, across environments the proposed CB method (CV00_BC) returned a SMSE of 7.3 days and a Pearson correlation of 0.93; within environments, the average SMSE and Pearson correlation were 7.2 days and 0.86. The EMD ranged between -6.4 and 4.1 days.

Table 1. Environmental means, Square root of the Mean Squared Error (SMSE) and the Pearson correlation between predicted and observed days to heading (DTH) for 112 rice genotypes tested in 51 environments in Japan. Three cross-validation schemes were implemented for mimicking realistic prediction scenarios. CV2 resembles incomplete field trails where some lines are tested in some environments but not in others. CV0 represents the scenario of predicting tested genotypes in unobserved environments. CV00 considers the prediction of untested genotypes in unobserved environments. CV2 was implemented using the conventional implementation (GS, E+G) leaving one phenotype out at a time. For CV0 the C method was used combining phenotypic and day length information (DL) of each genotype (one at a time) already tested in other environments. For CV00 two methods were considered: (*i*) the conventional implementation (GS, E+G) leaving one genotype out across environments and deleting any phenotypic information from the target environment; and (*ii*) the proposed CB method, which combines phenotypic and DL information from the tested genotypes (training set) with the genomic BLUP values ($\hat{\boldsymbol{g}}$) of the untested genotypes obtained in (*i*).

	Mean					Square root of the Mean Square Error				Correlation			
	Mean	CV00_CB	CV00_GS	CV0_C	CV2_GS	CV00_CB	CV00_GS	CV0_C	CV2_GS	CV00_CB	CV00_GS	CV0_C	CV2_GS
Tsukubamirai 2004 Late	79.4	76.7	85.9	76.3	79.4	7.0	12.6	4.6	2.5	0.86	0.70	0.99	0.98
Tsukubamirai 2004 Early	87.5	93.2	84.1	92.2	87.5	10.0	12.9	6.4	4.2	0.84	0.55	0.97	0.97
Tsukubamirai 2005 Early	98.5	102.1	87.1	101.0	98.5	6.9	14.0	3.5	2.0	0.87	0.75	0.98	0.99
Tsukubamirai 2005 Late	76.7	74.8	84.8	73.7	76.7	5.7	12.4	3.4	2.2	0.86	0.68	0.99	0.99
Kasai 2006	87.4	83.8	88.3	82.8	87.4	7.9	8.8	6.8	4.7	0.85	0.75	0.93	0.94
Fukuyama 2006	80.6	83.0	82.5	82.0	80.6	5.6	13.4	2.0	2.2	0.87	0.29	0.99	0.99
Fukuyama 2007	84.1	83.4	86.6	82.5	84.1	5.0	6.1	2.9	3.0	0.87	0.84	0.98	0.98
Tsukuba 2008	104.5	107.7	87.2	107.3	104.5	9.1	19.2	4.0	4.7	0.87	0.85	0.98	0.98
Kasai1st 2008	102.2	99.9	88.7	99.2	102.2	9.0	17.8	4.3	4.8	0.89	0.71	0.99	0.99
Fukuyama 2008	77.8	82.2	85.6	81.3	77.8	7.8	11.1	4.5	2.8	0.86	0.77	0.98	0.98
Tsukuba 2009	103.9	107.7	87.1	108.0	103.9	8.6	19.0	5.2	4.7	0.89	0.81	0.98	0.97
Kasai1st 2009	101.9	100.6	89.4	100.0	101.9	7.6	15.1	2.7	3.5	0.89	0.83	0.99	0.99
Fukuyama 2009	78.7	78.7	89.6	77.8	78.7	5.2	13.2	1.5	2.2	0.86	0.83	0.99	0.99
Tsukuba 2010 Late	83.5	89.1	88.9	88.4	83.5	8.4	11.1	5.9	3.0	0.86	0.66	0.97	0.97
Kasai1st 2010	100.6	100.7	89.5	100.1	100.6	7.1	14.3	1.9	2.9	0.88	0.77	0.99	0.99
Fukuyama 2010	76.2	79.5	88.2	78.1	76.2	5.9	15.3	2.6	2.8	0.86	0.50	0.99	0.99
Fukuyama 2010 Late	59.0	65.4	87.7	62.5	59.0	9.2	29.7	4.2	7.7	0.84	0.82	0.96	0.90
Tsukuba 2011 Early	105.9	109.9	85.5	109.9	105.9	9.9	23.9	5.3	5.6	0.87	0.66	0.98	0.97
Tsukuba 2011 Late	82.8	80.8	90.0	80.4	82.8	5.9	11.9	4.1	3.6	0.86	0.78	0.95	0.96
Tsukuba 2011 Middle	94.7	95.7	90.9	95.4	94.7	6.5	11.1	3.0	3.1	0.88	0.67	0.98	0.98
Kasai 2011	101.4	100.6	89.3	100.0	101.3	7.8	15.7	2.4	3.6	0.89	0.76	0.99	0.99
Fukuyama 2011	75.7	79.4	91.1	78.6	75.7	6.4	16.9	3.1	2.0	0.87	0.83	0.99	0.99
Fukuyama 2011 Late	66.4	72.7	97.9	66.2	66.4	8.9	32.4	2.9	5.7	0.68	0.58	0.95	0.86
Tsukuba 2012 Early	97.1	96.0	86.9	101.1	97.1	5.7	12.4	5.8	3.2	0.81	0.74	0.97	0.94
Tsukuba 2012 Late	84.3	81.6	88.2	81.0	84.3	6.1	7.5	4.3	3.5	0.86	0.80	0.96	0.97
Tsukuba 2012 Middle	92.6	95.4	88.4	94.8	92.6	6.3	8.3	3.7	2.9	0.87	0.82	0.97	0.97
Kasai 2012	103.8	102.7	86.4	102.1	103.8	7.9	21.3	2.7	3.8	0.89	0.60	0.99	0.99
Fukuyama 2012	75.2	78.8	89.9	78.0	75.2	6.2	16.3	3.1	2.0	0.87	0.83	0.99	0.99
Fukuyama 2012 Late	71.5	77.0	103.1	68.6	71.6	7.5	32.4	3.4	4.6	0.70	0.66	0.98	0.92
Fukuoka 2012	79.6	76.0	89.1	75.4	79.6	6.7	13.0	5.2	4.2	0.84	0.69	0.96	0.96
Akita 2012	113.1	115.9	90.1	114.5	113.1	8.3	26.1	2.8	4.4	0.86	0.54	0.98	0.97
Kasai 2013	104.8	104.1	87.3	103.5	104.8	8.0	20.1	3.2	4.7	0.90	0.80	0.98	0.98
Fukuyama 2013	74.9	80.1	89.4	79.3	74.9	7.4	16.3	4.6	1.8	0.87	0.82	0.99	0.99
Fukuyama 2013 Late	62.1	64.8	85.7	64.1	62.1	6.3	27.1	3.6	4.9	0.83	0.36	0.93	0.95
Fukuoka 2013	77.4	77.3	93.3	76.7	77.4	5.7	17.5	1.9	2.6	0.86	0.77	0.99	0.98
Akita 2013	116.6	115.8	91.1	115.1	116.5	8.5	28.1	2.4	4.9	0.88	0.66	0.99	0.98
Kasai 2014	102.4	105.1	88.9	104.4	102.3	8.4	18.0	2.8	4.4	0.90	0.66	0.99	0.99
Fukuyama 2014	79.6	82.1	87.9	81.3	79.6	5.9	12.1	2.1	1.3	0.88	0.68	0.99	0.99
Fukuyama 2014 Late	67.6	64.5	88.1	62.9	67.6	6.7	21.4	5.1	6.0	0.83	0.79	0.98	0.94
Fukuoka 2014	81.8	77.7	80.4	77.1	81.8	6.7	11.5	5.4	3.1	0.87	0.45	0.97	0.97
Akita 2014	115.2	116.2	89.6	115.6	115.2	9.2	28.5	2.2	5.8	0.88	0.68	0.99	0.98
Fukuyama 2015	76.3	72.6	88.1	71.9	76.3	6.5	14.4	4.6	3.3	0.86	0.63	0.99	0.98
Akita 2015	118.1	119.8	89.4	117.6	118.0	7.7	29.9	2.5	3.6	0.81	0.76	0.99	0.98
Kasai 2016	102.3	101.2	85.6	100.7	102.2	7.2	20.4	3.1	3.9	0.90	0.61	0.98	0.98
Fukuyama 2016	79.1	81.1	85.2	80.8	79.1	6.0	8.9	2.4	2.9	0.86	0.82	0.99	0.98
Fukuyama 2016 Late	67.5	65.2	89.6	64.6	67.5	6.8	26.1	4.5	6.3	0.79	0.63	0.95	0.91
Tsukubamirai 2016 Late	73.6	77.9	84.6	77.2	73.6	7.0	13.6	4.9	1.9	0.90	0.80	0.98	0.99
Kasai 2017	108.6	106.2	85.7	105.5	108.6	8.3	26.3	4.3	4.4	0.89	0.56	0.98	0.98
Fukuyama 2017	77.4	80.2	87.7	79.6	77.4	6.3	16.3	2.7	2.0	0.88	0.42	0.99	0.99
Fukuyama 2017 Late	68.4	67.7	88.8	67.0	68.4	6.0	22.9	3.0	4.9	0.82	0.69	0.97	0.95
Tsukubamirai 2017 Late	69.5	73.1	86.9	72.2	69.5	6.8	19.8	4.3	1.6	0.87	0.65	0.98	0.99
Mean						**7.2**	**17.5**	**3.7**	**3.7**	**0.86**	**0.69**	**0.98**	**0.97**

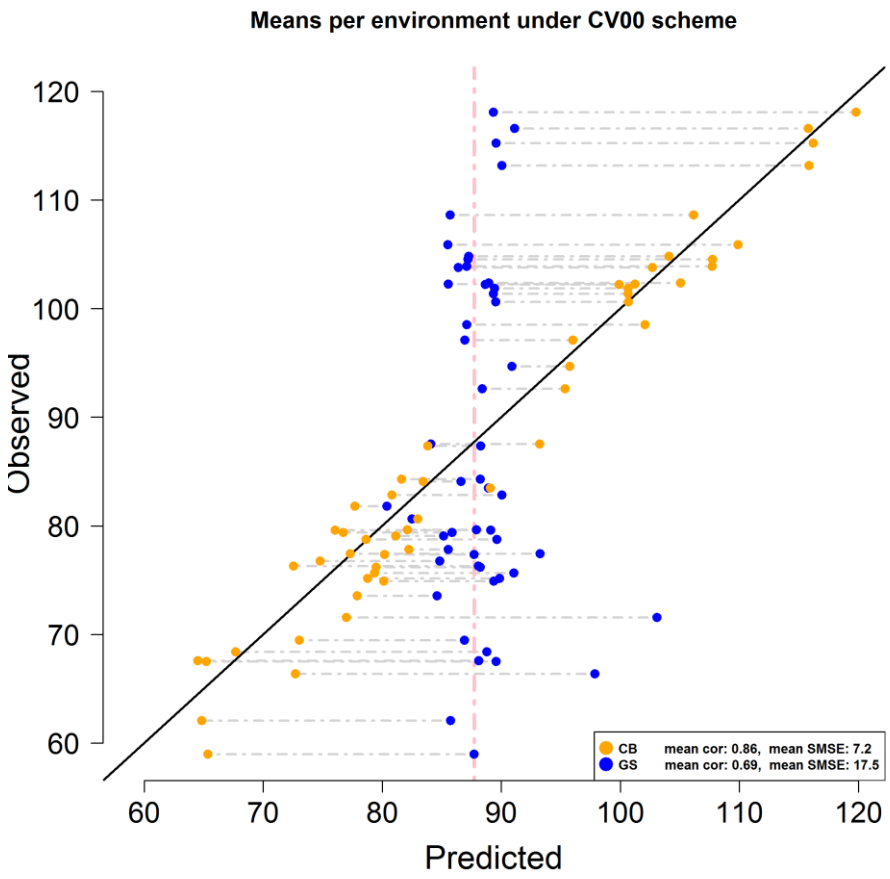
Only the results of CB and GS methods used under CV00 scheme are directly comparable between them. Figure 1 depicts the scatter plot between predicted and observed DTH values for all genotypes across environments. The top panel contains the results obtained with the proposed CB method, while the bottom panel shows the results derived from the conventional implementation (GS, E+G). The blue and black lines correspond to the fitted and the on target predictions with respect to the observed DTH values. In general, across and within environments (Table 1) the results obtained with CB method showed a better fit than those derived from the GS model.



1. Figure 1. Scatter plot of predicted (*x-axis*) and observed (*y-axis*) DTH values using two different methods (CB and GS) for predicting untested genotypes in unobserved environments (CV00 scheme) for a rice data set comprising 112 genotypes tested in 51 environment in Japan between 2005 and 2017. The top panel presents the

results obtained with the proposed CB method. It combines the predicted environmental means (E-mean) obtained using the proposed C method with the genomic BLUPs ($\hat{g}$) derived from the conventional implementation (GS, E+E) under the CV00 scheme. The bottom panel displays the results obtained with the conventional implementation (GS, E+G GS). The blue line represents the fitted line between predicted and observed DTH values while the black line show the exact match (on target) of the predicted values on the observed values.

The advantages of considering the proposed CB method for predicting DTH under the CV00 scheme are more clearly appreciated analyzing the predicted environmental means (E-means). Figure 2 displays the predicted (*x-axis*) and observed (*y-axis*) E-means. The predicted E-means obtained with the CB and GS methods are indicated in orange and blue colors, respectively. As expected, when no information from the target environment was available for calibrating models the conventional implementation (GS, E+G) regresses the predicted values towards the overall mean (~87.74) observed in training sets. On the other hand, the CB method consistently returned closer values to the diagonal black line, which represents the exact match.



2. Figure 2. Scatter plot of predicted and observed environmental means (E-means) using the proposed CB method (orange) and the conventional implementation (GS, E+G) (blue) under the CV00 scheme (predicting untested genotypes in unobserved environments). The vertical pink line indicates the position of the overall phenotypic mean across environments (87.74 days). The diagonal black line shows the exact match between predicted and observed means. SMSE and cor stand for square root of the mean square error and the Person correlation between predicted

and observed DTH values. These were computed within environments; and the mean SMSE and mean cor correspond to the across environments means.

Since the objective of this study is to develop a method for predicting DTH for untested genotypes in unobserved environments, the predicted values obtained with the CB method were used for fitting the DTH curves for each one of the 112 genotypes across environments. Figure S3 shows the fitted curves for the predicted (orange) and observed (gray) DTH values across the 51 environments. In principle, these fitted curves can be used for predicting DTH for other environments not included in this study. The information needed for defining a target environment are the location and planting date to allow the computation of the theoretical DL values using different methods[16]. Once these theoretical values are obtained the C method is implemented to find the estimated DTH E-means. In general, the fitted curves (orange) derived from the predicted DTH values were very similar to those derived from the observed DTH values (blue).

Discussion

The boxplot (Figure S1) showed a high heterogeneity among the environments in the study with those with "late" planting date exhibiting smaller phenotypic variability and shorter occurrence time for DTH in comparison with the "early" planting environments. Conventional GS models assume homogeneity of variances among environments and the violation of this assumption may induce serious inconsistencies in the results especially when no genotype-by-environment parameters are included in models.

The results obtained with the CV2_GS strategy can be viewed as the threshold of the predictive ability levels that can be reached with the conventional implementation when complete information about genotypes and environments is available. On the other hand, the results from the CV00_GS strategy provide information of the model performance of the conventional implementation when no phenotypic information is available for connecting genotypes in target environments. This cross-validation scheme is among the three cases (CV2: predicting tested genotypes in observed environments; CV0: predicting tested genotypes in unobserved environments; and CV00: predicting untested genotypes in unobserved environments) the most interesting, useful and challenging for breeders. Dealing with traits where the selection process is based in ranks rather than in the *per se* estimated values, the conventional implementation (GS) usually provides good results (moderate to high correlations) under this scheme (CV00).

However, in traits related with important growth stages like DTH where the *per se* estimated values are crucial for designing, planning and establishing future experiments (i.e., CV0 and CV00 situations) the implementation of the conventional implementations (GS) might not be the best choice. When no phenotypic information of the target environment is available, the E-mean of the target environment is determined as the weighted average of those

environments in the training set forcing it towards a general mean. Hence, the final prediction is composed of this general mean plus deviations due to genomic effects. Since these deviations are conceptualized to be centered in zero, under CV0 and CV00 schemes the GS methods would produce similar environmental means (i.e., centered on an overall mean determined by the environments in the calibration set).

The obtained results showed that the C method provides good estimations of DTH for the CV0 scheme. Across environments, the SMSE was of 3.9 days and a Pearson correlation of 0.98 was obtained. These results were very similar to those obtained using the conventional GS models when complete information of the genotypes and environments were available (i.e., CV2 scheme). Under CV2, the SMSE and Pearson correlation across environment were of 3.8 days and 0.98, respectively.

As mentioned before, it is expected to obtain similar environmental means using either CV0 or CV00 schemes under the conventional GS implementations. The CV00_GS strategy returned, across environments, a SMSE of 18.1 and a Pearson correlation of 0.41. The EMD between predicted and observed values ranged between -31.5 and 28.7 days while for the CV0_C strategy these ranged between -4.95 and 4.67 days. In this case, the CV0_C strategy used only the phenotypic information of the genotype to be predicted but observed in other environments. Since the CV0 and CV00 schemes produce similar E-means when no information of the target environment is available using the conventional implementations (GS) it is also expected that the proposed C method will produce good results compared with the CV0_GS strategy not considered in this study.

As was pointed before, the ultimate goal of this research was to develop a method able to deliver small SMSE when predicting unobserved genotypes in unobserved environments (CV00 scheme). The BC method seems to produce good results under this scenario. The SMSE across environments was 7.3 days with a Pearson correlation of 0.86. These results are very encouraging compared with those obtained using the conventional implementation (GS) under this scheme (SMSE: 18.1 and cor: 0.41). In this case, the BC method doubled the accuracy of the conventional GS method and reduced the SMSE by 60%. However, the advantages of this method were more evident when comparing the range of the differences between the predicted and observed environmental means. Under the conventional implementation (CV00_GS) these differences ranged between -31.5 and 28.7 days while with the proposed CB method (CV00_BC) these ranged between -6.4 and 4.1 days. The proposed CB method reduced the extremes (left and right sides) of this range between 80 and 85%, respectively. These improvements can be observed in Figure 2 were the environmental means of the proposed BC method (orange dots) were always closer to the diagonal line, representing the exact match with true means, than those means obtained using the conventional GS implementation (blue dots).

At the beginning, the implementation of the proposed methods (C and BC) requires enough information for an accurate characterization of the genotype's DL curves. For this, sets of genotypes are tested in several environments. Once this information is available, potentially any untested genotypes could be accurately predicted in any unobserved environment. Thus, the

obtained results should hold not only for the set of analyzed genotypes and environments of this study but also for others.

The developed methods here presented (C and BC) showed that the use of DL information significantly improves the predictive ability of the conventional implementations (GS) also reducing the bias between the predicted and observed DTH values. An advantage of the C method is that if phenotypic information of the same genotype is available for a set of environments, it is easy to estimate the expected DTH in an unobserved environment without needing molecular marker information or phenotypic information from other genotypes. The BC method improved the prediction of DTH of untested genotypes in unobserved environments by coupling the marker profiles of the untested genotypes with the phenotypic and DL information from other genotypes.

Both of the proposed methods require the day length (DL) information of the environments (training and testing) in their construction. The convenience of the proposed method resides in the fact that the theoretical DL information can be obtained in advance even before establishing the experiments and thus it can be exploited during the planning and designing stages. The proposed methods capture the genotype-by-environment interaction (G×E) component via the fitted curves thus the explicitly inclusion of this term is not needed. The conventional implementation (GS) returns the same estimated genomic effect across environments for a particular genotype when the G×E component is absent. The curves precisely capture the sensitivity experienced by genotypes under different environments driving specific responses to each environment. Thus, although the proposed BC method is composed of the genomic component computed without explicitly considering the G×E term the final prediction accounts for this interaction term via the curves. In addition, the formal modeling of heterogeneity of variances is not necessary because these patterns are also accounted with the current method via the different curves.

Materials and Methods

Phenotypic data. The phenotypic data set comprises the seeding and heading date of 112 rice cultivars from Japan tested in 51 locations (North to South of Japan) from 2004 to 2017 (Table S1). This multi-environmental experiment was established for investigating heading date related genes.

The DTH was calculated as the difference between the heading date and the seeding date. The heading date was determined when more than 50% of the individuals in the plot reached the heading stage. The cultivars that did not reach the heading stage at the end of the experiment were removed from the analysis.

Marker genotype data. For each variety, total DNA was extracted using the CTAB method[18]. Next-generation sequencing data was obtained using the Illumina HiSeq 2000, HiSeq 4000 and HiseqX systems via paired-end sequencing (Illumina, Inc., San Diego, CA, USA). The sequence

data have been deposited in the DNA Data Bank of Japan (DDBJ) Sequence Read Archive, under Submission DRA008071. Data sets deposited in the DDBJ Sequence Read Archive (SRA106223, ERA358140, DRA000158, DRA000307, DRA000897, DRA000927, DRA007273, DRA007256) were reanalyzed. Adapters and low-quality bases were removed from paired reads using the Trimmomatic v0.36 program[18]. The preprocessed reads were aligned to the Os-Nipponbare-Reference-IRGSP-1.0[19] by using the bwa-0.7.12 mem algorithm with the default options[20].

SNP calling was based on alignment using the Genome Analysis Toolkit[21] (GATK, 3.7-0-gcfedb67) and Picard package V2.5.0 (http://broadinstitute.github.io/picard). The mapped reads were realigned using RealignerTargetCreator and indelRealigner of GATK software. SNPs and InDel were called at the population level using the UnifiedGenotyper of GATK with the -glm BOTH option.
Heterozygous loci were treated as missing, owing to the assumption that almost all markers were well fixed in rice cultivars after several generations of self-pollination. We used only biallelic non-missing sites over all cultivars with a minor allele frequency (MAF) ≥ 0.025.

Day length data. Although there are several definitions for day length (DL) according to the Smithsonian Institution[22] roughly speaking it can defined as the amount of time in a day with sunlight. The theoretical DL values were computed by the default method[16] included in the geosphere (v1.5-5) R package[23].

GBLUP model. The Genomic Best Linear Unbiased Predictor model (Genomic BLUP or GBLUP) is a main effects model used in GS applications[24,8]. It attempts to explain the DTH value of the ith genotype (i=1, 2,…, I) observed in the jth environment (j=1, 2, …, J) as the sum of a common mean, random environmental and genomic effects, and an error term and it is represented as follows:

$$Y_{ij} = \mu + E_j + g_i + \varepsilon_{ij} \qquad [1]$$

where μ is a common effect to all genotypes across all environments; E_j are IID (independent and identically distributed) outcomes from a $N(0, \sigma_E^2)$ representing the effect of the jth environment; $\boldsymbol{g} = \{g_i\} \sim N\left(\mathbf{0}, \mathbf{G}\sigma_g^2\right)$ is the vector of genomic effects, with $\mathbf{G} = \frac{\mathbf{XX}'}{p}$ as the genomic relationship matrix whose entries describe the genetic similarities between pairs of individuals, $\mathbf{X}$ is the standardized (by columns) matrix for p molecular markers; ε_{ij} are IID outcomes from a $N(0, \sigma_\varepsilon^2)$ representing measure errors; and σ_E^2, σ_g^2 and σ_ε^2 are the corresponding associated variance components of the described random effects.

Usually, when METs are analyzed the environmental component explains the larger proportion of the phenotypic variability across environments[25,26]. Hence, from [1] is concluded that the predicted value of the ith genotype in the jth environment strongly depends on the

environmental mean, which is composed by a common mean across environments plus an environmental deviation that corresponds to the effect of jth environment.

Conceptually, the common mean effect can be viewed as the average performance of all genotypes across environments. The environmental deviation corresponds to the mean of the specific affectations of the current environmental *stimuli* in the set of the tested genotypes beyond the common mean. Thus, these environmental effects can be considered deviations around the common mean. Hence, by construction the contribution of the genomic component to the idealized phenotype is much smaller in comparison with the overall mean and the environmental effects.

If phenotypic information (DTH) of the target environment is available, good approximations of the environmental effects can be computed. Otherwise, the environmental effects are estimated using genotypes observed in other environments. Thus, the environmental component is estimated as the weighted average of the environmental effects in the training set. Since these random effects are conceptualized to be centered on zero, the weighted average will be close to zero. Hence, the predicted DTH value will be compose mainly of a common mean and a small genomic effect shrinking the predicted values towards the common mean. Although this method usually delivers high correlations between predicted and observed values, it also might induce a large bias (MSE) affecting the prediction of time related traits.

For solving this issue, two methods (C and CB) that incorporate DL information in the prediction process are proposed. The idea for using DL information in the prediction process was motivated examining the relationship between CDL and DTH of all genotypes across environments (Figure S2). A simple linear regression of DTH on CDL returned an R-squared value of 0.99 exhibiting a strong relationship. It is concluded that if CDL values are known accurate estimations of DTH may be obtained for any genotype in any environment. However, in order to compute CDL the number of days between planting date and the heading occurrence (DTH) should be known beforehand complicating the estimation process.

The analysis of DL values at DTH across environments showed relationship patterns. For illustrating this, Figure 3 is used as example. It corresponds to genotype 56 observed in 50 environments out of 51 (1-46, and 48-51 but not in 47). The goal is estimate DTH for this genotype in the unobserved environment (Tsukubamirai 2016 Late). For each environment, the planting date was considered as day 1 (*x-axis*). Then, the daily DL (*y-axis*) values were plotted (pink lines) from day 1 (planting) until genotype 56's DTH (blue point). In Figure 3, a pattern relating DTH on DL at the occurrence day can be identified. This gradient seems not to be strongly affected by the initial DL values at the different planting dates across environments. However, it appears to be linked with the initial DL's changing rate at the planting day. Those environments where DL's changing rates decreasing at the planting time (late cultivars) showed shortest DTH values and *vice-versa*.

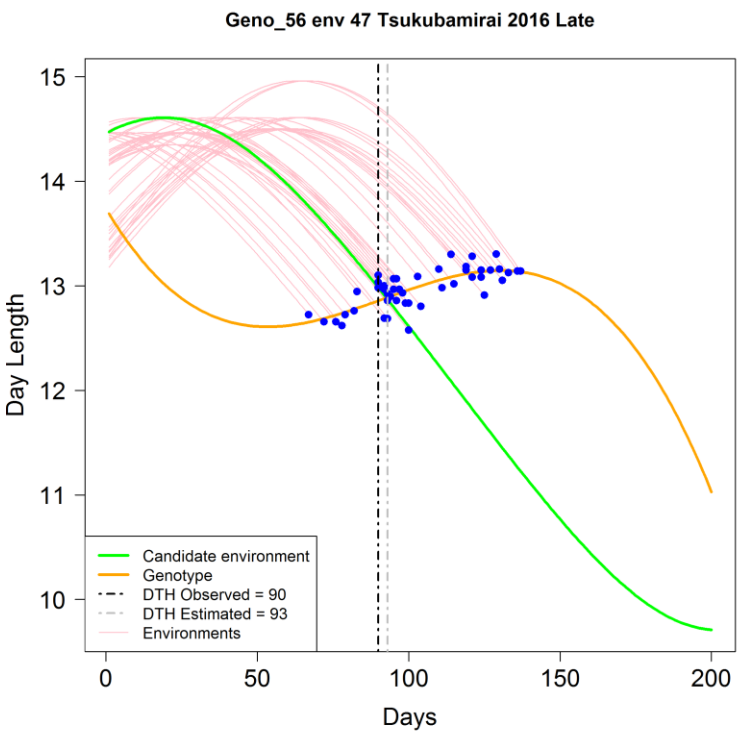
For CV00 scheme, the proposed method requires: (*i*) the estimation of DTH for a set of tested (training) genotypes (one at a time) yet to be observed in a target/unobserved environment (CV0); (*ii*) to compute the expected environmental mean of the unobserved environment using

the estimated values from (*i*); (*iii*) to add the resulting BLUPs of the untested genotypes derived from conventional GS methods to the estimated environmental mean. Further details of these steps are given next.

C method. The DTH estimation of a tested genotype in an unobserved environment (CV0) requires phenotypic and DL information across several environments. The first step considers the characterization of DL at the heading occurrence time across environments. For this, a third degree polynomial equation is used for fitting DL as function of DTH (orange curve in Figure 3) as follows:

$dl_{ij} = b_0 + b_1 t + b_2 t^2 + b_3 t^3$, where dl_{ij} represents the observed DL of the ith genotype in the jth environment; t is the time independent variable; b_0, b_1, b_2 and b_3 are the model parameters; and e_{ij} is the error term and it is assumed to be IID $N(0, \sigma^2)$. This model is referred as C_1.

Similarly, the daily DL information of the target environment (green curve) is modeled. The resulting fitted model is referred as C_2. The solution to this system of two equations (C_1, C_2) will produce an estimate of DTH of the genotype in the target environment. In order to find the intersection of these two curves numerical evaluation was implemented (day 93 in Figure 3; vertical dotted gray line).



3. Figure 3. Graphical representation of the curve's C method for estimating days to heading (DTH, *x-axis*) of genotype 56 in environment 47 (Tsukubamirai 2016 Late). Pink curves represent daily day length in hours (DL, *y-axis*) values from planting day (day 1) until heading time for genotype 56 in 50 environments (1-46, and 48-51). The blue points indicate the corresponding DL values when genotype 56 reached heading time. The green curve (C_2) represents the target environment (Environment 47) where no genotypes have been observed yet. The orange curve (C_1) depicts the fitted line for genotype 56 using a third degree polynomial equation relating DL at the occurrence time and DTH across environments. The vertical dotted gray line shows the intersection (93 days) between the green and the orange curves (C_1, C_2). This value is the estimated DTH for genotype 56 in the unobserved environment

"Tsukubamirai 2016 Late". The vertical dotted black line corresponds to the observed DTH (90 days) in environment 47.

CB Method. Since the ultimate goal is to predict DTH for untested genotypes in unobserved environments, the first step is to accurately estimate the expected DTH mean of the target environment. Then, the final predicted value is composed of the expected DTH mean of the target environment plus the genomic estimated breeding value ($\hat{g}_i$) of the untested genotype obtained via conventional GS methods (E+G model). For estimating the DTH mean, the C method was implemented and the predicted values were averaged.

Using a toy example consisting of four genotypes (G_1, G_2, G_3, and G_4) and three environments (E_1, E_2, and E_3) the procedure for finding the environmental mean of the target environment is described. The top panel of Figure 4 contains the observed values; the middle panel shows the predicted values presumably obtained with the C method; and the bottom panel displays the conditional (given the tested genotypes) estimated DTH environmental means for each genotype at each environment. Once the DTH means are computed the genomic effects obtained using model [1] under CV00 scheme are added.

		G_1	G_2	G_3	G_4	Mean
Observed	E_1	65	69	71	66	67.75
	E_2	80	85	87	90	85.50
	E_3	100	105	108	110	105.75
Predicted	E_1	68	69	73	70	70.00
	E_2	77	83	85	88	83.25
	E_3	104	103	110	114	107.75
E-Mean	E_1	70.7	70.3	69	70	69.98
	E_2	85.3	83.3	82.7	81.7	83.24
	E_3	109	109	107	106	107.65

4. Figure 4. Representation of the procedure for estimating the conditional DTH environmental mean for untested genotypes in unobserved environments. The top panel contains the observed values for four genotypes (G_1, G_2, G_3 and G_4) observed in three environments (E_1, E_2, and E_3). The middle panel contains the predicted values presumably obtained with the C method assuming that these genotypes were not tested in a particular environment. This method attempts predicts DTH of tested genotypes in unobserved environments using the phenotypic information of these genotypes but observed in other environments (one genotype at a time). For example, the predicted value for G_2 in E_1 is 69 and it was computed using the information of G_2 tested in E_2 (85) and E_3 (105) via the C method. The bottom panel contains the estimated DTH means (E-Mean). These are specific to each genotype and were computed as the mean of the predicted values of those genotypes in the training set for the current environment. For example, G_2 in E_1 (70.3) was computed as the mean of the predicted values of G_1 (68), G_3 (73) and G_4 (70) in E_1.

Data availability. Genome data can be accessed from DDBJ under submission DRA008071.

REFERENCES

1. FAO, IFAD, UNICEF, WFP and WHO. 2018. The State of Food Security and Nutrition in the World 2018. Building climate resilience for food security and nutrition. Rome, FAO

2. Bernardo R (2002) Breeding for Quantitative Traits in Plants. Stemma Press, USA.

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

4. Heffner HL, Sorrells ME, Jannink JL (2009) Genomic selection for crop improvement. Crop Science 49: 1–2.

5. Jannink JL, Lorenz AJ, and Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief. Funct. Genomics 9: 166–177.

6. Crossa, J., G. de los Campos, P. Pérez, D. Gianola, J. Burgueño, J.L. Araus, D. Makumbi, R.P. Singh, S. Dreisigacker, J. Yan, V. Arief, M. Banziger, and H.-J. Braun. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724.

7. Perez-de-Castro, A.M., S. Vilanova, J. Canizares, L. Pascual, J.M. Blanca, M.J. Diez, J. Prohens, and B. Pico. 2012. Application of genomic tools in plant breeding. Curr. Genomics 13:179–195. doi:10.2174/138920212800543084.

8. de los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M.P. L. Calus, 2013 Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193: 327–345.

9. Voss-Fels KP, Cooper M, Hayes BJ (2019) Accelerating crop genetic gains with genomic selection. Theor Appl Genet 132:669–686.

10. Yin X, Kropff MJ, Aggarwal PK, Peng S, Horie T (1997) Optimal preflowering phenology of irrigated rice for high yield potential in three Asian environments: a simulation study. Field Crop Res 51:19–27.

11. Hori K, Matsubara K, Yano M (2016) Genetic control of flowering time in rice: integration of Mendelian genetics and genomics. Theor Appl Genetics 129:2241–2252.

12. Onogi A, Ideta O, Inoshita Y, Ebana K, Yoshioka T, Yamasaki M, Iwata H (2015) Exploring the areas of applicability of whole-genome prediction methods for Asian rice (Oryza sativa L.). Theor Appl Genet 128:41–53.

13. Spindel J, Begum H, Akdemir D, Virk P, Collard B, Redoña, Atlin G, Jannink JL, McCouch SR (2015) Genomic selection and association mapping in rice (Oryza sativa): effect of trait genetic architecture, training population composition, marker number and statistical

model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLoS Genetics 11: e1004982.

14. Li, X., Guo, T., Mu, Q., Li, X. & Yu, J. Genomic and environmental determinants and their interplay underlying phenotypic plasticity. Proc. Natl Acad. Sci. USA 115, 6679–6684 (2018).

15. Onogi A, Watanabe M, Mochizuki T, Hayashi T, Nakagawa H, Hasegawa T, Iwata H (2016) Toward integration of genomic selection with crop modelling: the development of an integrated approach to prediction rice heading dates. Theor Appl Genet 129:805–817.

16. Forsythe, William C., Edward J. Rykiel Jr., Rand al S. Stahl, Hsin-i Wu and Robert M. Schoolfield, 1995. A model comparison for daylength as a function of latitude and day of the year. Ecological Modeling 80:87-95.

17. Murray, M. G., and Thompson, W. F. 1980, Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res., 8, 4321–5.

18. Bolger, A. M., Lohse, M., and Usadel, B. 2014, Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics, 30, 2114–20.

19. Kawahara, Y., de la Bastide, M., Hamilton, J. P., et al. 2013, Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice, 6, 1–10.

20. Li, H. 2012, Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. Bioinformatics, 28, 1838–44.

21. McKenna, A., Hanna, M., Banks, E., et al. 2010, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res., 20, 1297–303.

22. Smithsonian Institution, 1939. Smithsonian Meteorological Tables. Lord Baltimore Press, Baltimore, MD, pp. lxxvilxxviii, pp. 211-224.

23. Hijmans, R.J. (2017). geosphere: Spherical Trigonometry. R package version 1.5-5. https://CRAN.R-project.org/package=geosphere

24. VanRaden, P.M. Efficient methods to compute genomic predictions. J. Dairy Sci. 2008, 91, 4414–4423.

25. Crossa, J., D. Jarquín, J. Franco, P. Pérez-Rodríguez, J. Burgueñoet al., 2016.Genomic prediction of gene bank wheat landraces. G3 (Bethesda) 6:1819–1834.

26. Jarquín, D., C. Lemes da Silva, R.C. Gaynor, J. Poland, A. Fritz, R. Howard, S. Battenfield, and J. Crossa. 2017. Increasing Genomic-Enabled Prediction Accuracy by Modeling Genotype × Environment Interactions in Kansas Wheat. Plant Genome 10:0. doi:10.3835/plantgenome2016.12.0130

Acknowledgements

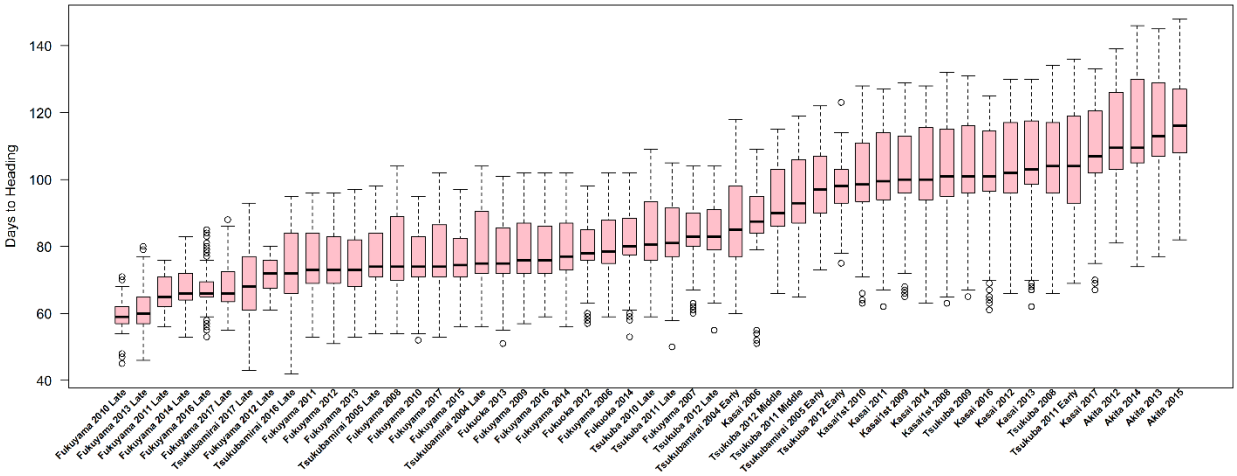
We are grateful to Dr. Vikas Belamkar for their valuable feedback and intellectual contribution during the elaboration of this manuscript.

Author Contributions

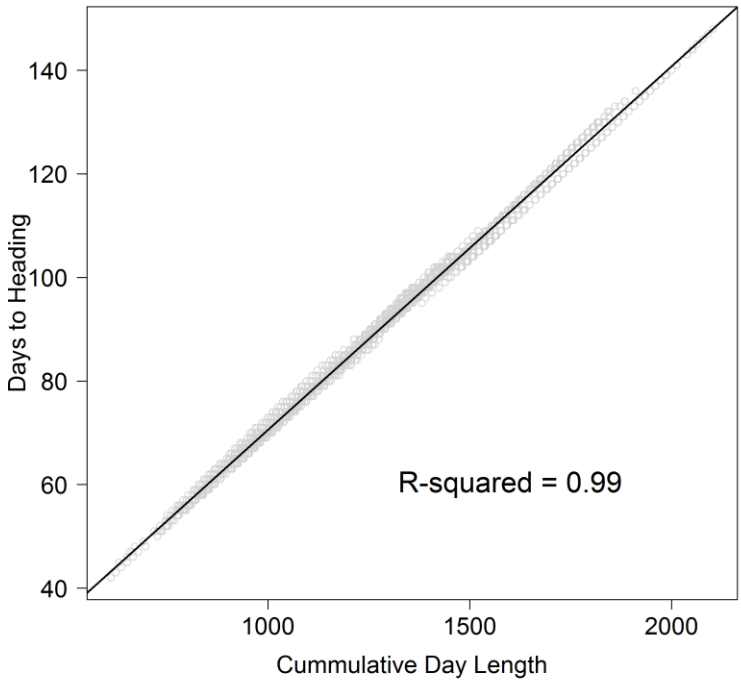
D.J. and H.I. conceptualized the study; H.K.K., S.Y., H.N., and M.Y. performed phenotyping; H.K.K. and C.T. generated and processed molecular marker, environmental and day length data sets; D.J., C.T., R.P. and H.I. analyzed the data; D.J., H.I., R.P., C.T., H.K.K and J.Y. prepared the draft M.S. and all authors contributed to finalize; all authors have read and approved the M.S.

Competing Interests: The authors declare no competing interests.

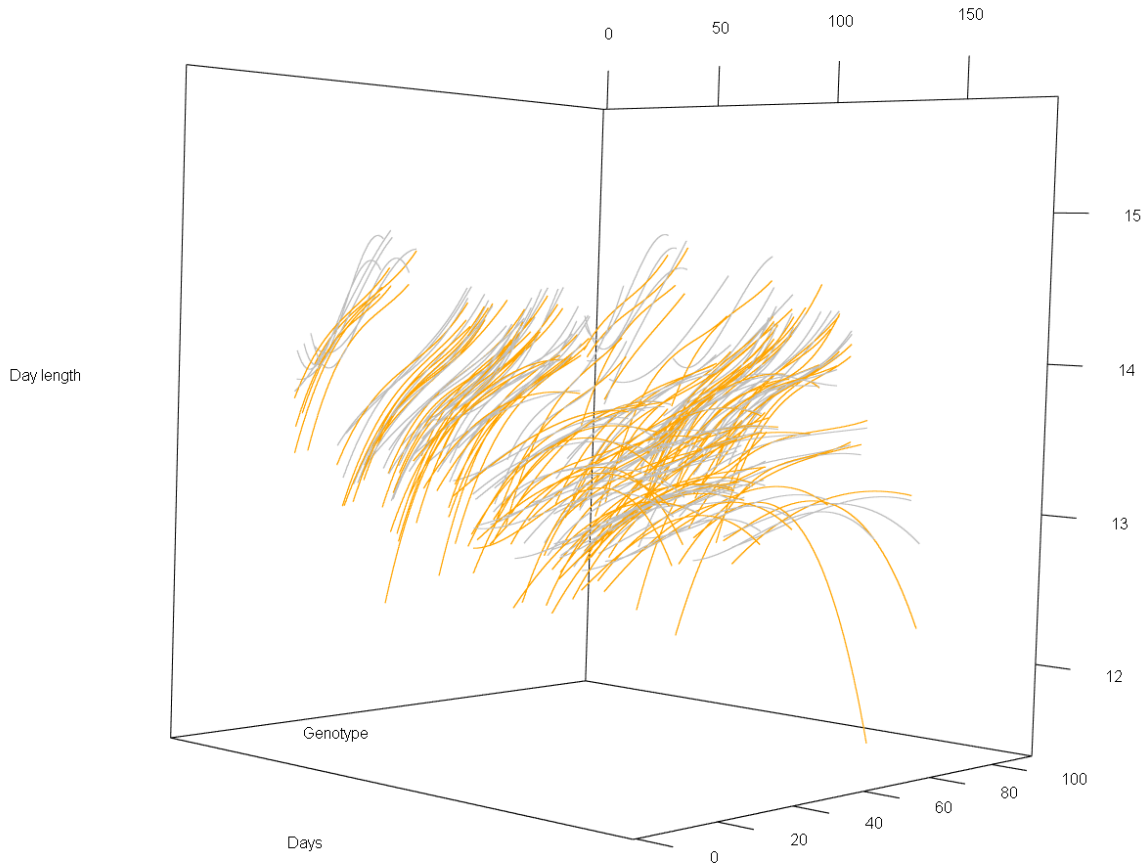
Supplementary Figures



1. Figure S1. Boxplot for days to heading (DTH) for a rice data set comprising 112 genotypes tested in 51 environments in Japan between 2005 and 2017. From the total of all possible genotype-environment combinations (112×51=5712), 94% (5365) of these were observed. Environments were ordered based on the environmental medians.



2. Figure S2. Scatter plot of days to heading (DTH) as function of the cumulative day length (CDL) from planting date until the occurrence event for a rice data set comprising 112 genotypes observed in 51 environments (in Japan between 2005 and 2017). From the total of all possible genotype-environment combinations (112×51=5712), 94% (5365) of these were observed. The black line represents the regression line between DTH and CDL (R-squared of 0.99).



3. Figure S3. Fitted curves for observed (gray) and predicted (orange) DTH (Day length-*axis*) values for a rice data set comprising 112 genotypes (Genotype-*axis*) tested in 51 environments in Japan between days 1 (planting date) and 180 (Days-*axis*). The orange color curves were fitted using the predicted values obtained with the CB method under the CV00 scheme (predicting untested genotypes in unobserved environments) while the gray color curves considered the observed data.

Supplementary Tables

4. Table S1. Rice trials for studying days to heading for 112 genotypes tested in 79 environments (site-by-year-by-planting date combinations) in multiple locations in Japan. In locations with more than 1 experiment per year different seeding times were tested.

Location	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	Latitude	Altitude
Fukuoka									1	1	1				33.21	130.49
Fukuyama			1	1	1	1	2	2	2	2	2	1	2	2	34.5	133.39
Kasai			1		1	1	1	1	1	1	1		1	1	34.88	134.86
Tsukubamirai	2	2											1	1	36.01	140.02
Tsukuba					1	1	1	3	3						36.02	140.11
Akita									1	1	1	1			39.48	140.49