Genomic prediction of agronomic traits in sugarcane using machine learning and quantitative genetics

Minoru Inamori
The University of Tokyo Graduate School of Agricultural and Life Sciences Faculty of Agriculture: Tokyo Daigaku Daigakuin Nogaku Seimei Kagaku Kenkyuka Nogakubu

Tatsuro Kimura
Toyota Motor Corporation

Masaaki Mori
Toyota Motor Corporation

Yusuke Tarumoto
National Agriculture and Food Research Organization Kyushu Okinawa Agricultural Research Center Subtropical Farming Research Team: Nogyo Shokuhin Sangyo Gijutsu Sogo Kenkyu Kiko Kyushu Okinawa Nogyo Kenkyu Center Seisan Kankyo Kenkyu Ryoiki Itoman

Taiichiro Hattori
National Agriculture and Food Research Organization Kyushu Okinawa Agricultural Research Center Subtropical Farming Research Team: Nogyo Shokuhin Sangyo Gijutsu Sogo Kenkyu Kiko Kyushu Okinawa Nogyo Kenkyu Center Seisan Kankyo Kenkyu Ryoiki Itoman

Michiko Hayano
National Agriculture and Food Research Organization Kyushu Okinawa Agricultural Research Center Subtropical Farming Research Team: Nogyo Shokuhin Sangyo Gijutsu Sogo Kenkyu Kiko Kyushu Okinawa Nogyo Kenkyu Center Seisan Kankyo Kenkyu Ryoiki Itoman

Makoto Umeda
National Agriculture and Food Research Organization Kyushu Okinawa Agricultural Research Center Subtropical Farming Research Team: Nogyo Shokuhin Sangyo Gijutsu Sogo Kenkyu Kiko Kyushu Okinawa Nogyo Kenkyu Center Seisan Kankyo Kenkyu Ryoiki Itoman

Hiroyoshi Iwata (✉ aiwata@mail.ecc.u-tokyo.ac.jp)
University of Tokyo https://orcid.org/0000-0002-6747-7036

Research Article

Keywords: Genomic selection, GRAS-Di, BLUPs, machine learning, random forest, simulation annealing ensemble (SAE)
Genomic prediction of agronomic traits in sugarcane using machine learning and quantitative genetics

Minoru Inamori¹, Tatsuro Kimura², Masaaki Mori³, Yusuke Tarumoto⁴, Taiichiro Hattori⁴⁵, Michiko Hayano⁶⁶, Makoto Umeda⁴, Hiroyoshi Iwata¹*

1. Laboratory of Biometry and Bioinformatics, Department of Agricultural and Environmental Biology, Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo, Tokyo 113-8657, Japan
2. Toyota Motor Corporation, Agriculture & Biotechnology Business Division, Turf Business Promotion Group, 1, Toyota-cho, Aichi, 471-8571, Japan
3. Toyota Motor Corporation, Environment Affairs and Engineering Management Division, CN Advanced Engineering Development Center, 1-4-18, Koraku, Bunkyo-ku, Tokyo 112-8701, Japan
4. NARO Kyushu Okinawa Agricultural Research Center, Tanegashima Sugarcane Breeding Site, 1742-1 Anno, Nishinoomote, Kagoshima 891-3102, Japan
5. NARO Kyushu Okinawa Agricultural Research Center, Head quarters, 2421 Suya, Koshi, Kumamoto 861-1192, Japan
6. NARO Institute for Agro-Environmental Science, 3-1-3 Kannondai, Tsukuba, Ibaraki 305-8604, Japan

*Corresponding Author:

Hiroyoshi Iwata

hiroiwata@g.ecc.u-tokyo.ac.jp
ORCID

Minoru Inamori

Hiroyoshi Iwata 0000-0002-6747-7036

Acknowledgments

This research was supported by cooperative research funds from the Toyota Motor Corporation to the University of Tokyo and partly by cooperative research funds from the Toyota Motor Corporation to the NARO Kyushu Okinawa Agricultural Research Center and was partially supported by JSPS KAKENHI Grant Number JP22H02306.
Abstract

To improve the efficiency of sugarcane breeding, introduction of genomic selection (GS) into breeding programs is anticipated. For sugarcane cultivars with highly heterozygous polyploid genomes, non-additive genetic effects must be considered in genomic prediction models. In this context, the use of machine learning techniques is an effective approach to model nonadditive genetic effects. In addition, pedigree information that tracks sugarcane breeding lineages can be used to calculate the degree of association among genotypes and as input for predictive models. In the present study, we used the breeding population used in a KARC/NARO sugarcane breeding program in Japan as experimental material and verified the following attributes. Specifically, the prediction accuracy of BLUP was validated based on genomic and pedigree information. In addition, a new machine learning method that can explicitly process the interaction between whole-genome markers and genetic background was developed, and comparative verification was performed using the conventional random forest method. Pedigree information showed high predictive accuracy, and when columns of the numerator relationship matrices were used as input for the machine learning models, their accuracy was identical or superior to that of BLUP depending on traits. Finally, we confirmed the effectiveness of the proposed machine learning method using interactions among markers as a model. Overall, these GS methods are expected to improve sugarcane breeding efficiency.

Keywords: Genomic selection, GRAS-Di, BLUPs, machine learning, random forest, simulation annealing ensemble (SAE)

Key message

Using numerator relationship matrix columns as input, the prediction accuracy of our machine
learning model was identical or superior to that of BLUP. Our method can explicitly model marker interactions.
Introduction

Sugarcane is a major industrial crop cultivated in the tropical and subtropical regions of the world. Sugarcane accounts for nearly 80% of the global sugar production and is a major source of sugar (https://www.isosugar.org/sugarsector/sugar). In recent years, sugarcane has attracted attention not only as a food source but also as a sustainable source of bioenergy. Research has progressed from first-generation ethanol production using sucrose to second-generation ethanol production through saccharification of cell wall sugars (Kandel et al. 2018; Calderan-Rodrigues et al. 2021). In addition, sugarcane biomass can be used for the production of high-value compounds, such as hydrogen, cellulose, lignin-derived products, organic acids, and bioplastics (Klein et al. 2019). Despite the potential and importance of sugarcane, however, recent improvements in its productivity have been trivial compared with those in other major crops. The rate of improvement in genetic gains from sugarcane breeding programs has stagnated because of narrow genetic bases, poor synchronization and fertility of flowers, long breeding cycle, low heritability of key agronomic traits, and high complexity and large size of the sugarcane genome (Kandel et al. 2018; Yadav et al. 2020). Thus, the efficiency of sugarcane breeding must be improved to fully exploit its potential and realize novel applications, such as biomass and energy production.

Genomic selection (Meuwissen et al. 2001) has recently garnered considerable attention as a method for improving the efficiency of plant and animal breeding (Hickey et al. 2017; R2D2 Consortium et al. 2021). In GS, the breeding value (or genotypic value) of a target trait is predicted based on genome-wide marker polymorphisms and plants are selected based on the predicted value. Although field trials are essential to construct and update predictive models for the genetic values of genotypes (lines/individuals), these are expected to shorten the breeding cycle because field trials are not required at the time of selection (Heffner et al. 2009; Jannink et al. 2010; Lorenz
et al. 2011). GS is known to be suitable for improving traits with low or moderate heritability that are dominated by a large number of genes. As mentioned above, a long breeding cycle and low heritability of major agronomic traits in sugarcane hinder breeding efficiency. The advantages of GS over conventional selection methods are significant, and the introduction of GS into sugarcane breeding programs is expected to significantly contribute to the genetic improvement of this crop (Yadav et al. 2020; Voss-Fels et al. 2021; Mahadevaiah et al. 2021).

However, the highly heterozygous polyploid genomes of sugarcane, which have been a barrier to conventional breeding, remain a hindrance in the application of GS in breeding programs. Such genomic characteristics render the application of high-throughput genotyping difficult; thus, careful data analysis with special algorithms is imperative (García et al. 2013). Even if genotyping is performed accurately, the mode of inheritance of a trait of interest may be complex. In addition to additive ones, nonadditive genetic factors may dominate the trait owing to the dominance and/or epistatic effects (Yadav et al. 2020, 2021). Therefore, technologies and methods must be developed to address these issues when applying GS in sugarcane breeding programs. To date, several studies have assessed the potential of GS in sugarcane production. For instance, in a previous study (Gouy et al. 2013), genomic predictions of 10 traits were generated using 167 genotypes from each of the two genetic resource panels representing the genetic diversity of sugarcane worldwide. A total of 1,499 diversity arrays technology (DArT) markers were used in that study, and the accuracy of genomic prediction was approximately 0.1–0.6. In another study (Deomano et al. 2020), prediction models were constructed for two yield traits using three panels (467, 1146, and 738 genotypes) from breeding populations. Approximately 50,000 single-nucleotide polymorphisms (SNPs) from the Affymetrix Axiom array were used as markers, and the prediction accuracy was approximately 0.2–0.5. Furthermore, Hayes et al. (2021) constructed models to predict yield- and flowering-related traits using 3,984 genotypes. The
authors used 26,000 markers from the Affymetrix Axiom array and achieved a prediction accuracy of 0.2–0.5. Yadav et al. (Yadav et al. 2021) constructed a prediction model for three yield traits based on 3,006 genotypes using 26,000 markers from the Affymetrix Axiom array and obtained a prediction accuracy of 0.2–0.5. Islam et al. (2021) constructed models to predict susceptibility to two major diseases based on 432 genotypes using 8,825 coding region-based SNPs and achieved a prediction accuracy of 0.1–0.5. Although most of the above results were obtained for different materials and traits, all studies highlighted the potential for genomic prediction. In addition, Aono et al. (Aono et al. 2020) constructed models to classify brown rust-resistant and -susceptible groups based on 219 full-sib genotypes using 14,540 SNPs; the classification accuracy was up to 95%, with 131 SNPs selected using feature selection methods. Voss-Fels et al. (Voss-Fels et al. 2021) simulated GS breeding in sugarcane and examined its effectiveness by considering the economic value of genetic improvement and cost of genotyping. The authors noted that introducing GS into sugarcane breeding produced a significant effect. Collectively, these reports suggest high potential for genomic selection in sugarcane.

In sugarcane, which possesses a highly heterozygous polymorphic genome, nonadditive genetic effects, that is, dominant and epistatic effects, must be considered in genomic prediction models (Yadav et al. 2021). Yadav et al. (2021) and Islam et al. (2021) showed that the accuracy of models that considered nonadditive genetic effects, albeit only for some traits, was higher than that of models that considered additive effects alone. Islam et al. (2021) demonstrated that the use of machine learning methods is an effective way to model nonadditive genetic effects (Bayer et al. 2021). Typically, machine learning methods possess excellent potential to incorporate nonadditive effects and may be useful for genomic selection in sugarcane. However, only a few studies (e.g., Islam et al. 2021) have applied machine learning methods for genomic prediction in sugarcane. In the present study, we examined the effectiveness of a novel machine learning
method, in addition to a conventional method (random forest).

Sugarcane offers certain advantages over other crops in terms of genomic selection. As sugarcane breeding rarely uses inbred lines derived from repeated selfing and backcrossing, which are generally used in autogamous crops, the pedigree of breeding lines can be easily trace based on the relationship between parents and offspring. By tracing pedigrees, the degree of relatedness between genotypes can be calculated. Such pedigree relationships have been described for various crops to provide additional information for predicting the breeding (or genotypic) values and improving the accuracy of GS (de los Campos et al. 2009; Robertsen et al. 2019; Howard et al. 2019). In sugarcane, Deomano et al. (2020) conducted genomic prediction with and without pedigree information and demonstrated that this information can improve the accuracy of GS. Although pedigree information may be useful for improving the accuracy of GS, several innovations are necessary to use this information in a machine learning method. This is because in genomic prediction, pedigree information is typically used as a matrix representing the relationships among genotypes (i.e., numerator relationship matrix) and not in the form of ordinary multivariate data. A simple yet effective approach to incorporate pedigree data may be to consider the relationship between a genotype and the others, that is, to consider each column of the numerator relationship matrix as a variable and use it as an input for the machine learning model. In other words, for a single genotype (line/individual) for which the breeding (or genotypic) value is to be predicted, a vector with the same dimension as the number of genotypes in the training population is provided, which becomes the input vector for the machine learning model. Another possible method is to consider the numerator relationship matrix as a Gram matrix, estimate the arrangement of genotypes in a multidimensional space that realizes the Gram matrix, and use the coordinates in the multidimensional space as input for the machine learning model. These methods may be useful for combining pedigree information with genome-wide marker
genotypes in machine learning-based genomic prediction model; however, their potential has not been evaluated to date.

Of note, in genomic selection, different target populations present different levels of genetic variance and patterns of linkage disequilibrium (LD), resulting in different prediction accuracies, even for the same trait (Schopp et al. 2017). To date, the accuracy of genomic prediction in sugarcane breeding populations has only been evaluated in studies using breeding populations in Australia (Deomano et al. 2020; Hayes et al. 2021; Yadav et al. 2021). Thus, one of the objectives of the present study was to determine whether GS could be used in a Japanese sugarcane breeding program, as described in detail in the Materials and Methods section. Although high sugar content and high yield remain the most important breeding goals worldwide, improving adaptability to cultivation on the southwestern islands of Japan is also an important breeding goal. In this case, adaptation to the natural and commercial environments of the islands must be taken into account. Even within the same island, there are diverse soil types, and depending on the soil type, the same genotype may exhibit different phenotypes. In addition to confirming the potential of GS under the island environment, the usefulness of data routinely collected in the current breeding programs to build genomic prediction models must be examined.

Based on the perspectives described above, we examined the following attributes using a breeding population from the third round of selection in the KARC/NARO breeding program as the experimental material: (1) validation of the prediction accuracy of BLUP based on genomic and pedigree information of the population; (2) investigation of the use of pedigree information in machine learning models; and (3) proposal of a novel machine learning method that can explicitly handle the interaction between genome-wide markers and pedigree-based genetic backgrounds. Finally, we summarize the results of these studies and discuss the effectiveness of genomic selection in future sugarcane breeding.
Materials and Methods

A regular KARC/NARO sugarcane breeding program

In the present study, we applied genomic prediction to breeding populations in the third round of selection in a KARC/NARO sugarcane breeding program. Selection in the breeding program was conducted in multiple rounds using the steps described below.

- **First round of selection**: In this round, approximately 30,000 individuals (seedlings) derived from several crosses were selected. Preliminary selection was conducted when the individuals grew into mature plants. Approximately 3,000 individuals were selected after removing those exhibiting poor growth. In the main selection, 1,000–1,200 individuals were selected based on the Brix value.

- **Second round of selection**: Individuals selected in the first round were treated as lines from this round. Approximately 10 buds from each individual selected in the first round were planted. Stalk count and soluble solid content are quantitatively evaluated two times a year in a field trial. Subjective qualitative evaluations of stalk length, count, and diameter were conducted two or three times a year. Based on this evaluation, approximately 150 lines were selected.

- **Third round of selection**: Field trials were conducted at two fields with different soil conditions on the Tanegashima Island and at one field on the Tokunoshima Island. The lines were evaluated as explained in the next subsection. Approximately 50 lines were selected with this experiment. As described later, in the present study, we analyzed the data of breeding populations evaluated at the two fields on the Tanegashima Island in 2019 and 2020.

- **Fourth round of selection**: Lines selected in the third round were tested for their yield potential and quality of the harvested product. A preliminary test for disease resistance was
conducted in parallel during this round.

- Adaptability test of the selected lines: Fifteen lines selected in the fourth round were tested at six locations, including the Tanegashima Island, for 1–2 years.
- Performance test for recommended varieties: A few lines that were superior in the adaptability test were assessed to determine whether they should be recommended for commercial production through several cycles of 2-year field trials.

**Field trials for the third round of selection**

As described above, in the present study, we analyzed phenotypic and marker genotype data collected from lines tested in the third round of selection of the KARC/NARO sugarcane breeding program.

Field trials of the lines were conducted at the Tanegashima Sugarcane Breeding Site of KARC/NARO (30°43′N, 131°04′E, altitude 45 m) from March 2018 to January 2019 (trial 1, T1) and from April 2019 to February 2020 (trial 2, T2). The lines evaluated in T1 were the third-round selection lines from the KARC/NARO breeding program and control cultivars (‘NCo310’, ‘NiF3’, ‘NiF8’, ‘NiTn18’, and ‘Kurokaido’). Approximately 150 third-round selection lines in T1 and T2 were designated as K15 and K16, respectively. Approximately 20 buds were planted in a plot of 2.5 m × 1.1 m size (73,000 plants·ha⁻¹) with one plot for each soil condition.

In T1, basal fertilization and planting were performed on March 30, 2018. Fertilizer was applied two times: during mid-June and mid-July. In T2, base fertilization and planting were performed on April 1, 2019. Additional fertilization was applied on May 30 and June 10. In both trials, the number of stalks in each plot was measured during the vigorous growth (July), autumn (October–November), and harvest (late January–early February) periods. Stalk length was measured in five medium-sized stalks in each plot during the vigorous growth period and at the
time of harvest. Stalk diameter was measured in five medium-sized stalks in each plot at the time of harvest. Soluble solid content, measured as Brix level, was measured in the middle and upper parts of a stalk (SSC.M and SSC.U) on three medium-sized plants in each plot using a hand-held refractometer (Master, Atago, Japan) during the autumn period and at the time of harvest.

**Soil types**

The third round of selection was performed in two fields with different soil types: “Kuroboku” and “Akahoya.” Soil distributed in the Tanegashima Island is Andosol (“Kuroboku”), which is a type of volcanic ash soil, while soil distributed in Japanese sugarcane-producing regions is non-volcanic ash soil (Inoue 2018). “Kuroboku” is characterized by high productivity due to high organic matter content. “Akahoya” is another type of volcanic ash soil but presents a lower organic matter content than “Kuroboku”. The stalk length of plants growing in “Akahoya” was consistent with that of plants growing on the Tokunoshima Island, where non-volcanic ash soil is distributed (Sugimoto et al. 1999).

**Genotyping system**

Sugarcane leaf samples (80–100 mg) were cut and used for DNA extraction. The sampled leaf tissues were frozen in liquid nitrogen and placed in 2 mL microcentrifuge tubes with three tungsten carbide beads (3 mm in diameter). The tissues were then crushed using a TissueLyser (QIAGEN) two times for 30 s each at 27.0 vibrations per second. DNA was extracted from the obtained tissue powder using DNeasy Plant Mini Kit (QIAGEN).

Genotyping was performed using GRAS-Di (Genotyping by Random Amplicon Sequencing, Direct), which was developed by Toyota Motor Corporation (Toyota, Japan). Details of the method are provided in the patent (Enoki and Takeuchi 2019).
Libraries for next-generation sequencing (NGS) were prepared using two rounds of PCR with template DNA from each sugarcane strain. The first PCR was performed using a primer mix of 12 random primers comprising 13-mer sequences with 15 ng · μL⁻¹ of template DNA. The second PCR was performed in 50 μL reaction mixture containing 1.5 μL of the first PCR product from each strain and primers for the index sequence. The second PCR products were mixed at an equal volume in a tube, of which 50 μL was purified using the MinElute PCR Purification Kit (QIAGEN) and used as a sequence library.

The sequence library was analyzed for each amplicon sequence using a next-generation sequencer (HiSeq 4000, Illumina) under paired-end conditions, with a read length of 100 bases. Genotyping of single nucleotide polymorphisms (SNPs) was performed from NGS-derived sequences using GRAS-Di software (Toyota Motor Corporation, Aichi, Japan), which is currently commercially available. SNP markers with a quality of D or higher were selected using the software, and markers showing the same genotype were integrated, resulting in 33,149 markers.

**Genetic relationships among lines**

Genetic relationships among lines were calculated using a pedigree-based additive relationship matrix (A) and a genome-wide marker-based additive relationship matrix (G). We calculated matrix A based on the sugarcane pedigree database (Tarumoto et al. 2016). The calculation of matrix A requires the pedigree information of lines in the breeding populations in the third round of selection as well as their ancestors. We calculated matrix A with 3,892 lines, including 3,595 ancestral lines and 297 lines in the breeding population. As sugarcane has a high ploidy level but is allopolyploid, we considered the lines to be diploid in the calculation of matrix A. We used the Amatrix function of R AGHmatrix (Amadeu et al. 2016) for calculations. The “ploidy” option of the Amatrix function was set to 2 such that the genome was considered diploid.
Matrix G was calculated as $G = \frac{XX'}{M}$, where $X$ is the standardized genotype matrix (see below) and $M$ is the number of SNPs (i.e., number of columns in matrix $X$). The elements of $X$ in the $i^{th}$ column are the genotypic scores of marker $i$, which were standardized to have zero mean and unit variance, that is, $\sum_{j}^{N} x_{ij} = 0$ and $\sum_{j}^{N} x_{ij}^2 / N = 1$ over $N$ lines.

To summarize and visualize genetic relationships among the lines, we applied principal component analyses based on matrices A and G. We assumed matrices A and G to be Gram matrices, that is, matrices whose $(i, i')$-element is the inner product of vectors representing $i$ and $i'$ lines, multiplied by a constant. The $(i, i')$-element of matrix G can be represented as $g_{ii'} = \frac{x_i x_{i'}}{M}$. In other words, matrix G is a Gram matrix multiplied by $1/M$. With the eigenvalue decomposition of matrix G, we can perform principal component analysis based on $X$. Matrix G can be obtained by eigenvalue decomposition as $G = U_g \Lambda_g U_g'$, where $U_g$ is a matrix of the eigenvectors ($i^{th}$ column is eigenvector $u_{gi}$) and $\Lambda_g$ is a diagonal matrix with eigenvalues as diagonal elements ($i^{th}$ diagonal element is eigenvalue $\lambda_{gi}$). The $i^{th}$ principal component score obtained from the principal component analysis of $X$ can be expressed as $\sqrt{M \lambda_{gi}} u_{i}$. As opposed to matrix G, matrix A is not a Gram matrix multiplied by a constant. However, by considering matrix A as a Gram matrix, eigenvalue decomposition was performed as $A = U_a \Lambda_a U_a'$. The principal component scores were calculated as $\sqrt{\lambda_{ai}} u_{ai}$, where $u_{ai}$ is the $i^{th}$ column of $U_a$ and $\lambda_{ai}$ is the $i^{th}$ diagonal element of $\Lambda_a$.

**Estimation of genotypic values, ANOVA, and heritability**

The genotypic values of 297 lines were estimated using a mixed model. Since differences in soil conditions are known significantly affect the phenotypes of sugarcane (Martini et al. 2020), two methods were used to estimate genotypic values: one for estimating genotypic
values in each soil type (Andosoil or Akahoya) and the other for estimating average genotypic values over two soil types. The mixed model was set as follows:

\[ y_{ijkl} = m + g_i + t_j + s_k + g_{sk} + e_{ijkl} \]

where \( y_{ijkl} \) is the phenotypic value of observation \( l \) (\( l = 1 \) for stalk number and \( l = 1, 2, \ldots, 5 \) for other traits) of genotype \( i \) (\( i = 1, 2, \ldots, 297 \)) in yearly trial \( j \) (\( j = 1, 2 \)) in soil type \( s \) (\( s = 1, 2 \)); \( m \) is the overall average; \( g_i \) is the random effect of genotype \( i \); \( t_j \) is the fixed effect of yearly trial \( j \); \( s_k \) is the fixed effect of soil type \( k \); \( g_{sk} \) is the random effect of the interaction between genotype \( i \) and soil type \( k \); and \( e_{ijkl} \) is the random error in observation \( l \) of genotype \( i \) in soil type \( k \) in trial \( j \). The genotypic value of genotype \( i \), \( \eta_i \), was estimated as \( \hat{\eta}_i = \hat{m} + \hat{g}_i \), where \( \hat{m} \) is the best linear unbiased estimate (BLUE) of \( m \) and \( \hat{g}_i \) is the best linear unbiased predictor (BLUP) of \( g_i \). Furthermore, we estimated the genotypic value of genotype \( i \) for soil type \( k \), \( \eta_{ik} \), as \( \hat{\eta}_{ik} = \hat{m} + \hat{g}_i + \hat{g}_{sk} \), where \( \hat{g}_{sk} \) is the BLUP of \( g_{sk} \).

Moreover, we applied ANOVA to evaluate the effects of genotype and the interaction effect between genotype and soil type. We analyzed data from each yearly trial separately and assumed that the genotypic effect and interaction effect between genotype and soil type were both fixed effects. The ANOVA model was set as follows:

\[ y_{ijkl} = m_j + g_{ij} + s_{jk} + g_{ijk} + e_{ijkl} \]

where \( m_j \) is the overall average in trial (year) \( j \); \( g_{ij} \) is the fixed effect of genotype \( i \) in trial \( j \); \( s_{jk} \) is the fixed effect of soil type \( k \) in trial \( j \); \( g_{ijk} \) is the fixed effect of the interaction between genotype \( i \) and soil type \( k \) in trial \( j \); and \( e_{ijk} \) is the random error of genotype \( i \) for soil type \( k \) in trial \( j \). For stalk number, the number of observations was one per plot; thus, the interaction term was used as the error term to test the main effects of genotype and soil type.

Heritability was estimated based on the mixed model described above. After fitting the model, we obtained variance estimates for \( g_i \), \( g_{sk} \), and \( e_{ijkl} \), that is, \( \sigma_g \), \( \sigma_{gs} \), and \( \sigma_e \),
respectively. Heritability was estimated as follows:

\[ h^2 = \frac{\hat{\sigma}_g}{\hat{\sigma}_g + \hat{\sigma}_{gs}/K + \hat{\sigma}_e/KL} \]

where \( K \) is the number of soil types (\( K = 2 \)) and \( L \) is the number of observations per plot (\( L = 1 \) for stalk number and \( L = 5 \) for other traits). Additionally, we calculated pedigree- and genomic-based heritability. In the calculation of the former, we assumed that the random genotypic effect had a variance structure proportional to matrix \( A \), that is, \( \mathbf{g} \sim N(\mathbf{0}, \mathbf{A}\sigma_g^2) \), where \( \mathbf{g} = (g_1, ..., g_K)^T \) and \( \sigma_g^2 \) are the variances of the genotypic effect; meanwhile, we assumed that the genotypic effect was independent among lines, that is, \( \mathbf{g} \sim N(\mathbf{0}, \mathbf{I}\sigma_g^2) \), in the mixed model mentioned above. In the calculation of the latter, we assumed that the genotypic effect had a variance structure proportional to matrix \( G \), that is, \( \mathbf{g} \sim N(\mathbf{0}, \mathbf{G}\sigma_g^2) \).

**Prediction methods and models**

**A and G BLUP**

Models for predicting the genotypic values \( \eta_l \) and \( \eta_{ik} \) were constructed using BLUP and two machine learning methods: random forest (RF) and simulation annealing ensemble (SAE).

The BLUP model was set as follows:

\[ \eta = \mu + Z\alpha + Z\beta + Z\gamma + \epsilon \]

where \( \eta \) is the vector of genotypic values whose \( i \)th element is \( \eta_l \) or \( \eta_{ik} \); \( \mu \) is the vector whose elements are the overall mean; \( \alpha \) is the vector of genetic effects accounted by genomic-based relationships, that is, \( \alpha \sim N(\mathbf{0}, \mathbf{G}\sigma_a^2) \), where \( \sigma_a^2 \) is the variance of the genomic-based effect; \( \beta \) is the vector of genetic effects accounted by pedigree-based relationships, that is, \( \beta \sim N(\mathbf{0}, \mathbf{A}\sigma_a^2) \), where \( \sigma_a^2 \) is the variance of the pedigree-based effect; \( \gamma \) is the vector of genetic effects accounted by the interaction between genomic- and pedigree-based relationships, that is, \( \gamma \sim N(\mathbf{0}, \mathbf{G} \otimes \mathbf{A}\sigma_{\alpha\gamma}^2) \), where \( \sigma_{\alpha\gamma}^2 \) is the variance of the interaction effects and \( \otimes \) is the
element-wise product (Hadamard product) of matrices; \( Z \) is the design matrix for the genetic effects; and \( \epsilon \) is the vector of residuals assuming \( \epsilon \sim N(0, 1\sigma^2_e) \), where \( \sigma^2_e \) is the residual variance. This model incorporating the interaction terms between genomic and pedigree relationships was first proposed by Howard et al. (Howard et al. 2019) and is hereafter referred to as the G×P-BLUP model. In addition, we applied the following three reduced models of G×P-BLUP:

G-BLUP:
\[
\eta = \mu + Z\alpha + \epsilon
\]

P-BLUP:
\[
\eta = \mu + Z\beta + \epsilon
\]

GP-BLUP:
\[
\eta = \mu + Z\alpha + Z\beta + \epsilon
\]

To estimate model parameters and calculate BLUP values, we used the R package BGLR (Pérez and de los Campos 2014). Hyperparameters for the estimation and calculation were set to the default settings of the BGLR function. The number of Markov chain Monte Carlo (MCMC) iterations for the total and burn-in periods was set to 5,000 and 15,000, respectively.

RF

RF (Breiman 2001) is an ensemble algorithm based on randomized regression trees. In RF, each regression tree is constructed using randomly selected explanatory variables based on randomly selected samples with replacement (i.e., bootstrap samples) from the training dataset. The final prediction is computed by averaging the predictions of all regression trees in the forest. In the present study, we constructed an RF regression model using a relationship matrix computed from genome-wide marker or pedigree data (G or A) as the input. Specifically, we considered one
column of matrix G or A as an explanatory variable and input it into the RF model that predicted the genotype value of a trait. This model predicted the genotype value based on the genetic relationships between the target (i.e., lines whose genotypic values were to be predicted) and other (i.e., lines in training dataset) lines. For comparison, we also constructed an RF model using principal component scores obtained from the eigenvalue decomposition of matrix G or A. Principal component scores are described in the previous section. The principal component scores computed from matrix G or A were used as input for the RF model. For genome-wide genotype marker data, we also assigned the marker genotype score $X$ as an input to the RF; this is an usual method of using RF for genomic prediction. To estimate the parameters of the RF models, we used the R package “ranger” (Wright and Ziegler 2017).

**SAE**

In modeling epistasis, a possible approach is to explicitly model the interactions between markers. However, when the number of markers is high, such models must consider epistasis from a massive number of marker combinations, even if they are limited to second-order additive–additive epistasis. For instance, in the present study, approximately 30,000 DNA markers were used, and the number of combinations of these markers was approximately 4.5 billion. From such a large number of candidate variables, we developed a model to narrow down and select useful combinations for prediction and build a prediction model based on them.

In this method, we must first determine the number of variables that represent the main effect and the number of variables that represent the interaction. In the present study, the number of each variable group (main effect or interaction) was set to 10. Subsequently, a simulated annealing algorithm was applied to select important variables. Specifically, we repeated the cycle involving the following steps.
Step 1: We randomly selected 20 variables from all. Of the 20 variables, 10 represented the main effect and 10 represented the interaction effect.

Step 2: We calculated the energy of the combination of variables selected in step 1 as the normalized residual sum of squares of multi-linear regressions with the 20 selected variables. Specifically, \[ E = \frac{10n \| \hat{\mu} - \hat{\gamma} \|^2}{y_w^2n}, \] where \( n \) is the number of samples and \( y_w = \max(y) - \min(y) \).

Step 3: At each step, we randomly selected a variable to be removed from the selected variables and another variable from the unselected variables as the substitute of the removed variable. When the removed variable represented the main effect, the substitute variable was selected to represent the main effect. When the removed variable represented the interaction effect, the substitute variable was selected to represent the interaction effect.

Note that there are two types of interaction. For instance, if the variables are \( x \) and \( y \), there are two types of interactions: \( xy \), which is the product of the variables, and \( x + y - xy \), which includes the main effect. Here, the values of \( x \) and \( y \) are either 0 or 1, such that they are both the same; however, this assumption is to consider the case in which the variables have continuous values.

Step 4. We determined the initial temperature such that the probability of state transition was 1/2 when the energy of the next state was higher than that of the initial state. Specifically, we randomly obtained 1,000 energy differences between the initial and next states and obtained the median of energy differences from trials on a set of positive (increasing) energy differences. Here, the median is represented by \( dE \).
The probability of the state transition is \( \exp\left(-\frac{dE}{T}\right) \). Then \( T_{init} = \frac{dE}{\log 2} \).

Step 5. We reduced the temperature such that it was proportional to difference from 0° and reached 1° at the 5,000th step.

\[
T = T_{init} \exp\left(-\log T_{init} \frac{t}{t_{max}}\right)
\]

The time \( t \) progresses at increments of 1 from 0 to \( t_{max} = 5000 \).

Step 5a. We completed the step when the variables did not change in 500 steps.

Step 6. Finally, we extracted the set of variables with the minimum energy.

These steps of simulated annealing were repeated 1,000 times.

Predictions were generated through multiple regression on each variable set obtained using simulated annealing. The predictions were then weighted and averaged. Here, weighting can be performed using any method. In the present study, the coefficient of determination of each multiple regression was used. The advantage of this weighting method is that reliable results can be obtained without adjusting hyperparameters.

Let \( p_i \) be the result of multiple regression on each set of variables; then, the prediction value \( p \) is calculated as follows:

\[
p = \frac{p_i r_i^2}{\sum_i r_i^2}
\]
Validation of prediction accuracy

To evaluate the prediction accuracy of each method, cross-validation was performed according to two schemes. The first cross-validation scheme was based on random partitioning of all genotypes into 10-folds (10-fold cross-validation). In this scheme, nine folds were used to build the prediction model, and the remaining fold was predicted using the obtained model and used to evaluate the prediction accuracy. This process was repeated such that all folds could be subjected to prediction and evaluation. Since random splitting causes stochastic fluctuations in the values of prediction accuracy indices (see below), the same split was applied in all methods. In addition, five repetitions of random split were applied to deal with fluctuations.

The second scheme was based on leave-one-out cross-validation. The purpose of this cross-validation scheme was to evaluate prediction accuracy for the genetic ability of offspring resulting from new mating combinations of parents. As described above, 297 genotypes belonged to 87 families (Fig. S1). We built a prediction model using data from 86 families, excluding one, and evaluated prediction accuracy for genotypes included in the left-out family using the obtained model. Since partitioning was performed deterministically and no stochastic fluctuation occurred in this scheme, this cross-validation was performed only once.

Prediction accuracy was evaluated based on Pearson’s product moment correlation coefficient between the predictions and BLUPs of genotypic values. The predicted values calculated in one repetition of cross-validation were aggregated for all genotypes, and the aggregated values were compared with the corresponding BLUPs of the genotypic values. Thus, in the 10-fold cross-validation, five values of the correlation coefficient were calculated for five repetitions, whereas one value was calculated in the leave-one-out cross-validation. For the 10-fold cross-validation, we calculated the average of five values to summarize the results over five repetitions.
Results

Genetic relationships among individuals in the breeding population

In the present study, we calculated a genomic relationship matrix based on polymorphisms of genome-wide markers as well as a pedigree-based relationship matrix based on parent–offspring relationship information of 297 genotypes used in the analysis. These matrices are expected to represent the genetic relationships among genotypes from different perspectives. Figure 1 shows a scatter plot of principal component scores based on the genomic and pedigree-based relationship matrices; families containing more than 10 individuals are colored. Genotypes belonging to 87 different families were derived from combinations of parents with various genetic backgrounds. Of note, in principal component analysis based on the pedigree-based relationship matrix, genotypes belonging to each family are represented as a single point.

Pearson’s product-moment correlation coefficients ($r$) between the first five principal components of both matrices were calculated to assess genetic relationships. The proportion of variation explained by the first five principal components was 63.8% in the genomic relationship matrix and 40.9% in the pedigree-based relationship matrix. Consequently, although the correspondence of one principal component to another was not clear, there was an association between the two groups of principal components. Moderate ($0.4 < |r| \leq 0.7$) to strong ($|r| > 0.7$) correlations were noted in various combinations (Fig. 2). For instance, the first principal component based on the pedigree-based relationship matrix (PPC1) was correlated with the first principal component of the genomic relationship matrix (GPC1) at $|r| = 0.52$. In addition, PPC2 was correlated with GPC1, GPC2, and GPC3 at $|r|$ values of 0.40, −0.53, and 0.46 respectively. PPC3 was correlated with GPC3 at $|r| = 0.77$, and PPC5 was correlated with GPC4 at $|r| = 0.83$. These results suggest that the two relationship matrices captured the
common patterns of genetic relationships among genotypes, although they were not completely consistent and likely captured different aspects as well.

**Analysis of variance and heritability**

As shown in Table 1, analysis of variance with genotype and soil type as the main effects showed that the genotype effect was highly significant for all traits in both trials (T1 and T2). In particular, SD and SL showed relatively large $F$ values in both trials, suggesting large genetic variations in these traits. The soil-type effect was significant for all traits except one (SSC.M in T18). The effect was on SL particularly significant, which showed larger $F$ values in both trials, suggesting that this trait is sensitive to soil type. Moreover, the genotype $\times$ soil type interaction was highly significant for all traits. A significance test on the interaction could not be performed for SN because the trait was measured on a plot basis, and the interaction term was set as the error term.

Heritability was calculated by applying a linear mixed model with the genotype effect as the random effect and trial and soil effects as fixed effects. Variance of the genotype effect was estimated under the assumption that BLUPs of the genotype effect independently follow an identical normal distribution and that BLUPs have a covariance structure that follows the pedigree-based or genomic relationship matrix. Similar estimates of heritability were obtained under both assumptions (Fig. 3). Heritability estimated based on the genomic relationship matrix showed the highest values for SSC.U, SD, and SN and the lowest value for SL. The heritability of stem-related traits (SD, SL, and SN) exceeded 0.7 and was estimated to be approximately 0.8 in most cases. In contrast, the heritability of SSC was lower than 0.7 for both SSC.M and SSC.U.
Accuracy of genomic and pedigree-based predictions

Comparison BLUPs based on G and A matrices

To evaluate the usefulness of information in matrix A for genomic prediction, we evaluated the accuracy of BLUPs based on the G, A, G+A, and G×A models. First, we describe the results of 10-fold cross-validation (Fig. 4a). For SSC.M and SSC.U, the accuracy of BLUPs based on matrix A was higher than that of BLUPs based on matrix G. For the size and number of stalks (SD, SL, and SN), BLUP based on matrix G was more accurate than that based on matrix A. BLUP based on both matrices G and A (G+A and G×A) showed higher accuracy than that based solely on either matrix for all traits, except SN. For four of the five traits, G×A showed equal or higher accuracy than G+A.

In the leave-one-out cross-validation, the accuracy of BLUP based on matrix G was higher than that of BLUP based on matrix A for all traits, except SSC.U (Fig. 4b). In addition, among BLUPs based on both matrices G and A (G+A and G×A), G×A showed higher accuracy for SSC.M and SSC.U than either A or G. Thus, the usefulness of matrix A was limited in leave-one-out cross-validation.

Comparison of prediction accuracy between BLUP and machine learning methods

In the present study, we investigated several methods of using matrices A and G as inputs for a machine learning model. The results of 10-fold cross-validation are presented in Figure 5. First, comparison among models using matrix G as the input (indicated by red bars in Fig. 5) showed that machine learning models showed the same or better accuracy than BLUP. We explored three methods of inputting genome-wide marker data into RF models: using matrix G directly (RFM), using principal component scores based on eigenvalue decomposition of matrix
G (RFP), and using marker scores (RFX). RFM showed higher accuracy than RFP and RFX for all traits, except SL. In addition, the accuracy of SAE was roughly the same as that of RF, although slightly worse for all traits, except one. In SAE, genomic information was input as a marker score. SAE outperformed RFX for two of the five traits.

Next, comparison among models with matrix A as the input (indicated by green bars in Fig. 5) showed that machine learning models achieved the same or better accuracy than BLUP. For all traits, except SL, the machine learning models were more accurate than BLUP. Of note, in many cases, the accuracy was higher when matrix A was directly used as the input (RFM) than when principal component scores based on the eigenvalue decomposition of matrix A were used as the input (RFP). RF outperformed SAE for four of the five traits.

Finally, comparison between models with both A and G matrices as the input (indicated by blue bars in Fig. 5) showed that machine learning models achieved the same or better accuracy than BLUP. For RF, the accuracy was higher when both matrices G and A were used as the input than when only G or A matrix was used, although the difference was not large. RF outperformed SAE for three of the five traits.

The results of leave-one-out cross-validation are presented in Figure 6. First, comparison between models with matrix G as the input (indicated by red bars in Fig. 6) showed that for four of the five traits, machine learning models achieved the same or higher accuracy than BLUP based on matrix G in most cases. The exceptional model was RF with principal component scores based on eigenvalue decomposition of matrix G as the input (RFP), which showed much lower accuracy than BLUP based on matrix G for four of the five traits. On the contrary, for SSC.M and SSC.U, the accuracy of RF directly using matrix G as the input (RFM) was much higher than that of BLUP based on matrix G (relative accuracy of 1.5 or higher).

Next, comparison between models using matrix A as the input (indicated by green bars
in Fig. 6) showed that for four of the five traits, the machine learning models achieved the same or higher accuracy than BLUP in most cases. Similar to that in the case of models using matrix G as the input, an exceptional model was RFP using principal component scores based on the eigenvalue decomposition of matrix A as the input, which showed lower accuracy than BLUP based on matrix A for three of the five traits. In particular, it showed a negative correlation with SD. Notably, the prediction accuracy of SAE for SSC.M and SSC.U (especially the latter) was significantly higher than that of BLUP.

Finally, comparison between models with both matrices G and A as the input (indicated by blue blot in Fig. 6) showed that for four of the five traits, the machine learning models were almost always as accurate as or more accurate than BLUP. The exceptional model was RFP (RFP) using principal component scores based on the eigenvalue decomposition of matrices A and G as the input, which showed much lower accuracy than BLUP based on matrices A and G for three of the five traits. Notably, the prediction accuracy of RFM for SSC.M and SSC.U (especially the latter) was significantly higher than that of BLUP.

**Effect of soil type on prediction accuracy**

We observed a non-significant effect of soil type on prediction accuracy (Fig. 7). Moreover, the effect of soil type on prediction accuracy showed a similar pattern regardless of the method used in both 10-fold and leave-one-out cross-validation.

There was no significant difference in prediction accuracy when the data of both soil types were integrated and the average genotype values were estimated to build the model or when the genotype values in each soil type were estimated to build the model based on the data of individual soils. In some cases, prediction accuracy decreased depending on the trait.
Discussion

Genomic prediction accuracy

Genomic prediction using GBLUP showed a prediction accuracy of approximately 0.1–0.5 for all traits. However, large difference was noted depending on the cross-validation methods, and the prediction accuracy significantly decreased in leave-one-out cross-validation. This may be because in leave-one-out cross-validation, the information of full-sibs is lost in the training data and cannot be borrowed for prediction. However, 10-fold cross-validation is not an unrealistic validation scheme. In Japanese sugarcane breeding, same parental combinations are used repeatedly; therefore, in such cases, full-sib data are included in the training data.

Moreover, the prediction accuracy greatly differed depending on traits and cross-validation methods. For instance, the highest prediction accuracy in 10-fold cross-validation was recorded for SSC.M, whereas the highest accuracy in leave-one-out cross-validation was recorded for SL, LD, and SN. In addition, the heritability of SD, SL, and SN was higher than that of SSC.M and SSC.U, suggesting that the prediction accuracy better reflected heritability in leave-one-out cross-validation than in 10-fold cross-validation.

Furthermore, the usefulness of pedigree information, that is, the relative accuracy of pedigree-based prediction to genomic prediction, depends on traits and cross-validation methods. In 10-fold cross-validation, the usefulness of pedigree information was high, that is, pedigree-based prediction was as accurate as or slightly better than genomic prediction depending on the combination of traits and methods. In contrast, in leave-one-out cross-validation, the relative superiority of the accuracy of genomic prediction over that of pedigree-based prediction varied greatly depending on the combination of traits and methods. For instance, for SL, the accuracy of pedigree-based prediction was extremely low when RF was used. The advantage of simultaneously using both genomic and pedigree information for prediction is not very high and
the obtained accuracy is approximately the same as that of genomic or pedigree-based prediction.

Meanwhile, the difference in prediction accuracy between machine learning methods and BLUP was not significant; however, for some traits, machine learning methods were more accurate than BLUP. However, the lack of difference between the two methods suggests that machine learning methods using pedigree information may achieve the same accuracy as BLUP based on genomic or pedigree information.

Comparison of genomic prediction accuracy with previous reports

The accuracy of genomic prediction varies depending on the test population. In the present study, we evaluated the accuracy of genomic predictions using Japanese breeding populations. Deomano et al. (2020), Hayes et al. (2021), and Yadav et al. (2021) used Australian breeding populations, while Islam et al. (2021) used an American breeding population. In those analyses, genetic variation was approximately 0.1–0.5, similar to the prediction accuracy noted in the present study. Although prediction accuracy naturally differs among traits with different genetic systems, prediction accuracy for traits evaluated in the present study was similar to that for yield and disease resistance traits evaluated in previous studies.

Previous studies did not observe significant differences among the modeling methods applied. For instance, in a study by Gouy et al. (2013), four methods were used, although their accuracies were similar. Hayes et al. (2021) noted improved prediction accuracy with methods that used both pedigree and genomic information, although the difference was not significant. In contrast, Yadav et al. (2021) applied GBLUP that considered nonadditive effects (dominance effect and additive–additive epistasis) in the model and noted improved prediction accuracy for one of the three traits. Likewise, Islam et al. (2021) observed improved accuracy with nonadditive machine learning-based models. In the present study, RF and SAE showed higher accuracy than
additive GBLUP for SSC.M and SSC.U in leave-one-out cross-validation, indicating the contribution of nonadditive effects on these traits. However, the improvement in accuracy with nonadditive machine learning models was not large, suggesting that prediction using additive GBLUP models does not result in a significant loss of accuracy.

Notably, prediction accuracy varied depending on the cross-validation method. Hayaes et al. (2021) and Yadav (2021) used “forward prediction” rather than cross-validation. In those studies, prediction accuracy was evaluated in a forward manner, that is, in the previous year (Hayes et al. 2021) or two years (Yadav et al. 2021) of available breeding data. Such prediction may be more difficult than cross-validation of data from the same year because of the genotype × environment (year) effect (Hayes et al. 2021) and/or genetic shifts in the breeding populations over years (i.e., genetic differentiation among populations of different years). In the present study, we did not use forward prediction; however, since prediction accuracy in leave-one-out cross-validation was significantly reduced in our study, genetic shift between years may reduce the accuracy of forward prediction.

**Pedigree information used in machine learning models**

RFs have been frequently used for genomic prediction and have been shown to be as accurate as models based on quantitative genetic methods (Ogutu et al. 2011; Heslot et al. 2012; Onogi et al. 2015; Spindel et al. 2015; Blondel et al. 2015). However, it is difficult to use information that specifies the covariance structure between phenotypes, such as a numerator relationship matrix, simultaneously with genomic information. In the present study, we showed that a numerator relationship matrix can be used to model RFs using the corresponding column (i.e., column corresponding to the target genotype) of the matrix as the input vector. To the best of our knowledge, no such method has been proposed for general application other than RF
models for genomic prediction. Furthermore, a genomic relationship matrix could be constructed in the same way using the corresponding column as the input vector. The accuracy of the model was comparable to that of a RF model that directly uses genomic data. When the number of markers is massive, it may be more effective to use the corresponding column of the genomic relationship matrix as the input rather than directly using genomic data. Both numerator and genomic relationship matrices can be represented as multidimensional data through spectral decomposition. However, prediction accuracy decreases following spectral decomposition. Meanwhile, our method of using the corresponding column of the numerator relationship matrix as the input vector achieved excellent accuracy with SAE, suggesting that this method may be applied to various common machine learning models.

RFs have been shown to be effective in the genomic prediction of disease resistance in sugarcane (Aono et al. 2020; Islam et al. 2021). In addition to genomic information, pedigree information can be used in such models. In addition, Rutkoski et al. (Rutkoski et al. 2012) demonstrated the effectiveness of RFs in which phenotypes of secondary traits were used as the input in addition to genomic data. According to the authors (Rutkoski et al. 2012), the usefulness of RFs lies in their flexible modeling capability. With the feasibility of using pedigree information for modeling, the possibilities of using RF may expand.

**Efficacy of SAE over other methods**

In the present study, we propose SAE—a method for variable selection that explicitly incorporates interactions. SAE showed comparable or even superior prediction accuracy to RF in some cases depending on traits. Thus, SAE may be an effective method for modeling nonlinear relationships among markers. In addition, similar to RF, SAE may be effective for modeling using pedigree information.
Effect of soil type on the accuracy of genomic prediction

In the present study, cultivation trials were conducted in two soil types. ANOVA with soil type and genotype as the main effects showed that the effects of both soil type and the interaction between genotype and soil type were significant, indicating the importance of selecting genotypes that can adapt to a given soil type. To evaluate the accuracy of genomic selection, we built prediction models with average genotype values estimated over the soil types. Comparison of these models showed that the former (trained based on average genotype values) achieved similar or lower prediction accuracy than the latter (trained based on genotype values estimated for each soil type). Although the number of measured plants was almost double for the former model, the prediction accuracy did not differ between the two models. This may be due to the interaction between soil type and genotype. When selecting lines for each soil separately, it may be better to build a prediction model for a given soil type. Alternatively, it may be most desirable (i.e., may provide higher accuracy) to consider genotypic values in different soil conditions as multiple traits with among-trait genetic correlations (Falconer 1952; Meyer 2009) and to apply genomic prediction models considering genetic correlations (Burgueño et al. 2012; Guo et al. 2013). In the present study, we evaluated the accuracy of genomic prediction using only cultivation data from different soil types on the Tanegashima Island. The differences in growing environments across the Nansei Islands were larger than those between soil types evaluated in the present study. Models targeting multiple islands are another important research topic for implementing genomic selection in Japanese sugarcane breeding. The development of such models and evaluation of their accuracy will be addressed in future research.
Unresolved issues for future research

The following points remain as issues for future research: First, sugarcane has a polyploid genome but was analyzed using a diploid model in the present study. The potential and problems of “pseudo-diploid modeling,” in which a predictive model is built for a polyploid organism as a diploid, have been discussed in the application of GS for sugarcane (Yadav et al. 2021). Statistical genetic methods that treat polyploid crops as polyploids have been proposed (Endelman et al. 2018; de Bem Oliveira et al. 2019), including the use of a marker system that can identify the dosage of each allele. The accuracy of a model considering polyploid inheritance with a marker system that can identify dosage should be examined in the future. Another problem is the prediction of ratoon traits. In the present study, we evaluated phenotypic values of individuals grown from plant crops. Since the evaluation of ratoon traits is time-consuming, GS for ratoon traits is considered particularly effective. Traits of ratoon crops are expected to present certain genetic correlations with traits of individuals grown from plant crops. Therefore, a predictive model that can leverage genetic correlations should be constructed. Finally, application of a multi-trait genomic prediction model, which is expected to be effective in the case of correlated traits (Calus and Veerkamp 2011; Jia and Jannink 2012), must be explored in the future.

Conclusion

The present study demonstrated the effectiveness of genomic prediction and selection in sugarcane, as suggested by previous studies (Gouy et al. 2013; Deomano et al. 2020; Hayes et al. 2021; Yadav et al. 2021). In addition, the use of pedigree information was shown to be effective for prediction. In previous studies, pedigree information was used to represent relationships among lines as numerator relationship matrices and then as the covariance structure of variable effects in mixture models. In the present study, we showed that using the columns of a numerator
relationship matrix as the input for machine learning models can yield the same or even higher accuracy as BLUP depending on the traits. Although simple, our approach is expected to expand the use of pedigree information in GS for various common machine learning methods. In addition, we proposed a novel machine learning method that can explicitly incorporate the interaction among markers into the model and confirmed its effectiveness. The proposed method showed the highest accuracy for some traits. Our results are expected to contribute to the improvement of sugarcane breeding efficiency in the future.

**Funding**

This research was supported by cooperative research funds from the Toyota Motor Corporation to the University of Tokyo and partly by cooperative research funds from the Toyota Motor Corporation to the NARO Kyushu Okinawa Agricultural Research Center.

**Competing Interests**

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

**Author Contributions**

MI and HI developed the methods, analyzed data, and wrote the manuscript. YT, TH, MH, and MU performed cultivation experiments and measured phenotypic data. TK and MM genotyped the plants to measure their phenotypic data. TK, MM, YT, TH, MU, and HI designed this study. TK, MM, YT, TH, MH, and MU assisted in manuscript writing. All authors have read and approved the final manuscript.
Data Availability

Not applicable.
References


Genotyping Calls in Phenotype Prediction. G3 9:1189–1198


Hayes BJ, Wei X, Joyce P, et al (2021) Accuracy of genomic prediction of complex traits in
sugarcane. Theor Appl Genet 134:1455–1462


Meyer K (2009) Factor-analytic models for genotype x environment type problems and structured covariance matrices. Genet Sel Evol 41:21


Figure 1

Scatter plot of principal component scores of genotypes. (a) The first and second components based on genomic relationship matrix (GPC1 and 2). (b) The first and second components based on pedigree relationship matrix (PPC1 and 2). Seven families (F1–F7) containing more than 10 genotypes are colored.
Figure 2

Correlations among principal components based on the genomic and principal pedigree relationship matrices of 297 genotypes.
Figure 3

Estimates of the heritability of five target traits under three different assumptions of random genetic effects: (I) effects independently follow identical normal distribution; (P) heritability assuming the among-line covariance structure according to matrix A; and (G) heritability assuming the among-line covariance structure according to matrix G. This heritability is called “genomic heritability.”
Figure 4

Prediction accuracy of BLUP based on matrices A and G. (a) Prediction accuracy in 10-fold cross-validation. (b) Prediction accuracy in leave-one-out cross-validation. The prediction accuracy was calculated as Pearson’s product moment correlation coefficient between the estimated and predicted genotypic values.
Figure 5

Accuracy of prediction based on BLUP and machine learning in 10-fold cross-validation. Models G, A, and GA represent models using matrix G, matrix A, and both matrices G and A, respectively. Method names are as follows: BLUP, best linear unbiased predictors; RFM, random forest directly using a matrix (or matrices) as the input; RFP, random forest using principal component scores based on eigenvalue decomposition of a matrix (or matrices) as the input; RFX, random forest using marker scores as the
input; SAE, a novel machine learning method proposed in the present study. The dotted line represents the prediction accuracy of BLUP based on matrix G (GBLUP).

Figure 6

names are as follows: BLUP, best linear unbiased predictors; RFM, random forest directly using a matrix (or matrices) as the input; RFP, random forest using principal component scores based on eigenvalue decomposition of a matrix (or matrices) as the input; RFX, random forest using marker scores as the input; SAE, a novel machine learning method proposed in the present study. The dotted line represents the prediction accuracy of BLUP based on matrix G (GBLUP).

![Graphs showing 10-fold CV and Leave-one-family-out CV predictions for different traits and soils using BLUP and RFM]

**Figure 7**
Prediction accuracy in genotypic values estimated in different soil types and genotypic values averaged over the soil types. (a) Ten-fold cross-validation with BLUP based on matrices $A$, $G$, and $A \times G$ and random forest model (RFM) using matrices $A$ and $G$ as the input. (b) Leave-one-out cross-validation. The modelling methods (BLUP and RFM) are same as in (a).

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementary20221017.docx](#)