genes

MDPI

Article

Genomic Selection for Optimum Index with Dry Biomass Yield, Dry Mass Fraction of Fresh Material, and Plant Height in Biomass Sorghum

Ephrem Habyarimana [1,*], Marco Lopez-Cruz [2] and Faheem S. Baloch [3]

[1] CREA Research Center for Cereals and Industrial Crops, via di Corticella 133, 40128 Bologna, Italy
[2] Crop, Soil, and Microbial Sciences Department, Michigan State University, 1066 Bogue St, East Lansing, MI 42824, USA; lopezcru@msu.edu
[3] Department of Field Crops, Faculty of Agricultural and Natural Sciences, Abant Izzet Baysal University, 14030 Bolu, Turkey; balochfaheem13@gmail.com
* Correspondence: ephrem.habyarimana@crea.gov.it

Received: 2 December 2019; Accepted: 1 January 2020; Published: 5 January 2020

Abstract: Sorghum is one of the world's major crops, expresses traits for resilience to climate change, and can be used for several purposes including food and clean fuels. Multiple-trait genomic prediction and selection models were implemented using genotyping-by-sequencing single nucleotide polymorphism markers and phenotypic data information. We demonstrated for the first time the efficiency genomic selection modelling of index selection including biofuel traits such as aboveground biomass yield, plant height, and dry mass fraction of the fresh material. This work also sheds light, for the first time, on the promising potential of using the information from the populations grown from seed to predict the performance of the populations regrown from the rhizomes—even two winter seasons after the original trial was sown. Genomic selection modelling of the optimum index selection including the three traits of interest (plant height, aboveground dry biomass yield, and dry mass fraction of fresh mass material) was the most promising. Since the plant characteristics evaluated herein are routinely measured in cereal and other plant species of agricultural interest, it can be inferred that the findings can be transferred in other major crops.

Keywords: *Sorghum bicolor*; *Sorghum halepense*; genomic selection; genomic prediction; optimum index; index selection; biomass; yield; plant height; GBS SNP

1. Introduction

Relying on fossil fuels is a major challenge to a world struggling to adapt and mitigate climate change [1]. In this context, biomass sorghum is a cereal crop that can play an important role for sustainable and environment friendly farming, as it is particularly resilient to drought stress [2] and is more energy efficient than most plant species of agricultural interest, including maize and sugarcane [1]. Sorghum biomass can be used to produce several types of green fuels including biogas and bioethanol, with reduced greenhouse emissions, which are less polluting for the environment relative to fossil fuels [3]. Biomass yield is the primary trait in biomass sorghum, as it measures productivity and profitability of the farmer. Sorghum biomass yields are positively correlated with plant height, maturity, and the concentration of the dry mass [1]. Breeding efforts to increase biomass production should therefore mostly focus on these plant characteristics.

Efficient breeding requires also knowledge-based selection strategies that efficiently exploit available phenotypic and genotypic information that existed in the crop of interest. Since the economic value of the final product in a breeding program depends on several traits [4,5], it can be inferred that selecting the parents for the next generation based on several different plant characteristics

can improve genetic gain. Genomic selection (GS) showed good results in the selection of complex quantitative traits like yields [6,7] and it has been successfully implemented in plant breeding and animal husbandry [8,9]. The main features of the GS approach is the use of algorithms that combine phenotypes and high-density marker data to predict genetic merit upon which superior unphenotyped candidates are selected [10,11]. This attribute is particularly interesting as it can reduce the costs associated with evaluating trials, shorten generation interval, and bypass the extensive field progeny testing that are otherwise required to select parental lines to be used in crossing blocks [12]. Several genomic selection methods have been developed and successfully implemented in plant breeding programs and in animal husbandry [13–15]. Despite the GS success stories, studies on the application of genomic models in sorghum are limited compared to other cereal crops, such as maize, wheat, and rice [16]. For instance, a first genomic selection study in grain sorghum was reported by Hunt et al. [17] for prediction of test-cross performance in individual trials. Velazco et al. [12] investigated different genomic models including pedigree information for across-environment prediction of parental breeding values in productivity and adaptability traits [10].

Most of genomic selection algorithms implemented thus far were based on the analysis of single traits, while selection index has had limited use in actual plant breeding programs [18–20]. Nonetheless, available results indicate that selection index for improving a single trait would not outperform direct selection for the trait itself, whereas selecting simultaneously for more than one trait in selection index might outperform selecting for a single trait [20]. With the use of selection indices, individuals with very high merit in some traits are saved for breeding, even when they are slightly inferior in other traits [19,21], which can not only sustain productivity but can also safeguard genetic diversity. The selection index represents a joint analysis of multiple traits and can increase the accuracy of genetic evaluations in comparison with the single-trait analysis as it exploits the information from correlated traits [22].

The genomic selection index (GSI) is a linear combination of genomic estimated breeding values (GEBVs) used to predict the individual net genetic merit upon which individual candidates are selected from a nonphenotyped testing population as parents of the next selection cycle [19]. The efficiency of applying selection index in breeding depends on the strength of genetic and environmental correlations between the characters of interest. According to Thompson and Meyer [23], the benefit of selection index increases for lowly heritable traits, when analyzed together with strongly correlated traits of higher heritability. Another selection index advantage is represented by the possibility to reduce selection bias or culling bias introduced by contemporary or sequential selection on correlated traits, which are ignored by single-trait approaches [24]. The importance of selection index in genomic selection was demonstrated in empirical and simulation studies [3,25,26]. It was shown that genomic selection index models can efficiently be used to integrate information from correlated traits and from relatives. For this purpose, a breeder interested in response to selection for a single target trait, can incorporate other auxiliary traits in the index to provide additional information on the primary trait.

The efficiency of genomic selection index (GSI) models was shown in other cereal crops including maize, rice, and wheat [27–29]. In sorghum, GSI was implemented only in advanced breeding lines of grain sorghum [12] and in biomass-type genotypes using a pre-breeding population [3]. In the later work, the objective was to apply the GSI on auxiliary characters to indirectly predict the genomic estimated breeding value corresponding to the primary trait. In this work, we present, for the first time, a study we conducted on the potential of exploiting selection index for genomic selection in a panel of 380 biomass sorghum genotypes consisting of a mixture *Sorghum bicolor* landraces and lines and *S. bicolor* × *S. halepense* advanced inbred lines. Our objectives were to: (1) investigate if the use of a genomic selection index made up of aboveground dry biomass yield, dry mass fraction of the fresh mass material, and plant height can improve prediction accuracy relative to a single trait genomic selection index, and (2) investigate the efficiency of genomic selection indices in *S. bicolor* × *S. halepense* regrown from the rhizomes (overwintered testing set) using the populations grown from seeds as training set.

2. Materials and Methods

2.1. Phenotypic and Genotypic Data

Plant materials evaluated in this work belonged to a panel of 369 biomass sorghum genotypes of which 180 *Sorghum bicolor* landraces and lines and 189 *S. bicolor* × *S. halepense* advanced recombinant inbred lines beyond the F_7 filial progeny. The two populations were evaluated at the same experimental site. Field trials covered four years (2014–2017 for *S. bicolor* and 2015–2018 for *S. bicolor* × *S. halepense*) for each population and were run side-by-side, except for in 2014 where only a *Sorghum bicolor* trial was planted. For the *S. bicolor* × *S. halepense* trials, the entire population was evaluated each year except in 2015 where only half the population was sown owing to scarce seed availability. Overall, there were four trials for *S. bicolor* population and nine trials for *S. bicolor* × *S. halepense* population. Of the nine trials of the later population, six were plants regrown from overwintered rhizomes, while three were trials grown from seeds. The list and the sizes (number of tested genotypes excluding checks) of the trials evaluated for each trait are presented in Table 1.

Table 1. Trials and respective sizes of populations and the traits evaluated.

Trials	PH	DMC	DMY [1]
IT14	174	123	123
IT15	179	179	179
IT16	180	NA	180
IT17	168	168	168
US15_DS	90	90	90
US15_RG16	89	89	89
US15_RG17	85	85	85
US15_RG18	85	85	85
US16_DS	189	189	189
US16_RG17	189	189	189
US16_RG18	189	189	189
US17_DS	189	189	189
US17_RG18	189	189	189

[1] IT and US, respectively, denote *S. bicolor* and *S. bicolor* × *S. halepense* trials. DS, RG, PH, DMC, DMY, respectively, denote trials grown from seeds (direct sowing trials), trials regrown from overwintered rhizomes (regrowth trials), plant height, dry mass fraction of the fresh material, and aboveground dry biomass yield. Number following IT and US are the years of direct sowing trials, while the numbers following RG are the years of the regrowth trials. "NA" indicates that the data was not scored.

All the experiments were open-field trials and were established at CREA Research Center for Cereal and Industrial Crops, in the experimental station of Cà Rossa in Anzola (Bologna, Italy). The augmented randomized complete block design was used with six controls (checks) and six blocks [30] except US15 trials which had four checks and 4 blocks. Elementary plots were single 5 m long rows distant 0.75 m, and were thinned to homogeneously distributed 50 plants per plot. We evaluated open field morpho-physiological data on aboveground dry biomass yields (DMY, t ha^{-1}), plant height (PH, cm), and dry mass fraction of the fresh material (DMC, %), as suggested by IBPGR [31]. Plant height was measured one week before harvest as the mean height of the elementary plot using a 5 m telescopic rod (Stanley 5 m grade rod aluminium) placed vertically on the ground in the middle of the row. Aboveground dry biomass yield and the dry mass fraction of the fresh material were measured as follows. The entire plot was machine chopped and fresh weight plot yield scored. Immediately after a plot was weighed, a sample was taken from each plot then weighed before and after oven drying at 80 °C to constant weight to determine moisture content. Dry mass fraction of the fresh material (DMC%) = (sample dry weight/sample wet weight) × 100. Aboveground dry biomass yield in metric tons per hectare (DMY t ha^{-1}) = ((total plot wet weight (kg) × (sample dry weight/sample wet weight))/ plot area (m^2)) × 10.

2.2. Phenotypic Data Analysis

Data from single trials were analyzed in two steps. In the first step, the adjusted means were calculated as suggested by Federer [30] to account for the variability of soil properties. In the second step, adjusted means from each trial were jointly analyzed to estimate genotype means across environments. The model fitted was as follows: $y_{ik} = \mu + G_i + E_k + GE_{ik} + \varepsilon_{ik}$, where y_{ik} is the best linear unbiased estimation (BLUE) of i-th genotype in the k-th environment, which was fitted by a random genotype effect (G_i), a fixed environmental effect (E_k), and the genotype × environment interaction (GE_{ik}). Given that genotype effects were considered random, the GE interaction involving G_i was random. All random effects were assumed independent homoscedastic and normally distributed with zero mean. The best linear unbiased estimates were used in the subsequent processes of fitting the genomic selection models.

2.3. Molecular Data

DNA extraction and whole-genome genotyping procedures were amply described in Habyarimana and Lopez-Cruz [10]. The molecular information used in this work consisted of genotyping-by-sequencing single nucleotide polymorphisms (GBS SNPs) produced by BGI Hong Kong Company Limited. To prepare the library, the ApeKI, a methylation-sensitive restriction enzyme, was used, and GBS was carried out on an Illumina HiSeq X Ten platform. For variants discovery, the sequencing reads were aligned to the sorghum reference genome (*Sorghum_bicolor* NCBIv3). The SNP datasets were filtered using VCF tools to extract marker data responding to high quality standards such as biallelic SNPs only, minor allele frequency (MAF) ≥ 0.05, site quality or the Phred-scaled probability that reference/alternative alleles polymorphism exists at a given site given the sequencing data Q ≥ 40 (i.e., base call accuracy ≥ 99.99%), and missing genotypes (NA) ≤ 20%. The final size of the high quality-controlled marker dataset matrix was 61,976 SNPs which were used in downstream steps in this work for genomic prediction and selection analytics.

2.4. Construction of Genomic Selection Indices

In matrix notation, an optimum phenotypic selection index (PSI) [32] takes the following form [20] $I_i = \sum_{j=1}^{p} \beta_j x_{ij} = \beta' x_i$ where $\beta' = \left[\beta_1 \beta_2 \ldots . \beta_p\right]$ is a vector of coefficients, p is the number of traits on I_i, and $x_i = \left[x_{i1}, \ldots, x_{ip}\right]$ is a vector of p measured phenotypic values which are centered with respect to their respective means. The linear genomic selection index for individual i is represented by the aggregate genotypes H and was defined as $H_i = \sum_{j=1}^{t} \alpha_j g_{yij} = \beta' g_{yi}$ where $\acute{g}_{yi} = \left[g_{yi1} g_{yi2} \ldots . g_{yit}\right]$ is a vector of the genotypic values of t selection targets y_i and $\alpha' = [\alpha_1 \alpha_2 \ldots ., \alpha_t]$ a vector of known and fixed economic weights [19]. Under the breeding perspective, economic values are used to reflect the relative importance of the traits of interest. The economic value is the increase in profit achieved by improving a particular trait by one unit [33,34]. In case of several traits, the total economic value is a linear combination of the breeding values of the traits weighted by their respective economic values as in the above equation [19,32], and this is called the net genetic merit (or aggregate genotype, selection target) of one individual.

To be used in the optimum indices, the β_j are derived such that I_i is maximally correlated with H_i, the solution of which is found to be the following matrix equation [20,35] $\hat{\beta} = P_x^{-1} G_{x,y} \alpha$. The matrices $G_{x,y}$ and P_x are, respectively, the genotypic covariance between the measured phenotypes and the selection targets, and the phenotypic variance-covariance among the measured phenotypes. On the other hand, $\hat{\beta}$ is the best linear unbiased predictor (BLUP) of β_j, while α is as described above [32,36,37]. From the above equations, the following statistics were derived as suggested in [18,19]: (1) heritability of the index $h_I^2 = \beta' G_x \beta / \beta' P_x \beta$, where G_x is the genotypic variance-covariance matrix among the measured phenotypes, (2) genetic correlation between the index and the selection target

$gencor = cor(g_I, g_H) = \beta' G_{x,y}\alpha / \sqrt{\beta' G_y \alpha}\sqrt{\beta' G_x \beta}$, where G_y is the genotypic variance-covariance matrix among the selection targets, and (3) accuracy of selection defined as the correlation between the index and the genotypic value of the selection target i.e., $acc = cor(I, g_H) = cor(g_I, g_H)h_I$. The accuracy of selection was used to evaluate the performance of the genomic prediction model performance.

2.5. Genomic Selection Models

In the genomic selection index modeling, phenotypic and marker data are scored in the training population and fitted into appropriate algorithm to produce individuals' whole-genome marker effects. The marker effects are used in subsequent cycles of selection to compute the genomic estimated breeding values (GEBVs) that are used as predictors of breeding values in a testing unphenotyped population. The genomic estimated breeding values are obtained as a product of the estimated marker effects in the training population and the coded marker values obtained in the testing population. To apply genomic selection index, GEBVs are obtained in the selection candidates and then used to predict and rank the net genetic merit of the candidates for selection.

In this work, the genomic selection analyses were implemented in the multiple-trait model (MTM) software [38] that uses a Bayesian approach [39]. The routines built in the MTM package allow the calculation of the phenotypic and genotypic variance-covariance matrices. The performance of the genomic selection models was assessed using Monte Carlo (repeated hold-out) cross-validation approach [40,41] applying 70% and 30%, respectively, as training and validation (test) sets. In a standard hold-out cross-validation, the test set represents new, unseen data to the model. To obtain a more robust performance estimate that was less dependent on how the data was split into training and validation sets, the holdout method was repeated 100 times using different random seeds. The hundred repetitions were then used to calculate the average prediction performance. In comparison to the standard holdout validation method, the repeated hold-out procedure implemented in this work provides a better estimate of the model prediction ability when a random test set is used [41]. The repeated hold-out procedure provides also the information about the stability of the model (produced by a learning algorithm) across training set splits. The parameters of the models were estimated in the training set before the models were validated in the testing set. The performance of the models was measured using the accuracy of selection and the genetic correlation between the index and the selection target as described previously [18,19].

The selection index algorithms were implemented for different targets of prediction considering $H_i = g_{yij}$ for each single trait in the target set, and then $H_i = \sum_{j=1}^{t} \alpha_j g_{yij}$ for multi-trait genomic selection index, with $\alpha' = [1, \ldots, 1]$ representing the economic weights of the t traits for which we expressed equal preference [32,35]. In the box below (Figure 1) is the example of a code snippet used in this work to instruct the creation of a training and testing sets in R:

The models were implemented using R software, version 3.5.3 (R Core Team, Vienna, Austria) [42] and the package MTM [1,38] by applying default rules for selecting hyperparameters. The Gibbs sampler was used and our analyses were based on 30,000 samples from the posterior distribution obtained after the first 5000 iterations were discarded as burn-in [1]. The visualization algorithms and statistical inferences used to present the genomic selection models' output were implemented using routines called from the R software. The magnitude and direction of the Pearson correlation coefficients were interpreted according to Gomez and Gomez [43] as follows: 0–0.1, 0.1–0.5, 0.5–0.8, and 0.8–1, 1, respectively, zero, low, medium, high, and perfect.

```
SETS <-list(
      list(
      trn = expand.grid(c("IT"),c(14),c(NA)),
      tst = expand.grid(c("IT"),c(15),c(NA))
      ),
      list(
      trn =
expand.grid(c("IT"),c(14,15,16),c(NA)),
      tst = expand.grid(c("IT"),c(17),c(NA))
      ),
      list(
      trn =
expand.grid(c("US"),c(15,16),c("DS")),
      tst =
expand.grid(c("US"),c(17),c("DS"))
      )
)
```

Figure 1. Example of a code snippet instructing the creation of training and testing sets. The list called 'SETS' contains three different training-testing (TRN-TST) sets. The IT14, IT15, IT16, and IT17 are four S. bicolor trials containing same genotypes evaluated in 2014, 2015, 2016, and 2017, respectively. The first set will train the model using as TRN set IT14 to predict IT15. In the second scenario, the model will be trained in IT14 + IT15 + IT16 to predict IT17, while in the third scenario, the model will be trained in US15DS + US16DS to predict US17DS.

3. Results

3.1. Comparison of Traits, Genetic Metrics, and Genomic Selection Approaches

The Pearson correlation was low and negative ($r = -0.35$) between the dry mass fraction of the fresh material and the plant height, low and positive ($r = 0.23$) between dry mass fraction of the fresh material and the aboveground dry biomass yield, and medium and positive ($r = 0.60$) between plant height and the aboveground dry biomass yield (Figure 2). On the other hand, the Pearson correlation was higher ($r = 0.94$) between accuracy and genetic correlation, followed by the correlation between accuracy and heritability ($r = 0.87$), and between heritability and genetic correlation ($r = 0.84$) (Figure 3). The heritability of all single traits and the genomic selection index came from same distribution with statistically comparable means ($h^2 = 0.59$–0.71) (Figure 4). Genetic correlation was higher (gencor = 0.6–0.63) and comparable in genomic selection index, aboveground dry biomass yield and plant height, while it was lower (gencor = 0.46) for the dry mass fraction of the fresh weight (Figure 5). The accuracy showed the same pattern as the genetic correlation. The accuracy was higher (*acc* = 0.52–0.59) and comparable in genomic selection index, aboveground dry biomass yield and plant height, while it was lower (*acc* = 0.36) for the dry mass fraction of the fresh weight (Figure 6).

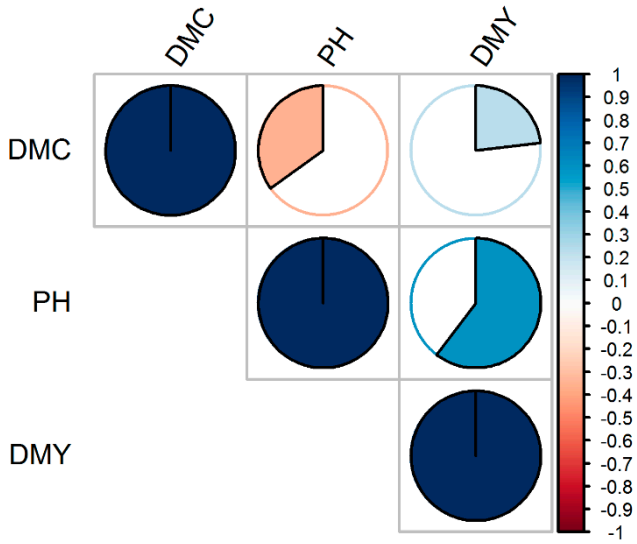

Figure 2. Pearson correlation coefficients among the evaluated traits. DMC, DMY, and PH, respectively, denote dry mass content, dry biomass yield, and plant height. The filled-in areas of the circles show the absolute value of corresponding correlation coefficients. The scale on the right hand side is colored from red (negative correlation) to blue (positive correlation); with the intensity of color scaled 0%–100% in proportion to the magnitude of the correlation. Refer to text for the description of the traits.

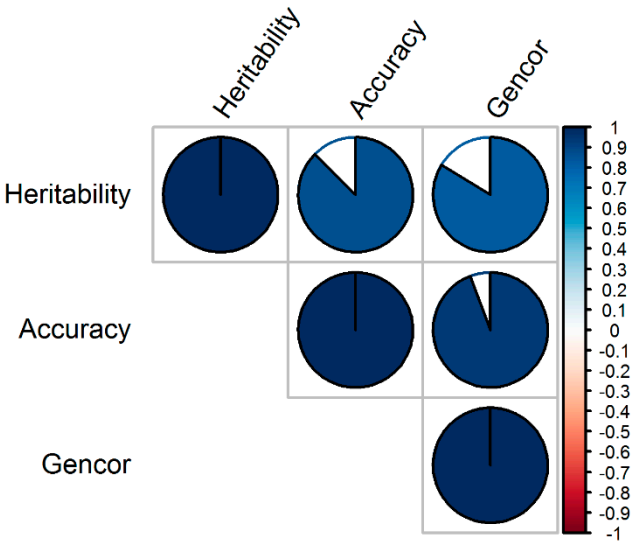

Figure 3. Pearson correlation coefficients among the genetic metrics. Gencor genetic correlation between the phenotypic and the genomic selection indices. The filled-in areas of the circles show the absolute value of corresponding correlation coefficients. The scale on the right-hand side is colored from red (negative correlation) to blue (positive correlation); with the intensity of color scaled 0%–100% in proportion to the magnitude of the correlation. Refer to text for the description of the genetic metrics.

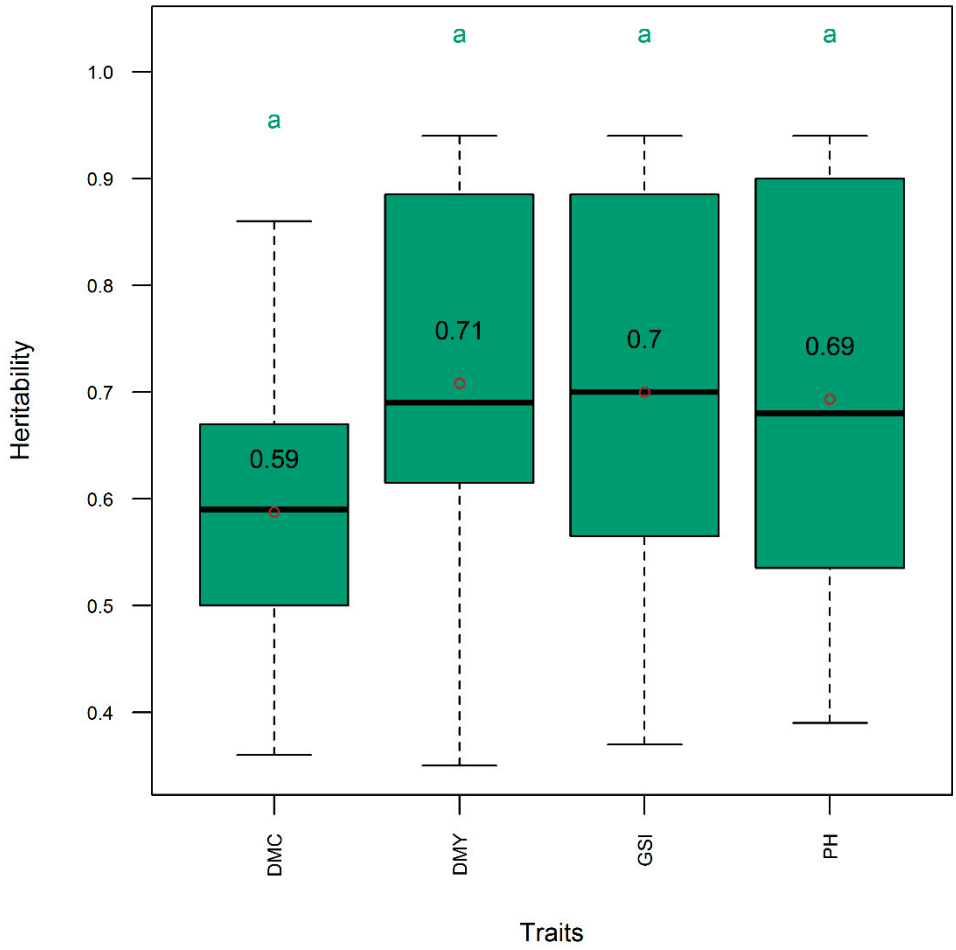

Figure 4. Distribution (boxplot) of narrow-sense heritability for single trait and three-trait selection indices in the entire panel. DMC, DMY, GSI, and PH, respectively, denote selection indices relative to dry mass fraction of fresh material, aboveground dry biomass yield, three-traits (DMC, DMY, and PH), and plant height. Means are indicated by open dots and are included within the boxplot. Means with same letter are not significantly different at the 5% level using the Tukey's HSD (honestly significant difference) test. Refer to text for the description of the GS models.

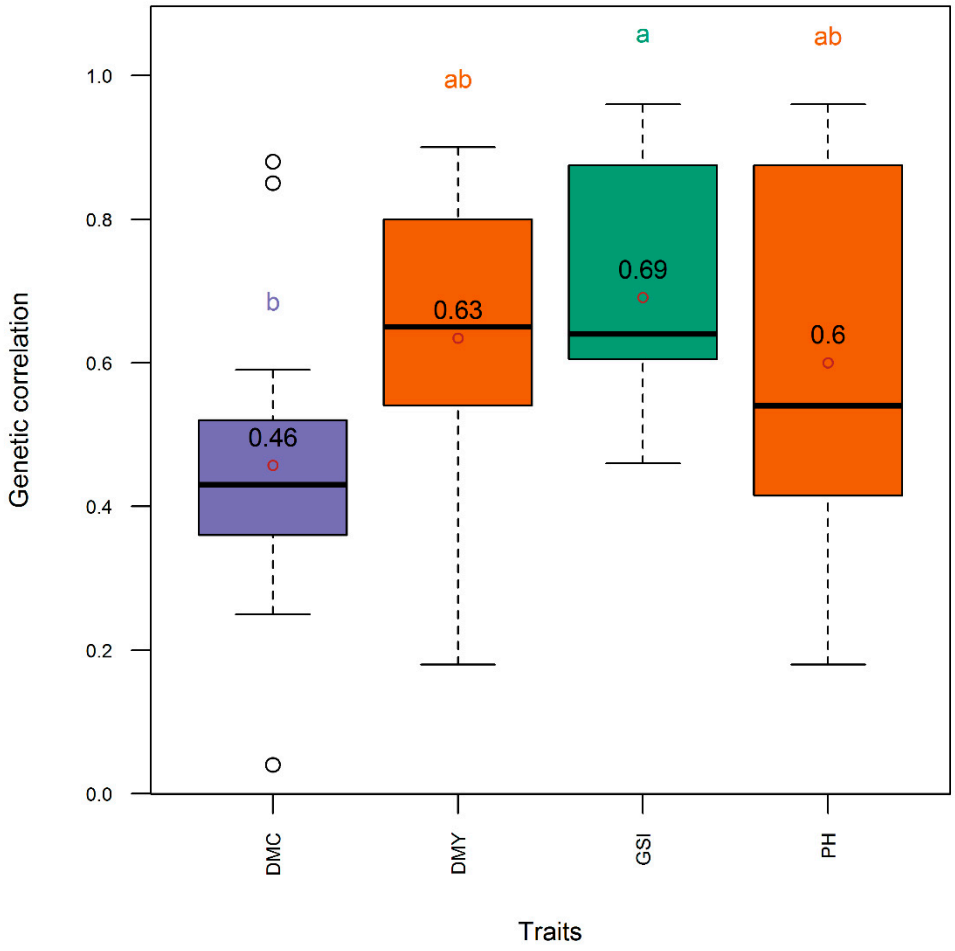

Figure 5. Distribution (boxplot) of genetic correlation between the phenotypic indices and the net genetic merit in the entire panel. DMC, DMY, GSI, and PH, respectively, denote selection indices relative to dry mass fraction of fresh material, aboveground dry biomass yield, three-trait (DMC, DMY, PH), and plant height. Means are indicated by open dots and are included within the boxplot. Means with same letter are not significantly different at the 5% level using the Tukey's HSD (honestly significant difference) test. Refer to text for the description of the GS models.

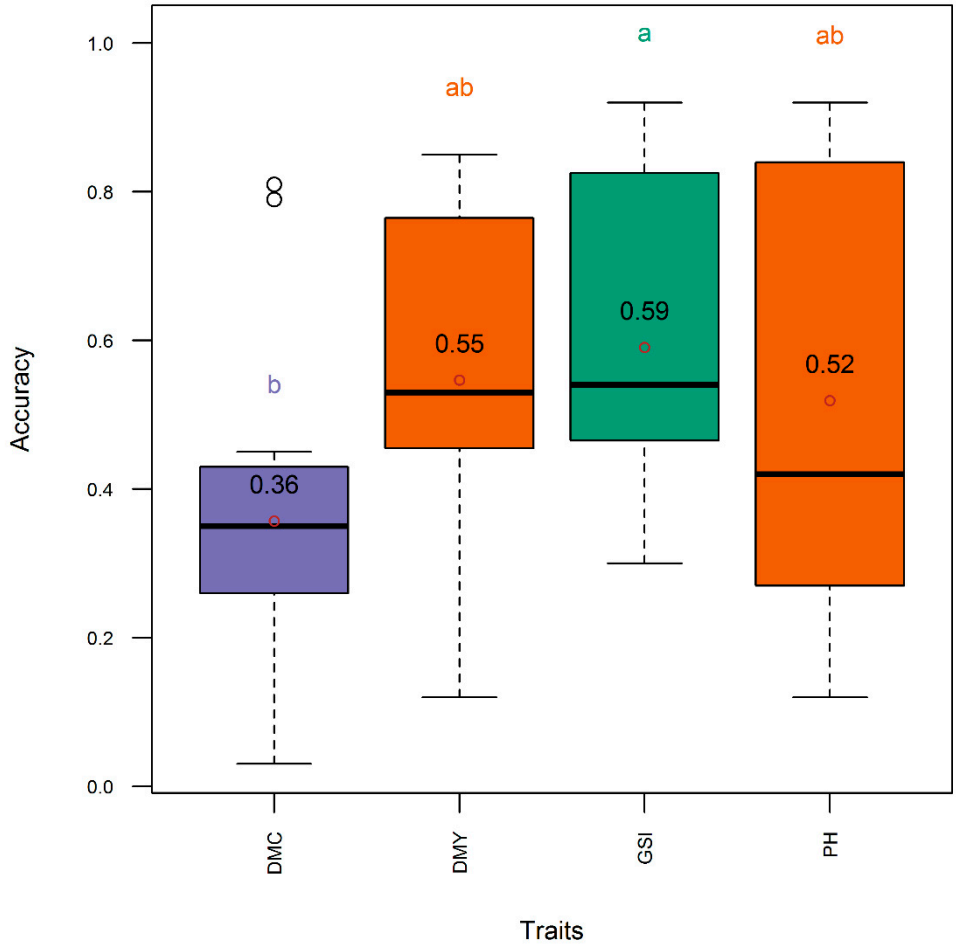

Figure 6. Distribution (boxplot) of genomic selection index accuracy using single traits and all three traits of interest simultaneously in the entire panel. DMC, DMY, GSI, and PH, respectively, denote selection indices relative to dry mass fraction of fresh material, aboveground dry biomass yield, all the three traits simultaneously, and plant height. Means are indicated by open dots and are included within the boxplot. Means with same letter are not significantly different at the 5% level using the Tukey's HSD (honestly significant difference) test. Refer to text for the description of the GS models.

The heritability and the genetic correlation were higher in *Sorghum bicolor* than in *S. bicolor* × *S. halepense*. Genetic correlation in *Sorghum bicolor* vs. *S. bicolor* × *S. halepense* was 0.86 vs. 0.40, 0.82 vs. 0.54, 0.92 vs. 0.43, and 0.91 vs. 0.57, respectively for the dry mass fraction of the fresh material, aboveground biomass yield, plant height, and the three-trait genomic selection index. Heritability in *Sorghum bicolor* vs. *S. bicolor* × *S. halepense* was 0.86 vs. 0.55, 0.90 vs. 0.60, 0.92 vs. 0.57, and 0.91 vs. 0.59, respectively for the dry mass fraction of the fresh material, aboveground biomass yield, plant height, and the three-trait genomic selection index. Accuracy in the *Sorghum bicolor* subpopulation was higher than in the *S. bicolor* × *S. halepense* subpopulation for all traits and the genomic selection index (Figure 7). In *S. bicolor*, accuracy was comparable (*acc* = 0.78–0.88) among single traits and the genomic selection index, while in in *S. bicolor* × *S. halepense*, the pattern followed that of the entire diversity panel with higher and comparable accuracy (*acc* = 0.33–0.44) in genomic selection index, aboveground dry biomass yield and plant height, while the accuracy was lower (*acc* = 0.30) for the dry mass fraction of the fresh weight (Figure 7).

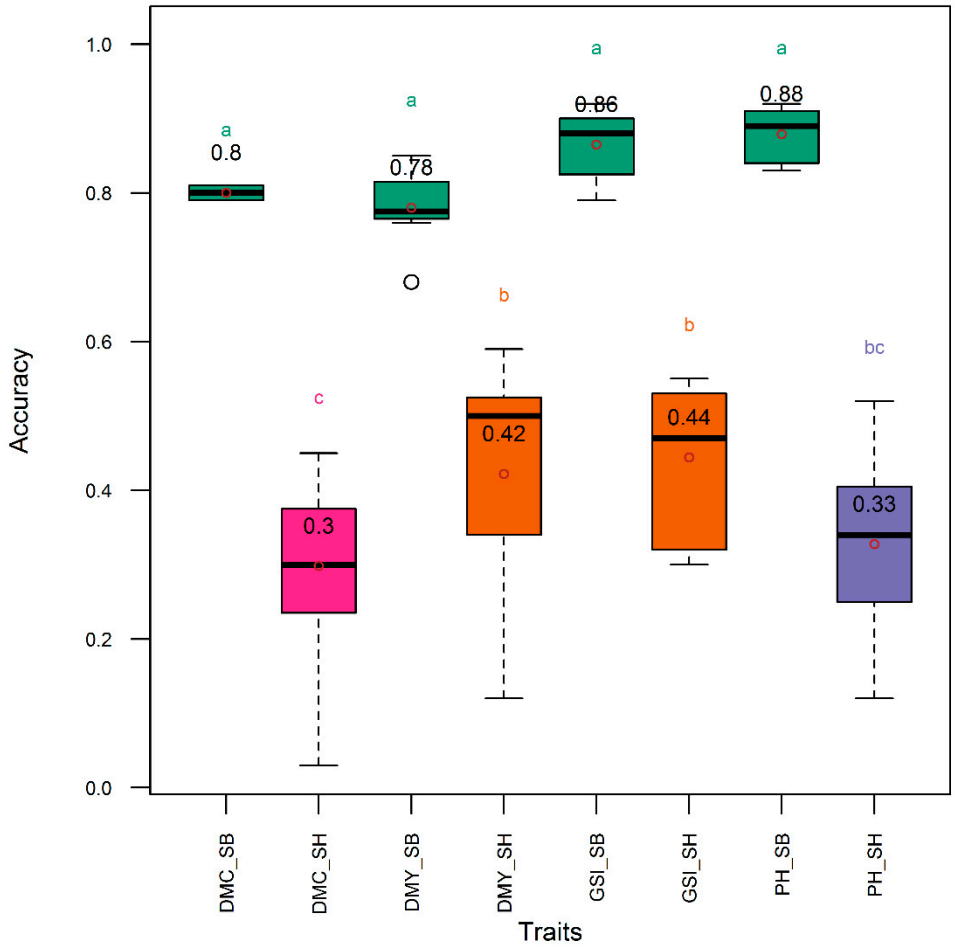

Figure 7. Distribution (boxplot) of genomic selection accuracy using single traits and all traits simultaneously in *Sorghum bicolor* and *S. bicolor* × *S. halepense* lines. DMC, DMY, GSI, and PH, respectively, denote selection indices relative to dry mass fraction of fresh mass material, aboveground dry biomass yield, all the three traits simultaneously, and plant height. Traits suffixed with "_SB" and "_SH", respectively, were collected from *Sorghum bicolor* and *S. bicolor* × *S. halepense* lines. Means are indicated by open dots and are included within the boxplot. Means with same letter are not significantly different at the 5% level using the Tukey's HSD (honestly significant difference) test. Refer to text for the description of the GS models.

3.2. Predicting Regrowth Performance in Perennial Sorghum Bicolor × Sorghum Halepense

The information from the *Sorghum bicolor* × *Sorghum halepense* trial sown in 2016 was used to predict the performance of the overwintered (regrowth) populations in 2017 and 2018 (Figure 8). For plant height, genetic correlation and accuracy were 0.58 and 0.47, respectively, in 2017 and decreased by 48% and 47%, respectively, in 2018. For the dry mass fraction of the fresh mass material, genetic correlation and accuracy were 0.43 and 0.35, respectively, in 2017 and decreased by 37% and 40%, respectively, in 2018. For the aboveground dry biomass yield, the genetic correlation and accuracy remained stable from 2017 to 2018 with respective ranges of 0.53–0.55 and 0.45–0.46. The heritability of the above three traits remained stable from 2017 to 2018 decreasing or increasing by one to five hundredths. The genetic correlation and accuracy obtained with the genomic selection index were higher than the best values obtained with a single trait. On the other hand, the heritability obtained with the genomic selection index was comparable to that obtained with the aboveground dry biomass and higher than the heritability realized in other traits.

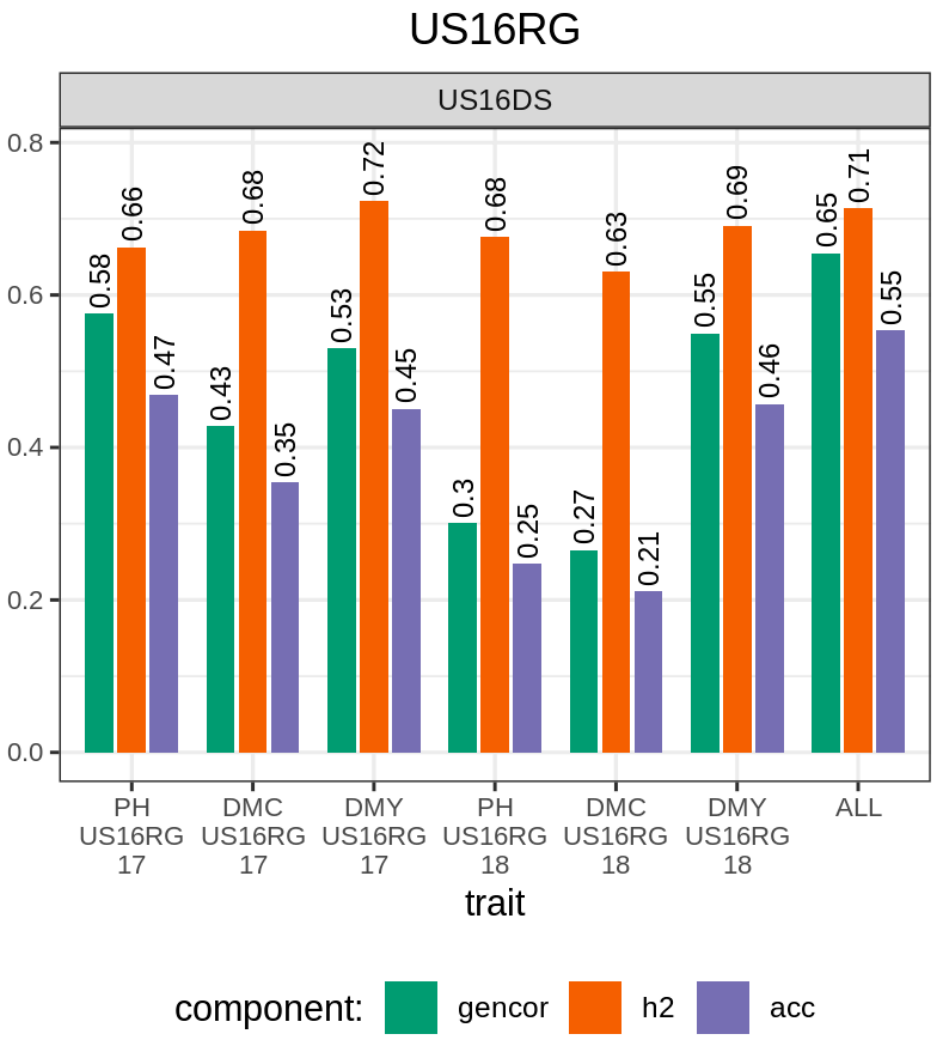

Figure 8. Accuracy of genomic prediction of the performance of *S. bicolor* × *S. halepense* regrown from rhizomes (testing set) using trials grown from seeds as training set. US16DS *S. bicolor* × *S. halepense* sown and evaluated in 2016; US16RG17 *S. bicolor* × *S. halepense* sown in 2016 and regrown and evaluated in 2017; US16RG18 *S. bicolor* × *S. halepense* sown in 2016 and regrown and evaluated in 2018. DMC, DMY, ALL, and PH, respectively, denote selection indices relative to dry mass fraction of the fresh material, aboveground dry biomass yield, all the three traits simultaneously, and plant height. Numbers on the top of the bars are mean accuracies. Refer to text for the description of the GS models.

4. Discussion

A diversity panel made up of a mixture of *Sorghum bicolor* lines and landraces and *Sorghum bicolor* × *Sorghum halepense* advanced recombinant inbred lines was used in this work in order to set up the groundwork upon which to build future germplasm improvement and cultivar development programs. A similar panel was used previously in a genome-wide linkage disequilibrium investigation in sorghum, and in genomic prediction and selection for antioxidant production in sorghum [10,44]. In these previous studies, mixing *Sorghum bicolor* and *Sorghum bicolor* × *Sorghum halepense* genotypes was motivated mainly by the observed weak structure of the resulting diversity panel. In addition, in these investigations and in the present work, *Sorghum bicolor* relevant information was used as the molecular marker information used was derived by aligning the sequencing reads to the sorghum reference genome (Sorghum_bicolor NCBIv3) to enable variants discovery. It was also shown that the use of *Sorghum bicolor* × *Sorghum halepense* recombinant inbred lines in the diversity panel brought novel useful polymorphism [44].

The correlation observed among the evaluated traits was not in full agreement with Habyarimana et al. [1] except for the relationship between plant height and the aboveground dry biomass yield. The differences between the two works can be attributed to different types of populations evaluated. In this work a panel of *Sorghum bicolor* and *S. bicolor* × *S. halepense* was evaluated, while the correlation reported in Habyarimana et al. [1] referred only to *S. bicolor* × *S. halepense*. The high pairwise correlation between plant height and the aboveground dry biomass yield implied that the proportion of variance shared by these traits was mostly explained by genetic causes. A perfect correlation between plant characteristics implies that genetic effects on the traits of interest are identical, which can indicate the existence of either linkage disequilibrium, pleiotropy or causal overlap, or ascertainment bias deriving from biased sampling [10].

The lower correlation coefficients observed in this work between the dry mass fraction of the fresh material and the plant height, on the one hand, and the aboveground dry biomass yield on the other hand implies that dry mass fraction of the fresh material can be improved independently of plant height and aboveground dry biomass yield. This can have important implications in terms of sustainability because high-yielding genotypes can be bred that contain less moisture in biomass at harvest, which means less energy would be spent on biomass conversion and transportation from the field to the bioreactor.

When faced with the necessity to simultaneously improve more than one trait, a breeder can use three approaches: tandem selection, independent culling levels, and index selection [45]. In tandem selection, only one character is selected in each cycle; in independent culling levels, all genotypes with a phenotypic value below the culling threshold for at least one characteristic are discarded; the selection index aims at improving several traits simultaneously in such a way as to make the biggest possible improvement in overall genetic merit [35]. In this work, we implemented the Optimum selection Index of Smith [32] the performance of which was demonstrated in previous studies [35,37]. In our optimum index selection, both desirable and undesirable (e.g., plant height vs. dry mass fraction of the fresh material) correlations were observed between traits (Figure 2) but, as Bradshaw [35] put it, these were accommodated by the index accounting for the simultaneous improvement of the traits on the index. In the process of computing the optimum index selection, equal weights in terms of phenotypic standard deviations ($1/\sigma_P$) were used as suggested by Bradshaw [35] and supported by Saeidnia et al. [46]. The later authors used optimum index and compared several economic weights including unit, phenotypic correlation, genotypic correlation, heritability, direct effects in path analysis and first factor loading in factor analysis. They found out that using unit coefficient in the optimum selection index allowed the highest genetic advance for all traits making up the index. In the same work the selection index with equal weights showed high correlation with the net genetic merit.

The accuracy was more associated with genetic correlation than heritability because heritability was generally high and did not show high variability across trials. This relationship among heritability, genetic correlation and accuracy of selection was consistently observed both in Pearson correlation analysis (Figure 3) and in post hoc analytics through mean separation (Figures 2–4). From the high heritability values of the index selection it can be inferred that the indices described in this work can be effectively used in breeding programs without significant environmental noise. The genetic correlation and accuracy were statistically comparable between the three-trait index selection, aboveground dry biomass yield and plant height, but these metrics were significantly lower for the dry mass fraction of the fresh material (Figures 3 and 4). It can therefore be inferred that the use of the three traits in the index selection can simultaneously improve the accuracy for selecting aboveground dry biomass yield, plant height, and particularly, the dry mass fraction of fresh material. Indeed, this is the inherent characteristics of a linear selection index as it is expected to allow extra merit in one trait to offset defects that existed in another. As Hazel and Lush [19,21] showed, by the use of a linear selection index, individuals with very high merit in one trait are saved for breeding, even when they are inferior in other traits.

The higher accuracy of selection observed in *Sorghum bicolor* relative to *S. bicolor* × *S. halepense* can be explained by the lower genetic variability in the *S. bicolor* × *S. halepense* materials as confirmed by their observed lower heritability of the index and lower genetic correlation between the index and the net genetic merit. The low genetic variability in *S. bicolor* × *S. halepense* lines might have resulted from the low number of parents used during early hybridizations [47] that led to a relatively narrow genetic base in the current progeny. On the other hand, higher genetic variability in *S. bicolor* was expected as these genotypes were derived from African and Asian landraces, and are expected to harbor a high level of genetic diversity for breeding purposes inasmuch as Africa and Asia represent, respectively, the primary and secondary sorghum centers of diversity [2].

The results from the regrowth trials were encouraging. Heritability was consistently higher across years for all selection indices, implying that effective selection can be carried out even several overwintering generations after the original seed sown trials. Among single trait genomic selection indices, the aboveground dry biomass yield showed better accuracy relative to other traits, and maintained the good accuracy across years. The accuracy for the dry mass fraction of the fresh material and the accuracy for plant height decreased over years. For these traits, the accuracy in regrowth trials can probably be improved by either re-training the models including the information from the immediately precedent generation or integrating the single traits of interest in a multi-trait index selection. The observed higher accuracy for the three-trait genomic selection holds therefore good promise for improving aboveground dry biomass yields and its auxiliary traits like plant height and the dry mass fraction of the fresh material in *S. bicolor* × *S. halepense*.

5. Conclusions

In this work, extensive experimental breeding data were used to demonstrate for the first time that the optimum index selection can be implemented in genomic selection predictive analytics for index selection including aboveground dry biomass yield, plant height, and dry mass fraction of the fresh material in biomass sorghum crop. Furthermore, this work shed light for the first time on the promising potential of using the information from the trial grown from seed to predict the performance of the populations regrown from the rhizomes even two winter seasons after the original trial was sown. For these particular populations established from regrowths, using multi-trait index selection was the recommended option to improve traits such as plant height and the dry mass fraction of the fresh material that were weakly predicted when the selection target was regrown from the rhizomes. Since the plant characteristics evaluated herein are routinely measured in cereal and other plant species of agricultural interest, it can be inferred that our findings can be harnessed in other major crops as well.

Author Contributions: Conceptualization, E.H.; methodology, E.H. and M.L.-C.; software, E.H. and M.L.-C.; formal analysis, E.H.; investigation, E.H., M.L.-C., and F.S.B.; data curation, E.H.; writing—original draft preparation, E.H.; writing—review and editing, E.H., M.L.-C., and F.S.B.; visualization, E.H.; supervision, E.H.; project administration, E.H.; funding acquisition, E.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Union, grant number 732064 (H2020-ICT-2016-1-innovation action) and the APC was funded by the European Union through the project Data-driven Bioeconomy (www.databio.eu) and the Ministero delle Politiche Agricole, Alimentari, Forestali e del Turismo (Rome, Italy) through the project Risorse GeneticheVegetali (RGV/FAO) 2014–2016.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Habyarimana, E.; Lorenzoni, C.; Marudelli, M.; Redaelli, R.; Amaducci, S. A meta-analysis of bioenergy conversion relevant traits in sorghum landraces, lines and hybrids in the Mediterranean region. *Ind. Crops Prod.* **2016**, *81*, 100–109. [CrossRef]

2. Habyarimana, E.; Lorenzoni, C.; Redaelli, R.; Alfieri, M.; Amaducci, S.; Cox, S. Towards a perennial biomass sorghum crop: A comparative investigation of biomass yields and overwintering of *Sorghum bicolor* × *S. halepense* lines relative to long term *S. bicolor* trials in northern Italy. *Biomass Bioenergy* **2018**, *111*, 187–195. [CrossRef]
3. Fernandes, S.B.; Dias, K.O.G.; Ferreira, D.F.; Brown, P.J. Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum. *Theor. Appl. Genet.* **2018**, *131*, 747–755. [CrossRef] [PubMed]
4. Bernardo, R. *Breeding for Quantitative Traits in Plants*; Stemma Pr: Woodbury, MN, USA, 2002; ISBN 978-0-9720724-0-3.
5. Lynch, M.; Walsh, B. *Genetics and Analysis of Quantitative Traits*, 1st ed.; Sinauer Associates is an Imprint of Oxford University Press: Sunderland, MA, USA, 1998; ISBN 978-0-87893-481-2.
6. Habyarimana, E. Genomic prediction for yield improvement and safeguarding genetic diversity in CIMMYT spring wheat (*Triticum aestivum* L.). *Aust. J. Crop Sci.* **2016**, *10*, 127–136.
7. Habyarimana, E.; Parisi, B.; Mandolino, G. Genomic prediction for yields, processing and nutritional quality traits in cultivated potato (*Solanum tuberosum* L.). *Plant Breed.* **2017**, *136*, 245–252. [CrossRef]
8. Meuwissen, T.; Hayes, B.; Goddard, M. Genomic selection: A paradigm shift in animal breeding. *Anim. Front.* **2016**, *6*, 6–14. [CrossRef]
9. Crossa, J.; Pérez-Rodríguez, P.; Cuevas, J.; Montesinos-López, O.; Jarquín, D.; de Los Campos, G.; Burgueño, J.; González-Camacho, J.M.; Pérez-Elizalde, S.; Beyene, Y.; et al. Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends Plant Sci.* **2017**, *22*, 961–975. [CrossRef]
10. Habyarimana, E.; Lopez-Cruz, M. Genomic Selection for Antioxidant Production in a Panel of Sorghum bicolor and *S. bicolor* × *S. halepense* Lines. *Genes* **2019**, *10*, 841. [CrossRef]
11. Meuwissen, T.H.; Hayes, B.J.; Goddard, M.E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **2001**, *157*, 1819–1829.
12. Velazco, J.G.; Jordan, D.R.; Mace, E.S.; Hunt, C.H.; Malosetti, M.; van Eeuwijk, F.A. Genomic Prediction of Grain Yield and Drought-Adaptation Capacity in Sorghum Is Enhanced by Multi-Trait Analysis. *Front. Plant Sci.* **2019**, *10*, 997. [CrossRef]
13. De los Campos, G.; Gianola, D.; Rosa, G.J.M.; Weigel, K.A.; Crossa, J. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genet. Res.* **2010**, *92*, 295–308. [CrossRef]
14. Pérez, P.; de los Campos, G. Genome-wide regression and prediction with the BGLR statistical package. *Genetics* **2014**, *198*, 483–495. [CrossRef]
15. Gianola, D. Priors in Whole-Genome Regression: The Bayesian Alphabet Returns. *Genetics* **2013**, *194*, 573–596. [CrossRef] [PubMed]
16. Kulwal, P.L. Association Mapping and Genomic Selection—Where Does Sorghum Stand. In *The Sorghum Genome*; Rakshit, S., Wang, Y.-H., Eds.; Compendium of Plant Genomes; Springer International Publishing: Cham, Switzerland, 2016; pp. 137–148. ISBN 978-3-319-47789-3.
17. Hunt, C.H.; van Eeuwijk, F.A.; Mace, E.S.; Hayes, B.J.; Jordan, D.R. Development of Genomic Prediction in Sorghum. *Crop Sci.* **2018**, *58*, 690–700. [CrossRef]
18. Ceron-Rojas, J.J.; Crossa, J.; Arief, V.N.; Basford, K.; Rutkoski, J.; Jarquín, D.; Alvarado, G.; Beyene, Y.; Semagn, K.; DeLacy, I. A Genomic Selection Index Applied to Simulated and Real Data. *G3* **2015**, *5*, 2155–2164. [CrossRef] [PubMed]
19. Céron-Rojas, J.J.; Hiriart, J.C. *Linear Selection Indices in Modern Plant Breeding*; Springer International Publishing: Cham, Switzerland, 2018; ISBN 978-3-319-91222-6.
20. Safari, P.; Honarnejad, R.; Esfahani, M. Indirect selection for increased oil yield in peanut: Comparison selection indices and biplot analysis for simultaneous improvement multiple traits. *Int. J. Biosci.* **2013**, *3*, 87–96.
21. Hazel, L.N.; Lush, J.L. The efficiency of three methods of selection. *J. Hered.* **1942**, *33*, 393–399. [CrossRef]
22. Henderson, C.R.; Quaas, R.L. Multiple Trait Evaluation Using Relatives' Records. *J. Anim. Sci.* **1976**, *43*, 1188–1197. [CrossRef]
23. Thompson, R.; Meyer, K. A review of theoretical aspects in the estimation of breeding values for multi-trait selection. *Livest. Prod. Sci.* **1986**, *15*, 299–313. [CrossRef]
24. Mrode, R.A. *Linear Models for the Prediction of Animal Breeding Values*, 3rd ed.; CABI: Boston, MA, USA, 2014.

25. Calus, M.P.; Veerkamp, R.F. Accuracy of multi-trait genomic selection using different methods. *Genet. Sel. Evol.* **2011**, *43*, 26. [CrossRef]
26. Jia, Y.; Jannink, J.-L. Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics* **2012**, *192*, 1513–1522. [CrossRef]
27. Wang, X.; Li, L.; Yang, Z.; Zheng, X.; Yu, S.; Xu, C.; Hu, Z. Predicting rice hybrid performance using univariate and multivariate GBLUP models based on North Carolina mating design II. *Heredity* **2017**, *118*, 302–310. [CrossRef] [PubMed]
28. Schulthess, A.W.; Zhao, Y.; Longin, C.F.H.; Reif, J.C. Advantages and limitations of multiple-trait genomic prediction for Fusarium head blight severity in hybrid wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* **2018**, *131*, 685–701. [CrossRef] [PubMed]
29. dos Santos, J.P.R.; de Castro Vasconcellos, R.C.; Pires, L.P.M.; Balestre, M.; Von Pinho, R.G. Inclusion of Dominance Effects in the Multivariate GBLUP Model. *PLoS ONE* **2016**, *11*, e0152045. [CrossRef] [PubMed]
30. Federer, W.T.; Cornell University, Biometrics Unit; Cornell University, Dept. of Biometrics; Cornell University, Dept. of Biological Statistics and Computational Biology. Augmented (or Hoonuiaku) Designs. *Biom. Unit Tech. Rep.* **1956**, *33*, 1.
31. Descriptors for Sorghum [*Sorghum bicolor* (L.) Moench]. Available online: https://www.bioversityinternational.org/e-library/publications/detail/descriptors-for-sorghum-sorghum-bicolor-l-moench/ (accessed on 24 November 2019).
32. Smith, H.F. A Discriminant Function for Plant Selection. *Ann. Eugen.* **1936**, *7*, 240–250. [CrossRef]
33. Tomar, S.S. Restricted selection index in animal system—A review. *Agric. Rev.* **1983**, *4*, 109–118.
34. Cartuche Macas, L. Economic weights in rabbit meat production. *World Rabbit Sci.* **2014**, *22*, 165–177. [CrossRef]
35. Bradshaw, J.E. Plant breeding: Past, present and future. *Euphytica* **2017**, *213*, 60. [CrossRef]
36. Kang, M.S. *Applied Quantitative Genetics*; Kang, M.S., Ed.; MS Kang: Baton Rouge, LA, USA, 1994; ISBN 978-0-9642970-4-3.
37. Baker, R.J. *Selection Indices in Plant Breeding*; CRC Press: Boca Raton, FL, USA, 1986; ISBN 978-0-8493-6377-1.
38. de Los Campos, G.; Grüneberg, A. QuantGen/MTM: MTM Version 1.0.0 from GitHub. Available online: https://rdrr.io/github/QuantGen/MTM/ (accessed on 24 November 2019).
39. Montesinos-López, O.A.; Montesinos-López, A.; Luna-Vázquez, F.J.; Toledo, F.H.; Pérez-Rodríguez, P.; Lillemo, M.; Crossa, J. An R Package for Bayesian Analysis of Multi-environment and Multi-trait Multi-environment Data for Genome-Based Prediction. *G3* **2019**, *9*, 1355–1369. [CrossRef]
40. Scutari, M.; Mackay, I.; Balding, D. Using Genetic Distance to Infer the Accuracy of Genomic Prediction. *PLoS Genet.* **2016**, *12*, e1006288. [CrossRef] [PubMed]
41. Raschka, S. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. *arXiv* **2018**, arXiv:1811.12808.
42. R Core Team. R: A Language and Environment for Statistical Computing. Available online: https://www.r-project.org/ (accessed on 27 September 2019).
43. Gomez, K.A.; Gomez, A.A. *Statistical Procedures for Agricultural Research*, 2nd ed.; Wiley-Interscience: New York, NY, USA, 1984; ISBN 978-0-471-87092-0.
44. Habyarimana, E.; Dall'Agata, M.; De Franceschi, P.; Baloch, F.S. Genome-wide association mapping of total antioxidant capacity, phenols, tannins, and flavonoids in a panel of Sorghum bicolor and *S. bicolor* × *S. halepense* populations using multi-locus models. *PLoS ONE* **2019**, *14*, e0225979. [CrossRef] [PubMed]
45. Wricke, G.; Weber, E. *Quantitative Genetics and Selection in Plant Breeding*, Reprint 2010 ed.; De Gruyter: Berlin, Germany, 1986; ISBN 978-3-11-007561-8.
46. Saeidnia, M.; Emami, H.; Honarnejad, R.; Esfahani, M. Comparing Economical Coefficients to Select the Best Optimum Selection Index in Peanut. *Am. Eurasian J. Agric. Environ. Sci.* **2012**, *12*, 393–398.
47. Piper, J.; Kulakow, P. Seed yield and biomass allocation in Sorghum bicolor and F1 and backcross generations of *S. bicolor* × *S. halepense* hybrids. *Can. J. Bot.* **2011**, *72*, 468–474. [CrossRef]