Optimizing genomic prediction for complex traits via investigating multiple factors in switchgrass

Peipei Wang[1,2,3]*, Fanrui Meng[1,3], Christina B Del Azodi[3], Kenia Segura Abá[1,4], Michael D. Casler[5], and Shin-Han Shiu[1,3,4,6]*

[1] DOE Great Lakes Bioenergy Research Center, Michigan State University, East Lansing, MI 48824, USA

[2] Kunpeng Institute of Modern Agriculture at Foshan, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, Guangdong 518124, China

[3] Department of Plant Biology, Michigan State University, East Lansing, MI 48824, USA

[4] Genetics and Genome Sciences Program, Michigan State University, East Lansing, MI 48824, USA

[5] Department of Plant and Agroecosystem Sciences, University of Wisconsin, Madison, WI 53706, USA

[6] Department of Computational Mathematics, Science, and Engineering, Michigan State University, East Lansing, MI 48824, USA

*: Corresponding authors

Abstract

Genomic prediction, the use of genetic information for predicting traits, has accelerated the breeding processes and provided mechanistic insights into the genetic bases of complex traits. While substantial efforts have been devoted to optimize genomic prediction, there remain areas that need to be further explored, including the impact of genome assemblies, genotyping approaches, variant types, allelic complexities, polyploidy levels, and population structures. Here, we assess the impact of these factors on the prediction of 20 complex traits in switchgrass (*Panicum virgatum* L.), a perennial biofuel feedstock. We found that short read-based genome assembly perform comparably to or even better than long read-based assembly in trait prediction; exome capture-based models have higher prediction accuracy than genotyping-by-sequencing-based models for 13 traits; bi-allelic insertion/deletions are as useful as bi-allelic single nucleotide polymorphisms in trait prediction, whereas multi-allelic variants outperform bi-allelic ones for 15 traits. Models built for tetraploids have higher prediction accuracy than those for octoploids for most traits. Traits of individuals with higher within-population genetic distances tend to have higher prediction accuracy. Finally, integrating different types of variants can improve the prediction accuracy. By exploring these factors, anthesis date prediction models built using multi-allelic insertion/deletions derived from exome capture led to the largest number of orthologs of benchmark flowering time genes compared to other models. Our study provides insights into the factors influencing genomic prediction outcomes that inform best practices for future studies and for improving agronomic traits in switchgrass and other species through selective breeding.

Key words: exome capture, flowering time, genomic prediction, genotyping-by-sequencing, multi-allelic variants, switchgrass

Introduction

Before the genomic era, the breeding process relied heavily on natural variation, and hybridization or mutagenesis, and was time- and labor-consuming. In the genomic era, the molecular variant-assisted breeding technologies, especially genomic selection or genomic prediction (GP), have dramatically accelerated the breeding processes for both animals and plants (Crossa et al., 2017; Alemu et al., 2024; Tade and Melesse, 2024). Learning from a training population with both genetic variants (e.g., single nucleotide polymorphisms, SNPs) and phenotypic data (e.g., flowering time), GP establishes statistical models to build the connection between the genetic variants and phenotypic data, which can be applied to a new population that has been genotyped but with unmeasured phenotypes. This can be done without planting all the seedlings and waiting for a whole growing season, therefore resulting in shortened breeding cycles (Estopa et al., 2023; Alemu et al., 2024).

Substantial effort has been made to understand the effects of different factors on the accuracy of GP (Norman et al., 2018; Zhang et al., 2019; Kriaridou et al., 2020; Batista et al., 2022; Alemu et al., 2024; Zheng et al., 2024), including variant density (Kriaridou et al., 2020), population size (Akdemir et al., 2019; Wu et al., 2023), population structure (Norman et al., 2018), relationship between training and test populations (Isidro et al., 2015), functional genomic annotations (Zheng et al., 2024), marker ploidy (Aalborg and Nielsen, 2024), variants besides SNPs (e.g., structural variants, Liu et al., 2024) and statistical methods (Azodi et al., 2019; Zhang et al., 2019; Wang et al., 2023b; Alemu et al., 2024), etc. For example, the prediction accuracy increases along with the increase of variant density to a certain point (Norman et al., 2018; Kriaridou et al., 2020). Regarding the statistical methods used in GP, a relatively simple linear model, ridge regression Best Linear Unbiased Prediction (rrBLUP), performed similarly or better compared with other “state-of-the-art” approaches for most scenarios, such as random forest, support vector machine, gradient boosting, artificial neural net, and convolutional neural net (Azodi et al., 2019; Azodi et al., 2020; Wang et al., 2024). Beyond the above, there are additional factors with their effects on GP accuracy remaining largely unexplored, such as genome assemblies, genotyping approaches, variant allelic complexities, and so on. With reduced sequencing cost and improvement in assembly algorithms, a number of plant species have genome assemblies improved from scaffold-level to chromosome-level or even telomere-to-telomere (T2T) (Chen et al., 2023; Lu et al., 2024). However, the higher the quality of the

assembly, the more the cost needed, and it remains unknown whether a better assembly is necessarily to improve the GP accuracy since few studies have focused on the effects of assembly quality on GP accuracy. A blueberry study showed that the chromosome-level assembly 'Draper' led to comparable prediction accuracy for most traits as the scaffold-level assembly 'W8520' with only half numbers of probes (Benevenuto et al., 2019), but it was still unknown whether the chromosome-level assembly led to improved GP accuracy or not when the genome-wide variants were used.

A number of approaches have been developed for genotyping, including SNP array, genotyping by random amplicon sequencing-direct system, genotyping-by-sequencing, exome capture, whole genome sequencing (WGS), and others (Scheben et al., 2017; Ayalew et al., 2022; Boatwright et al., 2022; Wang et al., 2023a; Minamikawa et al., 2024). WGS covers the majority of genetic variants across the whole genome, whereas the other approaches only cover a portion of variants but with relatively lower costs. It remains unclear how different genotyping approaches impact the GP accuracy and which approaches would be a better option with different aims of the studies.

The genetic variants can be both bi-allelic and multi-allelic: out of the SNPs called for the Arabidopsis 1001G data using the 250k SNP-array from Kim et al. (2007), 7% were multi-allelic (Alonso-Blanco et al., 2016); out of the SVs identified using 32 *A. thaliana* ecotypes, multi-allelic SVs (29.65 Mb) were much more than bi-allelic SVs (7.51 Mb, Kang et al., 2023). However, the majority of past genome-wide association study (GWAS) and GP studies mainly used bi-allelic variants (Crossa et al., 2017; Misra et al., 2017; Song et al., 2018; Wang et al., 2023b). It is important to know how the multi-allelic variants contribute to complex trait prediction to make the best of the ever increasing genomic data, such as graph-based pan-genomes that are used as references to call all types of genetic variants with high quality.

Here, using switchgrass (*Panicum virgatum* L.)—a perennial grass with both tetra- and octo-ploids and a key species as a bioenergy feedstock (McLaughlin and Adams Kszos, 2005)—as a model system, we aim to optimize GP by assessing the impacts on GP accuracy due to differences in: 1) genome assembly versions, 2) genotyping strategies, 3) genetic variant types, 4) numbers of variant alleles, and 5) polyploids and population structures. Both short-read and long-read based assemblies are available for switchgrass, and there are two genotyping datasets for a switchgrass diversity panel with 486 individuals: GBS (Lu et al., 2013) and EC (Evans et al.,

2015). We also assess how different types of genetic variants impact our ability to uncover the genetic basis underlying anthesis date by interpreting GP models.

Results

Baseline models predicted trait values with variable accuracy

There are 20 phenotypic trait values available for the switchgrass diversity panel we used (**Table S1**), including 7 morphological and 13 biochemical traits (**Table S2**, Lipka *et al.*, 2014). Among the 20 traits, some are highly associated with each other, such as three biochemical traits, namely cell wall concentration, etherified ferulates and *in vitro* dry matter digestibility, and two flowering time-related traits, heading date and anthesis date (**Fig. 1A**). To determine how well these 20 phenotypic traits can be predicted in switchgrass, we first built GP models using GBS bi-allelic SNPs that were mapped to the version 5 genome assembly of the lowland tetraploid AP13 (Lovell et al., 2021). The rrBLUP approach was used to build the models due to its relatively better or similar prediction accuracy compared with other methods, and its interpretability for underlying molecular mechanisms for target traits (Azodi et al., 2019; Azodi et al., 2020; Wang et al., 2024). Information of 20% of the individuals in the diversity panel were held out as the test set before model building to serve as an independent dataset for evaluating the performance of final models. The data of the remaining 80% individuals were used to train the models with a five-fold cross-validation scheme (see **Methods**). To compare prediction accuracy among different traits, we also established models using the population structure (p; defined as the first five principal components of the corresponding genetic variants [g], see **Methods**) to determine the baseline prediction accuracy for each trait (Azodi et al., 2020). The r^2 of the Pearson Correlation Coefficient (PCC) between the true and predicted trait values for individuals were first calculated for models built using genetic variants (r^2_g) and population structure (r^2_p), separately. Then the improvement in r^2 ($r^2_i = r^2_g - r^2_p$) was used to measure the performance of the GP models. This r^2_i was calculated both for individuals in the cross-validation (CV) set ($r^2_{i,CV}$) and the test set ($r^2_{i,test}$).

We found that $r^2_{i,CV}$ was positively correlated with $r^2_{i,test}$ (PCC=0.68, *p*-value=9.89e-04, **Fig. 1B,C**), suggesting the relative robustness of our models. In general, $r^2_{i,CV}$ for morphological traits (average $r^2_{i,CV}$=0.176) were higher than those for biochemical traits (average $r^2_{i,CV}$=0.055, *p*-value

of Wilcoxon signed rank test=4.66e-03), and the two flowering time related traits—anthesis date and heading date—had the highest $r^2_{i,CV}$ (0.289 and 0.250, respectively). These results indicate that morphological traits tended to have higher heritability, whereas the prediction accuracy (r^2_g) for most biochemical traits were mainly confounded by population structure. Thus, cautions need to be taken in the GWAS and GP practices for those biochemical traits.

Better genome assembly did not provide better trait prediction

To test whether the genome assembly with improved quality (i.e., v5, Lovell et al., 2021) would lead to higher prediction accuracy for switchgrass traits than short read-based assembly (e.g., v1), we also mapped GBS sequences to the v1 assembly and called the genetic variants using the same methods and parameter settings as those for v5-based variants (see **Methods**). The assembly sizes of v1 (1,230 Mb) and v5 assemblies (1,129 Mb) were similar, and there were almost equivalent numbers of variants for v1 (11,042) and v5 (11,020) assemblies, with similar distribution of variants across different gene functional regions (**Fig. S1A,B** and **Table S3**). In addition, most of the GBS bi-allelic SNPs were shared between these two assemblies, as 5,283 (69.92%) out of 7,556 v1-based bi-allelic SNPs had corresponding v5-based ones (see **Methods** and **Table S4**). Beyond these similarities, N50 for v5 assembly (N50=5.5 Mb) is nearly 100 times higher than v1 (N50=54 Kb). However, v5-based models outperformed v1-based models only for anthesis date, with the average difference in $r^2_{i,CV}$ between two models for 20 traits was only 0.006 (the "All" column in **Fig. 2A, Fig. 2B**; for the $r^2_{p,CV}$, $r^2_{g,CV}$, $r^2_{i,CV}$, $r^2_{p,test}$, $r^2_{g,test}$, and the $r^2_{i,test}$, see **Fig. S2-7**, respectively). This result suggests that the improvement in contig size may not be relevant to GP because genetic variants are treated as independent variables in GP algorithms.

In fact, although the effect sizes (differences in $r^2_{i,CV}$) were small, models built with variants derived from v1 had significantly better $r^2_{i,CV}$ than v5-based models for seven traits (the "All" column in **Fig. 2A**). These differences in $r^2_{i,CV}$ can be partially explained by the assembly-specific variants: models built using common variants (shared between v1 and v5 assemblies) had similar performance (the "Common" column in **Fig. 2A**, **Fig. 2C**), while assembly-specific variants-based models tended to have better performance when they were v1-based (the "Assembly-specific" column in **Fig. 2A**, **Fig. 2D**). A possible explanation of this finding that needs to be investigated further is that, while v5 assembly had improved contig sizes, some genomic regions present in the v1 assembly were not assembled from long-reads and thus were no longer present in the v5

assembly. These results suggest that, in switchgrass, short-read-based assembly is equally competent as or even superior to long-read-based assembly for the purpose of GP. Nevertheless, since the v5 assembly is used by the community, our subsequent analyses are based on genetic variants called using the v5 assembly.

Exome capture SNPs led to models better than those based on GBS

EC and GBS are two commonly used genotyping methods to capture the genetic variants with a much lower cost than WGS, but use different strategies: EC identifies genetic variants that may alter protein sequences, while GBS captures variants around restriction sites of (a) given restriction enzyme(s). We found that when using bi-allelic SNPs, EC-based models had significantly higher $r^2_{i,CV}$ than GBS-based models for 15 traits but lower $r^2_{i,CV}$ for only two traits (the "Unbalanced" column, **Fig. 3A**). Considering that there were ~72 times more EC variants (526,705, **Fig. S1C**) than GBS ones (7,357), to eliminate the influence of variant numbers on prediction accuracy, we randomly down-sampled EC bi-allelic SNPs to the same number of GBS bi-allelic SNPs, and conducted this down-sampling 100 times (**Methods**). The median $r^2_{i,CV}$ of the 100 models built with these down-sampled EC variants still had significantly higher $r^2_{i,CV}$ than GBS-based models for 13 traits, but had significantly lower $r^2_{i,CV}$ than GBS-based models for only three traits (the "Balanced" column, **Fig. 3A**). These results suggest that, regardless of the numbers of variants, the distribution of variants may be the main factor influencing GP accuracy when comparing EC and GBS-based models.

We found that EC variants tended to be located in genic regions or in intergenic regions that are closer to genes (**Fig. 3B, Fig. S7**), and contained more genic variants (86.4%) than GBS variants (55.5%, **Table S3**). We further established GP models using only genic GBS and EC variants (down-sampled to match the number of genic GBS variants) and found that genic GBS-based models still had significantly lower $r^2_{i,CV}$ than genic EC-based models for 15 of 20 traits (the "Genic" column, **Fig. 3A**). When examining the GBS bi-allelic SNPs and the 100 down-sampled subsets of EC bi-allelic SNPs in detail, we found significantly fewer genes that contained GBS variants (3,296) than contained EC variants (median across 100 replicates=3,754; *p*-value=9.31e-21, **Fig. 3C**). These findings indicate that having higher gene coverage, rather than more genic variants, may explain the higher GP accuracy of EC-based models.

Furthermore, we also built models using intergenic GBS and EC bi-allelic SNPs, and found that GBS-based models had significantly smaller $r^2_{i,CV}$ than EC-based models (balanced for variant numbers) for 10 of 20 traits, but significantly higher $r^2_{i,CV}$ only for five traits (the "Intergenic" column, **Fig. 3A**). Similar to genic variants, there were fewer genes adjacent to (3.5 Kb up- or downstream the variants) intergenic GBS variants (2,630, *p*-value=2.36e-18, **Fig. 3D**) than intergenic EC variants (down-sampled to match intergenic GBS, the median number of genes among 100 replicates was 2,944). This result suggests that, beyond higher gene coverage, including variants that are adjacent to more genes may also improve the GP accuracy, which is potential due to the linkage between these intergenic variants and genes. Another potential explanation for the better performance of models built using intergenic EC variants was because the intergenic EC variants were closer to gene bodies than the intergenic GBS variants (**Fig. 3B**). If this was true, we would expect negative correlations between the distance to the gene body and the absolute coefficient of variants that was used as a proxy for the contribution of variants to trait prediction. However, we did not see such correlation for leaf width models as an example (Spearman's rank correlation coefficients = -7.7e-03 and 7.2e-03 for models built with intergenic GBS and EC bi-allelic SNPs, respectively), for which the EC-based model had significantly higher $r^2_{i,CV}$ than the GBS-based model (the "Intergenic" column, **Fig. 3A**). Taken together, those results suggest that EC variants are superior to GBS variants in predicting switchgrass traits, and gene coverage by variants has a significant impact on how well a trait can be predicted.

Multi-allelic variants outperformed bi-allelic variants in trait prediction

Up to this point, all the genetic variants we have explored are bi-allelic SNPs, which are commonly used in GWAS and GP studies. Besides bi-allelic SNPs, bi-allelic indels and multi-allelic variants also contribute to phenotypic variation, and are informative for trait prediction (Veerkamp et al., 2016; Biová et al., 2024), but with the degree to which they contribute to GP accuracy largely unexplored. For the GBS data, compared with the 7,357 bi-allelic SNPs, 1,827 bi-allelic indels were identified (**Table S3**). Models built using bi-allelic indels had significantly lower $r^2_{i,CV}$ than SNP-based models for 15 traits (without down-sampling, the 1st column, **Fig. 4A)**, but had comparable performance as balanced SNP-based models (with lower $r^2_{i,CV}$ for 6 traits and higher $r^2_{i,CV}$ for another 6 traits, the 2nd column, **Fig. 4A**). These results suggest that,

although occur less frequently than SNPs, indels are comparably useful as SNPs when the numbers are balanced.

In addition to bi-allelic indels, there were 1,836 multi-allelic variants identified for the GBS data, including multi-allelic SNPs (350), multi-allelic indels (604) and multi-allelic SNPs/indels (882) (before encoding, **Table S3**). To simplify the comparison, we built models using only the 350 GBS multi-allelic SNPs. Multi-allelic SNP-based models had significantly lower $r^2_{i,CV}$ than models based on bi-allelic SNPs (without down-sampling) for 16 traits (the 3rd column in **Fig. 4A)**, but had significantly higher $r^2_{i,CV}$ for 15 traits than models built using down-sampled bi-allelic SNPs (the 4th column in **Fig. 4A**). Consistent with our earlier results indicating the importance of gene coverage by variants in GP, we found 0.83 times more genes covered by multi-allelic SNPs (632) than balanced bi-allelic SNPs (median: 346, *p*-value = 9.87e-24, **Fig. 4C**), but slightly fewer genes covered by bi-allelic indels (1,635, compared with median 1,711 genes covered by the corresponding balanced bi-allelic SNPs, *p*-value = 1.02e-09, **Fig. 4B**). Taken together, these results indicate that multi-allelic SNPs outperform the commonly used bi-allelic SNPs in trait prediction, particularly in their contribution to improve the coverage of genes, and should be included in future studies.

Next, we asked whether the prediction can be improved by taking advantage of all the types of variants. We built new GP models using all the GBS variants, including 13,265 bi- and multi-allelic SNPs and indels in total. This was not conducted for EC variants, due to the large number of EC variants (2,552,214 variants in total) and the extremely high computing resource requirement to establish models using all the EC variants. We found that five and seven traits had significantly improved $r^2_{i,CV}$ and $r^2_{i,test}$ in the integrated models compared with models that were built using individual types of GBS variants separately, respectively (**Fig. 5A**,**B** for $r^2_{i,test}$, r^2_p and r^2_g, see **Fig. S8**). These results suggest that different types of genetic variants should be integrated to improve the GP accuracy for switchgrass traits, and potentially traits in other plant species as well.

Models built for octoploids had lower trait prediction accuracy than those for tetraploids

Thus far, the GP models were built for all individuals, including both tetraploids and octoploids, where the variants for octoploid individuals were treated using a pseudo-diploid

genotyping strategy (see **Methods**), which is commonly used to handle variants for individuals with different ploidy levels in previous studies. Considering the relatively higher complexity of variants for octoploid individuals, we asked whether the same variant set (GBS or EC) would have different prediction accuracy for tetraploid and octoploid individuals. When GBS bi-allelic SNPs were used, models built for tetraploids had significantly higher $r^2_{i,CV}$ than octoploid models for 16 traits, but significantly lower $r^2_{i,CV}$ for only 3 traits (the "GBS" column, **Fig. 6A**). One possible explanation of the lower $r^2_{i,CV}$ for octoploid models is that the variants called for octoploids were not as good as those for tetraploids in GBS data, due to the relatively low read depth (RD) of GBS bi-allelic SNPs (average at 4.41, left panel in **Fig. 6B**), and potentially a need for deeper RD for octoploids than tetraploids in variant calling. In contrast, the EC data had an average RD at 21.10 (left panel in **Fig. 6C**), and much lower proportions of missing data for both tetraploids (3.1% [16,237], middle panel in **Fig. 6C**, compared with 12.5% [916] for GBS, **Fig. 6B**) and octoploids (2.4% [13,023], right panel in **Fig. 6C,** compared with 12.1% [893] for GBS, **Fig. 6B**). EC-based models for tetraploids still had significantly higher $r^2_{i,CV}$ than octoploid models for 12 traits (for the GBS-based models, this number was 16), but significantly lower $r^2_{i,CV}$ for the other 6 traits (3 for the GBS-based models) (the "Exome capture" column, **Fig. 6A**). These results indicate that the low RD in GBS data might, but only partially, explain the lower GP accuracy of octoploid models.

Another potential explanation for lower GP accuracy for octoploids is the different population structures between octoploids and tetraploids, since both $r^2_{p,CV}$ (**Fig. 6D**) and $r^2_{g,CV}$ (**Fig. 6E**) for octoploids were generally smaller than those values for tetraploids, regardless of whether GBS or EC data were used. One previous study (Evans et al., 2015) showed that the within-subpopulation genetic distances in two octoploid subpopulations were shorter than those in three tetraploid subpopulations (see **Fig. 4** in Evans et al., 2015). We further built a model for each of four subpopulations with ≥85 individuals each (**Table S1**) using EC bi-allelic SNPs, and found that subpopulations with larger within-subpopulation genetic distances tended to have higher GP accuracy (either $r^2_{p,CV}$, $r^2_{g,CV}$ or $r^2_{i,CV}$) than those with shorter distances (**Fig. S9**). This is consistent with findings in diverse panels of rice and maize that within-subpopulation genetic variance dominated predictions (Guo et al., 2014), but inconsistent with findings in Barley that GP accuracy decreased with the increase of genetic distances in the training population (Lorenz and

Smith, 2015). Taken together, these results indicate that read depth for genetic variants and population structure are potentially another two factors influencing GP accuracy, especially for octoploids in switchgrass.

Insights of molecular mechanisms underlying trait determination by interpreting GP models

Besides the prediction of complex traits, the GP models can be used to get insights of molecular mechanisms underlying trait variations as well, by examining the absolute coefficients of genetic variants in the trained models. Variants with absolute coefficients ranked above 95th or 99th percentiles were considered as important to trait prediction, and genes harboring or nearby the important variants (<3.5 Kb) were considered as important genes. We focused on the anthesis date models since the current knowledge for genetic bases of flowering time is the most abundant among all the 20 traits studied here, and anthesis date had the highest $r^2_{i,CV}$ (**Fig. S3**) and one of the highest $r^2_{i,test}$ (**Fig. S6**). To test how well anthesis date models recover flowering time genes, we collected 23 maize and 378 Arabidopsis flowering time genes (**Table S5**) to identify switchgrass orthologs that were used as benchmark flowering time genes (hereafter referred to as FT-genes, see **Methods**, **Table S6**).

We found that the number of FT-genes identified based on model coefficients is not correlated with model performance ($r^2_{i,CV}$), either at the 95th (PCC=0.56, *p*-value=0.15) or 99th percentile (PCC=0.45, *p*-value=0.27) threshold, suggesting that models with better anthesis date prediction do not necessarily had more FT-genes identified as important. In contrast, the number of FT-genes identified by a model was positively correlated with the number of variants used in the model (**Fig. 7A**, **Fig. S10**, **Table S7-S15**). Specifically, GBS bi-allelic SNPs, bi-allelic indels, multi-allelic SNPs and multi-allelic indels identified only 4 (1), 2 (0), 0 (0) and 1 (0) FT-genes as important, respectively, whereas EC bi-allelic SNPs, bi-allelic indels, multi-allelic SNPs and multi-allelic indels identified 187 (51), 102 (13), 162 (37) and 314 (106) FT-genes as important, respectively, when 95th (99th) percentile was used as a threshold (**Fig. 7B**, **Table S15**). Only models built using EC bi-allelic SNPs and EC multi-allelic indels identified significantly more FT-genes than random guessing (*p*-value for fisher's exact test < 0.034, red bars in **Fig. 7B**) and had the highest odds ratios among other models.

Out of the seven FT-genes identified by GBS variants as important genes when the 95th percentile was used as the threshold, six were also identified by EC variants (**Fig. S10, Table S15**). For example, the GBS bi-allelic SNP Chr01K_35695336, which ranked 26th in terms of absolute coefficient, was located within the genic region of *Pavir.1KG328000*, an ortholog of Arabidopsis *EARLY FLOWERING IN SHORT DAYS* (*EFS*), which represses flowering in the autonomous promotion pathway (Soppe *et al.*, 1999). Within *Pavir.1KG328000*, there were also several EC bi-allelic SNPs (e.g., Chr01K_35693499), bi-allelic indels (Chr01K_35687142), multi-allelic SNPs (Chr01K_35692786_T) and multi-allelic indels (Chr01K_35690136_TC) with absolute coefficients ranking above 95th percentile in the corresponding models. Individuals with homozygous reference alleles (1) for these loci flowered significantly differently as individuals with homozygous alternative (-1) and heterozygous alleles (0) (**Fig. 7C-G**). Besides the seven FT-genes identified by GBS variants, all the other identified FT-genes were EC variant-specific, further suggesting that EC variants are superior to GBS variants in either trait prediction or identification of potential contributing genes for traits in question.

Furthermore, we investigated in detail how multi-allelic variants contributed to trait prediction, by taking the locus Chr08K_52608078 as an example. This locus was located 67 bp upstream of *Pavir.8KG380900*, an ortholog of the Arabidopsis flowering time gene *VERNALIZATION INDEPENDENCE 3* (*VIP3*), which functions as an activator of the flowering-repressor gene *FLOWERING LOCUS C* (*FLC*) (Zhang et al., 2003). This locus had six alleles in the EC data among individuals studied here: the reference allele is CAA, and five alternative alleles included one SNP: AAA, and four indels: C, CA, CAAAA, and * (CAA was completely absent). In the model built using EC multi-allelic indels, Chr08K_52608078_* (ranked 15th, **Fig. 7H**), Chr08K_52608078_CA (ranked 102nd, **Fig. 7I**), and Chr08K_52608078_C (ranked 35069th, **Fig. 7J**) had absolute coefficients above the 95th percentile, and individuals with different alleles of these three indels had significantly different anthesis date. Interestingly, when examining different alternative alleles for this single locus, individuals with homozygous reference alleles (see **Methods**) flowered significantly earlier (**Fig. 7H**) or later (**Fig. 7I,J**) than individuals with heterozygous alleles and homozygous alternative alleles, suggestive of the complexity of multi-allelic variants in trait determination. In addition, when examining the original allele compositions of this locus among individuals (**Fig. 7K**), we found that there were 39 combinations of all the six alleles. Tetraploid individuals with homozygous or heterozygous alleles of C flowered earliest,

and tetraploid individuals with */*, AAA/CA and CA/* flowered latest, whereas all the other tetraploids and all the octoploids flowered medially. This finding suggests that promoter sequences with a C at this locus may interrupt the expression of *Pavir.8KG380900*, leading to decreased expression of orthologs of *FLC* and resulting earlier flowering phenotypes, if the regulatory rules of *VIP3* and *FLC*-like genes in flowering are conserved between Arabidopsis and switchgrass. This type of analysis highlights the valuable insights into trait prediction provided by multi-allelic variants that were normally neglected by previous studies, and the need to include multi-allelic variants in future studies of GWAS and GP.

Discussion

Using switchgrass as a model system, our study showed the differences of two types of variants (e.g., GBS vs EC SNPs) for a given factor (e.g., genotyping approach) in the prediction accuracy for 20 traits, and proposed suggestions that can be potentially taken in future GP practices. Generally, different versions of genome assemblies (short read-based vs. long read-based) had similar GP accuracy; genotyping approaches that capture variants with higher gene coverages should be considered with limited budgets; multi-allelic variants should be included in GP practices both for improving GP accuracy and identifying potential contributing variants for the target traits; and different types of variants should be integrated to improve the GP accuracy.

Although these conclusions summarized above were generally true for most traits, there were a few exceptions. For example, EC-based models tended to outperform GBS-based models for most traits, but the plant height (base of longest flowering stem to the node at the base of the panicle) and total plant height (base of the longest flowering stem to the tip of the panicle) were predicted with consistently higher accuracy by GBS-based models, no matter whether the balanced, unbalanced, genic or intergenic variants were used (**Fig. 3A**). This finding indicates that variations in plant height might be determined more by variants in genomic elements that are located within intergenic regions than those within genic regions. In addition, models built for tetraploids tended to have better prediction accuracy for most traits than those for octoploids, but the two plant height traits and ethanal (ethanol/g dry forage) were better predicted in octoploid models, both when GBS and EC variants were used (**Fig. 5A**). These exceptions, together with findings in Wang et al. (2023a) that whether Deep Learning approaches outperformed classical

approaches depends on the datasets used and traits in question, suggest that the optimal GP practices should be considered with caution on a case-by-case basis.

Based on our findings, in switchgrass, short read-based assembly led to comparable (for 12 traits) or even significantly better prediction accuracy (for 7 traits) for switchgrass complex traits, whereas long read-based assembly only led to better prediction accuracy for anthesis date, for which the difference in $r^2_{i,CV}$ between short read- and long read-based models was only 0.0163. Those results suggest that, for a plant species or lineage with only reference genome assembly based on short reads, it might not be necessary to update the assembly using long reads if only for improving the GP accuracy. However, if both the short and long read-based assemblies are available, long read-based assembly will still be a better option since ⅰ) the differences in $r^2_{i,CV}$ were relatively small (max = 0.0324); ⅱ) long read-based assembly tend to be more commonly used by the community, and is much more competent than short read-based assembly for calling SVs, another genetic variants that were not studied in this study but have been shown great power in GP and GWAS practices.

In addition, our findings suggested that multi-allelic variants outperformed bi-allelic ones in trait prediction and identification of potential important genes that contribute to the formation of traits in question. Multi-allelic tandem repeats (He et al., 2024) and multi-allelic copy number variations (Handsaker et al., 2015) have been shown to contribute to the rice agronomic traits and gene expression dosage variations in humans, respectively. Although ~10% (Jiang et al., 2020) or even a higher proportion (Kang et al., 2023) of genetic variants are multi-allelic, previous studies generally used bi-allelic variants only. Besides the strategy we used in this study to encode the multi-allelic variants, the BCFtools offered a model recently to handle loci with multiple alternative alleles (Danecek et al., 2021), which would help us make the best use of all the genetic variants in GWAS and GP practices.

Our results indicate that gene coverage by variants had a significant effect on GP accuracy, which gives recommendations for the selection of genotyping approaches and designation of SNP chips in future studies. A pan-genome study of 69 *A. thaliana* accessions showed that 18% of the 32,986 gene families were private to a single accession (Lian et al., 2024), indicating potential missing heritability by using a single reference genome. However, in most previous studies and our study, genetic variants were generally called by referencing a single genome assembly. Recently, more and more graph-based pan-genomes have been constructed to

capture the missing heritability in plant species (Zhou et al., 2022; Kang et al., 2023; Yan et al., 2023; Liu et al., 2024). For example, including SVs identified based on pan-genome has been shown to significantly improve the estimated heritability for grape traits (Liu et al., 2024). It is important to know whether including other hidden genomic variants (e.g., SNPs and indels) revealed by pan-genome would also help improve GP accuracy for plant traits, especially for traits enabling a small group of ecotypes/cultivars/individuals to adapt to particular environments or endowing them certain horticultural characteristics to be distinguished from others.

In this study we showed that models built for tetraploids tended to have better prediction accuracy for most traits than those for octoploids, potentially due to relatively longer genetic distances among tetraploids. Another potential reason that can not be ruled out is that the allele dosage in octoploids were much more complex than those in tetraploids, whereas we took the strategy of pseudo-diploid genotyping that grouped all heterozygotes (e.g., AT, ATTT, AATT, AAAT) into a single class. Some studies treated heterozygotes in an allele-dosage way (Yadav et al., 2024) or under different effect assumptions (namely additive, duplex dominant, and simplex dominant, Rosyara et al., 2016; Wilson et al., 2021) when dealing with variants for polyploids. Considering that it is unlikely that genes/loci with different allele compositions across the whole genome have a same effect pattern, and that multiple ploidy levels may coexist in some crops (e.g., diploids, triploids, tetraploids and aneuploids coexist in *Phalaenopsis* orchids), tools that can detect the potential effect pattern for each locus based on prior knowledges or integrate all the possible effect patterns of alleles, and handle with multiple levels of ploidies simultaneously, are needed for crop selective breeding. The prior knowledge can be that, under which assumption, the statistical significance between genetic variants and traits of individuals with different allele compositions is the highest.

Materials and Methods

Genomic and phenomic data

Two versions of switchgrass genome assemblies (v1 and v5) were downloaded from Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html). The GBS data was obtained from (Lu et al., 2013), and the raw data were downloaded from NCBI under the project PRJNA201059 (**Table S1**). The GBS barcodes for individuals in each pooled GBS library were obtained from the

description of each SRA sample in the NCBI website. The EC data was from (Evans et al., 2015), and the raw data were downloaded from NCBI under the project PRJNA280418 (**Table S1**). Phenotypic data were from the diversity panel which consisted of 540 individuals from 66 populations (Lipka et al., 2014), including seven morphological and 13 biochemical traits (**Table S2**), where the cumulative growing degree days was used to record the anthesis date and heading date. After filtering out individuals with low quality of GBS data, ploidy of 6, and the ones with information missing from any of the 20 traits, we retained 486 individuals in this study (**Table S2**), including 263 tetraploids and 223 octoploids.

SNP calling

Reads from the GBS and EC data were trimmed using Trimmomatic (Bolger et al., 2014) with parameters: 2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:35. Trimmed reads with scores > 20 and longer than 35 bp were kept for the following analysis. GBS reads were mapped to both two assemblies using bwa/0.7.12.r1044 (Li, 2013) with default parameter setting, while EC reads were only mapped to the v5 assembly. The alignment after mapping was sorted using picardTools/1.113 (http://broadinstitute.github.io/picard).

SNP calling was conducted using GATK/3.5.0 (Auwera and O'Connor, 2020). For EC data, duplication alignments, which may be amplified during library preparation progress, were marked using MarkDuplicates module. No treatment was processed to GBS alignments, because reads in GBS data always had the same start point (the same restriction size, *Ape*KI site) (Lu et al., 2013) and stop point. SNPs and indels were identified using HaplotypeCaller module, with parameters of -stand_call_conf 40 -stand_emit_conf 10 --max_alternate_alleles 8 -ploidy x, where x = 2 for tetraploids (since the reference genome was also a tetraploid) while 4 for octoploids. Base-quality recalibration was conducted using BaseRecalibrator and IndelRealigner for SNPs and indels, respectively. Then the HaplotypeCaller module was used for the second run of SNP and indel calling with the same parameter setting and the output files were saved at GVCF format. All GVCF files for individuals were merged together using the CombineGVCFs module. The resulting SNPs were filtered using VariantFiltration with parameters of "QD < 2.0 || MQ < 40.0 || FS > 60.0 || MQRankSum < –12.5 || ReadPosRankSum < –8.0 –clusterSize 3 –clusterWindowSize 10", while the indels were filtered with "QD < 2.0 || FS > 200.0 ||

ReadPosRankSum < –20.0”. Finally, SNPs and indels with minor allele frequency (MAF) > 0.05, and missing data < 20% were kept in the final list.

Encoding of multi-allelic variants

Multi-allelic variants were encoded using the method proposed by Zhan et al. (2016)with modifications, where m columns were encoded for m alternative alleles (m≥1). For example, for a tetra-allelic variant (e.g, the reference allele is A, three alternative alleles are T, G, AT), there would be three columns (**Table 1**): A/A, A/T or T/T for the first alternative allele, A/A, A/G or G/G for the second, and A/A, A/AT/ or AT/AT for the third. The first two columns were used as multi-allelic SNPs, and the third column was used as a multi-allelic indel. After encoding the multi-allelic variants, there were 1,628 GBS multi-allelic SNPs, 2,454 GBS multi-allelic indels, 444,102 EC multi-allelic SNPs, and 1,341,337 EC multi-allelic indels.

Table 1. Encoding for multi-allelic variants with A as the reference allele, and T, G and AT as the alternative alleles.

	Sequence	Column_1 (T)	Column_2 (G)	Column_3 (AT)
ID_1	A/A	1 (A/A)	1 (A/A)	1 (A/A)
ID_2	A/T	0 (A/T)	1 (A/A)	1 (A/A)
ID_3	T/G	0 (A/T)	0 (A/G)	1 (A/A)
ID_4	G/AT	1 (A/A)	0 (A/G)	1 (A/AT)
ID_5	AT/AT	1 (A/A)	1 (A/A)	-1 (AT/AT)

Imputation of missing data

For bi-allelic SNPs in tetraploids, the missing data were imputed using fastPHASE (Scheet and Stephens, 2006). Variants called for octoploids were treated as those called for tetraploids. For example, a variant with homozygous reference allele AAAA and a variant with homozygous alternative allele TTTT were treated as AA and TT, respectively, while all three types of

heterozygous alleles — ATTT, AATT and AAAT — were treated as AT. For the bi-allelic indels, the reference allele was converted to “R”, and the alternative allele was converted to “A”, and then the missing data were imputed using fastPHASE. The multi-allelic variants were first encoded to bi-allelic ones, and then were imputed using fastPHASE.

Genomic prediction model

To build GP models, homozygous reference alleles and alternative alleles were encoded as 1 and -1, respectively, while heterozygous variants were encoded as 0. The R package “rrBLUP” (Endelman, 2011) was used to build the GP models because it is one of the methods which have relatively higher prediction accuracy compared with others in general (Lipka et al., 2014; Azodi et al., 2019). For each trait, all the 486 individuals were first split into training (389, 80%) and test (97, 20%) sets using the stratified strategies to make sure that individuals in training and test sets had similar trait value distribution. The training set was then used to establish predictive models, using a five-fold cross-validation scheme: (1) all the training individuals were first split into five folds; (2) individuals in four folds (referred to as training subset) were used to build models, while individuals in the other fold (referred to as validation subset) was used to evaluate the model performance; (3) the second step was conducted five times to make sure each fold will be used as validation subset once; (4) the r^2 of PCC between true and predicted trait values of individuals in all five folds was calculated and was referred to as r^2_{CV}. This five-fold cross-validation scheme was repeated 10 times, and the median r^2_{CV} value of 10 replicate runs was calculated. During the model training, the coefficient of each variant was calculated for each cross-validation step and was averaged across five cross-validations. Then the absolute mean coefficient of a variant was used as the measure of importance for this variant in a model.

To balance the numbers of variants used between two models (e.g., models built using GBS bi-allelic SNPs vs those using EC bi-allelic SNPs), the one with more variants (i.e., EC bi-allelic SNPs) was down sampled to the same number as the one with fewer variants (GBS bi-allelic SNPs). This random down-sampling was conducted 100 times, resulting in 100 down-sampled data. The median prediction accuracy across the 100 models building using these 100 down-sampled data was used to measure the model performance.

Association of GBS bi-allelic SNPs between two assemblies

To assess which GBS bi-allelic SNPs were shared between v1 and v5 assemblies, two approaches were explored in this study. First, bi-allelic SNPs from two assemblies were associated by aligning sequences of these two assemblies using MUMmer/4.0.0beta2 (Kurtz et al., 2004). Sequences of the v1 chromosomes and contigs were aligned with those of v5 chromosomes and scaffolds, using the function “NUCmer” with the option of --maxmatch. The function “mummerplot” was used to output the coordinates and the sequence identities of matched regions. Only regions with sequence identity > 95% were kept for downstream analysis. For each pair of matched regions, if the v1 region contains a SNP, then search for SNPs on the corresponding v5 region. A v1 SNP was associated with a v5 SNP if sequences 20bp upstream or downstream of the v1 SNP matched to those of the v5 SNP (**Table S4**).

The second approach is to align the v1 SNP in question and its neighbor sequences (500 bp up- and 500 bp downstream) to the v5 genome sequence using BLASTN (Altschul et al., 1990). If the best matched v5 region for a v1 1001 bp region contains a SNP at the same relative coordinate as the v1 SNP, then the v1 SNP and the v5 SNP are associated.

By comparing results from these two approaches, out of 7556 v1 bi-allelic SNPs, 1502 (19.9%) were associated with v5 SNPs from both approaches, 3777 (50.0%) and four (0.05%) were associated with v5 SNPs from only blastn and Mummer methods, respectively. The corresponding v5 sequences of 2118 v1 SNPs (28%) were not identified as having SNPs. This may be because SNPs were filtered by applying a hard threshold on continuous scores (i.e., parameter setting during SNP calling), thus SNPs with scores close to the threshold may be filtered out in one assembly but not in the other one. The flank regions of five v1 SNPs had no matched sequences in v5, 83 v1 SNPs had no matches in v5 assembly near the SNPs, and 67 had different alleles between v1 and v5 assemblies (**Table S4**).

Orthologs to known flowering time genes

Known genes involved in flowering time in maize were gained from (Azodi et al., 2020) and the Maize Genetics and Genomics Database (https://www.maizegdb.org/). Flowering time genes in *Arabidopsis thaliana* were gained from the Flowering Interactive Database (http://www.phytosystems.ulg.ac.be/florid/, Bouché *et al.* 2016). Orthologs in switchgrass to these

known flowering time genes were identified using the software OrthoFinder (Emms and Kelly, 2019) with the default settings. All the known flowering time genes and their orthologs in switchgrass are shown in **Table S5,S6**. Genetic variants that were located in intergenic regions were associated with the nearby genes which are closest to the variants and with distances to the variants < 3.5 Kb.

Acknowledgments

This work was supported by the U.S. Department of Energy Great Lakes Bioenergy Research Center (BER DE-SC0018409 to SHS), the National Science Foundation (DGE-1828149 to SHS and KSA; IOS- 2107215, IOS-2218206, and MCB-2210431 to SHS), the National Natural Science Foundation of China (32370241 to PW) and the Scientific Research Foundation of and the Major Scientific Research Tasks from Kunpeng Institute of Modern Agriculture at Foshan (KIMAQD2022003 and KIMA-ZDKY2022004 to PW).

Author contributions

PW and SHS conceived and designed this study. MDC provided data and advice on study design and interpretation. PW and FM conducted the read mapping and SNP/indel calling analyses. PW conducted all the other analyses with help from CBA and KSA. PW and SHS wrote the manuscript with inputs from all authors. All authors read and approved the final manuscript.

Data availability

All the scripts used in this study are available on Github at: https://github.com/ShiuLab/Manuscript_Code/tree/master/2022_GP_in_Switchgrass.

Figure legends

Fig. 1 Correlation between the values of 20 traits and the prediction accuracy for these 20 traits. (**A**) Heatmap showing the correlation between the values of 20 traits. Gray font: 13 biochemical traits; black font: seven morphological traits. Color scale: rho of the spearman's rank correlation coefficient. (**B**,**C**) Prediction accuracy (r^2) for 20 traits in cross-validation (**B**) and test (**C**) subsets for models built using the bi-allelic single nucleotide polymorphisms (SNPs) called by mapping the genotyping-by-sequencing (GBS) data to v5 genome assembly. The models were

built on all individuals including tetraploids and octoploids. X-axis: 20 traits; y-axis: median r^2 between true and predicted trait values among 10 replicate runs. Blue (r^2_p): r^2 for models built with the population structure, which was defined as the first five principal components from the genetic variants; orange (r^2_g): r^2 of models built with genetic variants.

Fig. 2 Prediction accuracy differences between models based on v1 and v5 assemblies. (**A**) Differences in $r^2_{i,CV}$ between v1- and v5-based models. The column "All": models built using all the v1- and v5-based GBS bi-allelic SNPs; "Common": models built using the common GBS bi-allelic SNPs shared between v1 and v5 assemblies; "Assembly-specific": models built using GBS bi-allelic SNPs that are specific to v1 and v5 assemblies, separately. Values in the heatmap: $r^2_{i,CV}$ differences × 100; only the comparisons with statistical significance (*p* from Wilcoxon signed-rank test < 0.05) are indicated with colors; color scale: $-\log_{10}$ *p*-value; red and blue: significantly higher $r^2_{i,CV}$ was observed for the v1- and v5-based models, respectively. (**B**-**D**) Correlation between $r^2_{i,CV}$ of v1- (x-axis) and v5-based (y-axis) models. (**B**) Models built using all the v1- and v5-based GBS bi-allelic SNPs; (**C**) Models built using the common GBS bi-allelic SNPs shared between v1 and v5 assemblies; (**D**) Models built using GBS bi-allelic SNPs that are specific to v1 and v5 assemblies, separately.

Fig. 3 Better trait prediction in exome capture (EC)-based models than genotyping-by-sequencing (GBS)-based models. (**A**) Differences in $r^2_{i,CV}$ between EC- and GBS-based models. The column "Unbalanced": models built using all EC and GBS bi-allelic SNPs; "Balanced": models built using balanced EC (down-sampled to match the number of GBS variants) and GBS bi-allelic SNPs; "Genic": models built using balanced EC and GBS bi-allelic SNPs that were located within the gene bodies; "Intergenic": models built using balanced EC and GBS bi-allelic SNPs that were located within intergenic regions. Values in the heatmap: $r^2_{i,CV}$ differences × 100; only the comparisons with statistical significance (*p* from Wilcoxon signed-rank test < 0.05) are indicated with colors; color scale: $-\log_{10}$ *p*-value; red and blue: significantly higher $r^2_{i,CV}$ was observed for the EC- and GBS-based models, respectively. (**B**) Distribution of the GBS (blue) and EC (pink) bi-allelic SNPs in the genome. All the gene bodies were scaled to 3.5 Kb to make the locations of variants within the gene body comparable. The regions of 3.5 Kb up-stream and down-stream the gene bodies were also shown. (**C,D**) Comparison between the numbers of

genes harboring (**C**) or adjacent to (**D**) GBS bi-allelic SNPs (blue line) and the downsampled EC bi-allelic SNPs (pink line). z-score values of the GBS-based numbers in the distribution of EC-based numbers are shown in the figures.

Fig. 4 Performance of models built using bi-allelic indels and multi-allelic SNPs. (**A**) Differences in $r^2_{i,CV}$ between models built using GBS bi-allelic SNPs and models built using GBS bi-allelic indels (left two columns), and between GBS bi-allelic SNP-based models and GBS multi-allelic SNP-based models (right two columns). The "Unbalanced" and "balanced" columns: the bi-allelic SNPs were not and down-sampled to the same numbers of bi-allelic indels or multi-allelic SNPs, respectively. Values in the heatmap: $r^2_{i,CV}$ differences × 100; only the comparisons with statistical significance (*p* from Wilcoxon signed-rank test < 0.05) are indicated with colors; color scale: $-\log_{10}$ *p*-value; red: significantly higher $r^2_{i,CV}$ was observed for the bi-allelic SNP-based models; blue: significantly higher $r^2_{i,CV}$ was observed for models built using bi-allelic indels or multi-allelic SNPs. (**B**) Comparison between the numbers of genes harboring or adjacent to the down-sampled bi-allelic SNPs (pink line) and bi-allelic indels (blue line). (**C**) Comparison between the numbers of genes harboring or adjacent to the down-sampled bi-allelic SNPs (pink line) and multi-allelic SNPs (blue line).

Fig. 5 Improvement in prediction accuracy by integrating different types of GBS variants. (**A,B**) The $r^2_{i,CV}$ (**A**) and $r^2_{i,test}$ (**B**) for models built using GBS bi-allelic SNPs (dark blue), bi-allelic indels (blue), multi-allelic SNPs (dark green), multi-allelic indels (green), and all the GBS variants (yellow). Column with ** or ***: the integrating model had significantly improved prediction accuracy compared with all the models built using a single type of GBS variant; **: *p*-value from Wilcoxon signed-rank test < 0.01; ***: *p*-value < 0.001. Error bar: standard deviation.

Fig. 6 Models for octoploids had poorer prediction accuracy than models for tetraploids. (**A**) Differences in prediction accuracy between models for tetra- and octoploids when GBS and EC bi-allelic SNPs were used. Values in the heatmap: $r^2_{i,CV}$ differences × 100; only the comparisons with statistical significance (*p* from Wilcoxon signed-rank test < 0.05) are indicated with colors; color scale: $-\log_{10}$ *p*-value; red and blue: significantly higher $r^2_{i,CV}$ was observed for the tetraploid and octoploid model, respectively. (**B**) Read depth (left panel) and the number of

missing data (middle and right panels) of the GBS data. (**C**) Read depth (left panel) and the number of missing data (middle and right panels) of the EC data. (**D,E**) Comparison of the $r^2_{p,CV}$ (**D**) and $r^2_{g,CV}$ (**E**) between models for tetraploids (x-axis) and octoploids (y-axis) when GBS (left panel) or EC (right panel) bi-allelic SNPs were used.

Fig. 7 Analysis of important variants for genomic prediction models. (**A**) Correlation between the numbers of variants with absolute coefficient above 99th (left panel) or 95th (right panel) percentile (x-axis) and the corresponding numbers of switchgrass genes orthologous to known flowering time genes (y-axis). Blue: GBS-based models; pink: EC-based models. Dashed line: the diagonal line. (**B**) The number of identified flowering time ortholog genes by models (red bars) and the corresponding odds ratio (green bars) when 99th (left panel) and 95th (right panel) percentile was used as cut-off thresholds. *: *p*-value from Fisher's exact test < 0.05; **: *p*-value < 0.01; ***: *p*-value < 0.001. (**C-J**) Anthesis dates of individuals with different alleles. The type of variants which were used to build the models were listed above the figures. The absolute coefficient rank of the variant was listed in the parenthesis on the right of the variant. *EFS*: *EARLY FLOWERING IN SHORT DAYS*; *VIP3*: *VERNALIZATION INDEPENDENCE 3*. Yellow, red, and blue bars in (**C-J**): individuals with homologous alternative (-1), heterozygous (0) and homologous reference (1) alleles, respectively. *p*-values were from Wilcoxon signed-rank test. In (**K**), six types of alleles for the locus Chr08K_52608078 were indicated on the left of the table; values in the table indicate the numbers of the corresponding alleles; green: tetraploids; prasinous: octoploids; error bar: standard error; column with no error bar: there was only one individual containing the corresponding allele composition. Only columns with > 1 individual are indicated with letters indicating the statistical significance.

Supplemental figure legends

Fig. S1 Classification of different types of genetic variants. (**A**) Frequency of different types of variants called by mapping the GBS data to v1 assembly. (**B**) Frequency of different types of variants called by mapping the GBS data to v5 assembly. (**C**) Frequency of different types of variants called by mapping the EC data to v5 assembly.

Fig. S2 Prediction accuracy of models built using the population structure on the cross-validation sets. (**A**) $r^2_{p,CV}$ for models built on both tetraploids and octoploids using different types of variants. Table above the heatmap shows different factors the variants used may belong to. For example, the 1st and 4th column show $r^2_{p,CV}$ of models built using bi-allelic SNPs called by mapping the GBS data to the v1 and v5 assembly, respectively; the 12th and 13th column show $r^2_{p,CV}$ of models built using balanced subset (down-sampled to the same number of GBS bi-allelic SNPs) and all the v5-based EC bi-allelic SNPs, respectively. (**B**) $r^2_{p,CV}$ for models built for different ploidy levels or subpopulations using GBS or EC bi-allelic SNPs. Color scale in the heatmap: median $r^2_{p,CV}$ among 10 replicate runs.

Fig. S3 Prediction accuracy of models built using genetic variants on the cross-validation sets. (**A**) $r^2_{g,CV}$ for models built on both tetraploids and octoploids using different types of variants. Table above the heatmap shows different factors the variants used may belong to. (**B**) $r^2_{g,CV}$ for models built for different ploidy levels or subpopulations using GBS or EC bi-allelic SNPs. Color scale in the heatmap: median $r^2_{g,CV}$ among 10 replicate runs.

Fig. S4 Improvement of r^2 of models built using different genetic variants on the cross-validation sets. (**A**) $r^2_{i,CV}$ for models built on both tetraploids and octoploids using different types of variants. Table above the heatmap shows different factors the variants used may belong to. (**B**) $r^2_{i,CV}$ for models built for different ploidy levels or subpopulations using GBS or EC bi-allelic SNPs. Color scale in the heatmap: $r^2_{i,CV} \times 100$.

Fig. S5 Prediction accuracy of models built using the population structure on the test sets. (**A**) $r^2_{p,test}$ for models built on both tetraploids and octoploids using different types of variants. Table above the heatmap shows different factors the variants used may belong to. (**B**) $r^2_{p,test}$ for models built for different ploidy levels or subpopulations using GBS or EC bi-allelic SNPs. Color scale in the heatmap: median $r^2_{p,test}$ among 10 replicate runs.

Fig. S6 Prediction accuracy of models built using genetic variants on the test sets. (**A**) $r^2_{g,test}$ for models built on both tetraploids and octoploids using different types of variants. Table above the heatmap shows different factors the variants used may belong to. (**B**) $r^2_{g,test}$ for models

built for different ploidy levels or subpopulations using GBS or EC bi-allelic SNPs. Color scale in the heatmap: median $r^2_{g,test}$ among 10 replicate runs.

Fig. S7 Improvement of r^2 of models built using different genetic variants on the test sets. (**A**) $r^2_{i,test}$ for models built on both tetraploids and octoploids using different types of variants. Table above the heatmap shows different factors the variants used may belong to. (**B**) $r^2_{i,test}$ for models built for different ploidy levels or subpopulations using GBS or EC bi-allelic SNPs. Color scale in the heatmap: $r^2_{i,test} \times 100$.

Fig. S8 Prediction accuracy of models built using different GBS variants and models integrating all GBS variants. Color and value in (**A**-**D**) indicate $r^2_{p,cv}$, $r^2_{p,test}$, $r^2_{g,cv}$, and $r^2_{g,test}$, respectively.

Fig. S9 Prediction accuracy of models built for four subpopulations. (**A**-**C**) The $r^2_{p,CV}$ (**A**), $r^2_{g,CV}$ (**B**) and $r^2_{i,CV}$ (**C**) for models using EC bi-allelic SNPs, built for Lowland north (4X) (dark blue), Upland north (4X) (blue), Upland east (8X) (green) and Upland west (8X) (yellow) subpopulations. The color background indicates that the model built for the corresponding subpopulation had the highest prediction accuracy for the trait in question. Error bar: standard deviation.

Fig. S10 Flowering time orthologous genes associated with variants that had absolute coefficients above the 99th percentile in eight models. The type of genetic variants used to build the models were listed above the heatmap. Yellow, green and blue: variants with absolute coefficients >99th, >95th and ≤95th percentiles, respectively.

References

Aalborg T, Nielsen KL (2024) To be or not to be tetraploid—the impact of marker ploidy on genomic prediction and GWAS of potato. Front Plant Sci **15**: 1386837

Akdemir D, Beavis W, Fritsche-Neto R, Singh AK, Isidro-Sánchez J (2019) Multi-objective optimized genomic breeding strategies for sustainable food improvement. Heredity **122**: 672–683

Alemu A, Åstrand J, Montesinos-López OA, Isidro Y Sánchez J, Fernández-Gónzalez J, Tadesse W, Vetukuri RR, Carlsson AS, Ceplitis A, Crossa J, et al (2024) Genomic selection in plant breeding: key factors shaping two decades of progress. Mol Plant **17**: 552–578

Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KM, Cao J, Chae E, Dezwaan TM, Ding W, et al (2016) 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. Cell **166**: 481–491

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol **215**: 403–410

Auwera GAV de, O'Connor BD (2020) Genomics in the cloud: using Docker, GATK, and WDL in Terra, First edition. O'Reilly, Beijing Boston Farnham Sebastopol Tokyo

Ayalew H, Anderson JD, Krom N, Tang Y, Butler TJ, Rawat N, Tiwari V, Ma X-F (2022) Genotyping-by-sequencing and genomic selection applications in hexaploid triticale. G3 Bethesda **12**: jkab413

Azodi CB, Bolger E, McCarren A, Roantree M, de los Campos G, Shiu S-H (2019) Benchmarking parametric and machine learning models for genomic prediction of complex traits. G3 Bethesda **9**: 3691–3702

Azodi CB, Pardo J, VanBuren R, de los Campos G, Shiu S-H (2020) Transcriptome-based prediction of complex traits in maize. Plant Cell **32**: 139–151

Batista LG, Mello VH, Souza AP, Margarido GRA (2022) Genomic prediction with allele dosage information in highly polyploid species. Theor Appl Genet **135**: 723–739

Benevenuto J, Ferrão LFV, Amadeu RR, Munoz P (2019) How can a high-quality genome assembly help plant breeders? GigaScience **8**: giz068

Biová J, Kaňovská I, Chan YO, Immadi MS, Joshi T, Bilyeu K, Škrabišová M (2024) Natural and artificial selection of multiple alleles revealed through genomic analyses. Front Genet **14**: 1320652

Boatwright JL, Sapkota S, Jin H, Schnable JC, Brenton Z, Boyles R, Kresovich S (2022) Sorghum association panel whole‐genome sequencing establishes cornerstone resource for dissecting genomic diversity. Plant J **111**: 888–904

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**: 2114–2120

Bouché F, Lobet G, Tocquin P, Périlleux C (2016) FLOR-ID: an interactive database of flowering-time gene networks in *Arabidopsis thaliana*. Nucleic Acids Res **44**: D1167–D1171

Chen J, Wang Z, Tan K, Huang W, Shi J, Li T, Hu J, Wang K, Wang C, Xin B, et al (2023) A complete telomere-to-telomere assembly of the maize genome. Nat Genet **55**: 1221–1231

Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de los Campos G, Burgueño J, González-Camacho JM, Pérez-Elizalde S, Beyene Y, et al (2017) Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci **22**: 961–975

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al (2021) Twelve years of SAMtools and BCFtools. GigaScience **10**: giab008

Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol **20**: 238

Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome **4**: 250–255

Estopa RA, Paludeto JGZ, Müller BSF, De Oliveira RA, Azevedo CF, De Resende MDV, Tambarussi EV, Grattapaglia D (2023) Genomic prediction of growth and wood quality traits in *Eucalyptus benthamii* using different genomic models and variable SNP genotyping density. New For **54**: 343–362

Evans J, Crisovan E, Barry K, Daum C, Jenkins J, Kunde‐Ramamoorthy G, Nandety A, Ngan CY, Vaillancourt B, Wei C, et al (2015) Diversity and population structure of northern switchgrass as revealed through exome capture sequencing. Plant J **84**: 800–815

Guo Z, Tucker DM, Basten CJ, Gandhi H, Ersoz E, Guo B, Xu Z, Wang D, Gay G (2014) The impact of population structure on genomic prediction in stratified populations. Theor Appl Genet **127**: 749–762

Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, McCarroll SA (2015) Large multiallelic copy number variations in humans. Nat Genet **47**: 296–303

He H, Leng Y, Cao X, Zhu Y, Li X, Yuan Q, Zhang B, He W, Wei H, Liu X, et al (2024) The pan-tandem repeat map highlights multiallelic variants underlying gene expression and agronomic traits in rice. Nat Commun **15**: 7291

Isidro J, Jannink J-L, Akdemir D, Poland J, Heslot N, Sorrells ME (2015) Training set optimization under population structure in genomic selection. Theor Appl Genet **128**: 145–158

Jiang Y, Chen S, Wang X, Liu M, Iacono WG, Hewitt JK, Hokanson JE, Krauter K, Laakso M, Li KW, et al (2020) Association analysis and meta-analysis of multi-allelic variants for large-scale sequence data. Genes **11**: 586

Kang M, Wu H, Liu H, Liu W, Zhu M, Han Y, Liu W, Chen C, Song Y, Tan L, et al (2023) The pan-genome and local adaptation of *Arabidopsis thaliana*. Nat Commun **14**: 6259

Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D, Nordborg M (2007) Recombination and linkage disequilibrium in *Arabidopsis thaliana*. Nat Genet **39**: 1151–1155

Kriaridou C, Tsakiridou S, Houston RD, Robledo D (2020) Genomic prediction using low density marker panels in aquaculture: performance across species, traits, and genotyping platforms. Front Genet **11**: 124

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open software for comparing large genomes. Genome Biol **5**: R12

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997. doi: https://doi.org/10.48550/arXiv.1303.3997

Lian Q, Huettel B, Walkemeyer B, Mayjonade B, Lopez-Roques C, Gil L, Roux F, Schneeberger K, Mercier R (2024) A pan-genome of 69 *Arabidopsis thaliana* accessions reveals a conserved genome structure throughout the global species range. Nat Genet **56**: 982–991

Lipka AE, Lu F, Cherney JH, Buckler ES, Casler MD, Costich DE (2014) Accelerating the switchgrass (*Panicum virgatum* L.) breeding cycle using genomic selection approaches. PLoS ONE **9**: e112227

Liu Z, Wang N, Su Y, Long Q, Peng Y, Shangguan L, Zhang F, Cao S, Wang X, Ge M, et al (2024) Grapevine pangenome facilitates trait genetics and genomic breeding. Nat Genet **56**: 2804–2814

Lorenz AJ, Smith KP (2015) Adding genetically distant individuals to training populations reduces genomic prediction accuracy in barley. Crop Sci **55**: 2657–2667

Lovell JT, MacQueen AH, Mamidi S, Bonnette J, Jenkins J, Napier JD, Sreedasyam A, Healey A, Session A, Shu S, et al (2021) Genomic mechanisms of climate adaptation in polyploid bioenergy switchgrass. Nature **590**: 438–444

Lu D, Liu C, Ji W, Xia R, Li S, Liu Y, Liu N, Liu Y, Deng XW, Li B (2024) Nanopore ultra-long sequencing and adaptive sampling spur plant complete telomere-to-telomere genome assembly. Mol Plant **17**: 1773–1786

Lu F, Lipka AE, Glaubitz J, Elshire R, Cherney JH, Casler MD, Buckler ES, Costich DE (2013) Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. PLoS Genet **9**: e1003215

McLaughlin SB, Adams Kszos L (2005) Development of switchgrass (*Panicum virgatum*) as a bioenergy feedstock in the United States. Biomass Bioenergy **28**: 515–535

Minamikawa MF, Kunihisa M, Moriya S, Shimizu T, Inamori M, Iwata H (2024) Genomic prediction and genome-wide association study using combined genotypic data from different genotyping systems: application to apple fruit quality traits. Hortic Res **11**: uhae131

Misra G, Badoni S, Anacleto R, Graner A, Alexandrov N, Sreenivasulu N (2017) Whole genome sequencing-based association study to unravel genetic architecture of cooked grain width and length traits in rice. Sci Rep **7**: 12478

Norman A, Taylor J, Edwards J, Kuchel H (2018) Optimising genomic selection in wheat: effect of marker density, population size and population structure on prediction accuracy. G3 Bethesda **8**: 2889–2899

Rosyara UR, De Jong WS, Douches DS, Endelman JB (2016) Software for genome‐wide association studies in autopolyploids and Its application to potato. Plant Genome **9**: plantgenome2015.08.0073

Scheben A, Batley J, Edwards D (2017) Genotyping‐by‐sequencing approaches to characterize crop genomes: choosing the right tool for the right application. Plant Biotechnol J **15**: 149–161

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet **78**: 629–644

Song B, Mott R, Gan X (2018) Recovery of novel association loci in *Arabidopsis thaliana* and *Drosophila melanogaster* through leveraging INDELs association and integrated burden test. PLOS Genet **14**: e1007699

Soppe WJJ, Bentsink L, Koornneef M (1999) The early-flowering mutant *efs* is involved in the autonomous promotion pathway of *Arabidopsis thaliana*. Development **126**: 4763–4770

Tade B, Melesse A (2024) A review on the application of genomic selection in the improvement of dairy cattle productivity. Ecol Genet Genomics **31**: 100257

Veerkamp RF, Bouwman AC, Schrooten C, Calus MPL (2016) Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein–Friesian cattle. Genet Sel Evol **48**: 95

Wang H, Bernardo A, St. Amand P, Bai G, Bowden RL, Guttieri MJ, Jordan KW (2023a) Skim exome capture genotyping in wheat. Plant Genome **16**: e20381

Wang K, Abid MA, Rasheed A, Crossa J, Hearne S, Li H (2023b) DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants. Mol Plant **16**: 279–293

Wang P, Lehti-Shiu MD, Lotreck S, Segura Abá K, Krysan PJ, Shiu S-H (2024) Prediction of plant complex traits via integration of multi-omics data. Nat Commun **15**: 6856

Wilson S, Zheng C, Maliepaard C, Mulder HA, Visser RGF, Van Der Burgt A, Van Eeuwijk F (2021) Understanding the effectiveness of genomic prediction in tetraploid potato. Front Plant Sci **12**: 672417

Wu P-Y, Ou J-H, Liao C-T (2023) Sample size determination for training set optimization in genomic prediction. Theor Appl Genet **136**: 57

Yadav S, Ross EM, Wei X, Liu S, Nguyen LT, Powell O, Hickey LT, Deomano E, Atkin F, Voss‐Fels KP, et al (2024) Use of continuous genotypes for genomic prediction in sugarcane. Plant Genome **17**: e20417

Yan H, Sun M, Zhang Z, Jin Y, Zhang A, Lin C, Wu B, He M, Xu B, Wang J, et al (2023) Pangenomic analysis identifies structural variation associated with heat tolerance in pearl millet. Nat Genet **55**: 507–518

Zhan X, Hu Y, Li B, Abecasis GR, Liu DJ (2016) RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. Bioinformatics **32**: 1423–1426

Zhang H, Ransom C, Ludwig P, Van Nocker S (2003) Genetic analysis of early flowering mutants in Arabidopsis defines a class of pleiotropic developmental regulator required for expression of the flowering-time switch Flowering Locus C. Genetics **164**: 347–358

Zhang H, Yin L, Wang M, Yuan X, Liu X (2019) Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. Front Genet **10**: 189

Zheng Z, Liu S, Sidorenko J, Wang Y, Lin T, Yengo L, Turley P, Ani A, Wang R, Nolte IM, et al (2024) Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex traits within and between ancestries. Nat Genet **56**: 767–777

Zhou Y, Zhang Z, Bao Z, Li H, Lyu Y, Zan Y, Wu Y, Cheng L, Fang Y, Wu K, et al (2022) Graph pangenome captures missing heritability and empowers tomato breeding. Nature **606**: 527–534

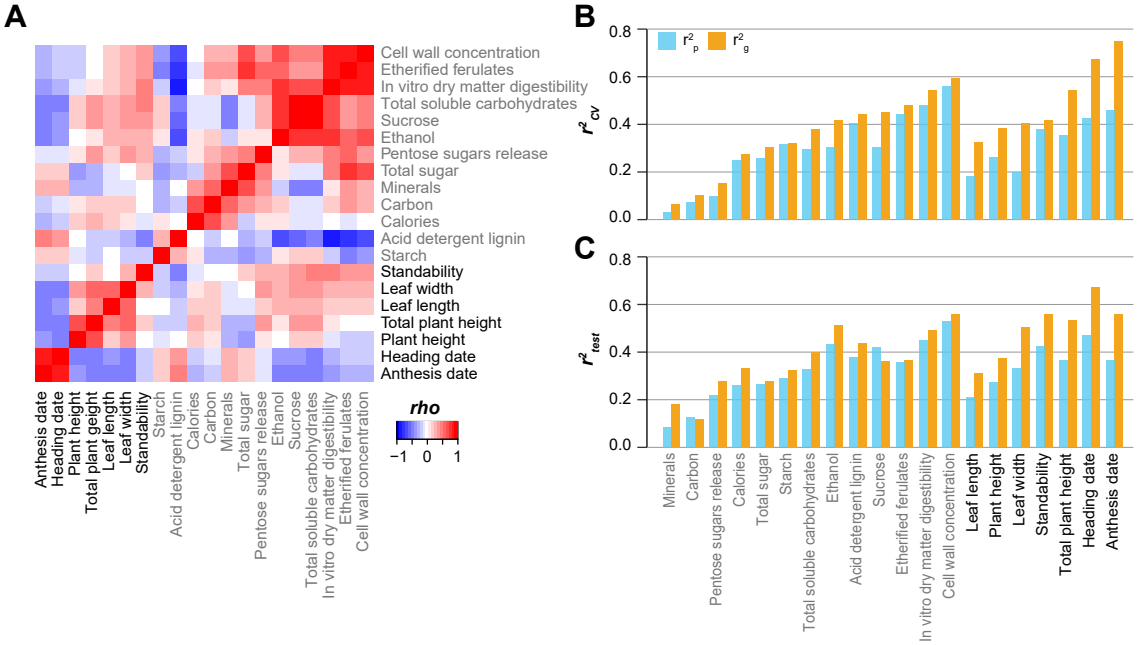

Figure 1. Correlation between the values of 20 traits and the prediction accuracy for these 20 traits. **A**) Heatmap showing the correlation between the values of 20 traits. Gray font: 13 biochemical traits; black font: seven morphological traits. Color scale: rho of the spearman's rank correlation coefficient. **B,C**) Prediction accuracy (r^2) for 20 traits in cross-validation (**B**) and test (**C**) subsets for models built using the bi-allelic single nucleotide polymorphisms (SNPs) called by mapping the genotyping-by-sequencing (GBS) data to v5 genome assembly. The models were built on all individuals including tetraploids and octoploids. X-axis: 20 traits; y-axis: median r^2 between true and predicted trait values among 10 replicate runs. Blue (r^2_p): r^2 for models built with the population structure, which was defined as the first five principal components from the genetic variants; orange (r^2_g): r^2 of models built with genetic variants.

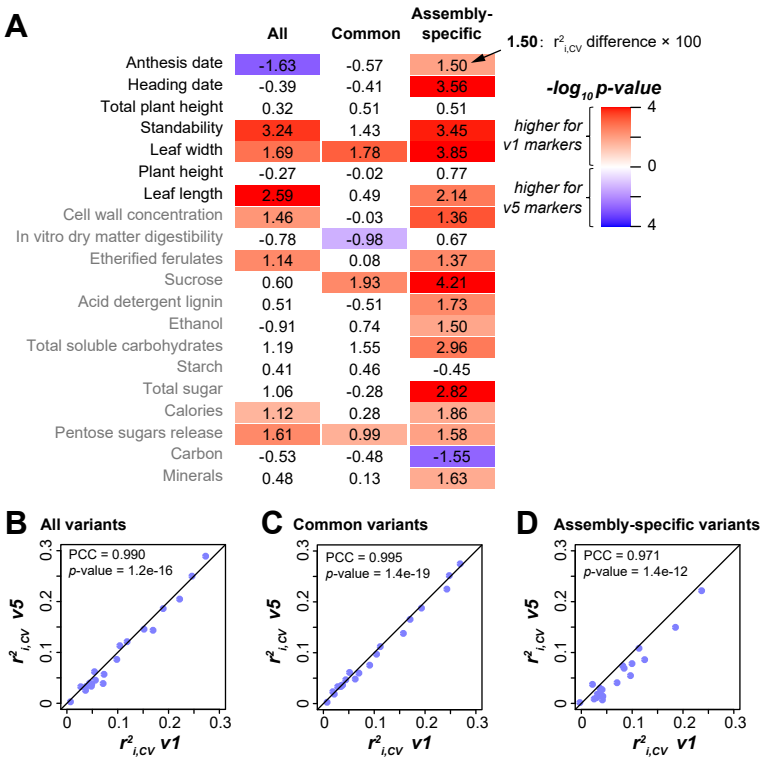

Figure 2. Prediction accuracy differences between models based on v1 and v5 assemblies. **A**) Differences in $r^2_{i,CV}$ between v1- and v5-based models. The column "All": models built using all the v1- and v5-based GBS bi-allelic SNPs; "Common": models built using the common GBS bi-allelic SNPs shared between v1 and v5 assemblies; "Assembly-specific": models built using GBS bi-allelic SNPs that are specific to v1 and v5 assemblies, separately. Values in the heatmap: $r^2_{i,CV}$ differences × 100; only the comparisons with statistical significance (*p* from Wilcoxon signed-rank test < 0.05) are indicated with colors; color scale: $-\log_{10}$ *p*-value; red and blue: significantly higher $r^2_{i,CV}$ was observed for the v1- and v5-based models, respectively. **B-D**) Correlation between $r^2_{i,CV}$ of v1- (x-axis) and v5-based (y-axis) models. **B**) Models built using all the v1- and v5-based GBS bi-allelic SNPs; **C**) Models built using the common GBS bi-allelic SNPs shared between v1 and v5 assemblies; **D**) Models built using GBS bi-allelic SNPs that are specific to v1 and v5 assemblies, separately.

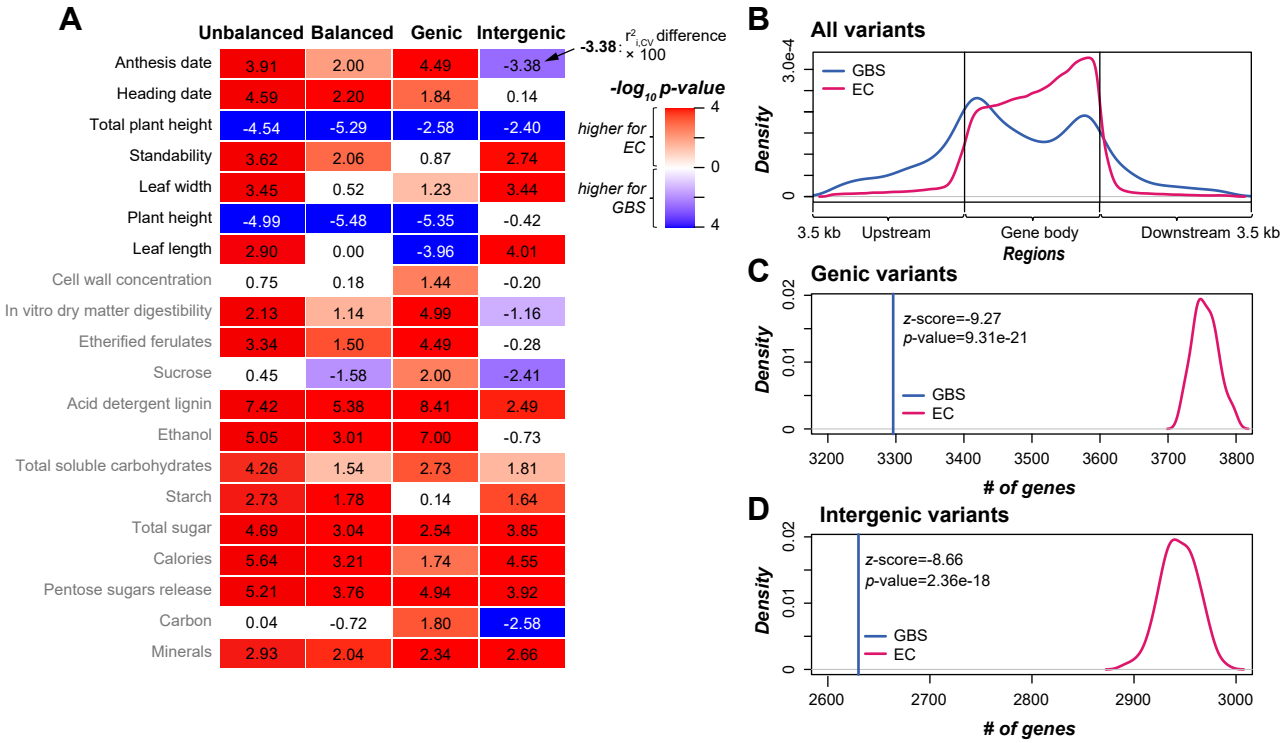

Figure 3. Better trait prediction in exome capture (EC)-based models than genotyping-by-sequencing (GBS)-based models. **A**) Differences in $r^2_{i,CV}$ between EC- and GBS-based models. The column "Unbalanced": models built using all EC and GBS bi-allelic SNPs; "Balanced": models built using balanced EC (down-sampled to match the number of GBS variants) and GBS bi-allelic SNPs; "Genic": models built using balanced EC and GBS bi-allelic SNPs that were located within the gene bodies; "Intergenic": models built using balanced EC and GBS bi-allelic SNPs that were located within intergenic regions. Values in the heatmap: $r^2_{i,CV}$ differences × 100; only the comparisons with statistical significance (*p* from Wilcoxon signed-rank test < 0.05) are indicated with colors; color scale: $-\log_{10}$ *p*-value; red and blue: significantly higher $r^2_{i,CV}$ was observed for the EC- and GBS-based models, respectively. **B**) Distribution of the GBS (blue) and EC (pink) bi-allelic SNPs in the genome. All the gene bodies were scaled to 3.5 Kb to make the locations of variants within the gene body comparable. The regions of 3.5 Kb up-stream and down-stream the gene bodies were also shown. **C**,**D**) Comparison between the numbers of genes harboring (**C**) or adjacent to (**D**) GBS bi-allelic SNPs (blue line) and the downsampled EC bi-allelic SNPs (pink line). *z*-score values of the GBS-based numbers in the distribution of EC-based numbers are shown in the figures.

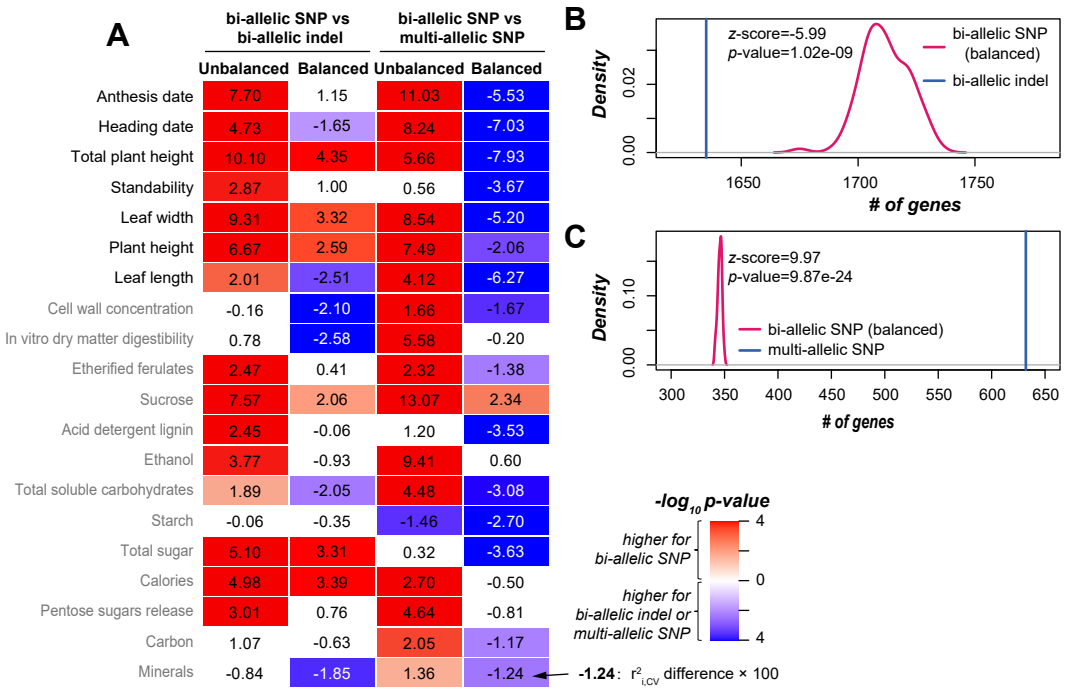

Figure 4. Performance of models built using bi-allelic indels and multi-allelic SNPs. **A**) Differences in $r^2_{i,CV}$ between models built using GBS bi-allelic SNPs and models built using GBS bi-allelic indels (left two columns), and between GBS bi-allelic SNP-based models and GBS multi-allelic SNP-based models (right two columns). The "Unbalanced" and "balanced" columns: the bi-allelic SNPs were not and down-sampled to the same numbers of bi-allelic indels or multi-allelic SNPs, respectively. Values in the heatmap: $r^2_{i,CV}$ differences × 100; only the comparisons with statistical significance (*p* from Wilcoxon signed-rank test < 0.05) are indicated with colors; color scale: $-\log_{10}$ *p*-value; red: significantly higher $r^2_{i,CV}$ was observed for the bi-allelic SNP-based models; blue: significantly higher $r^2_{i,CV}$ was observed for models built using bi-allelic indels or multi-allelic SNPs. **B**) Comparison between the numbers of genes harboring or adjacent to the down-sampled bi-allelic SNPs (pink line) and bi-allelic indels (blue line). **C**) Comparison between the numbers of genes harboring or adjacent to the down-sampled bi-allelic SNPs (pink line) and multi-allelic SNPs (blue line).

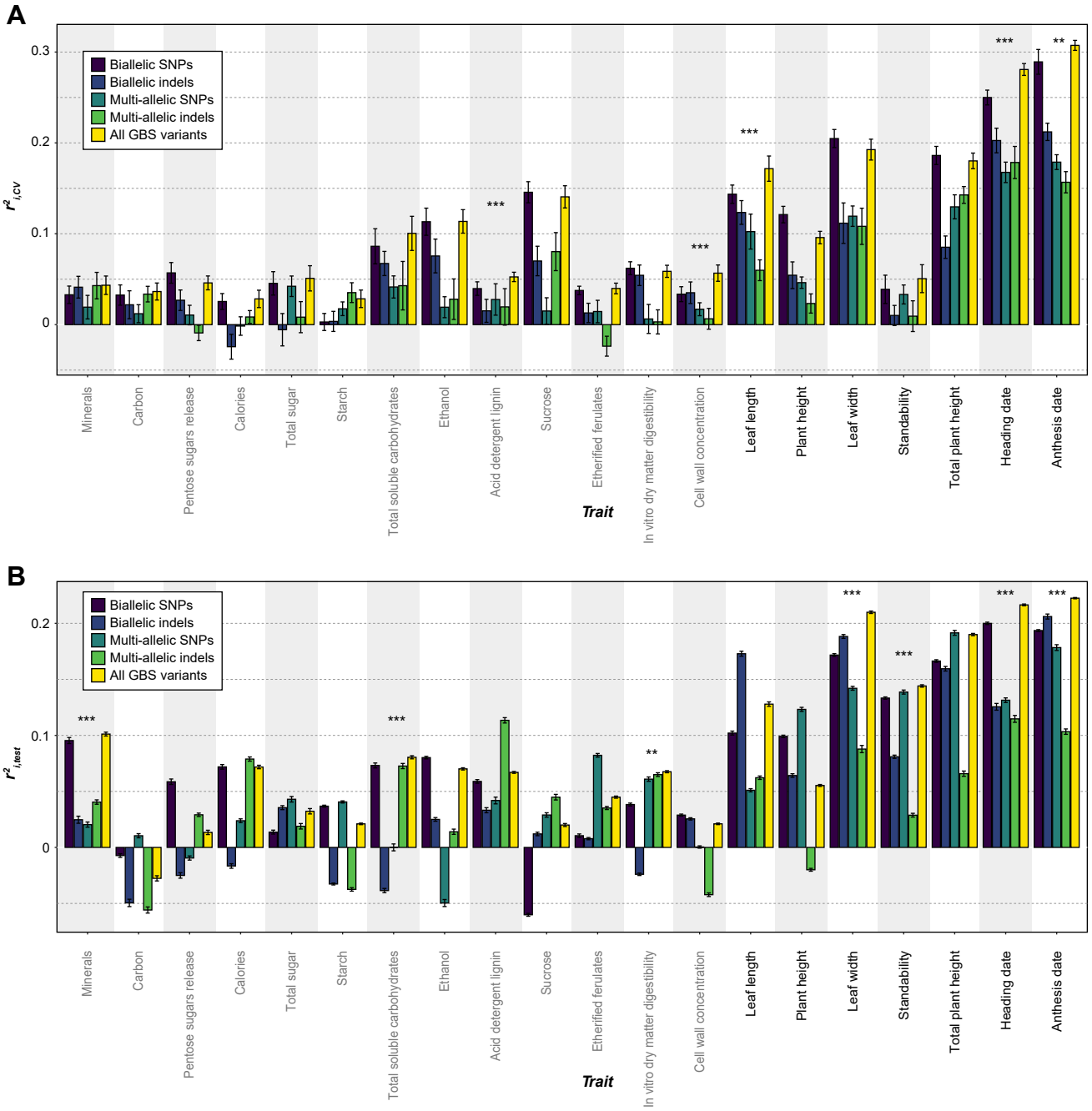

Figure 5. Improvement in prediction accuracy by integrating different types of GBS variants. **A,B**) The $r^2_{i,CV}$ (**A**) and $r^2_{i,test}$ (**B**) for models built using GBS bi-allelic SNPs (dark blue), bi-allelic indels (blue), multi-allelic SNPs (dark green), multi-allelic indels (green), and all the GBS variants (yellow). Column with ** or ***: the integrating model had significantly improved prediction accuracy compared with all the models built using a single type of GBS variant; **: *p*-value from Wilcoxon signed-rank test < 0.01; ***: *p*-value < 0.001. Error bar: standard deviation.

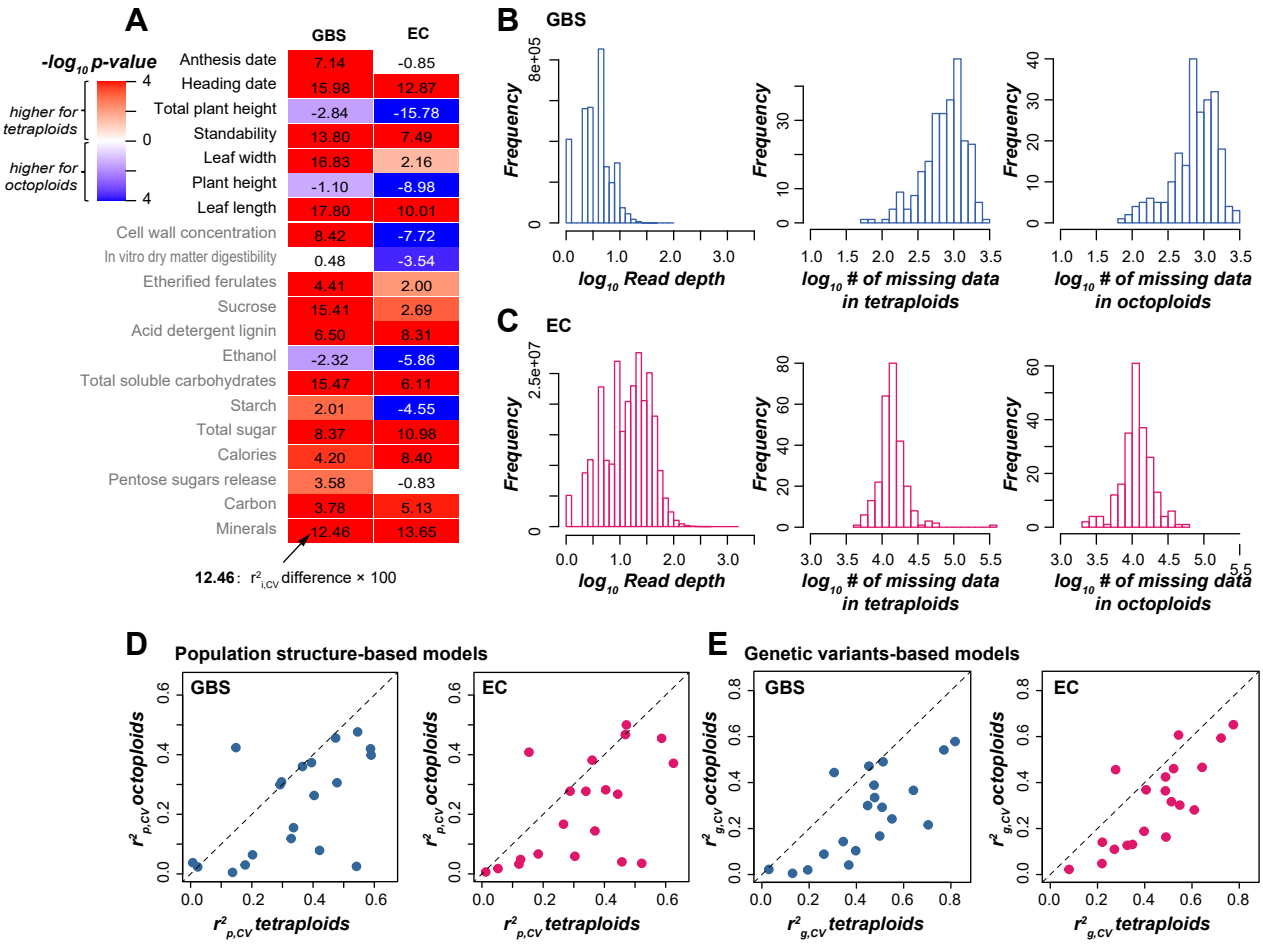

Figure 6. Models for octoploids had poorer prediction accuracy than models for tetraploids. **A**) Differences in prediction accuracy between models for tetra- and octoploids when GBS and EC bi-allelic SNPs were used. Values in the heatmap: $r^2_{i,CV}$ differences × 100; only the comparisons with statistical significance (p from Wilcoxon signed-rank test < 0.05) are indicated with colors; color scale: -$\log_{10}$ p-value; red and blue: significantly higher $r^2_{i,CV}$ was observed for the tetraploid and octoploid model, respectively. **B**) Read depth (left panel) and the number of missing data (middle and right panels) of the GBS data. **C**) Read depth (left panel) and the number of missing data (middle and right panels) of the EC data. **D,E**) Comparison of the $r^2_{p,CV}$ (**D**) and $r^2_{g,CV}$ (**E**) between models for tetraploids (x-axis) and octoploids (y-axis) when GBS (left panel) or EC (right panel) bi-allelic SNPs were used.

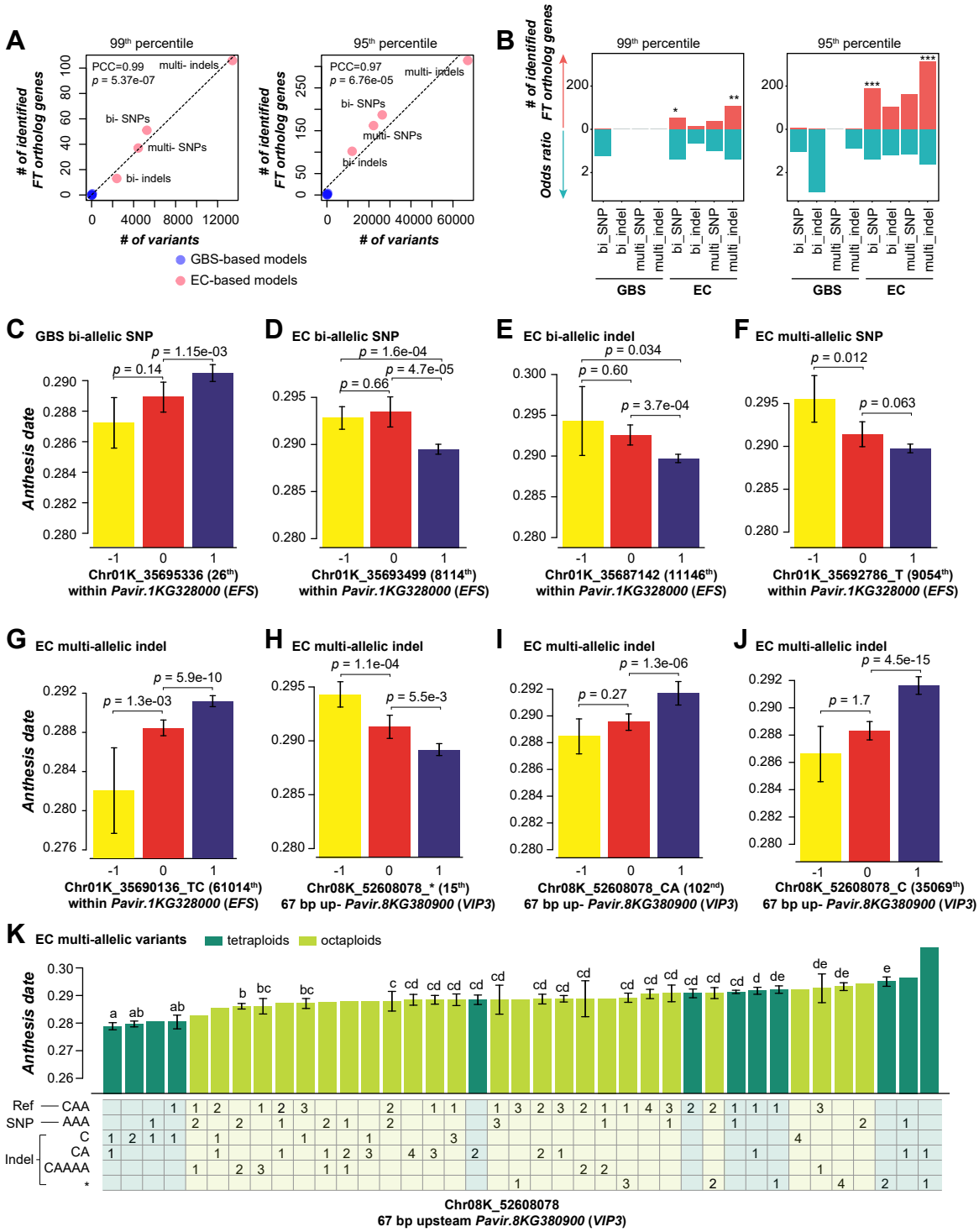

Figure 7. Analysis of important variants for genomic prediction models. **A**) Correlation between the numbers of variants with absolute coefficient above 99th (left panel) or 95th (right panel) percentile (x-axis) and the corresponding numbers of switchgrass genes orthologous to known flowering time genes (y-axis). Blue: GBS-based models; pink: EC-based models. Dashed line: the diagonal line. **B**) The number of identified flowering time ortholog genes by models (red bars) and the corresponding odds ratio (green bars) when 99th (left panel) and 95th (right panel) percentile was used as cut-off thresholds. *: *p*-value from Fisher's exact test < 0.05; **: *p*-value < 0.01; ***: *p*-value < 0.001. **C-H**) Anthesis dates of individuals with different alleles. The type of variants which were used to build the models were listed above the figures. The absolute coefficient rank of the variant was listed in the parenthesis on the right of the variant. *EFS*: *EARLY FLOWERING IN SHORT DAYS*; *VIP3*: *VERNALIZATION INDEPENDENCE 3*. Yellow, red, and blue bars in (**C-J**): individuals with homologous alternative (-1), heterozygous (0) and homologous reference (1) alleles, respectively. *p*-values were from Wilcoxon signed-rank test. In (**K**), six types of alleles for the locus Chr08K_52608078 were indicated on the left of the table; values in the table indicate the numbers of the corresponding alleles; green: tetraploids; prasinous: octoploids; error bar: standard error; column with no error bar: there was only one individual containing the corresponding allele composition. Only columns with > 1 individual are indicated with letters indicating the statistical significance.

Supporting Information

Article title: Optimizing genomic prediction for complex traits via investigating multiple factors in switchgrass

Authors: Peipei Wang, Fanrui Meng, Christina B Del Azodi, Kenia Segura Abá, Michael D. Casler, and Shin-Han Shiu

The following Supporting Information is available for this article:

Fig. S1 Classification of different types of genetic variants

Fig. S2 Prediction accuracy of models built using the population structure on the cross-validation sets

Fig. S3 Prediction accuracy of models built using genetic variants on the cross-validation sets

Fig. S4 Improvement of r^2 of models built using different genetic variants on the cross-validation sets

Fig. S5 Prediction accuracy of models built using the population structure on the test sets

Fig. S6 Prediction accuracy of models built using genetic variants on the test sets

Fig. S7 Improvement of r^2 of models built using different genetic variants on the test sets

Fig. S8 Prediction accuracy of models built using different GBS variants and models integrating all GBS variants

Fig. S9 Prediction accuracy of models built for four subpopulations

Fig. S10 Flowering time orthologous genes associated with variants that had absolute coefficients above the 99th percentile in eight models

Table S1. Genomic sequencing sample information (see separate file)

Table S2. 20 trait values for 486 individuals (see separate file)

Table S3. Summary of SNP/indel information detected using Genotyping-by-Sequencing (GBS) or exome capture sequencing (see separate file)

Table S4. Association between bi-allelic SNPs between two assemblies (see separate file)

Table S5. Benchmark flowering time genes in Arabidopsis and maize, and their orthologs in Switchgrass (see separate file)

Table S6. Ortholog groups for benchmark flowering time genes in three species (see separate file)

Table S7. Important variants above 95th (black) and 99th (red) percentiles for models built using GBS biallelic SNPs (see separate file)

Table S8. Important variants above 95th (black) and 99th (red) percentiles for models built using GBS biallelic indels (see separate file)

Table S9. Important variants above 95th (black) and 99th (red) percentiles for models built using GBS multiallelic SNPs (see separate file)

Table S10. Important variants above 95th (black) and 99th (red) percentiles for models built using GBS multiallelic indels (see separate file)

Table S11. Important variants above 95th (black) and 99th (red) percentiles for models built using exome biallelic SNPs (see separate file)

Table S12. Important variants above 95th (black) and 99th (red) percentiles for models built using exome biallelic indels (see separate file)

Table S13. Important variants above 95th (black) and 99th (red) percentiles for models built using exome multiallelic SNPs

Table S14. Important variants above 95th (black) and 99th (red) percentiles for models built using exome multiallelic indels

Table S15. Identified important variants that are associated with orthologs of known flowering time genes

Fig. S1 Classification of different types of genetic variants. (**A**) Frequency of different types of variants called by mapping the GBS data to v1 assembly. (**B**) Frequency of different types of variants called by mapping the GBS data to v5 assembly. (**C**) Frequency of different types of variants called by mapping the EC data to v5 assembly.

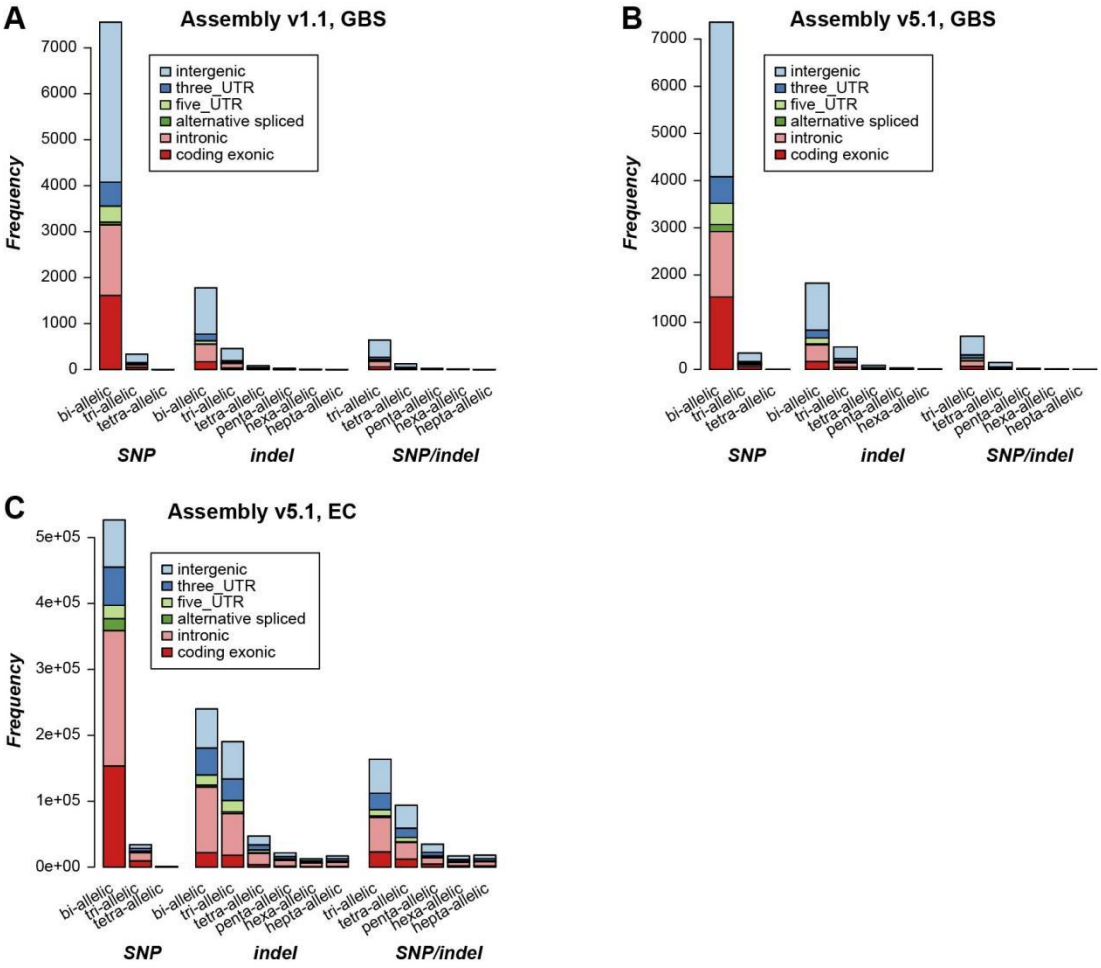

Fig. S2 Prediction accuracy of models built using the population structure on the cross-validation sets. (**A**) $r^2_{p,CV}$ for models built on both tetraploids and octoploids using different types of variants. Table above the heatmap shows different factors the variants used may belong to. For example, the 1st and 4th column show $r^2_{p,CV}$ of models built using bi-allelic SNPs called by mapping the GBS data to the v1 and v5 assembly, respectively; the 12th and 13th column show $r^2_{p,CV}$ of models built using balanced subset (down-sampled to the same number of GBS bi-allelic SNPs) and all the v5-based exome capture bi-allelic SNPs, respectively. (**B**) $r^2_{p,CV}$ for models built for different ploidy levels or subpopulations using GBS or exome capture bi-allelic SNPs. Color scale in the heatmap: median $r^2_{p,CV}$ among 10 replicate runs.

A

	GBS											EC					GBS			
	bi-allelic						bi-allelic		multi-	bi-	multi-	bi-allelic			multi-allelic		bi-allelic			
	SNP						indel		SNP		indel	SNP		indel	SNP	indel	SNP			
	v1			v5			v5					v5					v5			
	all	common	v1-specific	all	common	v5-specific	all	balanced for bi-allelic indels	all	balanced for multi-allelic SNPs	all	balanced for GBS bi-allelic SNPs	all	all	all	all	genic	intergenic	subset genic	subset intergenic
Anthesis date	0.48	0.47	0.45	0.46	0.47	0.46	0.46	0.45	0.44	0.42	0.46	0.45	0.45	0.48	0.51	0.5	0.47	0.44	0.46	0.48
Heading date	0.43	0.42	0.42	0.42	0.42	0.42	0.43	0.42	0.38	0.4	0.39	0.39	0.39	0.4	0.42	0.41	0.42	0.42	0.39	0.4
Total plant height	0.35	0.35	0.34	0.36	0.35	0.37	0.38	0.36	0.33	0.35	0.32	0.37	0.37	0.37	0.37	0.37	0.36	0.38	0.36	0.36
Standability	0.36	0.37	0.36	0.38	0.37	0.38	0.38	0.38	0.33	0.35	0.3	0.39	0.39	0.38	0.39	0.39	0.37	0.38	0.39	0.39
Leaf width	0.19	0.19	0.19	0.2	0.19	0.21	0.23	0.2	0.19	0.2	0.17	0.2	0.2	0.2	0.21	0.21	0.19	0.21	0.2	0.2
Plant height	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.27	0.25	0.26	0.24	0.29	0.3	0.3	0.29	0.29	0.27	0.29	0.29	0.29
Leaf length	0.17	0.17	0.17	0.18	0.17	0.17	0.17	0.18	0.18	0.16	0.21	0.19	0.19	0.19	0.21	0.22	0.16	0.19	0.19	0.19
Cell wall concentration	0.56	0.56	0.55	0.56	0.56	0.54	0.55	0.55	0.56	0.53	0.54	0.56	0.57	0.57	0.58	0.57	0.56	0.55	0.56	0.57
In vitro dry matter digestibility	0.49	0.49	0.47	0.48	0.48	0.47	0.47	0.48	0.47	0.45	0.46	0.47	0.48	0.49	0.5	0.49	0.5	0.48	0.47	0.48
Etherified ferulates	0.45	0.45	0.43	0.44	0.45	0.43	0.45	0.44	0.43	0.42	0.45	0.43	0.43	0.44	0.44	0.43	0.45	0.44	0.43	0.43
Sucrose	0.31	0.31	0.28	0.31	0.31	0.29	0.28	0.3	0.26	0.27	0.26	0.28	0.27	0.29	0.29	0.29	0.31	0.3	0.28	0.28
Acid detergent lignin	0.41	0.41	0.39	0.4	0.41	0.39	0.4	0.4	0.37	0.38	0.36	0.39	0.4	0.4	0.4	0.39	0.41	0.4	0.39	0.4
Ethanol	0.31	0.31	0.28	0.3	0.31	0.29	0.29	0.3	0.31	0.27	0.29	0.33	0.33	0.35	0.36	0.35	0.32	0.3	0.33	0.34
Total soluble carbohydrates	0.3	0.3	0.27	0.29	0.29	0.28	0.27	0.29	0.25	0.26	0.24	0.31	0.31	0.31	0.32	0.31	0.3	0.29	0.31	0.31
Starch	0.31	0.31	0.31	0.32	0.31	0.31	0.3	0.31	0.26	0.29	0.28	0.29	0.29	0.29	0.29	0.29	0.3	0.32	0.29	0.29
Total sugar	0.26	0.26	0.26	0.26	0.26	0.26	0.28	0.26	0.25	0.26	0.28	0.26	0.26	0.27	0.26	0.27	0.26	0.26	0.26	0.27
Calories	0.24	0.24	0.24	0.25	0.24	0.27	0.28	0.25	0.25	0.24	0.26	0.25	0.25	0.25	0.26	0.26	0.24	0.26	0.25	0.25
Pentose sugars release	0.09	0.09	0.09	0.1	0.09	0.09	0.09	0.1	0.09	0.09	0.1	0.1	0.1	0.1	0.09	0.09	0.09	0.11	0.1	0.1
Carbon	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.08	0.08	0.07	0.08	0.09	0.09	0.09	0.09	0.08	0.07	0.08	0.09	0.09
Minerals	0.04	0.03	0.04	0.03	0.03	0.04	0.04	0.04	0.03	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.03	0.04	0.04

$r^2_{p,CV}$: 0.6, 0.4, 0.2

B

	GBS		EC		EC				
	Tetra	Octo	Tetra	Octo	Tetra		Octo		T + O
	all	all	all	all	lowland north	upland north	upland east	upland west	all
Anthesis date	0.59	0.42	0.59	0.45	0.46	0.45	0.56	0.24	0.42
Heading date	0.55	0.48	0.47	0.47	0.32	0.54	0.38	0.16	0.39
Total plant height	0.54	0.03	0.52	0.04	0.43	0.07	0.24	0.03	0.39
Standability	0.15	0.42	0.15	0.41	0.19	0.01	0.03	0.17	0.39
Leaf width	0.4	0.26	0.37	0.14	0.43	0.05	0.38	0.01	0.22
Plant height	0.42	0.08	0.46	0.04	0.03	0.11	0.26	0.01	0.34
Leaf length	0.18	0.03	0.18	0.07	0.31	0.03	0.01	0.03	0.2
Cell wall concentration	0.59	0.4	0.63	0.37	0.04	0.15	0.09	0.01	0.57
In vitro dry matter digestibility	0.47	0.46	0.47	0.5	0.05	0.29	0.22	0.13	0.49
Etherified ferulates	0.48	0.31	0.44	0.27	0.05	0.12	0.04	0.03	0.42
Sucrose	0.3	0.31	0.29	0.28	0.06	0.24	0.07	0.01	0.27
Acid detergent lignin	0.39	0.37	0.36	0.38	0.04	0.26	0.19	0.03	0.41
Ethanol	0.37	0.36	0.4	0.28	0.07	0.32	0.23	0.01	0.34
Total soluble carbohydrates	0.29	0.3	0.34	0.28	0.01	0.12	0.06	0.01	0.31
Starch	0.34	0.16	0.3	0.06	0.06	0.03	0.01	0.05	0.29
Total sugar	0.33	0.12	0.27	0.17	0.08	0.04	0.01	0.04	0.27
Calories	0.14	0.01	0.13	0.05	0.3	0.05	0.02	0.04	0.26
Pentose sugars release	0.2	0.06	0.12	0.03	0.01	0.01	0.06	0.05	0.1
Carbon	0.01	0.04	0.01	0.01	0.02	0.03	0.01	0.03	0.11
Minerals	0.02	0.02	0.05	0.02	0.05	0.04	0.01	0.03	0.04

$r^2_{p,CV}$: 0.6, 0.4, 0.2

Fig. S3 Prediction accuracy of models built using genetic variants on the cross-validation sets. (**A**) $r^2_{g,CV}$ for models built on both tetraploids and octoploids using different types of variants. Table above the heatmap shows different factors the variants used may belong to. (**B**) $r^2_{g,CV}$ for models built for different ploidy levels or subpopulations using GBS or EC bi-allelic SNPs. Color scale in the heatmap: median $r^2_{g,CV}$ among 10 replicate runs.

A

	GBS											Exome capture					GBS		Exome capture	
	bi-allelic						bi-allelic		multi-	bi-	multi-	bi-allelic			multi-allelic		bi-allelic			
	SNP						indel		SNP		indel	SNP		indel	SNP	indel	SNP			
	v1			v5			v5					v5					v5			
	all	common	v1-specific	all	common	v5-specific	all	balanced for bi-allelic indels	all	balanced for multi-allelic SNPs	all	balanced for GBS bi-allelic SNPs	all	all	all	all	genic	intergenic	subset genic	subset intergenic
Anthesis date	0.75	0.74	0.68	0.75	0.74	0.68	0.68	0.67	0.62	0.55	0.61	0.76	0.78	0.78	0.78	0.77	0.72	0.73	0.75	0.73
Heading date	0.67	0.67	0.6	0.67	0.68	0.57	0.64	0.6	0.54	0.5	0.57	0.66	0.68	0.69	0.68	0.68	0.66	0.64	0.65	0.63
Total plant height	0.54	0.54	0.46	0.54	0.54	0.48	0.46	0.49	0.46	0.4	0.46	0.5	0.51	0.52	0.52	0.52	0.51	0.52	0.49	0.48
Standability	0.44	0.43	0.4	0.42	0.42	0.38	0.39	0.4	0.37	0.34	0.31	0.45	0.46	0.47	0.47	0.47	0.41	0.4	0.44	0.43
Leaf width	0.41	0.44	0.31	0.4	0.42	0.29	0.34	0.35	0.31	0.26	0.28	0.41	0.44	0.45	0.43	0.43	0.38	0.36	0.4	0.38
Plant height	0.38	0.37	0.34	0.38	0.37	0.34	0.32	0.35	0.29	0.29	0.26	0.36	0.37	0.38	0.38	0.38	0.39	0.35	0.35	0.34
Leaf length	0.34	0.34	0.27	0.33	0.34	0.25	0.29	0.28	0.28	0.2	0.27	0.33	0.36	0.36	0.35	0.35	0.33	0.28	0.32	0.31
Cell wall concentration	0.61	0.6	0.58	0.59	0.6	0.56	0.59	0.57	0.57	0.53	0.55	0.6	0.61	0.61	0.61	0.61	0.58	0.58	0.6	0.59
In vitro dry matter digestibility	0.55	0.54	0.51	0.54	0.55	0.5	0.53	0.51	0.48	0.45	0.46	0.55	0.57	0.57	0.56	0.56	0.52	0.54	0.54	0.53
Etherified ferulates	0.5	0.48	0.48	0.48	0.48	0.46	0.46	0.46	0.44	0.42	0.42	0.48	0.5	0.5	0.5	0.5	0.45	0.49	0.48	0.48
Sucrose	0.46	0.47	0.38	0.45	0.45	0.34	0.35	0.39	0.27	0.31	0.34	0.41	0.42	0.42	0.42	0.42	0.41	0.42	0.4	0.39
Acid detergent lignin	0.45	0.44	0.43	0.44	0.44	0.41	0.42	0.42	0.39	0.37	0.38	0.49	0.51	0.51	0.52	0.51	0.41	0.45	0.48	0.47
Ethanol	0.41	0.41	0.36	0.42	0.4	0.36	0.36	0.37	0.33	0.29	0.31	0.47	0.49	0.49	0.49	0.49	0.37	0.41	0.45	0.45
Total soluble carbohydrates	0.39	0.39	0.34	0.38	0.37	0.32	0.34	0.34	0.29	0.27	0.29	0.41	0.44	0.43	0.43	0.43	0.36	0.35	0.4	0.39
Starch	0.32	0.32	0.31	0.32	0.32	0.32	0.3	0.31	0.27	0.29	0.31	0.31	0.32	0.32	0.32	0.32	0.32	0.32	0.31	0.3
Total sugar	0.32	0.3	0.3	0.3	0.31	0.28	0.28	0.29	0.29	0.26	0.29	0.34	0.36	0.36	0.36	0.36	0.31	0.29	0.34	0.33
Calories	0.28	0.26	0.27	0.28	0.26	0.28	0.26	0.26	0.25	0.24	0.26	0.31	0.33	0.33	0.33	0.33	0.27	0.26	0.3	0.3
Pentose sugars release	0.16	0.16	0.12	0.15	0.15	0.11	0.12	0.13	0.1	0.09	0.09	0.2	0.21	0.2	0.21	0.22	0.13	0.15	0.19	0.18
Carbon	0.1	0.09	0.09	0.1	0.1	0.1	0.09	0.09	0.09	0.07	0.12	0.11	0.12	0.13	0.12	0.13	0.08	0.12	0.11	0.1
Minerals	0.07	0.07	0.07	0.07	0.07	0.05	0.08	0.06	0.05	0.04	0.08	0.1	0.11	0.11	0.11	0.12	0.07	0.05	0.09	0.09

$r^2_{g,cv}$: 0.8, 0.6, 0.4, 0.2

B

	GBS		Exome capture		Exome capture				
	Tetra	Octo	Tetra	Octo	Tetra		Octo		T + O
	all	all	all	all	lowland north	upland north	upland east	upland west	all
Anthesis date	0.82	0.58	0.78	0.65	0.65	0.72	0.71	0.57	0.74
Heading date	0.77	0.54	0.72	0.59	0.6	0.64	0.56	0.3	0.68
Total plant height	0.7	0.22	0.61	0.28	0.59	0.13	0.38	0.09	0.54
Standability	0.31	0.44	0.28	0.46	0.28	0.08	0.14	0.12	0.46
Leaf width	0.55	0.24	0.55	0.3	0.54	0.26	0.37	0.02	0.47
Plant height	0.5	0.17	0.49	0.16	0.22	0.11	0.23	0.02	0.4
Leaf length	0.37	0.04	0.35	0.13	0.4	0.11	0.08	0.04	0.37
Cell wall concentration	0.64	0.37	0.64	0.47	0.04	0.26	0.27	0.01	0.61
In vitro dry matter digestibility	0.51	0.49	0.54	0.61	0.15	0.36	0.44	0.29	0.58
Etherified ferulates	0.51	0.29	0.51	0.32	0.11	0.27	0.14	0.12	0.48
Sucrose	0.48	0.33	0.41	0.37	0.36	0.22	0.1	0.13	0.41
Acid detergent lignin	0.48	0.39	0.52	0.46	0.23	0.3	0.29	0.08	0.53
Ethanol	0.45	0.47	0.49	0.42	0.18	0.38	0.45	0.18	0.48
Total soluble carbohydrates	0.45	0.3	0.49	0.36	0.32	0.14	0.13	0.05	0.45
Starch	0.34	0.14	0.33	0.13	0.05	0.02	0	0.03	0.32
Total sugar	0.4	0.1	0.4	0.19	0.16	0.12	0.05	0.02	0.35
Calories	0.19	0.02	0.27	0.11	0.35	0.06	0.06	0.04	0.36
Pentose sugars release	0.26	0.09	0.22	0.14	0.06	0.34	0.01	0.1	0.18
Carbon	0.03	0.02	0.08	0.02	0.05	0.02	0.06	0.01	0.14
Minerals	0.13	0.01	0.22	0.05	0.26	0.05	0.01	0.03	0.16

$r^2_{g,cv}$: 0.8, 0.6, 0.4, 0.2

Fig. S4 Improvement of r^2 of models built using different genetic variants on the cross-validation sets. (**A**) $r^2_{i,CV}$ for models built on both tetraploids and octoploids using different types of variants. Table above the heatmap shows different factors the variants used may belong to. (**B**) $r^2_{i,CV}$ for models built for different ploidy levels or subpopulations using GBS or EC bi-allelic SNPs. Color scale in the heatmap: $r^2_{i,CV} \times 100$.

A

	GBS											EC					GBS		EC	
	bi-allelic						bi-allelic		multi-	bi-	multi-	bi-allelic			multi-allelic		bi-allelic			
	SNP						indel	SNP			indel	SNP		indel	SNP	indel	SNP			
	v1			v5			v5					v5					v5			
	all	common	v1-specific	all	common	v5-specific	all	balanced for bi-allelic indels	all	balanced for multi-allelic SNPs	all	balanced for GBS bi-allelic SNPs	all	all	all	all	genic	intergenic	subset genic	subset intergenic
Anthesis date	27.29	26.9	23.66	28.92	27.48	22.16	21.21	22.37	17.89	12.35	15.67	30.92	32.83	30.34	26.66	26.87	24.72	29.14	29.2	25.75
Heading date	24.6	24.76	18.52	25	25.16	14.96	20.26	18.61	16.76	9.73	17.84	27.2	29.59	28.56	26.5	26.74	23.85	22.35	25.7	22.49
Total plant height	18.94	19.29	11.35	18.62	18.78	10.84	8.52	12.87	12.96	5.04	14.27	13.33	14.08	14.91	14.62	14.48	15.31	14.58	12.73	12.18
Standability	7.13	6.26	4.14	3.89	4.82	0.69	1.01	2.01	3.32	−0.35	0.93	5.95	7.51	8.6	8.14	8.03	4.07	1.88	4.94	4.63
Leaf width	22.17	24.29	12.45	20.48	22.51	8.6	11.17	14.48	11.94	6.74	10.82	21	23.93	24.4	22.79	22.39	18.82	14.1	20.05	17.54
Plant height	11.84	11.18	8.15	12.11	11.2	7.38	5.44	8.03	4.62	2.56	2.33	6.63	7.12	7.99	8.87	8.65	11.77	6.29	6.41	5.87
Leaf length	16.94	17.05	9.97	14.35	16.56	7.82	12.34	9.84	10.24	3.97	5.99	14.35	17.25	16.87	13.79	13.59	17.08	8.42	13.12	12.43
Cell wall concentration	4.81	3.54	3.14	3.35	3.57	1.78	3.51	1.42	1.69	0.02	0.64	3.53	4.1	3.86	2.84	3.73	1.77	2.79	3.21	2.6
In vitro dry matter digestibility	5.42	5.14	3.65	6.21	6.12	2.98	5.43	2.85	0.63	0.43	0.3	7.35	8.34	7.79	6.14	7.6	1.78	6.64	6.77	5.48
Etherified ferulates	4.9	3.43	4.08	3.76	3.35	2.71	1.29	1.7	1.44	0.07	−2.37	5.26	7.1	6.64	6.23	6.61	0.18	4.74	4.67	4.46
Sucrose	15.17	15.73	9.67	14.57	13.8	5.46	7	9.06	1.5	3.84	8.03	12.99	15.02	13.63	13.11	13.18	9.94	12.64	11.94	10.23
Acid detergent lignin	4.48	2.81	3.35	3.97	3.32	1.63	1.52	1.45	2.77	−0.76	1.95	9.35	11.39	11.24	11.43	11.6	0.01	4.81	8.43	7.3
Ethanol	10.41	10.45	8.41	11.33	9.71	6.91	7.56	6.62	1.92	2.52	2.8	14.34	16.38	14.24	12.44	13.5	5.64	11.17	12.65	10.44
Total soluble carbohydrates	9.82	9.1	7.02	8.62	7.55	4.06	6.73	4.68	4.15	1.06	4.29	10.16	12.88	11.96	11.21	11.25	5.91	6.03	8.64	7.84
Starch	0.7	0.72	−0.29	0.29	0.26	0.16	0.35	0	1.75	−0.95	3.52	2.07	3.02	3.04	2.6	2.35	1.61	−0.56	1.74	1.08
Total sugar	5.6	4.4	4.23	4.54	4.68	1.41	−0.56	2.75	4.22	0.59	0.82	7.58	9.23	9.42	9.18	9.42	4.85	2.48	7.39	6.34
Calories	3.66	2.14	2.98	2.55	1.86	1.12	−2.43	0.96	−0.16	−0.65	0.84	5.75	8.19	7.99	7.13	7.36	2.83	0.36	4.57	4.92
Pentose sugars release	7.3	6.98	3.57	5.7	6	1.99	2.69	3.45	1.05	0.24	−0.91	9.46	10.91	10.63	11.93	12.3	3.72	4.28	8.67	8.2
Carbon	2.72	1.85	2.19	3.25	2.33	3.74	2.19	1.55	1.2	0.04	3.36	2.54	3.3	3.67	3.57	4.24	0.87	4.19	2.66	1.61
Minerals	3.76	3.77	2.51	3.28	3.64	0.87	4.12	2.27	1.92	0.68	4.29	5.32	6.22	7.01	6.87	7.49	2.86	2.01	5.21	4.67

$r^2_{i,cv} \times 100$: 30, 20, 10, 0

B

	GBS		EC		EC				
	Tetra	Octo	Tetra	Octo	Tetra		Octo		T + O
	all	all	all	all	lowland north	upland north	upland east	upland west	all
Anthesis date	23.02	15.88	18.83	19.68	19.44	26.66	14.93	33.17	32.11
Heading date	22.52	6.54	25.47	12.6	28.17	9.72	17.79	14.67	28.78
Total plant height	16.24	19.08	8.78	24.56	16.7	5.33	13.43	5.92	14.84
Standability	15.78	1.99	12.28	4.79	8.95	6.82	11.25	−4.3	7.39
Leaf width	14.72	−2.11	17.93	15.77	11.16	20.34	−0.04	0.25	24.83
Plant height	7.76	8.86	3.26	12.25	18.87	0.2	−3.74	1.06	5.86
Leaf length	18.95	1.14	16.41	6.39	9.67	8.77	6.87	0.29	17.35
Cell wall concentration	5.19	−3.23	1.78	9.49	−0.18	10.94	17.49	0.71	4.49
In vitro dry matter digestibility	3.93	3.45	7.13	10.67	9.94	7.07	21.94	16.84	9.61
Etherified ferulates	3.07	−1.33	6.93	4.93	5.6	15.34	9.93	8.77	5.27
Sucrose	18.02	2.61	11.79	9.09	29.17	−1.6	2.85	11.95	13.14
Acid detergent lignin	8.01	1.52	16.2	7.89	18.59	4.13	10.2	4.67	11.48
Ethanol	8.8	11.12	8.36	14.22	11.68	6.17	22.2	17.48	13.72
Total soluble carbohydrates	15.52	0.06	14.77	8.66	31.38	2.65	6.92	3.78	14.27
Starch	0.83	−1.18	2.24	6.79	−0.34	−1.47	−0.97	−2.3	3.18
Total sugar	6.88	−1.49	13.1	2.12	7.85	7.73	3.67	−2.09	8.07
Calories	5.72	1.51	14.52	6.12	4.79	0.54	3.17	0.41	10.18
Pentose sugars release	6.04	2.46	9.95	10.77	5.45	33.25	−4.71	4.48	8.07
Carbon	2.23	−1.55	6.72	1.59	2.6	−1.3	5.76	−1.89	3.22
Minerals	10.64	−1.82	16.64	2.99	21.87	0.76	0.45	−0.35	11.1

$r^2_{i,cv} \times 100$: 30, 20, 10, 0

Fig. S5 Prediction accuracy of models built using the population structure on the test sets. (**A**) $r^2_{p,test}$ for models built on both tetraploids and octoploids using different types of variants. Table above the heatmap shows different factors the variants used may belong to. (**B**) $r^2_{p,test}$ for models built for different ploidy levels or subpopulations using GBS or EC bi-allelic SNPs. Color scale in the heatmap: median $r^2_{p,test}$ among 10 replicate runs.

A

	GBS											EC					GBS			EC
	bi-allelic						bi-allelic		multi-	bi-	multi-	bi-allelic			multi-allelic		bi-allelic			
	SNP						indel	SNP			indel	SNP		indel	SNP	indel	SNP			
	v1			v5			v5					v5					v5			
	all	common	v1-specific	all	common	v5-specific	all	balanced for bi-allelic indels	all	balanced for multi-allelic SNPs	all	balanced for GBS bi-allelic SNPs	all	all	all	all	genic	intergenic	subset genic	subset intergenic
Anthesis date	0.37	0.37	0.36	0.37	0.37	0.36	0.36	0.35	0.34	0.33	0.27	0.33	0.32	0.34	0.36	0.35	0.36	0.35	0.34	0.35
Heading date	0.47	0.46	0.48	0.47	0.47	0.5	0.48	0.47	0.45	0.45	0.41	0.47	0.47	0.47	0.47	0.47	0.46	0.47	0.47	0.48
Total plant height	0.36	0.37	0.34	0.37	0.36	0.36	0.38	0.36	0.26	0.35	0.3	0.4	0.4	0.4	0.4	0.4	0.37	0.37	0.41	0.4
Standability	0.42	0.41	0.42	0.43	0.41	0.46	0.44	0.43	0.38	0.41	0.37	0.45	0.45	0.45	0.43	0.42	0.43	0.42	0.44	0.44
Leaf width	0.3	0.3	0.28	0.33	0.32	0.3	0.31	0.32	0.32	0.31	0.31	0.31	0.3	0.29	0.3	0.3	0.29	0.37	0.31	0.31
Plant height	0.3	0.29	0.29	0.27	0.28	0.29	0.29	0.29	0.29	0.28	0.29	0.31	0.32	0.32	0.33	0.33	0.33	0.27	0.3	0.31
Leaf length	0.17	0.18	0.15	0.21	0.19	0.19	0.21	0.2	0.2	0.18	0.24	0.22	0.21	0.21	0.23	0.26	0.18	0.23	0.22	0.21
Cell wall concentration	0.54	0.54	0.52	0.53	0.53	0.52	0.51	0.52	0.52	0.49	0.51	0.53	0.53	0.55	0.54	0.54	0.53	0.53	0.53	0.54
In vitro dry matter digestibility	0.44	0.45	0.4	0.45	0.45	0.43	0.41	0.43	0.38	0.39	0.39	0.44	0.43	0.45	0.45	0.44	0.43	0.45	0.45	0.45
Etherified ferulates	0.35	0.35	0.34	0.36	0.35	0.34	0.33	0.35	0.33	0.33	0.32	0.36	0.36	0.36	0.37	0.36	0.35	0.35	0.37	0.37
Sucrose	0.39	0.41	0.33	0.42	0.42	0.36	0.32	0.4	0.34	0.35	0.28	0.39	0.38	0.37	0.36	0.36	0.39	0.43	0.4	0.39
Acid detergent lignin	0.39	0.38	0.39	0.38	0.38	0.38	0.4	0.39	0.35	0.38	0.36	0.4	0.41	0.4	0.41	0.39	0.38	0.4	0.4	0.4
Ethanol	0.43	0.44	0.36	0.43	0.44	0.36	0.38	0.41	0.42	0.35	0.45	0.37	0.38	0.39	0.41	0.41	0.42	0.43	0.37	0.38
Total soluble carbohydrates	0.31	0.32	0.27	0.33	0.33	0.31	0.31	0.31	0.29	0.28	0.29	0.23	0.23	0.24	0.24	0.23	0.31	0.32	0.23	0.23
Starch	0.3	0.3	0.28	0.29	0.3	0.25	0.28	0.29	0.28	0.29	0.29	0.3	0.3	0.3	0.3	0.3	0.3	0.28	0.3	0.3
Total sugar	0.26	0.25	0.28	0.27	0.26	0.29	0.29	0.27	0.25	0.26	0.27	0.27	0.27	0.26	0.26	0.25	0.25	0.26	0.27	0.27
Calories	0.26	0.26	0.26	0.26	0.26	0.3	0.33	0.26	0.22	0.27	0.25	0.25	0.25	0.26	0.25	0.25	0.25	0.26	0.25	0.25
Pentose sugars release	0.24	0.24	0.24	0.22	0.23	0.24	0.25	0.23	0.22	0.22	0.22	0.15	0.16	0.17	0.17	0.16	0.25	0.21	0.14	0.15
Carbon	0.13	0.13	0.12	0.13	0.13	0.11	0.13	0.13	0.13	0.12	0.14	0.15	0.15	0.14	0.14	0.14	0.13	0.12	0.15	0.15
Minerals	0.08	0.08	0.11	0.09	0.08	0.09	0.09	0.08	0.1	0.08	0.09	0.09	0.09	0.08	0.08	0.09	0.07	0.09	0.09	0.09

$r^2_{p,test}$: 0.6, 0.4, 0.2

B

	GBS		EC		EC				
	Tetra	Octo	Tetra	Octo	Tetra		Octo		T + O
	all	all	all	all	lowland north	upland north	upland east	upland west	all
Anthesis date	0.57	0.52	0.51	0.56	0.54	0.42	0.57	0.53	0.44
Heading date	0.59	0.37	0.65	0.41	0.5	0	0.55	0.01	0.48
Total plant height	0.45	0	0.44	0.02	0.15	0.27	0.4	0	0.32
Standability	0.11	0.51	0.14	0.62	0.34	0.05	0.42	0.17	0.46
Leaf width	0.26	0.1	0.18	0.41	0.26	0	0.33	0	0.2
Plant height	0.32	0.01	0.2	0.11	0.06	0.19	0.43	0.05	0.16
Leaf length	0.22	0.05	0.16	0	0.26	0.09	0.06	0.04	0.19
Cell wall concentration	0.67	0.48	0.55	0.42	0.02	0.27	0.01	0.57	0.52
In vitro dry matter digestibility	0.56	0.67	0.41	0.44	0.1	0.44	0.2	0.01	0.43
Etherified ferulates	0.52	0.19	0.48	0.08	0.17	0.1	0.26	0.02	0.38
Sucrose	0.42	0.4	0.36	0.52	0	0.23	0.06	0.16	0.38
Acid detergent lignin	0.36	0.37	0.34	0.29	0.02	0.13	0.47	0.07	0.37
Ethanol	0.42	0.18	0.42	0.45	0.03	0.45	0.46	0.07	0.35
Total soluble carbohydrates	0.4	0.15	0.21	0.26	0.01	0.49	0.04	0.01	0.24
Starch	0.31	0.05	0.34	0.09	0.03	0.03	0.05	0	0.28
Total sugar	0.16	0.14	0.38	0	0.24	0.01	0.11	0.16	0.24
Calories	0.22	0.03	0.22	0.01	0.31	0.01	0.32	0.31	0.25
Pentose sugars release	0.07	0.17	0.18	0.32	0.03	0.05	0.01	0.1	0.23
Carbon	0	0	0	0.02	0.24	0.01	0.01	0	0.1
Minerals	0.18	0.08	0.1	0.01	0	0.09	0.17	0.02	0.09

$r^2_{p,test}$: 0.6, 0.4, 0.2

Fig. S6 Prediction accuracy of models built using genetic variants on the test sets. (**A**) $r^2_{g,test}$ for models built on both tetraploids and octoploids using different types of variants. Table above the heatmap shows different factors the variants used may belong to. (**B**) $r^2_{g,test}$ for models built for different ploidy levels or subpopulations using GBS or EC bi-allelic SNPs. Color scale in the heatmap: median $r^2_{g,test}$ among 10 replicate runs.

A

	GBS										EC					GBS		EC	
	bi-allelic						bi-allelic	multi-	bi-	multi-	bi-allelic			multi-allelic		bi-allelic			
	SNP						indel	SNP		indel	SNP		indel	SNP	indel	SNP			
	v1			v5			v5				v5					v5			
	all	common	v1-specific	all	common	v5-specific	all	balanced for bi-allelic indels	all	balanced for multi-allelic SNPs	all	balanced for GBS bi-allelic SNPs	all	all	all	all	genic	intergenic	subset genic	subset intergenic
Anthesis date	0.58	0.57	0.55	0.56	0.57	0.45	0.56	0.51	0.52	0.43	0.38	0.56	0.57	0.58	0.57	0.56	0.52	0.57	0.56	0.55
Heading date	0.68	0.67	0.62	0.67	0.66	0.6	0.61	0.61	0.58	0.52	0.52	0.73	0.74	0.75	0.74	0.75	0.63	0.68	0.71	0.7
Total plant height	0.54	0.53	0.45	0.53	0.51	0.49	0.54	0.48	0.45	0.41	0.37	0.57	0.59	0.59	0.59	0.6	0.54	0.5	0.57	0.56
Standability	0.54	0.52	0.51	0.56	0.53	0.55	0.52	0.51	0.52	0.44	0.4	0.58	0.6	0.6	0.59	0.58	0.56	0.51	0.57	0.54
Leaf width	0.53	0.5	0.5	0.5	0.51	0.45	0.5	0.46	0.46	0.38	0.4	0.52	0.54	0.55	0.55	0.56	0.5	0.46	0.51	0.51
Plant height	0.37	0.38	0.31	0.37	0.38	0.31	0.35	0.35	0.41	0.31	0.27	0.38	0.39	0.38	0.39	0.39	0.36	0.37	0.38	0.38
Leaf length	0.37	0.32	0.36	0.31	0.3	0.3	0.38	0.28	0.25	0.22	0.3	0.37	0.38	0.38	0.38	0.4	0.25	0.34	0.35	0.34
Cell wall concentration	0.55	0.54	0.55	0.56	0.54	0.56	0.54	0.54	0.52	0.5	0.46	0.58	0.59	0.59	0.58	0.57	0.54	0.55	0.58	0.57
In vitro dry matter digestibility	0.52	0.5	0.5	0.49	0.48	0.47	0.39	0.46	0.44	0.41	0.46	0.52	0.53	0.53	0.53	0.53	0.47	0.48	0.51	0.53
Etherified ferulates	0.37	0.36	0.37	0.37	0.37	0.34	0.34	0.36	0.42	0.33	0.36	0.42	0.43	0.44	0.44	0.45	0.36	0.36	0.42	0.42
Sucrose	0.36	0.35	0.34	0.36	0.34	0.38	0.33	0.34	0.37	0.31	0.33	0.42	0.43	0.42	0.42	0.42	0.37	0.32	0.41	0.41
Acid detergent lignin	0.47	0.47	0.43	0.44	0.47	0.36	0.43	0.43	0.39	0.41	0.48	0.5	0.51	0.51	0.51	0.5	0.46	0.41	0.49	0.48
Ethanol	0.5	0.49	0.48	0.51	0.5	0.45	0.4	0.46	0.37	0.37	0.46	0.47	0.49	0.48	0.49	0.48	0.49	0.49	0.46	0.44
Total soluble carbohydrates	0.37	0.38	0.33	0.4	0.38	0.38	0.27	0.37	0.29	0.31	0.36	0.32	0.34	0.35	0.34	0.34	0.39	0.36	0.32	0.32
Starch	0.32	0.33	0.3	0.33	0.32	0.32	0.25	0.31	0.32	0.29	0.25	0.33	0.33	0.32	0.33	0.32	0.33	0.3	0.33	0.32
Total sugar	0.27	0.26	0.29	0.28	0.27	0.3	0.33	0.28	0.29	0.27	0.29	0.31	0.32	0.33	0.32	0.32	0.26	0.3	0.31	0.3
Calories	0.31	0.34	0.26	0.33	0.35	0.3	0.31	0.32	0.25	0.29	0.33	0.3	0.31	0.31	0.32	0.31	0.32	0.34	0.3	0.3
Pentose sugars release	0.27	0.28	0.26	0.28	0.3	0.21	0.22	0.25	0.21	0.21	0.25	0.14	0.14	0.15	0.14	0.14	0.29	0.22	0.13	0.14
Carbon	0.11	0.12	0.1	0.12	0.12	0.11	0.08	0.12	0.14	0.1	0.09	0.16	0.17	0.16	0.16	0.15	0.15	0.08	0.16	0.16
Minerals	0.2	0.16	0.18	0.18	0.17	0.16	0.11	0.14	0.12	0.1	0.13	0.18	0.19	0.19	0.18	0.18	0.18	0.15	0.16	0.17

$r^2_{g,test}$: 0.8, 0.6, 0.4, 0.2

B

	GBS		EC		EC				
	Tetra	Octo	Tetra	Octo	Tetra		Octo		T + O
	all	all	all	all	lowland north	upland north	upland east	upland west	all
Anthesis date	0.67	0.62	0.66	0.77	0.84	0.63	0.72	0.58	0.68
Heading date	0.73	0.46	0.83	0.54	0.79	0.52	0.75	0.25	0.75
Total plant height	0.53	0.19	0.54	0.24	0.33	0.34	0.39	0.06	0.47
Standability	0.27	0.53	0.34	0.67	0.49	0.05	0.39	0.22	0.58
Leaf width	0.47	0.17	0.43	0.33	0.39	0.13	0.37	0.15	0.44
Plant height	0.42	0.07	0.29	0.19	0.37	0.15	0.52	0.15	0.26
Leaf length	0.29	0.08	0.45	0	0.2	0.22	0.01	0.03	0.35
Cell wall concentration	0.68	0.54	0.55	0.44	0.01	0.5	0.17	0.67	0.56
In vitro dry matter digestibility	0.64	0.68	0.42	0.45	0	0.48	0.18	0.28	0.48
Etherified ferulates	0.59	0.3	0.45	0.44	0.03	0.24	0.51	0.05	0.46
Sucrose	0.42	0.44	0.52	0.51	0.48	0.25	0.26	0.13	0.53
Acid detergent lignin	0.45	0.35	0.52	0.37	0.34	0.29	0.55	0.19	0.49
Ethanol	0.52	0.27	0.49	0.59	0.13	0.45	0.5	0.25	0.55
Total soluble carbohydrates	0.47	0.25	0.23	0.37	0.33	0.32	0.09	0.37	0.28
Starch	0.33	0.01	0.33	0.1	0.04	0.16	0.12	0.27	0.32
Total sugar	0.29	0.2	0.45	0.1	0.63	0.29	0.16	0.28	0.33
Calories	0.42	0.05	0.25	0.01	0.47	0.03	0.23	0.14	0.3
Pentose sugars release	0.16	0.18	0.28	0.27	0.01	0.03	0.11	0.22	0.28
Carbon	0.11	0.02	0.1	0.01	0.14	0.04	0	0.03	0.16
Minerals	0.31	0.12	0.15	0.04	0.18	0.02	0.15	0.11	0.11

$r^2_{g,test}$: 0.8, 0.6, 0.4, 0.2

Fig. S7 Improvement of r^2 of models built using different genetic variants on the test sets. (**A**) $r^2_{i,test}$ for models built on both tetraploids and octoploids using different types of variants. Table above the heatmap shows different factors the variants used may belong to. (**B**) $r^2_{i,test}$ for models built for different ploidy levels or subpopulations using GBS or EC bi-allelic SNPs. Color scale in the heatmap: $r^2_{i,test} \times 100$.

A

	GBS											EC					GBS		EC	
	bi-allelic						bi-allelic		multi-	bi-	multi-	bi-allelic			multi-allelic		bi-allelic			
	SNP						indel	SNP			indel	SNP		indel	SNP	indel	SNP			
	v1			v5			v5					v5					v5			
	all	common	v1-specific	all	common	v5-specific	all	balanced for bi-allelic indels	all	balanced for multi-allelic SNPs	all	balanced for GBS bi-allelic SNPs	all	all	all	all	genic	intergenic	subset genic	subset intergenic
Anthesis date	20.23	19.82	19.04	19.36	20.61	9.25	20.61	16.15	17.84	10.34	10.33	23.29	24.78	23.39	20.17	20.59	16.06	22.21	22.05	20.1
Heading date	20.83	20.67	14.13	20	19.41	10.59	12.55	14.25	13.14	6.51	11.47	25.38	27.34	27.48	27.11	27.86	16.46	21.12	23.97	22.09
Total plant height	18.58	16.65	10.62	16.64	14.96	13.42	15.95	11.71	19.15	6.01	6.58	17.04	19.09	19.81	19.93	19.77	16.54	12.96	16.23	15.61
Standability	12.29	10.6	9.51	13.33	12.11	9.24	8.09	8.05	13.87	2.92	2.88	13.03	14.31	15.32	15.79	15.71	12.28	9.12	12.84	9.17
Leaf width	22.93	19.42	21.85	17.17	18.62	15	18.83	14.08	14.2	7.65	8.78	21.75	24.27	25.45	25.08	25.72	20.74	8.93	19.76	20.28
Plant height	7.57	8.92	1.62	9.91	10.59	2.16	6.42	6.66	12.32	3.29	−2	6.97	6.67	6.1	6.47	6.29	2.92	9.51	7.81	7.11
Leaf length	20.32	14.9	20.72	10.21	11.51	11.21	17.29	7.61	5.1	4.02	6.22	14.91	16.89	17.16	14.52	14.12	7.32	10.58	13.46	12.72
Cell wall concentration	1.24	0.4	3.29	2.88	1.15	4.48	2.55	1.93	0.04	0.74	−4.22	5.19	5.82	4.61	3.73	3.31	1.08	2.16	4.49	3.21
In vitro dry matter digestibility	8.05	5.32	9.67	3.83	3.27	3.96	−2.4	2.51	6.1	1.74	6.5	7.72	10.2	8.5	8.53	9.18	3.29	3.78	5.64	7.63
Etherified ferulates	1.85	0.61	2.3	1.05	1.37	−0.01	0.78	0.76	8.22	−0.3	3.52	5.63	7.11	8.05	7.19	8.43	1.17	1.07	5.52	4.91
Sucrose	−2.97	−6.68	0.81	−6.01	−8.42	2.39	1.21	−5.45	2.89	−3.59	4.49	2.25	4.75	4.5	5.32	5.86	−1.78	−10.96	0.37	2.04
Acid detergent lignin	8.68	9.25	3.93	5.89	8.64	−2.1	3.32	3.74	4.19	2.55	11.34	9.32	10.59	11.07	9.96	11.08	7.91	0.63	8.79	7.64
Ethanol	7.6	5.09	11.72	8.02	6.5	9.01	2.5	4.91	−4.96	2.44	1.4	9.64	10.88	8.53	7.62	6.84	6.26	6.69	9.13	6.64
Total soluble carbohydrates	6.12	5.68	6.23	7.32	5.34	7.34	−3.85	5.78	0	3.66	7.26	9.28	10.95	10.46	10.1	10.32	8.51	3.71	8.63	8.68
Starch	2.74	2.48	2.11	3.68	2.35	6.78	−3.27	2.08	4.06	−0.34	−3.75	2.27	2.8	2.6	2.93	2.35	2.48	2.83	2.41	1.56
Total sugar	1.74	1.11	1.51	1.37	1.38	0.81	3.55	1.46	4.3	0.29	1.89	4.01	5.11	6.18	6.53	6.57	0.83	3.12	3.86	3.44
Calories	5.46	8.26	−0.64	7.19	9.08	0.47	−1.66	5.46	2.38	2.45	7.89	4.8	5.18	5.35	6.83	6.4	6.32	8.55	4.63	4.4
Pentose sugars release	3.12	3.98	1.14	5.86	7.79	−2.91	−2.5	2.45	−0.95	−0.73	2.91	−1.57	−2.2	−2.6	−2.63	−1.84	4.67	1.54	−1.07	−1.41
Carbon	−1.44	−1.21	−1.66	−0.73	−1.05	0.19	−4.95	−1.16	1.04	−1.77	−5.59	1.61	1.87	1.48	1.8	0.76	2.09	−4.26	1.18	0.81
Minerals	11.31	8.77	7.3	9.54	8.52	6.98	2.47	6.34	2.04	2.67	4.05	8.46	10.42	10.77	9.9	9.72	11.22	6.32	7.46	8.31

$r^2_{i,test} \times 100$: 40, 20, 0

B

	GBS		EC		EC				
	Tetra	Octo	Tetra	Octo	Tetra		Octo		T + O
	all	all	all	all	lowland north	upland north	upland east	upland west	all
Anthesis date	9.85	10.51	15.34	20.74	29.68	21.55	14.57	5.04	23.55
Heading date	14.43	8.87	18.46	12.43	28.39	52.22	20.63	23.4	27.03
Total plant height	8.62	19.05	10.56	22.26	18.81	6.8	−0.57	6.15	14.77
Standability	16.65	2.2	20.04	5.64	14.83	0.25	−2.98	4.27	12.01
Leaf width	20.42	7.07	24.82	−7.84	13.09	12.48	4.35	14.34	23.65
Plant height	10.37	6.23	8.98	7.91	31.23	−3.9	8.79	9.25	10.11
Leaf length	7.07	2.38	29.28	0.13	−5.5	13.08	−4.69	−1.49	15.95
Cell wall concentration	0.56	5.95	−0.32	1.82	−1.95	23.79	16.66	9.72	3.95
In vitro dry matter digestibility	7.57	0.54	0.85	1.14	−10.18	3.77	−2.1	26.85	4.65
Etherified ferulates	7.38	10.54	−3.14	35.94	−13.98	13.99	24.37	2.97	7.45
Sucrose	−0.21	4.55	15.74	−0.65	48.11	2.55	20.07	−2.98	14.46
Acid detergent lignin	9.2	−2.32	18.07	8.1	31.89	16.82	7.55	11.45	11.61
Ethanol	9.69	8.9	7.49	14.81	10.43	−0.09	4.53	18.61	19.71
Total soluble carbohydrates	6.29	9.14	2.37	11.49	32.11	−16.65	5.31	36.8	4.35
Starch	1.3	−3.47	−0.56	0.88	1.56	12.96	6.59	26.61	4.3
Total sugar	12.85	6.08	6.68	9.04	38.83	28.36	5.18	12.1	9.53
Calories	19.56	1.64	3.17	0.04	15.57	2.7	−9.31	−17.22	5.1
Pentose sugars release	8.88	0.98	10.31	−5.13	−1.83	−2.03	9.54	12	5.37
Carbon	10.54	1.66	9.19	−1.22	−10.32	3.12	−1.28	2.62	6.12
Minerals	13.15	4.24	5.04	2.71	18.39	−7.36	−2.61	8.58	2.41

$r^2_{i,test} \times 100$: 40, 20, 0

Fig. S8 Prediction accuracy of models built using different GBS variants and models integrating all GBS variants. Color and value in (**A-D**) indicate $r^2_{p,cv}$, $r^2_{p,test}$, $r^2_{g,cv}$, and $r^2_{g,test}$, respectively.

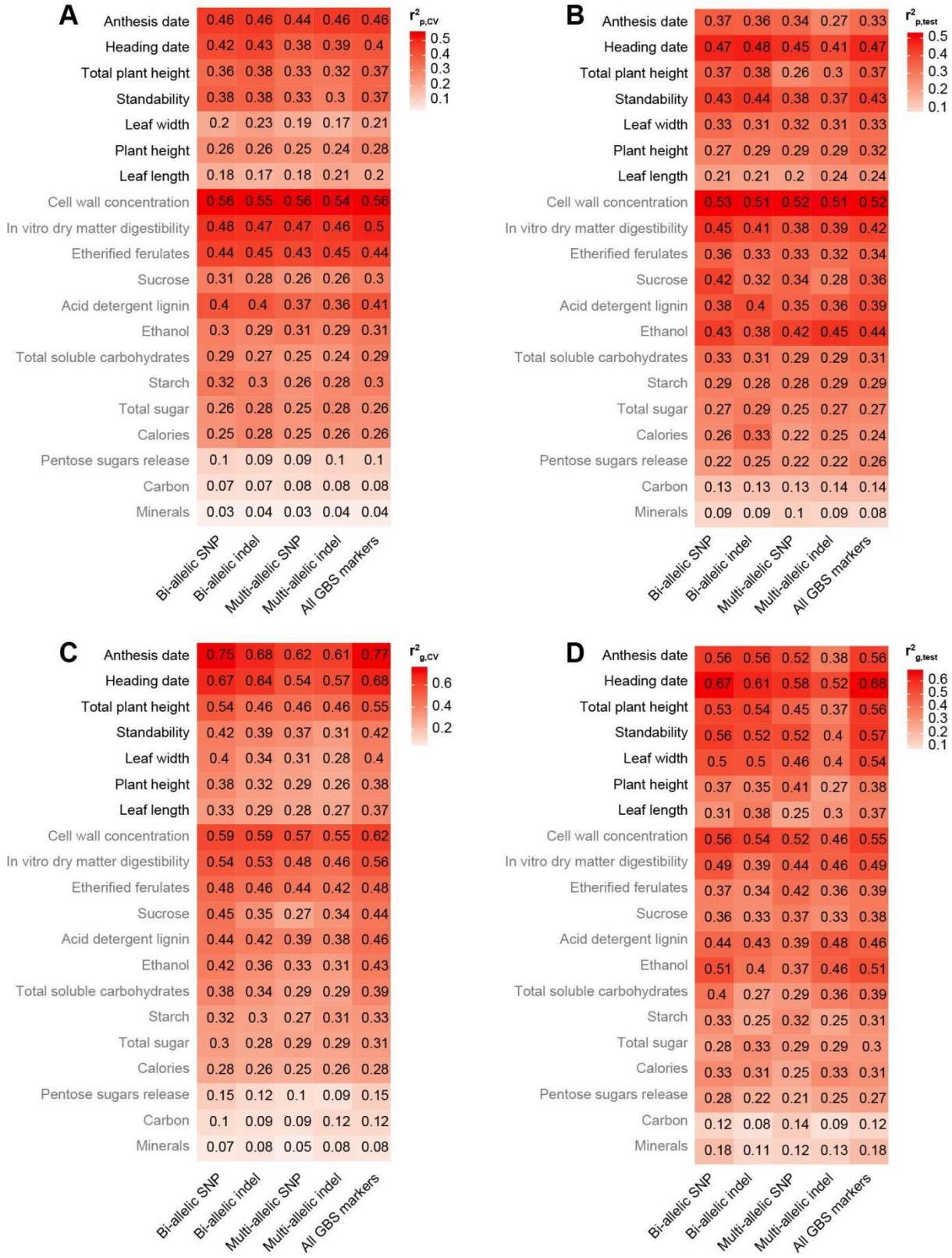

Fig. S9 Prediction accuracy of models built for four subpopulations. (**A-C**) The $r^2_{p,CV}$ (**A**), $r^2_{g,CV}$ (**B**) and $r^2_{i,CV}$ (**C**) for models using EC bi-allelic SNPs, built for Lowland north (4X) (dark blue), Upland north (4X) (blue), Upland east (8X) (green) and Upland west (8X) (yellow) subpopulations. The color background indicates that the model built for the corresponding subpopulation had the highest prediction accuracy for the trait in question. Error bar: standard deviation.

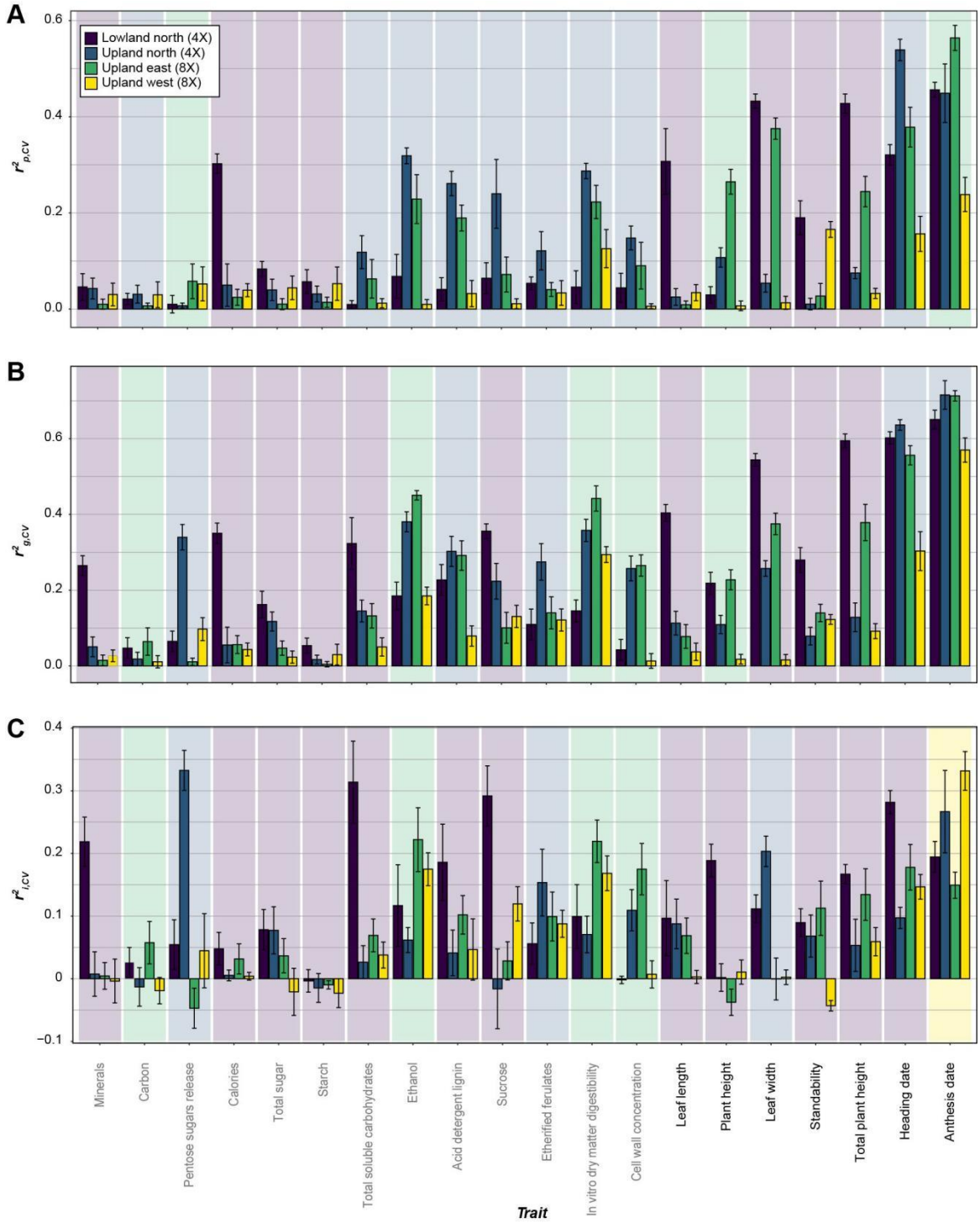

Fig. S10 Flowering time orthologous genes associated with variants that had absolute coefficients above the 99th percentile in eight models. The type of genetic variants used to build the models were listed above the heatmap. Yellow, green and blue: variants with absolute coefficients >99th, >95th and ≤95th percentiles, respectively.