Impact of genome assemblies, genotyping methods, variant types, ploidy levels and population structures on genomic prediction in switchgrass

Peipei Wang[1,2,3]*, Fanrui Meng[1,3], Christina B Del Azodi[4], Kenia Segura Abá[1,5], Michael D. Casler[6], and Shin-Han Shiu[1,3,5,7]*

[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] St. Vincent's Institute of Medical Research, Fitzroy 3065, Victoria, Australia
[5] Genetics and Genome Sciences Program, Michigan State University, East Lansing, MI 48824, USA
[6] Department of Plant and Agroecosystem Sciences, University of Wisconsin, Madison, WI 53706, USA
[7] Department of Computational Mathematics, Science, and Engineering, Michigan State University, East Lansing, MI 48824, USA

*: Corresponding authors

Abstract

Genomic prediction, where genotype information is used to predict phenotypes, has accelerated the breeding processes and can provide mechanistic insights into phenotypes of interest. Switchgrass (*Panicum virgatum* L.) is a perennial biofuel feedstock with multiple traits targeted for accelerated breeding using genomic prediction approaches. To optimize switchgrass genomic prediction, we assessed the impact of genome assembly versions, sequencing strategies for variant calling, variant types, variant allelic complexities, and polyploidy levels on the prediction of 20 traits in a switchgrass diversity panel with 486 individuals. We found that genome assembly has limited impact on prediction accuracy. Bi-allelic insertion/deletions and multi-allelic variants are as useful as bi-allelic single nucleotide polymorphisms. In addition, models built using exome capture-derived variants tend to have higher prediction accuracy than those using genotyping-by-sequencing variants. Sequencing depth, ploidy levels and population structures also have significant effects on prediction accuracy. The prediction accuracy can be improved by integrating different types of variants. We also show that the anthesis date prediction models based on exome capture variants, especially those using exome capture multi-allelic indels, identified the highest numbers of genes similar to known flowering time genes in other species. 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.

Introduction

With the rapid development of sequence techniques and phenome-wide phenotypic detection, genomic prediction (GP) or genomic selection is accelerating plant and animal breeding (Asoro *et al.*, 2011; de los Campos *et al.*, 2009; Crossa *et al.*, 2017; Lorenzana and Bernardo, 2009; VanRaden *et al.*, 2009). 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 (Crossa *et al.*, 2017). With the ever-changing updates of sequencing techniques and model building approaches, it is important to assess how best to do GP.

Substantial effort has been put on understanding the effects of different factors on the accuracy of GP (Desta and Ortiz, 2014; Ramstein *et al.*, 2016). For example, Kriaridou *et al.*, (2020) and Norman *et al.*, (2018) have shown that the prediction accuracy increases along with the increase of variant density to a certain point. Other factors influencing the GP accuracy include population size (Akdemir *et al.*, 2019; Jannink *et al.*, 2010), population structure (Norman *et al.*, 2018), relationship between training and test populations (Isidro *et al.*, 2015), and statistical methods (Azodi *et al.*, 2019; Desta and Ortiz, 2014; Zhang *et al.*, 2019), etc.

Beyond the above, there remain additional factors with unclear effects on GP accuracy. The first is genome assembly quality—with reduced sequencing cost and improvement in assembly algorithms, assembly versions change frequently. In a blueberry study, high-quality genome assembly-based models do not necessarily have higher prediction accuracy than those based on low-quality genomes (Benevenuto *et al.*, 2019). It remains an open question whether this finding is common in other species or is study-specific. The second is genotyping approaches. In addition to traditional approaches such as restriction fragment length polymorphism and random amplified polymorphism, genetic variants that are generated through genotyping-by-sequencing (GBS), exome capture or whole genome sequencing (WGS) are now commonly used in population genetics and GP applications (Lopez *et al.*, 2021; Scheben *et al.*, 2017; H. Wang *et al.*, 2023). However, it is unclear how different genotyping approaches impact the GP accuracy. The third factor is the types of genetic variants used for building GP models.

The majority of early genome-wide association study (GWAS) and GP studies mainly used bi-allelic SNPs (Crossa *et al.*, 2017), with a few exceptions using indels (Misra *et al.*, 2017; Song *et al.*, 2018). However, the genetic variants can be both bi-allelic and multi-allelic. Due to the higher complexity of multi-allelic variants compared with biallelic ones, how the former contribute to GP accuracy remains largely unexplored. Finally, the fourth factor is polyploidy. There are multiple rounds of whole genome duplications in the evolution of land plants and most plant species have both diploid and polyploid individuals that coexist (Fawcett and Van De Peer, 2010; Ramsey and Schemske, 2002). To simplify the problems associated with polyploidy, past studies mostly used either dominant variants or variants converted to diploid status (Batista *et al.*, 2022; Dufresne *et al.*, 2014), or took individuals with different ploidies as if they have the same ploidy level (Lu *et al.*, 2013). Thus, whether the ploidy levels affect the GP accuracy remains unresolved.

Switchgrass (*Panicum virgatum* L.) is a perennial grass with both tetra- and octo-ploids and a key species as a bioenergy feedstock (McLaughlin and Adams Kszos, 2005). 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 exome capture (Evans *et al.*, 2015). Here, 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. Finally, we also assessed how different types of genetic variants impacted our ability to uncover the genetic basis underlying anthesis date by interpreting GP models.

Results and Discussion

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 mapped to the version 5 genome assembly of the lowland tetraploid AP13 (Lovell

et al., 2021). 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). 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 than biochemical traits studied here.

Better genome assembly does not provide better trait prediction

Next, we investigated the influence of the first factor, genome assembly, on GP accuracy. As the v5 assembly (N50=5.5 Mb) is long-read based, the v5-based GP models are expected to have higher accuracy than models based on the short-read based assembly (e.g., v1, N50=54 Kb) due to the improved assembly quality in v5 (Lovell *et al.*, 2021). To test this, the GBS sequences were also mapped to the v1 assembly, and genetic variants were called using the same methods and parameter settings as those for v5-based variants (see **Methods**).

Contrary to our expectation, models built using variants derived from two different versions of assemblies had similar performance, as the average difference in $r^2_{i,CV}$ for 20 traits was 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. S1-6**, respectively). 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. S7A,B** and **Table S3**). The assembly sizes of v1 (1,230 Mb) and v5 assemblies (1,129 Mb)

are also similar. In addition, most of the GBS bi-allelic SNPs were shared between these two assemblies, as 5283 (69.92%) out of 7556 v1-based bi-allelic SNPs had corresponding v5-based ones (see **Methods** and **Table S4**). Beyond these similarities, N50 for v5 assembly is nearly 100 times higher than v1. But this 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 (difference 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, and significantly worse $r^2_{i,CV}$ only for anthesis date (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 tend to have better performance when they are 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 has 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 indicate that 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

Next, we explored the effect of genotyping approaches on trait predictions using bi-allelic SNPs. Exome capture-based models had better $r^2_{i,CV}$ than GBS-based models for 15 traits (the "Unbalanced" column, **Fig. 3A**). Since there were ~72 times more exome capture variants (526,705) than GBS variants (7,357), to eliminate the influence of variant numbers on prediction accuracy, we randomly down-sampled exome capture bi-allelic SNPs to the same number of GBS bi-allelic SNPs. This down-sampling was conducted 100 times, and the median $r^2_{i,CV}$ of the 100 replicate runs was calculated (**Methods**). Models built with these down-sampled exome capture variants still had significantly higher $r^2_{i,CV}$ than GBS bi-allelic SNP-based models for 13 traits (the "Balanced" column, **Fig. 3A**).

When comparing the distributions of variants, exome capture variants tend to be located in genic regions or in intergenic genes that are closer to genes compared with GBS variants (**Fig.**

3B, Fig. S7). Thus, we first tested the hypothesis that the better performance of exome capture-based models is because they include more genic variants (86.4%) than GBS-base (55.5%, **Table S3**). We established GP models using only genic GBS and exome capture variants (down-sampled to match the number of genic GBS variants) and found that genic GBS-based models still had significantly smaller $r^2_{i,CV}$ for 15 of 20 traits than genic exome capture-based models (the "Genic" column, **Fig. 3A**). While we used the same numbers of variants from GBS and exome capture, we found significantly fewer genes that contained GBS variants (3,296) than contained exome capture variants (median across 100 replicates=3,754; *p*-value=9.31e-21, **Fig. 3C**). Our findings indicate that having more genes with variant coverage tends to improve GP accuracy.

We also compared models built using GBS and exome capture intergenic variants. Models built using GBS intergenic variants had significantly smaller $r^2_{i,CV}$ for 10 of 20 traits than the corresponding exome capture-based models (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 exome capture variants (down-sampled to match intergenic GBS, the median number of genes among 100 replicates was 2,944). We also tested the hypothesis that better exome capture-based model performance was due to higher proportion of intergenic variants closer to gene bodies than intergenic GBS variants. If this was true, we would expect negative correlations between the distance to the gene body and how important variants were in explaining trait variances which can be estimated with the absolute coefficient of variants. However, we did not see such correlation for leaf width models (Spearman's Rank Correlation Coefficients = -7.7e-03 and 7.2e-03 for models built with intergenic GBS and exome capture variants, respectively) as an example, for which intergenic exome capture variants-based models had higher $r^2_{i,CV}$ than intergenic GBS variants-based models (the "Intergenic" column, **Fig. 3A**). Taken together, those results suggest that exome capture variants are superior to GBS variants in GP of switchgrass traits, and gene coverage by variants has a significant impact on how well a trait can be predicted.

Different types of genetic variants have similar trait prediction accuracy

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 GP (Biová *et al.*,

2024; Veerkamp *et al.*, 2016). For the GBS data, 1,827 bi-allelic indels were identified (**Table S3**). The models built using bi-allelic indels had significantly lower $r^2_{i,CV}$ than models based on all bi-allelic SNPs (without down-sampling, the 1st column, **Fig. 4A**) for 15 traits, but had lower $r^2_{i,CV}$ for only 6 traits and higher $r^2_{i,CV}$ for another 6 traits (the 2nd column, **Fig. 4A**) than models built using balanced bi-allelic SNPs (downsampled to 1,827 to be the same as bi-allelic indels, repeated 100 times).

In addition to bi-allelic indels, there were 1,836 multi-allelic variants identified for the GBS data, including multi-allelic SNPs, multi-allelic indels and multi-allelic SNPs/indels (before encoding). To simplify the question, we only compared the models built using GBS multi-allelic SNPs (350) and bi-allelic SNPs (7,357) as an example. Models built using GBS multi-allelic SNPs had significantly lower $r^2_{i,CV}$ than models based on all GBS bi-allelic SNPs (without down-sampling) for 16 traits (the 3rd column in **Fig. 4A)**. In contrast, when the number of variants were controlled for, multi-allelic SNP-based models had significantly higher $r^2_{i,CV}$ for 15 traits than models 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 slightly fewer genes covered by or adjacent to bi-allelic indels (1,635) than balanced bi-allelic SNPs (median: 1,711, *p*-value = 1.02e-09, **Fig. 4B**), but 0.83 times more genes by multi-allelic SNPs (632) than balanced bi-allelic SNPs (median: 346, *p*-value = 9.87e-24, **Fig. 4C**). Taken together, these results indicate that variants besides the commonly used bi-allelic SNPs can contribute to GP models and should also be further explored, particularly in their contribution to improve the coverage of genes.

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. For genomic prediction, the variants for octoploid individuals with e.g., AAAA and TTTT genotypes were converted to AA and TT, respectively, while ATTT, AATT and AAAT genotypes were converted to AT (see **Methods**). Considering the relatively higher complexity of variants for octoploid individuals, we asked whether the same variant set (GBS or exome capture) 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. 5A**).

One possible explanation of the lower $r^2_{i,CV}$ for octoploid models is that the variants called for octoploids are not as good as those for tetraploids in GBS data. This is because the GBS bi-allelic SNPs only had average read depth at 4.41 (left panel in **Fig. 5B**), and had high proportions of missing data for both tetraploids (average=916, 12.5%, middle panel) and octoploids (893, 12.1%, right panel in **Fig. 5B**). To test this, we built models separately for tetraploids and octoploids using exome capture bi-allelic SNPs, as that exome capture data had average read depth at 21.10 (left panel in **Fig. 5C**), and lower proportion of missing data for tetraploids (16,237, 3.1%, middle panel) and octoploids (13,023, 2.4%, right panel in **Fig. 5C**). Exome capture-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. 5A**). These results indicate that the lower variant quality in GBS data might, but only partially, explain the lower GP accuracy of octoploid models.

Another potential explanation is that the trait heritability of octoploids is lower than those of tetraploids, because both $r^2_{p,CV}$ (**Fig. 5D**) and $r^2_{g,CV}$ (**Fig. 5E**) for octoploid-based models were generally smaller than those values based on tetraploids, in both GBS and exome capture-based models. Considering that there are mainly three subpopulations for tetraploids and two for octoploids, and the genetic distances between individuals in three tetraploid subpopulations are larger than those in two octoploid subpopulations (see **Fig. 4** in Evans *et al.*, 2015), we further assessed the impact of genetic distances among individuals on GP accuracy. To do so, we built a model for each of four subpopulations with ≥85 individuals each (**Table S1**) using exome capture bi-allelic SNPs. The four subpopulations are, from high to low average within-subpopulation genetic distances: lowland north (4X) > upland north (4X) > upland east (8X) > upland west (8X) (Evans *et al.*, 2015). We found that subpopulations with larger intra-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. S8**). This is inconsistent with previous findings that GP accuracy in Barley decreased with the increase of genetic distances in the training population (Lorenz and Smith, 2015).

Besides the above two potential explanations, we should also emphasize that our encoding strategies for octoploid variants does not fully capture the genetic variations harbored in octoploid individuals. For example, individuals with ATTT, AATT and AAAT were treated the same as AT currently. More complicated allelic combinations may have more complicated dosage effects on gene function and plant phenotypes (Batista *et al.*, 2022). Thus, in future studies, exploration of different encoding strategies that better capture the allelic complexity in polyploids is warranted.

Integrating different types of genetic variants can improve GP accuracy

In the above section, we have discussed the different effects of multiple factors on GP accuracy. Next, we asked whether the prediction can be improved by taking advantage of all the GBS and exome capture variants. However, due to the large number of exome capture variants (2,552,214 variants in total, including bi-allelic SNPs and indels, multi-allelic SNPs and indels), there was an extremely high computing resource requirement to establish models using all the features. Instead, we built new GP models using all the GBS variants, including 13,265 bi- and multi-allelic SNPs and indels in total. We found that three morphological and two chemical traits had significantly improved $r^2_{i,CV}$ in the integrated models compared with all four models built using GBS bi-allelic SNPs, bi-allelic indels, multi-allelic SNPs, and multi-allelic indels separately (**Fig. S9A**, for $r^2_{i,test}$, r^2_p and r^2_g, see **Fig. S9B** and **Fig. S10**). These results show that the GP accuracy can be improved for some traits by integrating different types of genetic variants.

Insights of underlying molecular mechanisms by interpreting GP models

To get insights of molecular mechanisms underlying trait variations, we examined the absolute values of the coefficients of genetic variants in the trained GP models. We focused on the flowering time-related trait anthesis date 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**). For testing how well GP 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 by a model was positively correlated with the number of variants used in the model (**Fig. 6A**, **Fig. S11**, **Table S7-S15**). Due to the large differences in the number of variants, exome capture-based models identified dramatically larger numbers of FT-genes than GBS-based models, regardless of whether top variants above 99th or 95th percentile were investigated (**Fig. 6A**). The models built using exome capture multi-allelic indels identified the largest number of FT-genes (106 and 314 genes when 99th and 95th percentile were used as cutoff thresholds, respectively). Only models built using exome capture bi-allelic SNPs and exome capture multi-allelic indels identified significantly more FT-genes than random guessing (*p*-value for fisher's exact test < 0.034, red bars in **Fig. 6B**). Furthermore, we found that the number of FT-genes identified from model coefficients is not correlated with model performance ($r^2_{i,CV}$), at the 95th (PCC=0.56, *p*-value=0.15) nor 99th percentile (PCC=0.45, *p*-value=0.27) threshold. Taken together, these results suggest that the exome capture variants were superior to GBS variants not only in predicting trait values, but also in identifying variants within or nearby genes which are potential key regulators of the trait in question.

Example flowering time genes that are important for predicting anthesis date

Next, we explored several example variants in detail, which were important to the prediction of anthesis date and were associated with FT-genes or not. In the model built using GBS bi-allelic SNPs, there were four SNPs located within or nearby FT-genes and with the absolute coefficient above 95th percentile (**Table S7**). The most important FT-gene-associated SNP (Chr01K_35695336) was located within the genic region of *Pavir.1KG328000*, an ortholog of Arabidopsis *EARLY FLOWERING IN SHORT DAYS, EFS*), and ranked 26th in terms of absolute coefficient. The Arabidopsis *EFS* represses flowering in the autonomous promotion pathway (Soppe *et al.*, 1999). Individuals with homozygous reference alleles for Chr01K_35695336 flowered significantly later than individuals with homozygous alternative and heterozygous alleles (**Fig. 6C**). In contrast, individuals with different alleles of the SNP Chr05K_49855616, whose absolute coefficient ranked 1st, didn't show significant differences in anthesis date (**Fig. 6D**). More examples for the top FT-gene-associated variants and the variants ranked 1st for other models were exemplified in **Fig. S12**.

An example multi-allelic variant important for GP is the locus Chr08K_52608078, which had six alleles in the exome capture data among individuals studied here: the reference allele is

CAA, and five alternative alleles include one SNP: AAA, and four indels: C, CA, CAAAA, and * (CAA is completely absent). The locus Chr08K_52608078 was located 67 bp upstream of *Pavir.8KG380900*, an ortholog of the Arabidopsis flowering time gene *VERNALIZATION INDEPENDENCE 3, VIP3*). *VIP3* functions as an activator of the flowering-repressor gene *FLOWERING LOCUS C* (Zhang *et al.*, 2003). In the model built using exome capture multi-allelic indels, Chr08K_52608078_* (ranked 15th, **Fig. 6E**), Chr08K_52608078_CA (ranked 102nd, **Fig. 6F**), and Chr08K_52608078_C (ranked 35069th, **Fig. 6G**) had absolute coefficients above the 95th percentile, and individuals with different alleles of these three indels had significantly different anthesis date. When examining the original allele compositions of this locus among individuals (**Fig. 6H**), we found that there were 39 combinations of all these 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 analysis highlights the genetic complexity in switchgrass polyploid that can be revealed through interpreting GP models, and multi-allelic variants should also be more thoroughly evaluated to gain better understanding of the molecular mechanism underlying trait variations.

Conclusions

Before the genomic era, the breeding process relied heavily on natural variation and hybridization, and was time-consuming. In the genomic era, the molecular variant assisted breeding technologies, especially genomic selection or genomic prediction, have dramatically accelerated the breeding processes for both animals and plants (Asoro *et al.*, 2011; Crossa *et al.*, 2017). Considering the relatively higher cost for WGS for a large breeding population, GBS and exome capture are commonly used to gain the genome-wide variants among individuals. To optimize the utility of these two types of genetic data in the GP applications for switchgrass, we evaluated the effects of multiple factors on the GP accuracy for 20 traits in a switchgrass diversity panel.

Our study showed that the genome assembly had minor effects on GP, similar to what has been found in another published paper (Benevenuto *et al.*, 2019). Exome capture variants were superior to GBS variants in terms of both GP accuracy and ability to identify potential contributing genes underlying trait variations, which is potentially due to the larger number of exome capture variants and the higher proportion of exome capture variants that are located within or proximity

to genic regions. How well the exome capture variants perform compared with WGS variants in GP applications needs further exploration for populations with both types of genomic data. The bi-allelic indels and multi-allelic variants are comparable or even superior to bi-allelic SNPs in GP when balanced for the number of variants. In addition, due to the large number of exome capture multi-allelic indels in this study, models built using exome capture multi-allelic indels identified the largest number of FT-genes in switchgrass, suggesting the necessity to explore this type of variants in GP practices. We also found that populations with different ploidy levels had diverse population structures and GP accuracy, suggesting that caution should be taken when building GP models for populations with mixed ploidy levels. Furthermore, the GP accuracy can be improved by integrating different types of variants, as exemplified by GBS variants in this study.

In summary, we assessed the effects of several factors on GP accuracy in switchgrass. Besides the factors investigated in this study, there remain several other factors that have been assessed somewhere else. These factors include structural variation (Chen *et al.*, 2021), other omics data besides genome (Wang *et al.*, 2023), genomic prediction algorithms (Azodi *et al.*, 2019), training population size (Alemu *et al.*, 2024), etc. These findings provided insights on the applications of GP in future crop breeding practices.

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 exome capture 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 exome capture 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 exome capture 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 exome capture 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. (https://github.com/zhanxw/rvtests) 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 exome capture multi-allelic SNPs, and 1,341,337 exome capture multi-allelic indels.

Table 1. Encoding for multi-allelic variants with A as the reference allele.

	Sequence	Column_1	Column_2	Column_3
ID_1	A/A	A/A	A/A	A/A
ID_2	A/T	A/T	A/A	A/A
ID_3	T/G	A/T	A/G	A/A
ID_4	G/AT	A/A	A/G	A/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 (Azodi *et al.*, 2019; Lipka *et al.*, 2014). 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 exome capture bi-allelic SNPs), the one with more variants (i.e., exome capture 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 Scientific Research Foundation of and the Major Scientific Research Tasks from Kunpeng Institute of Modern Agriculture at Foshan (KIMAQD2022003 and KIMA-ZDKY2022004 to PW), and National Natural Science Foundation of China (32370241 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; color scale in the heatmap: $-\log_{10}$ *p*-value from Wilcoxon signed-rank test; white color: *p*-value ≥ 0.05; red and blue: 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-based models than GBS-based models. (**A**) Differences in $r^2_{i,CV}$ between exome capture- and GBS-based models. The column "Unbalanced": models built using all exome capture and GBS bi-allelic SNPs; "Balanced": models built using balanced exome capture (down-sampled to match the number of GBS variants) and GBS bi-allelic SNPs; "Genic": models built using balanced exome capture and GBS bi-allelic SNPs that were located within the gene bodies; "Intergenic": models built using balanced exome capture and GBS bi-allelic SNPs that were located within intergenic regions. Values in the heatmap: $r^2_{i,CV}$ differences × 100; color scale in the heatmap: $-\log_{10}$ *p*-value from Wilcoxon signed-rank test; white color: *p*-value ≥ 0.05; red and blue: higher $r^2_{i,CV}$ was observed for the exome capture- and GBS-based models, respectively. (**B**) Distribution of the GBS (blue) and exome capture (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 exome capture bi-allelic SNPs (pink line). z-score values of the GBS-based numbers in the distribution of exome capture-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; color scale in the heatmap: $-\log_{10}$ *p*-value from Wilcoxon signed-rank test; white color: *p*-value ≥ 0.05; red: higher $r^2_{i,CV}$ was observed for the bi-allelic SNP-based models; blue: 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 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 exome capture bi-allelic SNPs were used. Values in the heatmap: $r^2_{i,CV}$ differences × 100; color scale in the heatmap: $-\log_{10}$ *p*-value from Wilcoxon signed-rank test; white color: *p*-value ≥ 0.05; red and blue: 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 exome capture 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 exome capture (right panel) bi-allelic SNPs were used.

Fig. 6 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 variants based models; pink: exome capture-based models. (**C**) 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. Black font: the top variant that was located within or adjacent to switchgrass genes orthologous to known flowering time genes; gray font: non-flowering time gene-associated variant that had the highest absolute coefficient in the model. The absolute coefficient rank of variants were listed in the parenthesis on the right of the variants. *EFS*: *EARLY FLOWERING IN SHORT DAYS*; *VIP3*: *VERNALIZATION INDEPENDENCE 3*. Yellow, red, and blue bars in (**C-G**): individuals with homologous alternative (-1), heterozygous (0) and homologous reference (1) alleles, respectively. *p*-values were from Wilcoxon signed-rank test. In (**H**), 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.

Supplemental figure legends

Fig. S1 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.

Fig. S2 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 exome capture bi-allelic SNPs. Color scale in the heatmap: median $r^2_{g,CV}$ among 10 replicate runs.

Fig. S3 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 exome capture bi-allelic SNPs. Color scale in the heatmap: $r^2_{i,CV} \times 100$.

Fig. S4 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 exome capture bi-allelic SNPs. Color scale in the heatmap: median $r^2_{p,test}$ among 10 replicate runs.

Fig. S5 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 exome capture bi-allelic SNPs. Color scale in the heatmap: median $r^2_{g,test}$ among 10 replicate runs.

Fig. S6 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 exome capture bi-allelic SNPs. Color scale in the heatmap: $r^2_{i,test} \times 100$.

Fig. S7 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 exome capture data to v5 assembly.

Fig. S8 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 exome 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. S9 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. S10 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. S11 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.

Fig. S12 Anthesis dates of individuals with different alleles. The type of variants which were used to build the models were listed above the figures. Black font: the top variant that was located within or adjacent to switchgrass genes orthologous to known flowering time genes; gray font: non-flowering time gene-associated variant that had the highest absolute coefficient in the model. The absolute coefficient rank of variants were listed in the parenthesis on the right of the variants. Yellow, red, and blue bars: individuals with homologous alternative (-1), heterozygous (0) and homologous reference (1) alleles, respectively. *p*-values were from Wilcoxon signed-rank

test. Error bar: standard error. *COL5*: *CONSTANS-LIKE 5*; *DCL1*: *DICER-LIKE 1*; *VIP4*: *VERNALIZATION INDEPENDENCE 4*; *PRMT10*: *PROTEIN ARGININE METHYLTRANSFERASE 10*; *FBH4*: *FLOWERING BHLH 4*.

References

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

Alemu, A., Åstrand, J., Montesinos-López, O.A., Isidro Y Sánchez, J., Fernández-Gónzalez, J., Tadesse, W., et al. (2024) Genomic selection in plant breeding: Key factors shaping two decades of progress. *Molecular Plant*, **17**, 552–578.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Asoro, F.G., Newell, M.A., Beavis, W.D., Scott, M.P., and Jannink, J. (2011) Accuracy and Training Population Design for Genomic Selection on Quantitative Traits in Elite North American Oats. *The Plant Genome*, **4**, 132–144.

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

Azodi, C.B., Bolger, E., McCarren, A., Roantree, M., de los Campos, G., and Shiu, S.-H. (2019) Benchmarking Parametric and Machine Learning Models for Genomic Prediction of Complex Traits. *G3 Genes|Genomes|Genetics*, **9**, 3691–3702.

Azodi, C.B., Pardo, J., VanBuren, R., de los Campos, G., and Shiu, S.-H. (2020) Transcriptome-Based Prediction of Complex Traits in Maize. *The Plant Cell*, **32**, 139–151.

Batista, L.G., Mello, V.H., Souza, A.P., and Margarido, G.R.A. (2022) Genomic prediction with allele dosage information in highly polyploid species. *Theoretical and Applied Genetics*, **135**, 723–739.

Benevenuto, J., Ferrão, L.F.V., Amadeu, R.R., and Munoz, P. (2019) How can a high-quality genome assembly help plant breeders? *GigaScience*, **8**, giz068.

Biová, J., Kaňovská, I., Chan, Y.O., Immadi, M.S., Joshi, T., Bilyeu, K., and Škrabišová, M. (2024) Natural and artificial selection of multiple alleles revealed through genomic analyses.

Frontiers in Genetics, **14**, 1320652.

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

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

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., et al. (2009) Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree. *Genetics*, **182**, 375–385.

Chen, L., Pryce, J.E., Hayes, B.J., and Daetwyler, H.D. (2021) Investigating the Effect of Imputed Structural Variants from Whole-Genome Sequence on Genome-Wide Association and Genomic Prediction in Dairy Cattle. *Animals*, **11**, 541.

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

Desta, Z.A. and Ortiz, R. (2014) Genomic selection: genome-wide prediction in plant improvement. *Trends in Plant Science*, **19**, 592–601.

Dufresne, F., Stift, M., Vergilino, R., and Mable, B.K. (2014) Recent progress and challenges in population genetics of polyploid organisms: an overview of current state‐of‐the‐art molecular and statistical tools. *Molecular Ecology*, **23**, 40–69.

Emms, D.M. and Kelly, S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, **20**, 238.

Endelman, J.B. (2011) Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome*, **4**, 250–255.

Evans, J., Crisovan, E., Barry, K., Daum, C., Jenkins, J., Kunde‐Ramamoorthy, G., et al. (2015) Diversity and population structure of northern switchgrass as revealed through exome capture sequencing. *The Plant Journal*, **84**, 800–815.

Fawcett, J.A. and Van De Peer, Y. (2010) Angiosperm polyploids and their road to evolutionary success. *Trends in Evolutionary Biology*, **2**, 3.

Isidro, J., Jannink, J.-L., Akdemir, D., Poland, J., Heslot, N., and Sorrells, M.E. (2015) Training set optimization under population structure in genomic selection. *Theoretical and Applied*

Genetics, **128**, 145–158.

Jannink, J.-L., Lorenz, A.J., and Iwata, H. (2010) Genomic selection in plant breeding: from theory to practice. *Briefings in Functional Genomics*, **9**, 166–177.

Kriaridou, C., Tsairidou, S., Houston, R.D., and Robledo, D. (2020) Genomic Prediction Using Low Density Marker Panels in Aquaculture: Performance Across Species, Traits, and Genotyping Platforms. *Frontiers in Genetics*, **11**, 124.

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

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997 [q-bio]*.

Lipka, A.E., Lu, F., Cherney, J.H., Buckler, E.S., Casler, M.D., and Costich, D.E. (2014) Accelerating the Switchgrass (Panicum virgatum L.) Breeding Cycle Using Genomic Selection Approaches. *PLoS ONE*, **9**, e112227.

Lopez, B.I.M., An, N., Srikanth, K., Lee, S., Oh, J.-D., Shin, D.-H., et al. (2021) Genomic Prediction Based on SNP Functional Annotation Using Imputed Whole-Genome Sequence Data in Korean Hanwoo Cattle. *Frontiers in Genetics*, **11**, 603822.

Lorenz, A.J. and Smith, K.P. (2015) Adding Genetically Distant Individuals to Training Populations Reduces Genomic Prediction Accuracy in Barley. *Crop Science*, **55**, 2657–2667.

Lorenzana, R.E. and Bernardo, R. (2009) Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theoretical and Applied Genetics*, **120**, 151–161.

Lovell, J.T., MacQueen, A.H., Mamidi, S., Bonnette, J., Jenkins, J., Napier, J.D., et al. (2021) Genomic mechanisms of climate adaptation in polyploid bioenergy switchgrass. *Nature*, **590**, 438–444.

Lu, F., Lipka, A.E., Glaubitz, J., Elshire, R., Cherney, J.H., Casler, M.D., et al. (2013) Switchgrass Genomic Diversity, Ploidy, and Evolution: Novel Insights from a Network-Based SNP Discovery Protocol. *PLoS Genetics*, **9**, e1003215.

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

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

Norman, A., Taylor, J., Edwards, J., and Kuchel, H. (2018) Optimising Genomic Selection in Wheat: Effect of Marker Density, Population Size and Population Structure on Prediction Accuracy. *G3 Genes|Genomes|Genetics*, **8**, 2889–2899.

Ramsey, J. and Schemske, D.W. (2002) Neopolyploidy in Flowering Plants. *Annual Review of Ecology and Systematics*, **33**, 589–639.

Ramstein, G.P., Evans, J., Kaeppler, S.M., Mitchell, R.B., Vogel, K.P., Buell, C.R., and Casler, M.D. (2016) Accuracy of Genomic Prediction in Switchgrass (Panicum virgatum L.) Improved by Accounting for Linkage Disequilibrium. *G3 (Bethesda, Md.)*, **6**, 1049–1062.

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

Scheet, P. and Stephens, M. (2006) A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. *The American Journal of Human Genetics*, **78**, 629–644.

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

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

VanRaden, P.M., Van Tassell, C.P., Wiggans, G.R., Sonstegard, T.S., Schnabel, R.D., Taylor, J.F., and Schenkel, F.S. (2009) Invited Review: Reliability of genomic predictions for North American Holstein bulls. *Journal of Dairy Science*, **92**, 16–24.

Veerkamp, R.F., Bouwman, A.C., Schrooten, C., and Calus, M.P.L. (2016) Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein–Friesian cattle. *Genetics Selection Evolution*, **48**, 95.

Wang, H., Bernardo, A., St. Amand, P., Bai, G., Bowden, R.L., Guttieri, M.J., and Jordan, K.W. (2023) Skim exome capture genotyping in wheat. *The Plant Genome*, **16**, e20381.

Wang, P., Lehti-Shiu, M.D., Lotreck, S., Abá, K.S., Krysan, P.J., and Shiu, S.-H. (2023) Prediction of plant complex traits via integration of multi-omics data. *bioRxiv*, DOI: 10.1101/2023.11.14.566971.

Zhang, H., Ransom, C., Ludwig, P., and 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., and Liu, X. (2019) Factors Affecting the Accuracy of Genomic Selection for Agricultural Economic Traits in Maize, Cattle, and Pig Populations. *Frontiers in Genetics*, **10**, 189.

Fig. 1

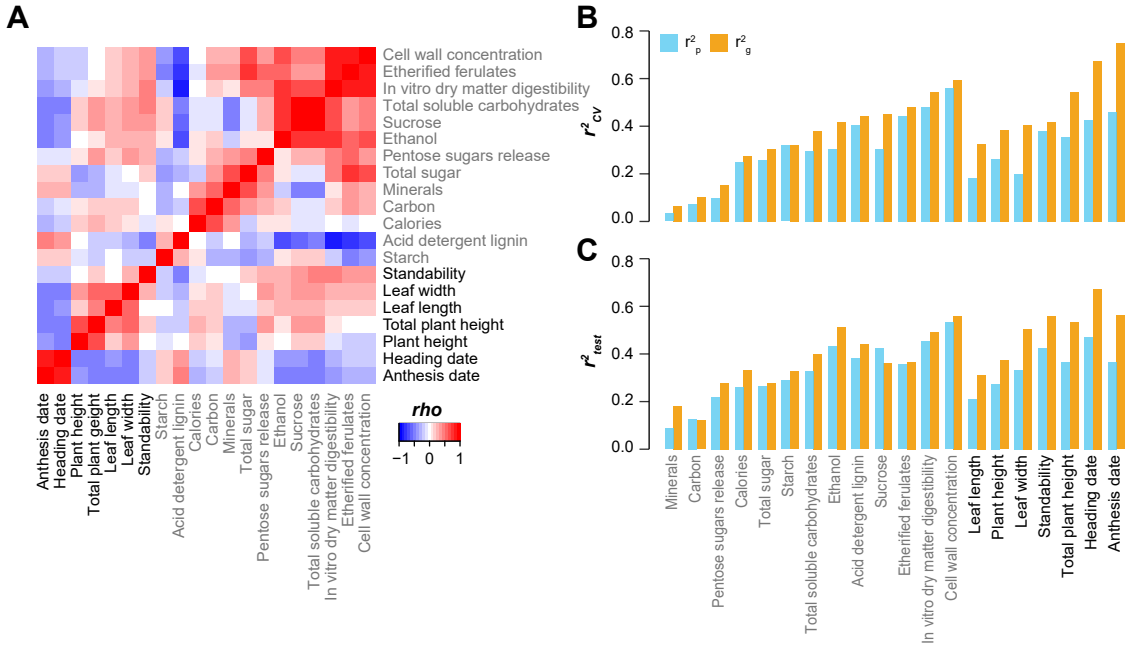

A
Cell wall concentration
Etherified ferulates
In vitro dry matter digestibility
Total soluble carbohydrates
Sucrose
Ethanol
Pentose sugars release
Total sugar
Minerals
Carbon
Calories
Acid detergent lignin
Starch
Standability
Leaf width
Leaf length
Total plant height
Plant height
Heading date
Anthesis date
rho
−1 0 1
B
r^2_p r^2_g
r^2_{CV}
0.0 0.2 0.4 0.6 0.8
C
r^2_{test}
0.0 0.2 0.4 0.6 0.8
Minerals
Carbon
Pentose sugars release
Calories
Total sugar
Starch
Total soluble carbohydrates
Ethanol
Acid detergent lignin
Sucrose
Etherified ferulates
In vitro dry matter digestibility
Cell wall concentration
Leaf length
Plant height
Leaf width
Standability
Total plant height
Heading date
Anthesis date

Fig. 2

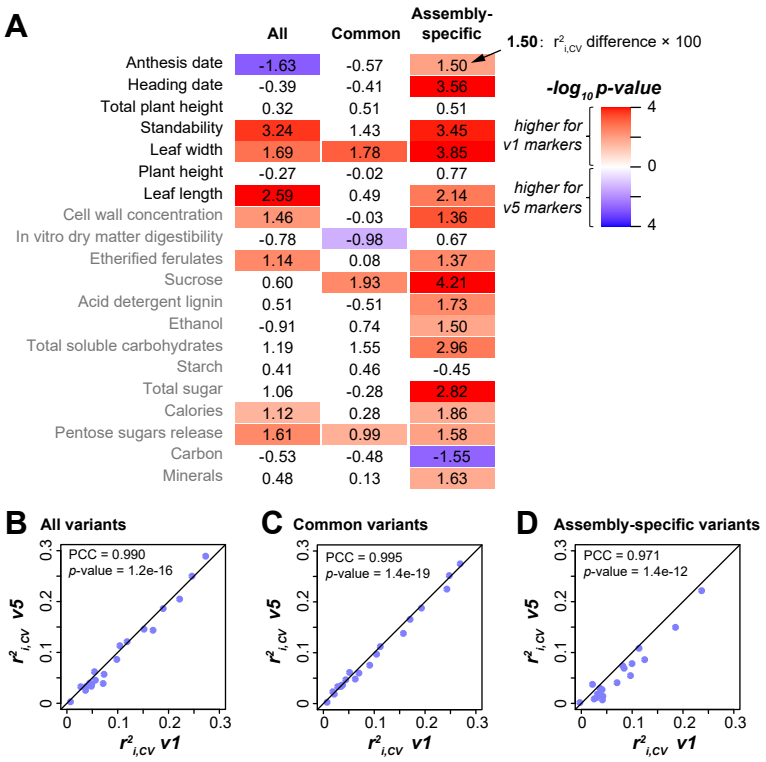

A
All
Common
Assembly-specific
1.50: r²i,CV difference × 100
Anthesis date -1.63 -0.57 1.50
Heading date -0.39 -0.41 3.56
Total plant height 0.32 0.51 0.51
Standability 3.24 1.43 3.45
Leaf width 1.69 1.78 3.85
Plant height -0.27 -0.02 0.77
Leaf length 2.59 0.49 2.14
Cell wall concentration 1.46 -0.03 1.36
In vitro dry matter digestibility -0.78 -0.98 0.67
Etherified ferulates 1.14 0.08 1.37
Sucrose 0.60 1.93 4.21
Acid detergent lignin 0.51 -0.51 1.73
Ethanol -0.91 0.74 1.50
Total soluble carbohydrates 1.19 1.55 2.96
Starch 0.41 0.46 -0.45
Total sugar 1.06 -0.28 2.82
Calories 1.12 0.28 1.86
Pentose sugars release 1.61 0.99 1.58
Carbon -0.53 -0.48 -1.55
Minerals 0.48 0.13 1.63
-log10 p-value
higher for v1 markers
higher for v5 markers
4
0
4
B All variants
PCC = 0.990
p-value = 1.2e-16
C Common variants
PCC = 0.995
p-value = 1.4e-19
D Assembly-specific variants
PCC = 0.971
p-value = 1.4e-12
r²i,cv v5
r²i,cv v1
0
0.1
0.2
0.3

Fig. 3

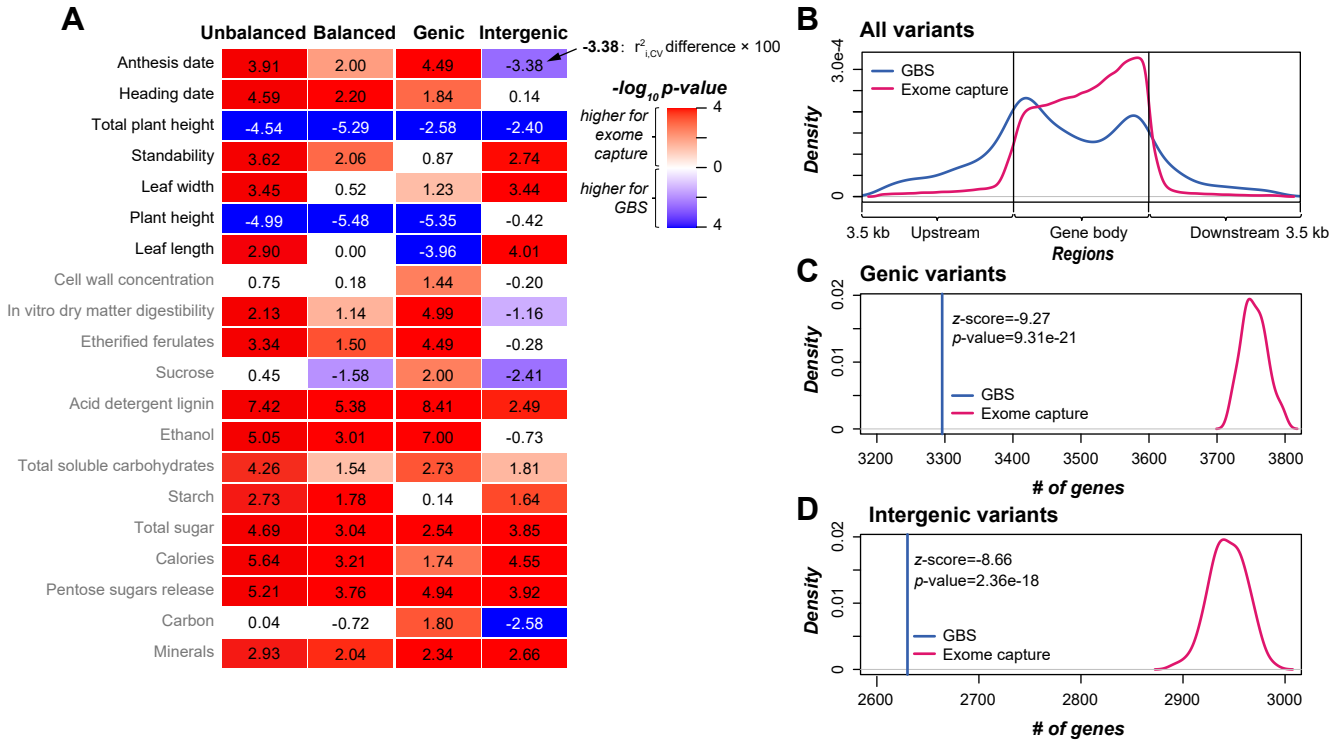

A
Unbalanced Balanced Genic Intergenic
Anthesis date 3.91 2.00 4.49 -3.38
Heading date 4.59 2.20 1.84 0.14
Total plant height -4.54 -5.29 -2.58 -2.40
Standability 3.62 2.06 0.87 2.74
Leaf width 3.45 0.52 1.23 3.44
Plant height -4.99 -5.48 -5.35 -0.42
Leaf length 2.90 0.00 -3.96 4.01
Cell wall concentration 0.75 0.18 1.44 -0.20
In vitro dry matter digestibility 2.13 1.14 4.99 -1.16
Etherified ferulates 3.34 1.50 4.49 -0.28
Sucrose 0.45 -1.58 2.00 -2.41
Acid detergent lignin 7.42 5.38 8.41 2.49
Ethanol 5.05 3.01 7.00 -0.73
Total soluble carbohydrates 4.26 1.54 2.73 1.81
Starch 2.73 1.78 0.14 1.64
Total sugar 4.69 3.04 2.54 3.85
Calories 5.64 3.21 1.74 4.55
Pentose sugars release 5.21 3.76 4.94 3.92
Carbon 0.04 -0.72 1.80 -2.58
Minerals 2.93 2.04 2.34 2.66
-3.38: r²i,cv difference × 100
-log10 p-value
higher for exome capture
higher for GBS
4
0
4
B
All variants
GBS
Exome capture
Density
3.0e-4
0
3.5 kb
Upstream
Gene body
Downstream 3.5 kb
Regions
C
Genic variants
z-score=-9.27
p-value=9.31e-21
GBS
Exome capture
Density
0.02
0.01
0
3200 3300 3400 3500 3600 3700 3800
# of genes
D
Intergenic variants
z-score=-8.66
p-value=2.36e-18
GBS
Exome capture
Density
0.02
0.01
0
2600 2700 2800 2900 3000
# of genes

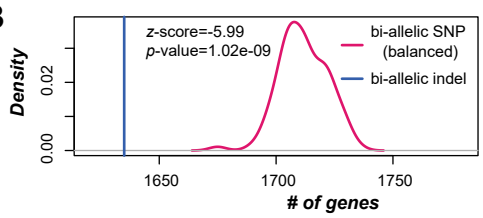
Fig. 4

A

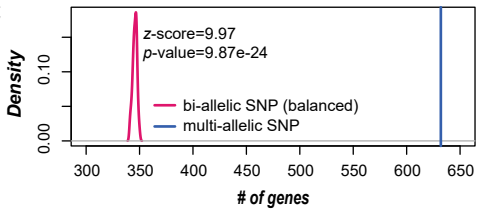
	bi-allelic SNP vs bi-allelic indel		bi-allelic SNP vs multi-allelic SNP	
	Unbalanced	**Balanced**	**Unbalanced**	**Balanced**
Anthesis date	7.70	1.15	11.03	-5.53
Heading date	4.73	-1.65	8.24	-7.03
Total plant height	10.10	4.35	5.66	-7.93
Standability	2.87	1.00	0.56	-3.67
Leaf width	9.31	3.32	8.54	-5.20
Plant height	6.67	2.59	7.49	-2.06
Leaf length	2.01	-2.51	4.12	-6.27
Cell wall concentration	-0.16	-2.10	1.66	-1.67
In vitro dry matter digestibility	0.78	-2.58	5.58	-0.20
Etherified ferulates	2.47	0.41	2.32	-1.38
Sucrose	7.57	2.06	13.07	2.34
Acid detergent lignin	2.45	-0.06	1.20	-3.53
Ethanol	3.77	-0.93	9.41	0.60
Total soluble carbohydrates	1.89	-2.05	4.48	-3.08
Starch	-0.06	-0.35	-1.46	-2.70
Total sugar	5.10	3.31	0.32	-3.63
Calories	4.98	3.39	2.70	-0.50
Pentose sugars release	3.01	0.76	4.64	-0.81
Carbon	1.07	-0.63	2.05	-1.17
Minerals	-0.84	-1.85	1.36	-1.24

-1.24: $r^2_{i,CV}$ difference × 100

B



C



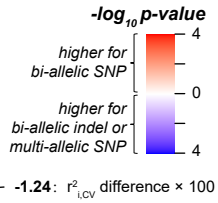


Fig. 5

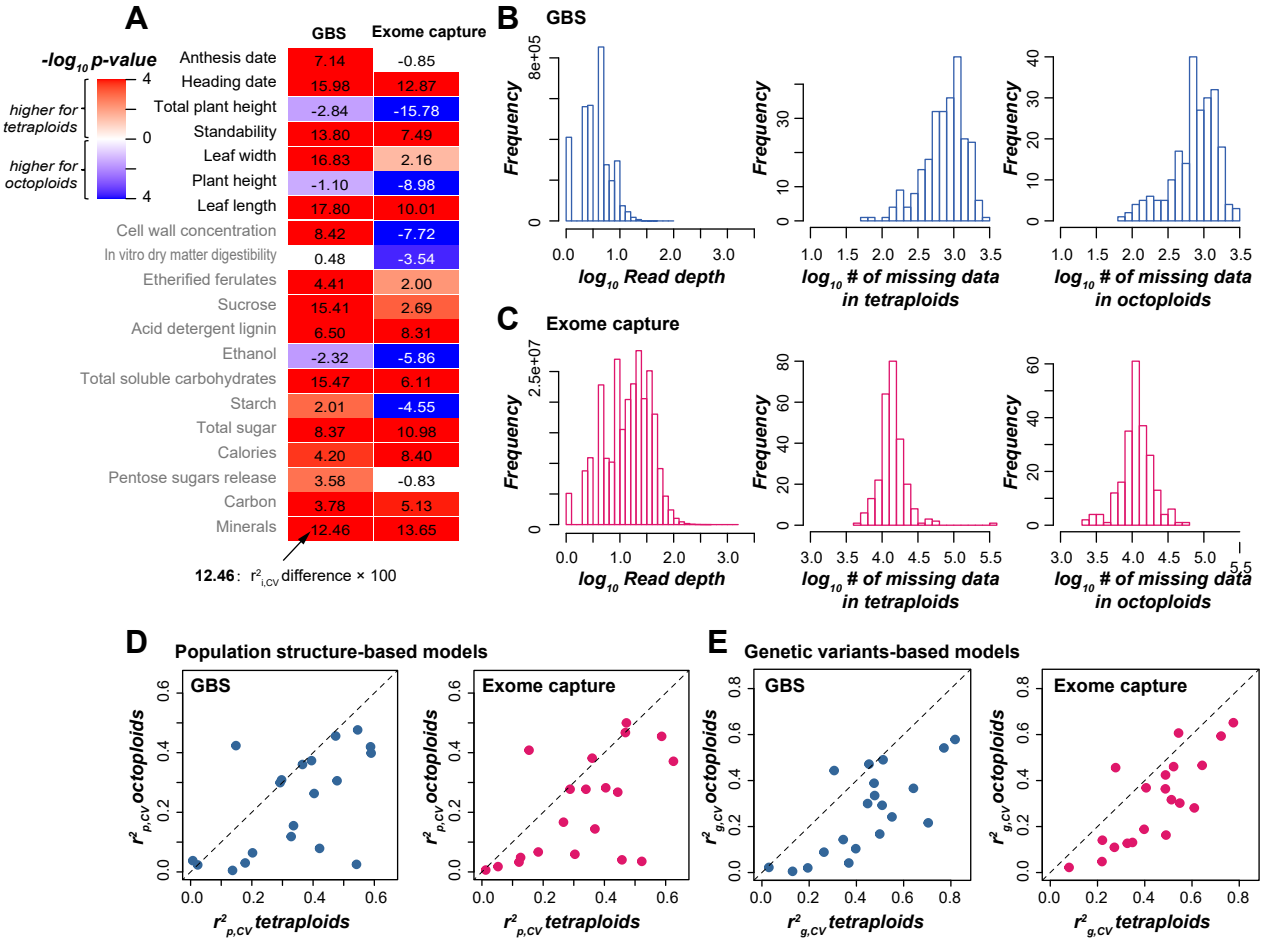
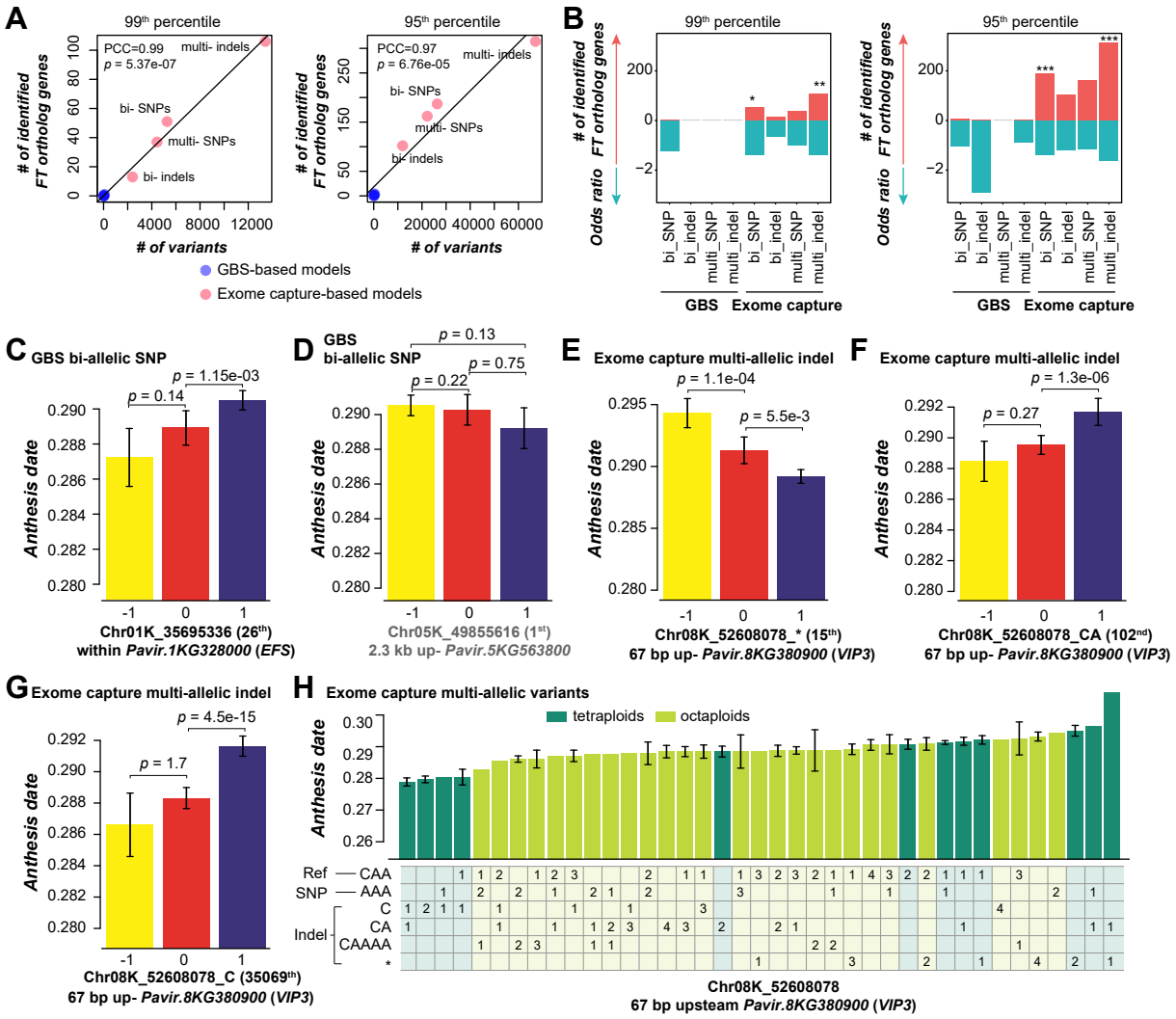

A
-log10 p-value
higher for tetraploids
higher for octoploids
GBS
Exome capture
Anthesis date 7.14 -0.85
Heading date 15.98 12.87
Total plant height -2.84 -15.78
Standability 13.80 7.49
Leaf width 16.83 2.16
Plant height -1.10 -8.98
Leaf length 17.80 10.01
Cell wall concentration 8.42 -7.72
In vitro dry matter digestibility 0.48 -3.54
Etherified ferulates 4.41 2.00
Sucrose 15.41 2.69
Acid detergent lignin 6.50 8.31
Ethanol -2.32 -5.86
Total soluble carbohydrates 15.47 6.11
Starch 2.01 -4.55
Total sugar 8.37 10.98
Calories 4.20 8.40
Pentose sugars release 3.58 -0.83
Carbon 3.78 5.13
Minerals 12.46 13.65
12.46: r2 i,CV difference × 100
B GBS
Frequency
log10 Read depth
log10 # of missing data in tetraploids
log10 # of missing data in octoploids
C Exome capture
D Population structure-based models
GBS
Exome capture
r2 p,CV octoploids
r2 p,CV tetraploids
E Genetic variants-based models
r2 g,CV octoploids
r2 g,CV tetraploids

Fig. 6


A
99th percentile
PCC=0.99
p = 5.37e-07
multi- indels
bi- SNPs
multi- SNPs
bi- indels
# of identified FT ortholog genes
# of variants
95th percentile
PCC=0.97
p = 6.76e-05
GBS-based models
Exome capture-based models
B
Odds ratio
bi_SNP
bi_indel
multi_SNP
multi_indel
GBS
Exome capture
C GBS bi-allelic SNP
p = 1.15e-03
p = 0.14
Anthesis date
Chr01K_35695336 (26th)
within Pavir.1KG328000 (EFS)
D GBS bi-allelic SNP
p = 0.13
p = 0.75
p = 0.22
Chr05K_49855616 (1st)
2.3 kb up- Pavir.5KG563800
E Exome capture multi-allelic indel
p = 1.1e-04
p = 5.5e-3
Chr08K_52608078_* (15th)
67 bp up- Pavir.8KG380900 (VIP3)
F Exome capture multi-allelic indel
p = 1.3e-06
p = 0.27
Chr08K_52608078_CA (102nd)
67 bp up- Pavir.8KG380900 (VIP3)
G Exome capture multi-allelic indel
p = 4.5e-15
p = 1.7
Chr08K_52608078_C (35069th)
67 bp up- Pavir.8KG380900 (VIP3)
H Exome capture multi-allelic variants
tetraploids
octaploids
Ref
SNP
Indel
CAA
AAA
C
CA
CAAAA
*
Chr08K_52608078
67 bp upsteam Pavir.8KG380900 (VIP3)