Wu *et al. BMC Genomics* (2022) 23:200
https://doi.org/10.1186/s12864-022-08337-7

BMC Genomics

RESEARCH Open Access

Improvement of prediction ability by integrating multi-omic datasets in barley

Po-Ya Wu[1], Benjamin Stich[1,2], Marius Weisweiler[1], Asis Shrestha[1], Alexander Erban[3], Philipp Westhoff[2,4] and Delphine Van Inghelandt[1*]

Abstract

Background: Genomic prediction (GP) based on single nucleotide polymorphisms (SNP) has become a broadly used tool to increase the gain of selection in plant breeding. However, using predictors that are biologically closer to the phenotypes such as transcriptome and metabolome may increase the prediction ability in GP. The objectives of this study were to (i) assess the prediction ability for three yield-related phenotypic traits using different omic datasets as single predictors compared to a SNP array, where these omic datasets included different types of sequence variants (full-SV, deleterious-dSV, and tolerant-tSV), different types of transcriptome (expression presence/absence variation-ePAV, gene expression-GE, and transcript expression-TE) sampled from two tissues, leaf and seedling, and metabolites (M); (ii) investigate the improvement in prediction ability when combining multiple omic datasets information to predict phenotypic variation in barley breeding programs; (iii) explore the predictive performance when using SV, GE, and ePAV from simulated 3'end mRNA sequencing of different lengths as predictors.

Results: The prediction ability from genomic best linear unbiased prediction (GBLUP) for the three traits using dSV information was higher than when using tSV, all SV information, or the SNP array. Any predictors from the transcriptome (GE, TE, as well as ePAV) and metabolome provided higher prediction abilities compared to the SNP array and SV on average across the three traits. In addition, some (di)-similarity existed between different omic datasets, and therefore provided complementary biological perspectives to phenotypic variation. Optimal combining the information of dSV, TE, ePAV, as well as metabolites into GP models could improve the prediction ability over that of the single predictors alone.

Conclusions: The use of integrated omic datasets in GP model is highly recommended. Furthermore, we evaluated a cost-effective approach generating 3'end mRNA sequencing with transcriptome data extracted from seedling without losing prediction ability in comparison to the full-length mRNA sequencing, paving the path for the use of such prediction methods in commercial breeding programs.

Keywords: Barley, Deleterious SV, Transcriptome, Metabolome, Genomic prediction, Omic prediction

Background

Barley (*Hordeum vulgare* L.) is the fourth most important cereal crop in the world (FAOSTAT, http://www.fao.org/faostat/en/) and is used for human nutrition and animal feed [1]. In the context of a growing global population [2], producing sufficient food is a big challenge for agriculture [3]. In addition, climate change is expected to negatively impact global crop production by increasing extreme temperatures and altering rainfall patterns [4]. Thus, high and stable yield in barley is one of the most important breeding goals. However, in addition to directly breeding for yield, the consideration of yield-related characters during the breeding processes proved successful [5]. Leaf angle (LA) e.g. is one of the most important canopy architecture

*Correspondence: inghelan@hhu.de
[1]Institute of Quantitative Genetics and Genomics of Plants, Heinrich Heine University, 40225 Düsseldorf, Germany
Full list of author information is available at the end of the article

parameters that influence the efficiency of photosynthesis and further affect yield production [6]. In addition, the control of plant height (PH) can be used to reduce yield loss arising from lodging and adaption to variable environments through heading time (HT) alteration impacts yield [7]. Therefore, the use of approaches that help breeders to reliably select for yield and yield-related quantitative traits increases the gain of selection.

Genomic prediction (GP) has emerged as a powerful tool to increase selection gain for complex traits in both livestock and plant breeding programs [8, 9]. This method is based on the idea that the performance of individuals can be predicted from genotypic information using the GP model which was trained on those individuals with both phenotypic and genotypic information. Thus, the genotyped individuals can be preselected before their phenotypes are measured in the field to shorten the breeding cycle as well as to reduce the cost of phenotyping [10].

Typically, single nucleotide polymorphisms (SNP) serve as predictors in GP [11–13]. SNP in gene coding regions can be classified into non-synonymous (nsSNP) and synonymous SNP (sSNP), which differ in their property to change or not the amino acid sequence of a protein. Therefore, these two SNP classes may have different influence on phenotypes. In previous studies, the advantage of using these classes of SNP in comparison to randomly selected SNP for GP was explored in pig [14]. However, they observed that predictive performance of neither nsSNP nor sSNP did significantly differ from those of random SNP for most traits. In addition, Heidaritabar et al. [15] observed that nsSNP did not enhance the performance of GP in chicken. On the other hand, a protein may be able to tolerate an amino acid change due to a nsSNP and still keep its function normal [16]. Therefore, SNP can be grouped using the SIFT algorithm [17] into (1) tolerant SNP (tSNP), which can keep a protein's function normal; and (2) deleterious SNP (dSNP), which will affect a protein's function. To the best of our knowledge, the use of tSNP or dSNP as predictor of the phenotypic variation has not yet been compared.

Complex biological processes such as transcription, translation, and biochemical cascades resulting in various metabolites occur between DNA sequence and phenotypes [11], which hamper the predictive power of SNP. In addition, higher-order epistatic effects may contribute to the genetic variance of complex traits [18], which can in most of cases not directly be captured using SNP information [13, 19]. Therefore, prediction ability of phenotypic variation using SNP information for quantitative traits still leaves room for improvement. In the last years, molecular technologies were developed, which allow a cheap and high-throughput gene expression and metabolite profiling [20]. Such data can act as bridge to shorten the biological distance between genotypes and phenotypes and may even capture higher-order epistatic interactions for the prediction of phenotypic variation [21, 22].

Transcription is the first downstream processes after the DNA sequence and, thus, more likely affects the variation of traits compared to SNP. Recently, thanks to technological developments, several studies have proposed to use gene expression (GE) variation as predictor of phenotypic variation in maize [11, 21], rice [22] and barley [23]. While Schrag et al. [21] and Hu et al. [22] used GE assessed from microarray experiments for GP and showed that a considerable proportion of phenotypic variation can be explained by such information, Guo et al. [11] and Weisweiler et al. [23] used mRNA sequencing datasets to predict the performance of phenotypic traits. The advantage of mRNA sequencing compared to microarray experiment is the possibility to extract SNP and small insertions/deletions (INDEL) called sequence variants (SV hereafter), in addition to the quantification of transcript abundance. Furthermore, a single gene can often produce more than one transcript through alternative splicing, which can generate various proteins to regulate the complexity of pathways [24]. These different transcripts of the same gene can be identified using full-length mRNA sequencing. To our knowledge, transcript expression (TE) as predictor in GP has not yet been compared to GE.

Compared to the two previous levels of molecular information (DNA sequence and GE), metabolites (M) have the closest relationship to the expressed phenotype because they are the end-points of upstream biochemical processes [25], and, thus, have a high potential as predictors for GP. Previous studies on the use of metabolites to predict phenotypic traits in *Arabidopsis thaliana*, maize, wheat, and barley reported lower or higher prediction abilities compared to SNP information, depending on the traits and species [11, 21, 26–29]. Gemmer et al. [29] recommended that metabolites cannot be used alone in barley for phenotype prediction. However, the integration of expression and metabolite datasets with SNP information improved prediction abilities in comparison to the benchmark using SNP information in maize [11, 21]. Thus, the integration of several layers of omic datasets such as SV, GE, TE, and M as predictors could outperform benchmark methods and should be evaluated in GP of phenotypic traits in barley.

The objectives of our study were to (i) assess the prediction ability for three yield-related phenotypic traits (LA, PH, and HT) using different omic datasets as single predictors compared to a SNP array, where these omic datasets included different types of sequence variants (SV, dSV, and tSV), different types of transcriptome (expression presence/absence variation-ePAV, GE, and TE) sampled from two tissues, leaf and seedling, and metabolites (M); (ii) investigate the improvement in prediction ability

when combining multiple omic datasets information to predict phenotypic variation in barley breeding programs; (iii) explore the predictive performance when using SV, GE, and ePAV from simulated 3'end mRNA sequencing of different lengths as predictors.

Results

Heritability

The three phenotypic traits (LA, PH, and HT) were measured for 23 spring barley inbreds in seven environments. The adjusted entry means of the 23 inbreds ranged from 2.52 to 7.07 for LA, 48.75 to 79.75 cm for PH, and 57.31 to 82.23 days for HT (Suppl. Table S1). Heritabilities on an entry mean basis (H^2) were high and similar for LA (0.91) and HT (0.90) and with 0.83 slightly lower for PH. A total of 192 chemical entities were annotated (Suppl. Table S2) and after filtering (see methods), 144 metabolites remained for which the relative abundances were used for further analyses. A total of 101 metabolites were found in databases and, thus, it was possible to assign them according to their chemical features to 12 compound classes, while the remaining 43 metabolites were unknown (Suppl. Table S3). The heritabilities of the metabolites on an entry mean basis ranged from 0 to 0.98 with an average of 0.62 (Suppl. Fig. S1). The classification of the metabolic predictors using different degrees of heritability (0.1 to 0.8 in increments of 0.1) resulted in eight groups with 133, 128, 121, 117, 109, 93, 72 and 45 metabolites, respectively. These groups were then considered for the omic prediction described below.

Correlation and genetic dissimilarity analyses

Positive correlations between the three phenotypic traits were observed (Suppl. Fig. S2). Particularly, LA was highly and significantly correlated with HT (0.685***), where the correlation coefficients between PH and HT as well as between PH and LA were with about 0.45 considerably lower. Many metabolites were significantly ($P < 0.05$) negatively associated with the assessed phenotypic traits (Fig. 1). For instance, a cluster of some acids, amino acids, and several unknown metabolites was strongly negatively correlated with the three traits. Interestingly, we found that the same metabolites that were significantly correlated with LA were also correlated with HT. This was consistent with the phenotypic correlations between both traits (Fig. 1 and Suppl. Fig. S2).

To assess similarity/dissimilarity between these omic datasets, we performed generalized procrustes analysis (GPA) [30] on the resulting principal component analysis (PCA) obtained from each omic dataset. The dissimilarity measurements from GPA were used for principal coordinates analysis (PCoA). The first two PCo accounted for 71.86% and 20.72% of the total variability, respectively (Fig. 2). The first PCo separated the metabolites from the other features while the second PCo tended to differentiate the two tissues, leaf (l) and seedling (s). GE, TE, and ePAV datasets were similar to each other within the same tissue. This can be explained thereby that the ePAV dataset was derived from GE dataset and the GE dataset was derived from the TE dataset. $ePAV_{ls}$ was, as expected, centered between the ePAV from the individual tissues. Although SNP array, SV, dSV, and tSV clustered together, SNP array was more distant from the cluster of dSV, tSV, and SV which almost overlapped. This was due to that dSV and tSV are a subset of SV. This finding indicated that SNP, expression and metabolite features would provide different layers of biological information and might contribute differently and complementarily to the phenotypic variation.

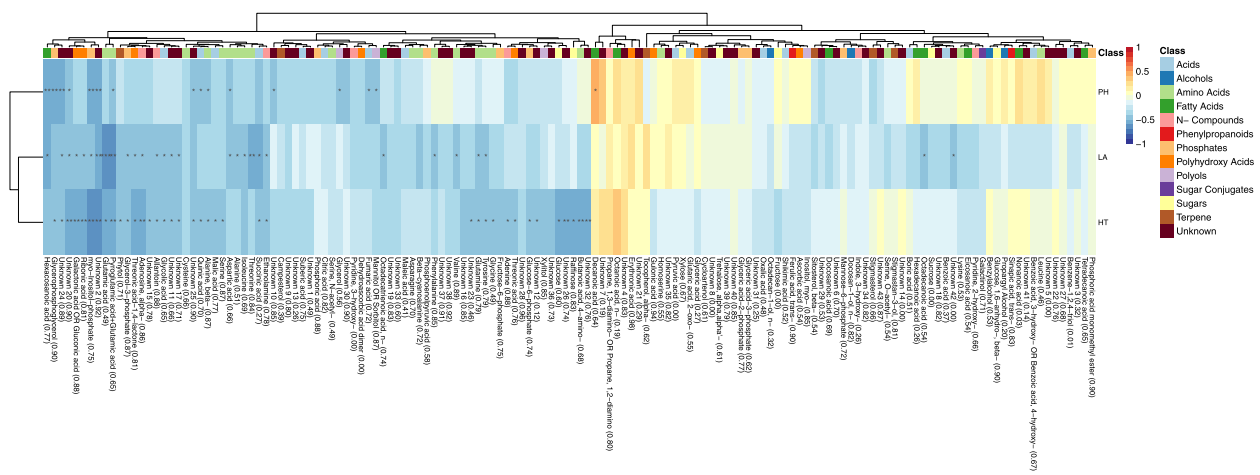

Fig. 1 Heatmap of Pearson correlation coefficients calculated between all pairs of the three phenotypic traits and the 144 metabolites. The three phenotypic traits are leaf angle (LA), plant height (PH) and heading time (HT). Correlations marked with *, **, and *** were significant at $P < 0.05$, 0.01, and 0.001, respectively. The heritability of each metabolite is given in parentheses after each metabolite's name

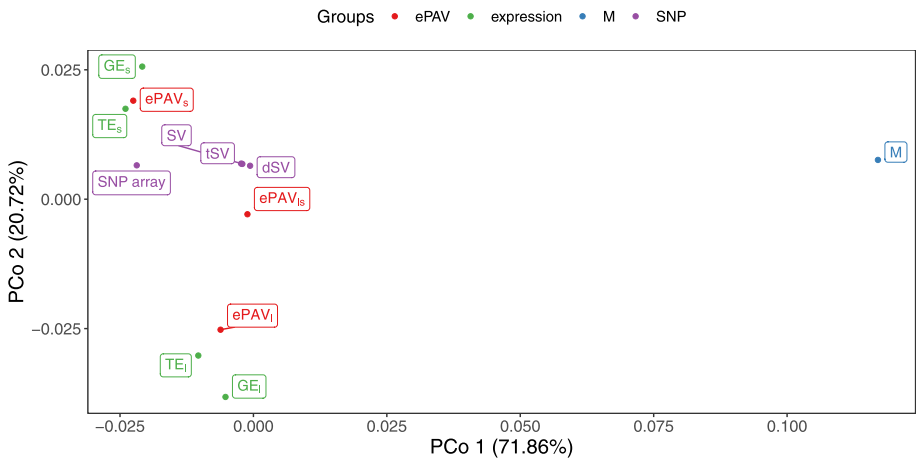

Fig. 2 Plot of the first two axes of the principal coordinate analysis for comparison of the similarity between different omic datasets based on generalized procrustes analysis. The omic datasets include SNP array, sequence variants (SV), deleterious sequence variants (dSV), tolerant sequence variants (tSV), gene expression in seedling and leaf (GE_s and GE_l), transcript expression in seedling and leaf (TE_s and TE_l), expression presence/absence variation in seedling, leaf and combining both tissues ($ePAV_s$, $ePAV_l$, and $ePAV_{ls}$), and metabolites (M). The colors show the four groups of omic datasets used in a grid search for integration of multiple predictors (Figure 5). Red represents ePAV, green expression, blue metabolite, and purple SNP and SV predictors

Omic prediction

The prediction ability of the three phenotypic traits using different single predictors was examined through five-fold cross-validation. Regardless of the predictor, the prediction abilities were higher for traits with higher heritabilities (Fig. 3). Prediction abilities based on SV, GE, TE, ePAV, and M datasets were compared to that realized with the SNP array which was used as baseline predictor. The observed median prediction ability based on the SNP array dataset ranged from 0.185 (HT) to 0.590 (LA). The prediction ability of SV extracted from mRNA sequencing dataset was slightly higher than that of SNP array dataset across the three traits. Moreover, the dSV dataset slightly outperformed the SV extracted from mRNA sequencing and the tSV dataset (Fig. 3). Even higher prediction abilities were observed for ePAV, any expression datasets from seedling (GE_s and TE_s), and metabolite datasets (Fig. 3). The prediction abilities for the ePAV dataset were significantly different among *l*, *s* and *ls*, but not consistently across the three traits (data not shown). $ePAV_{ls}$ was chosen as the best compromise across the three traits for further analyses, as it was for none of the three traits in the significance group with the lowest prediction abilities. The TE datasets slightly outperformed the GE datasets for HT and LA, and TE_s resulted in the highest prediction ability as single predictor for these traits. In contrast, no difference between TE and GE was observed for PH.

To explore whether the heritability of a metabolite affects the prediction performance, eight classes of metabolites based on different degrees of heritabilities served as predictor. The prediction ability increased when the metabolites with lower heritability (< 0.1) were not considered (Fig. 3). However, the prediction ability didn't increase significantly and consistently across the three traits with increasing heritability of the considered metabolites (data not shown). Therefore, we selected the metabolite group for which the highest prediction ability was observed across the three traits ($M_{0.6}$) for further analyses.

Pearsons correlation coefficients between pairwise predicted values of different omic datasets were calculated, and the correlation-based distance was used for PCoA analysis for each trait. Across the three examined traits, the metabolite feature was clearly separated from the other omics features (Fig. 4), and the predicted values of M were less correlated with those values of the other omic datasets than the other omic datasets among themselves (Suppl. Fig. S3). A similar result was observed between the two tissues, seedling and leaf, which were clearly separated on Fig. 4. In contrast, the predicted values from features that clustered together on Fig. 4, especially SNP array, SV, dSV, tSV, $ePAV_{ls}$, were highly correlated (Suppl. Fig. S3).

In order to evaluate whether the prediction ability can be improved by combining several predictors, a joined weighted relationship matrix of the single predictors with the highest prediction ability was established and a grid search was used to identify those combinations of dSV, $ePAV_{ls}$, TE_s, and $M_{0.6}$ resulting in the highest prediction ability. For the three examined traits, the highest median prediction ability was observed when more than one predictor was used (Fig. 5). Furthermore, the optimal weights of the four predictors to reach the maximal prediction ability differed among the three traits, but the weights of

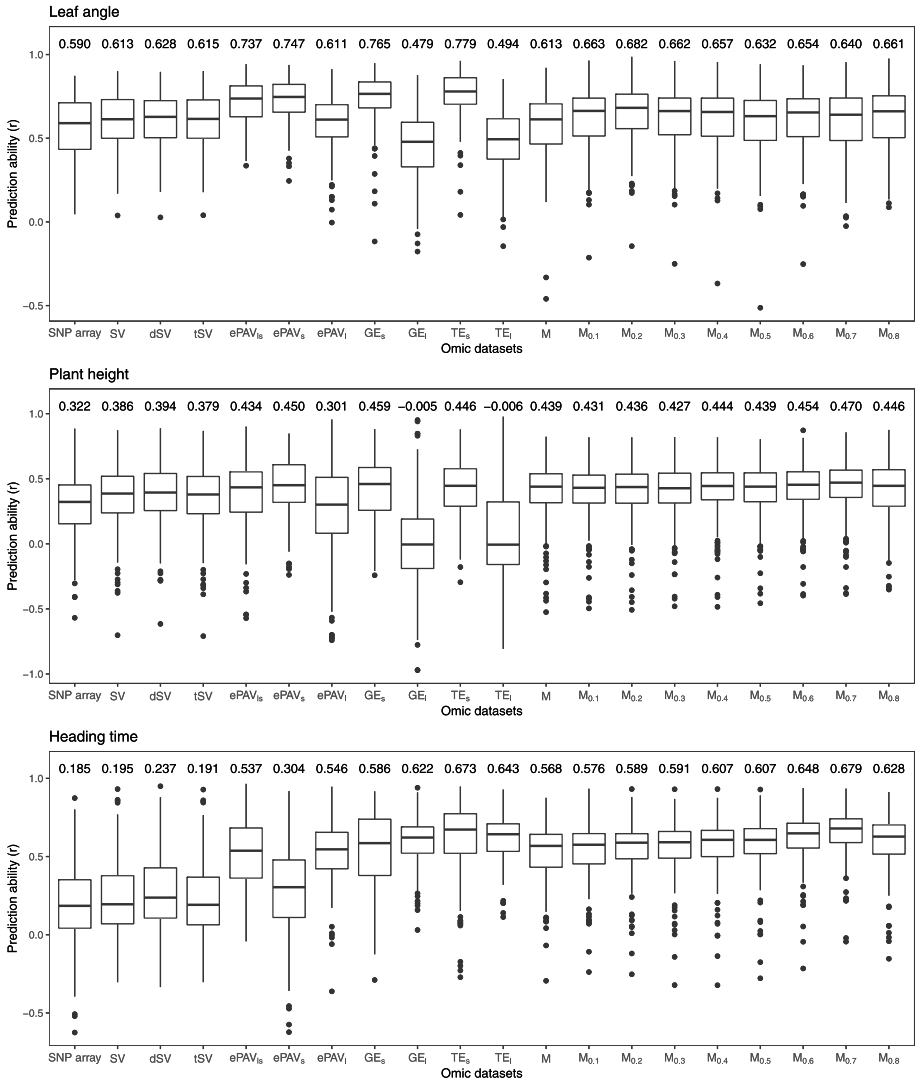

Fig. 3 Boxplot of prediction abilities for the three traits, leaf angle, plant height and heading time, based on 22 inbreds using different omic datasets as a single predictor across 200 five-fold cross-validation runs. The values given above each box represent the medians of 200 runs. The omic datasets include SNP array, sequence variants (SV), deleterious sequence variants (dSV), tolerant sequence variants (tSV), gene expression in seedling and leaf (GE_s and GE_l), transcript expression in seedling and leaf (TE_s and TE_l), expression presence/absence variation in seedling, leaf and combining both tissues ($ePAV_s$, $ePAV_l$, and $ePAV_{ls}$), metabolites filtered for their heritability (M, $M_{0.1}$, $M_{0.2}$, $M_{0.3}$, $M_{0.4}$ $M_{0.5}$, $M_{0.6}$, $M_{0.7}$, and $M_{0.8}$)

$ePAV_{ls}$ and TE_s were at least 10% and 50%, respectively. However, the optimal weight for M was, except for PH, 0, and the optimal weight for the dSV was 0 for the three traits.

We also assessed the prediction abilities of SV, GE, ePAV from 3'end mRNA sequencing that we simulated from our full-length mRNA sequencing dataset. Depending on the trait, a similar, slightly better or worse median of prediction abilities of SV, GE, ePAV were observed when considering 3'end mRNA sequencing compared to a full-length mRNA sequencing dataset as baseline (Fig. 6). Moreover, we did not observe a systematic trend on the prediction ability when increasing the length of the 3'end mRNA sequencing.

Discussion

Ability of different omic features to predict phenotypic traits

Genomic prediction has become a broadly used tool to improve the gain of selection in plant breeding [9]. The current standard procedure of genomic prediction is to use SNP markers generated from SNP array or genotyping by sequencing methods as predictors [12]. However, there are several complicated biological downstream pro-

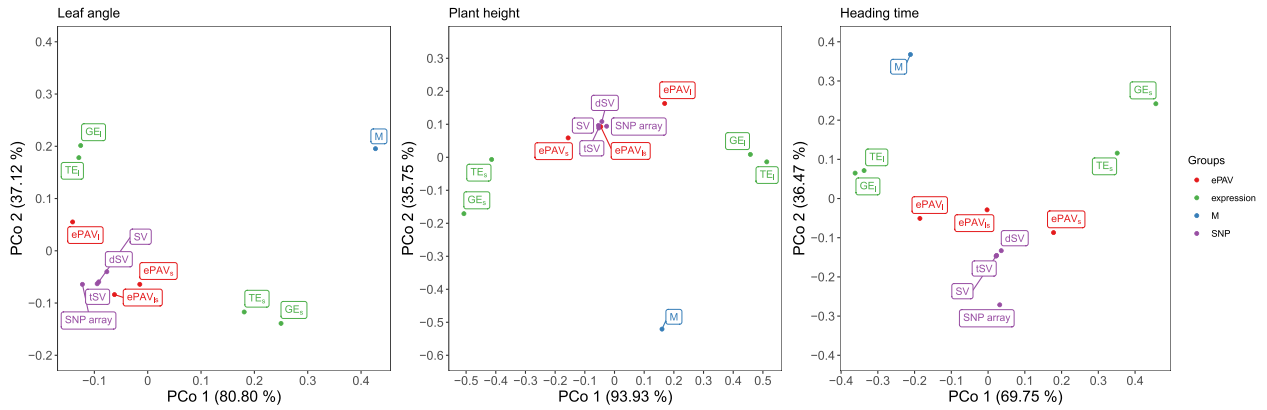

Fig. 4 Plot of the first two axes of principal coordinates analysis for comparison of predicted values from different omic datasets as a single predictor based on the median of correlation-based distance across 200 five-fold cross-validation runs for the three traits, leaf angle, plant height and heading time. The omic datasets include SNP array, sequence variants (SV), deleterious sequence variants (dSV), tolerant sequence variants (tSV), gene expression in seedling and leaf (GE_s and GE_l), transcript expression in seedling and leaf (TE_s and TE_l), expression presence/absence variation in seedling, leaf and combining both tissues ($ePAV_s$, $ePAV_l$, and $ePAV_{ls}$), and metabolites (M). The colors show the four groups of omic datasets used in a grid search for integration of multiple predictors (Figure 5). Red represents ePAV, green expression, blue metabolite, and purple SNP and SV predictors

cesses such as transcription, translation, and biochemical cascades resulting in various metabolites between DNA sequences and phenotypes [11]. Using predictors that are biologically closer to the phenotypes may increase the prediction ability in genomic predictions. With the development of high-throughput molecular technologies, the availability of such predictors from the genomic, transcriptomic, or metabolomic level is ensured [20]. In this pilot study, we aim to compare different types of omic datasets for their predictive performance in order to prioritize them for their later evaluation in large-scale experiments. We hold that this is true also with only 23 inbreds of our study, especially as these inbreds are representative of and cover most of the genotypic diversity of barley [23].

For the three examined traits, any of the SV information generated from mRNA sequencing (SV, dSV, as well as tSV) resulted in a higher prediction ability compared to the SNP data produced with the 50K SNP array (Fig. 3). This might be explained by the higher number of SV features, as increasing the number of predictors can increase the extent of linkage disequilibrium between SNP and quantitative trait loci (QTL) [23, 31]. In addition, INDEL information was included in the SV, which was not the case in the SNP array. INDEL are one type of genetic variation in living organisms that involve larger DNA fragments than single variants and have been identified in known genes (c.f. [32, 33]). Therefore, they are very usefull for the developpment of functional markers [34] and

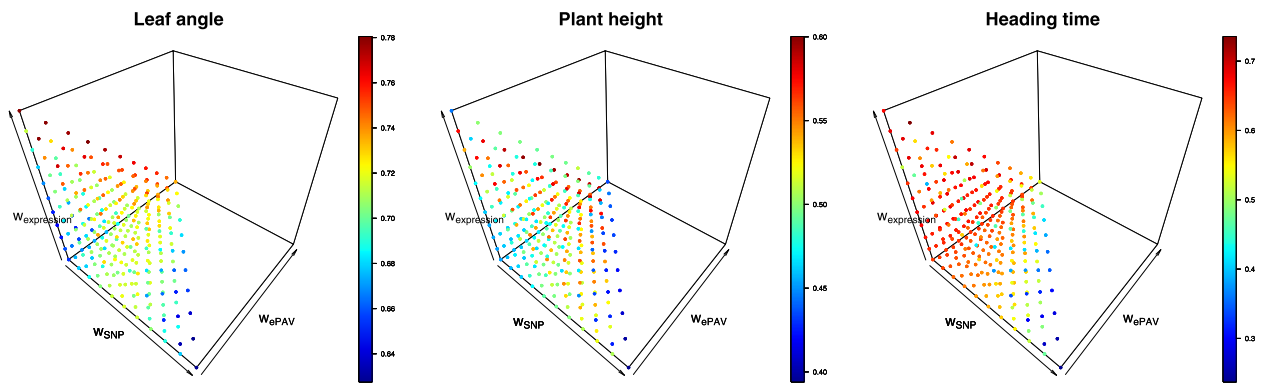

Fig. 5 Prediction ability for the three traits, leaf angle, plant height, and heading time, from 22 inbreds for 286 combinations of the joined weighted matrix which differ in the weights of four predictors, deleterious sequence variants (dSV), expression presence/absence variation in combined leaf and seedling ($ePAV_{ls}$), transcript expression in seedling (TE_s), and metabolite with a heritability on an entry mean basis >0.6 ($M_{0.6}$). Plotted values represent medians across 200 cross-validation runs

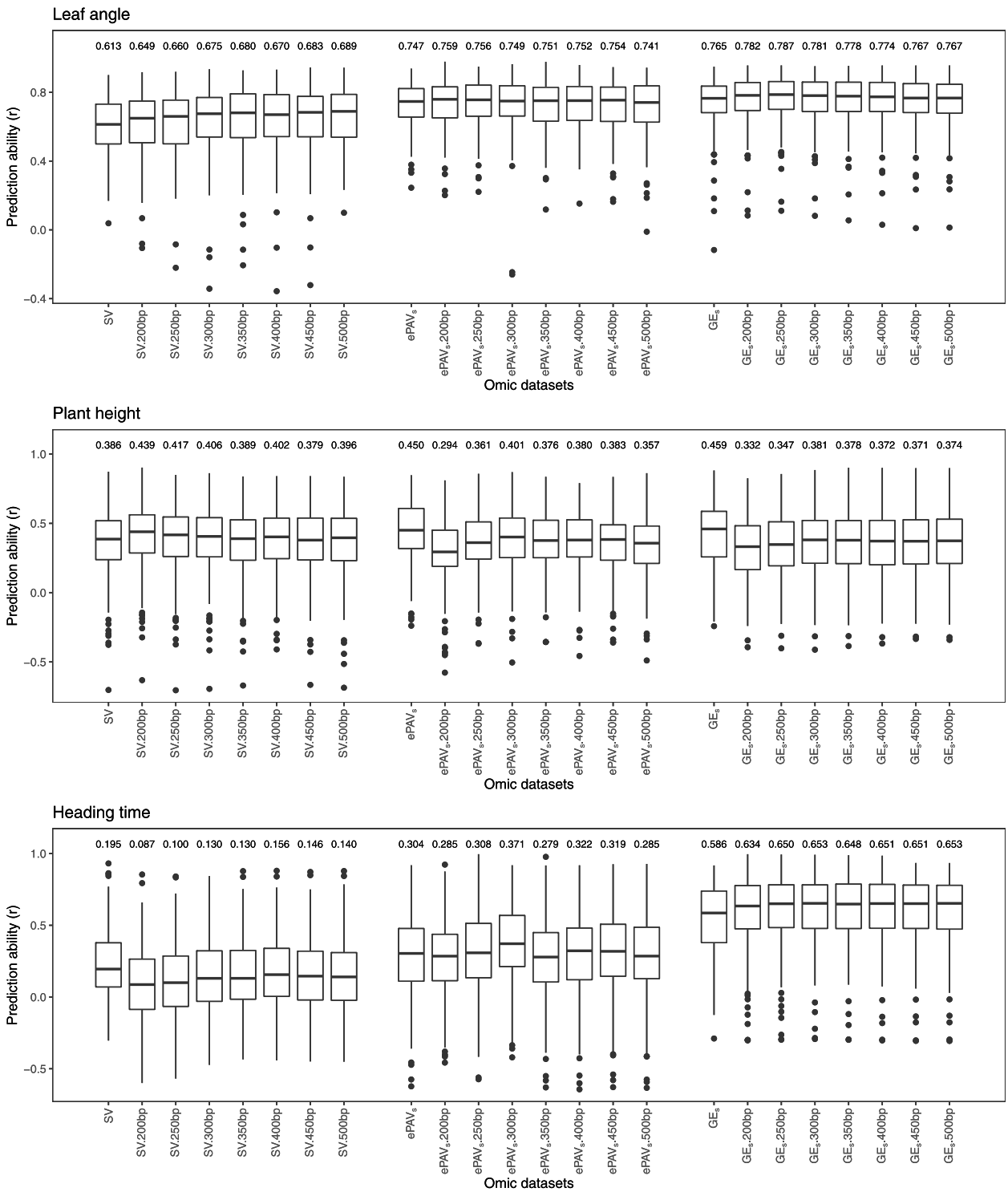

Fig. 6 Boxplot of prediction ability for the three traits, leaf angle, plant height and heading time, from 22 inbreds using different omic datasets from simulated 3'end mRNA sequencing with seven length categories (200, 250, 300, 350, 400,450, and 500 bp) as a single predictor across 200 cross-validation runs. The omic datasets from full-length mRNA sequencing are used as a baseline. The values given above each box represent medians of 200 prediction abilities. The omic datasets include sequence variants (SV), gene expression in seedling (GE_s), and expression presence/absence variation in seedling ($ePAV_s$)

are expected to cause extreme change in the phenotypes. This could be a further explaination why SV had better predictive performance than SNP array. Our observation is in agreement with the finding that the PCo 1 resulting from the GPA separated clearly SV and SNP array (Fig. 2), which indicates that SV and SNP array provide different information.

SV in gene coding regions can be classified into nsSV and sSV, where the former can change the amino acid sequence of proteins, but not the latter. However, not all amino acid changes lead to significant changes of the protein. This can be explored by the SIFT algorithm in classifying SV into dSV and tSV based on the conversion of amino acid sequences [16], where the former cause a loss of protein function but not the latter. Kono et al. [35] showed that known phenotype-altering variants were more frequently inferred as deleterious than the genome-wide average, and have a higher probability to contribute to phenotypic variation. Thus, we compared the prediction ability of dSV and tSV compared to that of SV across the three traits.

The predicted phenotypic values based on the three different classes of SV were highly correlated with each other (Suppl. Fig. S3), which can be expected because dSV and tSV are a subset of SV and clustered together in the GPA (Fig. 2). However, the prediction ability for the three phenotypic traits using dSV information was slightly higher than using tSV and all SV information, despite the fact that the number of dSV features was far smaller (15,868) than the number of tSV features (117,698) and the total number of SV. This trend of a higher prediction ability for dSV was even more pronounced when adjusting for differences in the number of features by resampling simulations (data not shown). Our finding is in discordance with the results of Do et al. [14] and Heidaritabar et al. [15], who observed no difference between the prediction performance of nsSNP and randomly sampled SNPs. A first explanation for our different findings could be that the former cited studies classified the SNP based on whether they may induce amino acid change or not, whereas our study distinguished tolerant/deleterious SNP. Secondly, the SNP used for GP by Heidaritabar et al. [15] were imputed for all genotypes from a 60K SNP array. This might have hampered the improvement of prediction ability in comparison to our study, which is based on real variant data for all inbreds (except few missing data that were mean-imputed). Our finding indicated that the preselection of variants based on their theoreticaly predicted protein function could improve prediction performance of traits, which can be of considerable importance for breeders.

The features derived from the transcriptome datasets (GE, TE, as well as ePAV) led to increased prediction abilities by 62.81% compared to SNP array and even SV on average across the three traits and two tissues. This finding was inconsistent with the results of previous studies [11, 21], who observed that the prediction abilities based on transcriptomic datasets were a little lower (5.30% and 0.03%) than those based on genomic information averaged across the examined traits. This difference might be caused by the complex genetic architectures of traits evaluated and tissue sampled in the studies cited above. However, the use of transcriptomic datasets as predictors still had reasonable prediction abilities in the former studies, which is in accordance with our results and can be explained by the fact that with such datasets expression levels can be quantified and physiological epistasis even captured.

A single gene can encode multiple distinct transcripts through alternative splicing, which allows organisms to increase the protein diversity based on the same set of genes [36], and therefore could lead to more phenotypic variation. As a consequence, a higher prediction ability could be expected for phenotypic traits predicted from TE compared to GE information. This was confirmed by our findings (Fig. 3), and suggests that TE information might be more efficient than GE information in predicting the performance of traits when the full-length mRNA sequencing has been performed.

All the datasets generated by mRNA sequencing from seedling were well separated from those from leaf (Fig. 2). Similarly, the correlation between predicted patterns based on the transcriptomic dataset of the two tissues was low (Fig. 4 and Suppl. Fig. S3) , which indicated that different types of tissue offer dissimilar information concerning the phenotypic variation and influence the prediction ability. In general, the prediction ability was considerably higher for the datasets from seedling in comparison with the datasets from leaf on average across the three traits (Fig. 3). This might be explained by the fact that more diverse genes are expressed in seedling than in leaf.

Only for HT, expression information from leaf (GE_l, TE_l) achieved the same level of prediction ability as that from seedling. One explanation for this finding might be that HT is triggered by environmental factors in later developmental stages and therefore the causal expression features for this trait are more likely to be revealed in leaf than in early developmental stages like seedling.

A total of 53 of the 144 metabolites quantified in our study were significantly correlated with at least one of the three phenotypic traits (Fig. 1). This suggests that the metabolites can be used for selection for phenotypes. In addition, the metabolite feature was clearly separated from the other features in the similarity/dissimilarity analysis (Fig. 2). More importantly, the correlations between the predicted values based on metabolic feature and other omic datasets were low, and lower than the correlation between different other omic datasets (Suppl. Fig. S3).

This finding suggested that the metabolites can provide another biological layer of information to capture the phenotypic variation. We observed across the three traits that prediction abilities based on metabolites were considerably higher compared to SNP or SV information (Fig. 3). This finding is in contradiction to results of previous studies [11, 29] who revealed considerably lower prediction ability using metabolites as predictor. This might be caused by the high accuracy of the metabolite assessment used in our study. The average heritability on an entry mean basis across 144 metabolites was with about 0.62 considerably higher than that observed by Guo et al. [11] with 0.49 and Gemmer et al. [29] with 0.26. This aspect was studied further by leaving out those metabolites with heritabilities < 0.1. This resulted in an increased prediction ability for all traits, which suggested that higher accuracy of metabolites can bring stable information in the prediction of phenotypes.

Generally, (di)-similarity between (1) different omic datasets (Fig. 2) and also between (2) the correlation between predicted phenotypic traits based on different omic datasets (Fig. 4 and Suppl. Fig. S3) was observed in our study. This suggested complementation between different biological perspectives to the phenotypic variation. Therefore, combining predictors covering different layers of biological information in an integrative model could have an advantage over the GP model based on single predictors, and was examined in our study.

Increasing prediction abilities by combining multiple predictors

In this study, a grid search was used to identify those combinations of dSV, $ePAV_{ls}$, TE_s, and $M_{0.6}$ in the joined weighted relationship matrix of GBLUP model maximizing the prediction ability. The highest prediction ability across the three examined traits was observed when more than one predictor was used and, for each of the three traits, without the contribution of the dSV (Fig. 5). This finding might be explained by the fact that transcriptome and metabolome information are closer to phenotypes than gene information according to the central dogma of molecular biology, and can capture together more genetic variation and physiological epistasis caused by complicated networks and interactions between genes than when using only one single predictor [11].

On the other hand, even if a higher prediction ability for all three examined traits was observed if more than one predictor was used (Fig. 5), the optimal weight of each component in the joined weighted relationship matrix depended highly on the traits. For instance, metabolite information was needed to obtain the highest prediction ability for PH, but not for the other traits. Transcriptome was the most important component, but the weight ranged from 0.5 to 0.9 across the three traits. From the physiological point of view, this might be explained by the different genetic architectures of the different traits and their exposure to different environments at different developmental stages and tissues. We observed the tendency that for traits with a lower heritability more different omic predictors were needed to result in the highest prediction ability. Further research on traits with high genetic complexity and low heritability such as yield is needed to test this hypothesis.

Summary: application in breeding programs

The results of our study suggested that combining the information of SV, expression, as well as metabolite dataset into genomic prediction models can improve the prediction ability of phenotypic traits. Especially, the expression datasets were the most important components for this improvement (Fig. 5). To be implemented in breeding programs, such datasets have to be created approximately at the costs of one traditional phenotyping unit (c.f. [37]). This implies that the datasets of SV, gene expression, and metabolite are sampled from one tissue, to avoid the cost of multiple sampling at several stages. The goal of this study was to compare predictors for their ability to predict phenotypic traits. The results of our study indicate that the higher and more stable predictive performance across traits can be achieved from gene and transcript expression gained on seedling samples. Seedling samples combine both aptitude in reaching a high prediction ability but can be also generated in a cost-effective and high-throughput manner. Thus, they are recommended as the best tissue to predict the variation of phenotypes in barley populations. However, for other crops such as tuber crops, different approaches and tissues might be needed, which requires further research.

The limited budget available in practical breeding programs for full-length mRNA sequencing hampers the use of such approaches. Instead, 3'end mRNA sequencing could be a cost-effective alternative method to obtain transcriptome information. For 3'end mRNA sequencing, only 50-800bp at the 3'end of the genes are sequenced. Interestingly, we observed that the prediction abilities of SV, GE, ePAV from simulated 3'end mRNA sequencing were on average across the three traits similar to those from the full-length mRNA sequencing (Fig. 6). Therefore, our finding suggested that transcriptome data can be generated from the 3'end mRNA sequencing without losing prediction ability in comparison to the full-length mRNA sequencing, paving the path for the use of such prediction methods in commercial breeding programs.

Although this study is based on a limited number of barley inbreds, it can be considered as a pilot research showing how different omic datasets can improve prediction of phenotypic variation and will open the path to

perform such analysis on a bigger scale, e.g. on segregating populations derived from the 23 inbreds [38].

Materials and methods

Plant materials and phenotypic data collection

This study was based on 23 spring barley inbreds which were selected from a worldwide collection [39] to maximize phenotypic and genotypic diversity [23]. The 23 inbreds were planted as replicated checks in a field experiment laid out as an augmented row-column design. The experiment was performed in seven agro-ecologically diverse environments (Cologne from 2017 to 2019, Mechernich and Quedlinburg from 2018 to 2019) in Germany in which the checks were replicated 10 to 21 times per environment. At each environment, three yield-related phenotypic traits were assessed. The leaf angle (LA) was scored on a scale from 1 (erect) to 9 (very flat) on four-week-old plants. The heading time (HT) was recorded as days after planting. Furthermore, the plant height (PH, cm) was measured after heading (only assessed in Cologne and Mechernich).

Omic datasets

Metabolite profiling

The metabolite profiling of our study was based on leaf samples collected for the 23 barley inbreds with quadruplicates in a greenhouse experiment, where no phenotypic traits were assessed. Seeds of the 23 spring barley inbreds were sown in controlled conditions with 16 hours light and eight hours dark at 22 °C. Plantlets were cultivated for two weeks and then moved to vernalisation in a growth chamber. After five weeks of vernalisation, the plants were repotted and returned to the greenhouse. After one week, one 3 x 1cm piece of the central part of the youngest fully developed leaf was harvested from two plants of the same inbred, pooled, and immediately flash frozen in liquid nitrogen. The collection of all samples was done within one hour to minimize the variation due to circadian rhythms. Each of the 92 samples was analyzed one time via gas chromatography-mass spectrometry (GC-MS) using an adapted protocol from Lisec et al. [40]. Metabolites were extracted from 45-55 mg frozen mortared samples with 1.5 ml of a 1:2.5:1 H_2O:methanol:chloroform (v:v:v) mixture pre-cooled to -20 °C, then mixed on a rotator for 10 min and centrifuged at 20,000 g for 2 min (both at 4 °C). A total of 30 μl of the supernatant were dried completely in a vacuum concentrator and derivatized in two steps via an MPS-Dual-head autosampler (Gerstel): (1) with 10 μl methoxyamine hydrochloride (Acros organics; freshly prepared at 20 mg/ml in pure pyridine (Sigma-Aldrich)) and shaking at 37 °C for 90 min, (2) adding 90 μl N-Methyl-N-(trimethylsilyl)trifluoroacetamide (MSTFA; Macherey-Nagel) and shaking at 37 °C for 30 min. After incubation for 2 hours at room temperature, 1 μl of derivatized compounds was injected at a flow of 1 ml/min with an automatic liner exchange system in conjunction with a cold injection system (Gerstel) in splitless mode (ramping from 50 °C to 250 °C at 12 °C/s) into the GC. Chromatography was performed using a 7890B GC system (Agilent Technologies) with a 30 m long, 0.25 mm internal diameter, HP-5MS column with 5% phenyl methyl siloxane film (Agilent 19091S-433). The oven temperature was held constant at 70°C for 2 min and then ramped at 12.5°C/min to 320°C at which it was held constant for 5 min; resulting in a total run time of 27 minutes.

Metabolites were ionized with an electron impact source at 70V and 200 °C source temperature and recorded in a mass range of m/z 60 to m/z 800 at 20 scans per second with a 7200 GC-QTOF (Agilent Technologies). Raw data files exported from MassHunter Qualitative (v b07, Agilent Technologies) in the mzData format (*mzdata.xml) were converted to the NetCDF format (*.cdf) and baseline-corrected via MetAlign (v 041012, [41]) using default parameters. Baseline-correction was visually inspected using OpenChrom (v 1.3.0, [42]). Quantitative analysis of GC-MS-based metabolite profiling experiments was then performed using TagFinder (v 4.1, [43]). After evaluating the uniqueness and linearity of each fragment, the aggregated fragment intensity was calculated as the average of the maximum scaled fragment intensity. For relative quantification, aggregated fragment intensities of the compounds were normalized to those of the internal standard ribitol (Sigma-Aldrich) which was added to the extraction buffer. Mass spectral annotation was manually supervised using the Golm Metabolome Database mass-spectral library (http://gmd.mpimp-golm.mpg.de/download/) after conversion of absolute time in retention indices [44]. The raw data, details of the quantification and annotation steps, and the processed metabolite profiles are available (https://www.ebi.ac.uk/metabolights/MTBLS1561). The compounds corresponding to contaminations, siloxane, ribitol, and dimethylphenylalanine were removed. Furthermore, if several compounds were identified as the same metabolite, the one with the greatest heritabilty, for which the calculation is described below, was retained.

SNP genotyping, RNA extraction, sequencing, and quantification of gene expression

The Illumina 50K barley SNP array [45] was used to genotype the 23 inbreds of our study [23]. This dataset is designated in the following as SNP array.

mRNA was extracted from leaf and seedling samples of the 23 inbreds as described earlier by Weisweiler et al. [23]. 46 polyA enriched RNA libraries were prepared at the Max Planck Genome Centre Cologne (https://mpgc.mpipz.mpg.de/home/). In addition, two tissue sam-

ples of one of the inbreds and one tissue sample of two other inbreds had to be removed during the data cleaning process. Reads were trimmed, adapter and low quality regions were removed. Afterwards, reads were mapped using HISAT2 (version 2.0.5) [46] to the Morex referene sequence version 1 [47]. Transcript calling was performed with StringTie (version 2.1.3) [48]. Newly identified and annotated genes were included to the dataset as described by Weisweiler et al. [23]. The expression data for the 23 inbreds was separated into gene expression and transcript expression data. The expression quantified as fragments per kilobase of exon model per million fragments mapped (FPKM) was measured for every transcript of a gene, resulting in one FPKM-value per gene and the corresponding FPKM-value for each transcript of a gene. The FPKM-values of genes and transcripts are designated in the following as GE and TE, where the indexes l and s were used to separate the leaf (GE_l, TE_l) and seedling (GE_s, TE_s) samples. For further details see Weisweiler et al. [23].

Determination of ePAV

For each tissue separately, a presence call was made for each inbred-gene combination in the matrix of presence/absence calls, if GE >0 and an absence call if GE = 0. No presence/absence call ("NA") was made for the inbreds with $0< GE <10\%$ of the maximum value of GE for a gene-tissue combination (cf. [49]). Tissue specific ePAV calls were combined to an across tissue ePAV call as described in detail by Weisweiler et al. [23]. The ePAV detection procedure resulted in three ePAV data sets, namely ePAV leaf ($ePAV_l$), ePAV seedling ($ePAV_s$), and one across both tissues ($ePAV_{ls}$).

Sequence variant calling

Variant calling of SNP and small INDEL and their filtering was performed with samtools (version 1.11) and bcftools (version 1.10.2) as described by Weisweiler et al. [23], and the dataset is designed in the following as SV. SIFT4G (version 2.4) was used to annotate and predict tolerant and deleterious variants. The prediction was done based on the conversion of amino acid sequences [16]. Amino acid substitutions were classified according to their effect on the protein functions and were predicted as tolerant if the score was >0.05 and as deleterious if the score was <= 0.05. The SIFT4G database was build based on the uniref 90 database (downloaded 2020/04/29) and the Morex reference sequence version 1 [47] with the tool SIFT4_Create_Genomic_DB.

Simulation of 3'end mRNA sequencing

For the simulation of 3'end mRNA sequencing, GE_s was only measured based on the last 200, 250, 300, 350, 400, 450, and 500 bp at the 3'end of each gene. To the same reduced set of sequence data, the ePAV detection procedure and the SV calling procedure has been applied resulting in seven different GE, ePAV, and SV datasets.

Statistical analyses

Adjusted entry means, variance components, and heritability

Based on visual inspections of quantile-quantile (Q-Q) plots of residuals as well as residuals vs. fitted values plots, phenotypic outliers were removed. Each of the phenotypic traits was then analysed across the environments using the following mixed model:

$$y_{ijk} = \mu + E_j + G_i + (G \times E)_{ij} + \varepsilon_{ijk}, \quad (1)$$

where y_{ijk} was the observed phenotypic value for the i^{th} genotype at the j^{th} environment within the k^{th} replication, μ the general mean, G_i the effect of the i^{th} inbred, E_j the effect of the j^{th} environment, $(G \times E)_{ij}$ the interaction between the i^{th} inbred and the j^{th} environment, and ε_{ijk} the random error. To estimate adjusted entry means for all inbreds, G_i was treated as fixed and the other effects as random. As the samples for metabolites were collected from one environment, the model [1] was reduced to:

$$y_{ik} = \mu + G_i + \varepsilon_{ik}, \quad (2)$$

where y_{ik} was the observed metabolite for the i^{th} inbred within the k^{th} replication, and ε_{ik} the random error. The resulting adjusted entry means of phenotypic traits and metabolites for each inbred were used in further analyses, where the adjusted entry means of metabolites were designated as M.

To estimate the genetic variance (σ_G^2), model (1) and (2) were used but considering G_i as random. The heritability on an entry mean basis for the phenotypic traits and metabolites was then calculated as $H^2 = \sigma_G^2/(\sigma_G^2 + \bar{v}/2)$, where $\bar{v}$ was the mean variance of difference between two adjusted entry means [50].

Prediction of phenotypic traits from multi-omic datasets

The performance to predict phenotypic variation of different types of predictors: (1) SNP array, (2) sequence variants (SV), (3) deleterious sequence variants (dSV), (4) tolerant sequence variants (tSV), (5) $ePAV_s$, (6) $ePAV_l$, (7) $ePAV_{ls}$, (8) gene expression in seedling (GE_s), (9) gene expression in leaf (GE_l), (10) transcript expression in seedling (TE_s), (11) transcript expression in leaf (TE_l), (12) metabolite (M), was compared based on the most stable and widely used model in GP, genomic best linear unbiased prediction (GBLUP) model [51], which can be described as

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zu} + \boldsymbol{\varepsilon} \quad (3)$$

where $\mathbf{y}$ is the vector of the adjusted entry means of the examined trait, $\mathbf{1}$ the unit vector, μ the general mean, $\mathbf{Z}$ the incidence matrix of genotypic effects, and $\mathbf{u}$ the vector of genotypic effects that are assumed be normal

distributed with $N(0, \mathbf{G}\sigma_u^2)$, in which $\mathbf{G}$ denotes the relationship matrix between inbreds and σ_u^2 the genetic variance. In addition, $\boldsymbol{\varepsilon}$ is the vector of residuals following a normal distribution $N(0, \mathbf{I}\sigma_e^2)$. In this study, only additive effects were modeled.

For each of the above mentioned omic dataset, the monomorphic features and the features with missing rates > 0.2 have been filtered out. $\mathbf{W}$ was defined as a matrix of feature measurements for the respective omic dataset that is designated in the following as predictor. The dimensions of $\mathbf{W}$ were the number of barley inbreds (n) times the number of features in the corresponding predictor (m) (Table 1). Because of genotyping problems for one of the inbreds, 22 inbred lines were used for further analyses ($n = 22$).

For each predictor, the additive relationship matrix $\mathbf{G}$ was defined as $\mathbf{G} = \frac{\mathbf{W}^*\mathbf{W}^{*T}}{m}$, where $\mathbf{W}^*$ is a matrix of feature measurement for the respective predictor, whose columns are centered and standardized to unit variance of $\mathbf{W}$, and $\mathbf{W}^{*\mathbf{T}}$ is the transpose of $\mathbf{W}^*$. In addition, to assess the impact of the heritability of a metabolite on the prediction performance, only those metabolites with a heritability on an entry mean basis higher than t, where t varied from 0.1 to 0.8 in increments of 0.1, were considered, and the datasets were designated as $M_{0.1}$, $M_{0.2}$, $M_{0.3}$, $M_{0.4}$ $M_{0.5}$, $M_{0.6}$, $M_{0.7}$ and $M_{0.8}$.

In order to understand whether the different omic datasets can capture similar genetic information, Pearsons correlation coefficients between pairwise predicted values of different omic datasets were calculated. Subsequently, 1 − the correlation coefficients among all pairs of predictors was used as the correlation-based distance in a PCoA. Furthermore, to investigate the performance of a joined weighted relationship matrix [21] to predict phenotypic variation, the matrices $\mathbf{G}$ in model (3) of four predictors were weighted and summed up to one joined weighted relationship matrix, where we varied:

1. the weight of SNP (w_{SNP}): the weight of the most representative SNP datasets was determined as the one from the SNP array, SV, tSV, or dSV which has the most stable prediction performance across the three traits (dSV).
2. the weight of ePAV (w_{ePAV}): the weight of the most representative ePAV datasets was determined as the one from $ePAV_{ls}$, $ePAV_s$, or $ePAV_l$ which has most stable prediction performance across the three traits ($ePAV_{ls}$).
3. the weight of expression ($w_{expression}$): the weight of the most representative of the expression datasets was determined as the one from GE_s, GE_l, TE_s, or TE_l which has most stable prediction performance across the three traits (TE_s).
4. the weight of metabolite (w_M, $1 - w_{SNP} - w_{ePAV} - w_{expression}$): the weight of the most representative metabolite datasets was determined as the one from M, $M_{0.1}$, $M_{0.2}$, $M_{0.3}$, $M_{0.4}$ $M_{0.5}$, $M_{0.6}$, $M_{0.7}$, or $M_{0.8}$ which has most stable prediction performance across the three traits ($M_{0.6}$).

A grid search, varying any weight (w) from 0 to 1 in increments of 0.1, resulted in 286 different combinations of joined weighted relationship matrix, where the summation of four weights in each combination must be equal to 1. In addition, the performance of SV, GE_s, and $ePAV_s$ from simulated 3'end mRNA sequencing of different length as described above was explored.

Five-fold cross-validation was used to assess the model performance. Prediction abilities were obtained by calculating Pearson correlations between observed (y) and predicted ($\hat{y}$) adjusted entry means in the validation set of each fold. The median prediction ability across the five folds within each replicate was calculated and the median of the median across the 200 replicates was used for further analyses.

Table 1 The number of features and the abbreviations for each omic dataset used in this study

Omic dataset	Abbreviation	Number of features
50K SNP array	SNP array	38,285
Sequence variants	SV	133,566
Deleterious sequence variants	dSV	15,868
Tolerant sequence variants	tSV	117,698
Expression presence/absence variation in seedling	$ePAV_s$	27,445
Expression presence/absence variation in leaf	$ePAV_l$	26,653
Expression presence/absence variation in combining leaf and seedling	$ePAV_{ls}$	36,235
Gene expression in seedling	GE_s	67,844
Gene expression in leaf	GE_l	60,888
Transcript expression in seedling	TE_s	250,490
Transcript expression in leaf	TE_l	220,749
Metabolites	M	144

Correlation and genetic similarity analyses

Correlations among the three phenotypic traits, and between the three phenotypic traits and the individual metabolites were measured as Pearson correlation coefficient. Principal component analysis (PCA) was performed on each omic dataset (SNP array, SV, dSV, tSV, $ePAV_s$, $ePAV_l$, $ePAV_{ls}$, GE_l, GE_s, TE_s, TE_l, and M). To evaluate similarity/dissimilarity among the various datasets, generalized procrustes analysis (GPA) [30] was performed

based on the PCA results. Subsequently, 1 − the procrustes similarity indexes among all pairs of omic datasets was used as dissimilarity measurements in a principal coordinates analysis (PCoA).

All analyses have been performed using the statistical software R [52].

Abbreviations

GP: Genomic prediction; GBLUP: Genomic best linear unbiased prediction; SNP: Single nucleotide polymorphisms; INDEL: Insertions/deletions; nsSNP: Non-synonymous single nucleotide polymorphisms; sSNP: Synonymous single nucleotide polymorphisms; tSNP: Tolerant single nucleotide polymorphisms; dSNP: Deleterious single nucleotide polymorphisms; SV: Sequence variants; dSV: Deleterious sequence variants; tSV: Tolerant sequence variants; *l*: Leaf; *s*: Seedling; ePAV: Expression presence/absence variation; $ePAV_s$: Expression presence/absence variation in seedling; $ePAV_l$: Expression presence/absence variation in leaf; $ePAV_{ls}$: Expression presence/absence variation in combining leaf and seedling; GE: Gene expression; GE_s: Gene expression in seedling; GE_l: Gene expression in leaf; TE: Transcript expression; TE_s: Transcript expression in seedling; TE_l: Transcript expression in leaf; M: Metabolites; QTL: Quantitative trait loci; GPA: Generalized procrustes analysis; PCA: Principal component analysis; PCoA: Principal coordinates analysis; LA: Leaf angle; PH: Plant height; HT: Heading time; GC-MS: Gas chromatography-mass spectrometry; FPKM: Fragments per kilobase of exon model per million fragments mapped; w_{SNP}: Weight of single nucleotide polymorphisms; w_{ePAV}: Weight of expression presence/absence variation; $w_{expression}$: Weight of expression; w_M: Weight of metabolite

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-022-08337-7.

Additional file 1: Supplemental Materials.
Additional file 2: Supplementary Table S1.
Additional file 3: Supplementary Table S2.

Acknowledgements

Computational infrastructure and support were provided by the Centre for Information and Media Technology at Heinrich Heine University Düsseldorf. We acknowledge the comments of Joachim Kopka (Department of Molecular Physiology, Max-Planck-Institute of Molecular Plant Physiology, 14476 Potsdam-Golm, Germany) and Dominik Brilhaus (Cluster of Excellence on Plant Sciences (CEPLAS), Heinrich Heine University, 40225 Düsseldorf, Germany) on an earlier version of this manuscript. The authors thank Florian Esser, and George Alskief for technical assistance with performing the field experiments at Cologne and Mechernich as well as Dr. Frankziska Wespel (Saatzucht Breun) and her team for realizing the field experiment at Quedlinburg. We acknowledge excellent technical assistance of Elisabeth Klemp, Katrin Weber, and Maria Graf for GC-MS measurements. We are grateful to Amaury de Montaigu for collecting and processing the metabolite samples.

Authors' contributions

DVI and BS designed and coordinated the project, MW processed genotypic and transcriptomic datasets, PW conducted GS-MS measurements, AE did datamining of GC-MS dataset, AS contributed to data interpretation, PYW performed the data analyses, and DVI, PYW, and BS wrote the manuscript. All authors read and approved the final manuscript.

Funding

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC 2048/1, Project ID: 390686111). The funders had no influence on study design, the collection, analysis and interpretation of data, the writing of the manuscript, and the decision to submit the manuscript for publication. Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

The sequencing datasets have been deposited in the NCBI Sequence Read Archive (SRA) under accession PRJNA534414. The metabolite dataset have been deposited in the MetaboLights (https://www.ebi.ac.uk/metabolights/MTBLS1561). The phenotypic dataset of the adjusted entry means for the three traits can be found in Supplementary Table S1. The annotation and abundance of metabolites can be found in Supplementary Table S2.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Institute of Quantitative Genetics and Genomics of Plants, Heinrich Heine University, 40225 Düsseldorf, Germany. [2]Cluster of Excellence on Plant Sciences (CEPLAS), Heinrich Heine University, 40225 Düsseldorf, Germany. [3]Department of Molecular Physiology, Max-Planck-Institute of Molecular Plant Physiology, 14476 Potsdam-Golm, Germany. [4]Institute of Plant Biochemistry, Heinrich Heine University, 40225 Düsseldorf, Germany.

Received: 17 September 2021 Accepted: 20 January 2022
Published online: 12 March 2022

References

1. Newton AC, Flavell AJ, George TS, Leat P, Mullholland B, Ramsay L, Revoredo-Giha C, Russell J, Steffenson BJ, Swanston JS, Thomas WTB, Waugh R, White PJ, Bingham IJ. Crops that feed the world 4. Barley: a resilient crop? Strengths and weaknesses in the context of food security. Food Secur. 2011;3(2):141–78. https://doi.org/10.1007/s12571-011-0126-3.
2. FAO. The Future of Food and Agriculture – Trends and Challenges. Rome. 2017. http://www.fao.org/3/i6583e/i6583e.pdf. Accessed on 7 May 2021.
3. Fróna D, Szenderák J, Harangi-Rákos M. The challenge of feeding the world. Sustainability (Switzerland). 2019;11(20):5816. https://doi.org/10.3390/su11205816.
4. Abberton M, Batley J, Bentley A, Bryant J, Cai H, Cockram J, Costa de Oliveira A, Cseke LJ, Dempewolf H, De Pace C, Edwards D, Gepts P, Greenland A, Hall AE, Henry R, Hori K, Howe GT, Hughes S, Humphreys M, Lightfoot D, Marshall A, Mayes S, Nguyen HT, Ogbonnaya FC, Ortiz R, Paterson AH, Tuberosa R, Valliyodan B, Varshney RK, Yano M. Global agricultural intensification during climate change: a role for genomics. Plant Biotechnol J. 2016;14(4):1095–8. https://doi.org/10.1111/pbi.12467.
5. Sreenivasulu N, Schnurbusch T. A genetic playground for enhancing grain number in cereals. Trends Plant Sci. 2012;17(2):91–101. https://doi.org/10.1016/J.TPLANTS.2011.11.003.
6. Mantilla-Perez MB, Salas Fernandez MG. Differential manipulation of leaf angle throughout the canopy: current status and prospects. J Exp Bot. 2017;68(21-22):5699–717. https://doi.org/10.1093/JXB/ERX378.
7. Bezant J, Laurie D, Pratchett N, Chojecki J, Kearsey M. Marker regression mapping of QTL controlling flowering time and plant height in a spring barley (Hordeum vulgare L.) cross. Heredity. 1996;77(1):64–73. https://doi.org/10.1038/hdy.1996.109.
8. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157(4):1819–29.
9. Desta ZA, Ortiz R. Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci. 2014;19(9):592–601. https://doi.org/10.1016/j.tplants.2014.05.006.
10. Xu Y, Liu X, Fu J, Wang H, Wang J, Huang C, Prasanna BM, Olsen MS, Wang G, Zhang A. Enhancing genetic gain through genomic selection: from livestock to plants. Plant Commun. 2020;1(1):100005. https://doi.org/10.1016/j.xplc.2019.100005.
11. Guo Z, Magwire MM, Basten CJ, Xu Z, Wang D. Evaluation of the utility of gene expression and metabolic information for genomic prediction in

maize. Theor Appl Genet. 2016;129(12):2413–27. https://doi.org/10.1007/s00122-016-2780-5.
12. 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, Dreisigacker S, Singh R, Zhang X, Gowda M, Roorkiwal M, Rutkoski J, Varshney RK. Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. 2017;22(11):961–75. https://doi.org/10.1016/j.tplants.2017.08.011.
13. Li Z, Gao N, Martini JWR, Simianer H. Integrating gene expression data Into genomic prediction. Front Genet. 2019;10(FEB):126. https://doi.org/10.3389/fgene.2019.00126.
14. Do DN, Janss LLG, Jensen J, Kadarmideen HN. SNP annotation-based whole genomic prediction and selection: an application to feed efficiency and its component traits in pigs. J Anim Sci. 2015;93(5):2056–63. https://doi.org/10.2527/jas.2014-8640.
15. Heidaritabar M, Calus MPL, Megens H-J, Vereijken A, Groenen MAM, Bastiaansen JWM. Accuracy of genomic prediction using imputed whole-genome sequence data in white layers. J Anim Breeding Genet. 2016;133(3):167–79. https://doi.org/10.1111/jbg.12199.
16. Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. Nat Protoc. 2016;11(1):1–9. https://doi.org/10.1038/nprot.2015.123.
17. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001;11(5):863–74. https://doi.org/10.1101/gr.176601.
18. Taylor MB, Ehrenreich IM. Higher-order genetic interactions and their contribution to complex traits. Trends Genet. 2015;31(1):34–40. https://doi.org/10.1016/j.tig.2014.09.001.
19. Wang X, Xu Y, Hu Z, Xu C. Genomic selection methods for crop improvement: current status and prospects. Crop J. 2018;6(4):330–40. https://doi.org/10.1016/j.cj.2018.03.001.
20. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. Bioinforma Biol Insights. 2020;14. https://doi.org/10.1177/1177932219899051.
21. Schrag TA, Westhues M, Schipprack W, Seifert F, Thiemann A, Scholten S, Melchinger AE. Beyond genomic prediction: combining different types of omics data can improve prediction of hybrid performance in maize. Genetics. 2018;208(4):1373–85. https://doi.org/10.1534/genetics.117.300374.
22. Hu X, Xie W, Wu C, Xu S. A directed learning strategy integrating multiple omic data improves genomic prediction. Plant Biotechnol J. 2019;17(10):2011–20. https://doi.org/10.1111/pbi.13117.
23. Weisweiler M, de Montaigu A, Ries D, Pfeifer M, Stich B. Transcriptomic and presence/absence variation in the barley genome assessed from multi-tissue mRNA sequencing and their power to predict phenotypic traits. BMC Genomics. 2019;20(1):787. https://doi.org/10.1186/s12864-019-6174-3.
24. Swarup R, Crespi M, Bennett MJ. One gene, many proteins: mapping cell-specific alternative splicing in plants. Dev Cell. 2016;39(4):383–5. https://doi.org/10.1016/j.devcel.2016.11.002.
25. Rattray NJW, Deziel NC, Wallach JD, Khan SA, Vasiliou V, Ioannidis JPA, Johnson CH. Beyond genomics: understanding exposotypes through metabolomics. Human Genomics. 2018;12(1):1–14. https://doi.org/10.1186/s40246-018-0134-x.
26. Meyer RC, Steinfath M, Lisec J, Becher M, Witucka-Wall H, Törjék O, Fiehn O, Eckardt Ä, Willmitzer L, Selbig J, Altmann T. The metabolic signature related to high plant growth rate in Arabidopsis thaliana. Proc Natl Acad Sci U S A. 2007;104(11):4759–64. https://doi.org/10.1073/pnas.0609709104.
27. Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisec J, Technow F, Sulpice R, Altmann T, Stitt M, Willmitzer L, Melchinger AE. Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat Genet. 2012;44(2):217–20. https://doi.org/10.1038/ng.1033.
28. Longin F, Beck H, Gütler H, Heilig W, Kleinert M, Rapp M, Philipp N, Erban A, Brilhaus D, Mettler-Altmann T, Stich B. Aroma and quality of breads baked from old and modern wheat varieties and their prediction from genomic and flour-based metabolite profiles. Food Res Int. 2020;129. https://doi.org/10.1016/j.foodres.2019.108748.
29. Gemmer MR, Richter C, Jiang Y, Schmutzer T, Raorane ML, Junker B, Pillen K, Maurer A. Can metabolic prediction be an alternative to genomic prediction in barley?,. PLoS ONE. 2020;15(6):0234052. https://doi.org/10.1371/journal.pone.0234052.
30. Gower JC. Generalized procrustes analysis. Psychometrika. 1975;40(1):33–51. https://doi.org/10.1007/BF02291478.
31. Goddard ME, Hayes BJ. Genomic selection. J Anim Breeding Genet. 2007;124(6):323–30. https://doi.org/10.1111/j.1439-0388.2007.00702.x.
32. Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, Devine SE. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res. 2006;16(9):1182–90. https://doi.org/10.1101/GR.4565806.
33. Clark TG, Andrew T, Cooper GM, Margulies EH, Mullikin JC, Balding DJ. Functional constraint and small insertions and deletions in the ENCODE regions of the human genome. Genome Biol. 2007;8(9):1–14. https://doi.org/10.1186/GB-2007-8-9-R180.
34. Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES. Dwarf8 polymorphisms associate with variation in flowering time. Nat Genet. 2001;28(3):286–9. https://doi.org/10.1038/90135.
35. Kono TJY, Fu F, Mohammadi M, Hoffman PJ, Liu C, Stupar RM, Smith KP, Tiffin P, Fay JC, Morrell PL. The role of deleterious substitutions in crop genomes. Mol Biol Evol. 2016;33(9):2307–17. https://doi.org/10.1093/molbev/msw102.
36. Black DL. Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. Cell. 2000;103(3):367–70. https://doi.org/10.1016/S0092-8674(00)00128-8.
37. Cobb JN, DeClerck G, Greenberg A, Clark R, McCouch S. Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype–phenotype relationships and its relevance to crop improvement. TAG Theor Appl Genet Theor Angew Genet. 2013;126(4):867. https://doi.org/10.1007/S00122-013-2066-0.
38. Casale F, Van Inghelandt D, Weisweiler M, Li J, Stich B. Genomic prediction of the recombination rate variation in barley - A route to highly recombinogenic genotypes. Plant Biotechnol J. 2021. https://doi.org/10.1111/PBI.13746.
39. Haseneyer G, Stracke S, Paul C, Einfeldt C, Broda A, Piepho H-P, Graner A, Geiger HH. Population structure and phenotypic variation of a spring barley world collection set up for association studies. Plant Breed. 2009;129(3):271–9. https://doi.org/10.1111/j.1439-0523.2009.01725.x.
40. Lisec J, Schauer N, Kopka J, Willmitzer L, Fernie AR. Gas chromatography mass spectrometry-based metabolite profiling in plants. Nat Protoc. 2006;1(1):387–96. https://doi.org/10.1038/nprot.2006.59.
41. Lommen A. Metalign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. Anal Chem. 2009;81(8):3079–86. https://doi.org/10.1021/ac900036d.
42. Wenig P, Odermatt J. OpenChrom: a cross-platform open source software for the mass spectrometric analysis of chromatographic data. BMC Bioinformatics. 2010;11. https://doi.org/10.1186/1471-2105-11-405.
43. Luedemann A, Strassburg K, Erban A, Kopka J. Data and text mining TagFinder for the quantitative analysis of gas chromatography-mass spectrometry (GC-MS)-based metabolite profiling experiments. Bioinformatics. 2008;24(5):732–7. https://doi.org/10.1093/bioinformatics/btn023.
44. Strehmel N, Hummel J, Erban A, Strassburg K, Kopka J. Retention index thresholds for compound matching in GC-MS metabolite profiling. J Chromatogr B Anal Technol Biomed Life Sci. 2008;871(2):182–90. https://doi.org/10.1016/j.jchromb.2008.04.042.
45. Bayer MM, Rapazote-Flores P, Ganal M, Hedley PE, Macaulay M, Plieske J, Ramsay L, Russell J, Shaw PD, Thomas W, Waugh R. Development and evaluation of a barley 50k iSelect SNP array. Front Plant Sci. 2017;8:1792. https://doi.org/10.3389/fpls.2017.01792.
46. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–60. https://doi.org/10.1038/nmeth.3317.
47. Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, Radchuk V, Dockter C, Hedley PE, Russell J, Bayer M, Ramsay L, Liu H, Haberer G, Zhang XQ, Zhang Q, Barrero RA, Li L, Taudien S, Groth M, Felder M, Hastie A, Šimková H, Stanková H, Vrána J, Chan S, Munõz-Amatriaín M, Ounit R, Wanamaker S, Bolser D, Colmsee C, Schmutzer T, Aliyeva-Schnorr L, Grasso S, Tanskanen J, Chailyan A, Sampath D, Heavens D, Clissold L, Cao S, Chapman B, Dai F, Han Y, Li H, Li X, Lin C, McCooke JK, Tan C, Wang P, Wang S, Yin S, Zhou G, Poland JA, Bellgard MI, Borisjuk L, Houben A, Doleael J, Ayling S, Lonardi S, Kersey P, Langridge P, Muehlbauer GJ, Clark MD, Caccamo M, Schulman

AH, Mayer KFX, Platzer M, Close TJ, Scholz U, Hansson M, Zhang G, Braumann I, Spannagl M, Li C, Waugh R, Stein N. A chromosome conformation capture ordered sequence of the barley genome. Nature. 2017;544(7651):427–33. https://doi.org/10.1038/nature22043.
48. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33(3):290–5. https://doi.org/10.1038/nbt.3122.
49. Jin M, Liu H, He C, Fu J, Xiao Y, Wang Y, Xie W, Wang G, Yan J. Maize pan-transcriptome provides novel insights into genome complexity and quantitative trait variation. Sci Rep. 2016;6(1):1–12. https://doi.org/10.1038/srep18936.
50. Piepho HP, Möhring J. Computing heritability and selection response from unbalanced plant breeding trials. Genetics. 2007;177(3):1881–8. https://doi.org/10.1534/genetics.107.074229.
51. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91(11):4414–23. https://doi.org/10.3168/jds.2007-0980.
52. R Core Team. R: A Language and Environment for Statistical Computing. 2019. https://www.r-project.org/. Accessed on 2 Sept 2019.

Publisher's Note