frontiers in Genetics

ORIGINAL RESEARCH
published: 20 January 2021
doi: 10.3389/fgene.2020.609117

Prior Biological Knowledge Improves Genomic Prediction of Growth-Related Traits in *Arabidopsis thaliana*

*Muhammad Farooq[1,2], Aalt D. J. van Dijk[1,3], Harm Nijveen[1], Mark G. M. Aarts[4], Willem Kruijer[3], Thu-Phuong Nguyen[4], Shahid Mansoor[2] and Dick de Ridder[1]**

[1] Bioinformatics Group, Wageningen University, Wageningen, Netherlands, [2] Molecular Virology and Gene Silencing Lab, Agricultural Biotechnology Division, National Institute for Biotechnology and Genetic Engineering (NIBGE), Punjab, Pakistan, [3] Biometris, Wageningen University, Wageningen, Netherlands, [4] Laboratory of Genetics, Wageningen University, Wageningen, Netherlands

OPEN ACCESS

Edited by:
Shiori Yabe,
Institute of Crop Science (NARO), Japan

Reviewed by:
Yongkang Kim,
University of Colorado Boulder, United States
Tian Qing Zheng,
Chinese Academy of Agricultural Sciences, China

***Correspondence:**
Dick de Ridder
dick.deridder@wur.nl

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: 22 September 2020
Accepted: 21 December 2020
Published: 20 January 2021

Citation:
Farooq M, van Dijk ADJ, Nijveen H, Aarts MGM, Kruijer W, Nguyen T-P, Mansoor S and de Ridder D (2021) Prior Biological Knowledge Improves Genomic Prediction of Growth-Related Traits in Arabidopsis thaliana. Front. Genet. 11:609117. doi: 10.3389/fgene.2020.609117

Prediction of growth-related complex traits is highly important for crop breeding. Photosynthesis efficiency and biomass are direct indicators of overall plant performance and therefore even minor improvements in these traits can result in significant breeding gains. Crop breeding for complex traits has been revolutionized by technological developments in genomics and phenomics. Capitalizing on the growing availability of genomics data, genome-wide marker-based prediction models allow for efficient selection of the best parents for the next generation without the need for phenotypic information. Until now such models mostly predict the phenotype directly from the genotype and fail to make use of relevant biological knowledge. It is an open question to what extent the use of such biological knowledge is beneficial for improving genomic prediction accuracy and reliability. In this study, we explored the use of publicly available biological information for genomic prediction of photosynthetic light use efficiency (Φ_{PSII}) and projected leaf area (PLA) in *Arabidopsis thaliana*. To explore the use of various types of knowledge, we mapped genomic polymorphisms to Gene Ontology (GO) terms and transcriptomics-based gene clusters, and applied these in a Genomic Feature Best Linear Unbiased Predictor (GFBLUP) model, which is an extension to the traditional Genomic BLUP (GBLUP) benchmark. Our results suggest that incorporation of prior biological knowledge can improve genomic prediction accuracy for both Φ_{PSII} and PLA. The improvement achieved depends on the trait, type of knowledge and trait heritability. Moreover, transcriptomics offers complementary evidence to the Gene Ontology for improvement when used to define functional groups of genes. In conclusion, prior knowledge about trait-specific groups of genes can be directly translated into improved genomic prediction.

Keywords: genomic prediction (GP), photosynthesis, phenomics data analysis, *Arabidopsis thaliana* (Arabidopsis), GBLUP, GFBLUP

INTRODUCTION

Due to breakthroughs in DNA sequencing technology over the past decade, high-throughput genotyping is now a routine practice in plant breeding (Rimbert et al., 2018). Phenotyping is undergoing a similar revolution: large phenomics facilities are being developed that can rapidly score large germplasm collections of plants in a range of different environments (Flood et al., 2016; Crain et al., 2018). These technological developments have made it possible to acquire datasets describing genotypes and phenotypes for large numbers of individuals at an extended temporal scale. Despite recent advances in phenomics it is still more expensive and laborious than genotyping. To make the most use of phenomic datasets, Genomic Selection (GS) based breeding programs aim to predict unobserved phenotypes of individuals based on genotypes alone. This has the twofold benefit of reducing breeding costs and speeding up breeding programs as plants can be genotyped in the seedling stage and selected accordingly, thus negating the need to grow large populations to maturity and scoring them all to obtain breeding values based on phenotypes. GS usually models the unobserved phenotypes as additive effects of all genetic markers (total additive genomic value or breeding value) in the test population using a genomic prediction (GP) model. This GP model is based on a reference population which has both been genotyped and phenotyped for the trait(s) of interest (Meuwissen et al., 2001). The performance of GP depends on many factors, including genetic architecture, reference population size and structure and heritability (Karaman et al., 2016). However, GP accuracy, usually defined as the correlation (Pearson's r) between observed phenotypes and predicted breeding values, is generally lower for complex traits than for simpler ones (Morgante, 2018). This is because such traits are affected by many loci with small to moderate effects, along with non-additive genetic (dominance, epistasis) and genotype-by-environment (GxE) interactions (Falconer and Mackay, 1996). Incorporating epistasis into GP models has been reported to improve performance in selfing plant species but may not work for outcrossing species; therefore, additive GP models are still the primary choice (Jiang and Reif, 2015).

In GP models, each individual's genetic or breeding value is modeled as the sum of additive marker effects. Despite advancements in phenomics, phenotyping data is still usually only available for a few traits of several hundreds of individuals (n), compared to millions of genetic markers (p). GP models tackle this curse of dimensionality ($p > n$) by regularization (Meuwissen et al., 2001). When marker effects are fixed, this comes in the form of a penalty term added to the log-likelihood, as in LASSO or ridge regression. More frequently, marker effects are considered random, and regularization is achieved through prior distributions on the marker effects. The variance in these priors is directly related to the heritability, and can be estimated either using REML, or a fully Bayesian approach. In the classical GBLUP-approach, a single normal distribution with equal variance is assumed for all marker effects (Vanraden, 2008). More recently, mixture distributions have been considered (Moser et al., 2015). The prior could e.g., be a mixture of Gaussian distributions with large and small variances, and a point mass at zero, allowing a marker to have respectively, large or small effects, or no effect at all (Macleod et al., 2016). Moreover, restrictions on the shape of the probability distribution, usually Gaussian, can be relaxed (e.g., t-distribution) to accommodate genetic architectures having a larger number of high to moderate effect sizes (Gianola, 2013) or another suitable distribution can be exploited instead. In spite of these refinements, it is usually impossible to find the true causal variants when $p > n$, which may lead to suboptimal prediction. Therefore, several authors suggested that *a priori* available biological knowledge may be incorporated in GP models, prioritizing likely causal markers, and ultimately improving prediction accuracy (Edwards et al., 2016; Ehsani et al., 2016; Wang et al., 2018).

Two types of biological knowledge have been considered in the literature: first, knowledge on biological properties of genes and their associated markers and second, knowledge in the form of secondary phenotypes. The latter typically concerns -omics data, and is modeled using additional relatedness matrices (Guo et al., 2016; Morgante, 2018; Azodi et al., 2020) or penalized selection indices (Lopez-Cruz et al., 2020). Although such -omics data can in principle be generated for the GP reference population, the use of more general publicly available information is often more feasible and cost-effective. We therefore focus on biological properties of genes and markers, such as Gene Ontology (GO) and post-GWAS QTL information. The GO provides a structured resource of functional classes of gene products based on orthology, represented into three biological domains, i.e., molecular function, cellular component and biological process (Ashburner et al., 2000). Similar functional groupings can be achieved from transcriptomic experiments based on the assumption that functionally related genes are expressed together. These clusters of co-expressed genes may be enriched in multiple GO terms or pathways. Such information can be incorporated by allowing the GP model to put more weight on either certain individual markers (Legarra and Ducrocq, 2012; Macleod et al., 2016) or groups of markers (Edwards et al., 2016). Various modeling approaches have been proposed to enable use of such data (Zhang et al., 2010; Speed and Balding, 2014; Edwards et al., 2016; Ehsani et al., 2016; Guo et al., 2016; Fragomeni et al., 2017). Here we use the Genomic Feature Best Linear Unbiased Predictor (GFBLUP) approach proposed by Edwards et al., 2016. GFBLUP extends GBLUP by partitioning the total genomic variance into two sub-components to weigh different genomic regions differently. This allows incorporating prior biological knowledge about groups of variants by treating each region as a separate random genetic effect with different variance. Subsequently, researchers applied this approach to various traits (Sarup et al., 2016; Fang et al., 2017; Rohde et al., 2017; Gebreyesus et al., 2019). While prior biological knowledge has thus been used to improve GP accuracy, the question remains what type of knowledge is most useful and how much the genetic architecture impacts the potential for improvement of particular traits.

In this study, we investigate improvement in GP performance using two sources of publicly available biological knowledge, i.e., Gene Ontology (GO) and clusters of co-expressed genes

(COEX). This information was incorporated using the GFBLUP modeling approach, grouping markers in genes according to either their predicted function or co-expression, respectively. As complex traits of study, we focused on photosynthetic light use efficiency of photosystem II (Φ_{PSII}) and projected leaf area (PLA) in *Arabidopsis thaliana*. Both of these traits are related, in the sense that the Φ_{PSII} directly illustrates the photosynthetic light use efficiency and can capture the most immediate physiological and regulatory response to varying irradiance levels (Van Rooijen et al., 2015), whereas growth in PLA is the net outcome of unit leaf photosynthetic capacity over time (Weraduwage et al., 2015; Liu et al., 2020).

RESULTS

Genomic Prediction of Complex Growth Related Traits

Previously, Van Rooijen et al. (2017) conducted a GWAS on *A. thaliana* photosynthesis. In particular, they measured the light use efficiency of photosystem II electron transport (Φ_{PSII}) for 344 accessions of the Arabidopsis HapMap population, switching from low light (100 μmol m^{-2} s^{-1}) to high light (550 μmol m^{-2} s^{-1}) irradiance at the onset of day 25. In total, they took 6 measurements before and 12 after applying light stress to identify potential QTLs during acclimation to high light. As we intend to use this population to explore the utility of biological knowledge in genomic prediction, we combined projected leaf area (PLA), another indicator of plant growth, with Φ_{PSII}. We first assessed whether GP works with reasonable performance for these complex traits. For this purpose, a classical Genomic Best Linear Unbiased Prediction (GBLUP) model was constructed to assess how well the infinitesimal modeling assumptions fit and to calculate markers-based heritability. In this model (Equation 2), all marker effects are treated as arising from a single normal distribution $N(0, \mathbf{G}\sigma_g^2)$ having one random genetic component, to regress each individual phenotype measurement over all markers simultaneously. At low light (LL) levels, mean prediction accuracy for Φ_{PSII} is lower (Pearson's r between predicted and observed phenotypic values ranging from 0.16 ± 0.02 to 0.22 ± 0.01) than at high light (HL, Pearson's r ranging from 0.40 ± 0.01 to 0.48 ± 0.01), as shown in **Figure 1A**. Prediction accuracy for PLA (**Figure 1B**) ranges from 0.06 ± 0.01 to 0.17 ± 0.01 and rises with the increase in plant size and simultaneously decreases with increase in phenotypic coefficient of variation. Genomic heritability (h^2_{GBLUP}) for Φ_{PSII} ranged from 0.08 to 0.13 under LL and 0.56 to 0.87 under HL, and 0.05 to 0.17 for PLA (**Supplementary Figure 1**). Differences in prediction accuracy for Φ_{PSII} between LL and HL are in line with differences in genomic heritability, in accordance with the observation that genomic prediction accuracy is generally positively correlated with heritabilities (Hayes et al., 2009). Moreover, for ∼1.2% of the GBLUP models for PLA, h^2_{GBLUP} was zero because of undetermined genomic variance, whereas for Φ_{PSII} ∼7% of genomic variances were estimated to be 100% ($h^2_{GBLUP} = 1$), which is clearly an over-estimation (**Supplementary Figure 2**). As reported by Kruijer et al., 2015, it was expected (based on 5000 simulated traits) that ∼10-15% of GBLUP models could have variance components that cannot be estimated for this population, so we discarded these models from our analysis.

An extension of GBLUP is MultiBLUP (Speed and Balding, 2014), using multiple random genetic components in the model (Equation 4), thus allowing differential weighting of groups of genomic markers, each having a separate kinship matrix derived from that group. We applied MultiBLUP using adjacent overlapping chromosomal partitions of 10 kb (yielding best performance when testing window sizes of 1 to 100 kb) to check if multiple kinship matrices or genomic variance decomposition improve prediction. The results (**Supplementary Figure 3**) indicate that performance was close to that of GBLUP and could not be improved further. This could be because most models ended up with only one background kinship matrix during cross-validation and many of these genomic regions did not meet the significance threshold ($p_{\text{Bonferroni}} < 0.05$) during association testing. In summary, these results show that predictive performance for these complex traits is low and there may be room for improvement by incorporating prior biological knowledge, decomposing the total genomic variance into biologically relevant subsets.

High-Level Biological Knowledge Does Not Necessarily Improve Genomic Prediction

The next question is whether predictive performance can be improved by using only markers residing within genes that are known to be linked to the traits of interest. The idea comes from previous studies, in which a subset of markers was associated to biological relevant genes and achieved a genomic value similar to the total genomic value achieved when using all SNPs (Vanraden et al., 2017; Li et al., 2018). Here, we selected 7,242 photosynthesis related genes, referred to as *PSGENES* in the text, from public repositories (see M&M) and constructed a GBLUP model based only on these. The Genomic Relationship Matrix (GRM) was constructed from all markers within the ORFs of *PSGENES*, leaving ∼17% of the total genotyped markers after filtering. Interestingly, the models performed equally well (**Figure 1**) as the GBLUP based on all markers for both traits, with a slight improvement in predictive ability for PLA (max. ∼6% increase in accuracy). Subsequently, to assess whether this pre-selected subset of markers can improve results if they are weighted differently than the rest of markers, we constructed another model using the GFBLUP modeling approach (Edwards et al., 2016) (Equation 3) having two genomic components. In this model, the markers within *PSGENES* were treated as one genomic component and the remaining markers as a second genomic component. Again, this model showed similar predictive performance as GBLUP, with some reduction in variability for PLA, but could not improve the accuracy further (**Figure 1**). From this, we conclude that prior biological knowledge-based selection of functionally relevant genes is potentially useful, but an optimal grouping may be important to improve GP further.

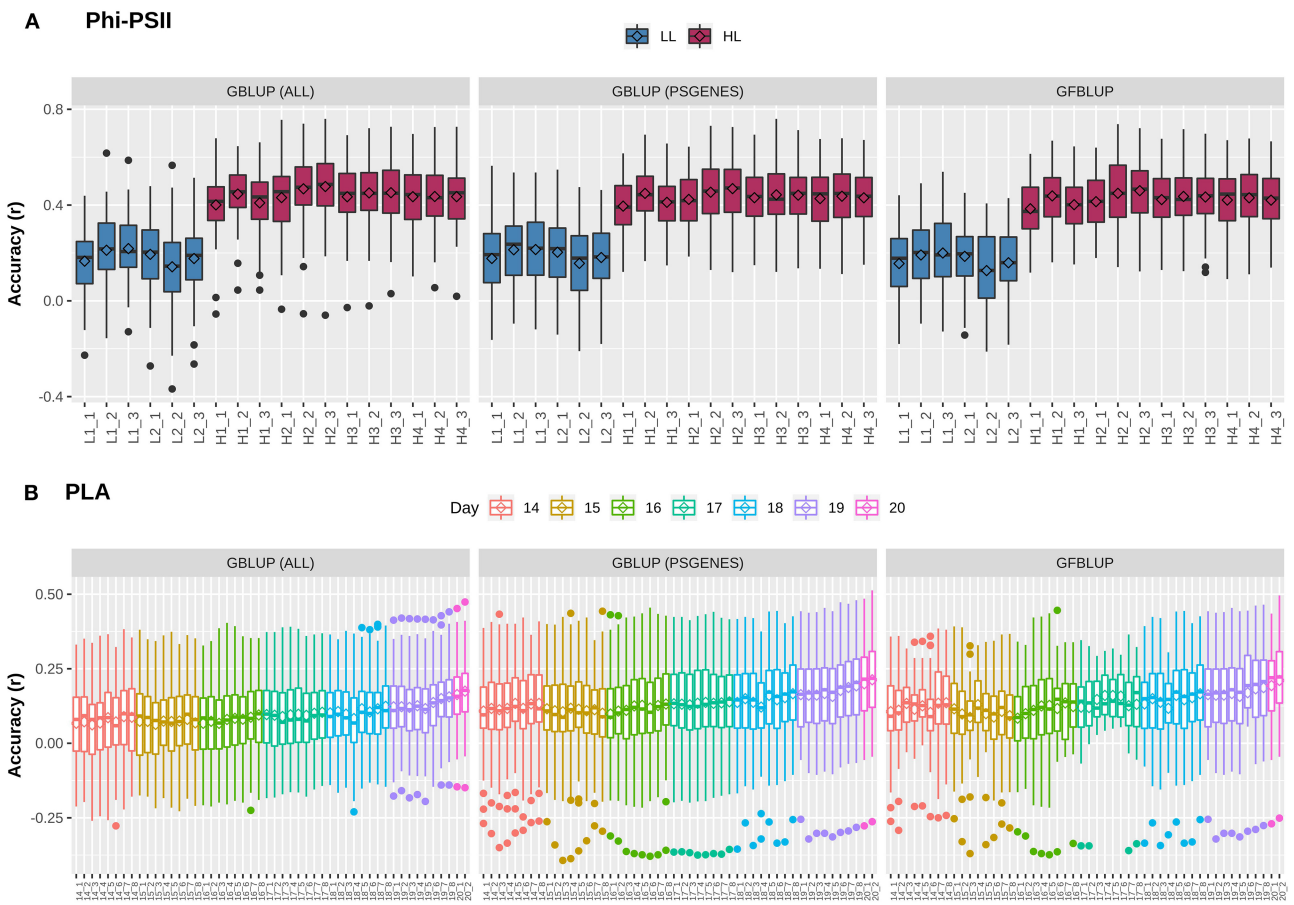

FIGURE 1 | Genomic prediction accuracy for photosynthetic light efficiency and projected leaf area. Cross-validation based assessment of prediction accuracy as Pearson's *r* between true and predicted values using three models: (i) GBLUP using all genomic markers (ALL), (ii) GBLUP using only pre-selected photosynthesis related genic markers (*PSGENES*) and (iii) GFBLUP using *PSGENES* as a genomic feature in two genomic components. **(A)** Accuracy for Φ_{PSII} for two days (6 time points) under low light irradiance levels (LL) and four days (12 time points) under high light irradiance (HL). **(B)** Accuracy for PLA measured 8 times per day from day 14 after sowing to day 20, where day 20 has only two measurements.

More Fine-Grained Biological Knowledge Is Useful for Improving Genomic Prediction

To assess whether prior information from publicly available resources can help improve GP performance, we tested grouping of genes based on Gene Ontology (GO) terms and previously reported clusters of co-expressed genes (COEX) of *Arabidopsis thaliana* in multiple tissues and developmental stages (Movahedi et al., 2011). Each of the three GO sub-ontologies, Biological Process (BP), Molecular Function (MF) and Cellular Component (CC), was used. The corresponding groups of markers in a GO or COEX group, called a genomic feature (GF), were used in GFBLUP (Equation 3) using a separate model for each feature with two genomic components, i.e., one with markers from the GF and the other with the remaining markers (rGF). The predictive performance was compared to that of the GBLUP benchmark using all markers with identical sets of 8-fold cross-validation test populations. Each group of markers based on GO or COEX was treated as a separate random effect in its respective GFBLUP model for which its contribution to the total genomic variance was calculated (see M&M). For each GF, the effects of all corresponding markers were assumed to follow a normal distribution with equal variance, but different from the remaining markers; that is, the markers in the GF are differentially weighted and prioritized from the rest.

In total, 7,297 GO terms and 12,419 disjoint COEX gene groups were linked to at least one marker. The total number of genes ranged between 1 and 24,998 for the GO features and between 1 and 3,384 for the COEX groups (**Supplementary Figure 4**, **Supplementary Table 4**); the number of markers ranged between 0 and 109,723 for the GO features and 4 and 19,621 for the COEX groups. Due to the hierarchical GO structure, the 95th percentile of the total number of genes within GO features was lower (496) as compared to COEX (2,466). Note that both GO and COEX groups may overlap, i.e., a gene can be in multiple functionally related GO/COEX groups. In the following results, the improvement in genomic prediction has been quantified in terms of percent gain in accuracy compared to the GBLUP benchmark, GFBLUP model's goodness

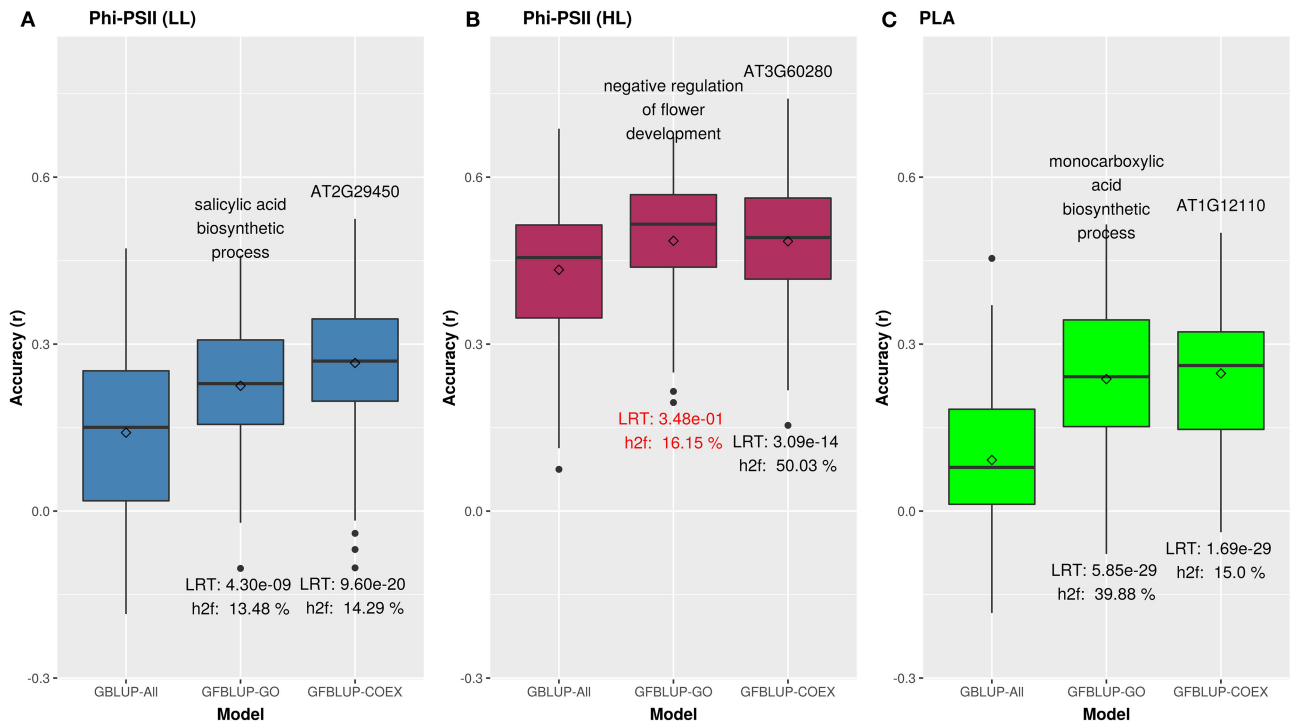

FIGURE 2 | Biological priors can help improve genomic prediction accuracy for growth related traits. Prediction accuracy of the best overall GFBLUP models using Gene Ontology (GO) and co-expression (COEX) gene groups, compared to the GBLUP benchmark (without prior biological knowledge). Since the GBLUP model was evaluated for each measurement point, so the GBLUP here is shown for the corresponding time point where improvement by GFBLUP was observed. Accuracy was calculated as Pearson's *r* between true vs. predicted values. The GBLUP-ALL model uses all markers in GBLUP; GFBLUP-GO and GFBLUP-COEX models use the top GO terms and COEX (see text for details). **(A)** Accuracy for Φ_{PSII} under low light irradiance levels (LL). **(B)** Accuracy for φ_{PSII} under high light irradiance (HL). Here, despite showing some improvement, the GFBLUP-GO model did not pass all of our model evaluation criteria (see Model Performance Evaluation). **(C)** Accuracy for PLA.

of fit measured using likelihood ratio test (LR), and genomic heritability (h^2_{GBLUP}) and proportion of genomic heritability explained by a genomic feature (h^2_f).

GO Informed Prediction

7,297 GO terms were tested with repeated 8-fold cross-validation at multiple measurements of a trait, leading to a total of ∼10 million GFBLUP model accuracies for Φ_{PSII} and ∼29 million for PLA (**Supplementary Figure 5**). The models for which variance was apparently over-estimated ($h^2_f > 0.99$) or undetermined ($h^2_f < 0.01$) were not considered for subsequent analysis. This was the case for ∼50% of the models for both traits, indicating that only selected biological groups are potentially helpful.

We initially analyzed the highest gain in prediction performance obtained by any GO term at any time point. For Φ_{PSII}, "salicylic acid biosynthesis" (BP) provided the highest increase in accuracy (∼60%), for Φ_{PSII} measurements under low light on the second day (**Figure 2**, **Supplementary Table 2A**). For the GO sub-ontologies CC and MF, "organelle outer membrane" and "phosphatase activity," respectively yielded highest gains in these categories under low light (∼43 and 37%, respectively; **Supplementary Table 2A**). None of the GO terms yielded a significant improvement after high light stress; however, some GO terms, e.g., "protein containing complex" yielded an increase in accuracy higher than the benchmark but not passing our model evaluation criteria wholly (**Figure 3**). For PLA, the largest improvement (∼197%) was obtained by the biological process "monocarboxylic acid biosynthesis" (**Figure 2**, **Supplementary Table 2B**). The best performing MF and CC terms for PLA were "exopeptidase activity" and "chloroplast part" (∼185 and ∼178%, respectively; **Figure 3**, **Supplementary Table 2B**). Interestingly, these best CC terms for both traits are directly related to photosynthesis, which lends credibility to the usefulness of the GO terms to capture relevant prior biological knowledge.

In total, 43 GO terms (BP:34, CC:6, MF:3) were potentially informative (i.e., Wilcoxon–Mann–Whitney test *p*-values < 0.05, without multiple testing correction), showing a tendency to improve Φ_{PSII} traits and yielding a significant increase in GFBLUP model accuracy (**Supplementary Figures 6A**, **7**, **Supplementary Table 2A**) compared to GBLUP. The overall gain in accuracy for these informative GO features ranged between 23 and 60%. The GO terms' hierarchical redundancy was removed using GO trimming (Jantzen et al., 2011) and the remaining 40 informative terms fell broadly into six biological clusters (**Figure 4**, **Supplementary Figure 9**): (i) hormonal regulation; (ii) cellular development; (iii) transport; (iv)

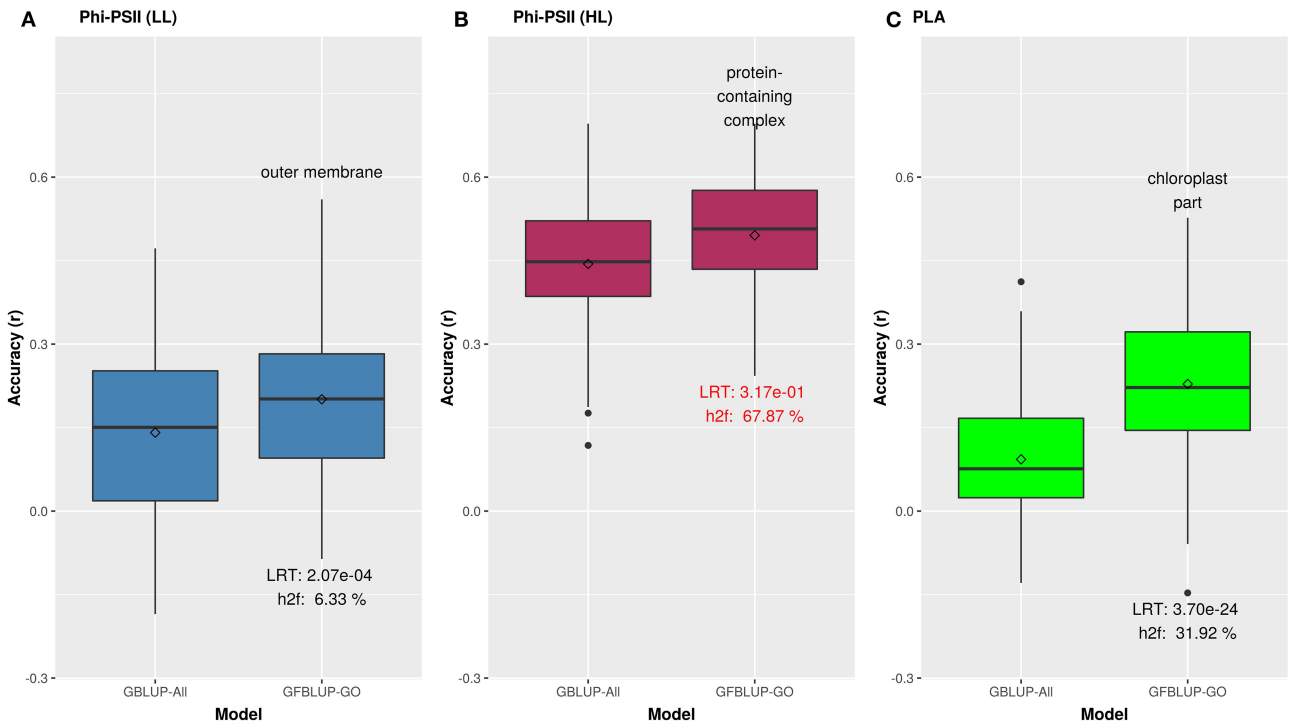

FIGURE 3 | Biological priors based on top GO cellular components improving genomic prediction accuracy for growth related traits. Prediction accuracy of the best GFBLUP models using Gene Ontology (GO) cellular components gene groups, compared to the GBLUP benchmark (without prior biological knowledge). The accuracy of benchmark model may differ within corresponding figures of **Figures 2**, **3**, because it is calculated from the corresponding time point, where improvement by GFBLUP was observed. Accuracy was calculated as Pearson correlation between true vs. predicted values. The GBLUP-ALL model uses all markers in GBLUP; GFBLUP-GO models use the top GO cellular component terms mentioned in the text above. The text in the bottom of boxplots shows the likelihood ratio test *p*-value (LRT) and proportion of genomic heritability explained (h_f^2) by corresponding GO model. **(A)** Accuracy for Φ_{PSII} under low light irradiance levels (LL). **(B)** Accuracy for Φ_{PSII} under high light irradiance (HL). Similar to **Figure 2**, the GFBLUP-GO model did not pass all of our model evaluation criteria (see Model Performance Evaluation), though showing some improvement. **(C)** Accuracy for PLA.

metabolism; (v) catabolism and (vi) macromolecular complex assembly, organization, and biogenesis. The cellular component terms were semantically clustered into organellar membranes and photosynthesis machinery sub-compartments, whereas molecular function terms were related to transmembrane transport and phosphatase activities.

For PLA, 52 GO terms (BP:41, CC:6, MF:5) resulted in significant improvement ($p_{\text{FDR}} < 0.05$) in predictive ability (**Figure 5**, **Supplementary Figure 6C**, **Supplementary Table 2B**) and the gain in accuracy ranged between 104 and 197%. After removal of hierarchical redundancy, semantic grouping of the remaining 45 GO terms showed that they involved a number of growth and developmental processes. Biological process GO terms fell into ∼8 clusters (**Figure 6**, **Supplementary Figure 10**) related to development, defense response, stress response, cell cycle regulation, metabolism, molecular biosynthesis, cellular component organization, and transport. The molecular function terms were clustered into two groups including exopeptidase and methyltransferase activities. The cellular component terms included the photosynthesis machinery (i.e., chloroplast) and endoplasmic reticulum. Comparison of average accuracy over multiple folds of GO models (**Supplementary Figures 6A,C**) indicate that many models performed better than GBLUP. Some of these passed our significance threshold (see model evaluation criteria, M&M) at a particular trait measurement but appeared to improve prediction performance for other measurement points as well.

The maximum number of genes annotated with the informative GO terms for Φ_{PSII} and significant GO terms for PLA were 1,358 and 1,245, respectively. These GO terms appeared at multiple levels of the GO hierarchical structures, including parent and child terms closely related to photosynthesis and growth (**Table 1**). Moreover, many genes were common with the pre-selected photosynthesis related *PSGENES*: 42 and 58% for Φ_{PSII} and PLA respectively, significantly more than what expected by chance ($p_{\chi^2_{df:1}} < 0.05$). Total genomic heritability (h^2_{GBLUP}) was negatively correlated with predictive gain ($r_{\Phi PSII} = -0.77$, $r_{\text{PLA}} = -0.5$). The genomic heritability explained individually (h_f^2) by the informative GO terms ranged between 6 and 31% for Φ_{PSII} and between 3 and 43% for PLA (**Supplementary Tables 2A,B**). Interestingly, the markers associated with these GO terms constituted only 0.1–3.3% of the total markers for Φ_{PSII} and 0.005–2.8% for PLA. This indicates that to improve predictive ability, genomic variance can be decomposed based on biologically meaningful sets of genes

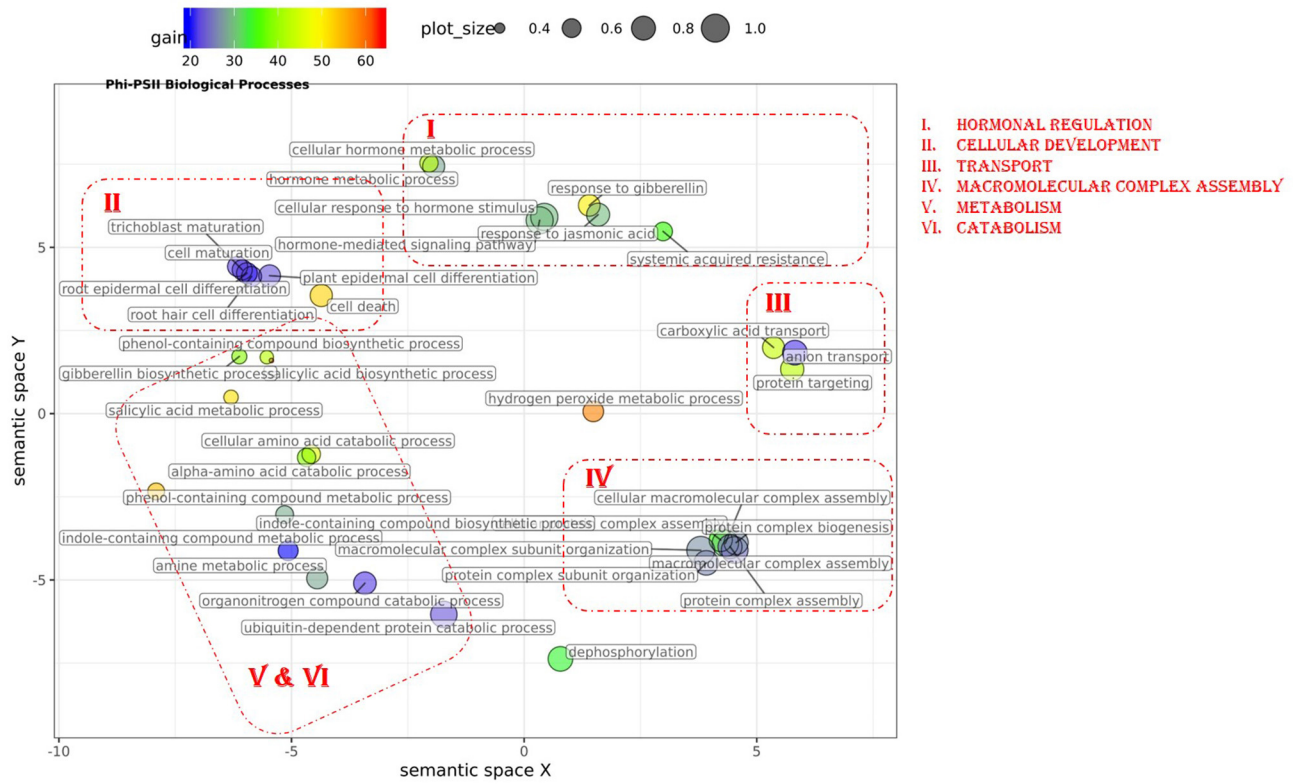

FIGURE 4 | Semantic clustering of GO informed prediction for Φ_{PSII}. Multidimensional scaling (MDS) plot of the representative subset (i.e., terms remaining after the redundancy reduction) of biological process GO terms informative for Φ_{PSII}. Semantically similar GO terms are clustered based on the *"SimRel"* semantic similarity measure using *Revigo*. Dot size is proportional to the number of genes annotated with the GO term, such that more general GO terms have larger circles. The *x* and *y* coordinates indicate relative cluster distances in 2 dimensions. The %gain of a particular GO term is indicated by the circle color.

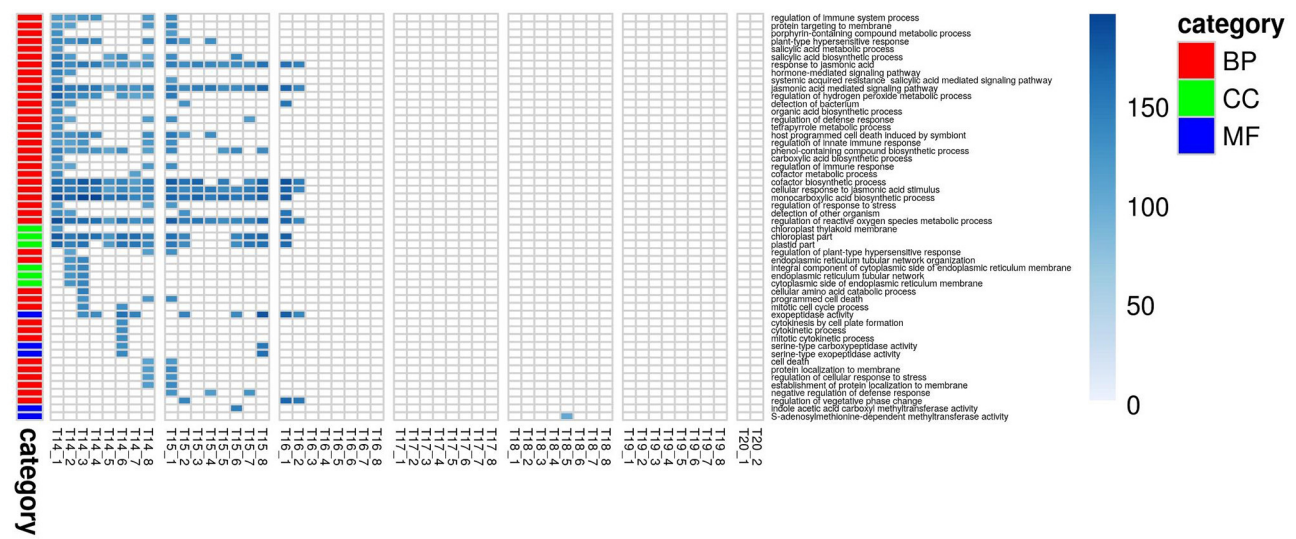

FIGURE 5 | Improvement in genomic predictive performance using GO for PLA. All GO terms that significantly improve GFBLUP models for PLA with %gain in accuracy (*r*) over GBLUP. Each GO term has a separate model for individual measurements indicated as T{day}_{Number of measurement}. The color bar identifies the GO terms as Biological Process (BP), Cellular Component (CC) and Molecular Function (MF).

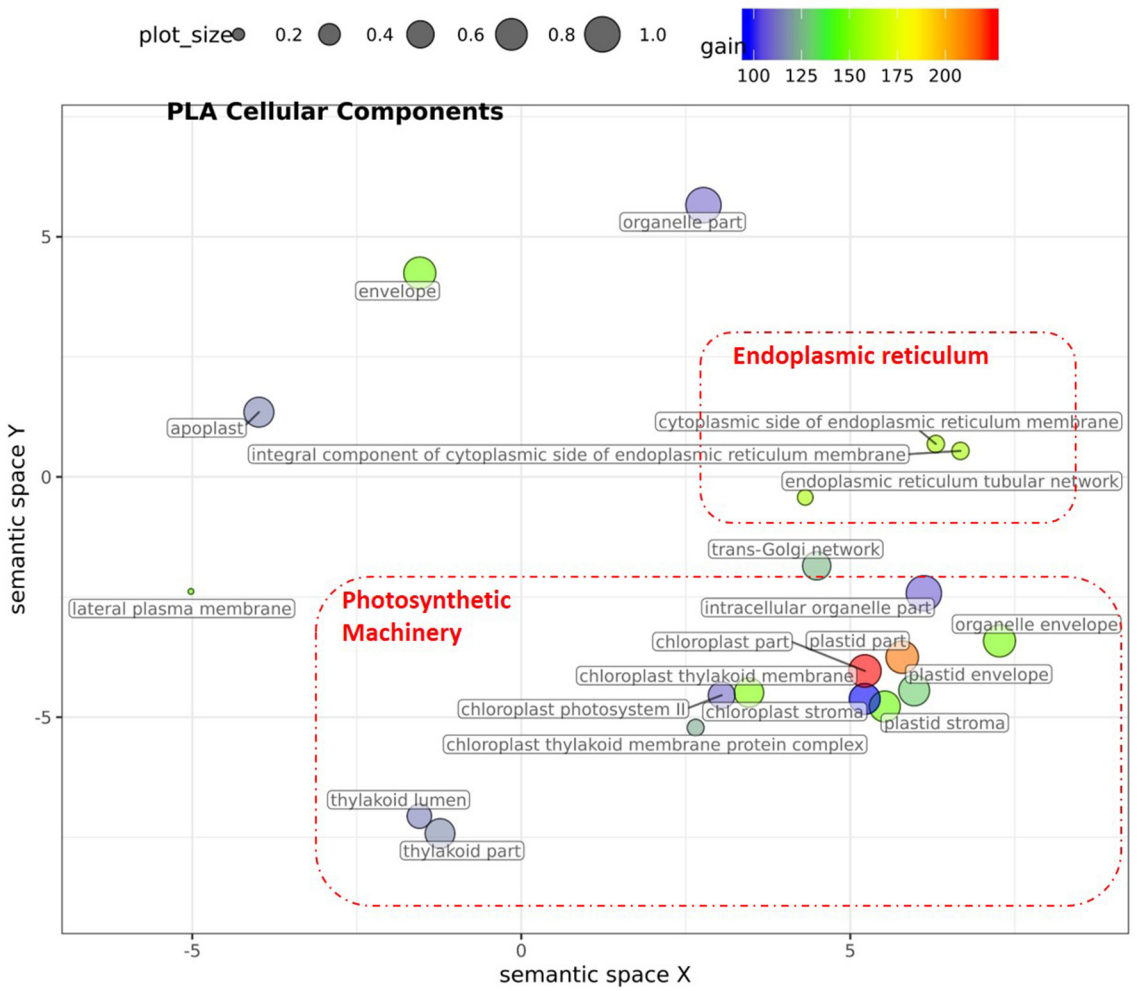

FIGURE 6 | Semantic clustering of GO informed prediction for PLA. Multidimensional scaling (MDS) plot of the representative subset (i.e., terms remaining after the redundancy reduction) of cellular component GO terms informative for PLA. Semantically similar GO terms are clustered based on the *"SimRel"* semantic similarity measure using *Revigo* (Supek et al., 2011). Dot size is proportional to the number of genes annotated with the GO term, such that more general GO terms have larger bubbles. The *x* and *y* coordinates indicate relative virtual cluster distances in 2 dimensions. The %gain of a particular GO term is indicated by the bubble color.

scattered over the genome rather than lie in adjacent regions such as in the MultiBLUP analysis above. Moreover, h_f^2 is positively correlated with GO gene group size ($r_{\Phi\mathrm{PSII}} = 0.87$, $r_{\mathrm{PLA}} = 0.77$) as well as with the likelihood ratio ($r_{\Phi\mathrm{PSII}} = 0.60$, $r_{\mathrm{PLA}} = 0.65$) of both trait models, indicating that incorporating meaningful prior subsets into the GFBLUP model improves goodness of fit.

From this we infer that GO-based prior knowledge can improve GP performance. The improvement is most prominent for traits with low heritability, where some of the GO terms appeared more frequently for PLA than Φ_{PSII} at multiple measurement times.

COEX Informed Prediction

Similar to genomic features based on GO, we made subsets of markers based on COEX clusters by selecting the markers within the ORFs of genes which were part of a given COEX cluster. Similar to GO based models, COEX models with zero and with 100% variance explained were discarded (**Supplementary Figure 5**). In general, more COEX models pass our model evaluation threshold (**Supplementary Figures 6B,D**) and they have a higher likelihood ratio than GO based models. This could be due to the genic overlap between groups and the enrichment of multiple related GO terms within a group.

For Φ_{PSII} we found 172 informative COEX gene groups potentially improving predictive ability, one of which was statistically significant ($p < 0.05$) after correcting for multiple testing using FDR (**Supplementary Figures 6B**, **8**). 355 COEX groups significantly improved predictive ability for PLA (**Figure 7**, **Supplementary Figure 6D**, **Supplementary Tables 3A,B**). The gain in accuracy was higher for PLA (80 to 243%) than for Φ_{PSII} (7 to 89%) and was negatively correlated with genomic heritability ($r_{\Phi\mathrm{PSII}}$

TABLE 1 | Known trait-specific GO terms improving genomic prediction performance for both traits.

GO ID	Ontology	Type	h^2_f	LR	*p*-value (unadj)	#gene	#marker	%gain	Cor(G_f,G_r)	h^2_{GBLUP}
Φ_{PSII}										
GO: 0009543	chloroplast thylakoid lumen	CC	0.07	10.53	1.48×10^{-2}	71	218	33	0.59	0.09
GO: 0031968	organelle outer membrane	CC	0.06	12.47	4.3×10^{-3}	72	345	40	0.61	0.08
GO:0044429	mitochondrial part	CC	0.14	47.05	2.3×10^{-3}	298	1069	38	0.81	0.09
GO:0005740	mitochondrial envelope	CC	0.13	8.43	2.7×10^{-2}	255	914	25	0.79	0.12
GO ID	**Ontology**	**Type**	h^2_f	**LR**	***p*-value (adj)**	**#gene**	**#marker**	**%gain**	**Cor(G_f,G_r)**	h^2_{GBLUP}
PLA										
GO:0044434	Chloroplast part	CC	0.32	101	5.26×10^{-5}	1211	5658	178	0.94	0.07
GO:0009535	chloroplast thylakoid membrane	CC	0.14	10	4.9×10^{-2}	322	1139	121	0.81	0.07
GO:0000911	cytokinesis by cell plate formation	BP	0.15	34	9.6×10^{-3}	204	1465	134	0.81	0.07
GO:0010090	trichome morphogenesis	BP	0.04	30	8.3×10^{-4}	31	65	154	0.40	0.06
GO:0010321	regulation of vegetative phase change	BP	0.14	18	4.9×10^{-3}	425	1512	106	0.84	0.07
GO:0048366	leaf development	BP	0.10	48	1.96×10^{-5}	99	487	187	0.62	0.06
GO:0090698	post-embryonic plant morphogenesis	BP	0.04	7	8.3×10^{-7}	4	11	207	0.20	0.06

The proportion of explained genomic heritability (h^2_f) by a GO term, likelihood ratio (LR) between GFBLUP and GBLUP models, Wilcoxon–Mann–Whitney test p-value, total number of genes and markers, %gain in accuracy (r), correlation between genomic relationship matrices based on GO term markers (G_f) and remaining markers (G_r) and total genomic heritability (h^2_{GBLUP}), for different trait specific GO terms that are common to both GO and COEX based analyses. For GO terms, the type is indicated—molecular function (MF), biological process (BP) and cellular component (CC).

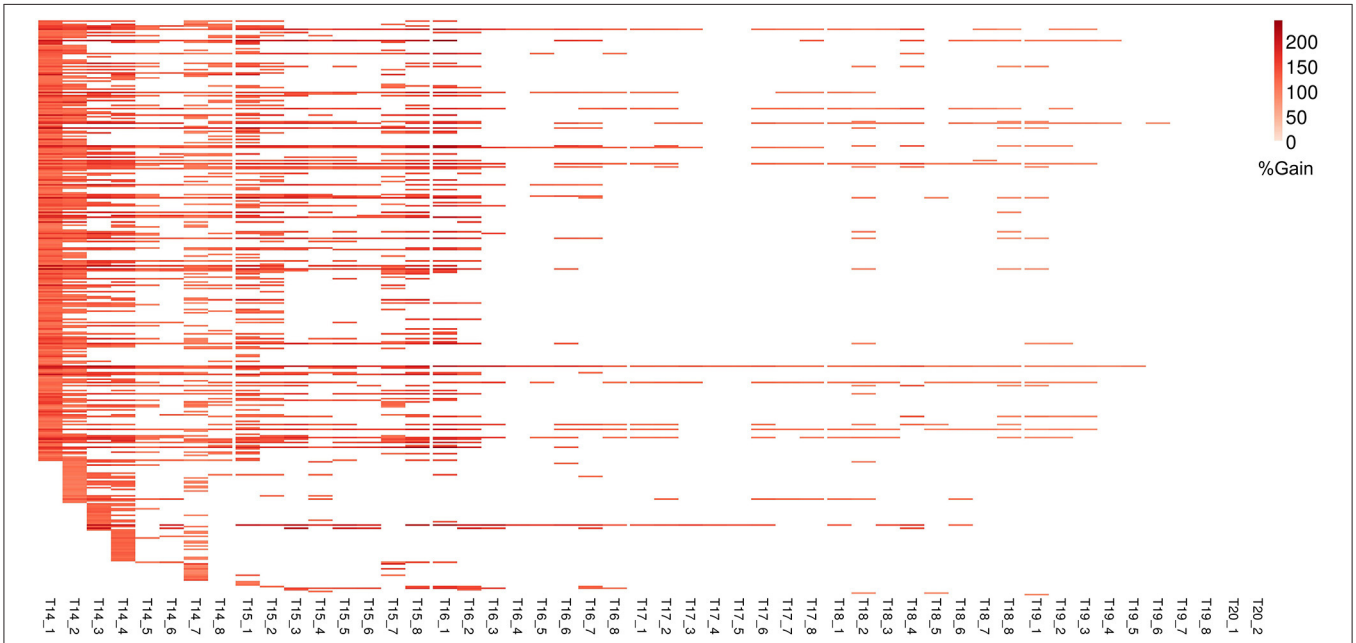

FIGURE 7 | Improvement in genomic prediction performance using co-expressed gene clusters for PLA. All COEX clusters that significantly improve GFBLUP models for PLA with %gain in accuracy (*r*) over GBLUP. Each COEX cluster has a separate model for individual measurements indicated as T{day}_{Number of measurement}. The clusters are ordered according to "cluster_sr_no" column in **Supplementary Table 3B**.

$= -0.86$, $r_{PLA} = -0.56$), like for GO informed prediction. This improvement was attributed to a maximum of only ∼5% of the total genomic markers in all groups. Interpretation of COEX gene groups is not as straightforward as of GO terms, which by nature carry an informative name. Interestingly, ∼90% of genes were common in the COEX groups for both traits, possibly due to the relatedness of the traits. To attach biological meaning to these groups we performed GO enrichment analysis on all groups together. We found 113 BP, 29 MF, and 24 CC most specific GO terms enriched in these clusters. The top 10 GO terms with highest fold enrichment include photosynthesis machinery, i.e., chloroplast stroma (GO:0009570), chloroplast

envelope (GO:0009941) cellular components; ATPase activity coupled with transmembrane ion transport (GO:0015662); and glucose metabolic process (**Supplementary Figure 11**, **Supplementary Table 5**). These results indicate that trait-specific co-expressed gene functional groups can also help improve prediction performance and that these groups capture biologically relevant functions.

Similar to GO informed prediction, ∼34% of COEX genes were common to the pre-selected photosynthesis related genes (*PSGENES*) for both traits, but here this is close to what we expect by chance. This indicates that, even though the COEX groups contain only a limited subset of all genes, they are not biased toward photosynthesis genes. The gain in predictive ability and explained genomic heritability (h_f^2) for Φ_{PSII} by the top COEX gene group was higher (89% resp. 14%) than those for the top GO feature (60% resp. 13%). Similarly, for PLA the top COEX gene group achieved a higher accuracy gain (242%) than the top GO group (197%), as shown in **Figure 2**. Notwithstanding these differences, we observed that many genes were common between GO and COEX based prediction for both traits (21 and 19% of all models passing the evaluation criteria for Φ_{PSII} and PLA resp.). These common genes in COEX based prediction were mainly enriched for many fundamental photosynthesis and growth related GO terms (**Supplementary Tables 7A,B**), e.g., light harvesting in photosystem I and photosynthetic electron transport in photosystem II (BP), chloroplast (CC), and ATP binding (MF).

The largest informative COEX groups for Φ_{PSII} and for PLA only differ slightly in sizes (3,176 and 2,840 genes, respectively), but on average, COEX groups were larger than the GO groups for both traits. The 95th percentile of genomic heritability explained individually by the COEX groups (h_f^2) was 70% for Φ_{PSII} and 39% for PLA, indicating that some Φ_{PSII} models could be over-estimated. Analogous to GO, h_f^2 was positively correlated with COEX gene group sizes ($r_{\Phi\mathrm{PSII}} = 0.88$, $r_{\mathrm{PLA}} = 0.40$) and likelihood ratio ($r_{\Phi\mathrm{PSII}} = 0.27$, $r_{\mathrm{PLA}} = 0.22$), indicating that incorporating meaningful prior subsets into the COEX model improved goodness of fit.

Together, our results illustrate that both of the meaningfully specific GO terms and more general COEX groups of genes with interrelated functions may improve GP predictive performance.

DISCUSSION

Predicting Photosynthesis

In this work, we aimed at improving GP performance by exploiting publicly available biological knowledge to group genes in three different ways: using our knowledge about the trait, using the Gene Ontology and using co-expression. Instead of developing new methodology, we focused on using existing BLUP methods, widely used in animal and plant breeding, to explore new sources of biological prior knowledge, e.g., clusters of co-expressed genes. The GFBLUP methodology was initially proposed for *Drosophila melanogaster* using Gene Ontology data as biological prior knowledge (Edwards et al., 2016). We also investigated to what extent different traits benefit from and the use of prior knowledge. Our results support a strong influence of different trait genetic architectures, since performance improvement was more evident for leaf area phenotypes than for Φ_{PSII}.

The approach can be generally applied to complex traits, but here we focused on photosynthesis and plant size. Besides serving as a case study, photosynthesis is also interesting in its own right, for two reasons. First, the genetic architecture of photosynthesis, though well-studied over the previous decades, is still poorly described in the quantitative genetic context (Van Rooijen et al., 2017). Secondly, it is an important target for improvement in crop breeding (Long et al., 2015). Modest improvements in photosynthesis efficiency by engineering photorespiratory pathways have demonstrated enormous yield gains (Kromdijk et al., 2016; South et al., 2019). The yield model of Monteith (Monteith, 1977) suggests that increased light use efficiency of photosystem II holds great potential to meet global food challenges by increasing the conversion efficiency of intercepted irradiance into biomass (ε_c) (Van Bezouw et al., 2019). Another determinant of plant growth rate is leaf area growth, involving precise regulation of photosynthesis machinery and growth hormones such as auxin (Zhang et al., 2017). Leaf area measurements from fluorescence based non-destructive optical phenotyping systems, can be efficiently used to screen plants at different growth stages with varying levels of photosynthetic rates (Weraduwage et al., 2015). Therefore, improved GP models for these traits could have impact in future crop breeding.

Following Edwards et al. (2016), we studied accuracy on internal test sets within the HapMap population. Further work is needed for data-driven selection of the most relevant terms for prediction on external test sets. For example, a possible strategy may be to select the feature with highest genomic variance explained, or with lowest p-value in the LRT we described. Our results indicate that biological priors driven GP models can be used to rank groups of genes potentially associated to the trait of interest along with improving prediction performance. The GWAS conducted on the same HapMap population for photosynthetic light use efficiency of photosystem II identified that the *A. thaliana* "Yellow Seedling 1" gene is involved in photosynthesis acclimation response (Van Rooijen et al., 2017). This *YS1* gene is annotated with GO Cellular Component terms chloroplast, intracellular membrane-bounded organelle and mitochondrion and GO Biological Process terms thylakoid membrane organization and photosystem II assembly. Our results using GO and COEX GP (**Table 1**) clearly demonstrate that these GO terms were most prevalent to improve the prediction and explain a large amount of genomic heritability. This indicates that genomic prediction and GWAS support each other as potentially useful tools for forward genetics.

The gain of predictive accuracy of the GP models compared to the base-model is trait-specific and negatively correlates with genomic heritability, which is promising for breeding at low h^2. This inverse relation may be due to the fact that we deal with highly polygenic, complex traits: many physiological and regulatory biological processes are involved in Φ_{PSII} under high light stress, e.g., PSII repair, ROX etc. Our models, testing groups

of genes individually, may not be able to improve performance for such cases. Another potential explanation lies in the ability of GFBLUP to capture small genetic variance at low h^2 in a separate random component, potentially including known causal genes, which is not possible in GBLUP.

Exploiting Biological Knowledge to Improve Genomic Prediction

With recent technological advances in both field and controlled environment high-throughput phenotyping systems, phenotypes can be measured at unprecedented scales. Phenotypes can vary in space and time due to genetics and environment alone, genotype-by-environment (GxE) interactions as well as stochastic and development effects. Component variances due to these factors can be calculated by precise modeling. If multiple measurements are available, GP models can be developed on individual measurements, treated as individual phenotypes, or on derived parameters, e.g., growth curves. We found that at each measurement timepoint, at least some GO (in particular cellular component terms) or COEX group could help to improve performance, and some were more frequent (**Figure 4**, **Supplementary Figure 7**). For example, for Φ_{PSII} no single GO or COEX gene group was capable of improving GP accuracy for all time points (either LL or HL separately), but a number of gene groups were able to improve PLA at multiple measurements (although not always meeting the threshold for significance). Phenotyping at an extended scale and GP modeling thus provides an opportunity to obtain biological insights. As an alternative to modeling at each timepoint separately, a whole time series or growth curve can be used instead. We did not pursue this here, as time series data is not generally available in most practical scenarios and we were interested to learn whether performance improvement was specific to growth stages and conditions e.g., models for Φ_{PSII} behaved differently under low and high light conditions.

Here, we mainly investigated two approaches to incorporate publicly available trait-specific biological information into GP, i.e., pre-selecting a list of genes and selecting sets or groups of genes based on predicted functional (i.e., GO) or expression (COEX) information. The approach using predicted functional information proved to be more useful in this context, but more approaches and sources of information can also be incorporated with a focus on prioritizing biologically related genomic regions. Moreover, knowledge from multiple heterogeneous sources can be combined to further pinpoint potential QTLs, termed as poly-omics GP models (Wheeler et al., 2014; Uzunangelov et al., 2020). These information sources may include (i) predicted variants effects, (ii) gene functions e.g., GO, COEX, (iii) networks of gene-gene and protein-protein interactions, stored in public resources like STRING (Mering et al., 2003), GeneMANIA (Warde-Farley et al., 2010); (iv) pathways, in which genes are grouped e.g., KEGG (Kanehisa and Goto, 2000); (v) previously generated GWAS and QTL results which indicate involvement of particular regions for specific traits e.g., AraGWAS (Togninalli et al., 2020), AraQTL (Nijveen et al., 2017), (vi) known connections to phenotypes and (vii) endophenotypes, usually measured using -omics data at different stages of genetic information flow toward phenotypes. The reliability of these sources of information is an important factor for credible analysis. Information describing the (un)certainty of annotations is generally available in the form of a score (e.g., for gene functions based on GO evidence scores or reliability scores generated by a prediction method). It remains an open question how to incorporate such scores in the process of using the biological knowledge for GP.

Our first approach, pre-selecting a gene list, seems to be naive but can be useful as a baseline for comparison with more complex statistical procedures. The group based approach is usually based on gene function, but this heavily depends on computational prediction, as for most of the genes in plants and animals, no experimental function annotation is available (Radivojac et al., 2013). Function prediction is often based on sequence similarity, which works well for predicting molecular functions but less so for biological processes. Using expression compendia based on multiple experiments poses an interesting alternative, since genes with similar expression patterns are more likely functionally related, hence more likely involved in the same biological process(es) (Kourmpetis et al., 2011). Alternatives are to define phenotype associated genomic regions based on differential gene expression levels (Fang et al., 2017) or metabolite levels and metabolic fluxes (Tong et al., 2020), or to construct haplotypes in genic regions based on their ontology information (Gao et al., 2018). The GP requiring genomics inferred relationship matrices (GRM), e.g., GBLUP and its variants, can make use of information derived from these sources to construct a population variance-covariance structure (Zhang et al., 2010, 2011; Fragomeni et al., 2017). A simple approach is to include multiple random effects for each knowledge source yielding its own variance-covariance structure for the population under study, in the mixed model equations (Guo et al., 2016). One way to combine multiple omics datasets is to prepare a Composite Relationship Matrix (CRM) as a linear combination of Genomic Relationship Matrices (GRMs), Expression Relationship Matrices (XRMs), Metabolome Relationship Matrices (MRMs), MicroRNA Relationship Matrices (miRMs) etc. (Wheeler et al., 2014).

Alternative Models for Genomic Prediction

Linear mixed model (LMM)-based genomic prediction, as used in this work, makes use of raw genotypes and parameter regularization to estimate thousands of SNP marker effects using only a few hundred observations ($p >> n$), employing different prior statistical assumptions on these parameters. This makes the approach fairly simple and interpretable; therefore, biological knowledge can be incorporated straightforwardly by employing these statistical assumptions. But with the increase in the ratio between markers and available phenotypes, serious overfitting problems may be encountered in these models (González-Recio et al., 2014), leading to a need to use prior knowledge in regularization. A more general set of statistical learning methods are Machine Learning (ML) methods for prediction and classification, capable of dealing with the dimensionality problem in a more flexible manner. In these methods, phenotypes

are regressed on nonlinear functions of genotypes rather than raw genotype values, compromising model interpretability but potentially improving prediction performance. Several studies have reported the use of Support Vector Machines (SVM), Reproducing Kernel Hilbert Spaces Regression (RKHS), Neural Networks (NN), Random Forests (RF), and boosting (De Los Campos et al., 2010; Ogutu et al., 2011) for genomic prediction. Still, low prediction accuracy remains a problem for complex traits. It will be interesting to further explore how biological knowledge can be incorporated into ML approaches for GP. One way could be to involve a knowledge driven regularization-based approach as demonstrated for disease prediction in human (Deng and Runger, 2013).

CONCLUSION

The wealth of publicly available transcriptomics and Gene Ontology based prior biological knowledge can be incorporated for genomic prediction of photosynthetic light use efficiency of photosystem II electron transport (Φ_{PSII}) and PLA. Significant improvement in prediction accuracy over the benchmark GBLUP model was obtained for several GO terms and COEX groups. This improvement is trait-specific and negatively correlates with genomic heritability; whereas, for projected leaf area we found more added value than for Φ_{PSII}. Many known photosynthesis-specific GO terms lead to improvements, providing evidence of the potential usefulness of this approach in future breeding practice. We foresee incorporation of heterogeneous prior biological information into machine learning algorithms as an active area of research in future.

MATERIALS AND METHODS

Datasets

Genotype Data

Genotype data of the 360 natural accessions in the core set of the *Arabidopsis thaliana* HapMap population, representing its global diversity, was obtained using Affymetrix 250k SNP array (Zhang and Borevitz, 2009; Baxter et al., 2010). The HapMap accessions were chosen as most accessions are more or less equally interrelated, so modeling is not heavily affected by population structure. Phenotypes of 344 accessions were available, so 16 accessions were removed from the analysis (CS76104, CS76112, CS76254, CS76257, CS76121, CS28051, CS28108, CS28808, CS28631, CS76086, CS76138, CS76212, CS76196, CS76110, CS76117, CS76118). Genotype data were subjected to quality control and all genotypes with a missing call in any accession were removed. Only 510 (0.24%) markers had minor allele frequency (MAF) <0.01 and 14,824 (6.9%) had MAF <0.05 (**Supplementary Figure 12**). To incorporate the effects of rare alleles along with common alleles in the GP model, the MAF filtering threshold was set at 0.01. Of subsequent markers in a window of 50bp with a Pearson correlation coefficient (r) < 0.999, one was removed, using PLINKv1.9 (Purcell et al., 2007). In total, 214,051 SNPs passed quality filtering, 213,541 remained after MAF filtering and 207,981 SNPs were available after LD pruning for the analyses. The resulting minimal distance between SNPs was found to be ∼550 bp.

Phenotype Data

The light use efficiency of Photosystem II electron transport (Φ_{PSII}) dataset was obtained from Van Rooijen et al. (2017), who measured it using chlorophyll fluorescence via NIR imaging at 790 nm. In this dataset, Φ_{PSII} was recorded three times a day; under 100 μmol m^{-2} s^{-1} (low light) for 2 days and for four continuous days after induction of high light stress at 550 μmol m^{-2} s^{-1} to study the photosynthetic acclimatory response. We measured PLA every 3 h starting from the afternoon of day 22 after sowing until early morning of day 29 using the *"Phenovator"* high-throughput automated phenotyping system (Flood et al., 2016), which results in total of 54 timepoints for this trait (**Supplementary Table 8**). Technical mis-match errors between the imaging system and the coordination of image analysis software were identified for some replicates at some time points for a small number of genotypes, but these were not found to influence overall results and the data was thus retained. Data of timepoints on day 22 was excluded from the analyses due to their relatively low coefficient of variation.

The *Phenovator* system has been designed to screen Arabidopsis plants for photosynthesis and growth on a larger temporal scale in a carefully controlled environment with minimal noise. The plants are grown over a table, spatially arranged into sowing blocks, imaged using a moveable monochrome camera recording 12 plants per image, and processed using an image processing software (available on demand from the authors). The system design allows spatial uniformity and temporal reproducibility by minimizing the design parameter variances. Therefore, we expected low variances of interactions between genotype and the design parameters; whereas, within image position and sowing position could have larger main effects and thus could be corrected for. Phenotypic values were taken as the average of one to four replicates of Best Linear Unbiased Estimators (BLUE) using the linear mixed model adjusted for experimental design factors (**Supplementary Table 9**) that were described in Flood et al. (2016). For this experiment, the important design factors are spatial row (x) and column (y) coordinate, the image position and the sowing block. Thus, the BLUE for phenotypic mean is calculated based on this equation, implemented in R with the *lmer* function (supplemental R script) using the lme4 package (Bates et al., 2007):

$$Y = Genotype + x + y + Image\_position + Sowing\_block + error \quad (1)$$

where *Genotype* is used as fixed effect and the other factors are defined as random effects.

Both traits, at all measurement times, showed approximately normal distributions (**Supplementary Figures 13**, **14**). The distributions are leptokurtic and left skewed for both traits (except for a few measurements for PLA on day 14 and day 15). The coefficients of variation under low light conditions for Φ_{PSII} ranged from 1.95 to 2.30% and 2.92 to 7.58% under high

light and 18.73 to 27.04% for PLA (**Supplementary Table 1**). Correlation between subsequent measurement times was high ($r > 0.9$) for both traits, except between measurements under low vs. high light conditions of Φ_{PSII}; therefore, these were analyzed separately.

Biological Priors

Co-expressed gene groups were obtained from the Arabidopsis expression compendium by Movahedi et al. (2011). GO data was retrieved using the R package "org.At.tair.db" (Carlson, 2019b) and genes were annotated using "GO.db" (Carlson, 2019a) irrespective of evidence codes. The set of genes in GO terms were up-propagated along the GO tree, such that each GO group in our analysis comprised of a set of all those genes attributed to itself or to all of its child terms. The up-propagated sets of genes were retrieved using the "GO2ALLTAIRS" method in the "org.At.tair.db" package. Markers in genes linked to a specific GO term or COEX cluster were used in the analyses.

Moreover, a set of 7,242 photosynthesis related genes was manually compiled (**Supplementary Table 6**) using four publicly available sources: KEGG (Kanehisa, 2001) pathways related to photosynthesis (i.e., ath00195, ath00197, ath00710); the Arabidopsis pathway database AraCyc for four photosynthesis pathways (i.e., Calvin cycle, photorespiration, oxygenic, light reaction); genes annotated with GO terms directly related to photosynthesis machinery; and all 51 priority genes selected for GWAS of photosynthesis acclamatory response identified by for this HapMap population.

Statistical Analysis

Linear Mixed Models

The Linear Mixed Model (LMM) with one random genomic component was used as baseline. This model (Equation 2), known as Genomic Best Linear Unbiased Prediction (GBLUP) (Habier et al., 2007; Vanraden, 2008) was used to predict marker effects, calculate genomic heritability (h^2_{GBLUP}) and the total additive genomic values, which is the sum of all marker effects:

$$\tilde{\mathbf{y}} = \boldsymbol{\mu} + \mathbf{g} + \boldsymbol{\varepsilon} \tag{2}$$

Here, $\tilde{\mathbf{y}}$ is an nx1 vector of adjusted phenotypes as described in section 5.1.2, $\boldsymbol{\mu}$ is the overall mean, $\mathbf{g}$ is an nx1 vector of genomic values captured by all genomic markers such that $\mathbf{g} = \hat{\mathbf{g}}$ and $\boldsymbol{\varepsilon}$ is an n-vector of residuals. The random genomic values $\mathbf{g}$ and residuals were assumed to be independent, normally distributed as $\mathbf{g} \sim N(0, \mathbf{G}\sigma^2_g)$, $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}\sigma^2_e)$. Here $\mathbf{G}$ is the genomic relationship matrix (GRM), providing variance-covariance structure of genotypes calculated from all genomic markers and $\mathbf{I}$ is the identity matrix.

Accordingly, for each GO and COEX gene groups, another linear mixed model similar to GBLUP but with two random genomic components (Equation 3), known as Genomic Feature Best Linear Unbiased Predictor (GFBLUP) (Edwards et al., 2016) was applied:

$$\tilde{\mathbf{y}} = \boldsymbol{\mu} + \mathbf{f} + \mathbf{r} + \boldsymbol{\varepsilon} \tag{3}$$

This model differs from GBLUP in that the total estimated genomic value ($\hat{\mathbf{g}} = \mathbf{f}+\hat{\mathbf{r}}$) is partitioned into genomic value captured by markers in a GO/COEX group ($\mathbf{f}$) and by the remaining markers ($\hat{\mathbf{r}}$), such that $\mathbf{f} \sim N(0, \mathbf{G_f}\sigma^2_f)$, $\mathbf{r} \sim N(0, \mathbf{G_r}\sigma^2_r)$ and $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}\sigma^2_e)$. For both GBLUP and GFBLUP, total genomic value $\hat{\mathbf{g}}$ of the test population was predicted conditional on observed phenotypes of the training population, using the approach mentioned by Edwards et al. (2016). The genomic relationship matrix $\mathbf{G}$ in the GBLUP model was constructed based on all genomic markers such that $\mathbf{G} = \frac{\mathbf{WW'}}{m}$, where $\mathbf{W}$ is an $n \times m$ genotype matrix (n genotypes and m markers), centered and scaled such that its i^{th} column $\mathbf{w_i} = \frac{(\mathbf{z_i} - 2p_i)}{\sqrt{2p_i(1-p_i)}}$, where $\mathbf{z_i}$ is the i^{th} column vector of $\mathbf{Z}$ having minor allele counts (0, 1, or 2) as entries and p_i is the MAF of the i^{th} marker. In our case, all genotypic locations were homozygous, so genotypes are coded as 0 or 2. For the GFBLUP model, the genomic relationship matrix $\mathbf{G_f}$ for each GO or COEX group was calculated from the markers linked to that group; $\mathbf{G_r}$ was constructed from the remaining markers.

The MultiBLUP model (Equation 4) was constructed according to the Adaptive MultiBLUP strategy proposed by (Speed and Balding, 2014). Briefly, the total genome was divided into adjacent but 50% overlapping regions of 10 kb. The genomic markers within these regions were tested as a group to estimate their association with the phenotype ($p < 10^{-5}$) and adjacent regions were merged if $p_{\text{Bonferroni}} < 0.05$. Subsequently, separate covariance matrices $\mathbf{K}_1$, $\mathbf{K}_2$,..., $\mathbf{K}_M$ were constructed for each region (M regions in total) based on its markers and genomic values $\mathbf{g}_1$, $\mathbf{g}_2$,..., $\mathbf{g}_M$ were estimated. The GRM based on all markers (equivalent to GBLUP) was used if no region was found significant. The total genomic value is $\hat{\mathbf{g}} = \sum_{m=1}^{M} \hat{\mathbf{g}}_m$ with i.i.d. $\mathbf{g}_m \sim N(0, \mathbf{K}_m\sigma^2_m)$ and $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}\sigma^2_e)$:

$$\tilde{\mathbf{y}} = \boldsymbol{\mu} + \sum_{\boldsymbol{m}=1}^{\boldsymbol{M}} \mathbf{g}_{\boldsymbol{m}} + \boldsymbol{\varepsilon} \tag{4}$$

Variance components in all of these LMMs were estimated using the average information restricted maximum-likelihood (REML) procedure (Johnson and Thompson, 1995) implemented in the *greml* method of the R package *qgg* (Rohde et al., 2020) for GBLUP/GFBLUP, using a maximum of 100 iterations at a tolerance level of 10^{-5}; and LDAK v5.1 (http://dougspeed.com/) for MultiBLUP.

Total additive genomic value was predicted using 8-fold cross-validation. This involved training the model using 301 (78%) genotypes and using the remaining 43 for testing in each fold. The exact same accessions were used for both GBLUP and GFBLUP during each split to enable a fair comparison. Prediction accuracy of models was defined as Pearson correlation (r) between observed phenotypic values and predicted genomic values of the test population in each fold. The procedure was repeated 10 times, thus modeled predictive ability distributions consisted of 80 correlations or fewer if variances were over- or underestimated as described earlier by simulation studies (Kruijer et al., 2015). For comparison between models, the median of these correlations was used, and significance of the difference was tested using the non-parametric Wilcoxon–Mann–Whitney test

for assessing significant differences in median accuracy between GBLUP and GFBLUP. Subsequently, *p*-values were adjusted for multiple-testing correction by calculating False Discovery Rate (FDR) based on total number of GO/COEX groups multiplied by total number of time points (Edwards et al., 2016). For Φ_{PSII} we also analyzed results without FDR adjustment, which are referred as "informative" as opposed to "significant" throughout the text.

Model Performance Evaluation

GFBLUP models were compared to the benchmark GBLUP based on their goodness of fit, predictive ability and estimated genomic parameters. Using the likelihood ratio test (LRT) we tested the null-hypothesis $\sigma_f^2 = 0$. LRT *p*-values were based on the asymptotic distribution of the LRT-statistic, which is a mixture of a point mass at 0 and a χ^2-distribution with 1 degree of freedom (d.o.f.) (Edwards et al., 2015). The significantly improved GFBLUP models ($p_{\text{LRT}} < 0.05$) having predictive abilities greater than the benchmark GBLUP (i.e., *p-value* of Wilcoxon-Mann-Whitney tests < 0.05) were filtered for subsequent analysis. Genomic parameters were calculated from variance estimates of both models to analyze only models passing the abovementioned filtering criteria. This includes total genomic heritability explained ($h^2_{GBLUP} = \frac{\sigma_g^2}{(\sigma_g^2+\sigma_\varepsilon^2)}$) and proportion of genomic heritability explained by an individual GO/COEX group in GFBLUP models ($h^2_f = \frac{\sigma_f^2}{(\sigma_f^2+\sigma_r^2+\sigma_\varepsilon^2)}$). In order to check if we obtained a higher number of *PSGENES* in GO/COEX groups than expected by chance, we used the chi-square test with 1 d.o.f. to compare the observed vs. expected frequencies of *PSGENES* in these groups.

Semantic Clustering of GO Terms

Informative GO terms were clustered based on their semantic similarity using the *Revigo* (Supek et al., 2011) web server with *"SimRel"* semantic similarity metric equal to 0.7. The resulting GO clusters were plotted using a Multidimensional Scaling (MDS) plot in R, where maximum %gain in accuracy by each GO term was used to color the bubbles. GO terms enriched in COEX groups were found using the PANTHER classification system (Mi et al., 2019). Fisher's exact test was used for calculating enrichment *p*-values followed by multiple testing correction using the FDR, reporting enrichment at $p < 0.05$. These enriched GO terms were sorted in order of their GO hierarchical tree such that a child term was below its parent; thus, the most specific GO terms are the child GO terms in the bottom of that tree, were used for subsequent analysis.

DATA AVAILABILITY STATEMENT

All data and scripts have been uploaded to the Wageningen University & Research git server (https://git.wur.nl/faroo002/pub1).

AUTHOR CONTRIBUTIONS

MA and T-PN provided the genotype and phenotype datasets. MF performed the analyses. DR, AD, and HN were involved in designing the analyses and interpreting the results. WK helped with statistical analysis. MF wrote the manuscript with DR, AD, HN, and SM. All authors read the final manuscript.

FUNDING

MF was supported by the sandwich Ph.D. programme of Wageningen University and Research (WUR). The authors are grateful for the support of both WUR and NIBGE to conduct this study.

ACKNOWLEDGMENTS

We are thankful to Pádraic J Flood of Plant Breeding, Wageningen University and Research for reviewing the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.609117/full#supplementary-material

Supplementary Figure 1 | Relation between genomic heritability and GBLUP predictive ability. GBLUP prediction accuracy is directly proportional to genomic heritability for both traits. **(A)** shows the relation between heritability and accuracy under low light (LL) and high light (HL) irradiance levels for Φ_{PSII}, **(B)** shows the same for PLA.

Supplementary Figure 2 | GBLUP accuracy (*r*) vs. genomic variance (h^2_{GBLUP}). Each dot corresponds to prediction accuracy (*r*) of GBLUP (*y*-axis) for each split of the data during cross-validation. The genomic variance explained by the model (*x*-axis) ranges from 0 to 1 and calculated as $h^2_{GBLUP} = \frac{\sigma_g^2}{(\sigma_g^2+\sigma_\varepsilon^2)}$. Models at different measurement times are colored differently. **(A)** represents GBLUP models for Φ_{PSII} and contains two separate clouds of dots, representing LL (left) and HL (right) models with different heritability ranges. **(B)** represents GBLUP models for PLA.

Supplementary Figure 3 | MultiBLUP predictive ability. The boxplots show the prediction accuracy (*r*) of MultiBLUP applied to 18 measurements of Φ_{PSII} and 50 measurements of PLA. The average accuracy is slightly lower than the average GBLUP accuracy (white star) for both traits. **(A)** shows the prediction accuracy under low light (LL) and high light (HL) irradiance levels for Φ_{PSII} whereas, **(B)** shows the same for PLA.

Supplementary Figure 4 | Number of genes and markers in GO and COEX features. Total number of genes and markers associated with those genes for both types of genomic features, i.e., GO (left) and COEX (right).

Supplementary Figure 5 | GFBLUP accuracy (*r*) vs. genomic variance (h^2_f) explained by a GO/COEX group. Each dot corresponds to prediction accuracy (*r*) of GFBLUP (y-axis) for each split of data during cross-validation for a particular GO **(A,C)** and COEX **(B,D)** group. The genomic variance explained by the particular GO/COEX (*x*-axis) ranges from 0 to 1. **(A,B)**: GFBLUP models for Φ_{PSII}; **(C,D)**: GFBLUP models for PLA.

Supplementary Figure 6 | GBLUP vs. GFBLUP predictive ability. Average prediction accuracy (*r*) of GBLUP vs. GFBLUP using GO terms **(A,C)** and COEX clusters **(B,D)** for Φ_{PSII} **(A,B)** and PLA **(C,D)**. The average was calculated over 80 splits of the data (8-fold cross-validation repeated 10 times), excluding models

where variance was undetermined). Red dots indicate models that passed our model evaluation criteria (see M&M).

Supplementary Figure 7 | Improvement in genomic prediction performance using informative GO terms for $\varphi_{PSII.}$ All informative GO terms with %gain in accuracy (*r*) of GFBLUP over GBLUP at multiple Φ_{PSII} measurement times, indicated by {Low|High light}{day}_{Number of measurement}. The color bar identifies GO terms as Biological Process (BP), Cellular Component (CC) or Molecular Function (MF).

Supplementary Figure 8 | Improvement in genomic prediction performance using informative COEX groups for φ_{PSII}. All informative COEX clusters with %gain in accuracy (*r*) of GFBLUP over GBLUP at multiple Φ_{PSII} measurement times, indicated by {Low|High light}{day}_{Number of measurement}.

Supplementary Figure 9 | Semantic clustering of GO informed prediction for Φ_{PSII}. Multidimensional scaling (MDS) plot of representative subset (i.e., terms remaining after the redundancy reduction) of informative GO terms molecular functions and cellular components, capable of improving predictive ability of GFBLUP models for Φ_{PSII}. Semantically similar GO terms are clustered based on the "SimRel" semantic similarity measure using *Revigo*. Dot size is proportional to the number of genes annotated with a GO term in the TAIR9 reference genome annotation. The *x* and *y* coordinates indicate relative cluster distances in 2 dimensions. The %gain of a particular GO term is indicated by the bubble color.

Supplementary Figure 10 | Semantic clustering of GO informed prediction for PLA. Multidimensional scaling (MDS) plot of representative subset (i.e., terms remaining after the redundancy reduction) of informative GO terms molecular functions and cellular components, capable of improving predictive ability of GFBLUP models for PLA. Semantically similar GO terms are clustered based on the "*SimRel*" semantic similarity measure using *Revigo*. Dot size is proportional to the number of genes annotated with a GO term in the TAIR9 reference genome annotation. The *x* and *y* coordinates indicate relative cluster distances in 2 dimensions. The %gain of a particular GO term is indicated by the bubble color.

Supplementary Figure 11 | Top 10 enriched GO terms in COEX clusters for Φ_{PSII} and PLA. Top 10 most specific GO terms enriched in 172 informative COEX clusters for the Φ_{PSII} and 355 for PLA traits. The horizontal axis measures the fold enrichment, i.e., the observed fraction of genes annotated with a particular GO term divided by the expected fraction in the reference genome of *Arabidopsis thaliana*. Enrichment *p*-values were found using Fisher's exact test with multiple testing correction using False Discovery Rate (FDR); only terms with $p_{FDR} < 0.05$ are shown.

Supplementary Figure 12 | Minor allele frequency spectrum (MAF). MAF distribution of all 214,051 chip markers. The orange bar represents all markers having MAF<5%, the red bar rare alleles with MAF<1%.

Supplementary Figure 13 | φ_{PSII} phenotypic data distributions using Best Linear Unbiased Estimates (BLUE). Distributions of genotypic means of BLUE values of genotypes in the dataset.

Supplementary Figure 14 | PLA phenotypic data distributions using Best Linear Unbiased Estimates (BLUE). Distributions of genotypic means of BLUE values of genotypes in the dataset.

Supplementary Table 1 | Best Linear Unbiased Estimated Phenotypic data statistics.

Supplementary Table 2a | Informative GO terms increasing GFBLUP prediction accuracy for Φ_{PSII}.

Supplementary Table 2b | GO terms significantly increasing GFBLUP prediction accuracy for PLA.

Supplementary Table 3a | Informative COEX improving GFBLUP prediction accuracy for Φ_{PSII}.

Supplementary Table 3b | COEX significantly improving GFBLUP prediction accuracy for PLA.

Supplementary Table 4 | Genomic features statistics.

Supplementary Table 5 | Enriched Go terms in Φ_{PSII} and PLA COEX analysis.

Supplementary Table 6 | List of genes used in GBLUP based on only photosynthesis genes markers.

Supplementary Table 7a | GO Enrichment of common genes between GO and COEX based analysis for Φ_{PSII}.

Supplementary Table 7b | GO Enrichment of common genes between GO and COEX based analysis for PLA.

Supplementary Table 8 | Raw measurements of Projected Leaf Area.

Supplementary Table 9 | Average best linear unbiased estimates (BLUE) of Projected Leaf Area.

REFERENCES

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics* 25, 25–29. doi: 10.1038/75556

Azodi, C. B., Pardo, J., Vanburen, R., De Los Campos, G., and Shiu, S.-H. (2020). Transcriptome-based prediction of complex traits in maize. *Plant Cell* 32, 139–151. doi: 10.1105/tpc.19.00332

Bates, D., Sarkar, D., Bates, M. D., and Matrix, L. (2007). *The lme4 Package*. R package version 2, 74.

Baxter, I., Brazelton, J. N., Yu, D., Huang, Y. S., Lahner, B., Yakubova, E., et al. (2010). A coastal cline in sodium accumulation in Arabidopsis thaliana is driven by natural variation of the sodium transporter AtHKT1; 1. *PLoS Genetics* 6:e1001193. doi: 10.1371/journal.pgen.1001193

Carlson, M. (2019a). *GO.db: A Set of Annotation Maps Describing the Entire Gene Ontology*. R package version 3.10.10.

Carlson, M. (2019b). *org.At.tair.db: Genome Wide Annotation for Arabidopsis*. R package version 3.10.10.

Crain, J., Mondal, S., Rutkoski, J., Singh, R. P., and Poland, J. (2018). Combining high-throughput phenotyping and genomic information to increase prediction and selection accuracy in wheat breeding. *Plant Genome* 11:43. doi: 10.3835/plantgenome2017.05.0043

De Los Campos, G., Gianola, D., Rosa, G. J., Weigel, K. A., and Crossa, J. (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genet. Res.* 92, 295–308. doi: 10.1017/S0016672310000285

Deng, H., and Runger, G. (2013). Gene selection with guided regularized random forest. *Pattern Recogn.* 46, 3483–3489. doi: 10.1016/j.patcog.2013.05.018

Edwards, S. M., Sorensen, I. F., Sarup, P., Mackay, T. F., and Sorensen, P. (2016). Genomic prediction for quantitative traits is improved by mapping variants to gene ontology categories in *Drosophila melanogaster*. *Genetics* 203, 1871–1883. doi: 10.1534/genetics.116.187161

Edwards, S. M., Thomsen, B., Madsen, P., and Sørensen, P. (2015). Partitioning of genomic variance reveals biological pathways associated with udder health and milk production traits in dairy cattle. *Genet. Select. Evol.* 47:60. doi: 10.1186/s12711-015-0132-6

Ehsani, A., Janss, L., Pomp, D., and Sørensen, P. (2016). Decomposing genomic variance using information from GWA, GWE and eQTL analysis. *Anim. Genet.* 47, 165–173. doi: 10.1111/age.12396

Falconer, D., and Mackay, T. (1996). *Introduction to Quantitative Genetics*. Harlow: Longmans Green 3.

Fang, L., Sahana, G., Ma, P., Su, G., Yu, Y., Zhang, S., et al. (2017). Exploring the genetic architecture and improving genomic prediction accuracy for mastitis and milk production traits in dairy cattle by mapping variants to hepatic transcriptomic regions responsive to intra-mammary infection. *Genet. Select. Evol.* 49:44. doi: 10.1186/s12711-017-0319-0

Flood, P. J., Kruijer, W., Schnabel, S. K., Van Der Schoor, R., Jalink, H., Snel, J. F. H., et al. (2016). Phenomics for photosynthesis, growth and reflectance in *Arabidopsis thaliana* reveals circadian and long-term fluctuations in heritability. *Plant Methods* 12:14. doi: 10.1186/s13007-016-0113-y

Fragomeni, B. O., Lourenco, D. A. L., Masuda, Y., Legarra, A., and Misztal, I. (2017). Incorporation of causative quantitative trait nucleotides in single-step GBLUP. *Genet. Select. Evol.* 49:59. doi: 10.1186/s12711-017-0341-2

Gao, N., Teng, J., Ye, S., Yuan, X., Huang, S., Zhang, H., et al. (2018). Genomic prediction of complex phenotypes using genic similarity based relatedness matrix. *Front. Genet.* 9:364. doi: 10.3389/fgene.2018.00364

Gebreyesus, G., Bovenhuis, H., Lund, M. S., Poulsen, N. A., Sun, D., and Buitenhuis, B. (2019). Reliability of genomic prediction for milk fatty acid composition by using a multi-population reference and incorporating GWAS results. *Genet. Select. Evol.* 51:16. doi: 10.1186/s12711-019-0460-z

Gianola, D. (2013). Priors in whole-genome regression: the bayesian alphabet returns. *Genetics* 194, 573–596. doi: 10.1534/genetics.113.151753

González-Recio, O., Rosa, G. J., and Gianola, D. (2014). Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. *Livestock Sci.* 166, 217–231. doi: 10.1016/j.livsci.2014.05.036

Guo, Z., Magwire, M. M., Basten, C. J., Xu, Z., and Wang, D. (2016). Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. *Theor. Appl. Genet.* 129, 2413–2427. doi: 10.1007/s00122-016-2780-5

Habier, D., Fernando, R. L., and Dekkers, J. C. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177, 2389–2397. doi: 10.1534/genetics.107.081190

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., and Goddard, M. E. (2009). Invited review: genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.* 92, 433–443. doi: 10.3168/jds.2008-1646

Jantzen, S. G., Sutherland, B. J. G., Minkley, D. R., and Koop, B. F. (2011). GO Trimming: Systematically reducing redundancy in large Gene Ontology datasets. *BMC Res. Notes* 4:267. doi: 10.1186/1756-0500-4-267

Jiang, Y., and Reif, J. C. (2015). Modeling epistasis in genomic selection. *Genetics* 201, 759–768. doi: 10.1534/genetics.115.177907

Johnson, D., and Thompson, R. (1995). Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. *J. Dairy Sci.* 78, 449–456. doi: 10.3168/jds.S0022-0302(95)76654-1

Kanehisa, M. (2001). Prediction of higher order functional networks from genomic data. *Pharmacogenomics* 2, 373–385. doi: 10.1517/14622416.2.4.373

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Karaman, E., Cheng, H., Firat, M. Z., Garrick, D. J., and Fernando, R. L. (2016). An upper bound for accuracy of prediction using GBLUP. *PLoS ONE* 11:e161054. doi: 10.1371/journal.pone.0161054

Kourmpetis, Y. A., Van Dijk, A. D., Van Ham, R. C., and Ter Braak, C. J. (2011). Genome-wide computational function prediction of Arabidopsis proteins by integration of multiple data sources. *Plant Physiol.* 155, 271–281. doi: 10.1104/pp.110.162164

Kromdijk, J., Głowacka, K., Leonelli, L., Gabilly, S. T., Iwai, M., Niyogi, K. K., et al. (2016). Improving photosynthesis and crop productivity by accelerating recovery from photoprotection. *Science* 354, 857–861. doi: 10.1126/science.aai8878

Kruijer, W., Boer, M. P., Malosetti, M., Flood, P. J., Engel, B., Kooke, R., et al. (2015). Marker-based estimation of heritability in immortal populations. *Genetics* 199, 379–398. doi: 10.1534/genetics.114.167916

Legarra, A., and Ducrocq, V. (2012). Computational strategies for national integration of phenotypic, genomic, and pedigree data in a single-step best linear unbiased prediction. *J. Dairy Sci.* 95, 4629–4645. doi: 10.3168/jds.2011-4982

Li, B., Zhang, N., Wang, Y. G., George, A. W., Reverter, A., and Li, Y. (2018). Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. *Front. Genet.* 9:237. doi: 10.3389/fgene.2018.00237

Liu, P.-C., Peacock, W. J., Wang, L., Furbank, R., Larkum, A., and Dennis, E. S. (2020). Leaf growth in early development is key to biomass heterosis in Arabidopsis. *J. Exp. Botany* 71, 2439–2450. doi: 10.1093/jxb/eraa006

Long, S. P., Marshall-Colon, A., and Zhu, X.-G. (2015). Meeting the global food demand of the future by engineering crop photosynthesis and yield potential. *Cell* 161, 56–66. doi: 10.1016/j.cell.2015.03.019

Lopez-Cruz, M., Olson, E., Rovere, G., Crossa, J., Dreisigacker, S., Mondal, S., et al. (2020). Regularized selection indices for breeding value prediction using hyper-spectral image data. *Sci. Rep.* 10:8195. doi: 10.1038/s41598-020-65011-2

Macleod, I. M., Bowman, P. J., Vander Jagt, C. J., Haile-Mariam, M., Kemper, K. E., Chamberlain, A. J., et al. (2016). Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genom.* 17:144. doi: 10.1186/s12864-016-2443-6

Mering, C. V., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003). STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* 31, 258–261. doi: 10.1093/nar/gkg034

Meuwissen, T. H. E., Hayes, B., and Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. Available online at: https://www.genetics.org/content/157/4/1819.long

Mi, H., Muruganujan, A., Ebert, D., Huang, X., and Thomas, P. D. (2019). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic acids Res.* 47, D419–D426. doi: 10.1093/nar/gky1038

Monteith, J. L. (1977). Climate and the efficiency of crop production in Britain. *Phil. Trans. R. Soc. London. Biol. Sci.* 281, 277–294. doi: 10.1098/rstb.1977.0140

Morgante, F. (2018). *Genetic Analysis and Prediction of Complex Traits in Drosophila melanogaster*. Ph.D. Thesis, North Carolina State University.

Moser, G., Lee, S. H., Hayes, B. J., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2015). Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. *PLoS Genet.* 11:e1004969. doi: 10.1371/journal.pgen.1004969

Movahedi, S., Van De Peer, Y., and Vandepoele, K. (2011). Comparative network analysis reveals that tissue specificity and gene function are important factors influencing the mode of expression evolution in Arabidopsis and rice. *Plant Physiol.* 156, 1316–1330. doi: 10.1104/pp.111.177865

Nijveen, H., Ligterink, W., Keurentjes, J. J., Loudet, O., Long, J., Sterken, M. G., et al. (2017). Ara QTL-workbench and archive for systems genetics in *Arabidopsis thaliana*. *Plant J.* 89, 1225–1235. doi: 10.1111/tpj.13457

Ogutu, J. O., Piepho, H. P., and Schulz-Streeck, T. (2011). A comparison of random forests, boosting and support vector machines for genomic selection. *BMC Proceed.* 5:S11. doi: 10.1186/1753-6561-5-S3-S11

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., et al. (2013). A large-scale evaluation of computational protein function prediction. *Nat. Methods* 10, 221–227. doi: 10.1038/nmeth.2340

Rimbert, H., Darrier, B., Navarro, J., Kitt, J., Choulet, F., Leveugle, M., et al. (2018). High throughput SNP discovery and genotyping in hexaploid wheat. *PLoS ONE* 13:e0186329. doi: 10.1371/journal.pone.0186329

Rohde, P. D., Demontis, D., Børglum, A., and Sørensen, P. (2017). "Improved prediction of genetic predisposition to psychiatric disorders using genomic feature best linear unbiased prediction models," in *50th European Society of Human Genetics Conference: Posters* (Copenhagen).

Rohde, P. D., Fourie Sørensen, I., and Sørensen, P. (2020). qgg: an R package for large-scale quantitative genetic analyses. *Bioinformatics* 36, 2614–2615. doi: 10.1093/bioinformatics/btz955

Sarup, P., Jensen, J., Ostersen, T., Henryon, M., and Sorensen, P. (2016). Increased prediction accuracy using a genomic feature model including prior information on quantitative trait locus regions in purebred Danish Duroc pigs. *BMC Genet.* 17:11. doi: 10.1186/s12863-015-0322-9

South, P. F., Cavanagh, A. P., Liu, H. W., and Ort, D. R. (2019). Synthetic glycolate metabolism pathways stimulate crop growth and productivity in the field. *Science* 363:77. doi: 10.1126/science.aat9077

Speed, D., and Balding, D. J. (2014). MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.* 24, 1550–1557. doi: 10.1101/gr.169375.113

Supek, F., and Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* 6:e21800. doi: 10.1371/journal.pone.0021800

Togninalli, M., Seren, Ü., Freudenthal, J. A., Monroe, J. G., Meng, D., Nordborg, M., et al. (2020). AraPheno and the AraGWAS Catalog 2020: a major database update including RNA-Seq and knockout mutation data for Arabidopsis thaliana. *Nucleic Acids Res.* 48, D1063–D1068. doi: 10.1093/nar/gkz925

Tong, H., Küken, A., and Nikoloski, Z. (2020). Integrating molecular markers into metabolic models improves genomic selection for Arabidopsis growth. *Nature Commun.* 11, 1–9. doi: 10.1038/s41467-020-16279-5

Uzunangelov, V., Wong, C. K., and Stuart, J. (2020). Highly accurate cancer phenotype prediction with AKLIMATE, a stacked kernel learner integrating multimodal genomic data and pathway knowledge. *bioRxiv [Preprint]*. doi: 10.1101/2020.07.15.205575

Van Bezouw, R. F. H. M., Keurentjes, J. J. B., Harbinson, J., and Aarts, M. G. M. (2019). Converging phenomics and genomics to study natural variation in plant photosynthetic efficiency. *Plant J.* 97, 112–133. doi: 10.1111/tpj.14190

Van Rooijen, R., Aarts, M. G. M., and Harbinson, J. (2015). Natural genetic variation for acclimation of photosynthetic light use efficiency to growth irradiance in Arabidopsis. *Plant Physiol.* 167, 1412–1429. doi: 10.1104/pp.114.252239

Van Rooijen, R., Kruijer, W., Boesten, R., Van Eeuwijk, F. A., Harbinson, J., and Aarts, M. G. M. (2017). Natural variation of YELLOW SEEDLING1 affects photosynthetic acclimation of *Arabidopsis thaliana*. *Nat. Commun.* 8:1421. doi: 10.1038/s41467-017-01576-3

Vanraden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980

Vanraden, P. M., Tooker, M. E., O'connell, J. R., Cole, J. B., and Bickhart, D. M. (2017). Selecting sequence variants to improve genomic predictions for dairy cattle. *Genet. Select. Evol.* 49:32. doi: 10.1186/s12711-017-0307-4

Wang, J., Zhou, Z., Zhang, Z., Li, H., Liu, D., Zhang, Q., et al. (2018). Expanding the BLUP alphabet for genomic prediction adaptable to the genetic architectures of complex traits. *Heredity*. 121, 648–662. doi: 10.1038/s41437-018-0075-0

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., et al. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* 38, W214–W220. doi: 10.1093/nar/gkq537

Weraduwage, S. M., Chen, J., Anozie, F. C., Morales, A., Weise, S. E., and Sharkey, T. D. (2015). The relationship between leaf area growth and biomass accumulation in *Arabidopsis thaliana*. *Front. Plant Sci.* 6:167. doi: 10.3389/fpls.2015.00167

Wheeler, H. E., Aquino-Michaels, K., Gamazon, E. R., Trubetskoy, V. V., Dolan, M. E., Huang, R. S., et al. (2014). Poly-omic prediction of complex traits: OmicKriging. *Genetic Epidemiol.* 38, 402–415. doi: 10.1002/gepi.21808

Zhang, M., Hu, X. L., Zhu, M., Xu, M. Y., and Wang, L. (2017). Transcription factors NF-YA2 and NF-YA10 regulate leaf growth via auxin signaling in Arabidopsis. *Sci. Rep.* 7:1475. doi: 10.1038/s41598-017-01475-z

Zhang, X., and Borevitz, J. O. (2009). Global analysis of allele-specific expression in *Arabidopsis thaliana*. *Genetics* 182, 943–954. doi: 10.1534/genetics.109.103499

Zhang, Z., Ding, X., Liu, J., De Koning, D. J., and Zhang, Q. (2011). Genomic selection for QTL-MAS data using a trait-specific relationship matrix. *BMC Proceed.* 5:S15. doi: 10.1186/1753-6561-5-S3-S15

Zhang, Z., Liu, J., Ding, X., Bijma, P., De Koning, D. J., and Zhang, Q. (2010). Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. *PLoS ONE* 5:e12648. doi: 10.1371/journal.pone.0012648

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.