G3 Genes | Genomes | Genetics

QTG-Finder2: A Generalized Machine-Learning Algorithm for Prioritizing QTL Causal Genes in Plants

Fan Lin, Elena Z. Lazarus, and Seung Y. Rhee[1]
Department of Plant Biology, Carnegie Institution for Science, Stanford, California 94305
ORCID IDs: 0000-0003-3742-5291 (F.L.); 0000-0001-7572-2516 (E.Z.L.); 0000-0002-7572-4762 (S.Y.R.)

ABSTRACT Linkage mapping has been widely used to identify quantitative trait loci (QTL) in many plants and usually requires a time-consuming and labor-intensive fine mapping process to find the causal gene underlying the QTL. Previously, we described QTG-Finder, a machine-learning algorithm to rationally prioritize candidate causal genes in QTLs. While it showed good performance, QTG-Finder could only be used in Arabidopsis and rice because of the limited number of known causal genes in other species. Here we tested the feasibility of enabling QTG-Finder to work on species that have few or no known causal genes by using orthologs of known causal genes as the training set. The model trained with orthologs could recall about 64% of Arabidopsis and 83% of rice causal genes when the top 20% ranked genes were considered, which is similar to the performance of models trained with known causal genes. The average precision was 0.027 for Arabidopsis and 0.029 for rice. We further extended the algorithm to include polymorphisms in conserved non-coding sequences and gene presence/absence variation as additional features. Using this algorithm, QTG-Finder2, we trained and cross-validated *Sorghum bicolor* and *Setaria viridis* models. The *S. bicolor* model was validated by causal genes curated from the literature and could recall 70% of causal genes when the top 20% ranked genes were considered. In addition, we applied the *S. viridis* model and public transcriptome data to prioritize a plant height QTL and identified 13 candidate genes. QTL-Finder2 can accelerate the discovery of causal genes in any plant species and facilitate agricultural trait improvement.

KEYWORDS
Sorghum bicolor
Setaria viridis
machine learning
quantitative trait loci
causal genes

Improving crop production to address the rapid increase in the global food demand, combined with increasing limitations of arable land, remains a major challenge. The world population is expected to exceed 9 billion by 2050 and will require a 70% increase in global food production (FAO 2009). Between 1985 and 2005 the world's croplands increased by only about 2.4% (Foley *et al.* 2011). Therefore, a significant increase of crop yield is required to feed the growing population, especially when there are increasing uncertainties of the changing climate.

Two common approaches used to improve crop yield and other agriculturally important traits are plant breeding and genome editing. Although plant breeding contributed significantly to crop yield improvement in the past century, it is facing obstacles such as limited sources of genetic variation and time-consuming phenotyping and germplasm evaluation (Rodríguez-Leal *et al.* 2017). With the advancement of CRISPR/Cas9 technology, genome editing has become a much faster way to enhance crop traits and it is possible to introduce novel alleles for a single gene via targeted mutagenesis (Rodríguez-Leal *et al.* 2017). However, genome editing will require identification of the trait-associated genes or the causal variants (Ramstein *et al.* 2019). Many trait-associated causal genes in quantitative trait loci (QTL) have been validated by mutational analysis and functional complementation experiments (Weigel and Nordborg 2005).

As one of the most commonly used genetic mapping methods, thousands of QTL mapping studies have been conducted on many crops (Yonemaru *et al.* 2010; Mace *et al.* 2019) but the causal genes for most of these QTLs have not been identified or validated by experiments. For example, there are less than one hundred curated causal genes that have been cloned and validated by complementation experiments in Arabidopsis or rice, and there are even fewer known causal genes in other plant species (Martin and Orgogozo 2013). Identifying causal genes from hundreds to thousands of genes in a

doi: https://doi.org/10.1534/g3.120.401122
Manuscript received February 3, 2020; accepted for publication May 12, 2020; published Early Online May 18, 2020.

Supplemental material available at figshare: https://doi.org/10.25387/g3.11789646.
[1] Corresponding author: Department of Plant Biology, Carnegie Institution for Science, Stanford, California 94305. E-mail: srhee@carnegiescience.edu.

QTL region usually requires a great amount of time and effort to fine map the causal gene (Huang *et al.* 2016a). Therefore, a computational method to predict or prioritize causal genes will be helpful for accelerating the discovery of novel trait-associated genes in QTLs.

Previously we developed a machine-learning based algorithm, named QTG-Finder, to prioritize causal genes in QTLs (Lin *et al.* 2019). The algorithm uses additional information such as polymorphisms from re-sequencing data, function annotation, co-function network, gene essentiality and paralog copy number to prioritize causal genes in QTLs. We trained models for *Arabidopsis thaliana* (Arabidopsis) and *Oryza sativa* (rice) with curated causal genes in each species. Based on validation using an independent set of newly curated genes, the models could recall about 64% of Arabidopsis and 79% of rice causal genes when the top 20% ranked genes in a QTL were considered. However, the models were only developed for Arabidopsis and rice and were trained on a relatively small number of known causal genes from each species. There were insufficient numbers of known causal genes in other plants to develop such predictive models.

To devise an algorithm that would work even on species with few or no training data available, we wondered whether orthology could be used to create or extend training data. This idea was based on several factors. First, many causal genes identified by linkage mapping are evolutionary hotspots (Martin and Orgogozo 2013). Second, in plants and animals, some genes have repeatedly been major components of phenotypic variation of similar traits (Gompel and Prud'homme 2009; Kopp 2009). For example, *Flowering Time* (*FT*) has been reported to be a causal gene of flowering time QTLs in Arabidopsis (Kojima *et al.* 2002; Schwartz *et al.* 2009), barley (Yan *et al.* 2006), wheat (Yan *et al.* 2006), sunflower (Blackman *et al.* 2010) and ryegrass (Skøt *et al.* 2011). There are many other examples of conservation in causal genes for the same trait across species (Martin and Orgogozo 2013). Therefore, we hypothesized that the orthologs of causal genes are also likely to be causal genes.

We tested this hypothesis by training models in Arabidopsis and rice with orthologs of known causal genes. The performance indicated the feasibility of this approach. We further tested the approach by training models for *Sorghum bicolor* (sorghum) and *Setaria viridis* (Setaria), which have only few known causal genes. We validated the sorghum model by testing QTLs with known causal genes curated from the literature. We also demonstrated the usage of the Setaria model by combining the prioritization results with published transcriptome data to obtain 13 causal gene candidates for a Setaria height QTL.

MATERIALS AND METHODS

The orthologs of known causal genes

The list of causal genes used for orthology analysis was based on a list of causal alleles previously published (Martin and Orgogozo 2013). Since the original list only provided the gene name of those causal genes, we first curated their gene ID or UniProt ID from the references cited. When the ID was not available in the papers, we searched the gene name in genome annotation databases such as RAP-DB (https://rapdb.dna.affrc.go.jp), maizeGDB (https://www.maizegdb.org), soyKB (http://soykb.org/) or the UniProt database (https://www.uniprot.org). The gene ID or UniProt ID was used as a query to search the EggNOG database (v4.5.1) (Huerta-Cepas *et al.* 2016) to obtain the ortholog group to which it belongs and its fine-grained orthologs. Fine-grained orthologs in EggNOG are defined as orthologs derived from a pairwise orthology between members of two species in an orthologous group based on phylogenic analysis. For genes that were not found in EggNOG, we obtained their protein sequences from UniProt or Genbank and used a HMMER-based sequence search (http://eggnogdb.embl.de/#/app/seqscan) to find the ortholog group. When available, the fine-grained orthologs were used as the orthologs of causal genes. When fine-grained orthologs were not available, all members in the ortholog group were used as orthologs. We examined the orthology in major crops and model organisms of eudicots and monocots: *Arabidopsis thaliana*, *Solanum lycopersicum*, *Brassica rapa*, *Glycine max*, *Oryza sativa japonica*, *Oryza sativa indica*, *Setaria italica*, *Sorghum bicolor*, *Brachypodium distachyon*, *Hordeum vulgare* and *Zea mays* (Supplemental Table S1).

We obtained the ortholog list for *Setaria viridis* in a different way because the EggNOG database only includes *S. italica* genes and not *S. viridis* genes. Since *S. italica* is a domesticated line derived *from S. viridis* and has excellent collinearity with *S. viridis* (Supplemental Figure S1), (Bennetzen *et al.* 2012), we used DAGChainer (Haas *et al.* 2004) in CoGe to identify collinear gene pairs that fall in contiguous chains between *S. viridis* and *S. italica*. The DAGChainer results allowed us to convert *S. italica* gene IDs to *S. viridis* gene IDs (Supplemental Table S1).

Building new features based on polymorphisms in conserved non-coding regions and gene presence/absence for Arabidopsis and rice models

The conserved non-coding regions, the Conserved Elements (CE) and the Transcription Factor Binding Sites (TFBS), were downloaded from the Plant Transcriptional Regulatory Map (PlantRegMap, last modified on 2019-10-11, http://plantregmap.cbi.pku.edu.cn/). The CEs were identified based on the genome alignments of plants (Jin *et al.* 2014). The TFBSs were based on the correlation between frequencies in binding motifs and conservation scores (Tian *et al.* 2019). The TFBSs and CEs were assigned to genes that were located within 1kb upstream or downstream of the genes. We incorporated the TFBSs and CEs as regulatory annotations in SnpEff (v 4.3r) and identified SNPs and Indels in these TFBSs or CEs. We counted the number of SNPs and Indels in these conserved non-coding regions for each gene and built four features: CE_snp, TFBS_snp, CE_indel, TFBS_indel.

The gene presence/absence data were obtained from published studies. The Arabidopsis gene presence/absence data were based on 80 *A. thaliana* accessions with 10 to 24x sequencing coverage (Tan *et al.* 2012). The rice gene presence/absence data were based on 453 *O. sativa* accessions with sequencing depth of over 20x (Hu *et al.* 2018). The percentage of absence across the sequenced accessions was calculated for each gene and used as the feature "percent_absence".

Features and model training of sorghum and Setaria models

Features for sorghum and Setaria models were generated as follows. Polymorphism features were extracted in the same way as previously described (Lin *et al.* 2019). Briefly, SNP data were annotated by SIFT4G (v 2.4) and SnpEff (v 4.3r) and assigned to each gene in the genome. The polymorphism features were mostly binary features that represent the presence of a specific type of SNP for each gene. For example, if a gene contained any deleterious non-synonymous SNPs, the "is_nonsyn_deleterious" feature was set to 1, otherwise it was set to 0. The sorghum SNP data were downloaded from Sorghum Genome SNP Database (SorGSD) (Luo *et al.* 2016), which provides SNP data for a diverse panel of 48 sorghum lines. The Setaria SNP data and gene presence/absence data are based on a diverse panel of 598 *S. viridis* accessions (Huang *et al.* 2019). Since there is no

Table 1 Curated *S. bicolor* QTL causal genes and external validation results

QTL Trait	Gene name	Gene ID	Genes in QTL	Percent rank	Reference
Light sensitivity	*phyB*	Sobic.001G394400	144	1%	Yang *et al.* 2014
Brown midrib	*bmr2*	Sobic.004G062500	27	3%	Saballos *et al.* 2012
Amylose	*Wx*	Sobic.010G022600	706	3%	Boyles *et al.* 2017
Fungus resistance	*Ds1*	Sobic.005G065000	389	15%	Kawahigashi *et al.* 2011
Plant height	*Dw2*	Sobic.006G067700	335	15%	Hilley *et al.* 2017
Seed shattering	*Sh1*	Sobic.001G199200	117	18%	Lin *et al.* 2012
Aluminum tolerance	*MATE*	Sobic.003G403000	26	19%	Magalhaes *et al.* 2007
Light sensitivity	*ghd7*	Sobic.006G004400	115	61%	Murphy *et al.* 2014
Pollen fertility	*PPR*	Sobic.002G057050	26	69%	Jordan *et al.* 2010
Flowering time	*PRR37*	Sobic.006G057900	22	77%	Murphy *et al.* 2011

pre-built SnpEff database for *S. viridis*, we built a *S. viridis* database using the .gff file of *S. viridis* v2.1 downloaded from Phytozyme12 (https://phytozome.jgi.doe.gov).

The Gene Ontology (GO) annotations for sorghum and Setaria were obtained from PLAZA4.0 (https://bioinformatics.psb.ugent.be/plaza/versions/plaza_v4_monocots/) (Van Bel *et al.* 2018). We used GOslim (http://current.geneontology.org/ontology/subsets/goslim_metagenomics.obo) to aggregate the molecular function GO terms to higher-level terms such as is_transporter and is_transcription_factor. For genes encoding an enzyme, we further determined their metabolic domains based on Plant Metabolic Network databases (PMN, release 12.5) (Schläpfer *et al.* 2017).

The paralog copy number of each gene was determined by OrthoFinder (v2.3.3) (Emms and Kelly 2019). We used the DIAMOND algorithm and the default setting of OrthoFinder. Protein sequences of thirteen species or subspecies were downloaded from PLAZA4.0, including *Arabidopsis thaliana*, *Brachypodium distachyon*, *Brassica rapa*, *Glycine max*, *Hordeum vulgare*, *Oryza sativa japonica*, *Oryza sativa indica*, *Populus trichocarpa*, *Sorghum bicolor*, *Setaria viridis*, *Setaria italica*, *Solanum lycopersicum* and *Zea mays*. The paralogs counted for each species in the orthologous group were determined by OrthoFinder.

The Setaria and sorghum models were trained with orthologs of causal genes from any other plant species. To train the models, we used the random forest algorithm, which is an ensemble learning method that fits a number of decision trees on various sub-sampled datasets (Ho 1998). Random forest integrates the votes from these trees to improve accuracy and reduce the chance of over-fitting. Model parameters including the number of trees, maximum number of features to consider when seeking for the best split, the minimum number of samples required to split a node in a decision tree and the ratio of positives and negatives in the training set were used to optimize the models to maximize cross-validation AUC-ROC (Supplemental Figures S2 and S3).

Cross-validation, external validation and feature importance analysis

The methods for model training, cross-validation, external validation and feature importance analysis were the same as previously described (Lin *et al.* 2019). Briefly, we used random forest as the machine learning algorithm backbone. We split the data into training and testing sets using a 5-fold cross-validation and validated the model with an independent, external dataset. Feature importance was measured by the reduction of AUC-ROC after removing each feature from the models. The causal genes used for external validation were not used for training the QTG-Finder models. However, some of them are orthologs of known causal genes in other species, which could have been used for training QTG-Finder2 models. We therefore excluded these orthologs from the training set of QTG-Finder2 to avoid over-estimation of model performance.

The external validation of the sorghum model was conducted on a set of causal genes curated from the literature (Table 1) (Magalhaes *et al.* 2007; Jordan *et al.* 2010; Kawahigashi *et al.* 2011; Murphy *et al.* 2011; Lin *et al.* 2012; Saballos *et al.* 2012; Murphy *et al.* 2014; Yang *et al.* 2014; Boyles *et al.* 2017; Hilley *et al.* 2017). We applied the model to all genes in the QTL regions, which were defined by the literature.

We used Fisher's exact test for the pairwise comparison of external validation results. Each gene in the external validation set was assigned to one of two classes: (1) the gene was included in the prioritized fraction (*e.g.*, top 5%, 10% or 20%), or (2) the gene was not included in the prioritized fraction. The number of genes in the two classes (prioritized *vs.* not prioritized) was used for Fisher's exact tests.

Sequence alignment for candidate genes

Multiple sequence alignments were performed using Clustal Omega (v1.2.4, https://www.ebi.ac.uk/Tools/msa/clustalo/) across grass species including *Brachypodium distachyon*, *Panicum virgatum*, *Oryza sativa*, *Setaria viridis*, *Setaria italica*, *Sorghum bicolor* and *Zea mays*. Homologous protein sequences in these species were obtained from Phytozyme12. Human and yeast RIO2 sequences were obtained from NCBI (https://www.ncbi.nlm.nih.gov).

For promoter sequence comparison, we used pairwise global sequence alignment (EMBOSS Needle, https://www.ebi.ac.uk/Tools/psa/emboss_needle/). The promoter sequences were defined as 1kb upstream of the coding sequence (CDS) and downloaded from Phytozome12. We further examined the putative Transcription Factor Binding Sites (TFBS) in the promoters. TFBSs were predicted by Plant Transcriptional Regulatory Map tool (PlantRegMap, http://plantregmap.cbi.pku.edu.cn/binding_site_prediction.php).

Data availability

The source code and training data for Arabidopsis, rice, sorghum and Setaria are available at Github (https://github.com/carnegie/QTG_Finder). The pre-trained QTG2-Finder2 models are available at Dryad (https://doi.org/10.5061/dryad.hhmgqnkdj). All models were trained and tested using Python 3.7.3 and scikit-learn 0.21.2. Supplemental material available at figshare: https://doi.org/10.25387/g3.11789646

RESULTS

Incorporating an orthology approach in the QTG-Finder2 algorithm

The QTG-Finder algorithm described previously only used known causal genes of a single species to train a model of that species (Figure

1A). For QTG-Finder2, we trained models with not only the known causal genes in the target species but also the orthologs of causal genes from other species (Figure 1B).

With 253 curated causal genes from any plant species (Figure 2), we identified their orthologs in 12 species and subspecies: *Arabidopsis thaliana*, *Solanum lycopersicum*, *Brassica rapa*, *Glycine max*, *Oryza sativa japonica*, *Oryza sativa indica*, *Setaria italica*, *Setaria viridis*, *Sorghum bicolor*, *Brachypodium distachyon*, *Hordeum vulgare* and *Zea mays* (Supplemental Table S1). The 12 species included major crops and model organisms of eudicots and monocots. Each species had orthologs for about 60% of the causal genes (Supplemental Figure S4A). Some causal genes had multiple orthologs and the average number of orthologs varied across species (Supplemental Figure S4B).

Testing Arabidopsis and rice models trained with orthologs

We asked if the models trained with orthologs would perform as well as models trained with only the known causal genes in the target species. To test this hypothesis, we trained models in Arabidopsis and rice using three different sets of positive training data: 1) only known causal genes in the target species, 2) only orthologs, and 3) known causal genes plus orthologs. For the Arabidopsis model, we used 60 known causal genes from Arabidopsis and 146 orthologs of causal genes from other species (Figure 3A). In the rice model, we used 45 known causal genes from rice and 206 orthologs of causal genes from other species. The negative train sets included genes that were randomly selected from the genome as described previously (Lin *et al.* 2019).

We first performed cross-validation to evaluate models that were trained with the three training sets (Figure 3B). We used the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) to evaluate the training performance of these models. For both species, models trained with any of the training sets was significantly higher than the models trained with randomly selected genes (Figure 3B, One-way ANOVA followed by Tukey HSD post-hoc test, p-value < 0.05). In Arabidopsis, the models trained with only the known causal genes had the highest average AUC-ROC score (0.86). The model trained with only orthologs had an average AUC-ROC of 0.82, which is comparable to the model trained with known causal genes. The model trained on both the orthologs and known causal genes had an average AUC-ROC of 0.81, which was not distinguishable from the model with just the orthologs. This indicates that orthologs have similar properties as known causal genes. Compared to these scores, the model trained with randomly selected genes had an average AUC-ROC of 0.52. In rice, the model trained with only orthologs had the highest average AUC-ROC of 0.84. This is not simply due to the sample size increasing since this trend was not observed in Arabidopsis where the sample size was also increased in the ortholog training data. Interestingly, the model trained with only the known genes showed the lowest score of 0.73. Combining the orthologs with the known causal genes for training gave a score of 0.81. Since the models trained with only orthologs had significantly higher AUC-ROC than models trained with random genes, these results indicate the orthologs by themselves will be useful for model training. The F1 scores were 0.23 for the Arabidopsis model (precision = 0.33 and recall = 0.19) and 0.05 for the rice model (precision = 0.036 and recall 0.22).

After optimizing model parameters from cross-validation results, we evaluated the final models with an external validation set, which contained independently curated causal genes that had not been seen by the models. The validation method and data set were the same as previously described (Lin *et al.* 2019). For each causal gene in the validation set, the models were applied to rank all genes in the QTL region where the causal gene is located. We applied the model to rank and prioritize the top 5%, 10% and 20% of genes in the QTL region and examined if the known causal gene was included in the prioritized gene list. All models performed significantly better than the models trained with randomly selected genes (Figure 4). The models trained with only orthologs not only performed significantly better than background, but also were not different from models trained with just the known causal genes. For Arabidopsis, the model trained with only orthologs could recall 27%, 36% and 64% of causal genes at the top 5%, 10% and 20% cutoffs, respectively (Figure 4). For rice, the model trained with only orthologs can recall 28%, 56% and 83% of causal genes at the top 5%, 10% and 20% cutoffs, respectively (Figure 4). At all three cut-offs, the Arabidopsis models that combined the Arabidopsis causal genes and orthologs performed better than models using either causal genes or orthologs alone, though not significantly (Fisher's exact test, p-value>0.05). For rice, all three models performed similarly. All the Arabidopsis and rice models performed significantly better than the background at 10% and 20% cutoffs (Fisher's exact test, p-value < 0.05) but not at 5% cutoff. The background was determined as the theoretical probability of

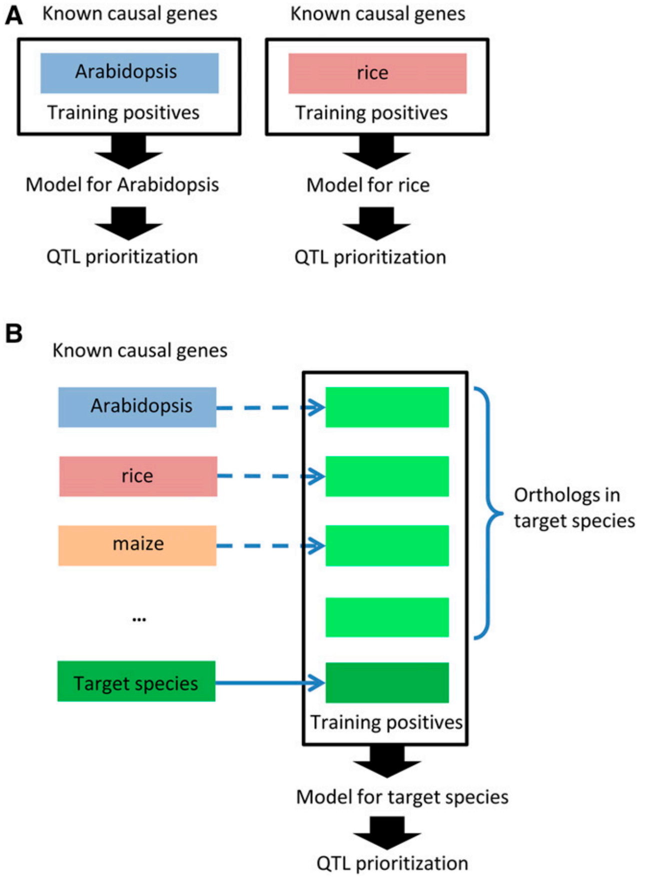

Figure 1 Incorporating an orthology approach to the QTG-Finder algorithm facilitates training models in other plant species (A) The original QTG-Finder algorithm. Only the known causal genes were used to train a model for a given species. (B) QTG-Finder2 algorithm. Orthologs of the known causal genes from any species were also used to train a model. This method allows QTG-Finder to be implemented in species without enough or any known causal genes.

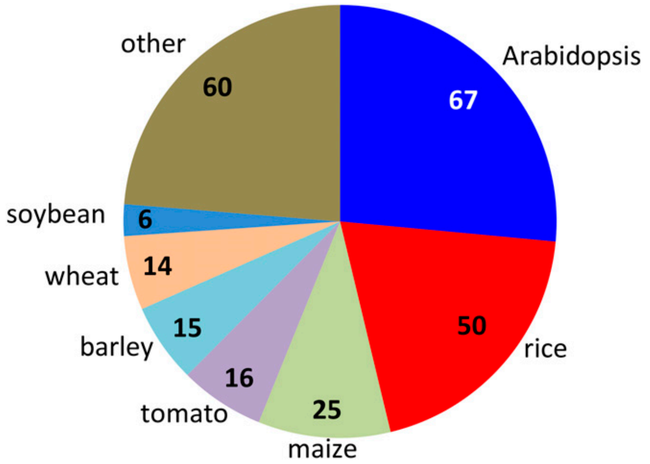

Figure 2 Most plant species do not have enough curated known causal genes to train a model as Arabidopsis or rice does. Numbers indicate the number of known causal genes in each species. Arabidopsis, *Arabidopsis thaliana*; rice, *Oryza sativa japonica*; maize, *Zea mays*; tomato, *Solanum lycopersicum*; barley, *Hordeum vulgare*; wheat, *Triticum aestivum*; soybean, *Glycine max*. Data from Martin and Orgogozo, 2013.

including the causal gene when we randomly selected 5%, 10% or 20% of the genes from the QTL region. For Arabidopsis, the average precisions were 0.073, 0.039, and 0.027 at the top 5%, 10%, and 20% cutoffs. For rice, the average precisions were 0.057, 0.037, and 0.029 at the top 5%, 10%, and 20% cutoffs (Supplemental Table S2).

To determine if different levels of orthology affected performance, we compared the orthology method described above with two alternative methods: 1) using taxon-constrained orthologs and 2) using only EggNOG's fine-grained orthologs defined as orthologs derived from a pairwise orthology between members of two species in an orthologous group based on phylogenic analysis (Huerta-Cepas *et al.* 2016). For the taxon-derived orthology method, we only considered the orthologs for species that are in the same lineage (monocot or eudicot) as the causal gene being queried. For example, if the known causal gene were identified in a monocot species, then we would only consider its orthologs in monocot species. Using the same external validation set, we compared the original orthology method with these two methods and found that their performance was similar to each other (Fisher's exact test, p-value>0.05) (Supplemental Figure S5).

Models trained with both the known causal genes in the species and orthologs from other species represent a more generalized model since it combines information from known causal genes in target species and information from causal genes in other species. The models trained with known causal genes plus orthologs performed similarly as the models trained with only orthologs in cross-validation and external validation (Figures 3 and 4). Therefore, we used models trained with known causal genes plus orthologs for subsequent analyses.

Exploring and adding new features to QTG-Finder2

We explored new features that may help distinguish causal genes from other genes such as polymorphisms in conserved non-coding regions and structural variations such as gene presence/absence. SNPs or Indels in some conserved non-coding sequences may disrupt transcription factor binding and influence gene expression patterns and traits. In addition, gene presence/absence variation has been linked to phenotypic variations. For example, causal genes like *RLM3* and *FRIGIDA* in Arabidopsis (Werner *et al.* 2005; Staal *et al.* 2008), *Sub1A* in rice (Xu *et al.* 2006) and *ZCCT1* and *ZCCT2* in barley (Yan *et al.* 2004) are absent in some accessions.

To generate features from polymorphisms in non-coding sequences, we used two types of predicted non-coding sequences in PlantRegMap (Tian *et al.* 2019): Conserved Elements (CE) and functional Transcription Factor Binding Sites (TFBS). In Arabidopsis and rice, there are significantly more SNPs and Indels in the CEs nearby causal genes than in the CEs nearby an average gene in the genome (Mann-Whitney U Test, p-value <0.05, Figure 5A, Supplemental Tables S3 and S4). However, the SNPs or Indels in TFBS were not significantly different between causal genes and non-causal genes (Mann-Whitney U Test, p-value >0.05, Figure 5A, Supplemental Tables S3 and S4).

In addition, we constructed a percent absence feature using previously published gene presence/absence analyses (Tan *et al.* 2012; Hu *et al.* 2018). In both Arabidopsis and rice, the causal genes had significantly higher percent absence than genome genes (Mann-Whitney U Test, p-value <0.05, Figure 5A, Supplemental Tables S3 and S4).

We were encouraged by the enrichment of the CEs and presence/absence variation in the causal genes and added them as new features. However, these new features did not change the model performance significantly (Fisher's exact test, p-value>0.05). We first compared the external validation results for models with or without the new features (Figure 5B). Then, we examined the feature importance of

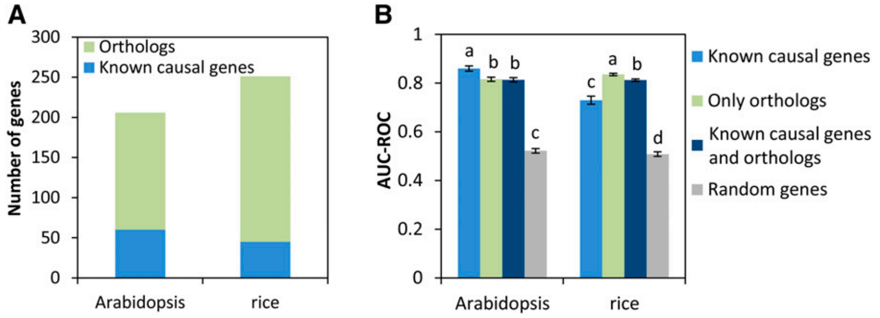

Figure 3 Models trained with orthologs have comparable performance as the models trained with known causal genes (A) The number of known causal genes and orthologs derived from causal genes in any plant species (B) Cross-validation of models trained with known causal genes, with orthologs, or with causal genes plus orthologs. AUC-ROC (Area Under the Curve - Receiver Operating Characteristic) was used to compare training performance of the models. Error bars represent standard deviation, N = 50 for each bar. One-way ANOVA followed by Tukey HSD post-hoc test was performed to determine the statistical difference ($P < 0.05$) among the groups as represented by letters.

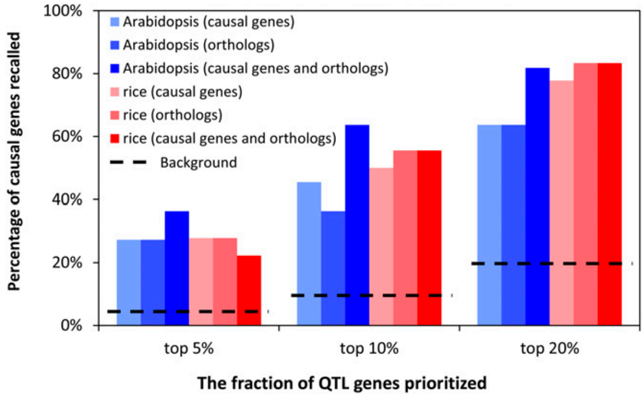

Figure 4 Models trained only with orthologs have similar performance as the models trained with known causal genes based on external validation. Model performance was evaluated when the top 5%, 10% or 20% of the ranked QTL genes were considered. Fisher's exact test was performed between the models trained only with known causal genes *vs.* the models trained with only orthologs or the models trained with causal genes plus orthologs. No statistical difference was detected ($P > 0.05$). The black dashed lines indicate the theoretical background estimated by the percentage of causal genes being recalled when we randomly selected 5%, 10% or 20% of the genes from the QTL region.

those new features by using a leave-one-out analysis (Supplemental Figure S6). Paralog copy number remains to be the most important feature, which is consistent with the previous version of QTG-Finder (Lin *et al.* 2019). The CE_snp feature was the fourth most important feature in the Arabidopsis model. The other new features were not within the top 5 most important features. Since these new features do not reduce model performance, we kept them in the algorithm for subsequent analyses.

Applying QTG-Finder2 to train sorghum and Setaria models

To demonstrate that the QTG-Finder2 algorithm can be used to train models for species that have few or no known causal genes, we trained and cross-validated models in *Setaria viridis* (Setaria) and *Sorghum bicolor* (sorghum) with the orthologs derived from causal genes in other species. *S. bicolor* is an important C4 photosynthesis crop with excellent drought resistance (Calviño and Messing 2012). *S. viridis* is a C4 photosynthesis model grass and the wild ancestor of foxtail millet (*Setaria italica*), an important crop in Asia and Africa (Huang *et al.* 2016b).

We first conducted cross-validation for the Setaria and sorghum models (Figure 6A). The AUC-ROCs were 0.79 (Setaria model) and 0.77 (sorghum model), respectively, which were reasonable, though lower than the Arabidopsis and rice models trained on causal genes. The precisions were 0.12 (Setaria model) and 0.1 (sorghum model) at a recall of 20%.

With the models that were trained only with causal gene orthologs, we performed external validation. Since there is insufficient data to perform external validation for Setaria, we performed external validation only for the sorghum model. We curated ten sorghum causal genes from the literature (Table 1). When the top 5%, 10% and 20% of the genes in the QTL region were prioritized by the sorghum model (Supplemental Table S5), 30%, 30% and 70% of the causal genes were recalled, respectively (Figure 6B). The precisions were 0.065, 0.041, and 0.044 for top 5%, 10%, and 20% (Supplemental Table S5). The sorghum model's performance was similar to the Arabidopsis model, which recalled 27%, 36%, and 64% of the causal genes when the top 5%, 10%, and 20% of the genes in the QTL were prioritized. While the performance of the sorghum model was lower than that for the rice model, there was no statistical difference in the performance between sorghum and Arabidopsis or rice models. All models performed significantly better than the background where the same number of genes were randomly prioritized at 10% and 20% cutoffs (Fisher's exact test, p-values > 0.05, Figure 5B) but not at the 5% cutoff.

Combining the Setaria model and transcriptome data to prioritize causal genes for a Setaria height QTL

Since there are no cloned QTL causal genes available for Setaria, we could not evaluate the Setaria model's performance with independent data. To demonstrate the usage of the Setaria model, we applied it to prioritize a well-determined plant height QTL in Setaria. This QTL is located on chromosome 5 and has been reported by two independent studies (Mauro-Herrera and Doust 2016; Feldman *et al.* 2017) and has large effects on height under many conditions such as different watering levels and density of planting (Feldman *et al.* 2017). There are 335 genes in the LOD1.5 interval of this major QTL (Feldman *et al.* 2017).

To select a testable number of candidates for this height QTL, we combined the Setaria model with published transcriptome data (Martin *et al.* 2016). We first applied the Setaria model to the QTL and prioritized 67 genes that ranked within the top 20%. Given that the experimental validation for this number of candidates would still constitute a large effort at this time, we incorporated transcriptome data to further narrow down the candidate gene list. Based on a transcriptome study on the developing internode of *S. viridis*, we selected genes that were up-regulated by more than 2-fold in the internode meristem or cell elongation zone relative to the maturation zone. We posited that genes that were up-regulated in these zones are more likely to be involved in internode elongation and therefore contribute to plant height. In the QTL interval, 60 genes were up-regulated either in the meristem or elongation zone relative to the maturation zone (Figure 7A). By comparing the top 20% of the prioritized genes with the up-regulated genes, we found 13 genes that met both criteria (Figure 7B, Supplemental Table S6).

In addition to the 13 candidates we prioritized, there is one gene (*Semidwarf*, *SD1*, Sevir.5G410400) in this QTL interval, which was suggested to be a putative causal gene, though it has not been experimentally validated. *SD1* encodes gibberellin20 oxidase2 in rice, involved in gibberellin biosynthesis, and a loss of function allele gives a dwarf phenotype in rice (Spielmeyer *et al.* 2002). The putative causal gene *SD1* has a percent rank of 24% according to the prediction of our Setaria model. Though not within the top 20% ranked genes, *SD1* is up-regulated in the meristem zone of *S. viridis* internode (Martin *et al.* 2016). Therefore, *SD1* could also be considered as a candidate gene.

We next examined if any of these candidate genes had changes in protein sequence or gene expression patterns in the parental lines, *S. viridis* and *S. italica*. SD1 and the proteins encoded by four candidate genes have differences in the protein sequence between *S. viridis* and *S. italica* (Supplemental Table S6). One candidate Sevir.5G413600 (its *S. italica* ortholog, Seita.5G407900) is particularly interesting because the encoded protein contains four amino acid replacements between *S. viridis* and *S. italica*, which change the physicochemical property in a conserved C-terminal domain (Supplemental Figures S7 and S8). The protein is most similar to RIO2 kinase/ATPases. RIO2 proteins are widely conserved from archaea to eukaryotes and are involved in the maturation of small ribosome

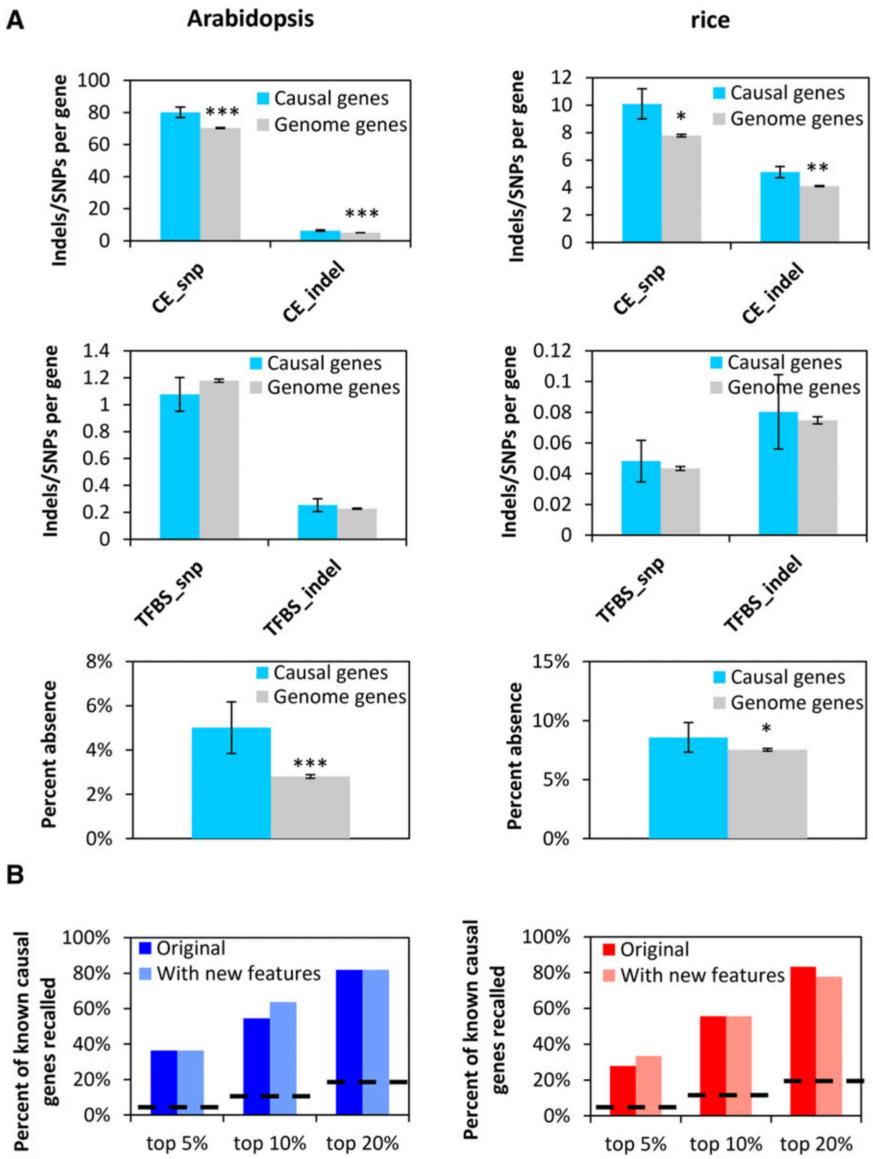

Figure 5 Exploring new features for QTG-Finder2 (A) Several new features we added were enriched for causal genes relative to an average genome gene. New features include the number of SNPs and Indels in conserved non-coding sequences flanking genes and the percent absence of a gene in the collection of natural variant accessions. The Mann-Whitney U Test was used to compare the statistical difference between causal genes and genome genes. Significance levels were defined as: *, $P < 0.05$, **, $P < 0.01$,***, $P < 0.001$. (B) External validation shows that the model performance did not change much after adding the new features. Independent validation sets were used to evaluate model performance. The black dashed lines indicate the theoretical background. Fisher's exact test was performed between models with and without new features. No statistical difference was detected ($P > 0.05$).

subunits during ribosome biogenesis through ATPase activity (LaRonde-LeBlanc and Wlodawer 2005; Knüppel *et al.* 2018). The protein has three domains (Ferreira-Cerca *et al.* 2012). While deletions in each of these domains render the protein non-functional and are lethal, a shorter truncation in the C-terminal domain is not lethal but leads to synthetic lethality with a non-essential ribosome factor called LTV1 (Ferreira-Cerca *et al.* 2012).

The SD1 protein sequence in *S. viridis* has two amino acid replacements (Supplemental Figure S9). The first substitution is a glutamate to aspartate change at position 157 of the *S. viridis* SD1 protein. This amino acid replacement occurs within a relatively conserved region across grass species (Supplemental Figure S9) but is a conservative replacement in the same physicochemical group and this change occurs in other grasses. The second substitution is an alanine to aspartate change at position 366 of the *S. viridis* SD1 protein. This amino acid replacement is a non-conservative replacement but the sequence nearby it is not conserved across grass species. Neither amino acid replacements are within the catalytic Fe^{2}OG dioxygenase domain of SD1.

We also examined gene expression differences between *S. viridis* and *S. italica* for the thirteen candidate genes and the *SD1* gene based on the RNAseq data available at Phytozome12 (Supplemental Table S6). One candidate gene, Sevir.5G394900, has lower expression in most *S. viridis* tissues than its *S. italica* ortholog Seita.5G389700 (Supplemental Figure S10). This gene is annotated as a gene encoding a ribosomal protein belonging to the L1P family (Byrne 2009). The expression difference may be caused by polymorphisms in the promoter region of this gene. We therefore compared the 1kb upstream sequence flanking this gene between *S. viridis* and *S. italica*. We identified five SNPs and one insertion in *S. viridis*, including a SNP located at a predicted MYB transcription factor binding site (Supplemental Figure S11 and Supplemental Table S7).

DISCUSSION

The QTG-Finder we previously developed relies on known causal genes as a training set and cannot be extended to other plant species with few or no known causal genes. Since nearly all plant species, including important crops, do not currently have a sufficient set of

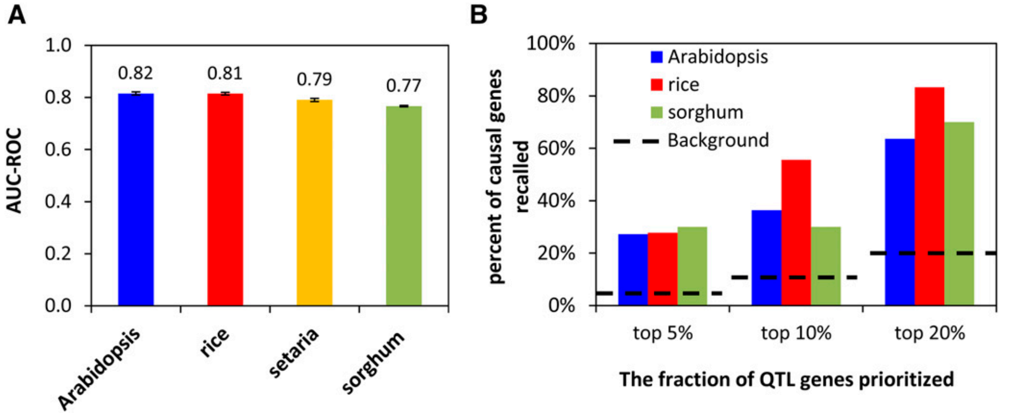

Figure 6 Performance of Setaria and sorghum models (A) The new *Sorghum bicolor* (sorghum) and *Setaria viridis* (Setaria) models trained by QTG-Finder2 algorithm have similar performance as the *Arabidopsis thaliana* (Arabidopsis) and *Oryza sativa* (rice) models. Cross-validation indicated by AUC-ROC (Area Under the Curve - Receiver Operating Characteristic). Error bars indicate standard deviation, N = 50 per species. (B) External validation of the models for Arabidopsis, rice and sorghum that had independent causal gene data available. Fisher's exact test was performed between the sorghum model *vs.* Arabidopsis or rice model, and no statistical difference was detected ($P > 0.05$). The black dashed lines indicate the theoretical background when the same fraction of genes in the QTL was randomly prioritized.

cloned causal genes, this algorithm could not be applied to species beyond Arabidopsis and rice. Here, we have developed QTG-Finder2, which solves this problem by using the orthologs of causal genes to train models in other species.

Some orthologous genes have been repeatedly found to cause variation in similar traits across species. There are more than 100 examples showing that mutations occur at orthologous loci and cause similar phenotypic variation (Martin and Orgogozo 2013). Therefore, we posited that the orthologs of causal genes might determine similar trait variation as the causal genes. Why these genes become genetic hotspots of trait variation is still unknown but there are two theories (Martin and Orgogozo 2013). The first theory is mutational bias. The hotspot genes may be more prone to ectopic changes due to being in unstable chromosomal regions or structures like repeat-rich regions (Chan *et al.* 2010; Martin and Orgogozo 2013). The second theory is optimal pleiotropy (Kopp 2009). The hotspot genes may be able to generate variations in a trait without interfering with other traits. These hypotheses remain to be rigorously tested. In the meantime, given that many known causal genes are genetic hotspots of trait variation, we hypothesized, tested and showed that we can use an orthology approach to transfer the information about causal genes between species.

The major advantage of QTG-Finder2 over QTG-Finder is that it facilitates building models for species that have a limited number of known causal genes, which currently represents almost all plant species, including all major crops except rice. We have shown that Arabidopsis and rice models that were trained on orthologs of causal genes from other species have similar performance as models trained with known causal genes in Arabidopsis and rice. This result indicates that the orthologs derived from known causal genes in other species contain information that can be used to train models. As proof of concept, we applied QTG-Finder2 to train new models for *Setaria viridis* and *Sorghum bicolor*. The sorghum model has a 70% chance to recall a real causal gene (unseen during training) when the top 20% of genes in a QTL are prioritized, which is a comparable performance to the Arabidopsis and rice models.

Sorghum is an important C4 crop with good drought resistance. There are 2605 sorghum QTLs identified by linkage mapping according to Sorghum QTL Atlas (Mace *et al.* 2019). For most of these QTLs, the causal genes have not been identified. The sorghum model can be used to prioritize candidate genes and accelerate the discovery of causal genes in these QTLs. *Setaria viridis* has been developed as a model grass due to advantages like short life span, small plant stature and small diploid genome (Huang *et al.* 2016b). High-throughput phenotyping techniques have been developed for both underground and above-ground traits (Fahlgren *et al.* 2015; Rellan-Alvarez *et al.* 2015; Sebastian *et al.* 2016), which facilitate not only more QTL mapping studies but also faster phenotype screening for characterizing mutants of candidate genes. The Setaria model will play an important role in this pipeline by refining the candidates identified by QTL mapping for the downstream validation and functional analyses.

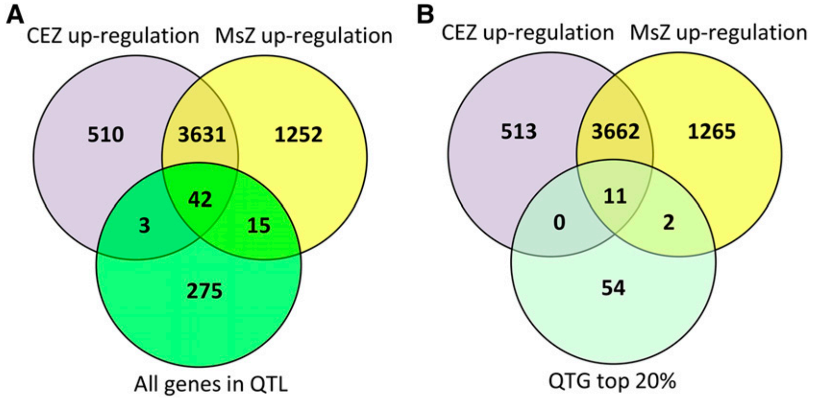

Figure 7 Candidate causal genes of a Setaria plant height QTL prioritized by QTG-Finder2 and transcriptome analysis (A) The overlap among genes up-regulated in the Meristem Zone (MsZ) and the Cell Elongation Zone (CEZ) relative to the maturation zone of the internode and all genes in the height QTL interval. (B) The overlap among genes up-regulated in the Meristem Zone (MsZ) and the Cell Elongation Zone (CEZ) relative to the maturation zone of the internode and the top 20% genes prioritized by QTG-Finder2 (QTG top 20%). Transcriptome data were obtained from Martin, 2016.

We have applied the Setaria model to prioritize genes in a height QTL and combined the results with published transcriptome, gene function and sequence data to generate a hypothesis for candidate genes. We prioritized 13 candidate genes including two strong candidates, a RIO2 kinase/ATPase gene (Sevir.5G413600) and an L1P ribosome protein gene (Sevir.5G394900). RIO2 has two domains, a winged helix (wHTH) and a kinase, which are conserved from archaea to eukaryotes, and a C-terminal extension domain that is conserved only in eukaryotes (LaRonde-LeBlanc and Wlodawer 2005; Ferreira-Cerca *et al.* 2012). While deletions in each of these domains render the protein non-functional and are lethal, a shorter truncation in the C-terminal domain was not lethal but led to synthetic lethality with a non-essential ribosome factor called LTV1 (Ferreira-Cerca *et al.* 2012). It is this region where there are four non-synonymous substitutions between *S. viridis* and *S. italica* (Supplementary Figures S7 and S8). RIO2 is found as a single-copy gene in most plants (Gao *et al.* 2018) but has not been functionally characterized in any plants to date.

The other candidate gene (Sevir.5G394900) encodes an L1P family ribosomal protein. L1P family ribosomal proteins are involved in binding and releasing de-acylated tRNA from the E site of ribosomes (Nikulin *et al.* 2003; Byrne 2009). Arabidopsis mutants of an L1P family ribosomal protein, *PGY1*, are not lethal and have subtle leaf phenotypes (Pinon *et al.* 2008). *PGY1* may function with proteins like ASYMMETRIC LEAVES1 (AS1) and REVOLUTA (REV) to affect plant development in different organs. For example, the *as1 pgy1* double mutant has ectopic leaf lamina outgrowth and the *rev pyg1* double mutant has inflorescence defects (Pinon *et al.* 2008). This candidate gene has higher expression in *S. italica* than *S. viridis* across leaf, shoot and root tissues (Supplemental Figure S10) and therefore may have a broad effect on development in Setaria. The expression difference of this gene is likely caused by a SNP in a putative MYB transcription factor binding site located in the promoter of this gene (Supplemental Table S7).

Though not prioritized as a top 20% gene, the *SD1* gene (Sevir.5G410400) is also a potential causal gene based on its function in other species and up-regulation in internode meristem zone relative to maturation zone. *SD1* gene encodes gibberellin20 oxidase2 in rice and a loss of function allele gives a dwarf phenotype in rice (Spielmeyer *et al.* 2002). However, the rice *SD1* has three orthologs in *S. viridis*: Sevir.3G242400, Sevir.5G410400, Sevir.7G114500, and we do not know if they are functionally redundant.

In summary, we developed the QTG-Finder2 algorithm by incorporating an orthology approach, which can be used to train models in species that have few or even no known causal genes. We have built new models using QTG-Finder2 for *S. bicolor* and *S. viridis* to accelerate the causal gene discovery in these cereal crops and models. The algorithm can also be potentially applied to other important crop species such as maize, barley and wheat to accelerate gene discovery and trait improvement.

FUNDING

This research was supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomic Science Program grant nos. DE- SC0018277 and DE-SC0008769.

ACKNOWLEDGMENTS

We thank Drs. L. Wang and S. Yang for sharing the rice gene presence/absence matrix. We acknowledge H. Nam for technical assistance. We thank all members in Rhee lab for useful discussions, especially Dr. J. Yen.

LITERATURE CITED

Bennetzen, J. L., J. Schmutz, H. Wang, R. Percifield, J. Hawkins *et al.*, 2012 Reference genome sequence of the model plant Setaria. Nat. Biotechnol. 30: 555–561. https://doi.org/10.1038/nbt.2196

Blackman, B. K., J. L. Strasburg, A. R. Raduski, S. D. Michaels, and L. H. Rieseberg, 2010 The role of recently derived FT paralogs in sunflower domestication. Curr. Biol. 20: 629–635. https://doi.org/10.1016/j.cub.2010.01.059

Boyles, R. E., B. K. Pfeiffer, E. A. Cooper, B. L. Rauh, K. J. Zielinski *et al.*, 2017 Genetic dissection of sorghum grain quality traits using diverse and segregating populations. Theor. Appl. Genet. 130: 697–716. https://doi.org/10.1007/s00122-016-2844-6

Byrne, M. E., 2009 A role for the ribosome in development. Trends Plant Sci. 14: 512–519. https://doi.org/10.1016/j.tplants.2009.06.009

Calviño, M., and J. Messing, 2012 Sweet sorghum as a model system for bioenergy crops. Curr. Opin. Biotechnol. 23: 323–329. https://doi.org/10.1016/j.copbio.2011.12.002

Chan, Y. F., M. E. Marks, F. C. Jones, G. Villarreal, Jr., M. D. Shapiro *et al.*, 2010 Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. Science 327: 302–305. https://doi.org/10.1126/science.1182213

Emms, D. M., and S. Kelly, 2019 OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20: 238. https://doi.org/10.1186/s13059-019-1832-y

Fahlgren, N., M. Feldman, M. A. Gehan, M. S. Wilson, C. Shyu *et al.*, 2015 A Versatile Phenotyping System and Analytics Platform Reveals Diverse Temporal Responses to Water Availability in Setaria. Mol. Plant 8: 1520–1535. https://doi.org/10.1016/j.molp.2015.06.005

FAO, 2009 How to feed the world in 2050. Available at: http://www.fao.org/fileadmin/templates/wsfs/docs/expert_paper/How_to_Feed_the_World_in_2050.pdf

Feldman, M. J., R. E. Paul, D. Banan, J. F. Barrett, J. Sebastian *et al.*, 2017 Time dependent genetic analysis links field and controlled environment phenotypes in the model C-4 grass Setaria. PLoS Genet. 13: e1006841. https://doi.org/10.1371/journal.pgen.1006841

Ferreira-Cerca, S., V. Sagar, T. Schafer, M. Diop, A. M. Wesseling *et al.*, 2012 ATPase-dependent role of the atypical kinase Rio2 on the evolving pre-40S ribosomal subunit. Nat. Struct. Mol. Biol. 19: 1316–1323. https://doi.org/10.1038/nsmb.2403

Foley, J. A., N. Ramankutty, K. A. Brauman, E. S. Cassidy, J. S. Gerber *et al.*, 2011 Solutions for a cultivated planet. Nature 478: 337–342. https://doi.org/10.1038/nature10452

Gao, Q., S. Xu, X. Zhu, L. Wang, Z. Yang *et al.*, 2018 Genome-wide identification and characterization of the RIO atypical kinase family in plants. Genes Genomics 40: 669–683. https://doi.org/10.1007/s13258-018-0658-4

Gompel, N., and B. Prud'homme, 2009 The causes of repeated genetic evolution. Dev. Biol. 332: 36–47. https://doi.org/10.1016/j.ydbio.2009.04.040

Haas, B. J., A. L. Delcher, J. R. Wortman, and S. L. Salzberg, 2004 DAGchainer: a tool for mining segmental genome duplications and synteny. Bioinformatics 20: 3643–3646. https://doi.org/10.1093/bioinformatics/bth397

Hilley, J. L., B. D. Weers, S. K. Truong, R. F. McCormick, A. J. Mattison *et al.*, 2017 Sorghum Dw2 Encodes a Protein Kinase Regulator of Stem Internode Length. Sci. Rep. 7: 4616. https://doi.org/10.1038/s41598-017-04609-5

Ho, T., 1998 The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20: 832–844. https://doi.org/10.1109/34.709601

Hu, Z., W. Wang, Z. Wu, C. Sun, M. Li *et al.*, 2018 Novel sequences, structural variations and gene presence variations of Asian cultivated rice. Sci. Data 5: 180079. https://doi.org/10.1038/sdata.2018.79

Huang, C., Q. Chen, G. Xu, D. Xu, J. Tian *et al.*, 2016a Identification and fine mapping of quantitative trait loci for the number of vascular bundle in maize stem. J. Integr. Plant Biol. 58: 81–90. https://doi.org/10.1111/jipb.12358

Huang, P., S. Mamidi, A. Healey, J. Grimwood, J. Jenkins *et al.*, 2019 The Setaria viridis genome and diversity panel enables discovery of a novel domestication gene. bioRxiv (Preprint posted on August 24, 2019) https://doi.org/10.1101/744557

Huang, P., C. Shyu, C. P. Coelho, Y. Y. Cao, and T. P. Brutnell, 2016b Setaria viridis as a Model System to Advance Millet Genetics and Genomics. Front Plant Sci 7: 1781. https://doi.org/10.3389/fpls.2016.01781

Huerta-Cepas, J., D. Szklarczyk, K. Forslund, H. Cook, D. Heller *et al.*, 2016 eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 44: D286–D293. https://doi.org/10.1093/nar/gkv1248

Jin, J., H. Zhang, L. Kong, G. Gao, and J. Luo, 2014 PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. Nucleic Acids Res. 42: D1182–D1187. https://doi.org/10.1093/nar/gkt1016

Jordan, D. R., E. S. Mace, R. G. Henzell, P. E. Klein, and R. R. Klein, 2010 Molecular mapping and candidate gene identification of the Rf2 gene for pollen fertility restoration in sorghum. [*Sorghum bicolor* (L.) Moench] Theor. Appl. Genet. 120: 1279–1287. https://doi.org/10.1007/s00122-009-1255-3

Kawahigashi, H., S. Kasuga, T. Ando, H. Kanamori, J. Wu *et al.*, 2011 Positional cloning of ds1, the target leaf spot resistance gene against Bipolaris sorghicola in sorghum. Theor. Appl. Genet. 123: 131–142. https://doi.org/10.1007/s00122-011-1572-1

Knüppel, R., R. H. Christensen, F. C. Gray, D. Esser, D. Strauss *et al.*, 2018 Insights into the evolutionary conserved regulation of Rio ATPase activity. Nucleic Acids Res. 46: 1441–1456. https://doi.org/10.1093/nar/gkx1236

Kojima, S., Y. Takahashi, Y. Kobayashi, L. Monna, T. Sasaki *et al.*, 2002 Hd3a, a rice ortholog of the Arabidopsis FT gene, promotes transition to flowering downstream of Hd1 under short-day conditions. Plant Cell Physiol. 43: 1096–1105. https://doi.org/10.1093/pcp/pcf156

Kopp, A., 2009 Metamodels and phylogenetic replication: a systematic approach to the evolution of developmental pathways. Evolution 63: 2771–2789. https://doi.org/10.1111/j.1558-5646.2009.00761.x

LaRonde-LeBlanc, N., and A. Wlodawer, 2005 A family portrait of the RIO kinases. J. Biol. Chem. 280: 37297–37300. https://doi.org/10.1074/jbc.R500013200

Lin, F., J. Fan, and S. Y. Rhee, 2019 QTG-Finder: A Machine-Learning Based Algorithm To Prioritize Causal Genes of Quantitative Trait Loci in Arabidopsis and Rice. G3 (Bethesda) 9: 3129–3138 (Bethesda). https://doi.org/10.1534/g3.119.400319

Lin, Z., X. Li, L. M. Shannon, C. T. Yeh, M. L. Wang *et al.*, 2012 Parallel domestication of the Shattering1 genes in cereals. Nat. Genet. 44: 720–724. https://doi.org/10.1038/ng.2281

Luo, H., W. Zhao, Y. Wang, Y. Xia, X. Wu *et al.*, 2016 SorGSD: a sorghum genome SNP database. Biotechnol. Biofuels 9: 6. https://doi.org/10.1186/s13068-015-0415-8

Mace, E., D. Innes, C. Hunt, X. Wang, Y. Tao *et al.*, 2019 The Sorghum QTL Atlas: a powerful tool for trait dissection, comparative genomics and crop improvement. Theor. Appl. Genet. 132: 751–766. https://doi.org/10.1007/s00122-018-3212-5

Magalhaes, J. V., J. Liu, C. T. Guimaraes, U. G. Lana, V. M. Alves *et al.*, 2007 A gene in the multidrug and toxic compound extrusion (MATE) family confers aluminum tolerance in sorghum. Nat. Genet. 39: 1156–1161. https://doi.org/10.1038/ng2074

Martin, A., and V. Orgogozo, 2013 The Loci of Repeated Evolution: A Catalog of Genetic Hotspots of Phenotypic Variation. Evolution 67: 1235–1250.

Martin, A. P., W. M. Palmer, C. Brown, C. Abel, J. E. Lunn *et al.*, 2016 A developing Setaria viridis internode: an experimental system for the study of biomass generation in a C-4 model species. Biotechnol. Biofuels 9: 45. https://doi.org/10.1186/s13068-016-0457-6

Mauro-Herrera, M., and A. N. Doust, 2016 Development and Genetic Control of Plant Architecture and Biomass in the Panicoid Grass, Setaria. PLoS One 11: e0151346. https://doi.org/10.1371/journal.pone.0151346

Murphy, R. L., R. R. Klein, D. T. Morishige, J. A. Brady, W. L. Rooney *et al.*, 2011 Coincident light and clock regulation of pseudoresponse regulator protein 37 (PRR37) controls photoperiodic flowering in sorghum. Proc. Natl. Acad. Sci. USA 108: 16469–16474. https://doi.org/10.1073/pnas.1106212108

Murphy, R. L., D. T. Morishige, J. A. Brady, W. L. Rooney, S. Yang *et al.*, 2014 Ghd7 (Ma6) Represses Sorghum Flowering in Long Days: Ghd7 Alleles Enhance Biomass Accumulation and Grain Production. Plant Genome 7: 1–10. https://doi.org/10.3835/plantgenome2013.11.0040

Nikulin, A., I. Eliseikina, S. Tishchenko, N. Nevskaya, N. Davydova *et al.*, 2003 Structure of the L1 protuberance in the ribosome. Nat. Struct. Biol. 10: 104–108. https://doi.org/10.1038/nsb886

Pinon, V., J. P. Etchells, P. Rossignol, S. A. Collier, J. M. Arroyo *et al.*, 2008 Three PIGGYBACK genes that specifically influence leaf patterning encode ribosomal proteins. Development 135: 1315–1324. https://doi.org/10.1242/dev.016469

Ramstein, G. P., S. E. Jensen, and E. S. Buckler, 2019 Breaking the curse of dimensionality to identify causal variants in Breeding 4. Theor. Appl. Genet. 132: 559–567. https://doi.org/10.1007/s00122-018-3267-3

Rellan-Alvarez, R., G. Lobet, H. Lindner, P. L. Pradier, J. Sebastian *et al.*, 2015 GLO-Roots: an imaging platform enabling multidimensional characterization of soil-grown root systems. eLife 4: e07597. https://doi.org/10.7554/eLife.07597

Rodríguez-Leal, D., Z.H. Lemmon, J. Man, M.E. Bartlett, and Z.B. Lippman, 2017 Engineering Quantitative Trait Variation for Crop Improvement by Genome Editing. Cell 171: 470–480.e8. https://doi.org/10.1016/j.cell.2017.08.030

Saballos, A., S. E. Sattler, E. Sanchez, T. P. Foster, Z. Xin *et al.*, 2012 Brown midrib2 (Bmr2) encodes the major 4-coumarate:coenzyme A ligase involved in lignin biosynthesis in sorghum (*Sorghum bicolor* (L.) Moench). Plant J. 70: 818–830. https://doi.org/10.1111/j.1365-313X.2012.04933.x

Schläpfer, P., P. Zhang, C. Wang, T. Kim, M. Banf *et al.*, 2017 Genome-Wide Prediction of Metabolic Enzymes, Pathways, and Gene Clusters in Plants. Plant Physiol. 173: 2041–2059. https://doi.org/10.1104/pp.16.01942

Schwartz, C., S. Balasubramanian, N. Warthmann, T.P. Michael, J. Lempe *et al.*, 2009 Cis-regulatory changes at FLOWERING LOCUS T mediate natural variation in flowering responses of Arabidopsis thaliana. Genetics 183: 723–732. https://doi.org/10.1534/genetics.109.104984

Sebastian, J., M. C. Yee, W. Goudinho Viana, R. Rellan-Alvarez, M. Feldman *et al.*, 2016 Grasses suppress shoot-borne roots to conserve water during drought. Proc. Natl. Acad. Sci. USA 113: 8861–8866. https://doi.org/10.1073/pnas.1604021113

Skøt, L., R. Sanderson, A. Thomas, K. Skot, D. Thorogood *et al.*, 2011 Allelic variation in the perennial ryegrass FLOWERING LOCUS T gene is associated with changes in flowering time across a range of populations. Plant Physiol. 155: 1013–1022. https://doi.org/10.1104/pp.110.169870

Spielmeyer, W., M. H. Ellis, and P. M. Chandler, 2002 Semidwarf (sd-1), "green revolution" rice, contains a defective gibberellin 20-oxidase gene. Proc. Natl. Acad. Sci. USA 99: 9043–9048. https://doi.org/10.1073/pnas.132266399

Staal, J., M. Kaliff, E. Dewaele, M. Persson, and C. Dixelius, 2008 RLM3, a TIR domain encoding gene involved in broad-range immunity of Arabidopsis to necrotrophic fungal pathogens. Plant J. 55: 188–200. https://doi.org/10.1111/j.1365-313X.2008.03503.x

Tan, S., Y. Zhong, H. Hou, S. Yang, and D. Tian, 2012 Variation of presence/absence genes among Arabidopsis populations. BMC Evol. Biol. 12: 86. https://doi.org/10.1186/1471-2148-12-86

Tian, F., D. C. Yang, Y. Q. Meng, J. Jin, and G. Gao, 2019 PlantRegMap: charting functional regulatory maps in plants. Nucleic Acids Res. 48: D1104–D1113.

Van Bel, M., T. Diels, E. Vancaester, L. Kreft, A. Botzki *et al.*, 2018 PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. Nucleic Acids Res. 46: D1190–D1196. https://doi.org/10.1093/nar/gkx1002

Weigel, D., and M. Nordborg, 2005 Natural variation in Arabidopsis. How do we find the causal genes? Plant Physiol. 138: 567–568. https://doi.org/10.1104/pp.104.900157

Werner, J. D., J. O. Borevitz, N. H. Uhlenhaut, J. R. Ecker, J. Chory *et al.*, 2005 FRIGID-Independent Variation in Flowering Time of Natural *Arabidopsis thaliana* Accessions. Genetics 170: 1197–1207. https://doi.org/10.1534/genetics.104.036533

Xu, K., X. Xu, T. Fukao, P. Canlas, R. Maghirang-Rodriguez *et al.*, 2006 *Sub1A* is an ethylene-response-factor-like gene that confers submergence tolerance to rice. Nature 442: 705–708. https://doi.org/10.1038/nature04920

Yan, L., D. Fu, C. Li, A. Blechl, G. Tranquilli *et al.*, 2006 The wheat and barley vernalization gene *VRN3* is an orthologue of *FT*. Proc. Natl. Acad. Sci. USA 103: 19581–19586. https://doi.org/10.1073/pnas.0607142103

Yan, L., A. Loukoianov, A. Blechl, G. Tranquilli, W. Ramakrishna *et al.*, 2004 The wheat *VRN2* gene is a flowering repressor down-regulated by vernalization. Science 303: 1640–1644. https://doi.org/10.1126/science.1094305

Yang, S., R. L. Murphy, D. T. Morishige, P. E. Klein, W. L. Rooney *et al.*, 2014 Sorghum phytochrome B inhibits flowering in long days by activating expression of SbPRR37 and SbGHD7, repressors of SbEHD1, SbCN8 and SbCN12. PLoS One 9: e105352. https://doi.org/10.1371/journal.pone.0105352

Yonemaru, J., T. Yamamoto, S. Fukuoka, Y. Uga, K. Hori *et al.*, 2010 Q-TARO: QTL Annotation Rice Online Database. Rice (N. Y.) 3: 194–203. https://doi.org/10.1007/s12284-010-9041-z

Communicating editor: A. Doust