QTG-Finder: a machine-learning based algorithm to prioritize causal genes of quantitative trait loci

Fan Lin, Jue Fan, Seung Y. Rhee*

Department of Plant Biology, Carnegie Institution for Science, Stanford, California 94305, USA

Contact: srhee@carnegiescience.edu
Phone: (650) 739-4251

Author Contributions

Conceptualization, S.Y.R.; Methodology, J.F., F.L., and S.Y.R.; Investigation, F.L. and J.F.; Formal Analysis, F.L. and J.F.; Writing–Original Draft, F.L.; Writing–Review & Editing, F.L., S.Y.R., and J.F.; Funding Acquisition, S.Y.R.; Resources, S.Y.R.; Supervision, S.Y.R.

Abstract

Linkage mapping is one of the most commonly used methods to identify genetic loci that determine a trait. However, the loci identified by linkage mapping may contain hundreds of candidate genes and require a time-consuming and labor-intensive fine mapping process to find the causal gene controlling the trait. With the availability of a rich assortment of genomic and functional genomic data, it is possible to develop a computational method to facilitate faster identification of causal genes. We developed QTG-Finder, a machine learning based algorithm to prioritize causal genes by ranking genes within a quantitative trait locus (QTL). Two predictive models were trained separately based on known causal genes in Arabidopsis and rice. With an independent validation analysis, we demonstrate the models can correctly prioritize about 65% and 60% of Arabidopsis and rice causal genes when the top 20% ranked genes were considered. The models can prioritize different types of traits though at different efficiency. We also identified several important features of causal genes including paralog copy number, being a transporter, being a transcription factor, and containing SNPs that cause premature stop codon. This work lays the foundation for systematically understanding characteristics of causal genes and establishes a pipeline to predict causal genes based on public data.

One sentence summary: We systematically analyzed the genomic characteristics of causal genes in QTLs and developed a novel computational tool to prioritize causal genes.

Keywords: Arabidopsis, causal gene, machine-learning algorithm, candidates, quantitative trait loci, rice

Introduction

As the world's population expands, food security faces a major challenge in the near future. By 2050, world population is projected to grow by 34%, which will require a 70% increase of global food production to meet the demand (FAO 2009). To catch up with the growing global food demand, it is important to improve the efficiency of arable land usage by developing better crops.

Many agriculturally and medically important traits are quantitative and controlled by multiple genetic loci. Examples include plant height, grain yield, and flowering time in plants and common disorders such as cancer, diabetes, and hypertension in humans. The variation in quantitative traits allows organisms to adapt to various environments (Baxter *et al.* 2010; Leinonen *et al.* 2013). Quantitative traits are determined by a combination of genetic complexity and environmental factors (Mackay 2001). The genetic complexity of quantitative traits comes from the involvement of multiple quantitative trait loci (QTL) and the non-additive interactions among them (Carlborg and Haley 2004; Mackay 2014). To better understand the evolutionary forces and molecular mechanisms that shape the genetic architectures of adaptive traits, we need to identify all the causal genes that contribute to most of the phenotypic variation of the traits and elucidate the molecular mechanisms of their actions. Achieving this goal will facilitate rational engineering of plant traits and more accurate prediction of the effects of the modifications on the engineered plant.

QTL linkage mapping and genome wide association study (GWAS) are two common approaches used to identify QTLs, each with its own strengths and limitations. Both mapping approaches are based on the co-segregation of a trait and genetic variants in a population. The population for linkage mapping is usually the progeny of parental plants that differ in a trait, such as an F2 population or recombinant inbred lines (Bergelson and Roux 2010). GWAS mapping uses a natural population that has a heritable variation of a trait. Compared to GWAS, linkage mapping does not suffer from issues like rare alleles and population structure (Bergelson, 2010). For example, the most significant seed dormancy QTL DOG1 identified by linkage mapping was not identified by GWAS, likely due to the rarity of the strong allele in the GWAS population (Bentsink *et al.* 2010; He 2014). Confounding population structure can cause a high false positive rate in GWAS, though some methods have been developed to ameliorate it (Price *et al.* 2010). However, efforts to correct it could result in a higher false negative rate (Brachi *et al.* 2010). Linkage mapping does not suffer from these issues, but it has a relatively lower mapping resolution and cannot identify QTLs of minor effects when the sample size is small (Martin and Orgogozo 2013; Otto and Jones 2000; Wellenreuther and Hansson 2016; Xu 2003).

For QTLs identified by linkage mapping, finding causal genes underlying them is still a big bottleneck (Bergelson and Roux 2010). In a typical rice linkage mapping, the size of a QTL can range from 200kb- 3Mb, which can harbor tens to hundreds of genes depending on the mapping population and gene density (Bargsten *et al.* 2014; Daware *et al.* 2017). Even in the post-genomic era where all the genes in the genome are uncovered, identifying QTL causal genes is not straightforward since many QTLs either contain no obvious candidate genes or too many genes relevant for the trait (Nuzhdin *et al.* 1999).

Therefore, despite the many QTLs that have been reported in plants, only a few have been studied at the molecular level.

Conventional fine mapping is a reliable but time-consuming and labor-intensive approach to narrow down the range of candidate genes in a QTL region. The basis of fine mapping is to create a population that has more recombination events within a QTL in order to identify a smaller genomic segment that co-segregates with the trait. However, the enormous time and labor required for creating and screening a population of progenies limits the usage of this method (Tuinstra *et al.* 1997). Depending on the frequency of recombination, thousands of progenies may need to be genotyped to get to a gene-scale resolution (Dinka *et al.* 2007). For example, 1,160 progenies were screened to identify the *Pi36* gene in rice and as many as 18,994 progenies were screened to identify the causal gene of Bph15 in rice (Liu *et al.* 2005; Yang *et al.* 2004). The high cost associated with genotyping and phenotyping makes it challenging to apply fine mapping to all QTLs.

Alternative approaches to refine the candidate list of causal genes include meta-analysis, joint linkage-association analysis, and other computational methods including machine-learning algorithms. The first two approaches require either the availability of many QTL studies on similar traits or an additional association mapping experiment (Buckler *et al.* 2009; Motte *et al.* 2014; Yin *et al.* 2017). Computational methods including machine-learning algorithms have been developed to prioritize disease associated genes and genetic variants in human (Hormozdiari *et al.* 2015; Kircher *et al.* 2014; Perez-Iratxeta *et al.* 2002; Ritchie *et al.* 2014). To distinguish disease-associated from non-associated variants, a variety of information has been used, including the effect of polymorphism (Gelfman *et al.* 2017; Kircher *et al.* 2014; Ng and Henikoff 2003), sequence conservation (Huang *et al.* 2017; Pollard *et al.* 2010), regulatory information (Deo *et al.* 2014), expression profile (Deo *et al.* 2014; Mordelet and Vert 2011), Gene Ontology (GO) (Mordelet and Vert 2011), KEGG pathway (Mordelet and Vert 2011), and publications (Perez-Iratxeta *et al.* 2002). In contrast, only two causal gene prioritization approaches are available for plants. One method was developed for GWAS in maize based on co-expression networks (Schaefer *et al.* 2018). Another method was developed for linkage mapping based on biological process GOs (Bargsten *et al.* 2014). To date, no machine-learning approaches using multiple data types have been developed to address this problem.

Here, we built a supervised learning algorithm to prioritize QTL causal genes using known causal genes in *Arabidopsis thaliana* (Arabidopsis) and *Oryza sativa* (rice) and a suite of publicly available genetic and genomic data. For each species, we trained a predictive model using features based on polymorphism data, function annotation, co-function network, and paralog copy number. By testing the models on an independent set of known causal genes, we demonstrated its efficiency in prioritizing causal genes.

Materials and methods

Data sources and features used in QTG-Finder

Twenty-eight features were extracted from published genome-scale data, which included polymorphism features, functional annotation features and other genomic and functional genomic features.

Arabidopsis polymorphism data of 1,135 accessions was downloaded from 1001 Genomes Project (https://1001genomes.org) (Consortium 2016) and rice polymorphism data of 3,010 cultivars was downloaded from Rice SNP-Seek Database (http://snp-seek.irri.org) (Mansueto *et al.* 2017). We used SIFT4G (v 2.4) (Ng and Henikoff 2003) and SnpEff (v 4.3r) (Cingolani *et al.* 2012) to annotate the raw polymorphism data. The number of non-synonymous SNP as annotated by SIFT4G was normalized to protein length and used as a numeric feature (normalized_nonsyn_SNP). Non-synonymous SNPs at conserved protein sequences were predicted to cause deleterious amino acid changes by SIFT4G. The presence of deleterious non-synonymous SNPs in a gene was used as a binary feature (is_nonsyn_deleterious). If a gene contained any deleterious non-synonymous SNPs, the "is_nonsyn_deleterious" feature was set to 1, otherwise it was set to 0. Other binary polymorphism features such as "is_start_lost" (start codon lost) and "is_start_gained" (start codon gained) were extracted from SnpEff annotations in the same way. For "is_SNP_cis", the Position Weight Matrices of cis-elements were downloaded from CIS-BP database (Build 1.02) (Weirauch *et al.* 2014) and mapped to 1kb upstream of all genes in the genome using FIMO (v 4.12.0) (Grant *et al.* 2011). The cis-elements with a matching score above 55 were imported into SnpEff library to annotate the SNPs. This matching score cutoff was determined by a cross-validation as described later.

Functional annotation features were binary features based on GO (Gotz *et al.* 2008; Jones *et al.* 2014) and Plant Metabolic Network (PMN) (Schlapfer *et al.* 2017). Arabidopsis and rice genes were annotated by Blast2GO (BLAST+ 2.2.29) and InterProScan (v 5.3-46.0). The molecular function GOs were then converted to high-level functional groups such as transcription factor, receptor, kinase, transporter, and enzyme to mitigate the effect of some inaccurate annotations (Jones *et al.* 2007). Genes annotated as enzymes were further classified into 13 PMN metabolic domains such as carbohydrate metabolism and nucleotide metabolism (Schlapfer *et al.* 2017). Unclassified genes in PMN were classified as "is_other_metabolism". Genes annotated as enzymes by GO but not present in PMN databases are enzymes involved in macromolecule metabolic process or enzymes that don't have a specific function assigned. Since a majority of them is involved in macromolecule metabolic process, we named this group as "is_macromolecule_metabolism".

Co-functional networks of Arabidopsis and rice were retrieved from AraNet and RiceNet (Lee *et al.* 2010; Lee *et al.* 2011). The sum of all the edge weights of a gene was used as the "network_weight" feature.

Paralog copy number (paralog_copy_number) and essential gene prediction (is_essential_gene) were taken from a previous publication (Lloyd *et al.* 2015).

Arabidopsis and rice causal genes used for training and independent validation

For model training and cross-validation, curated causal genes from Martin and Orgogozo were used as

positives for algorithm training (Martin and Orgogozo 2013). In total, 60 Arabidopsis and 45 rice causal genes were used as the initial training set. For literature validation, we performed a further literature curation and found eleven Arabidopsis and ten rice causal genes, which were not included in the Martin and Orgogozo list (Supplementary Methods).

Algorithm training and parameter optimization

The QTG-Finder algorithm was developed in Python (v 3.6) with the 'sklearn' package (v 0.19.0) (Pedregosa *et al.* 2011). We developed an extended 5-fold cross-validation framework (Fig. 1a) to evaluate training performance and optimize model parameters.

For the 5-fold cross validation, curated causal genes were used as positives and the other genes from the genome were used as negatives. The positives were randomly split into training and testing positives in a 4:1 ratio. Training and testing positives were combined with different sets of negative genes that were randomly selected from the rest of the genome. To increase the combination of positives and negatives, we re-split the positives 50 times randomly and selected negatives 50 times. This number of iterations ensured greater than 99% probability that every positive sample co-occurred with every negative at least once in the training or testing set during the cross-validation process. The probability of co-occurrence was calculated as Equation 1. P_{co} is the probability of co-occurrence of a positive and a negative in a testing or training set. N is the total number of negative samples. n is the number of negative samples selected as testing or training samples. R is the number of iterations used to re-split the positive set. C is the number of cross-validation folds that contains a positive sample. C was set to 4 for the training set and set to 1 for the testing test. S is the number of iterations to randomly select the negative set.

$$P_{co} = 1 - \left[\prod_{i=0}^{n}\left(1 - \frac{1}{N-i}\right)\right]^{R*C*S} \quad (1)$$

We tested different classifiers and parameters and optimized the model based on Area Under the Curve of the Receiver Operating Characteristic (AUC-ROC). The average AUC-ROC from all iterations was used to evaluate training performance. The three classifiers we tested were Random Forest, naïve Bayes, and Support Vector Machines (Cortes and Vapnik 1995; Tin Kam 1998; Zhang 2004)(Supplementary Fig. S1). For Random Forest, we tuned the number of trees and the maximum number of features for each tree based on AUC-ROC (Supplementary Fig. S2). We used 100 trees and a max_feature of 9 for Random Forest. For Support Vector Machines, RBF kernel was used and the C parameter was tuned. Random Forest was chosen for further analysis since its performance was slightly better than the other two classifiers. The ratio of positives and negatives in training data was also tuned to maximize cross-validation AUC-ROC (Supplementary Fig. S3). The best performing positives:negatives ratio was 1:20 for Arabidopsis and 1:5 for rice. For testing, a positives:negatives ratio of 1:200 was used since it is close to the average ratio of causal and non-causal genes in real QTLs.

The source code for cross-validation and any other analyses below are available at https://github.com/carnegie/QTG_Finder

Feature importance analysis

We implemented a leave-one-out analysis to evaluate feature importance. This method was based on the change of AUC-ROC (ΔAUC-ROC) when leaving out one feature from the models. The same cross-validation framework was used for this analysis. For each iteration, we calculated AUC-ROC on the original and the leave-one-out models developed with the same training and testing datasets. The ΔAUC-ROC was calculated by subtracting the leave-one-out AUC-ROC from the original AUC-ROC. With the results from all iterations, we calculated the average ΔAUC-ROC for each feature.

Independent literature validation

For validation, we applied the models to an independent set of causal genes that were curated from recent literature and not used for cross-validation. The models were trained by known causal genes from the initial list and negatives were randomly selected from the rest of the genome. Model training was repeated 5,000 times by resampling training negatives from the genome. With 5,000 iterations, there was >99% probability that each gene in the genome was selected at least once based on simulation. We applied the models to each of the independent causal gene and all other genes located within the QTL. All genes within the QTL were ranked based on the frequency of being predicted as a causal gene.

We calculated the probability of correctly prioritizing at least K causal genes when applying the models to a total of N QTLs with Equation 2. p is the probability to correctly prioritize a causal gene of a single QTL at a certain threshold. x is the number of causal genes being correctly prioritized.

$$P(x \geq K) = \sum_{x=K}^{N} \binom{N}{x} p^{x} (1-p)^{N-x} \quad (2)$$

Trait category analysis

The trait category analysis was performed in a similar way as the independent literature validation except using different training and testing sets. Each curated causal gene was tested once. For each round, one curated causal gene was removed from the training set. Then the model was trained and applied to rank the known causal gene and 200 flanking genes.

Results

QTG-Finder: a machine-learning algorithm to prioritize causal genes

We developed the QTG-Finder algorithm to find causal genes from QTL data and generated two predictive models in Arabidopsis and rice with the algorithm. These two species were selected for model training since they have the largest number of QTL causal genes (QTGs) that have been discovered by fine mapping and map-based cloning in plants (Martin and Orgogozo 2013). For model training, we selected 60 Arabidopsis and 45 rice causal genes as a positive set (Martin and Orgogozo, 2013, Supplementary Tables S1 and S2). The negative set was a subset of genes randomly selected from the rest of the genome. To train the models, we used 28 Arabidopsis features and 27 rice features, including polymorphisms, functional categories of genes, function interference from co-function networks, gene essentiality, and paralog copy number (Supplementary Tables S3, S4 and S5). These features were generally independent from each other

(most have a Pearson's correlation coefficient <0.2) (Supplementary Fig. S4).

We optimized the models with an extended cross-validation framework (Fig. 1a). In addition to a typical 5-fold cross-validation (Kuhn and Johnson 2013), iterations were applied to randomly select genes from the negative set and re-split the positive set in order to maximize the combinations of positives and negatives in the training and testing sets (See method).

With this framework, we evaluated the training performance with Area Under the Curve of Receiver Operating Characteristic (AUC-ROC) and optimized parameters. To find the optimal parameters, we compared the AUC-ROC of different machine-learning classifiers, modeling parameters, and the ratio of positive:negative genes in the training set (Supplementary Fig. S2, S3, and S4). Random Forest was selected as the classifier since it was less prone to over-fitting and performed better than the other classifiers tested (Supplementary Fig. S1). After optimization, AUC-ROC for the Arabidopsis and rice models were 0.86 and 0.73, respectively (Fig. 1b).

Since the positive training set used was relatively small, we also evaluated the relationship between training performance and size of the training set. The AUC-ROC increased as a larger training set was used. Interestingly, maximum gain in the AUC-ROC was achieved with 20 causal genes (Supplementary Fig. S5).

Important features for predicting causal genes

With the optimized models, we wanted to know which features were important for causal gene prediction. Since Random Forest uses features and their interactions for classification (Touw *et al*. 2013), the importance of a feature cannot be measured by simple enrichment or depletion of a single feature in causal genes. Therefore, we evaluated feature importance based on the change of ROC-AUC (**Δ**ROC-AUC) when excluding a feature from the model (Lloyd *et al*. 2015). When an important feature is excluded from the model, the ROC-AUC should decrease.

Here, we highlighted the six most important features out of a total of 28 features. The six most important features for Arabidopsis were paralog copy number, transporter, the number of non-synonymous SNPs normalized to protein length (normalized_nonsyn_SNP), receptor, transcription factor, and SNPs causing premature stop codon (is_stop_gained) (Fig. 2a). The six most important features for rice were paralog copy number, macromolecule metabolism, network weight sum, transcription factor, transporter, and SNPs causing premature stop codon. Four out of the six most important features were consistent between Arabidopsis and rice models, which were paralog copy number, transporter, transcription factor, and SNPs causing premature stop codon.

For the six most important features in Arabidopsis and rice, we examined their ratio in known causal genes versus randomly selected genes in the genome (Fig. 2b). Compared to other genes in the genome, the causal genes tended to have more paralogs, higher frequency of being a transporter or a transcription factor, and higher frequency of containing SNPs that cause premature stop codons in both species.

The rest of the features contributed less to, but did not impair, model performance to a large degree (**Δ**ROC-AUC< 0.02). Since there was no strong evidence that they impair prediction, we did not remove

250 them from the models for further analysis.

251 **Validating QTG-Finder by ranking an independent set of QTL genes**

252 To assess the predictability of QTG-Finder models, we searched the literature for a separate set of
253 known causal genes from the initial training set. We found eleven Arabidopsis and ten rice genes that are
254 likely causal genes underlying QTLs when interpreting linkage mapping with additional evidence such as
255 functional complementation, fine mapping, joint linkage-association analysis or genetic analyses
256 (Supplementary Table S6). These causal genes were not used for model training or cross-validation.

257 To examine model performance, we applied the QTG-Finder models to this new set of causal genes.
258 For each known causal gene, we ranked all genes in the QTL region based on the frequency of being
259 predicted as a causal gene from 5,000 iterations. Since the number of genes in a QTL region varies, we
260 used a gene's rank percentile for evaluation. The rank percentile of a gene indicates the percentage of QTL
261 genes that had higher ranks than the gene of interest.

262 Based on the rank of these known causal genes, we evaluated model performance at different cutoffs.
263 We calculated the percentage of known causal genes included in the top 5%, 10%, and 20% of the
264 prioritized genes within a QTL (Fig. 3a). The top 20% of the ranked genes included seven Arabidopsis
265 (~64%) and six rice (~60%) causal genes. With a more stringent cutoff of 5%, four Arabidopsis (~27%)
266 and three rice (~30%) causal genes were prioritized.

267 Most linkage mapping studies identify multiple QTLs. We therefore calculated a theoretical model
268 performance on identifying causal genes from multiple QTLs simultaneously, which we defined as the
269 probability of identifying at least X% of all causal genes when applying the model to all QTLs of a trait
270 (Fig. 3b and c). For example, assuming there were five QTLs of a trait identified by a linkage mapping
271 study and each QTL contained one causal gene. For the Arabidopsis model, the probability of identifying at
272 least one causal gene would be 99% when the top 20% genes of all QTLs were tested experimentally. The
273 probability of identifying all five causal genes would be 10% when the top 20% cutoff was used. We
274 further compared the performance of all three cutoffs, top 20%, top 10%, and top 5%. The probability of
275 identifying at least one out of five causal genes would be no less than 80% for all three cutoffs. The
276 probability to correctly prioritize at least four out of five causal genes would be 40% (for top 20%), 14%
277 (for top 10%), and 2% (for top 5%). Therefore, a less stringent cutoff (top 20%) performs much better than
278 a more stringent cutoff if one is interested in finding most of the causal genes or causal genes of a particular
279 QTL. However, if the goal is to identify any causal gene, then screening the top 5% of all QTLs may be a
280 more strategic approach since fewer candidate genes need to be tested experimentally.

281 **Trait type preference of QTG-Finder models**

282 Since the training set included genes for different types of traits at an imbalanced ratio, we wanted to
283 know how QTG-Finder models would work for each type of traits (Fig. 4a). The independent validation
284 described above was based on causal genes related to plant development and disease resistance
285 (Supplementary Table S6). However, this validation set was not large enough for a systematic analysis and
286 did not have any abiotic-stress-related causal genes. Therefore, we performed a rank analysis for different

trait categories using the known causal genes from the initial training set (60 for Arabidopsis and 45 for rice). For this rank analysis, each causal gene was taken out from the training set once and used for a rank test. The single causal gene and its 200 neighboring genes in the genome were used as a testing set. We applied the models to each testing set to obtain the rank for each causal gene. Then we calculated the average rank for the causal genes in the four trait categories: development, abiotic stress, biotic stress and "other". The "other" category included traits in seed hull color, oil composition, necrosis, etc.

Performance of the models was not the same for different trait categories. Both abiotic and biotic stress traits had better performance than developmental traits (Fig. 4b). In addition, the Arabidopsis model performed slightly better than the rice model for all trait categories. This trait category analysis can guide users to determine rank cutoffs when applying models to different types of traits.

Discussion

Linkage mapping is a useful tool to identify the genomic regions responsible for many agriculturally and medically important traits. However, it is not straightforward to identify the genes that cause the trait variation from these genome regions. The discovery of causal genes still requires time-consuming and labor-intensive fine mapping. In this study, we developed a machine-learning algorithm to reduce the number of candidates to be tested experimentally in order to accelerate the discovery of causal genes.

A machine-learning algorithm to prioritize QTL causal genes

Several causal variant or gene prioritization methods have been developed for human data but not many in plants (Bargsten *et al.* 2014; Jagadeesh *et al.* 2016; Kircher *et al.* 2014; Schaefer *et al.* 2018). Most prioritization methods have been developed for GWAS mapping in human, an organism where linkage mapping cannot be performed. However, linkage mapping can capture rare alleles and has been broadly used to study quantitative traits of livestock, crops, and model organisms. A causal gene prioritization is especially helpful for large QTLs identified by linkage mapping, which can constitute tens to hundreds of genes. One method has been developed in rice to prioritize causal genes for linkage mapping (Bargsten *et al.* 2014). This method is based on the hypothesis that causal genes from multiple QTLs of the same trait are more likely to have the same biological process GO terms, and therefore genes with overrepresented biological process GOs were prioritized as causal genes. However, this method gives no predictions for ~15% of traits and lack an unbiased performance evaluation since the same set of causal genes was used to determine cutoff and evaluate performance.

In this study, we built a supervised learning algorithm using multiple features and validated its efficacy with an independent dataset from the literature. The models could accelerate the discovery of causal genes by ranking all the genes in a QTL region and prioritizing the top 5%, 10%, or 20% genes, which are most likely to contain the causal gene, for experimental testing. Based on an assessment using independent data in the literature, we calculated the performance when applying the models to all QTLs of a trait and compared three cutoffs (top 5%, 10%, and 20%). The less stringent cutoff (top 20%) had a higher chance to find more causal genes (Fig. 3b and c) but yielded more candidates that needed to be tested by experiments.

The more stringent cutoff (top 5%) had a lower chance to find all causal genes but yielded a smaller set of candidates to test. The probability for the models to find at least one causal gene is high for all three cutoffs. If the goal were to find one or more causal genes for functional studies and the particular QTL regions did not matter, the 5% cutoff would be more efficient. If the goal were to discover all causal genes and understand the genetic architecture of a trait, the 20% cutoff would be better. Similarly, if a particular QTL were of interest for discovering the underlying causal gene, the 20% cutoff would be better.

There are several conceptual and practical advantages of QTG-Finder algorithm. First, this algorithm combines multiple types of publically available data including polymorphisms, function annotations, co-function network and other genomic data, which have not been applied to prioritize causal genes from linkage mapping studies. Second, models were trained on causal genes from various traits and can be applied to several types of traditional traits, though the prioritization efficiency was not equivalent. Third, validation from the literature provides guidance on what proportion of genes to prioritize in practice rather than arbitrarily selecting a threshold. Fourth, the models treat each QTL independently and have the flexibility to prioritize a specific QTL of interest.

Two limitations of this study are the small number of known causal genes in plants and the impurity of negative set used for model training. We used 60 Arabidopsis and 45 rice causal genes that have been verified by map-based-cloning as a positive dataset. Even though they are of high quality, this positive dataset may not be large enough to represent all the features of causal genes. There could still be other important features of causal genes that we were not able to capture with this small dataset. The negative set was composed of genes randomly selected from the rest of the genome. Though we excluded known causal genes, there could still be some uncharacterized causal genes. As a result of these limitations, 20% cutoff will still yield ~100 candidates for large QTLs, which is challenging for genetic characterization unless at least a medium-throughput phenotyping method is available. Fortunately, plant science is entering an era of high-throughput phenotyping with advances in automation, computation and sensor technology (Araus *et al.* 2018; Fahlgren *et al.* 2015). Our study establishes an extendable framework that can be easily updated with new training datasets and features. As more causal genes are uncovered, the new data can be easily incorporated to improve the models.

Important features for predicting QTL causal genes

Many causal genes were repeatedly found to cause phenotypic variation of similar traits, which is also known as genetic hotspots of phenotypic variation or gene reuse (Martin and Orgogozo 2013). By examining 1,008 causative alleles in animals, plants, and yeasts, Martin and Orgogozo found *de novo* mutations to occur repeatedly at certain genes or orthologous loci and causing similar phenotypic variations either among lineages or within a single lineage. The prevalence of gene reuse suggests that causal genes are likely to have some genetic and genomic characteristics that allow them to be repeatedly used for phenotypic variation. The mechanism for gene reuse is not clear but it may be influenced by factors such as the availability of standing genetic variation, mutation rate, pleiotropic constraint, and epistatic interactions of a gene (Conte *et al.* 2015; Conte *et al.* 2012).

While many QTL causal genes have been cloned, their features have not been systematically examined before. Instead of evaluating each feature individually, we trained Random Forest models and evaluated feature importance for all features by adopting the leave-one-out strategy. Several of the most important features were consistent between Arabidopsis and rice models: containing SNPs that cause a premature stop codon, paralog copy number, being a transporter, and being a transcription factor.

We extracted polymorphism features from re-sequencing data, which provide more information about the existence of standing genetic variation in the species than the polymorphism data used for linkage mapping, which typically comes from two parental lines. DNA polymorphisms such as nonsense SNPs, deleterious non-synonymous SNPs, SNPs at cis-regulatory elements, and SNPs at splice junctions have been used as features to classify causal and non-causal variants of human diseases (Jagadeesh *et al.* 2016; Kircher *et al.* 2014). These SNPs can directly affect the function or expression of a gene and therefore are more likely to be causal than the rest of the SNPs. Our results indicate Arabidopsis and rice causal genes were more likely to carry a SNP that causes premature stop codon (nonsense SNP) than an average gene in the genome. We also found Arabidopsis causal genes were more likely to have more non-synonymous SNPs than an average gene in the genome. Besides the high impact SNPs in coding regions, we also examined polymorphisms in non-coding regions since about 90% of human trait/disease-associated SNPs are located outside of coding regions (Hindorff *et al.* 2009). The SNPs at cis-regulatory elements did not show a high feature importance in our algorithm, although this might be due to limited exploration of non-coding sequences in plants. The CIS-BP database contains cis-elements of 44% of the transcription factors in Arabidopsis (Weirauch *et al.* 2014). Developing a more accurate and complete map of functional non-coding regions based on conserved noncoding sequences (Van de Velde *et al.* 2014) will potentially make non-coding polymorphism features more useful for prioritizing causal genes.

Paralogs contribute to the evolution of plant traits by providing functional divergence that gives plants the potential to adapt to complex environments (Panchy *et al.* 2016). Through evolution, genes involved in signal transduction and stress response have retained more paralogs while essential genes like DNA gyrase A have retained fewer paralogs (Lloyd *et al.* 2015; Panchy *et al.* 2016). By acquiring new functions or sub-functions, paralogs allow plants to sense and handle different environmental conditions in a more comprehensive and adjustable way. For example, the eight paralogous heavy metal ATPases (HMAs) in Arabidopsis are all involved in heavy metal transport but have different substrate preference, tissue expression patterns, and subcellular compartment locations (Takahashi *et al.* 2012). Three of them, HMA3, HMA4, HMA5, are known causal genes of QTLs identified by linkage mapping. The known causal genes we analyzed have more paralog copies than other genes in the genome. This may suggest that many plant causal genes are playing a role in providing more phenotypic tuning parameters to allow plants to adapt to complex environments.

Plant transporters are involved in nutrient uptake, response to abiotic stresses, pathogen resistance, and other plant-environment interactions (Conde *et al.* 2011; Doidy *et al.* 2012). Polymorphisms in transporters play an important role in local adaptation since many transporters are directly involved in environment

responses (Baxter *et al*. 2010; Turner *et al*. 2010). For example, in *Arabidopsis lyrata*, the polymorphisms most strongly associated with soil type are enriched in metal transporters (Turner *et al*. 2010). We observed a higher frequency of causal genes being transporters than the average gene in the genome. Causal transporters that contribute to trait variation may have a more important role in local adaptation than other transporters.

Transcription factors were enriched in causal genes not only in plants but also in other organisms (Martin and Orgogozo 2013). This enrichment may be due to an ascertainment bias since linkage mapping tends to identify genes with large effects (Martin and Orgogozo 2013). Since QTG-Finder focuses on prioritizing the causal genes identified by linkage mapping, this feature is useful in distinguishing them from other causal genes such as the medium-effect genes that can be detected by GWAS but not by linkage mapping.

Overall, QTG-Finder is a novel machine-learning pipeline to prioritize causal genes for QTLs identified by linkage mapping. We trained QTG-Finder models for Arabidopsis and rice based on known causal genes from each species, respectively. By utilizing information like polymorphisms, function annotations, co-function networks, and paralog copy numbers, the models can rank QTL genes to prioritize causal genes. Our independent literature validation demonstrates that the models can correctly prioritize about 65% of causal genes for Arabidopsis and 60% for rice when the top 20% of ranked QTL genes were considered. The algorithm is applicable to any traditional quantitative traits but the performance was different for each trait type. Since QTG-Finder is a machine-learning based pipeline, extending the training set and adding features can easily expand and improve the models. We envision that frameworks like QTG-Finder can accelerate the discovery of novel quantitative trait genes by reducing the number of candidate genes and efforts of experimental testing.

Acknowledgements

We thank Dr. John Lloyd and Dr. Shin-Han Shiu for sharing the data of rice essential gene prediction. We thank Kevin Radja for testing the source code and giving useful comments.

Funding

This work was supported by the United States Department of Energy's Biological and Environmental Research Program [DE-SC0008769, DE-SC0018277].

Conflict of interest

The authors declare no conflict of interest.

References

Araus JL, Kefauver SC, Zaman-Allah M, Olsen MS, Cairns JE (2018) Translating High-Throughput Phenotyping into Genetic Gain Trends Plant Sci 23:451-466 doi:10.1016/j.tplants.2018.02.001

Bargsten JW, Nap JP, Sanchez-Perez GF, van Dijk AD (2014) Prioritization of candidate genes in QTL regions based on associations between traits and biological processes BMC Plant Biol 14:330 doi:10.1186/s12870-014-0330-3

Baxter I, Brazelton JN, Yu D, Huang YS, Lahner B, Yakubova E, Li Y, Bergelson J et al. (2010) A Coastal Cline in Sodium Accumulation in *Arabidopsis thaliana* Is Driven by Natural Variation of the Sodium Transporter AtHKT1;1 PLoS Genet 6:e1001193 doi:10.1371/journal.pgen.1001193

Bentsink L, Hanson J, Hanhart CJ, Blankestijn-de Vries H, Coltrane C, Keizer P, El-Lithy M, Alonso-Blanco C et al. (2010) Natural variation for seed dormancy in Arabidopsis is regulated by additive genetic and molecular pathways Proc Natl Acad Sci U S A 107:4264-4269 doi:10.1073/pnas.1000410107

Bergelson J, Roux F (2010) Towards identifying genes underlying ecologically relevant traits in *Arabidopsis thaliana* Nat Rev Genet 11:867-879 doi:10.1038/nrg2896

Brachi B, Faure N, Horton M, Flahauw E, Vazquez A, Nordborg M, Bergelson J, Cuguen J et al. (2010) Linkage and association mapping of *Arabidopsis thaliana* flowering time in nature PLoS Genet 6:e1000940 doi:10.1371/journal.pgen.1000940

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S et al. (2009) The Genetic Architecture of Maize Flowering Time Science 325:714-718 doi:10.1126/science.1174276

Carlborg O, Haley CS (2004) Epistasis: too often neglected in complex trait studies? Nat Rev Genet 5:618-U614 doi:10.1038/nrg1407

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu XY et al. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w(1118); iso-2; iso-3 Fly 6:80-92 doi:10.4161/fly.19695

Conde A, Chaves MM, Geros H (2011) Membrane Transport, Sensing and Signaling in Plant Adaptation to Environmental Stress Plant Cell Physiol 52:1583-1602 doi:10.1093/pcp/pcr107

Consortium TG (2016) 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana Cell 166:481-491 doi:10.1016/j.cell.2016.05.063

Conte GL, Arnegard ME, Best J, Chan YF, Jones FC, Kingsley DM, Schluter D, Peichel CL (2015) Extent of QTL Reuse During Repeated Phenotypic Divergence of Sympatric Threespine Stickleback Genetics 201:1189-1200 doi:10.1534/genetics.115.182550

Conte GL, Arnegard ME, Peichel CL, Schluter D (2012) The probability of genetic parallelism and convergence in natural populations Proc Biol Sci 279:5039-5047 doi:10.1098/rspb.2012.2146

Cortes C, Vapnik V (1995) Support-vector networks Machine Learning 20:273-297 doi:10.1007/BF00994018

Daware AV, Srivastava R, Singh AK, Parida SK, Tyagi AK (2017) Regional Association Analysis of MetaQTLs Delineates Candidate Grain Size Genes in Rice Front Plant Sci 8:807 doi:10.3389/fpls.2017.00807

471 Deo RC, Musso G, Tasan M, Tang P, Poon A, Yuan C, Felix JF, Vasan RS et al. (2014) Prioritizing causal
472 disease genes using unbiased genomic features Genome Biol 15:534 doi:10.1186/s13059-014-
473 0534-8

474 Dinka SJ, Campbell MA, Demers T, Raizada MN (2007) Predicting the size of the progeny mapping
475 population required to positionally clone a gene Genetics 176:2035-2054
476 doi:10.1534/genetics.107.074377

477 Doidy J, Grace E, Kuhn C, Simon-Plas F, Casieri L, Wipf D (2012) Sugar transporters in plants and in their
478 interactions with fungi Trends Plant Sci 17:413-422 doi:10.1016/j.tplants.2012.03.009

479 Fahlgren N, Gehan MA, Baxter I (2015) Lights, camera, action: high-throughput plant phenotyping is
480 ready for a close-up Curr Opin Plant Biol 24:93-99 doi:10.1016/j.pbi.2015.02.006

481 FAO (2009) How to feed the world in 2050

482 Gelfman S, Wang Q, McSweeney KM, Ren Z, La Carpia F, Halvorsen M, Schoch K, Ratzon F et al. (2017)
483 Annotating pathogenic non-coding variants in genic regions Nat Commun 8:236
484 doi:10.1038/s41467-017-00141-2

485 Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talon M et al. (2008)
486 High-throughput functional annotation and data mining with the Blast2GO suite Nucleic Acids
487 Res 36:3420-3435 doi:10.1093/nar/gkn176

488 Grant CE, Bailey TL, Noble WS (2011) FIMO: scanning for occurrences of a given motif Bioinformatics
489 27:1017-1018 doi:10.1093/bioinformatics/btr064

490 He H (2014) Environmental Regulation of Seed Performance. Dissertation. Wageningen University

491 Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential
492 etiologic and functional implications of genome-wide association loci for human diseases and
493 traits Proc Natl Acad Sci U S A 106:9362-9367 doi:10.1073/pnas.0903103106

494 Hormozdiari F, Kichaev G, Yang WY, Pasaniuc B, Eskin E (2015) Identification of causal genes for
495 complex traits Bioinformatics 31:206-213 doi:10.1093/bioinformatics/btv240

496 Huang YF, Gulko B, Siepel A (2017) Fast, scalable prediction of deleterious noncoding variants from
497 functional and population genomic data Nat Genet 49:618-624 doi:10.1038/ng.3810

498 Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, Bernstein JA, Bejerano G
499 (2016) M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at
500 high sensitivity Nat Genet 48:1581-1586 doi:10.1038/ng.3703

501 Jones CE, Brown AL, Baumann U (2007) Estimating the annotation error rate of curated GO database
502 sequence annotations BMC Bioinform 8:170 doi:10.1186/1471-2105-8-170

503 Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J et al. (2014)
504 InterProScan 5: genome-scale protein function classification Bioinformatics 30:1236-1240
505 doi:10.1093/bioinformatics/btu031

506 Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J (2014) A general framework for
507 estimating the relative pathogenicity of human genetic variants Nat Genet 46:310-315
508 doi:10.1038/ng.2892

509 Kuhn M, Johnson K (2013) Applied Predictive Modeling. New York : Springer. doi:10.1007/978-1-4614-
510 6849-3

511 Lee I, Ambaru B, Thakkar P, Marcotte EM, Rhee SY (2010) Rational association of genes with traits using
512 a genome-scale gene network for *Arabidopsis thalian*a Nat Biotechnol 28:149-U114
513 doi:10.1038/nbt.1603

514 Lee I, Seo YS, Coltrane D, Hwang S, Oh T, Marcotte EM, Ronald PC (2011) Genetic dissection of the
515 biotic stress response using a genome-scale gene network for rice Proc Natl Acad Sci U S A
516 108:18548-18553 doi:10.1073/pnas.1110384108

517 Leinonen PH, Remington DL, Leppala J, Savolainen O (2013) Genetic basis of local adaptation and
518 flowering time variation in *Arabidopsis lyrata* Mol Ecol 22:709-723 doi:10.1111/j.1365-
519 294X.2012.05678.x

520 Liu XQ, Wang L, Chen S, Lin F, Pan QH (2005) Genetic and physical mapping of *Pi36(t)*, a novel rice
521 blast resistance gene located on rice chromosome 8 Mol Genet Genomics 274:394-401
522 doi:10.1007/s00438-005-0032-5

523 Lloyd JP, Seddon AE, Moghe GD, Simenc MC, Shiu S-H (2015) Characteristics of Plant Essential Genes
524 Allow for within- and between-Species Prediction of Lethal Mutant Phenotypes Plant Cell
525 27:2133-2147 doi:10.1105/tpc.15.00051

526 Mackay TFC (2001) The genetic architecture of quantitative traits Annu Rev Genet 35:303-339 doi:DOI
527 10.1146/annurev.genet.35.102401.090633

528 Mackay TFC (2014) Epistasis and Quantitative Traits: Using Model Organisms to Study Gene-Gene
529 Interactions Nat Rev Genet 15:22-33 doi:10.1038/nrg3627

530 Mansueto L, Fuentes RR, Borja FN, Detras J, Abriol-Santos JM, Chebotarov D, Sanciangco M, Palis K et
531 al. (2017) Rice SNP-seek database update: new SNPs, indels, and queries Nucleic Acids Res
532 45:D1075-D1081 doi:10.1093/nar/gkw1135

533 Martin A, Orgogozo V (2013) The Loci of repeated evolution: a catalog of genetic hotspots of phenotypic
534 variation Evolution 67:1235-1250 doi:10.1111/evo.12081

535 Mordelet F, Vert JP (2011) ProDiGe: Prioritization Of Disease Genes with multitask machine learning
536 from positive and unlabeled examples BMC Bioinform 12:389 doi:10.1186/1471-2105-12-389

537 Motte H, Vercauteren A, Depuydt S, Landschoot S, Geelen D, Werbrouck S, Goormachtig S, Vuylsteke M
538 et al. (2014) Combining linkage and association mapping identifies RECEPTOR-LIKE PROTEIN
539 KINASE1 as an essential Arabidopsis shoot regeneration gene Proc Natl Acad Sci U S A
540 111:8305-8310 doi:10.1073/pnas.1404978111

541 Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function Nucleic Acids
542 Res 31:3812-3814

543 Nuzhdin SV, Dilda CL, Mackay TF (1999) The genetic architecture of selection response. Inferences from
544 fine-scale mapping of bristle number quantitative trait loci in *Drosophila melanogaster* Genetics
545 153:1317-1331

546 Otto SP, Jones CD (2000) Detecting the undetected: Estimating the total number of loci underlying a
547 quantitative trait Genetics 156:2093-2107

548 Panchy N, Lehti-Shiu M, Shiu SH (2016) Evolution of Gene Duplication in Plants Plant Physiol 171:2294-
549 2316 doi:10.1104/pp.16.00523

550 Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P et al.
551 (2011) Scikit-learn: Machine Learning in Python J Mach Learn Res 12:2825-2830

552 Perez-Iratxeta C, Bork P, Andrade MA (2002) Association of genes to genetically inherited diseases using
553 data mining Nat Genet 31:316 doi:10.1038/ng895

554 Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A (2010) Detection of nonneutral substitution rates on
555 mammalian phylogenies Genome Res 20:110-121 doi:10.1101/gr.097857.109

556 Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-
557 wide association studies Nat Rev Genet 11:459-463 doi:10.1038/nrg2813

558 Ritchie GR, Dunham I, Zeggini E, Flicek P (2014) Functional annotation of noncoding sequence variants
559 Nat Methods 11:294-296 doi:10.1038/nmeth.2832

560 Schaefer R, Michno J-M, Jeffers J, Hoekenga OA, Dilkes BP, Baxter IR, Myers C (2018) Integrating co-
561 expression networks with GWAS to prioritize causal genes in maize Plant Cell
562 doi:10.1105/tpc.18.00299

563 Schlapfer P, Zhang P, Wang C, Kim T, Banf M, Chae L, Dreher K, Chavali AK et al. (2017) Genome-
564 Wide Prediction of Metabolic Enzymes, Pathways, and Gene Clusters in Plants Plant Physiol
565 173:2041-2059 doi:10.1104/pp.16.01942

566 Takahashi R, Bashir K, Ishimaru Y, Nishizawa NK, Nakanishi H (2012) The role of heavy-metal ATPases,
567 HMAs, in zinc and cadmium transport in rice Plant Signal Behav 7:1605-1607
568 doi:10.4161/psb.22454

569 Tin Kam H (1998) The random subspace method for constructing decision forests IEEE Transactions on
570 Pattern Analysis and Machine Intelligence 20:832-844 doi:10.1109/34.709601

571 Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, van Hijum SA (2013) Data mining
572 in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? Brief Bioinform
573 14:315-326 doi:10.1093/bib/bbs034

574 Tuinstra MR, Ejeta G, Goldsbrough PB (1997) Heterogeneous inbred family (HIF) analysis: a method for
575 developing near-isogenic lines that differ at quantitative trait loci Theor Appl Genet 95:1005-1011
576 doi:10.1007/s001220050654

577 Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin SV (2010) Population resequencing reveals
578 local adaptation of *Arabidopsis lyrata* to serpentine soils Nat Genet 42:260-263
579 doi:10.1038/ng.515

580 Van de Velde J, Heyndrickx KS, Vandepoele K (2014) Inference of Transcriptional Networks in
581 Arabidopsis through Conserved Noncoding Sequence Analysis Plant Cell 26:2729-2745
582 doi:10.1105/tpc.114.127001

583 Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA
584 et al. (2014) Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity
585 Cell 158:1431-1443 doi:10.1016/j.cell.2014.08.009

586 Wellenreuther M, Hansson B (2016) Detecting Polygenic Evolution: Problems, Pitfalls, and Promises
587 Trends Genet 32:155-164 doi:10.1016/j.tig.2015.12.004

588 Xu S (2003) Theoretical basis of the Beavis effect Genetics 165:2259-2268

589 Yang HY, You AQ, Yang ZF, Zhang F, He RF, Zhu LL, He G (2004) High-resolution genetic mapping at
590 the Bph15 locus for brown planthopper resistance in rice (*Oryza sativa* L.) Theor Appl Genet
591 110:182-191 doi:10.1007/s00122-004-1844-0

592 Yin ZG, Qi HD, Chen QS, Zhang ZG, Jiang HW, Zhu RS, Hu ZB, Wu XX et al. (2017) Soybean plant
593 height QTL mapping and meta-analysis for mining candidate genes Plant Breed 136:688-698
594 doi:10.1111/pbr.12500

595 Zhang H (2004) The Optimality of Naive Bayes Proc FLAIRS
596

597

598

599 **Figure Captions**

600 **Fig. 1** Model training and optimization based on cross-validation. (a) model training and cross-validation
601 framework. We randomly selected negatives from the genome and iterated to maximize the combinations
602 of training and testing data. (b) The ROC curve of Arabidopsis and rice models after parameter
603 optimization. True and false positive rates were based on the average of all iterations. The grey diagonal
604 line indicates the expected performance based on random guessing. The number in parentheses indicates
605 Area Under the ROC Curve (AUC-ROC)

606 **Fig. 2** Important features of causal genes and their enrichment or depletion relative to the genome
607 background (a) Feature importance as indicated by the change of AUC-ROC (ΔAUC-ROC) when
608 excluding each feature. The ΔAUC-ROC indicates the average value of all iterations. Error bars indicate
609 standard deviation. The features with a name that starts with "is_" are binary variables. (b) The enrichment
610 or depletion of the top 6 features in Arabidopsis and rice models. The enrichment/depletion were indicated
611 by the ratio of causal genes to genome background. ns, not shown because the feature is not one of the top
612 6 features in that species

613 **Fig. 3** Model performance at different thresholds (a) Percentage of correctly prioritized causal genes of a
614 single QTL at different rank thresholds. Dashed lines indicate the background of random selections. (b-c)
615 The probability of correctly prioritizing at least X% of causal genes when analyzing multiple QTLs
616 simultaneously

617 **Fig. 4** (a) Trait categories of known causal genes from the training set. (b) The rank percentile of causal
618 genes of different trait categories. Each causal gene and 200 neighboring genes were used as testing set
619 once. All other known causal genes were used as training set. Each dot indicates a known causal gene. The
620 grey dashed line indicates 20% rank percentile. The trait categories of causal genes are defined in Tables
621 S1 and S2
622

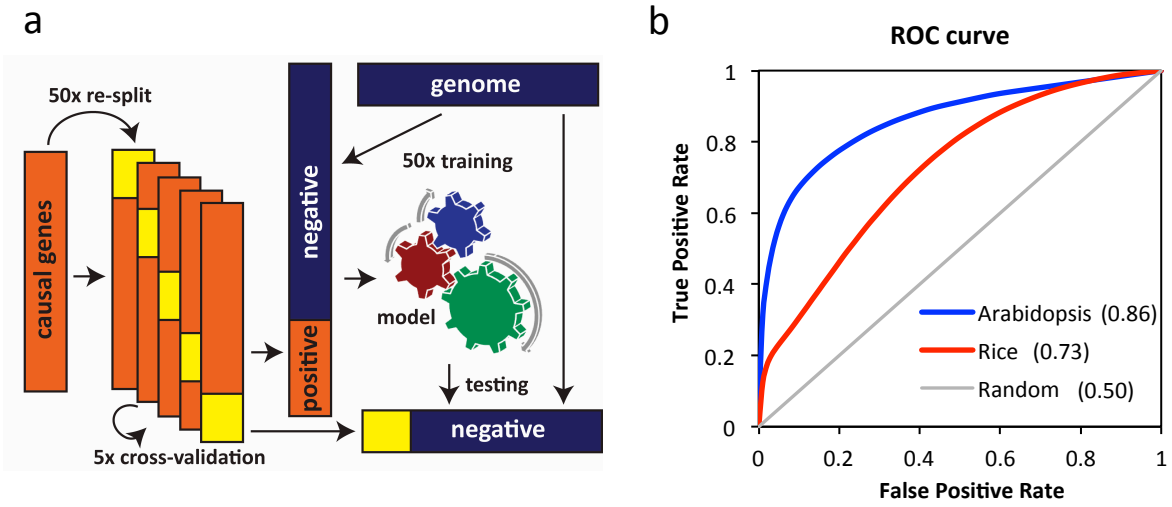

Fig. 1 Model training and optimization based on cross-validation. (a) model training and cross-validation framework. We randomly selected negatives from the genome and iterated to maximize the combinations of training and testing data. (b) The ROC curve of Arabidopsis and rice models after parameter optimization. True and false positive rates were based on the average of all iterations. The grey diagonal line indicates the expected performance based on random guessing. The number in parentheses indicates Area Under the ROC Curve (AUC-ROC)

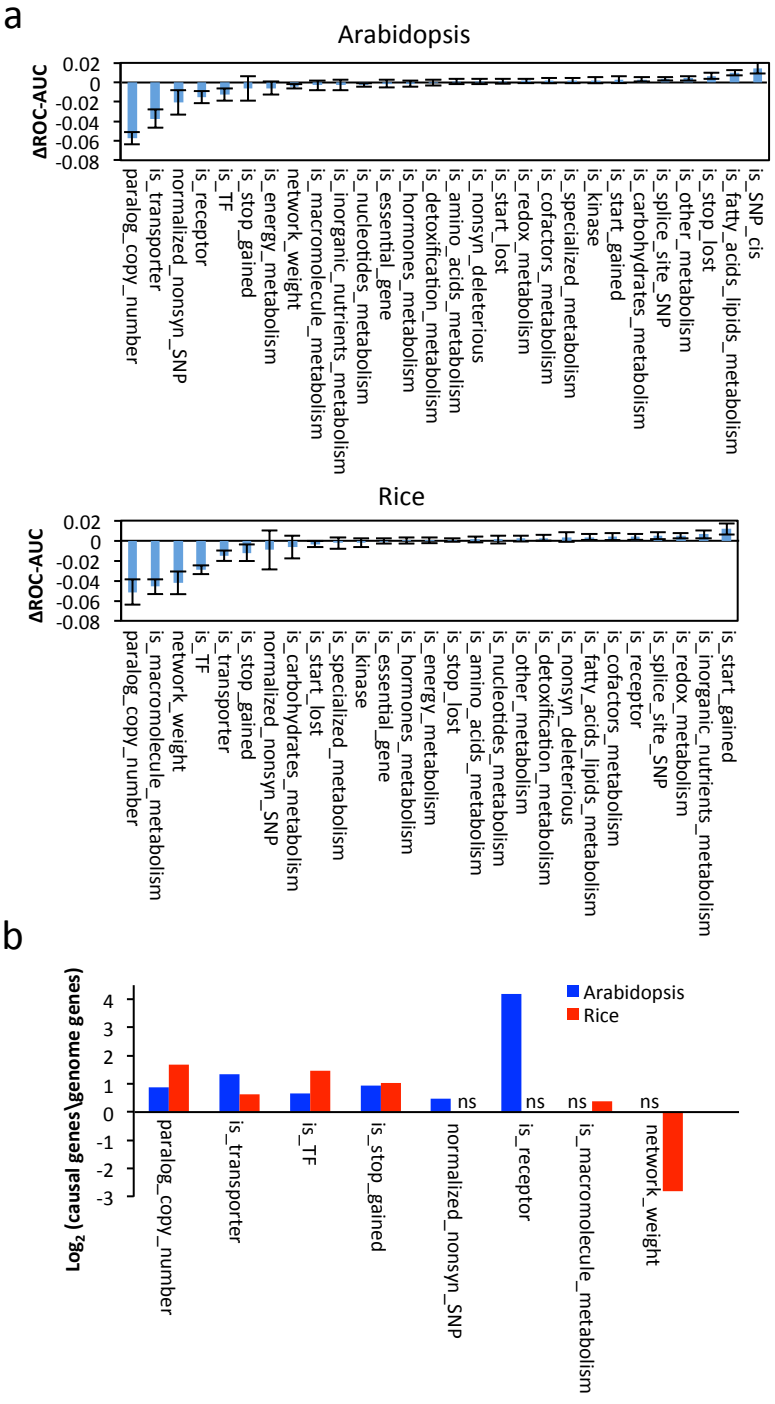

Fig. 2 Important features of causal genes and their enrichment or depletion relative to the genome background (a) Feature importance as indicated by the change of AUC-ROC (ΔAUC-ROC) when excluding each feature. The ΔAUC-ROC indicates the average value of all iterations. Error bars indicate standard deviation. The features with a name that starts with "is_" are binary variables. (b) The enrichment or depletion of the top 6 features in Arabidopsis and rice models. The enrichment/depletion were indicated by the ratio of causal genes to genome background. ns, not shown because the feature is not one of the top 6 features in that species

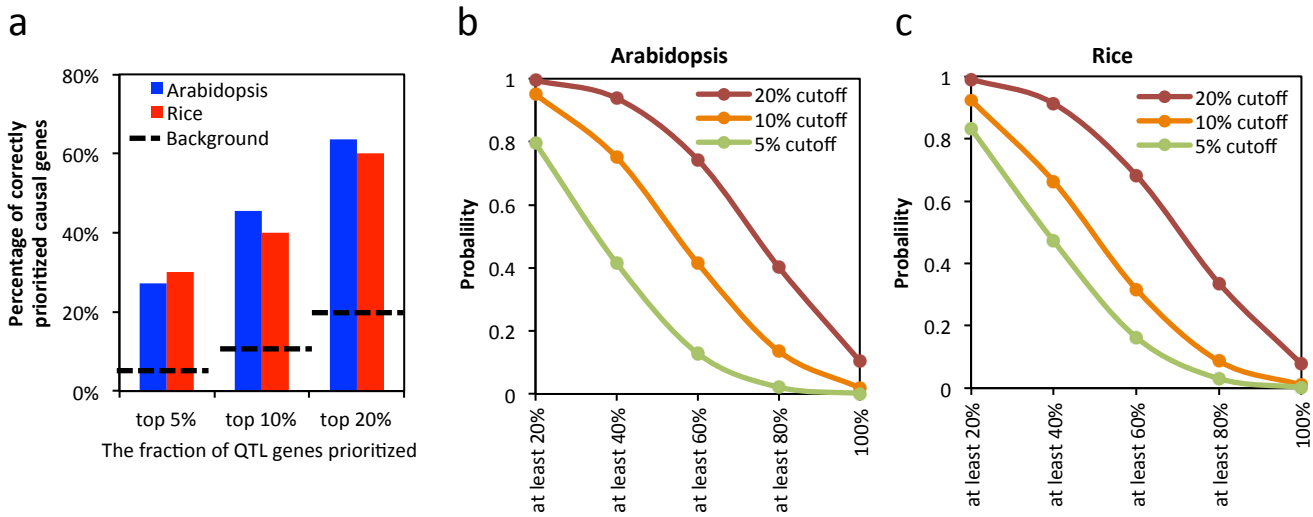

Fig. 3 Model performance at different thresholds (a) Percentage of correctly prioritized causal genes of a single QTL at different rank thresholds. Dashed lines indicate the background of random selections. (b-c) The probability of correctly prioritizing at least X% of causal genes when analyzing multiple QTLs simultaneously

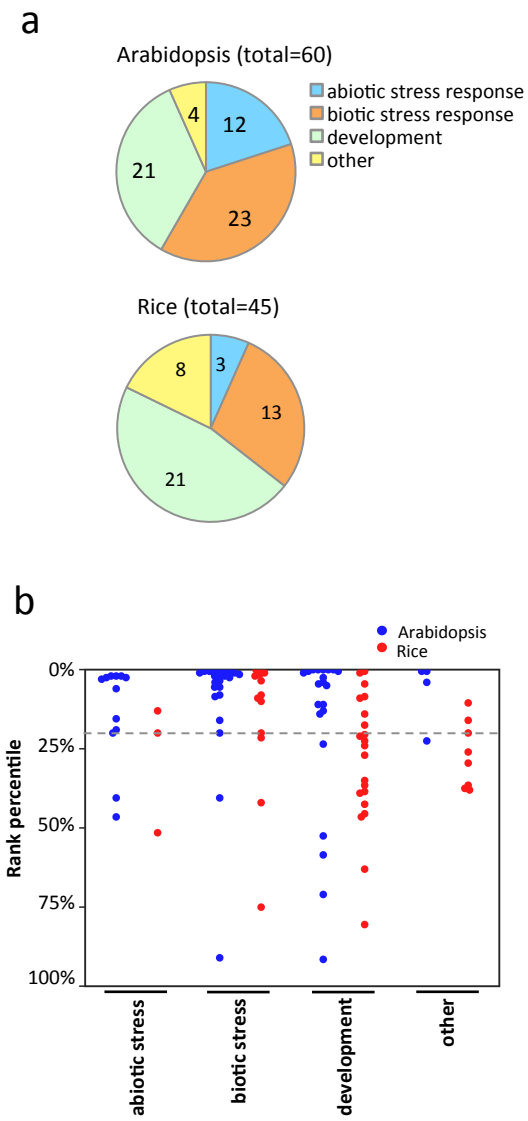

Fig. 4 (a) Trait categories of known causal genes from the training set. (b) The rank percentile of causal genes of different trait categories. Each causal gene and 200 neighboring genes were used as testing set once. All other known causal genes were used as training set. Each dot indicates a known causal gene. The grey dashed line indicates 20% rank percentile. The trait categories of causal genes are defined in Tables S1 and S2