QTG-Finder: a machine-learning algorithm to prioritize causal genes of quantitative trait loci in Arabidopsis and rice

Fan Lin, Jue Fan, Seung Y. Rhee*

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

Contact: srhee@carnegiescience.edu

Running title: QTL causal gene prioritization algorithm

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 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 80% and 55% 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 non-synonymous SNPs at conserved protein sequences, paralog copy number, and being a transporter. This work lays the foundation for systematically understanding characteristics of causal genes and establishes a pipeline to predict causal genes based on public data.

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

Introduction

As the world 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). The number of undernourished people in the world has been on the rise since 2014, reaching 821 million in 2017 (FAO, 2018). To catch up with the growing global food demand, it is important to improve the efficiency of arable lands 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 (e.g. epistasis) (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 a recombinant inbred line (Bergelson and Roux, 2010). GWAS mapping is based 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, though some methods have been developed to ameliorate it

65 (Price *et al.*, 2010). However, efforts to correct it could result in a higher false negative rate
66 (Brachi *et al.*, 2010). Linkage mapping does not suffer from these issues, but it has a relatively
67 lower mapping resolution and cannot identify QTLs of minor effects when the sample size is
68 small (Martin and Orgogozo, 2013; Otto and Jones, 2000; Wellenreuther and Hansson, 2016;
69 Xu, 2003).

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

78 Conventional fine mapping is a reliable but time-consuming and labor-intensive approach to
79 narrow down the range of candidate genes in a QTL region. The basis of fine mapping is to
80 create a population that has more recombination events within a QTL in order to identify a
81 smaller genomic segment that co-segregates with the trait. However, the enormous time and
82 labor required for creating and screening a population of progenies limits the usage of this
83 method (Tuinstra *et al.*, 1997). Depending on the frequency of recombination, thousands of
84 progenies may need to be genotyped to get to a gene-scale resolution (Dinka *et al.*, 2007). For
85 example, 1160 progenies were screened to identify the *Pi36* gene in rice and as many as 18,994
86 progenies were screened to identify the causal gene of Bph15 in rice (Liu *et al.*, 2005; Yang *et*
87 *al.*, 2004). The high cost associated with genotyping and phenotyping makes it challenging to
88 apply fine mapping to all QTLs.

89 Alternative approaches to refine the candidate list of causal genes include meta-analysis,
90 joint linkage-association analysis, and other computational methods including machine-learning
91 algorithms. The first two approaches require either the availability of many QTL studies on
92 similar traits or an additional association mapping experiment (Buckler *et al.*, 2009; Motte *et al.*,
93 2014; Yin *et al.*, 2017). Computational methods including machine-learning algorithms have
94 been developed to prioritize human disease associated genes and genetic variants (Hormozdiari
95 *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* and *Oryza sativa* 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.

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 52 Arabidopsis and 37 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 genome. The features used for model training include polymorphisms, functional categories of genes, function interference from co-function networks, and gene essentiality, and paralog copy number (Table 1, Supplementary Tables S3 and S4). These features were generally independent from each other (most have a Pearson's correlation coefficient <0.2) (Supplementary Figure S1).

We optimized the models with an extended cross-validation framework (Figure 1A). In

127 addition to a typical 5-fold cross-validation (Kuhn and Johnson, 2013), iterations were applied to
128 randomly select genes from the negative set (100 times) and re-split positive set (50 times) in
129 order to maximize the combinations of positives and negatives in the training and testing sets.
130 With this framework, we evaluated the training performance with Area Under the Curve of
131 Receiver Operating Characteristic (AUC-ROC) and optimized parameters. To find the optimal
132 parameters, we compared the AUC-ROC of different machine-learning classifiers, modeling
133 parameters, and the ratio of positive:negative genes in the training set (Supplementary Figures
134 S2, S3, and S4). Random Forest was selected as the classifier since it was less prone to over-
135 fitting and performed better than other classifiers tested (Supplementary Figure S2). After
136 optimization, AUC-ROC for the Arabidopsis and rice models were 0.82 and 0.74, respectively
137 (Figure 1B).
138
139 **Important features for predicting causal genes**
140 With the optimized models, we wanted to know which features were important for causal gene
141 prediction. Since Random Forest uses all features and their interactions for classification (Touw
142 *et al.*, 2013), the importance of a feature cannot be measured by simple enrichment or depletion
143 of a single feature in causal genes. Therefore, we evaluated feature importance based on the
144 change of ROC-AUC (**Δ**ROC-AUC) when excluding a feature from the model (Lloyd *et al.*,
145 2015). When an important feature is excluded from the model, the ROC-AUC should decrease.
146 Four out of five of the most important features were consistent between the Arabidopsis and
147 rice models (Figure 2A). The five most important features for Arabidopsis were paralog copy
148 number, transporter, network weight sum, receptor, and deleterious non-synonymous SNP at
149 conserved protein sequences. The five most important features for rice were network weight
150 sum, transporter, deleterious non-synonymous SNP, macromolecule metabolism, and paralog
151 copy number. Receptor and macromolecule metabolism were features unique to Arabidopsis and
152 rice, respectively.
153 For the five most important features in Arabidopsis and rice, we examined their enrichment
154 or depletion in known causal genes (Figure 2B). Compared to other genes in the genome, the
155 causal genes tended to have more paralogs, higher frequency of containing a deleterious non-
156 synonymous SNP and higher frequency of being a transporter in both species. For network
157 weight sum, the trend was not consistent between Arabidopsis and rice. In Arabidopsis, causal

genes had higher network weight sum than the rest of the genes in the genome. In contrast, rice causal genes had lower network weight sum than the rest of the genome.

The rest of the features contributed less to model performance but did not impair model performance to a large degree (ΔROC-AUC< 0.02). Since there was no strong evidence that they impair prediction, we did not remove them from the models for further analysis.

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

To assess the predictability of QTG-Finder models, we searched the literature for a separate set of known causal genes from the original training set. We found nine Arabidopsis and nine rice genes that had been identified as causal genes underlying QTLs by map-based cloning, joint linkage-association analysis or mutant study (Table 2). These known causal genes were not used for model training or cross-validation.

To examine model performance, we applied the QTG-Finder models to this new set of causal genes. For each known causal gene, we ranked all genes in the QTL region based on the frequency of being predicted as a causal gene from 5000 iterations. Since the number of genes in a QTL region varies, we used a gene's rank percentile for evaluation. The rank percentile of a gene indicates the percentage of QTL genes that had higher ranks than the gene of interest.

Based on the rank of these known causal genes, we evaluated model performance at different cutoffs. We calculated the percentage of known causal genes being included in the top 5%, 10%, and 20% prioritized genes within a QTL (Figure 3A). The top 20% of the ranked genes included seven Arabidopsis (~78%) and five rice (~56%) causal genes. With a more stringent cutoff of 5%, four Arabidopsis (~44%) and three rice (~33%) causal genes were prioritized.

Most linkage mapping studies identify multiple QTLs. We therefore calculated a theoretical model performance on identifying causal genes from multiple QTLs simultaneously, which we defined as the probability of identifying at least X% of all causal genes when applying the model to all QTLs of a trait (Figure 3B and C). For example, assuming there were five QTLs of a trait identified by a linkage mapping study and each QTL contained one causal gene. The probability of identifying at least one causal gene would be 99% when the top 20% genes of all QTLs were tested experimentally. The probability of identifying all five causal genes was 28% when the top 20% cutoff was used. We further compared the performance of all three cutoffs, top 20%, top

10%, and top 5%. For the Arabidopsis model, the probability to identify at least one out of five causal gene is no less than 95% for all three cutoffs. The probability to correctly prioritize at least four out of five causal genes is 69% (for top 20%), 26% (for top 10%), and 13% (for,top 5%). Therefore, a less stringent cutoff (top 20%) performs better than a more stringent cutoff if one is interested in finding most of the causal genes or causal genes of a particular QTL. However, if the goal is to identify any causal genes, then screening the top 5% of all QTLs may be a more strategic approach.

Trait type preference of QTG-Finder models

Since the training set included genes for different types of traits at an imbalanced ratio, we wanted to know how QTG-Finder models would work for each type of traits (Figure 4A). The independent validation described above was based on causal genes related to plant development and disease resistance (Table 2). However, this validation set was not large enough for a systematic analysis and 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 original training set (52 for Arabidopsis and 37 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 made into 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 Arabidopsis and rice models had similar performance for most trait categories except abiotic stress response traits. In Arabidopsis, abiotic stress response traits had better performance than the other traits. However, in rice, abiotic stress response traits performed worse than the other traits. The "other" category performed poorly for both species. The worst performing traits all have small number of samples in the training set (Figure 4A). This trait category analysis will 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 quantitative traits. However, 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

220 candidates to be tested experimentally in order to accelerate the discovery of causal genes.
221
222 **A machine-learning algorithm to prioritize QTL causal genes**
223 Several causal variant or gene prioritization methods have been developed for human data but
224 not many in plants (Bargsten *et al.*, 2014; Jagadeesh *et al.*, 2016; Kircher *et al.*, 2014; Schaefer *et*
225 *al.*, 2018). Most prioritization methods have been developed for GWAS mapping in human, an
226 organism where linkage mapping cannot be performed. However, linkage mapping can capture
227 rare alleles and has been broadly used to study quantitative traits of livestock, crops, and model
228 organisms. A causal gene prioritization is especially helpful for the large QTLs identified by
229 linkage mapping, which can constitute hundreds to thousands of genes. One method has been
230 developed in rice to prioritize causal genes for linkage mapping (Bargsten *et al.*, 2014). This
231 method is based on the hypothesis that causal genes from multiple QTLs of the same trait are
232 more likely to have the same biological process GO terms, and therefore genes with
233 overrepresented biological process GOs were prioritized as causal genes. However, there are
234 fundamental issues of using GO annotation as the only criterion because of annotation bias,
235 where a large proportion of annotations are for only the small proportion of well-studied genes
236 (Tomczak *et al.*, 2018). It is not known if this GO-based method favors identifying causal genes
237 with known functions over other types of causal genes. In addition, an unbiased performance
238 evaluation is still lacking for this method since the same set of causal genes was used to
239 determine cutoff and evaluate performance.
240 In this study, we built a supervised learning algorithm using multiple features and validated
241 its efficacy with an independent dataset from the literature. The models could accelerate the
242 discovery of causal genes by ranking all the genes in a QTL region and prioritizing the top 5%,
243 10%, or 20% genes, which are most likely to contain the causal gene, for experimental testing.
244 Based on an assessment using independent data in the literature, we calculated the performance
245 when applying the models to all QTLs of a trait and compared three cutoffs (top 5%, 10%, and
246 20%). The less stringent cutoff (top 20%) had a higher chance to find more causal genes (Figure
247 3B and C) but yielded more candidates that needed to be tested by experiments. The more
248 stringent cutoff (top 5%) had a lower chance to find all causal genes but yielded a smaller set of
249 candidates to test. The probability for the models to find at least one causal gene is high for all
250 three cutoffs. If the goal were to find one or more causal genes for functional studies and the

251 particular QTL regions did not matter, the 5% cutoff would be more efficient. If the goal were to
252 discover all causal genes and understand the genetic architecture of a trait, the 20% cutoff would
253 be better. Similarly, if a particular QTL were of interest for discovering the underlying causal
254 gene, the 20% cutoff would be better.

255 There are several conceptual and practical advantages of QTG-Finder algorithm. First, this
256 algorithm combines multiple types of publically available data including polymorphisms,
257 function annotation, co-function network and other genomic data, which have not been applied
258 to prioritize causal genes from linkage mapping studies. Second, models were trained on causal
259 genes from various traits and can be applied to several types of traditional traits, though the
260 prioritization efficiency was not equivalent. Third, validation from the literature provides
261 guidance on what proportion of genes to prioritize in practice rather than arbitrarily selecting a
262 threshold. Fourth, these models can be applied to either a single QTL or multiple QTLs.

263 Two limitations of this study are the small number of known causal genes in plants and the
264 impurity of negative set used for model training. We used 52 Arabidopsis and 37 rice causal
265 genes that have been verified by map-based-cloning as a positive dataset. Even though they are
266 of high quality, this positive dataset may not be large enough to represent all the features of
267 causal genes. There could still be other important features of plant causal genes that we were not
268 able to capture with this small dataset. The negative set was composed of genes randomly
269 selected from the rest of the genome. Though we excluded known causal genes, it could still
270 contain some uncharacterized causal genes. For large QTLs, 20% cutoff will still yield ~100
271 candidates, which is challenging for genetic characterization unless at least a medium-throughput
272 phenotyping method is available.

273 Fortunately, plant science is entering an era of high-throughput phenotyping with advances
274 in automation, computation and sensor technology (Araus *et al.*, 2018; Fahlgren *et al.*, 2015).
275 Our study establishes an extendable framework that can be easily updated with new training
276 datasets and features. As more causal genes are uncovered, the new data can be easily
277 incorporated to improve the models.

278

279 **Important features for predicting QTL causal genes**

280 While many QTL causal genes have been cloned, their features have not been systematically
281 examined before. Instead of evaluating each feature individually, we trained Random Forest

282 models and evaluated feature importance for all features by adopting the leave-one-out strategy.
283 The majority of the most important features were consistent between Arabidopsis and rice
284 models: deleterious non-synonymous SNP, paralog copy number, and being a transporter.

285 DNA polymorphism is the basis of genetic mapping and a major source of trait variation.
286 DNA polymorphisms such as deleterious non-synonymous SNPs, SNPs at cis-regulatory
287 elements, SNPs at splice junctions, and nonsense SNPs have been used as features to classify
288 causal and non-causal variants of human diseases (Jagadeesh *et al.*, 2016; Kircher *et al.*, 2014).
289 These SNPs can change the function or expression of a gene and therefore are more likely to be
290 causal than the rest of the SNPs. Our results indicate Arabidopsis and rice causal genes were
291 more likely to carry a deleterious non-synonymous SNP than an average gene in the genome.
292 This is consistent with a previous study, which showed that the trait/disease-associated SNPs
293 were enriched in non-synonymous sites, especially potentially deleterious sites (Hindorff *et al.*,
294 2009). Despite the enrichment in non-synonymous sites, about 90% of human trait/disease-
295 associated SNPs was located outside coding regions (Hindorff *et al.*, 2009). Therefore,
296 polymorphisms in non-coding regions should not be neglected when prioritizing causal genes.
297 Though the SNPs at cis-regulatory elements did not show high feature importance in our QTG-
298 Finder algorithm, this might be due to limited exploration of non-coding sequences in plants. The
299 CIS-BP database only covers cis-elements of 44% transcription factors in Arabidopsis (Weirauch
300 *et al.*, 2014). Developing a more accurate and complete map of functional non-coding regions
301 based on conserved noncoding sequences (Van de Velde *et al.*, 2014) will potentially make non-
302 coding polymorphism features more useful for prioritizing causal genes.

303 Paralogs contribute to the evolution of plant traits by providing functional divergence that
304 gives plants the potential to adapt to complex environments (Panchy *et al.*, 2016). Through
305 evolution, genes involved in signal transduction and stress response have retained more paralogs
306 while essential genes like DNA gyrase A have retained fewer paralogs (Lloyd *et al.*, 2015;
307 Panchy *et al.*, 2016). By acquiring new functions or sub-functions, paralogs allow plants to sense
308 and handle different environmental conditions in a more comprehensive and adjustable way. For
309 example, the eight paralogous heavy metal ATPases (HMAs) in Arabidopsis are all involved in
310 heavy metal transport but have different substrate preference, tissue expression patterns, and
311 subcellular compartment locations (Takahashi *et al.*, 2012). Three of them, HMA3, HMA4,
312 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 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 gene 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.

The network weight sum was an important feature in our Arabidopsis and rice models but had opposite trends in the two species. In Arabidopsis, causal genes had higher weight sum than other genome genes. In rice, causal genes had lower weight sum. Genes with higher weight sum interact with more genes and tend to influence multiple traits (Promislow, 2004). In natural populations, some pleiotropic genes help achieve the coordinated evolution of multiple, related traits such as the adaptive pleiotropy among floral trait loci (Smith, 2016). The high weight sum of Arabidopsis causal genes indicates that many of them may be pleiotropic and potentially adaptive. Rice is a crop plant that was mostly selected for agronomical traits and mutations with strong pleiotropic effects were not favored during human selection process (Swinnen *et al.*, 2016). This could be a reason that many causal genes of agronomical traits have low weight sum in rice.

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 additional information like polymorphisms, function annotation, co-function network, and paralog copy number, the models can rank QTL genes to prioritize causal genes. Our independent literature validation demonstrates that the models can correctly prioritize about 80% of causal genes for Arabidopsis and 55% 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 pipeline, to further improve its performance,

344 extending the training set and adding features can easily expand the models. We envision that
345 frameworks like QTG-Finder can accelerate the discovery of novel quantitative trait genes by
346 reducing the number of candidate genes and efforts of experimental tests.
347

348 Methods

349 Data sources and processing of features used in QTG-Finder

350 Thirty-two features were extracted from published genome-scale data, which included
351 polymorphism features, functional annotation features and other genomic and functional
352 genomic features.

353 Arabidopsis polymorphism data (Cao *et al.*, 2011; Ossowski *et al.*, 2008) was downloaded
354 from 1001 Genomes Project (https://1001genomes.org) and rice polymorphism data was
355 downloaded from dbSNP (ftp://ftp.ncbi.nlm.nih.gov/snp/, Build138). A stringent quality score
356 cutoff (QUAL<30) was applied to filter out low quality SNPs. We used SIFT4G (v 2.4) and
357 SnpEff (v 4.3r) to annotate the raw polymorphism data. The number of non-synonymous SNP as
358 annotated by SIFT4G was normalized to protein length and used as a numeric feature
359 (normalized_nonsyn_SNP). Non-synonymous SNPs at conserved protein sequences were
360 predicted to cause deleterious amino acid changes by SIFT4G. The deleterious non-synonymous
361 SNPs were used as a binary feature (is_nonsyn_deleterious). If a gene contained at least one
362 deleterious non-synonymous SNP, the "is_nonsyn_deleterious" feature was 1, otherwise was 0.
363 Other binary polymorphism features such as "is_strat_lost" (start codon lost) and
364 "is_start_gained" (start codon gained) were extracted from SnpEff annotations in the same way.
365 For "is_SNP_cis" and "is_INDEL_cis", the Position Weight Matrices of cis-elements were
366 downloaded from CIS-BP database (Build 1.02) (Weirauch *et al.*, 2014) and mapped to 1kb
367 upstream of all genome genes using FIMO (v 4.12.0) (Grant *et al.*, 2011). The cis-elements with
368 a matching score above 55 were imported into SnpEff to annotate SNP and INDEL. This
369 matching score cutoff was determined by the cross-validation results.

370 Functional annotation features were based on GO (Gotz *et al.*, 2008; Jones *et al.*, 2014) and
371 Plant Metabolic Network (PMN) (Schlapfer *et al.*, 2017). Arabidopsis and rice genes were
372 annotated by Blast2GO (BLAST+ 2.2.29) and InterProScan (v 5.3-46.0). The molecular function
373 GOs were then converted to high-level functional groups, such as transcription factor, receptor,

374 kinase, transporter, and enzyme. Genes annotated as enzymes were further classified into PMN
375 metabolic domains (PMN, Release 12, 2017) such as carbohydrate metabolism and nucleotide
376 metabolism. Unclassified genes in PMN were classified as "is_other_metabolism". Genes
377 annotated as enzymes by GO but not present in PMN database are enzymes involved in
378 macromolecule metabolic process or enzymes that don't have a specific function assigned. Since
379 a majority of them is involved in macromolecule metabolic process, we named this group as
380 "is_macromolecule_metabolism".
381 Co-functional networks of Arabidopsis and rice were retrieved from AraNet and RiceNet
382 (Lee *et al.*, 2010; Lee *et al.*, 2011). The sum of all the edge weights of a gene was used as the
383 "network_weight" feature.
384 Paralog copy number (paralog_copy_number) and essential gene prediction (is_essential_gene)
385 were data from a previous publication (Lloyd *et al.*, 2015).

386 **Arabidopsis and rice causal genes used for training and independent validation**

387 For model training and cross-validation, curated causal genes from Martin et al were used as
388 positives for algorithm training (Martin and Orgogozo, 2013). The curated causal genes only
389 included genes that have been studied by fine mapping and were cloned. Causal genes identified
390 from different parental lines were merged into one causal gene entry for training. In total, 52
391 Arabidopsis and 37 rice causal genes were used as the original training set.
392 For literature validation, a separate set of nine casual genes was used for each species. It
393 includes causal genes that have been cloned or indicated by joint linkage-association analysis or
394 indicated by mutant analysis (Chardon *et al.*, 2013; Fan *et al.*, 2016; Fukuoka *et al.*, 2014; Gao *et*
395 *al.*, 2016; Guo *et al.*, 2015; Hu *et al.*, 2008; Huang *et al.*, 2012; Liu *et al.*, 2017; Motte *et al.*,
396 2014; Oikawa *et al.*, 2015; Qi *et al.*, 2008; Sharma *et al.*, 2005; Yuan *et al.*, 2016; Zeng *et al.*,
397 2013)

398 **Algorithm training and parameter optimization**

399 The QTG-Finder algorithm was developed in Python (v 3.6) with the 'sklearn' package (v
400 0.19.0) (Pedregosa *et al.*, 2011). The framework of model training and 5-fold cross-validation is
401 demonstrated in Figure 1A. The curated causal genes were randomly split into training and
402 testing positives in a 4:1 ratio. Training and testing positives were combined with different sets
403 of negative genes that were randomly selected from the rest of the genome. To increase the

404 combination of positives and negatives, we split the positives 50 times randomly and selected
405 negatives 100 times. In total, the modeling was repeated 25000 times for the 5-fold cross-
406 validation. To evaluate training performance, the average AUC-ROC from all iterations was
407 calculated.

408 We tested different classifiers and parameters and optimized the model based on AUC-ROC.
409 The three classifiers we tested were Random Forest, naïve Bayes, and Support Vector Machines
410 (Ben-Hur *et al.*, 2002; Denisko and Hoffman, 2018; Murty and Devi, 2011) . For Random
411 Forest, we tuned the number of trees and the maximum number of features for each tree. For
412 Support Vector Machines, RBF kernel was used and the C parameter was tuned. Random Forest
413 was used for further analysis since its performance was slightly better than the other two
414 classifiers (Supplementary Figure S2). We used 200 trees and the "auto" max_features for
415 Random Forest. The ratio of positives and negatives in training data was also tuned. The best
416 performing positive:negative ratio was 1:20 for Arabidopsis and 1:5 for rice. For testing, a
417 positive:negative ratio of 1:200 was used since it is close to the average ratio of causal and non-
418 causal genes in real QTLs.

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

421 **Feature importance analysis**

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

429 **Independent literature validation**

430 For validation, we applied models to an independent set of nine causal genes that were curated
431 from recent literature and not used for cross-validation. The models were trained by known
432 causal genes from the original list and negatives randomly selected from the rest of the genome
433 (Figure 1A). Model training was repeated 5000 times by resampling training negatives from the

genome. 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 the frequency of being predicted as a causal gene from the 5000 iterations.

The probability to correctly prioritize a causal gene of a single QTL at a certain threshold was calculated based on the nine causal genes. Based on the probability of finding a causal gene when applying the algorithm to a single QTL, we calculated the probability of correctly prioritizing at least *K* causal genes when applying the models to a total of *N* QTLs with equation (1). Where p is the probability to correctly prioritize a causal gene of a single QTL at a certain threshold.

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

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 from the original list was tested once. For each round, one curated causal gene was removed from training set. Then the model was trained and applied to rank the known causal gene and 200 flanking genes.

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.

Acknowledgements

This work was supported by the United States Department of Energy's Biological and Environmental Research Program [DE-SC0008769, DE-SC0018277]. We thank Dr. John Lloyd and Dr. Shin-Han Shiu for sharing the features and score of rice essential gene prediction. The authors declare no conflict of interest. We thank Kevin Radja for testing the source code and giving useful comments.

References

Araus, J.L., *et al.* (2018) Translating High-Throughput Phenotyping into Genetic Gain. *Trends Plant Sci*, 23(5),451-466.

Bargsten, J.W., *et al.* (2014) Prioritization of candidate genes in QTL regions based on associations between traits and biological processes. *BMC Plant Biol*, 14,330.

Baxter, I., *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(11),e1001193.

Ben-Hur, A., *et al.* (2002) Support vector clustering. *J Mach Learn Res*, 2,125-137.

Bentsink, L., *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(9),4264-4269.

Bergelson, J. and Roux, F. (2010) Towards identifying genes underlying ecologically relevant traits in Arabidopsis thaliana. *Nat Rev Genet*, 11(12),867-879.

Brachi, B., *et al.* (2010) Linkage and association mapping of Arabidopsis thaliana flowering time in nature. *PLoS Genet.*, 6(5),e1000940.

Buckler, E.S., *et al.* (2009) The Genetic Architecture of Maize Flowering Time. *Science*, 325(5941),714-718.

Cao, J., *et al.* (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat Genet*, 43(10),956-963.

Carlborg, O. and Haley, C.S. (2004) Epistasis: too often neglected in complex trait studies? *Nat Rev Genet*, 5(8),618-U614.

Chardon, F., *et al.* (2013) Leaf Fructose Content Is Controlled by the Vacuolar Transporter SWEET17 in Arabidopsis. *Current Biology*, 23(8),697-702.

Conde, A., *et al.* (2011) Membrane Transport, Sensing and Signaling in Plant Adaptation to Environmental Stress. *Plant Cell Physiol*, 52(9),1583-1602.

Daware, A.V., *et al.* (2017) Regional Association Analysis of MetaQTLs Delineates Candidate Grain Size Genes in Rice. *Front Plant Sci*, 8,807.

Denisko, D. and Hoffman, M.M. (2018) Classification and interaction in random forests. *Proc Natl Acad Sci U S A*, 115(8),1690-1692.

Deo, R.C., *et al.* (2014) Prioritizing causal disease genes using unbiased genomic features. *Genome Biol*, 15(12),534.

Dinka, S.J., *et al.* (2007) Predicting the size of the progeny mapping population required to positionally clone a gene. *Genetics*, 176(4),2035-2054.

Doidy, J., *et al.* (2012) Sugar transporters in plants and in their interactions with fungi. *Trends Plant Sci*, 17(7),413-422.

Fahlgren, N., *et al.* (2015) Lights, camera, action: high-throughput plant phenotyping is ready for a close-up. *Curr Opin Plant Biol*, 24,93-99.

Fan, S.X., *et al.* (2016) Characterization and fine mapping of osh15(t), a novel dwarf mutant gene in rice (*Oryza sativa* L.). *Genes & Genomics*, 38(9),849-856.

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

FAO. (2018) The State of Food Security and Nutrition in the World 2018.

Fukuoka, S., *et al.* (2014) Multiple functional polymorphisms in a single disease resistance gene in rice enhance durable resistance to blast. *Sci Rep*, 4,4550.

Gao, Y., *et al.* (2016) QTL analysis for chalkiness of rice and fine mapping of a candidate gene for qACE9. *Rice*, 9.

Gelfman, S., *et al.* (2017) Annotating pathogenic non-coding variants in genic regions. *Nat Commun*, 8(1),236.

Gotz, S., *et al.* (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res*, 36(10),3420-3435.

Grant, C.E., *et al.* (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7),1017-1018.

Guo, J., *et al.* (2015) Target Enrichment Improves Mapping of Complex Traits by Deep Sequencing. *G3-Genes Genomes Genetics*, 6(1),67-77.

He, H. Wageningen University; 2014. Environmental Regulation of Seed Performance.

Hindorff, L.A., *et al.* (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*, 106(23),9362-9367.

Hormozdiari, F., *et al.* (2015) Identification of causal genes for complex traits. *Bioinformatics*, 31(12),206-213.

Hu, K.M., *et al.* (2008) Isolation and manipulation of quantitative trait loci for disease resistance in rice using a candidate gene approach. *Mol Plant*, 1(5),786-793.

Huang, X., *et al.* (2012) Epistatic natural allelic variation reveals a function of AGAMOUS-LIKE6 in axillary bud formation in Arabidopsis. *Plant Cell*, 24(6),2364-2379.

Huang, Y.F., *et al.* (2017) Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat Genet*, 49(4),618-624.

Jagadeesh, K.A., *et al.* (2016) M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet*, 48(12),1581-1586.

Jones, P., *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9),1236-1240.

Kircher, M., *et al.* (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*, 46(3),310-315.

Kuhn, M. and Johnson, K. Applied Predictive Modeling. New York : Springer; 2013.

Lee, I., *et al.* (2010) Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana*. *Nat Biotechnol*, 28(2),149-U114.

Lee, I., *et al.* (2011) Genetic dissection of the biotic stress response using a genome-scale gene network for rice. *Proc Natl Acad Sci U S A*, 108(45),18548-18553.

Leinonen, P.H., *et al.* (2013) Genetic basis of local adaptation and flowering time variation in *Arabidopsis lyrata*. *Mol Ecol*, 22(3),709-723.

Liu, L., *et al.* (2017) Fine Mapping and Candidate Gene Analysis of the Tiller Suppression Gene ts1 in Rice. *Plos One*, 12(1).

Liu, X.Q., *et al.* (2005) Genetic and physical mapping of Pi36(t), a novel rice blast resistance gene located on rice chromosome 8. *Mol Genet Genomics*, 274(4),394-401.

Lloyd, J.P., *et al.* (2015) Characteristics of Plant Essential Genes Allow for within- and between-Species Prediction of Lethal Mutant Phenotypes. *Plant Cell*, 27(8),2133-2147.

Mackay, T.F.C. (2001) The genetic architecture of quantitative traits. *Annu Rev Genet*, 35,303-339.

Mackay, T.F.C. (2014) Epistasis and Quantitative Traits: Using Model Organisms to Study Gene-Gene Interactions. *Nat Rev Genet*, 15(1),22-33.

Martin, A. and Orgogozo, V. (2013) The Loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. *Evolution*, 67(5),1235-1250.

Mordelet, F. and Vert, J.P. (2011) ProDiGe: Prioritization Of Disease Genes with multitask machine learning from positive and unlabeled examples. *BMC Bioinform*, 12,389.

Motte, H., *et al.* (2014) Combining linkage and association mapping identifies RECEPTOR-LIKE PROTEIN KINASE1 as an essential Arabidopsis shoot regeneration gene. *Proc Natl Acad Sci U S A*, 111(22),8305-8310.

Murty, M.N. and Devi, V.S. Pattern Recognition - An Algorithmic Approach. Springer-Verlag London; 2011.

Ng, P.C. and Henikoff, S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res*, 31(13),3812-3814.

Nuzhdin, S.V.*, et al.* (1999) The genetic architecture of selection response. Inferences from fine-scale mapping of bristle number quantitative trait loci in *Drosophila melanogaster*. *Genetics*, 153(3),1317-1331.

Oikawa, T.*, et al.* (2015) The Birth of a Black Rice Gene and Its Local Spread by Introgression. *Plant Cell*, 27(9),2401-2414.

Ossowski, S.*, et al.* (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res*, 18(12),2024-2033.

Otto, S.P. and Jones, C.D. (2000) Detecting the undetected: Estimating the total number of loci underlying a quantitative trait. *Genetics*, 156(4),2093-2107.

Panchy, N.*, et al.* (2016) Evolution of Gene Duplication in Plants. *Plant Physiol*, 171(4),2294-2316.

Pedregosa, F.*, et al.* (2011) Scikit-learn: Machine Learning in Python. *J Mach Learn Res*, 12,2825-2830.

Perez-Iratxeta, C.*, et al.* (2002) Association of genes to genetically inherited diseases using data mining. *Nat Genet*, 31,316.

Pollard, K.S.*, et al.* (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*, 20(1),110-121.

Price, A.L.*, et al.* (2010) New approaches to population stratification in genome-wide association studies. *Nat Rev Genet*, 11(7),459-463.

Promislow, D.E.L. (2004) Protein networks, pleiotropy and the evolution of senescence. *Proceedings of the Royal Society B: Biological Sciences*, 271(1545),1225-1234.

Qi, J.*, et al.* (2008) Mutation of the Rice Narrow *leaf1* Gene, Which Encodes a Novel Protein, Affects Vein Patterning and Polar Auxin Transport. *Plant Physiol*, 147(4),1947-1959.

Ritchie, G.R.*, et al.* (2014) Functional annotation of noncoding sequence variants. *Nat Methods*, 11(3),294-296.

Schaefer, R.*, et al.* (2018) Integrating co-expression networks with GWAS to prioritize causal genes in maize. *Plant Cell*.

Schlapfer, P., *et al.* (2017) Genome-Wide Prediction of Metabolic Enzymes, Pathways, and Gene Clusters in Plants. *Plant Physiol*, 173(4),2041-2059.

Sharma, T.R., *et al.* (2005) High-resolution mapping, cloning and molecular characterization of the Pi-k(h) supercript stop gene of rice, which confers resistance to Magnaporthe grisea. *Mol Genet Genomics*, 274(6),569-578.

Smith, S.D. (2016) Pleiotropy and the evolution of floral integration. *New Phytol*, 209(1),80-85.

Swinnen, G., *et al.* (2016) Lessons from Domestication: Targeting Cis-Regulatory Elements for Crop Improvement. *Trends Plant Sci*, 21(6),506-515.

Takahashi, R., *et al.* (2012) The role of heavy-metal ATPases, HMAs, in zinc and cadmium transport in rice. *Plant Signal Behav*, 7(12),1605-1607.

Tomczak, A., *et al.* (2018) Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations. *Sci Rep*, 8(1),5115.

Touw, W.G., *et al.* (2013) Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? *Brief Bioinform*, 14(3),315-326.

Tuinstra, M.R., *et al.* (1997) Heterogeneous inbred family (HIF) analysis: a method for developing near-isogenic lines that differ at quantitative trait loci. *Theor Appl Genet*, 95(5),1005-1011.

Turner, T.L., *et al.* (2010) Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. *Nat Genet*, 42(3),260-263.

Van de Velde, J., *et al.* (2014) Inference of Transcriptional Networks in Arabidopsis through Conserved Noncoding Sequence Analysis. *Plant Cell*, 26(7),2729-2745.

Weirauch, M.T., *et al.* (2014) Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell*, 158(6),1431-1443.

Wellenreuther, M. and Hansson, B. (2016) Detecting Polygenic Evolution: Problems, Pitfalls, and Promises. *Trends Genet*, 32(3),155-164.

Xu, S. (2003) Theoretical basis of the Beavis effect. *Genetics*, 165(4),2259-2268.

Yang, H.Y., *et al.* (2004) High-resolution genetic mapping at the Bph15 locus for brown planthopper resistance in rice (*Oryza sativa* L.). *Theor Appl Genet*, 110(1),182-191.

Yin, Z.G., *et al.* (2017) Soybean plant height QTL mapping and meta-analysis for mining candidate genes. *Plant Breed*, 136(5),688-698.

Yuan, W., *et al.* (2016) Extreme QTL mapping of germination speed in Arabidopsis thaliana. *Mol Ecol*, 25(17),4177-4196.

Zeng, Y.X., *et al.* (2013) Fine mapping and candidate gene analysis of LM3, a novel lesion mimic gene in rice. *Biologia*, 68(1),82-90.

Figure and Table Legends

Figure 1 Model training and optimization based on cross-validation. (A) model training and cross validation framework. We randomly selected negatives from genome and added iterations 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).

Figure 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. (B) The enrichment or depletion of the top 5 features in Arabidopsis and rice models. The fold enrichment indicates the ratio of averages of causal genes to genome background. Ni, not included as it is not a top 5 feature in that species.

Figure 3 Model performance at different thresholds (A) Percentage of correctly prioritized causal genes of a single QTL at different rank thresholds. (B-C) The probability of correctly prioritizing at least X% of causal genes when analyzing multiple QTLs simultaneously.

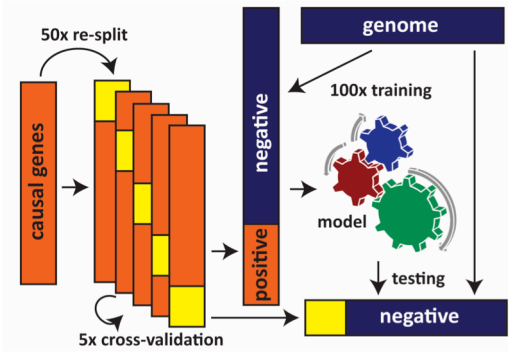
Figure 4 (A) Trait categories of known causal genes from the training set. (B) The rank percentile of causal genes of different trait categories. Each causal gene and 200 neighboring genes were used as testing set once. All other known causal genes were used as training set. Each dot indicates a known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

Table 1 Causal gene features and their frequency and value in causal genes and the genome background

Table 2 Evaluating model performance based on the rank of an independent set of known causal genes.

Figures and Tables

A



B

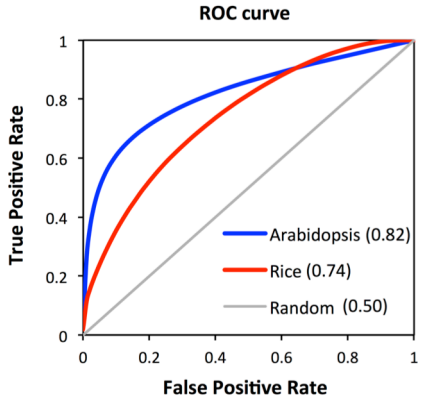


Figure 1 Model training and optimization based on cross-validation. (A) model training and cross validation framework. We randomly selected negatives from genome and added iterations to maximize the combinations of training and testing data. (B) The ROC curve of Arabidopsis and rice models after parameter 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).

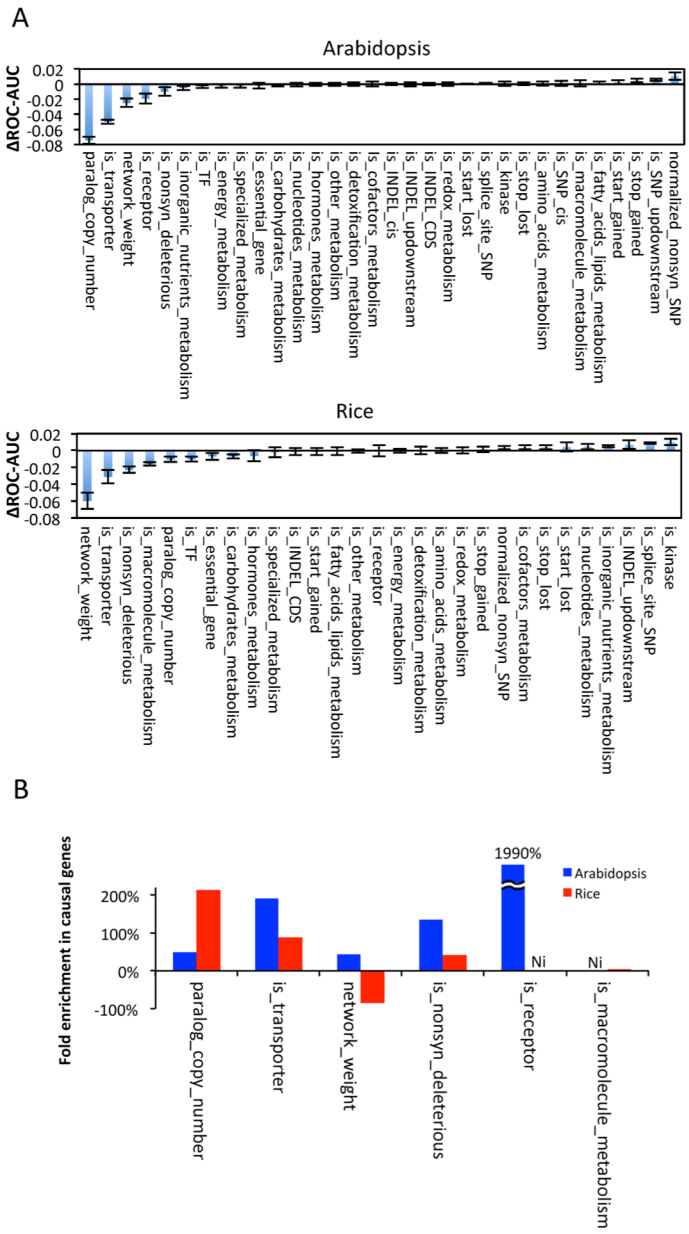


Figure 2 Important features of causal genes and their enrichment or depletion relative to the genome background (A) Feature importance as indicated by the change of AUC-ROC (ΔAUC-ROC) when excluding each feature. The ΔAUC-ROC indicates the average value of all iterations. Error bars indicate standard deviation. (B) The enrichment or depletion of the top 5 features in Arabidopsis and rice models. The fold enrichment indicates the ratio of averages of causal genes to genome background. Ni, not included as it is not a top 5 feature in that species.

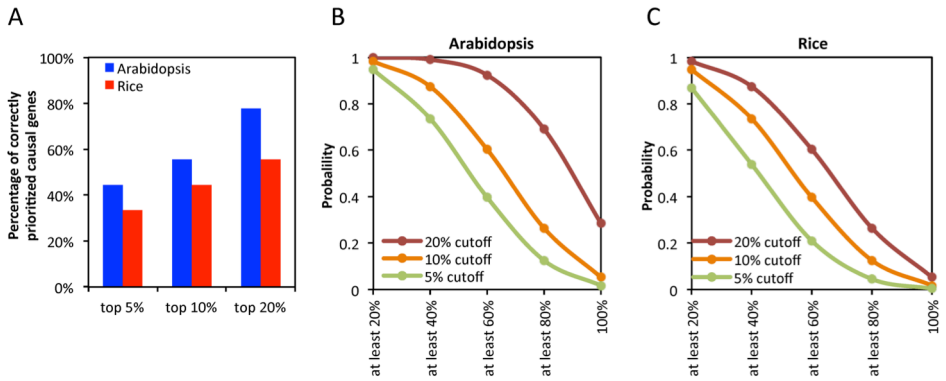


Figure 3 Model performance at different thresholds (A) Percentage of correctly prioritized causal genes of a single QTL at different rank thresholds. (B-C) The probability of correctly prioritizing at least X% of causal genes when analyzing multiple QTLs simultaneously.

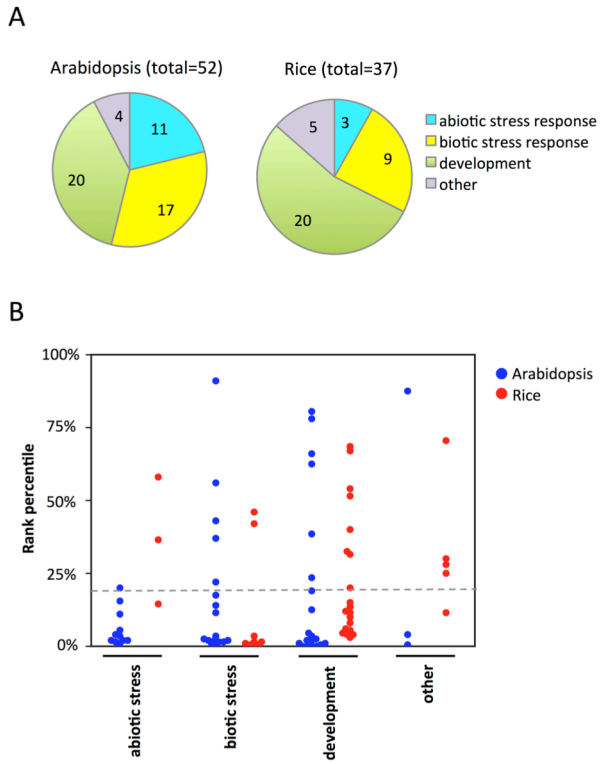


Figure 4. (A) Trait categories of known causal genes from the training set. (B) The rank percentile of causal genes of different trait categories. Each causal gene and 200 neighboring genes were used as testing set once. All other known causal genes were used as training set. Each dot indicates a known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

Table 1 Causal gene features and their frequency and value in causal genes and the genome background

Category	Feature name[a]	Arabidopsis			Rice		
		Causal gene	Genome gene	P value[b]	Causal gene	Genome gene	P value[b]
Enzyme class	is_carbohydrates_metabolism	1.9%	5.0%	5.2E-01	8.1%	5.2%	4.4E-01
Enzyme class	is_nucleotides_metabolism	0.0%	2.2%	6.3E-01	2.7%	2.5%	6.1E-01
Enzyme class	is_energy_metabolism	3.8%	1.9%	2.6E-01	0.0%	1.8%	1.0E+00
Enzyme class	is_fatty_acids_lipids_metabolism	3.8%	2.8%	6.6E-01	0.0%	3.2%	6.3E-01
Enzyme class	is_specialized_metabolism	5.8%	3.7%	4.4E-01	0.0%	4.4%	4.1E-01
Enzyme class	is_cofactors_metabolism	0.0%	2.1%	6.3E-01	2.7%	2.4%	6.0E-01
Enzyme class	is_other_metabolism	0.0%	1.3%	1.0E+00	2.7%	1.3%	3.8E-01
Enzyme class	is_hormones_metabolism	1.9%	1.2%	4.7E-01	8.1%	1.2%	1.1E-02
Enzyme class	is_inorganic_nutrients_metabolism	9.6%	1.4%	9.0E-04	2.7%	1.9%	5.2E-01
Enzyme class	is_detoxification_metabolism	0.0%	0.9%	1.0E+00	0.0%	1.1%	1.0E+00
Enzyme class	is_redox_metabolism	0.0%	0.7%	1.0E+00	0.0%	1.0%	1.0E+00
Enzyme class	is_amino_acids_metabolism	1.9%	1.7%	5.8E-01	0.0%	1.9%	1.0E+00
Enzyme class	is_macromolecule_metabolism[c]	13.5%	12.9%	8.4E-01	37.8%	33.9%	6.1E-01
GO	is_TF	7.7%	6.3%	5.7E-01	13.5%	4.8%	3.1E-02
GO	is_receptor	13.5%	0.6%	5.5E-08	0.0%	0.6%	1.0E+00
GO	is_kinase	13.5%	10.9%	5.1E-01	18.9%	13.3%	3.3E-01
GO	is_transporter	23.1%	7.9%	6.4E-04	13.5%	7.1%	1.9E-01
Co-function network	network_weight[d]	182.0	126.4	5.9E-02	11.8	76.5	1.2E-21
Polymorphism	is_start_gained	9.6%	4.1%	6.0E-02	10.8%	22.0%	1.1E-01
Polymorphism	is_start_lost	0.0%	0.3%	1.0E+00	8.1%	3.9%	1.8E-01
Polymorphism	is_stop_gained	5.8%	1.8%	6.4E-02	51.4%	44.1%	4.1E-01
Polymorphism	is_stop_lost	3.8%	0.5%	2.6E-02	5.4%	5.1%	7.1E-01
Polymorphism	is_SNP_splice_site	0.0%	0.8%	1.0E+00	24.3%	23.5%	8.5E-01
Polymorphism	is_SNP_updownstream[e]	88.5%	90.9%	4.7E-01	NA	NA	NA
Polymorphism	is_INDEL_CDS	0.0%	2.1%	6.3E-01	18.9%	16.9%	6.6E-01
Polymorphism	is_INDEL_updownstream[e]	21.2%	29.5%	2.2E-01	32.4%	30.2%	7.2E-01
Polymorphism	is_INDEL_cis[f]	0.0%	0.9%	1.0E+00	NA	NA	NA
Polymorphism	is_SNP_cis[f]	1.9%	7.3%	1.8E-01	NA	NA	NA
Polymorphism	is_nonsyn_deleterious[g]	30.8%	13.1%	7.4E-04	75.7%	53.2%	7.5E-03
Polymorphism	normalized_nonsyn_SNP[h]	0.006	0.005	4.7E-01	0.013	0.015	1.7E-01
Miscellaneous	is_essential_gene[i]	1.9%	10.0%	6.0E-02	0.0%	8.9%	7.4E-02
Evolution related	paralog_copy_number	45.6	30.4	7.8E-02	116.4	37.0	1.1E-02

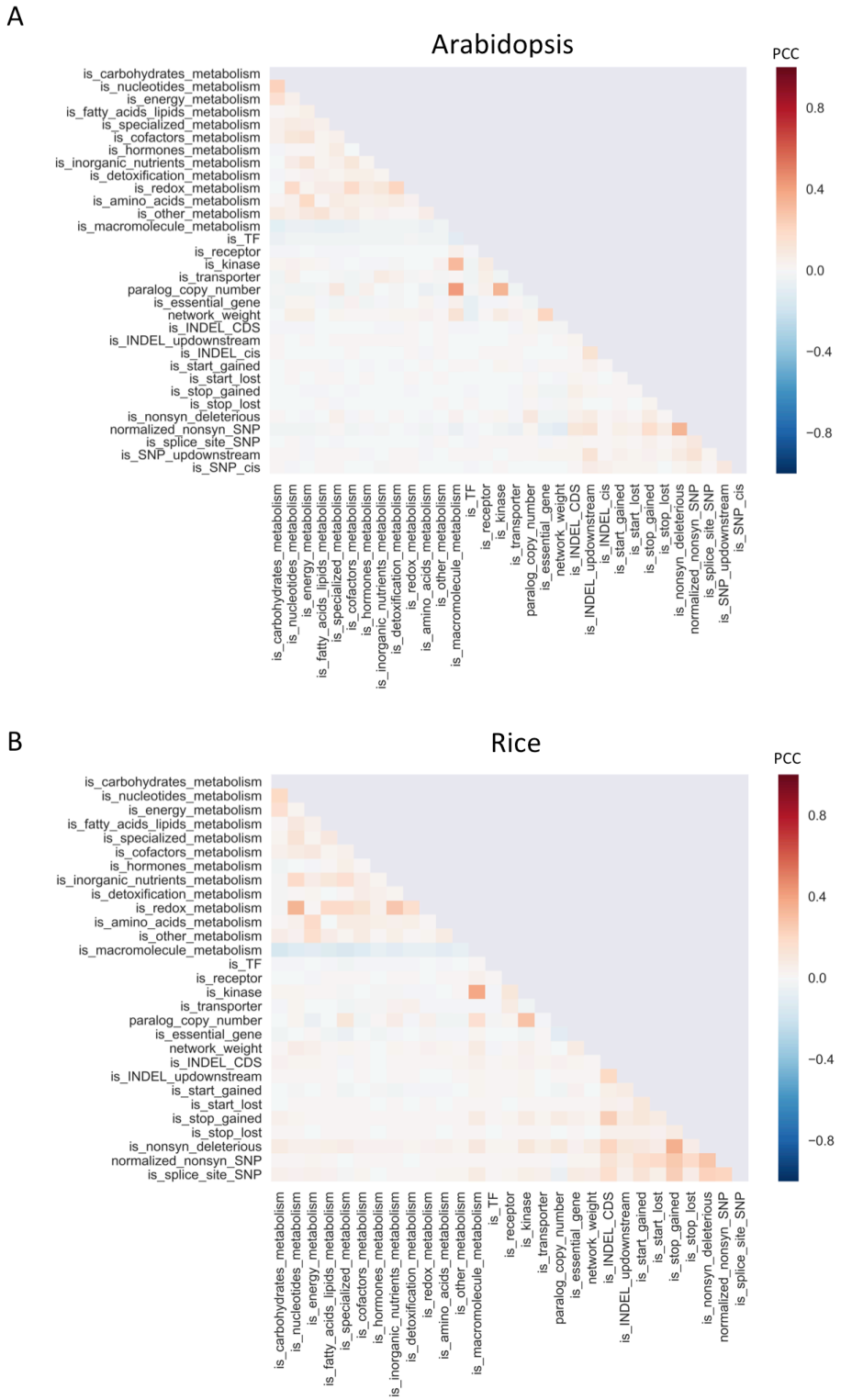
[a] The features with a name that starts with "is_" are binary variables.
[b] Fisher's exact test was used for all binary features. Welch's t-test was used for all continuous variables.
[c] This category contains genes that were annotated as enzyme by GO but not present in PMN. A majority of them is involved in macromolecule metabolism.
[d] "network_weight" is the sum weight of all edges connected to a gene based on AraNet and RiceNet (Lee et al., 2010; Lee et al., 2011).
[e] "is_SNP_updownstream" and "is_INDEL_updownstream" indicate there were polymorphisms present within 1kb upstream or downstream of the gene. Since all rice genes had at least one SNP present in upstream or downstream, "is_SNP_updownstream" was not used as a feature in the rice model.
[f] "is_SNP_cis" and "is_INDEL_cis" indicate whether polymorphisms are present in the cis-element of the gene (Weirauch, et al., 2014) . The cis-element data was only available for Arabidopsis and therefore not used in the rice model.
[g] "is_nonsyn_deleterious" indicates the presence of non-synonymous SNPs in highly conserved residues, which are likely to affect protein function (Ng et al., 2003)
[h] The number of non-synonymous SNPs was normalized to the length of the protein.
[i] Essential gene prediction was based on function annotation, duplication, expression levels and patterns, rate of evolution, cross-species conservation and other information (Lloyd et al., 2015)

Table 2 Evaluating model performance based on the rank of an independent set of known causal genes.

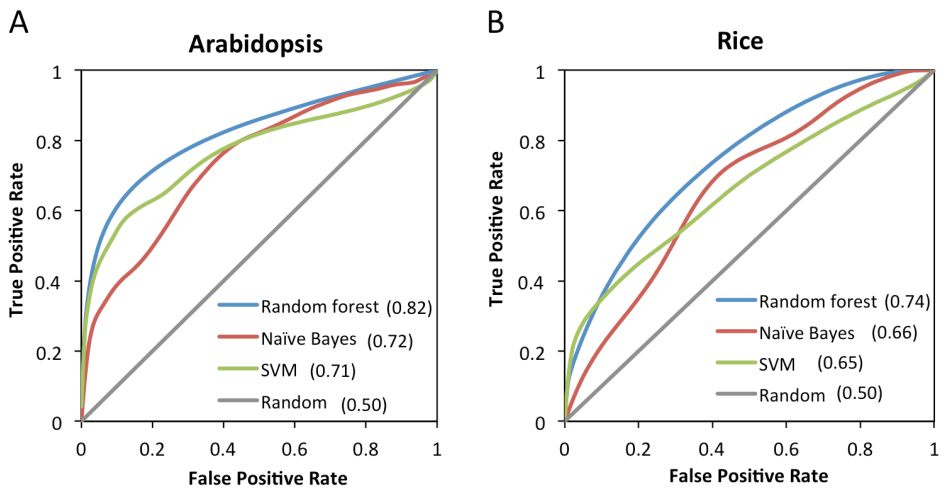
QTL trait	Gene name	Rank in a QTL[a]	Causality evidence[b]	Reference
Arabidopsis				
Seed size	*AHK3*	1/366 (<0.1%)	Indirect	Guo, 2016
Seed size	*AHK2*	2/293 (<0.1%)	Indirect	Guo, 2016
Stem branching	*AGL6*	10/457 (1%)	Direct	Huang, 2012
Stem branching	*MAX3*	13/457 (2%)	Indirect	Huang, 2013
Stem branching	*SCL27*	42/457 (8%)	Indirect	Huang, 2013
Seed germination	*FBA2*	17/111 (14%)	Indirect	Yuan,2016
Seed germination	*AZF2*	24/111 (20%)	Indirect	Yuan,2016
Fructose content in leaves	*Sweet17*	52/186 (27%)	Direct	Fabien, 2013
Shoot regeneration	*RPK1*	234/554 (42%)	Direct	Motte, 2014
Rice				
Blast resistance	Pi35/Pish	3/203 (0.5%)	Direct	Fukuoka, 2014
Blast resistance	Pi-k[h]	20/444 (4%)	Indirect	Sharma, 2005
Plant height	osh15	7/117 (5%)	Indirect	Fan, 2016
Leaf blade width	nal1	14/210 (6%)	Direct	Qi,2008
Grain chalkiness	*qACE9*	23/120 (18%)	Direct	Gao,2016
Tiller number	ts1	48/175 (26%)	Direct	Liu,2017
Blast and bacterial blight lesion	UROD	12/38 (30%)	Indirect	Zeng, 2013
Bacterial blight resistance	*WRKY13*	161/460 (34%)	Indirect	Hu,2008
Grain color	kala4	29/62 (45%)	Direct	Oikawa, 2015

[a] The numerator is the rank of the causal gene. The denominator is the total number of genes in the QTL region. The percentage in parentheses indicates rank percentile.
[b] Direct causality evidence includes functional complementation and haplotype analyses showing the association between trait variation with polymorphisms of the gene or with the expression of the gene. Indirect evidence indicates mutant phenotype analysis.

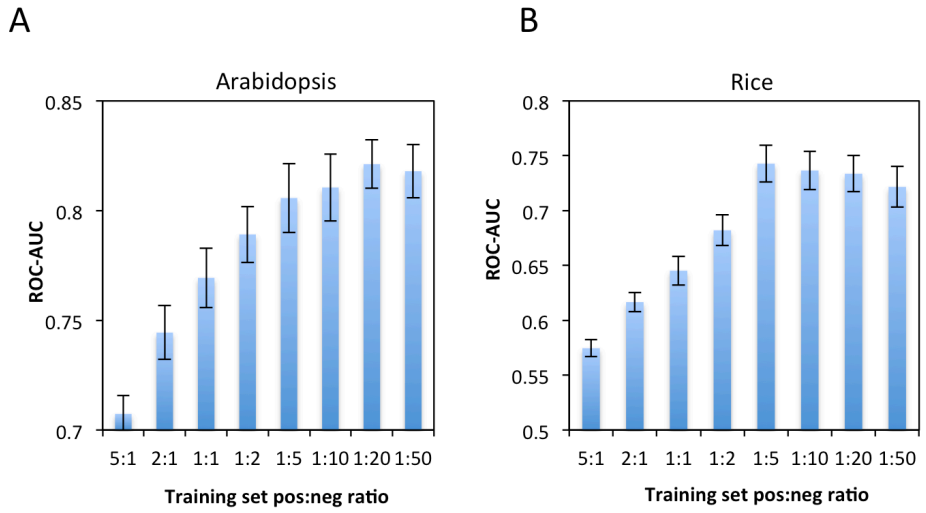
Supplemental Information



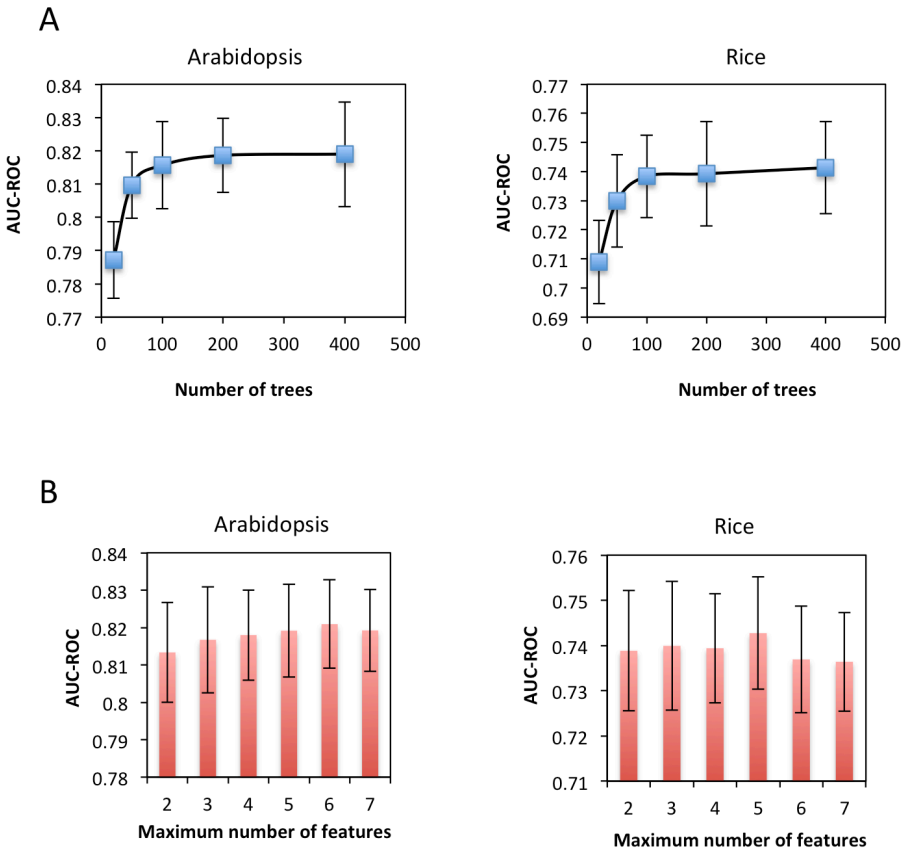
Supplemental Figure S1 Correlations among features used for the Arabidopsis and rice models. Related to Table1. Pearson's correlation coefficients (PCC) were used for heatmap (A) Arabidopsis (B) rice



Supplemental Figure S2 Comparing the Receiver Operating Characteristic curves (ROC) of Random forest, Naïve Bayes, and Support Vector Machine (SVM). Related to Figure 1. The ROC curves were based on cross-validation. Numbers in parentheses indicate Area Under Curve (AUC). The grey diagonal line indicates the expected training performance of a model based on random guessing. (A) Arabidopsis (B) rice



Supplemental Figure S3 The ratio of positive: negative in training set. Related to Figure 1. Area Under Curve of ROC (ROC-AUC) was based on cross-validation. Error bars indicate standard deviation (A) Arabidopsis (B) rice



Supplemental Figure S4 Testing two Random Forest parameters. Related to Figure 1. Area Under Curve of ROC was based on cross-validation. Error bars indicate standard deviation (A) Number of trees in the forest (B) The number of features to consider when looking for the best split.

Supplemental Table S1 Curated Arabidopsis causal genes used for model training and cross-validation.

Gene Name	Gene ID	Gene Function	Trait	Trait category	Reference PMID
APR2	AT1G62180	enzyme	shoot sulfate content	abiotic stress response	17589509
ACD6	AT4G14400	membrane protein	leaf initiation and necrosis	development	20520716, 20336072
AOP3	AT4G03050	enzyme	glucosinolate accumulation	biotic stress response	11251105, 19737743, 21857804
AT5G41740	AT5G41740	R-protein with leucine-rich repeats	necrosis	other	17803357
AT5G41750	AT5G41750	R-protein with leucine-rich repeats	necrosis	other	17803357

AtHKT1	AT4G10310	transporter ion	salt tolerance	abiotic stress response	17140289, 21085628
Brevis radix (BRX)	AT1G31880	TF	root length	development	15031265
Cryptochrome 2 (CRY2) EDI allele	AT1G04400	chromophore protein	flowering time and other pleiotropic effects	development	11726930, 15280248, 15248119, 14605225
CYCD5;1	AT4G37630	cyclin-dependent kinase	cell division	development	22392991
CYP81F2	AT5G57220	enzyme (P450)	glucosinolate metabolism	biotic stress response	19293369
DOG1 (DELAY OF GERMINATION 1)	AT5G45830	unknown	seed germination	development	17065317, 22231484, 20336072
EARLY FLOWERING 3(ELF3)	AT2G25930	circadian oscillator	flowering	development	20838594, 21857804, 20713464, 23129635
Epithiospecifier Modifier1 (ESM1)	AT3G14210	enzyme associated	glucosinolate	biotic stress response	16679459
Epithiospecifier protein (ESP)	AT1G54040	enzyme	glucosinolate	biotic stress response	11752388
ERECTA	AT2G26330	receptor (RTK-LRR)	plant and leaf architecture, transpiration	development	16007076, 20374533, 21368205
FLC (Flowering Locus C)	AT5G10140	TF (MADS)	flowering	development	22865739
FLM (MAF1)	AT1G77080	TF (MADS)	flowering	development	15695584
Flowering locus T (FT)	AT1G65480	RAF kinase inhibitor	flowering	development	17158798
FPN2	AT5G03570	transporter	nickel tolerance	abiotic stress response	19861554
FRD3 (FERRIC REDUCTASE DEFECTIVE3)	AT3G08040	transporter	iron transport	abiotic stress response	23236296
Frigida (FRI)	AT4G00650	nuclear regulatory protein	flowering time	development	11030654, 12140238, 12805638, 15908596
Frigida like 1 (FRL1)	AT5G16320	nuclear regulatory protein	flowering time	development	17056759
Frigida like 2 (FRL2)	AT1G31814	nuclear regulatory protein	flowering time	development	17056759
GIBBERELLIC ACID	AT4G02780	enzyme	plant morphology (flowers)	development	22510148

REQUIRING 1 (GA1)					
heavy metal atpase3 (HMA3)	AT4G30120	transporter	cadmium accumulation	abiotic stress response	22969436
heavy metal atpase4 (HMA4)	AT2G19110	transporter	cadmium accumulation	abiotic stress response	17434989, 18425111
heavy metal atpase5 (HMA5)	AT1G63440	transporter	copper accumulation	abiotic stress response	18701674
HUA2	AT5G23150	signalling RPR domain protein	flowering time, shoot morphology	development	17764945
KCS18	AT4G34520	enzyme	oil composition	other	23145136
MADS AFFECTING FLOWERING 2 (MAF2)	AT5G65050	TF (MADS)	flowering	development	20551443
MAM1	AT5G23010	enzyme	glucosinolate	biotic stress response	11706188, 19737743, 21857804, 23042895
metal tolerance protein1	AT2G46800	transporter	zinc concentration	abiotic stress response	15255871, 20419142
Molybdenum transporter1 (MOT1)	AT2G25680	transporter	molybdenum content	abiotic stress response	18454190
mucilage-modified 2 (mum2)	AT5G63800	enzyme	mucilage formation	other	18165330
phytochrome A (PHYA)	AT1G09570	chromophore protein	light sensitivity	development	11726931
phytochrome B (PHYB)	AT2G18790	chromophore protein	light sensitivity	development	18287016
phytochrome C (PHYC)	AT5G35840	chromophore protein	light sensitivity	development	16732287
phytochrome D (PHYD)	AT4G16250	chromophore protein	light sensitivity, increased petiole length, reduced cotyledon area and anthocyanin accumulation in seedling stems	development	9286109
RAS1	AT1G09950	microProtein	salt tolerance	abiotic stress response	20212128
resistant to methyl viologen 1 (RMV1)	AT5G05630	transporter	polyamine uptake	abiotic stress response	22492932
RLM1	AT1G64070	R-protein with leucine-rich repeats	pathogen resistance	biotic stress response	16623885
RLM3	AT4G16990	R-protein with leucine-rich repeats	pathogen resistance	biotic stress response	18397376
RPM1	AT3G07040	R-protein with leucine-rich repeats	pathogen resistance	biotic stress response	16623885

RPP13	AT3G46530	R-protein with leucine-rich repeats	pathogen resistance	biotic stress response	15082565
RPP4	AT4G16860	R-protein with leucine-rich repeats	pathogen resistance	biotic stress response	11846877
RPP5	AT4G16950	R-protein with leucine-rich repeats	pathogen resistance	biotic stress response	9212464, 11846877, 20479233
RPP8	AT5G43470	R-protein with leucine-rich repeats	pathogen resistance	biotic stress response	9811794 , 20479233
RPS2	AT3G03600	R-protein with leucine-rich repeats	pathogen resistance	biotic stress response	9874813
RPS4	AT5G45250	R-protein with leucine-rich repeats	pathogen resistance	biotic stress response	11846877
RPS5	AT1G12220	R-protein with leucine-rich repeats	pathogen resistance	biotic stress response	9212464, 11846877, 20479233
RRS1	AT5G45260	R-protein with leucine-rich repeats	pathogen resistance	biotic stress response	19519800, 19686535
WRR4	AT1G56510	R-protein with leucine-rich repeats	pathogen resistance	biotic stress response	18624640

Supplemental Table S2 Curated rice causal genes used for model training and cross-validation.

Gene name	ID	Gene function	Trait	Trait category	Reference PMID
Alk / Starch Synthase II	LOC_Os06g12450	enzyme	grain cooking texture	other	12579422, 16027975, 20972439
BADH2	LOC_Os08g32870	enzyme	fragrance	other	17129318, 19706531
Ehd1	LOC_Os10g32600	TF	flowering time	development	15078816
EARLY FLOWERING 3/Hd17	LOC_Os06g05060	circadian oscillator	flowering time	development	22399582
Ghd7	LOC_Os07g15770	TF (CO-like)	flowering time, plant morphology (inflorescence)	development	18454147
Hd6a	LOC_Os03g55389	protein kinase CK2	flowering time	development	11416158
PROG1	LOC_Os07g05900	TF	plant architecture	development	18820699, 18820696, 23034647
qPE9-1	LOC_Os09g26999	keratin-associated protein	plant and inflorescence architecture	development	19546322
qSH1	LOC_Os01g62920	TF BEL1-type homeobox gene (presumptive transription factor)	seed shattering	development	16614172

se5	LOC_Os06g40080	chromophore protein	flowering time	development	10849355
Shattering1 - OsSh1	LOC_Os03g44710	TF (YABBY-like)	seed shattering	development	22581231, 21695282, 22158310
shattering4	LOC_Os04g57530	TF (homeobox)	seed shattering	development	16527928
SKC1 =OsHKT1	LOC_Os06g48810	transporter (HKT-type)	salt homeostasis and salt tolerance	abiotic stress response	16155566
heavy metal atpase3 (HMA3)	LOC_Os07g12900	transporter	metal tolerance	abiotic stress response	20823253
Nramp aluminum transporter1	LOC_Os02g03900	transporter	metal tolerance	abiotic stress response	21829395
Pi-ta	LOC_Os12g18360	R-protein with leucine-rich repeats	pathogen resistance	biotic stress response	11090207, 21695282, 22158310
Pi2 (Nbs4-Pi2)	LOC_Os06g17920	R-protein with leucine-rich repeats	pathogen resistance	biotic stress response	17073304
Pi9 (= Nbs2-Pi9)	LOC_Os06g05359	R-protein with leucine-rich repeats	pathogen resistance	biotic stress response	16387888
Piz-t	LOC_Os06g17900	R-protein with leucine-rich repeats	pathogen resistance	biotic stress response	17073304
Pi37	LOC_Os01g57310	R-protein with leucine-rich repeats	pathogen resistance	biotic stress response	17947408
Pid3	LOC_Os06g22460	R-protein with leucine-rich repeats	pathogen resistance	biotic stress response	21621742
Xa1	LOC_Os04g53160	R-protein with leucine-rich repeats	pathogen resistance	biotic stress response	9465073
Xa21	LOC_Os01g56470	R-protein with leucine-rich repeats	pathogen resistance	biotic stress response	8525370
Xa26	LOC_Os04g13640	R-protein with leucine-rich repeats	pathogen resistance	biotic stress response	14756760
OsGA20ox1	LOC_Os03g63970	enzyme	seedling vigor	development	22481119
GRAIN INCOMPLETE FILLING 1	LOC_Os04g33740	enzyme	grain weight	development	18820698
GS5	LOC_Os05g06660	enzyme	grain size	development	22019783
GW2	LOC_Os02g14720	enzyme E3 ubiquitin ligase	grain size	development	17417637
OsCKX2=Gn1a	LOC_Os01g10110	enzyme	grain yield	development	15976269
OsPPKL1/qGL3	LOC_Os03g44500	signal transduction	grain size	development	23236132
OsSPL14 / WFP	LOC_Os08g39890	TF	grain yield	development	20495565, 20495564
OsSPL16	LOC_Os08g41940	TF	grain size and shape	development	22729225
qSW5	LOC_Os05g09520	unknown	grain size	development	18604208, 20972439

Sd1 (=GA20ox-2)	LOC_Os01g66100	enzyme	plant stature, dwarfism	development	11961544
OsC1	LOC_Os06g10350	TF (R2R3-MYB)	plant coloration (loss of apiculus color)	other	15514070, 23034647, 20972439
Rc	LOC_Os07g11020	TF (bHLH)	seed color	other	17696613, 16399804, 20972439
Waxy /GBSS	LOC_Os06g04200	enzyme	amylose content (glutinous rice)	other	7742858, 9718725, 9747848, 16547098

Supplemental Table S3 Features used for the Arabidopsis model. Known causal genes were labeled as 1 in the 'class' column. Other genes were labeled as 0.

See .csv files

Supplemental Table S4 Features used for the rice model. Known causal genes were labeled as 1 in the 'class' column. Other genes were labeled as 0.

See .csv files

The source code for model training and predictions are available at https://github.com/carnegie/QTG_Finder