Title: Prioritizing Candidate eQTL Causal Genes in Arabidopsis using Random Forests

Author names and affiliations

Margi Hartanto[1], Asif Ahmed Sami[1], Dick de Ridder[1], Harm Nijveen[1]

[1.] Bioinformatics Group, Wageningen University and Research, The Netherlands

Corresponding authors:

Margi Hartanto, Droevendaalsesteeg 1 6708PB Wageningen, phone number: +31630634476, email: margi.hartanto@wur.nl;

Harm Nijveen, Droevendaalsesteeg 1 6708PB Wageningen, phone number: +31317484706 , email: harm.nijveen@wur.nl

Running title: Candidate eQTL Causal Genes Prioritization

Abstract

Expression quantitative trait locus (eQTL) mapping has been widely used to study the genetic regulation of gene expression in *Arabidopsis thaliana*. As a result, a large amount of eQTL data has been generated for this model plant; however, only a few causal eQTL genes have been identified, and experimental validation is costly and laborious. A prioritization method could help speed up the identification of causal eQTL genes. This study extends the machine-learning-based QTG-Finder2 method for prioritizing candidate causal genes in phenotype QTLs to be used for eQTLs by adding gene structure, protein interaction, and gene expression. Independent validation shows that the new algorithm can prioritize sixteen out of twenty-five potential eQTL

causal genes within the 20% rank percentile. Several new features are important in prioritizing causal eQTL genes, including the number of protein-protein interactions, unique domains, and introns. Overall, this study provides a foundation for developing computational methods to prioritize candidate eQTL causal genes. The prediction of all genes is available in the AraQTL workbench (https://www.bioinformatics.nl/AraQTL/) to support the identification of gene expression regulators in Arabidopsis.

INTRODUCTION

One of the main objectives of genetic research is to link traits to genotypic variation. However, the path from genetics to observable traits is not straightforward; instead, it goes through a network of interconnecting intermediate phenotypes, such as gene expression, protein levels, and metabolite levels (Civelek and Lusis 2013). Studying the effect of the genetic perturbation on these intermediate phenotypes could improve our understanding of how a trait is regulated. Following recent advances in omics technology, the effect of multiple genetic perturbations can now be studied in a single experiment using linkage mapping or association studies. One example is genetical genomics, where variation in transcript levels is statistically associated with genetic variation in a population (Jansen and Nap 2001) to find so-called expression quantitative trait loci (eQTLs).

A mapped eQTL can be categorized as *cis* or *trans* based on its location relative to the affected gene. *Cis*-eQTLs are mapped close to the gene and are assumed to arise due to sequence polymorphisms in or near the gene itself, for instance, in *cis*-regulatory elements (*e.g.*, the promoter). In contrast, *trans*-eQTLs are mapped far away from the target gene and emerge due to polymorphisms in *trans*-acting factors (*e.g.*, transcription factors) called expression quantitative trait genes or eQTGs (Rockman and Kruglyak 2006; Brem et al. 2002). However,

a *trans*-eQTL typically spans a large genomic region with hundreds of candidate eQTGs. Experimental fine mapping to narrow down the region (*e.g.*, in Eshed and Zamir 1995) is costly and laborious. As a result, only a few causal genes have been identified in the thousands of eQTLs that have been mapped for *Arabidopsis thaliana*, using different populations and experimental conditions (Keurentjes et al. 2007; West et al. 2007; Cubillos et al. 2012; Snoek et al. 2012; Lowry et al. 2013; Hartanto et al. 2020). As an *in silico* alternative, a prioritization method can help to limit the number of candidate eQTGs for further validation.

Several network-based methods have been used to find eQTGs (*e.g.*, in Keurentjes et al. 2007; Jimenez-Gomez et al. 2010; Hartanto et al. 2020). These methods primarily aim to find master regulator(s) at loci where *trans*-eQTLs for many genes are collocated, known as eQTL hotspots (Breitling et al. 2008). In general, these methods utilize a coexpression network built using genes having an eQTL on the hotspot (called *targets*) and genes located in the hotspot (called *candidate eQTGs*). Candidates are then usually prioritized based on a network centrality measure, such as degree or closeness centrality (Serin et al. 2016). Several candidate eQTGs have been identified in this way, for example, *GIGANTEA* (Keurentjes et al. 2007), *ELF3* (Jimenez-Gomez et al. 2010), *ICE1*, and *DEWAX* (Hartanto et al. 2020). This approach, unfortunately, only works for eQTL hotspots, not for regions that only have a small number of eQTLs. Another limitation is the sole reliance on coexpression data: given the complexity of gene expression regulation, the expression of the regulator is not necessarily correlated to that of its targets, particularly in eukaryotes (Marbach et al. 2012; Lelli et al. 2012). Therefore, additional data sources should be considered to capture possible interactions between the regulator and its target.

Previously, a machine-learning-based method, QTG-Finder, was developed to prioritize candidate genes for phenotype QTLs in Arabidopsis (Lin et al. 2019). This method used features derived from various gene properties, such as paralog copy number, gene ontology, and the

number of SNPs, to rank the candidate genes in the QTL interval. The model could recall 64% of Arabidopsis QTGs when the top 20% ranked genes were considered. Further development of this method led to QTG-Finder2, which used orthology information and allowed for gene prioritization in species with no or few known QTGs (Lin et al. 2020). We were curious about the capability of this algorithm to prioritize eQTGs, given that some QTGs are involved in gene expression regulation, for example, *ELF3* (Jimenez-Gomez et al. 2010), *ERECTA* (Terpstra et al. 2010), *FRI* (Lowry et al. 2013), *MAM1* (Jansen et al. 2009), and *AOP2* (Jansen et al. 2009).

We propose eQTG-Finder, an extended version of QTG-Finder2 for eQTG prioritization, and apply the new algorithm to prioritize eQTGs in Arabidopsis. eQTG-Finder contains twelve new features based on protein-protein interaction, gene structure, and expression variation. These features significantly improve model performance, which is underscored by a feature importance analysis. We demonstrate the efficacy of this algorithm in prioritizing eQTGs using an independent test set. Finally, we use the new model to predict all Arabidopsis genes and make these available in our Arabidopsis eQTL analysis platform AraQTL (https://www.bioinformatics.nl/AraQTL/)(Nijveen et al. 2017) to help identify gene expression regulators.

MATERIALS AND METHODS

QTG-Finder2 was developed for prioritizing causal phenotype QTL genes (QTG) in Arabidopsis (Lin et al. 2020). This algorithm consists of 5,000 Random Forest classifiers (Ho 1998) trained using known QTGs and Arabidopsis orthologs of QTGs from other species as positives and other genes as negatives. QTG-Finder2 prioritizes candidate genes based on features generated from polymorphism data, functional annotation, co-function networks, and paralog copy numbers. Our method extends QTG-Finder2 with new features, and we train the

resulting model using the same sets of positive and negative genes. We evaluate the performance in prioritizing candidate causal eQTL genes (eQTGs) in Arabidopsis.

New features

We generate and include twelve new features in addition to the ones already used by QTG-Finder2. These new features are based on protein-protein interactions, gene expression, and gene/protein structure.

1. Protein-protein interaction feature

   Genes can be associated with other genes, for instance, because the encoded proteins participate in the same pathway or are mentioned in the same publication. The number of such interactions a gene has could measure its propensity to be an eQTL causal gene. We generate a network-based feature using Arabidopsis protein-protein interaction (PPI) data from STRING-DB (Szklarczyk et al. 2019). The data were downloaded from the download page of STRING-DB version 11 (https://string-db.org/cgi/download). We only keep high-confident interactions by removing those with STRING scores below 700. We count the number of interactions of each Arabidopsis gene as a feature.

2. Gene expression features

   The consequence of genetic variation in causal genes might be detected as early as in gene expression variability. We, therefore, generate features based on gene expression variation. We use the standard deviation of expression levels across different tissues from CoNekT (http://www.evorepro.plant.tools/) (Julca et al. 2020). We also use the average and standard deviation of *Arabidopsis thaliana* Columbia ecotype expression data from different samples as features. These data were retrieved from the Athrna-database (http://ipf.sustc.edu.cn/pub/athrna/) (Zhang et al. 2020).

3. Structural features

The structure of causal genes and encoded proteins might differ from the other genes. Therefore, we generate structural features: the numbers of introns, splice variants, total protein domains, unique protein domains, and splice variants per gene. Data were retrieved from https://www.arabidopsis.org/ (accessed May 2021). The number of introns and splice variants are counted in TAIR10's BLAST datasets. The other two features are generated from all.domains.txt by counting each Arabidopsis gene's total number of domains and the number of unique domains.

Hyperparameter tuning

Model evaluation is based on QTG-Finder (Lin et al. 2019) and QTG-Finder2 (Lin et al. 2020). Similar to QTG-Finder2, we use known QTGs and Arabidopsis orthologs of QTGs found in other species as positives and other genes as negatives. We use hyperparameter tuning to determine the best parameter combination (the number of trees, minimal samples split, and maximum number of features) using grid search and assess the area under the curve (AUC) of the receiver characteristics operator (ROC) curve in an extended version of the 5-fold cross-validation framework. In this framework, the positives are randomly re-split into a training and validation set in a 4:1 ratio iteratively. Next, each set is combined with randomly selected negatives. The ratio of positives and negatives is an optimized hyperparameter. This splitting of positives is done 50 times, and for each positive set random selection of the negatives was conducted 50 times. This extensive procedure (2,500 evaluations) makes that positive co-occurs with all negative at least once with high probability. All machine-learning model training and testing in this study is performed using Python’s scikit-learn library version 1.0.2.

Selection of candidate eQTL genes and independent validation of model performance

A list of candidate eQTGs in Arabidopsis is manually selected from the literature. These genes are categorized as confirmed/strong-candidate, hypothetical, or hypothetical-ortholog. Genes that have been through experimental validation or have strong evidence as eQTG are categorized into the confirmed/strong-candidate group, for example, *GIGANTEA* (Keurentjes et al. 2007; Snoek et al. 2012). Some confirmed/strong-candidate eQTGs are used as positive in QTG-Finder2, and we remove these from the positive instances to be used as validation genes. Meanwhile, genes that were not experimentally validated but are predicted to play a role as eQTG through *in silico* analysis (*e.g.*, network analysis) are categorized as hypothetical, for example, *ICE1* and *DEWAX* (Hartanto et al. 2020). If a gene's ortholog is considered an eQTG in another species, it is categorized as hypothetical-ortholog; for example, *NF-YC4* is found as an eQTG in potatoes (van Muijen et al. 2016). In total, this yields twenty-five candidate eQTGs in Arabidopsis: six confirmed/strong-candidate, four hypothetical, and fifteen hypothetical-ortholog genes (Supplementary Table 1). We ensure that these candidates are not used for hyperparameter tuning or cross-validation.

Independent validation is performed using the best combination of parameters (Supplementary Table 1). We train 5,000 Random Forest classifiers using all positives but different sets of negatives, with a positive: negative ratio of 1:200 to approximate the ratio of causal and non-causal genes in real eQTLs. The models are then applied to each candidate eQTG and other genes located 1 Mbp around it. These genes are ranked based on the average probability of being causal genes over 5,000 models.

Feature importance analysis

Feature importance is determined using a leave-one-out analysis. Iteratively, each feature is removed from the dataset, and a model is trained using the reduced dataset. The AUC difference in the full model (with all features) and the reduced model is then calculated and used to indicate the feature importance. We use the previous cross-validation framework and the best parameters to measure the model performance in this analysis.

Data analyses and code availability

Pairwise Pearson correlation coefficients between features are calculated using the Pandas (version 1.3.5) DataFrame.corr method in Python. Pearson Wilcoxon Rank Sum Test tests differences in the median between positive and negative genes for the twelve new features. The test is conducted in R using the base 'wilcox.test' function. Gene ontology enrichment analysis for the top and bottom 5% predicted causal genes is performed using TopGO in R (Alexa et al. 2006) using the algorithm's default 'weight01' parameter, which is the mixture of 'elim' and 'weight' methods. The Python version used for the analyses is 3.8.12, and the R version is 4.0.2. The source code and data are available at https://git.wur.nl/harta003/eqtg-finder.

RESULTS

The QTG-Finder2 algorithm could rank phenotype QTL causal genes higher than other genes in a cross-validation setting (AUC = 0.81) and recall 80% independent curated causal genes when the top 20% of genes in the QTL are considered (Lin et al. 2020). In this study, we extend QTG-Finder2 with a set of new features and evaluate its performance in prioritizing expression QTGs.

New features improve causal gene prediction performance

To improve model performance and better tailor it fit for eQTG prioritization, we added twelve new features based on gene expression, structure, and protein-protein interactions in the QTG-Finder2 algorithm. Most new features only show a low to moderate correlation with the existing ones (Supplementary Figure 1), indicating that we add new information to the model. Figure 1 shows feature distributions for the causal genes as the positive class (55 known QTGs and 145 Arabidopsis ortholog of QTGs from other species) and the other genes in the genome as the negative class (n=26,970). For most features, the causal genes' median value is significantly different from that of the other genes in the genome (see Supplementary Table 3). The expression of causal genes is more variable than that of other genes. Moreover, causal genes tend to have more and varied protein domains. Causal genes also have slightly more introns than other genes. These differences between the causal genes and the other genes in the genome provide a first indication of potential discriminating features for the machine learning model. We assess the performance of the model with and without new features using a cross-validation framework.

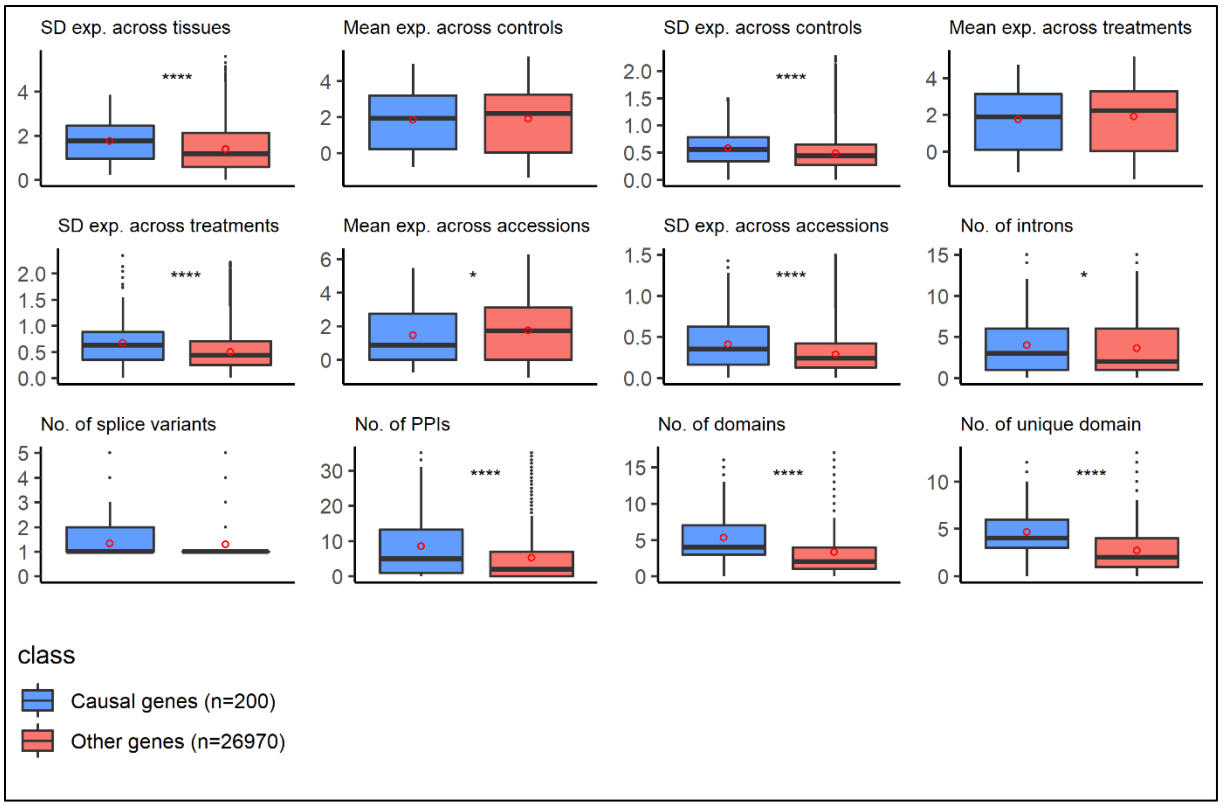


Figure 1. Distribution of twelve new features for known causal genes as the positive class (blue: n=200; 55 known QTGs and 145 orthologs of QTGs from other species) and the remaining genes in the genome as the negative class (red: n=26,970). Significance of differences in medians was assessed using the Wilcoxon Rank Sum Test (*: p <= 0.05; ****: p <= 0.0001). Red dots indicate means. SD = standard deviation. Exp. = gene expression. PPIs = protein-protein interactions.

To assess the contribution of new features to the model performance, we compare the area under the receiver-operating characteristic curve (AUC) between the original QTG-Finder2 with the extended model that we labeled eQTG-Finder, and for the extended model with the class labels permutated, as a control (Figure 2 left). The AUC was measured in an extended cross-validation setting over 2,500 different combinations of positive and negative gene sets. The results show that eQTG-Finder (AUC = 0.859 ± 0.008) performs better than QTG-Finder2 (AUC = 0.801 ±

0.01) and the control model (AUC = 0.502 ± 0.014). Adding new features thus allows the model to rank causal genes higher than the other genes. The next section analyzed model performance in prioritizing eQTG using selected candidate eQTGs.

To determine how the new features contribute to causal gene prediction, we calculate feature importance using a leave-one-out approach. Each feature is iteratively removed from the dataset, and the reduced model's performance is compared to that of the model containing all features. The drop in AUC indicates a feature's importance. A positive AUC drop means removing that feature decreases the model's predictive capability. The result shows that four of the most important features in the model are the new ones: the number of unique domains, the PPI count, the intron count, and the domain count. However, the large standard deviation for the domain count AUC drop indicates that the contribution of this feature is not consistent over different samples of positive and negative sets.

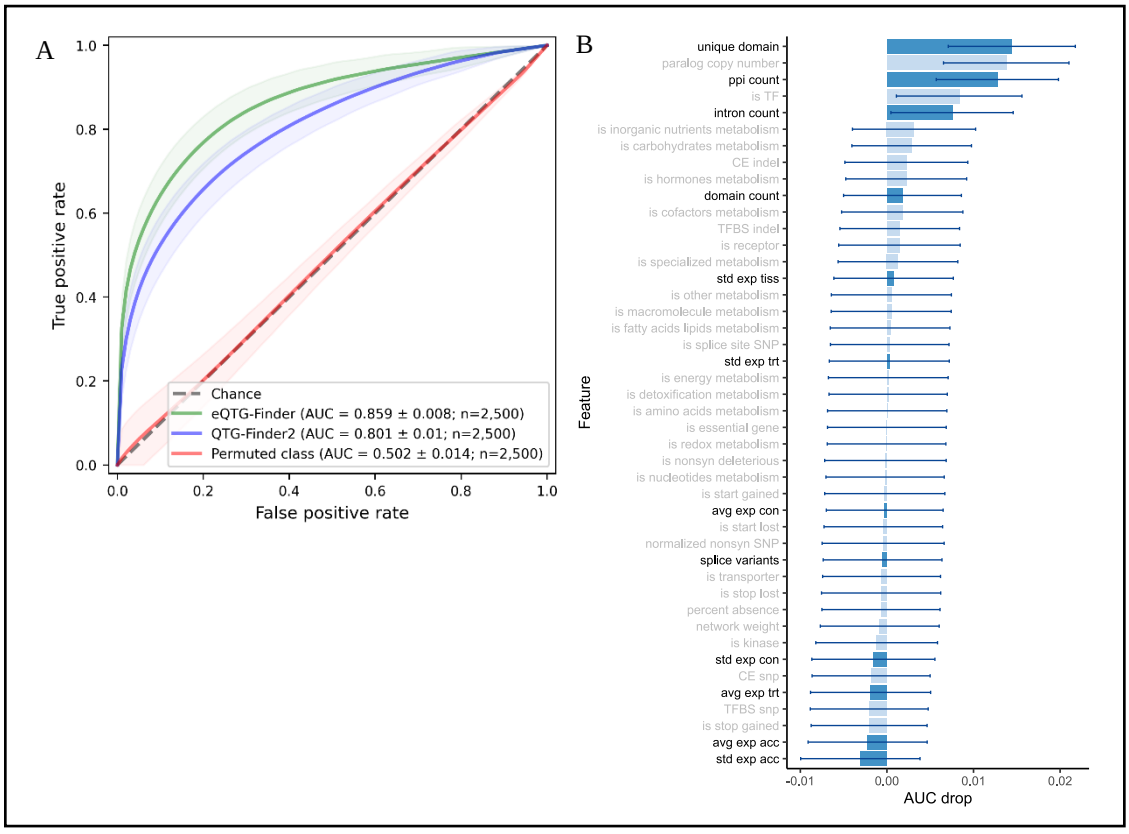


Figure 2. (A) Receiver operating characteristic (ROC) curves of the original QTG-Finder2 model (blue) and extended eQTG-Finder model (green), and eQTG-Finder trained with randomized class labels (red) as a control. Transparent areas indicate standard deviations over 2,500 repetitions. (B) Feature importance is measured using leave-one-out analysis. A positive AUC drop indicates that the removal of the feature reduces the model's predictive capability. Feature names in bold and with dark blue bars indicate new features. Error bars indicate standard deviations over 2,500 repetitions.

eQTG-Finder ranks most strong eQTG candidates better than QTG-Finder2

To evaluate eQTG prioritization performance, we again train the original QTG-Finder2 and the extended eQTG-Finder model and use them to rank selected potential eQTGs (Supplementary Table 1). Models are trained using all positives (known QTGs and Arabidopsis ortholog QTGs from other species). We repeated the training 5,000 times with different negative samples to select each negative gene at least once in training with high probability. These models rank each of the twenty-five potential eQTGs with their surrounding genes within a 2 Mbp window as a hypothetical eQTL region. These potential eQTGs are selected manually from the literature and grouped based on the evidence of being causal eQTL genes (see Methods for detail). Gene ranking is based on the average probability of a gene being causal, as predicted by the 5,000 models. We use the rank percentile to indicate the percentage of genes on the eQTL with higher ranks than the gene of interest (*i.e.,* a rank percentile of 0.1 indicates that 10% of genes in the eQTL region rank higher than the gene of interest). We predefine cutoffs of 5%, 10%, and 20%, in each of which we compare recall between QTG-Finder2 and eQTG-Finder. These recalls for different cutoffs can be used by researchers to decide the proportion of top prioritized genes for further experimental validation.

The QTG-Finder2 model recalls 16%, 28%, and 52% of eQTG candidates if the top 5%, 10%, and 20% ranked genes are considered (Figure 3). With added features, eQTG-Finder ranks eQTGs slightly better with percentages of 36%, 52%, and 64% respectively. The eQTGs vary in their evidence of being causal genes (see Methods). Four out of six strong eQTG candidates (*AOP2*, *ERECTA*, *GIGANTEA*, and *MAM1*) rank within the 5% rank percentile by eQTG-Finder compared to only one (*ERECTA*) by QTG-Finder2. The other two strong candidates, *FRI* and *ELF3*, were ranked at the 10.2% and 61.2% percentile by eQTG-Finder. The ranks of sixteen genes are improved by eQTG-Finder, eight are worse, and one stays the same (Supplementary Table 4). The rank of four out of six strong eQTG candidates improves, with *GIGANTEA* one of the most drastic improvements, moving from 53,7.7% to 4.2%. On the other hand, the rank of *ERECTA* drops (0.4% to 2.8%) but remained falls in the 5% rank percentile. Both models rank another strong eQTG candidate *ELF3* poorly (at 44% rank percentile by QTG-Finder2 and 61.2% by eQTG-Finder). Despite the decent overall performance in eQTG prioritization, we notice that eQTG-Finder performance in prioritizing phenotype QTGs is still inconsistent. Using the initial independent validation set, only seven out of eleven QTGs are ranked within the 20% rank percentile by eQTG-Finder, compared to nine by QTG-Finder2 (Supplementary Figure 2).

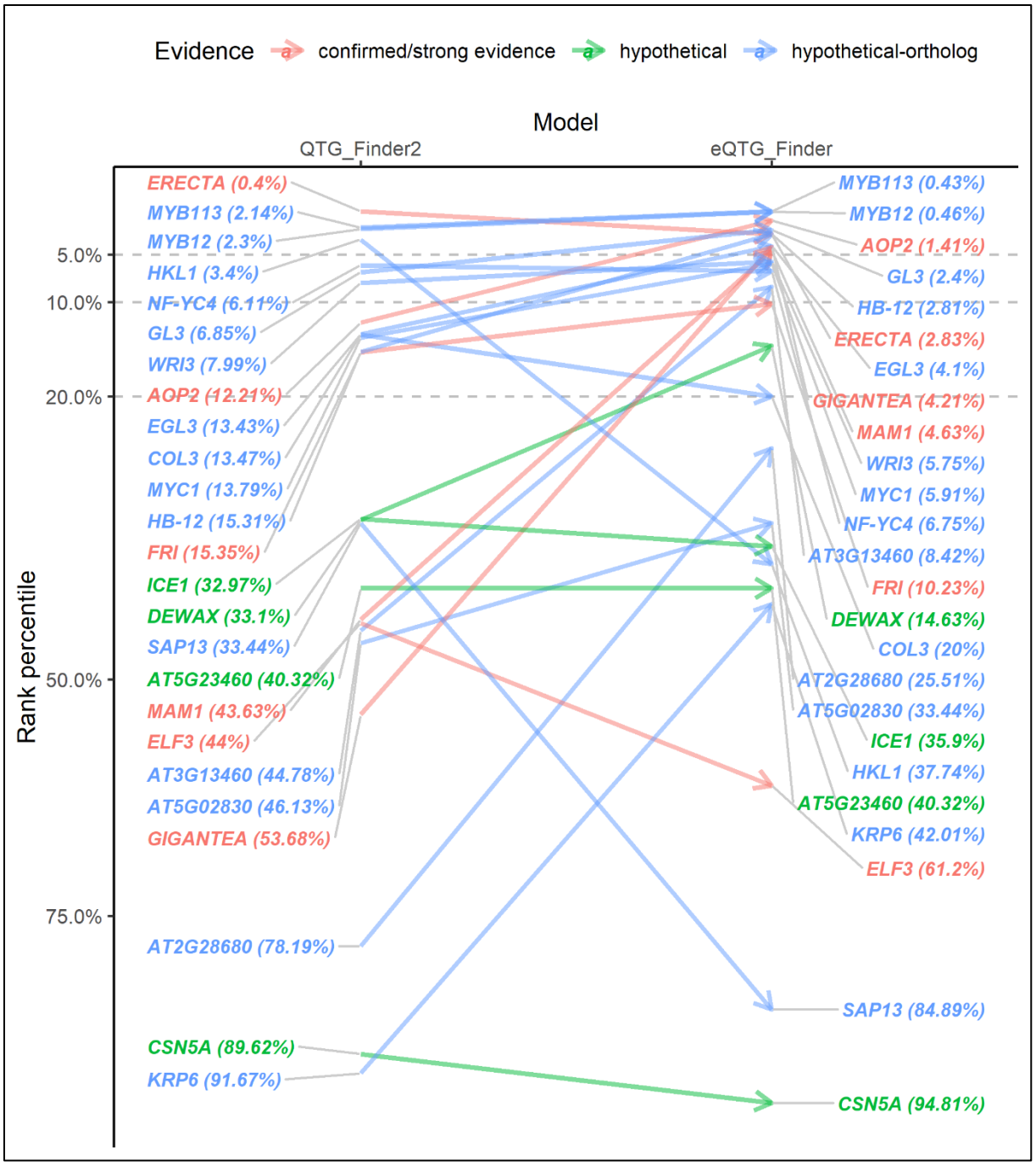


Figure 3. Rank percentile comparison of sixteen candidate eQTGs using the model with new features (eQTG-Finder) and the original model (QTG-Finder2).

To get an overview of eQTG-Finder predictions, we inspect the distribution of the average predicted probability of being causal for all Arabidopsis genes (Figure 4). This skewed towards a low value, with a median value of 0.007 (note that the *x*-axis of Figure 4 is on a $\log_{10}$ scale).

Twenty-one of the twenty-five genes in the validation set have a predicted probability higher than the median. *ELF3* (probability=0.0045) is the only strong eQTG candidate with a predicted probability lower than the median. A Gene Ontology (GO) enrichment analysis shows that the top 5% genes in the distribution are significantly enriched (FDR p-value < 0.05) for 67 GO terms (Supplementary Figure 5), most of which are related to response to abiotic and biotic stresses, such as "defense response to bacterium", "defense response to fungus", and "response to wounding". The term “regulation of transcription” is also enriched, suggesting that transcription factors are likely to be causal, consistent with the feature importance analysis result where is_TF is among the most important features. Meanwhile, the bottom 5% are not enriched for any term.

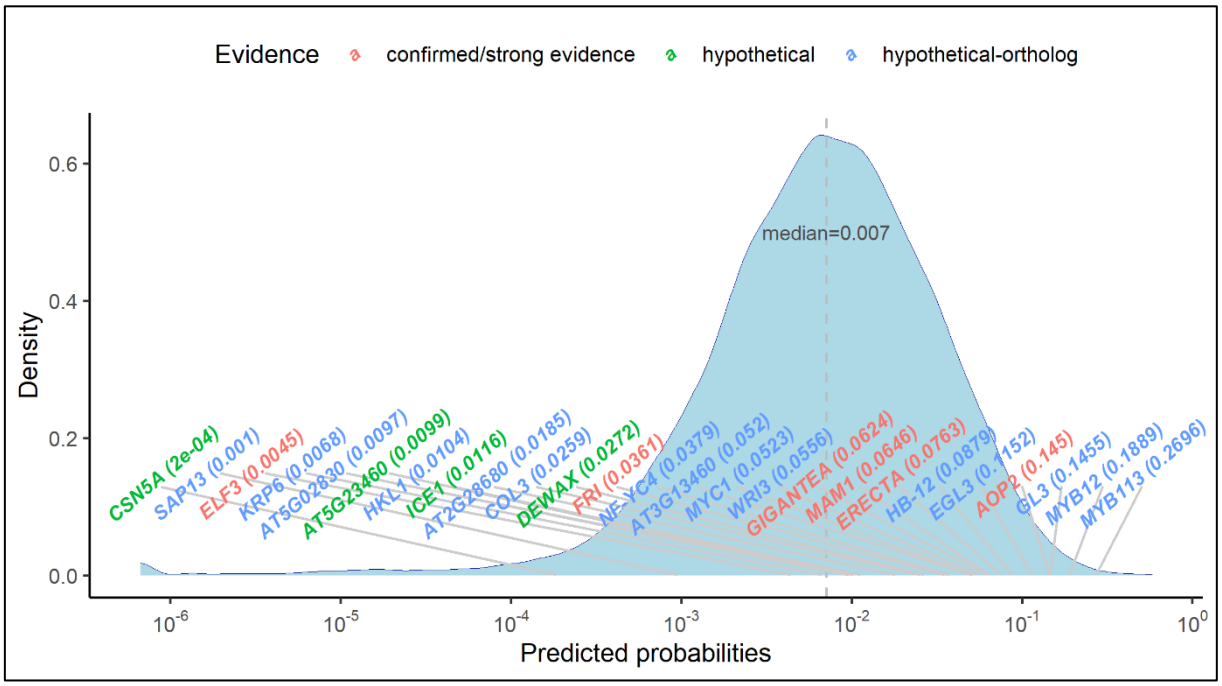


Figure 4. The density plot of probabilities of being causal predicted by eQTG-Finder for all Arabidopsis genes. Text labels point to the probability of the gene in the plot. The *x*-axis is on a $\log_{10}$ scale.

eQTG-Finder is available in AraQTL to support new hypotheses on the gene expression regulation

To make eQTG-Finder results easily accessible for researchers, we include predicted probabilities of causality (herewith referred to as eQTG-Finder score) for all Arabidopsis genes in AraQTL, our Arabidopsis eQTL data workbench (Nijveen et al. 2017). Prioritizing genes using QTG-Finder2 is not straightforward as it requires users to prepare a list of candidate genes and command-line usage skills. Integrating the eQTG-Finder score in AraQTL facilitates users to interactively identify gene expression regulators. For example, we here discuss a case on predicting a new potential regulator for *GLK2* using the eQTG-Finder score and other interaction evidence in AraQTL. *GLK2* is a GARP nuclear transcription factor involved in light-controlled signaling (Waters et al. 2009). Liu et al. (2021) recently found that *HY5* is the regulator of *GLK2* based on the fact that *HY5* is a well-known regulatory switch for light signaling in literature. The same conclusion can also be derived using the Serin et al. (manuscript in preparation) eQTL experiment and prior knowledge data in AraQTL. Another approach to finding potential regulators of *GLK2* can be made in AraQTL using the eQTG-Finder score. In a Kas x Tsu eQTL experiment on leaf tissue (Lowry et al. 2013), *GLK2* has an eQTL on the beginning of chromosome 1, indicating the location of the potential regulator(s) (Figure 5, top). As many as 257 candidate regulatory genes are present in the eQTL (Figure 5, bottom). We can filter out weak candidates by constructing a network of *GLK2* connected to its potential regulators on the eQTL based on prior knowledge, such as protein-protein interaction and gene annotation (Hartanto *et al.*, manuscript in preparation). Here, we threshold the eQTG-Finder score to remove weak candidates. Moreover, eQTG-Finder can prioritize the remaining fourteen genes by selecting the “Bipartite by eQTG-Finder score” network layout and ordering genes by their score. The result suggests some promising *GLK2* regulator candidates ranked at the top, for example, a transcription factor *LHY* in second place. Until now, *LHY* has not been

323 reported to regulate *GLK2*. However, this gene is a promising *GLK2* regulator candidate as the
324 network shows that it has a transcription factor binding site(s) on the *GLK2* promoter (O'Malley
325 et al. 2016). Moreover, *LHY* is involved in light signaling (Joo et al. 2017; Kim et al. 2003).
326 This example suggests that integrating the eQTG-Finder score in AraQTL can help infer new
327 regulatory interactions.

328

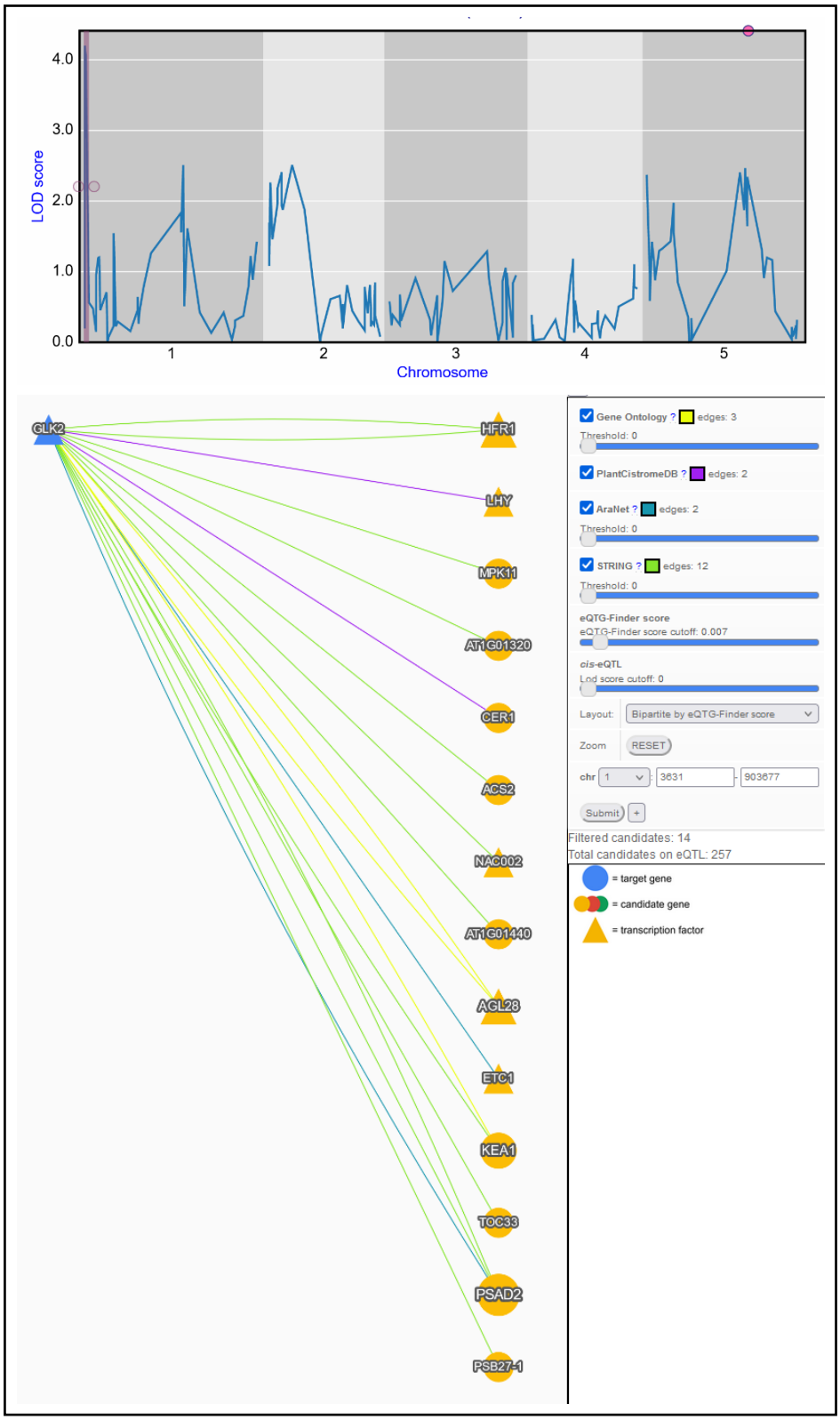


FIGURE 5. Prioritization of *GLK2* regulator using the eQTG-Finder score in AraQTL. (top) eQTL profile of *GLK2* from the Lowry et al. (2013) experiment. The eQTL region on chromosome 1 (shaded in pink)

pinpoints the location of potential *GLK2* regulator(s). (bottom) Prior-knowledge network connecting *GLK2* (blue node) with candidate regulators (yellow nodes) based on prior knowledge data. Here, the eQTG-Finder score is used to order candidates based on their probability of being causal.

DISCUSSION

The concept of genetical genomics was first coined two decades ago (Jansen and Nap 2001), and numerous Arabidopsis eQTL data sets have been published since then (Nijveen et al. 2017). The aim of genetical genomics is to pinpoint genomic regions associated with gene expression variation (eQTL) and ultimately unravel genes involved in expression regulation. However, identifying causal genes (eQTGs) is difficult because of the often large genomic regions they span, regularly harboring dozens or even hundreds of candidates. The regions can be narrowed down by experimental fine-mapping (Eshed and Zamir 1995), and the remaining candidate genes can then be validated using functional genomics methods (*e.g.*, using CRISPR-Cas9-mediated deletions as in Evans and Andersen 2020). However, performing these experiments for thousands of eQTLs is very costly. Using genomics and annotation data, a computational prioritization method can help identify candidate eQTGs. This study extends an existing machine-learning algorithm, QTG-Finder2, to address this issue and evaluates its performance for prioritizing eQTG. eQTG-Finder outperforms its predecessor in a cross-validation setting and independent validation test. We make eQTG-Finder scores available in AraQTL to help researchers interactively identify key regulators.

The key improvement of eQTG-Finder lies in the inclusion of twelve new features based on gene expression, structure, and interactions. Given the complexity of the resulting model, it is

not straightforward to assess how these features improve eQTG-Finder in gene prioritization (Petch et al. 2022). We calculated the contribution of each feature in the model using a leave-one-out feature importance analysis (see Materials and Methods). This showed that the number of unique protein domains, the number of protein-protein interactions (PPI), and the number of introns are in the top five most contributing features in the model. We showed that known causal genes tend to have more domains, protein-protein interaction partners, and introns than other genes (Figure 1). These new features may provide insight into what distinguishes causal and non-causal genes. For instance, since protein domains determine protein functions (Vogel et al. 2004; Enright and Ouzounis 2001), the presence of multiple domains in a causal gene could indicate involvement in a wide range of biological functions. The diverse functions of causal genes could also be reflected in their larger number of protein-protein interaction partners than non-causal as genes perform their function in concert with other genes (Ito et al. 2001). The number of introns reflects the number of exons in a gene. Several studies demonstrated that exons play a role in the evolution of domain architectures through exon-shuffling, leading to new combinations of domains with new functions.

Variation in phenotype can be traced back to variation in gene expression (Skelly et al. 2009; Albert and Kruglyak 2015). For this reason, we included features based on the standard deviation (SD) of gene expression across different Arabidopsis accessions and conditions. Even though the medians between causal and other genes are significantly different (Figure 1), features based on SD of expression have low importance in the model. A study showed that correlations between features decrease the importance to zero (Gregorutti et al. 2016). Given that three SD features are highly correlated (Supplementary Figure 1), their importance in the model might be underestimated. Nevertheless, we do not have evidence that these features negatively affect the prediction performance; hence, we kept them in the model.

eQTG-Finder uses known QTGs (*i.e.*, causal genes for a phenotype QTL) as positive instances for model training because of the limited number of known eQTGs. We argue that QTGs are relevant for prioritizing eQTG since variation at the molecular level (*e.g.*, in gene expression, metabolite, or protein level) can be propagated and cause variation at higher phenotypic levels (Fu et al. 2009; Civelek and Lusis 2013). For example, genetic variations in *AOP2* and *MAM1* cause *cis*-eQTLs for gene expression and metabolite QTLs for aliphatic glucosinolate biosynthesis, which confer insect resistance in Arabidopsis (Wentzell et al. 2007; Jansen et al. 2009). Both genes were prioritized in the top 5% rank percentile by eQTG-Finder. This result suggests that eQTG-Finder can identify QTLs for other molecular phenotypes, including metabolite and protein.

A lack of model interpretability may hamper a user's comprehensive evaluation and assessment of the prioritization results. Regardless of the good performance, it is difficult to precisely understand how eQTG-Finder classifies certain genes as causal and others as non-causal, a typical issue for a complex model like Random Forest (Petch et al. 2022). Instead, in AraQTL, we provide additional sources of evidence to support the eQTG-Finder prioritization results (Hartanto *et al.*, unpublished). For example, eQTG-Finder prioritizes transcription factor *LHY* as the regulator of *GLK2* (Figure 5). The network visualization in AraQTL showed that *LHY* is connected to *GLK2* by transcription factor binding site evidence, indicating that *LHY* may bind to the *GLK2* promoter and modulate its expression. Incorporating eQTG-Finder in the AraQTL web interface facilitates researchers to identify key regulators for genes of interest without the need for computational skills.

In the independent validation, some eQTG candidates were ranked poorly by eQTL-Finder (Figure 3). Low ranked assumed eQTG genes from the hypothetical and hypothetical-orthologs groups might not be actual eQTGs; however, the strong eQTG candidate ELF3 was also ranked poorly by both eQTG-Finder (61.2%) and QTG-Finder (44%). *ELF3* encodes a nuclear protein

and was demonstrated to regulate gene expression leading to shade-avoidance response (Jimenez-Gomez et al. 2010). The complexity of the eQTG-Finder algorithm makes it difficult to dissect the prediction for *ELF3*. We investigated two of the most important features and noticed that this gene only has one identified protein domain and one paralog copy number, which is lower than the median values of causal genes (four and seventeen, respectively).

Likely, some features associated with eQTG are still missing in our model or underrepresented in our set of positive instances. Since the regulator-target relationship is specific, we expect that features representing gene-gene/protein-protein relationships (for example, STRING scores (Szklarczyk et al. 2019), transcription factor binding sites (Tian et al. 2020), and gene ontology semantic similarity (Yu 2020)) are relevant for prioritizing eQTG. Including these would shift the prioritization of generic eQTGs based on gene properties to the prioritization of eQTGs for a specific target using features based on gene-pair relationships. This is similar to the approaches of Wong et al. (2004) and Pandey et al. (2010), who predicted genetic interaction using gene pair relationships in yeast. The number of positive examples (*i.e.*, confirmed eQTG-target pairs) is currently too small to properly train such a model for Arabidopsis. However, as data regarding genetic regulation is steadily increasing, we are optimistic that this strategy will be possible in the future.

Data availability

The code and data for the analysis and visualization is available at the Wageningen University GitLab repository (https://git.wur.nl/harta003/eqtg-finder). eQTG-Finder prioritization is available at AraQTL (https://www.bioinformatics.nl/AraQTL/; Nijveen et al. 2017)

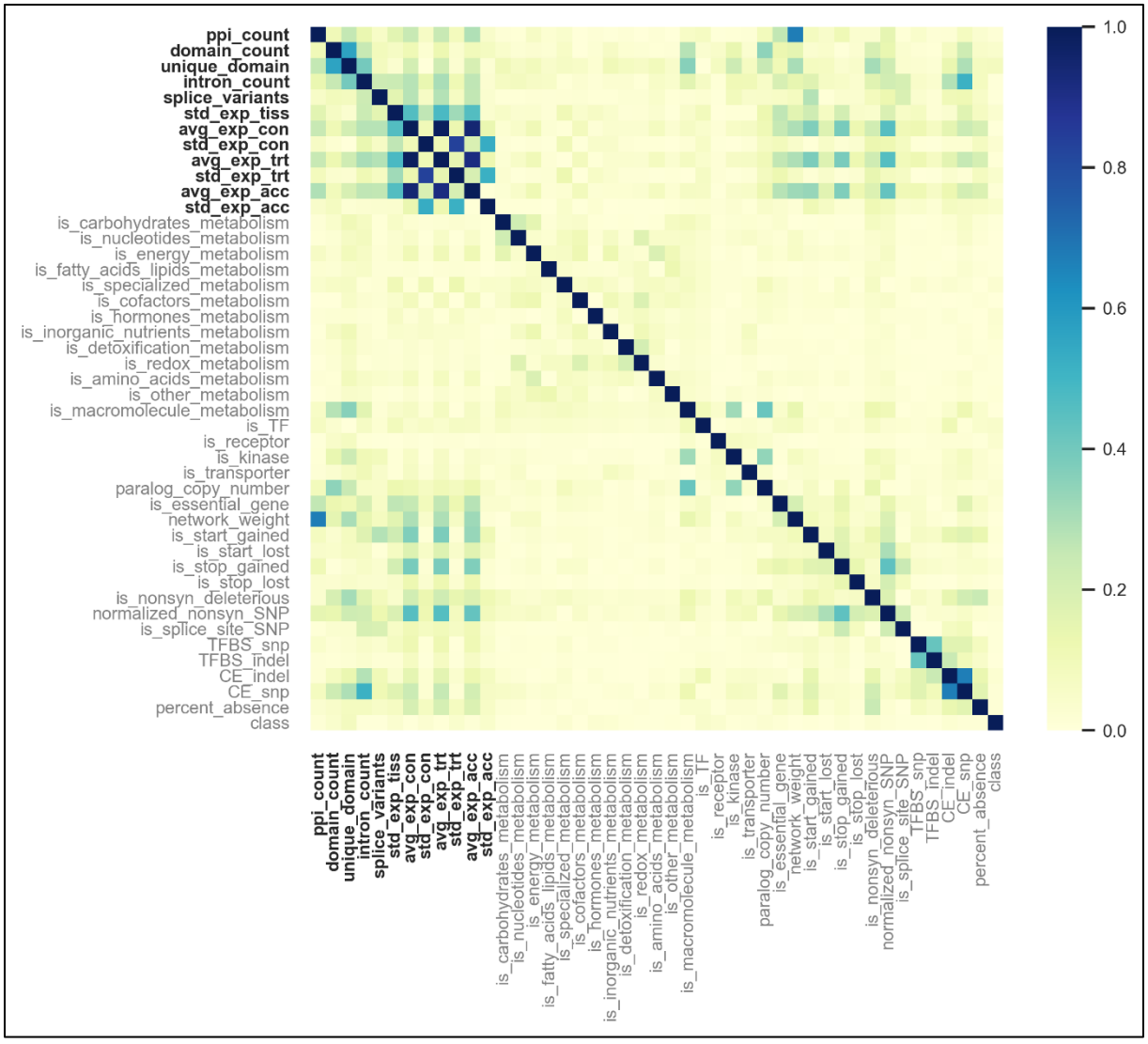
Acknowledgments

We thank members of the Bioinformatics Group, Wageningen University, for feedback and suggestions.

Conflict of Interest

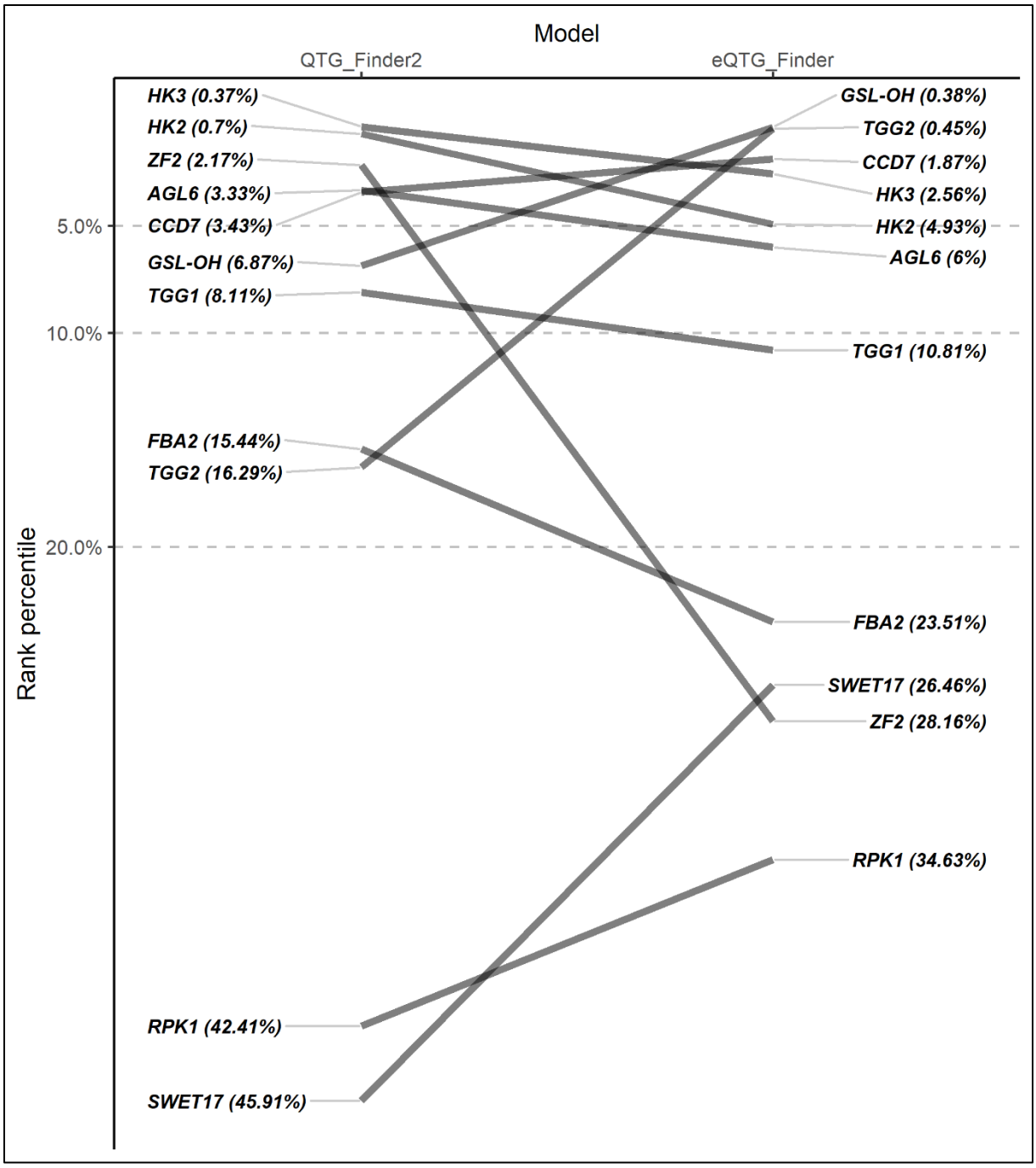
We declare no conflict of interest.

SUPPLEMENTARY FIGURES



Supplementary Figure 1. Correlation matrix of features used in the machine learning model.

New features are indicated in bold.



Supplementary Figure 2. Rank percentile comparison of eleven original validation gene sets using the model with new features and the original model (QTG-Finder2).

SUPPLEMENTARY TABLES

Supplementary Table 2. The list of candidate eQTGs used for independent validation

Gene ID	Gene name	evidence	Species origin	Label in species origin	Gene class	Associated trait	Trait category	Reference (PMID)
AT5G61590	DEWAX	hypothetical	*Arabidopsis thaliana*	DEWAX	transcription factor	seed germination	development	32963085
AT1G22770	GIGANTEA	confirmed/ strong evidence	*Arabidopsis thaliana*	GIGANTEA	nuclear protein	seed germination	development	32963085
AT3G26744	ICE1	hypothetical	*Arabidopsis thaliana*	ICE1	transcription factor	flowering	development	23335938/ 17237218
AT5G63470	NF-YC4	hypothetical-ortholog	*Solanum tuberosum*	NF-YC4	transcription factor	drought responce	abiotic stress	27353051
AT3G61890	HB-12	hypothetical-ortholog	*Solanum tuberosum*	HB-12	transcription factor	drought responce	abiotic stress	27353051
AT2G28680	-	hypothetical-ortholog	*Solanum tuberosum*	PGCRURSE5	RmlC-like cupins superfamily protein	tuber starch content	development	33051578
AT5G02830	-	hypothetical-ortholog	*Cucumis melo*	cmPPR1/ Melo3C003069	Tetratricopeptide repeat (TPR)-like superfamily protein	flesh color intensity	development	29385635
AT3G13460	-	hypothetical-ortholog	*Zea mays*	ECT2	YTH domain-containing protein	kernel size	development	30548709
AT1G50460	HKL1	hypothetical-ortholog	*Zea mays*	HEX9	hexokinase-like	glycolysis	development	29275164
AT1G63650	EGL3	hypothetical-ortholog	*Zea mays*	R1/COLORED1	transcription factor	flavonoid biosynthesis	development	32184350/ 29275164
AT5G41315	GL3	hypothetical-ortholog	*Zea mays/ Hordeum vulgare*	R1/COLORED1	transcription factor	flavonoid biosynthesis	development	21115826/ 32184350/ 29275164

AT4G00480	MYC1	hypothetical-ortholog	*Zea mays*	R1/COLORED1	transcription factor	flavonoid biosynthesis	development	32184350/ 29275164
AT2G47460	MYB12	hypothetical-ortholog	*Solanum lycopersicum/ Populus trichocarpa*	MYB12	transcription factor	flavonoid biosynthesis	development	33199703/ 6888306
AT3G57480	SAP13	hypothetical-ortholog	*Solanum lycopersicum*	PtSAP13	stress-associated protein	flavonoid biosynthesis	development	33199703
AT1G16060	WRI3	hypothetical-ortholog	*Solanum lycopersicum*	WRI3	transcription factor	lipid metabolism	development	33199703
AT1G22920	CSN5A	hypothetical	*Arabidopsis thaliana*	CSN5A	transcription coactivator	low-light response	abiotic stress	23335938
AT1G66370	MYB113	hypothetical-ortholog	*Ipomoea batatas*	IbMYB1-2	transcription factor	flavonoid biosynthesis	development	32528702
AT3G19150	KRP6	hypothetical-ortholog	*Gossypium hirsutum*	KRP6	kip-related protein	fibre-cell length	development	32017125
AT2G24790	COL3	hypothetical-ortholog	*Zea mays*	COL11	transcription factor	photosynthesis	development	32184350
AT5G23460	-	hypothetical	*Arabidopsis thaliana*	-	-	flowering	development	17237218
AT2G25930	ELF3	confirmed/ strong evidence	*Arabidopsis thaliana*	ELF3	nuclear protein	shade avoidance	abiotic stress	20838594
AT2G26330	ERECTA	confirmed/ strong evidence	*Arabidopsis thaliana*	ERECTA	kinase			20833726
AT4G00650	FRI	confirmed/ strong evidence	*Arabidopsis thaliana*	FRI	-		development	24045022
AT5G23010	MAM1	confirmed/ strong evidence	*Arabidopsis thaliana*	MAM1	methylthioalkylmalate synthase	insect resistance	biotic stress	19196544
AT4G03050	AOP2	confirmed/ strong evidence	*Arabidopsis thaliana*	AOP2	2-oxoglutarate-dependent dioxygenase	insect resistance	biotic stress	19196544

Supplementary Table 3. Wilcoxon Rank Sum Test statistics and p values for the difference in median of new features between causal genes and the other genes in the genome.

Feature	W statistics	p-value
std exp tiss	3312857	2.51×10^{8}
avg exp con	2624347	0.51
std exp con	3207362	3.86×10^{6}
avg exp trt	2531055	0.13
std exp trt	3354636	2.64×10^{9}
avg exp acc	2445143	0.02
std exp acc	3287936	8.19×10^{8}
intron count	2937908	0.03
splice variants	2805881	0.17
ppi count	3142925	3.75×10^{5}
domain count	3705449	3×10^{20}
unique domain	3944166	3.12×10^{30}

Supplementary Table 4. Candidate eQTL genes and their rank percentile based on the original QTG-Finder2 and eQTG-Finder.

ID	Name	Type	Evidence	Rank		Total gene in QTL	Rank percentile		Rank improvement
				QTG_Finder2	eQTG_Finder		QTG_Finder2	eQTG_Finder	
AT2G28680	-	eQTG	hypothetical ortholog	190	62	243	78.189	25.514	52.675
AT3G19150	KRP6	eQTG	hypothetical ortholog	264	121	288	91.667	42.014	49.653
AT1G22770	GIGANTEA	eQTG	confirmed/ strong evidence	153	12	285	53.684	4.211	49.474
AT5G23010	MAM1	eQTG	confirmed/ strong evidence	113	12	259	43.629	4.633	38.996
AT3G13460	-	eQTG	hypothetical ortholog	133	25	297	44.781	8.418	36.364
AT5G61590	DEWAX	eQTG	hypothetical	95	42	287	33.101	14.634	18.467
AT5G02830	-	eQTG	hypothetical ortholog	149	108	323	46.130	33.437	12.693
AT3G61890	HB-12	eQTG	hypothetical ortholog	49	9	320	15.313	2.813	12.500
AT4G03050	AOP2	eQTG	confirmed/ strong evidence	26	3	213	12.207	1.408	10.798
AT1G63650	EGL3	eQTG	hypothetical ortholog	36	11	268	13.433	4.104	9.328
AT4G00480	MYC1	eQTG	hypothetical ortholog	28	12	203	13.793	5.911	7.882
AT4G00650	FRI	eQTG	confirmed/ strong evidence	33	22	215	15.349	10.233	5.116
AT5G41315	GL3	eQTG	hypothetical ortholog	20	7	292	6.849	2.397	4.452
AT1G16060	WRI3	eQTG	hypothetical ortholog	25	18	313	7.987	5.751	2.236
AT2G47460	MYB12	eQTG	hypothetical ortholog	5	1	217	2.304	0.461	1.843
AT1G66370	MYB113	eQTG	hypothetical ortholog	5	1	234	2.137	0.427	1.709
AT5G23460	-	eQTG	hypothetical	100	100	248	40.323	40.323	0.000
AT5G63470	NF-YC4	eQTG	hypothetical ortholog	19	21	311	6.109	6.752	-0.643
AT2G26330	ERECTA	eQTG	confirmed/ strong evidence	1	7	247	0.405	2.834	-2.429
AT3G26744	ICE1	eQTG	hypothetical	90	98	273	32.967	35.897	-2.930
AT1G22920	CSN5A	eQTG	hypothetical	259	274	289	89.619	94.810	-5.190
AT2G24790	COL3	eQTG	hypothetical ortholog	33	49	245	13.469	20.000	-6.531
AT2G25930	ELF3	eQTG	confirmed/ strong evidence	110	153	250	44.000	61.200	-17.200
AT1G50460	HKL1	eQTG	hypothetical ortholog	9	100	265	3.396	37.736	-34.340
AT3G57480	SAP13	eQTG	hypothetical ortholog	104	264	311	33.441	84.887	-51.447
AT4G15920	SWET17	QTG	n/a	118	68	257	45.914	26.459	19.455
AT5G25980	TGG2	QTG	n/a	36	1	221	16.290	0.452	15.837
AT1G69270	RPK1	QTG	n/a	109	89	257	42.412	34.630	7.782
AT2G25450	GSL-OH	QTG	n/a	18	1	262	6.870	0.382	6.489
AT2G44990	CCD7	QTG	n/a	11	6	321	3.427	1.869	1.558
AT1G27320	HK3	QTG	n/a	1	7	273	0.366	2.564	-2.198
AT2G45650	AGL6	QTG	n/a	10	18	300	3.333	6.000	-2.667
AT5G26000	TGG1	QTG	n/a	18	24	222	8.108	10.811	-2.703
AT5G35750	HK2	QTG	n/a	1	7	142	0.704	4.930	-4.225
AT4G38970	FBA2	QTG	n/a	44	67	285	15.439	23.509	-8.070
AT3G19580	ZF2	QTG	n/a	6	78	277	2.166	28.159	-25.993

Supplementary Table 5. Gene ontology terms enriched in the top 5% genes predicted as causal.

GO ID	Term	Annotated	Significant	Expected	p-value	FDR
GO:0030154	cell differentiation	675	118	37.74	7.50×10^{30}	4.51×10^{26}
GO:0006468	protein phosphorylation	1037	141	57.98	3.00×10^{22}	9.02×10^{19}
GO:0051762	sesquiterpene biosynthetic process	25	19	1.4	1.80×10^{19}	3.61×10^{16}
GO:0006355	regulation of transcription, DNA-templat...	2053	279	114.78	2.60×10^{19}	3.91×10^{16}
GO:0007165	signal transduction	1429	167	79.89	7.80×10^{19}	9.38×10^{16}
GO:0048544	recognition of pollen	43	23	2.4	4.20×10^{18}	4.21×10^{15}
GO:0045893	positive regulation of transcription, DN...	509	96	28.46	1.40×10^{16}	1.20×10^{13}
GO:0042742	defense response to bacterium	430	69	24.04	1.70×10^{14}	1.28×10^{11}
GO:0009686	gibberellin biosynthetic process	29	14	1.62	3.00×10^{11}	2.01×10^{8}
GO:0070588	calcium ion transmembrane transport	45	16	2.52	2.70×10^{10}	1.62×10^{7}
GO:0050832	defense response to fungus	257	41	14.37	3.20×10^{10}	1.63×10^{7}
GO:0009753	response to jasmonic acid	189	37	10.57	3.70×10^{10}	1.63×10^{7}
GO:0009611	response to wounding	215	37	12.02	3.70×10^{10}	1.63×10^{7}
GO:0016114	terpenoid biosynthetic process	146	49	8.16	3.80×10^{10}	1.63×10^{7}
GO:0061408	positive regulation of transcription fro...	24	12	1.34	1.30×10^{9}	5.21×10^{7}
GO:0009414	response to water deprivation	381	52	21.3	1.50×10^{9}	5.64×10^{7}
GO:0045087	innate immune response	348	55	19.46	7.00×10^{9}	2.48×10^{6}
GO:0009617	response to bacterium	508	87	28.4	2.90×10^{8}	9.69×10^{6}
GO:0006952	defense response	1046	166	58.48	4.80×10^{8}	1.52×10^{5}
GO:2000652	regulation of secondary cell wall biogen...	28	13	1.57	5.60×10^{8}	1.68×10^{5}
GO:0045490	pectin catabolic process	96	21	5.37	5.90×10^{8}	1.69×10^{5}

GO:0002229	defense response to oomycetes	76	18	4.25	6.60×10^{8}	1.78×10^{5}
GO:0010200	response to chitin	141	26	7.88	6.80×10^{8}	1.78×10^{5}
GO:0046777	protein autophosphorylation	191	31	10.68	8.40×10^{8}	2.11×10^{5}
GO:0045487	gibberellin catabolic process	7	6	0.39	2.00×10^{7}	4.81×10^{5}
GO:0042545	cell wall modification	168	33	9.39	2.40×10^{7}	5.35×10^{5}
GO:0045944	positive regulation of transcription by ...	220	42	12.3	2.40×10^{7}	5.35×10^{5}
GO:0009735	response to cytokinin	104	20	5.81	2.60×10^{7}	5.59×10^{5}
GO:0080027	response to herbivore	15	8	0.84	4.20×10^{7}	8.71×10^{5}
GO:0019264	glycine biosynthetic process from serine	5	5	0.28	5.40×10^{7}	0.0001
GO:1904482	cellular response to tetrahydrofolate	5	5	0.28	5.40×10^{7}	0.0001
GO:0006565	L-serine catabolic process	5	5	0.28	5.40×10^{7}	0.0001
GO:0045892	negative regulation of transcription, DN...	296	35	16.55	1.00×10^{6}	0.0002
GO:0048481	plant ovule development	52	15	2.91	2.00×10^{6}	0.0004
GO:0009625	response to insect	30	10	1.68	3.10×10^{6}	0.0005
GO:0046655	folic acid metabolic process	6	5	0.34	3.10×10^{6}	0.0005
GO:0009416	response to light stimulus	741	90	41.43	7.80×10^{6}	0.001
GO:0009809	lignin biosynthetic process	49	13	2.74	8.10×10^{6}	0.001
GO:0007166	cell surface receptor signaling pathway	49	13	2.74	9.30×10^{6}	0.001
GO:0010114	response to red light	60	13	3.35	1.50×10^{5}	0.002
GO:0009620	response to fungus	327	57	18.28	1.80×10^{5}	0.003
GO:0005983	starch catabolic process	17	7	0.95	2.00×10^{5}	0.003
GO:0010093	specification of floral organ identity	13	7	0.73	2.10×10^{5}	0.003
GO:0010951	negative regulation of endopeptidase act...	12	6	0.67	2.10×10^{5}	0.003
GO:0009944	polarity specification of adaxial/abaxia...	23	8	1.29	2.10×10^{5}	0.003

GO:0009828	plant-type cell wall loosening	37	10	2.07	2.50×10^{5}	0.003
GO:0009693	ethylene biosynthetic process	30	9	1.68	2.50×10^{5}	0.003
GO:0006979	response to oxidative stress	453	57	25.33	4.30×10^{5}	0.005
GO:0055114	oxidation-reduction process	645	63	36.06	4.40×10^{5}	0.005
GO:0030574	collagen catabolic process	5	4	0.28	4.60×10^{5}	0.006
GO:0000266	mitochondrial fission	14	6	0.78	6.10×10^{5}	0.007
GO:0010087	phloem or xylem histogenesis	101	19	5.65	9.80×10^{5}	0.011
GO:0080086	stamen filament development	10	5	0.56	0.0001	0.012
GO:0070370	cellular heat acclimation	10	5	0.56	0.0001	0.012
GO:0009957	epidermal cell fate specification	6	4	0.34	0.0001	0.014
GO:0010106	cellular response to iron ion starvation	6	4	0.34	0.0001	0.014
GO:0009651	response to salt stress	461	47	25.77	0.0002	0.016
GO:0097054	L-glutamate biosynthetic process	3	3	0.17	0.0002	0.017
GO:0016099	monoterpenoid biosynthetic process	3	3	0.17	0.0002	0.017
GO:0009823	cytokinin catabolic process	3	3	0.17	0.0002	0.017
GO:1900386	positive regulation of flavonol biosynth...	3	3	0.17	0.0002	0.017
GO:0010311	lateral root formation	55	11	3.07	0.0002	0.019
GO:0031408	oxylipin biosynthetic process	17	6	0.95	0.0002	0.021
GO:0051301	cell division	258	26	14.42	0.0003	0.025
GO:0006826	iron ion transport	63	12	3.52	0.0003	0.03
GO:0009737	response to abscisic acid	541	55	30.25	0.0004	0.032
GO:0055072	iron ion homeostasis	93	16	5.2	0.0005	0.047

REFERENCES

Albert, F.W., and L. Kruglyak, 2015 The role of regulatory variation in complex traits and disease. *Nat Rev Genet* 16 (4):197-212.

Alexa, A., J. Rahnenführer, and T. Lengauer, 2006 Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* 22 (13):1600-1607.

Breitling, R., Y. Li, B.M. Tesson, J. Fu, C. Wu *et al.*, 2008 Genetical genomics: spotlight on QTL hotspots. *PLoS Genet* 4 (10):e1000232.

Brem, R.B., G. Yvert, R. Clinton, and L. Kruglyak, 2002 Genetic dissection of transcriptional regulation in budding yeast. *Science* 296 (5568):752-755.

Civelek, M., and A.J. Lusis, 2013 Systems genetics approaches to understand complex traits. *Nature Reviews Genetics* 15:34.

Cubillos, F.A., J. Yansouni, H. Khalili, S. Balzergue, S. Elftieh *et al.*, 2012 Expression variation in connected recombinant populations of Arabidopsis thaliana highlights distinct transcriptome architectures. *BMC Genomics* 13:117.

Enright, A.J., and C.A. Ouzounis, 2001 Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. *Genome Biol* 2 (9):RESEARCH0034.

Eshed, Y., and D. Zamir, 1995 An introgression line population of Lycopersicon pennellii in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. *Genetics* 141 (3):1147-1162.

Evans, K.S., and E.C. Andersen, 2020 The Gene scb-1 Underlies Variation in Caenorhabditis elegans Chemotherapeutic Responses. *G3 (Bethesda)* 10 (7):2353-2364.

Fu, J., J.J. Keurentjes, H. Bouwmeester, T. America, F.W. Verstappen *et al.*, 2009 System-wide molecular evidence for phenotypic buffering in Arabidopsis. *Nat Genet* 41 (2):166-167.

Gregorutti, B., B. Michel, and P. Saint-Pierre, 2016 Correlation and variable importance in random forests. *Statistics and Computing* 27 (3):659-678.

Hartanto, M., R.V.L. Joosen, B.L. Snoek, L.A.J. Willems, M.G. Sterken *et al.*, 2020 Network Analysis Prioritizes DEWAX and ICE1 as the Candidate Genes for Major eQTL Hotspots in Seed Germination of Arabidopsis thaliana. *G3 (Bethesda)* 10 (11):4215-4226.

Ho, T.K., 1998 The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20 (8):832-844.

Ito, T., T. Chiba, R. Ozawa, M. Yoshida, M. Hattori *et al.*, 2001 A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* 98 (8):4569-4574.

Jansen, R., and J. Nap, 2001 Genetical genomics: the added value from segregation. *Trends in Genetics* 17 (7):388-391.

Jansen, R.C., B.M. Tesson, J. Fu, Y. Yang, and L.M. McIntyre, 2009 Defining gene and QTL networks. *Curr Opin Plant Biol* 12 (2):241-246.

Jimenez-Gomez, J.M., A.D. Wallace, and J.N. Maloof, 2010 Network analysis identifies ELF3 as a QTL for the shade avoidance response in Arabidopsis. *PLoS Genet* 6 (9):e1001100.

Joo, Y., V. Fragoso, F. Yon, I.T. Baldwin, and S.G. Kim, 2017 Circadian clock component, LHY, tells a plant when to respond photosynthetically to light in nature. *J Integr Plant Biol* 59 (8):572-587.

Julca, I., C. Ferrari, M. Flores-Tornero, S. Proost, A.-C. Lindner *et al.*, 2020.

Keurentjes, J.J., J. Fu, I.R. Terpstra, J.M. Garcia, G. van den Ackerveken *et al.*, 2007 Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. *Proc Natl Acad Sci U S A* 104 (5):1708-1713.

Kim, J.Y., H.R. Song, B.L. Taylor, and I.A. Carre, 2003 Light-regulated translation mediates gated induction of the Arabidopsis clock protein LHY. *EMBO J* 22 (4):935-944.

Lelli, K.M., M. Slattery, and R.S. Mann, 2012 Disentangling the many layers of eukaryotic transcriptional regulation. *Annu Rev Genet* 46:43-68.

Lin, F., J. Fan, and S.Y. Rhee, 2019 QTG-Finder: A Machine-Learning Based Algorithm To Prioritize Causal Genes of Quantitative Trait Loci in Arabidopsis and Rice. *G3 (Bethesda)* 9 (10):3129-3138.

Lin, F., E.Z. Lazarus, and S.Y. Rhee, 2020 QTG-Finder2: A Generalized Machine-Learning Algorithm for Prioritizing QTL Causal Genes in Plants. *G3 (Bethesda)* 10 (7):2411-2421.

Liu, D., D. Zhao, X. Li, and Y. Zeng, 2021 AtGLK2, an Arabidopsis GOLDEN2-LIKE transcription factor, positively regulates anthocyanin biosynthesis via AtHY5-mediated light signaling. *Plant Growth Regulation* 96 (1):79-90.

Lowry, D.B., T.L. Logan, L. Santuari, C.S. Hardtke, J.H. Richards *et al.*, 2013 Expression quantitative trait locus mapping across water availability environments reveals contrasting associations with genomic features in Arabidopsis. *Plant Cell* 25 (9):3266-3279.

Marbach, D., J.C. Costello, R. Kuffner, N.M. Vega, R.J. Prill *et al.*, 2012 Wisdom of crowds for robust gene network inference. *Nat Methods* 9 (8):796-804.

Nijveen, H., W. Ligterink, J.J. Keurentjes, O. Loudet, J. Long *et al.*, 2017 AraQTL - workbench and archive for systems genetics in Arabidopsis thaliana. *Plant J* 89 (6):1225-1235.

O'Malley, R.C., S.C. Huang, L. Song, M.G. Lewsey, A. Bartlett *et al.*, 2016 Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell* 165 (5):1280-1292.

Pandey, G., B. Zhang, A.N. Chang, C.L. Myers, J. Zhu *et al.*, 2010 An integrative multi-network and multi-classifier approach to predict genetic interactions. *PLoS Comput Biol* 6 (9).

Petch, J., S. Di, and W. Nelson, 2022 Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology. *Can J Cardiol* 38 (2):204-213.

Rockman, M.V., and L. Kruglyak, 2006 Genetics of global gene expression. *Nat Rev Genet* 7 (11):862-872.

Serin, E.A., H. Nijveen, H.W. Hilhorst, and W. Ligterink, 2016 Learning from Co-expression Networks: Possibilities and Challenges. *Front Plant Sci* 7:444.

Skelly, D.A., J. Ronald, and J.M. Akey, 2009 Inherited variation in gene expression. *Annu Rev Genomics Hum Genet* 10:313-332.

Snoek, L.B., I.R. Terpstra, R. Dekter, G. Van den Ackerveken, and A.J. Peeters, 2012 Genetical Genomics Reveals Large Scale Genotype-By-Environment Interactions in Arabidopsis thaliana. *Front Genet* 3:317.

Szklarczyk, D., A.L. Gable, D. Lyon, A. Junge, S. Wyder *et al.*, 2019 STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 47 (D1):D607-D613.

Terpstra, I.R., L.B. Snoek, J.J. Keurentjes, A.J. Peeters, and G. van den Ackerveken, 2010 Regulatory network identification by genetical genomics: signaling downstream of the Arabidopsis receptor-like kinase ERECTA. *Plant Physiol* 154 (3):1067-1078.

Tian, F., D.C. Yang, Y.Q. Meng, J. Jin, and G. Gao, 2020 PlantRegMap: charting functional regulatory maps in plants. *Nucleic Acids Res* 48 (D1):D1104-D1113.

van Muijen, D., A.M. Anithakumari, C. Maliepaard, R.G. Visser, and C.G. van der Linden, 2016 Systems genetics reveals key genetic elements of drought induced gene regulation in diploid potato. *Plant Cell Environ* 39 (9):1895-1908.

Vogel, C., M. Bashton, N.D. Kerrison, C. Chothia, and S.A. Teichmann, 2004 Structure, function and evolution of multidomain proteins. *Curr Opin Struct Biol* 14 (2):208-216.

Waters, M.T., P. Wang, M. Korkaric, R.G. Capper, N.J. Saunders *et al.*, 2009 GLK transcription factors coordinate expression of the photosynthetic apparatus in Arabidopsis. *Plant Cell* 21 (4):1109-1128.

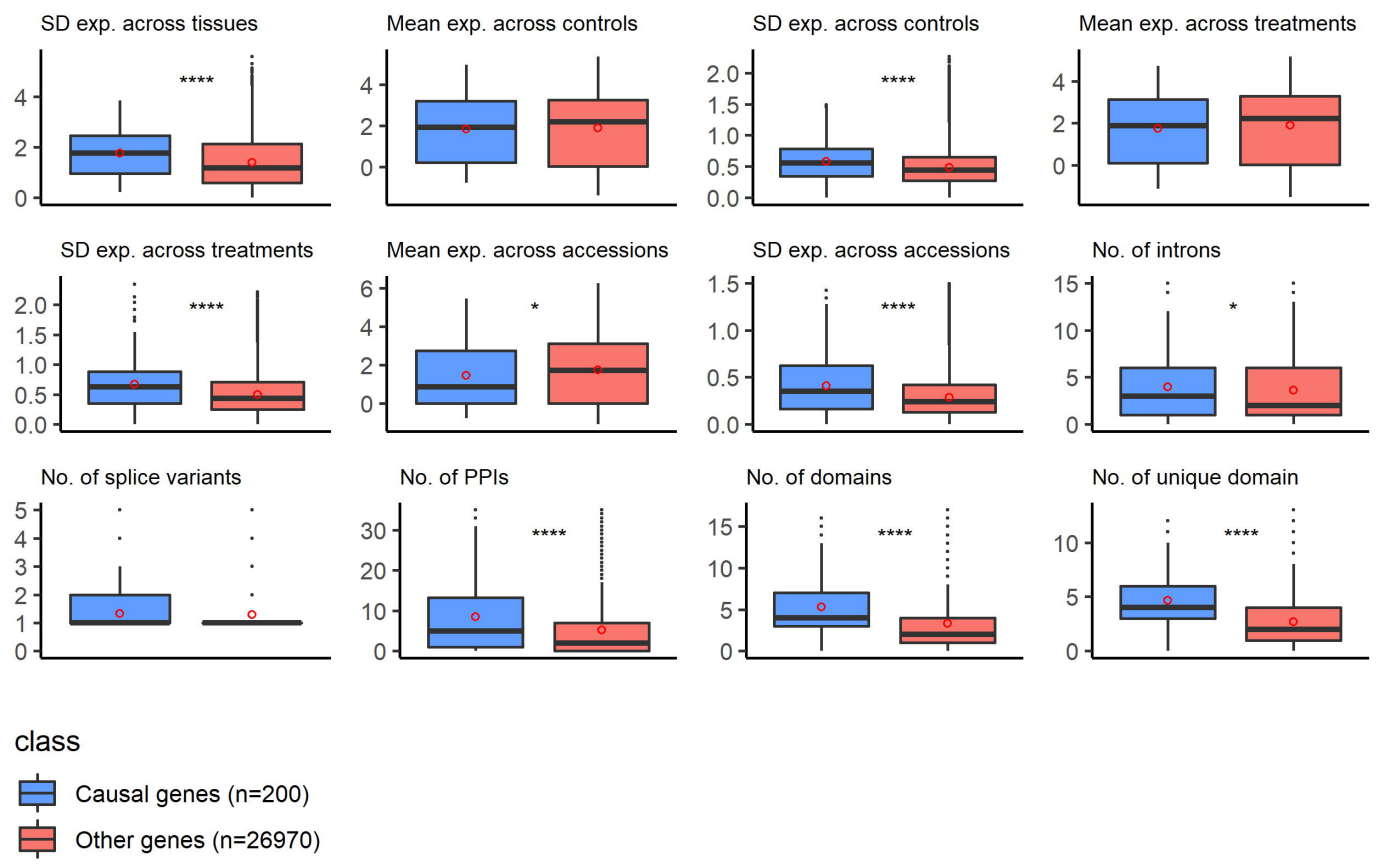
Wentzell, A.M., H.C. Rowe, B.G. Hansen, C. Ticconi, B.A. Halkier *et al.*, 2007 Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways. *PLoS Genet* 3 (9):1687-1701.

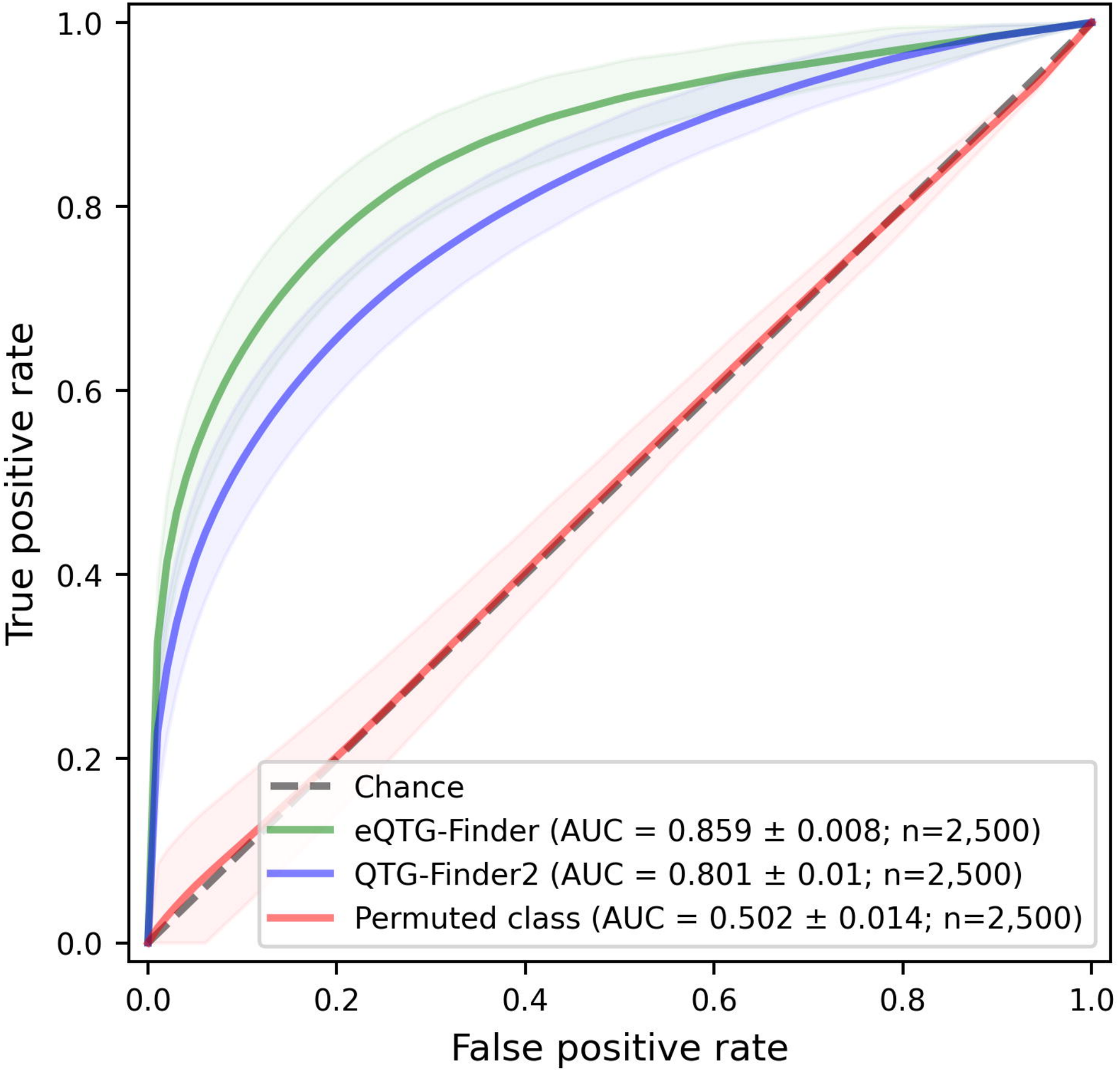
West, M.A., K. Kim, D.J. Kliebenstein, H. van Leeuwen, R.W. Michelmore *et al.*, 2007 Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. *Genetics* 175 (3):1441-1450.

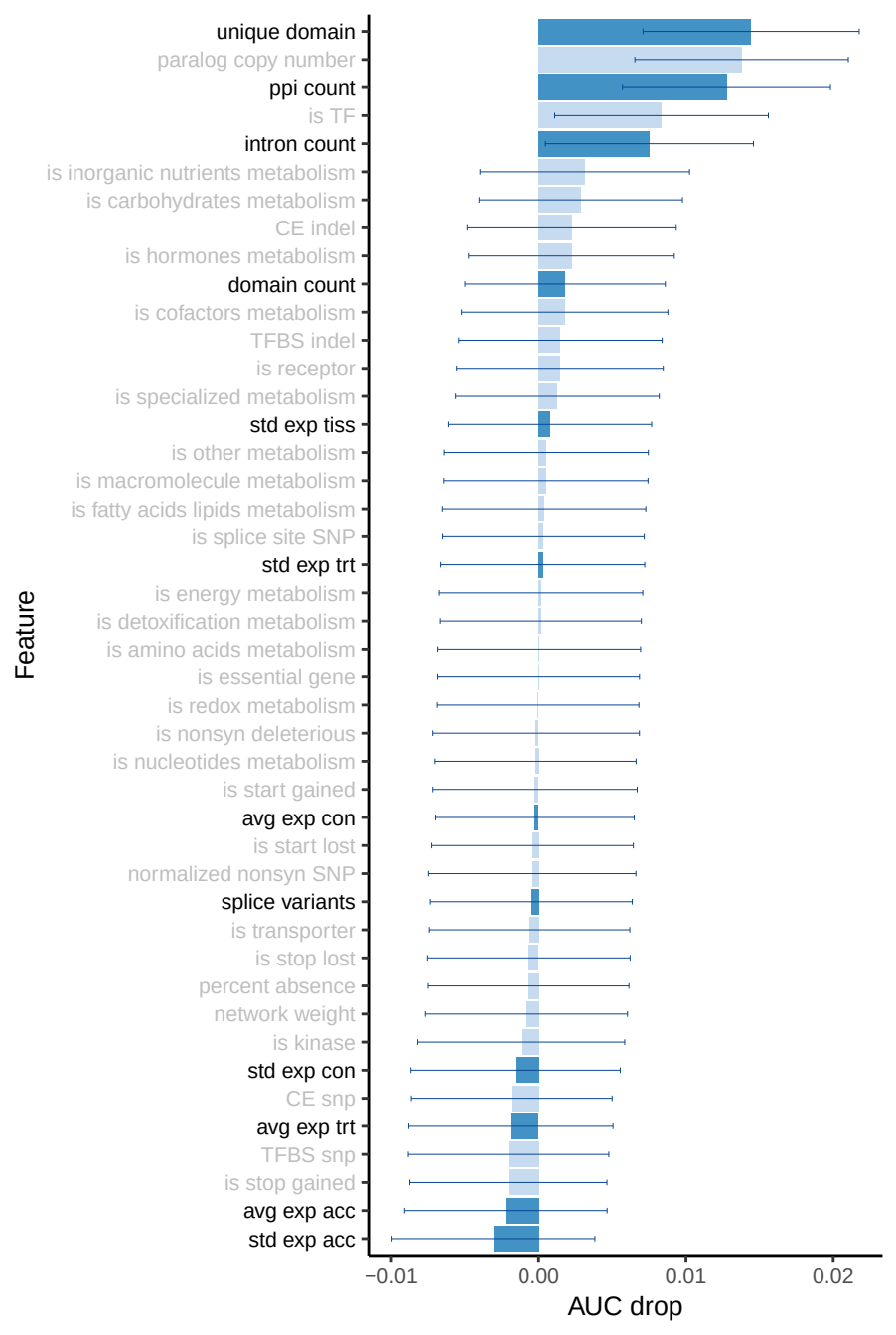
Wong, S.L., L.V. Zhang, A.H. Tong, Z. Li, D.S. Goldberg *et al.*, 2004 Combining biological networks to predict genetic interactions. *Proc Natl Acad Sci U S A* 101 (44):15682-15687.

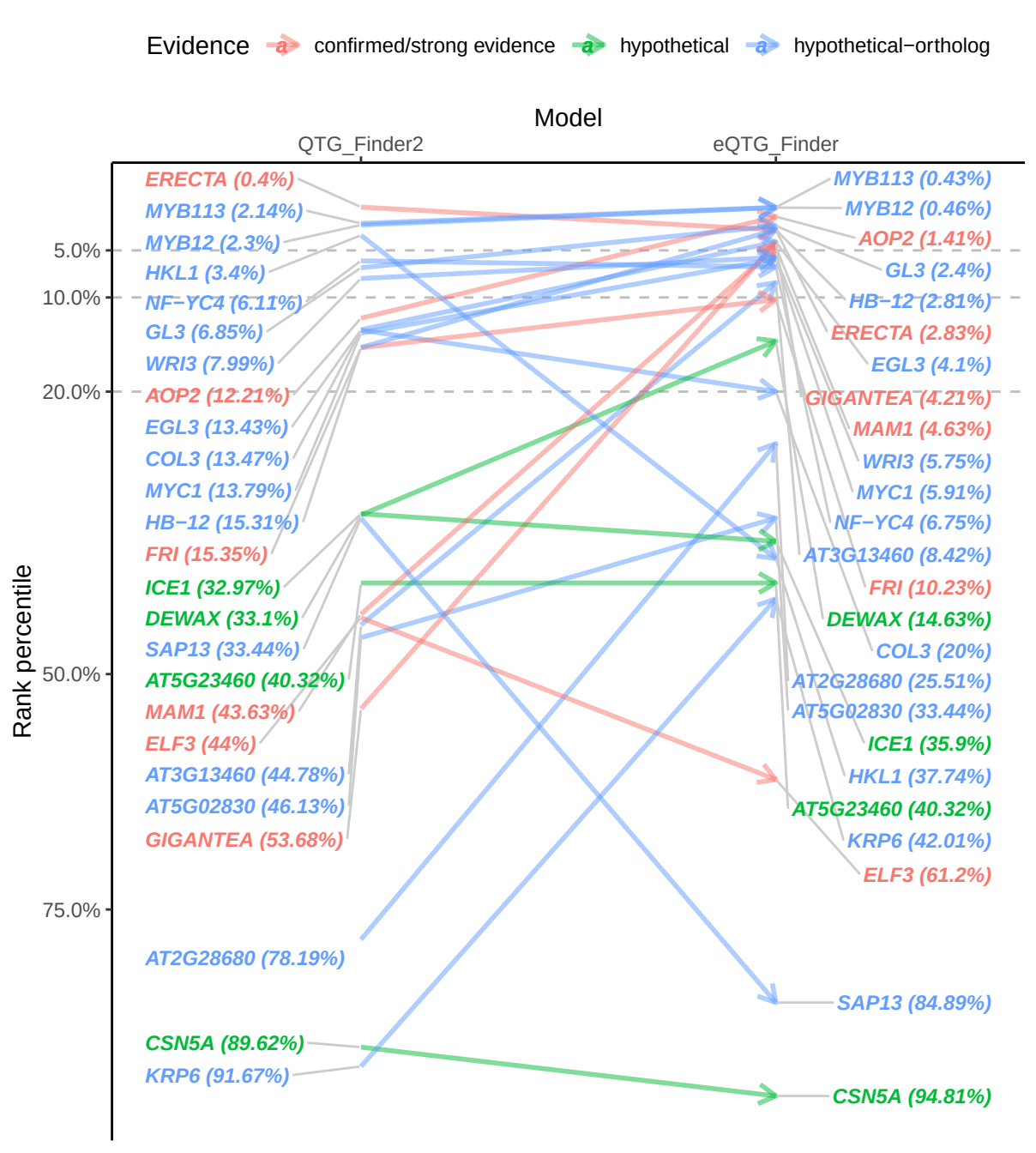
Yu, G., 2020 Gene Ontology Semantic Similarity Analysis Using GOSemSim. *Methods Mol Biol* 2117:207-215.

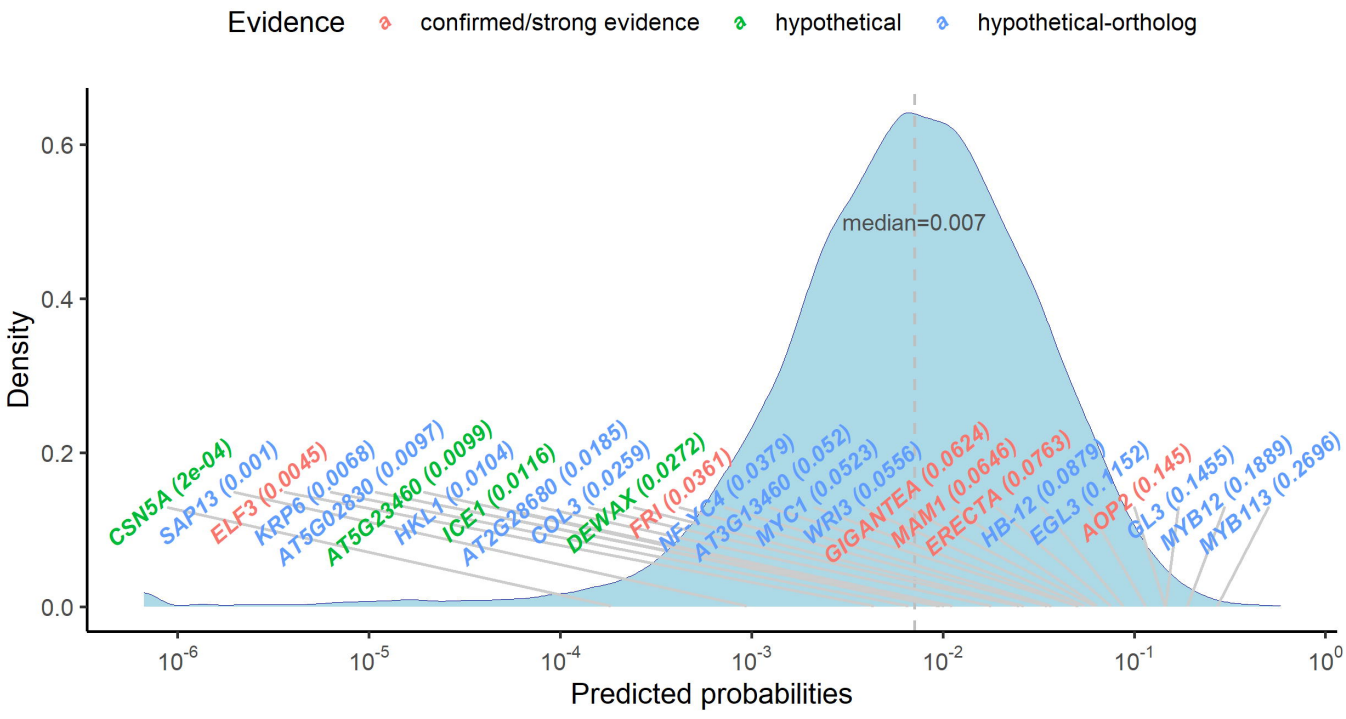
Zhang, H., F. Zhang, Y. Yu, L. Feng, J. Jia *et al.*, 2020 A Comprehensive Online Database for Exploring approximately 20,000 Public Arabidopsis RNA-Seq Libraries. *Mol Plant* 13 (9):1231-1233.


SD exp. across tissues
****
Mean exp. across controls
SD exp. across controls
****
Mean exp. across treatments
SD exp. across treatments
****
Mean exp. across accessions
*
SD exp. across accessions
****
No. of introns
*
No. of splice variants
No. of PPIs
****
No. of domains
****
No. of unique domain
****
class
Causal genes (n=200)
Other genes (n=26970)


1.0
0.8
0.6
0.4
0.2
0.0
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
Chance
eQTG-Finder (AUC = 0.859 ± 0.008; n=2,500)
QTG-Finder2 (AUC = 0.801 ± 0.01; n=2,500)
Permuted class (AUC = 0.502 ± 0.014; n=2,500)


unique domain
paralog copy number
ppi count
is TF
intron count
is inorganic nutrients metabolism
is carbohydrates metabolism
CE indel
is hormones metabolism
domain count
is cofactors metabolism
TFBS indel
is receptor
is specialized metabolism
std exp tiss
is other metabolism
is macromolecule metabolism
is fatty acids lipids metabolism
is splice site SNP
std exp trt
is energy metabolism
is detoxification metabolism
is amino acids metabolism
is essential gene
is redox metabolism
is nonsyn deleterious
is nucleotides metabolism
is start gained
avg exp con
is start lost
normalized nonsyn SNP
splice variants
is transporter
is stop lost
percent absence
network weight
is kinase
std exp con
CE snp
avg exp trt
TFBS snp
is stop gained
avg exp acc
std exp acc
Feature
−0.01
0.00
0.01
0.02
AUC drop


Evidence
confirmed/strong evidence
hypothetical
hypothetical−ortholog
Model
QTG_Finder2
eQTG_Finder
Rank percentile
5.0%
10.0%
20.0%
50.0%
75.0%
ERECTA (0.4%)
MYB113 (2.14%)
MYB12 (2.3%)
HKL1 (3.4%)
NF−YC4 (6.11%)
GL3 (6.85%)
WRI3 (7.99%)
AOP2 (12.21%)
EGL3 (13.43%)
COL3 (13.47%)
MYC1 (13.79%)
HB−12 (15.31%)
FRI (15.35%)
ICE1 (32.97%)
DEWAX (33.1%)
SAP13 (33.44%)
AT5G23460 (40.32%)
MAM1 (43.63%)
ELF3 (44%)
AT3G13460 (44.78%)
AT5G02830 (46.13%)
GIGANTEA (53.68%)
AT2G28680 (78.19%)
CSN5A (89.62%)
KRP6 (91.67%)
MYB113 (0.43%)
MYB12 (0.46%)
AOP2 (1.41%)
GL3 (2.4%)
HB−12 (2.81%)
ERECTA (2.83%)
EGL3 (4.1%)
GIGANTEA (4.21%)
MAM1 (4.63%)
WRI3 (5.75%)
MYC1 (5.91%)
NF−YC4 (6.75%)
AT3G13460 (8.42%)
FRI (10.23%)
DEWAX (14.63%)
COL3 (20%)
AT2G28680 (25.51%)
AT5G02830 (33.44%)
ICE1 (35.9%)
HKL1 (37.74%)
AT5G23460 (40.32%)
KRP6 (42.01%)
ELF3 (61.2%)
SAP13 (84.89%)
CSN5A (94.81%)


Evidence
confirmed/strong evidence
hypothetical
hypothetical-ortholog
median=0.007
Density
0.0
0.2
0.4
0.6
CSN5A (2e-04)
SAP13 (0.001)
ELF3 (0.0045)
KRP6 (0.0068)
AT5G02830 (0.0097)
AT5G23460 (0.0099)
HKL1 (0.0104)
ICE1 (0.0116)
AT2G28680 (0.0185)
COL3 (0.0259)
DEWAX (0.0272)
FRI (0.0361)
NF-YC4 (0.0379)
AT3G13460 (0.052)
MYC1 (0.0523)
WRI3 (0.0556)
GIGANTEA (0.0624)
MAM1 (0.0646)
ERECTA (0.0763)
HB-12 (0.0879)
EGL3 (0.1152)
AOP2 (0.145)
GL3 (0.1455)
MYB12 (0.1889)
MYB113 (0.2696)
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
Predicted probabilities