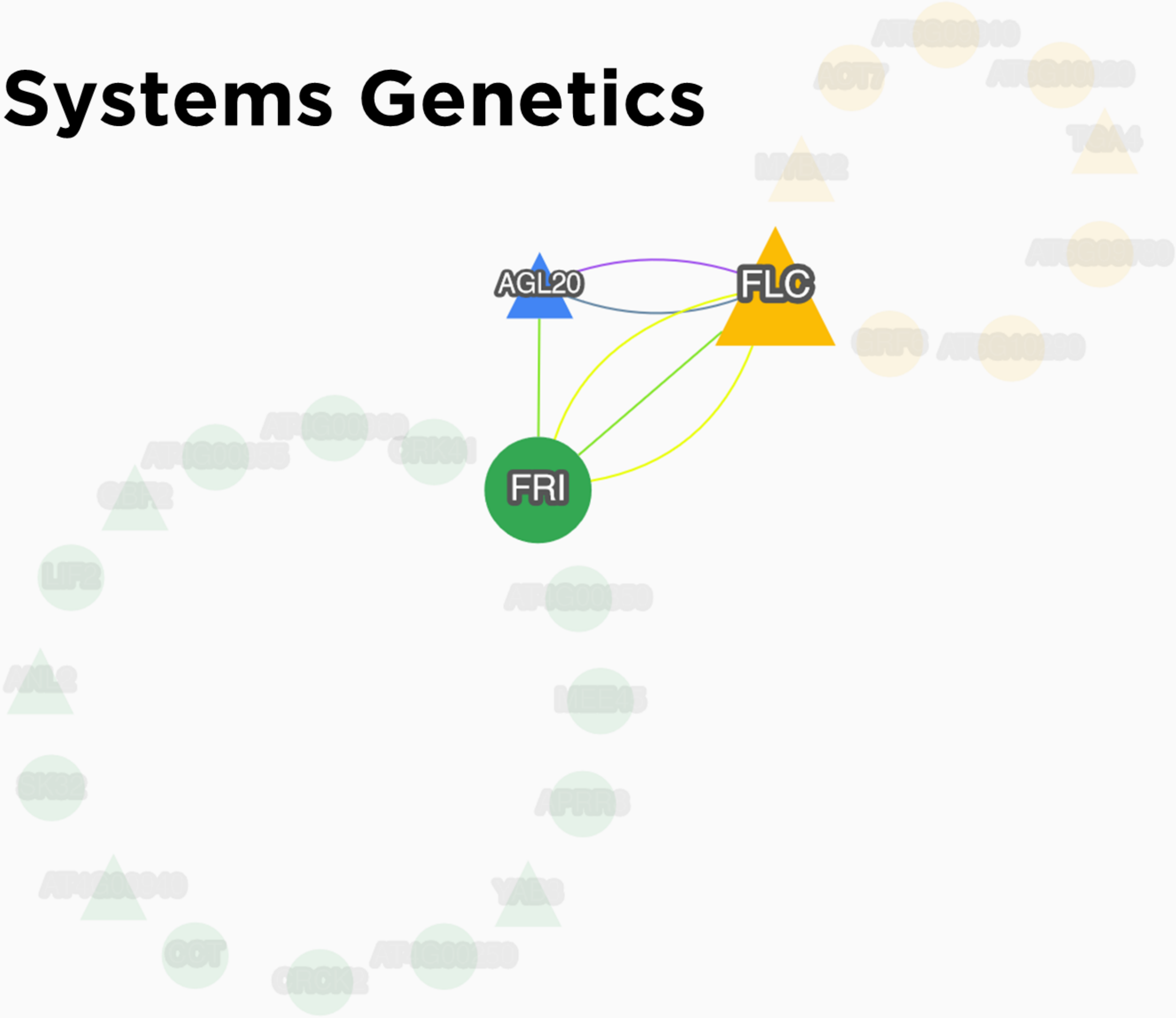
Linking Genes to Traits in Arabidopsis using Systems Genetics



Margi Hartanto

Propositions

1. Prior knowledge helps systems genetics studies deliver on their promise.
   (this thesis)
2. It is too early to apply systems genetics in plant breeding.
   (this thesis)
3. Independent repetition of a study is valuable and publishable.
4. Developing countries should focus on applied rather than basic research.
5. Replacing palm oil with other vegetable oil sources increases environmental impact.
6. The better way to clean the behind is by using water.

Propositions belonging to the thesis, entitled

Linking genes to traits in Arabidopsis using systems genetics

Margi Hartanto
Wageningen, 12 December 2023

Linking Genes to Traits in Arabidopsis Using Systems Genetics

Margi Hartanto

Thesis committee

Promotor

Prof. Dr D. de Ridder
Professor of Bioinformatics
Wageningen University & Research

Co-promotor

Dr H. Nijveen
Researcher, Bioinformatics Group
Wageningen University & Research

Other members

Prof. Dr L. Bentsink, Wageningen University & Research
Dr G.V. James, Rijk Zwaan
Prof. Dr J.E. Kammenga, Wageningen University & Research
Prof. Dr L.M. Trindade, Wageningen University & Research

This research was conducted under the auspices of the Graduate School Experimental Plant Sciences

Linking Genes to Traits in Arabidopsis Using Systems Genetics

Margi Hartanto

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Tuesday 12 December 2023
at 11 a.m. in the Omnia Auditorium.

Margi Hartanto

Linking genes to traits in Arabidopsis using systems genetics, 128 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2023)

With references, with a summary in English

ISBN 978-94-6447-961-4

DOI https://doi.org/10.18174/641121

Table of contents

Chapter 1. Introduction

1.1. Accelerating plant breeding through systems genetics

As the human population is expected to increase to 9 billion in 2050 (Godfray et al. 2010), food production should increase by 50-100% of the current rate to satisfy world demand (Royal Society of London 2009). Global food production has dramatically increased in the past decades thanks to efforts made during the Green Revolution (Khush 2001). However, the gains were achieved mainly through increases in land, water, and agrochemical use, which led to deforestation, nitrate pollution, and other major environmental impacts (FAO 2011). This approach is not sustainable because of the depletion of natural resources and the impact on the environment. In the meantime, agricultural production systems worldwide are facing serious challenges due to climate change. Global warming leads to temperature increases and longer, more severe drought periods, which cause environmental stress for crops (FAO 2016). Moreover, global warming is one of the drivers of crop pests and disease spread to new regions (Bebber et al. 2014; Bebber et al. 2013). These environmental impacts will eventually reduce crop yield and limit progress in meeting future food demand. A key solution to this problem is the development of improved crops through breeding.

Plant breeding is the science of developing plant cultivars by improving traits related to agricultural production and consumer demands, mainly crop yield and other agriculturally important traits, such as stress tolerance, nutritional qualities, and plant architecture. It has been practised since the beginning of agriculture by domesticating wild plants. Plant breeding plays an important role in addressing the human requirements for food and other plant-derived products.

Plant traits are improved through the collection or generation of trait variation and the subsequent selection of plants with desired characteristics. The main plant breeding approach has remained generally unchanged since it was first practised (Acquaah 2012). What has changed is the introduction of more cost-effective and efficient techniques due to scientific advancement. The emergence of genetics as a field of science played an essential role in revolutionising plant breeding techniques. Following the discovery of modern genetics by Gregor Mendel with his law of inheritance, researchers began to cross different plant varieties to produce hybrids with new combinations of traits. The next major milestone in plant breeding was the development of DNA-based markers that allow breeders to select plants more efficiently and accurately based on the genotypes (Moose and Mumm 2008).

Recently, the generation of massive genomics, transcriptomics, proteomics, metabolomics, and other omics data sets in plant research marks the emergence of a new branch of genetics: systems genetics (Civelek and Lusis 2013). However, plant breeding has yet to take full advantage of omics data and the opportunities of systems genetics. Recent reviews suggest that systems genetics has the potential to help us better understand traits at the molecular level by unravelling the complex gene and protein interaction networks governing polygenic traits (Lavarenne et al. 2018;

Pazhamala et al. 2021). Information about causal genes and their interactions could greatly help plant breeding, for example, to improve genomic prediction (Teng et al. 2020) or to provide targets for genome editing (Varshney et al. 2021). Causal gene discovery typically starts by finding an association between a trait and genetic variation in a plant population through quantitative trait locus (QTL) mapping.

1.2. QTL mapping to identify the genetic component of plant traits

Important target traits for plant breeding mostly show continuously distributed phenotypes. Such traits are called quantitative traits. A main characteristic of quantitative traits is that they are influenced by numerous genes (Hill 2010). For example, seed yield, the most important trait in oilseed rape (*Brassica napus*), was found to be associated with at least 85 distinct genomic locations (loci) (Shi et al. 2009). These trait-associated loci, individually known as quantitative trait locus (QTL), contain genes potentially controlling the trait.

The most widely used method to identify QTLs is called linkage mapping, or simply QTL mapping (Bazakos et al. 2017). This method requires a population developed by crossing two or more parental plants that vary in the trait of interest, thus having different variants (alleles) of genes controlling the trait. The first filial (F1) generation is self-fertilised to generate a segregated F2 population, where the individuals have a combination of alleles from the parents. Self-fertilisation can be continued up to the F6 or even later generation to produce recombinant inbred lines with very high levels of allelic homozygosity (Loudet et al. 2002; Alonso-Blanco et al. 1998).

In addition to a population, QTL mapping requires genetic markers to determine the genotypes of the parents and the offspring. Genetic markers are detectable DNA sequences located in the genome. These markers do not have to be located on the causal gene but could be in proximity so they can be inherited together (*i.e.,* are in genetic linkage). To be informative, markers should be able to discriminate allelic variants in the mapping population (polymorphic). There are many types of genetic markers, of which the most widely used today are single nucleotide polymorphisms (SNPs), variations of a single nucleotide at a specific position in the genome (Collard and Mackill 2008; Collard et al. 2005).

In general, QTL mapping analyses the correlation of the alleles of each marker with the corresponding trait values and defines QTL intervals based on markers showing significant correlation. From simple to more sophisticated, three common QTL mapping methods are single marker, simple interval, and composite interval mapping (Broman 2015). A QTL is usually presented as an interval indicating the location on the genome, together with the significance, the effect, and the most significant marker in the QTL. Plant breeders routinely use QTLs for marker-assisted selection. However, QTLs do not directly provide information about the biology of the trait. Subsequent identification of the QTL causal gene(s) (QTG) is necessary to learn more about the underlying genetic architecture, for instance, to provide a target for genome editing (Varshney et al. 2021) or to develop a genomic prediction model (Teng et al. 2020).

A QTL typically spans a large area on the genome containing dozens to hundreds of candidate QTGs. The traditional approach to identifying the actual QTG is fine-mapping, *i.e.,* increasing the resolution, using substitution mapping (Paterson et al. 1990). This approach involves crossing two lines with distinct genotypes at the QTL of interest to obtain new recombinant lines for that region. As a result, the QTL interval can be narrowed down, and the number of candidate QTGs can be reduced to one or a few (Eshed and Zamir 1995; Kooke et al. 2012). The remaining candidate QTGs can be validated using functional genomic methods, such as gene knockout or overexpression using CRISPR-Cas9 (Noman et al. 2016). Still, identifying the QTG is only a piece of the puzzle of understanding quantitative traits. A more systematic approach is needed to dissect the regulation of quantitative traits at a molecular level.

1.3. Systems genetics to study quantitative traits

The development of omics technologies in the past decade resulted in the emergence of a new field of genetics: systems genetics (Acquaah 2012). Systems genetics aims to understand the flow of genetic information from DNA sequences to traits (Figure 1.1). This new genetic approach models and integrates intermediate molecular traits, such as RNA, metabolite, and protein expression (Civelek and Lusis 2013). Expression levels can be used as quantitative traits in a QTL mapping approach, known as genetical genomics (Jansen and Nap 2001). The resulting QTL is assumed to contain a direct or indirect regulator of the molecule's expression. Integrating multiple levels of these expression QTLs may help us understand trait regulation at a molecular level. An example of a systems genetics study in Arabidopsis is the work of Wentzell et al. (2007), who mapped QTLs for transcript, protein, and metabolite levels and, based on the co-location of these different levels of QTLs on the genome and on pathway information, which forms a hypothesis that an insect resistance trait is caused by secondary metabolites called aliphatic glucosinolates produced by the *AOP* gene (Jansen et al. 2009).

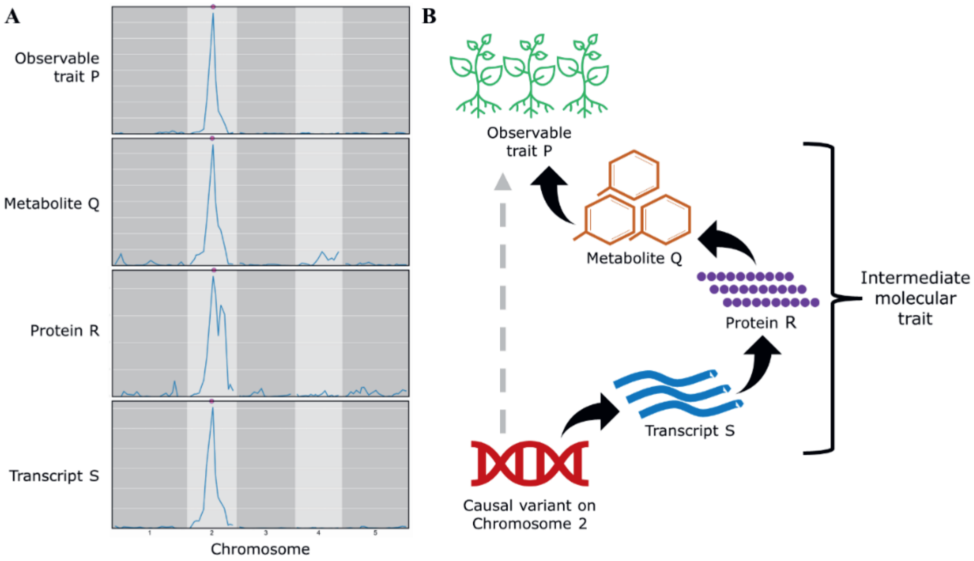


Figure 1.1. A systems genetics approach to identify causal genes and molecular mechanisms underlying a trait. This hypothetical example shows the collocation of the observable trait QTL with the molecular phenotype (metabolite, protein, and transcript) QTLs (A). Based on this case, a hypothesis, potentially supported by prior biological knowledge, can be made regarding the involvement of molecular traits in regulating the observable trait (B).

An example of genetical genomics is mapping a gene expression QTL (eQTL). Gene expression is analysed in high throughput using (previously) microarrays or (currently) RNA sequencing (RNA-seq). The eQTL mapping process is similar to that for observable or organism-level phenotypes (*e.g.,* yield, disease resistance, and seed germination): associating genetic markers with, in this case, transcript levels. The identified eQTLs can be categorised based on the location of the causal variants (Figure 1.2). A *cis*-eQTL is caused by variation in the expressed gene itself, for instance, in a *cis*-regulatory element (e.g., promoter), while for a *trans*-eQTL the variation is in a *trans*-regulatory element (*e.g.,* transcription factor) (Rockman and Kruglyak 2006; Brem et al. 2002). However, the low density of markers typically used in eQTL mapping and the low recombination frequency in the population prevents the precise identification of the actual causal variant. Therefore, the classification of an eQTL is usually based on its location relative to the gene encoding the transcripts, calling an eQTL local if it is near the expressed gene (± 1 Mb to the most significant marker) and distant if located elsewhere on the genome (Rockman and Kruglyak 2006; Brem et al. 2002).

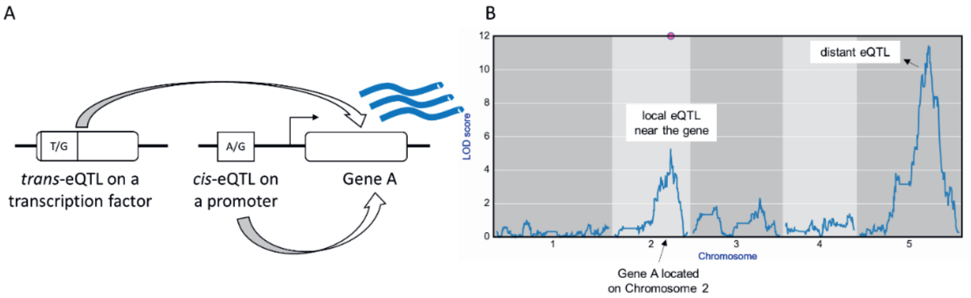


Figure 1.2. eQTLs can be categorised based on: A) the regulatory mechanism, into *cis*-eQTL (causal variant affecting *cis*-regulatory elements, such as a promoter) or *trans*-eQTL (causal variant affecting trans-regulatory elements, such as a transcription factor) and B) the location of the most significant marker: local eQTL if near the transcript-encoding gene (on chromosome 2) and distant if elsewhere (on chromosome 5). The LOD score on the y-axis shows the probability estimate that the eQTL is present in a particular genome location.

Most eQTL mapping studies in plants were done using the model plant *Arabidopsis thaliana* (Keurentjes et al. 2007; West et al. 2007; Cubillos et al. 2012; Snoek et al. 2012; Lowry et al. 2013; Hartanto et al. 2020). In addition to its small genome, short lifecycle, mutant availability, and accumulated body of knowledge (Lamesch et al. 2012; Arabidopsis Genome 2000; Koornneef and Meinke 2010), Arabidopsis is particularly favourable for genetical genomics studies due to the availability of recombinant inbred lines (RILs) as mapping populations (Loudet et al. 2002; Alonso-Blanco et al. 1998). Arabidopsis eQTL data are accessible in the AraQTL online analysis platform (https://www.bioinformatics.nl/AraQTL/, Nijveen et al. (2017)), allowing the research community to conduct follow-up analyses and explore potential gene regulatory interaction.

eQTL data potentially are a rich source of gene regulatory information useful to unravel molecular machinery underpinning plant traits. For example, Keurentjes et al. (2007) constructed a regulation network of genes likely involved in the transition to flowering using regulator-target pairs inferred from eQTL data based on gene expression correlation. However, the correlation-based method does not always lead to identifying the correct regulator(s), and the lack of reliable methods to identify causal genes obstructs further exploration of eQTL data. Like organism-level phenotype QTL analysis, eQTL analysis suffers from low mapping resolution, resulting in many candidate causal genes (eQTGs) per eQTL. For the model plant *Arabidopsis thaliana*, many eQTL data sets have been generated, each consisting of thousands of eQTLs. However, the actual eQTG was identified for only a few of these eQTLs (Jimenez-Gomez et al. 2010; Terpstra et al. 2010; Lowry et al. 2013). The same holds for other model species with a large number of eQTL studies, like *Caenorhabditis elegans* (Evans et al. 2021) and yeast (Albert et al. 2018). Since experimental fine-mapping thousands of eQTLs to extend the list of identified eQTG is currently not feasible, in this thesis, I explore *in silico* alternatives to prioritise candidate genes underlying eQTLs.

1.4. Candidate gene prioritisation using prior knowledge

Experimental validation to determine the causal gene is often costly and laborious, so starting with the most promising candidate genes in the region is advisable. To do this, researchers can rank the candidate genes based on defined criteria. This problem is known as gene prioritisation.

Researchers have developed many gene prioritisation methods, often accompanied by benchmarking studies to compare and evaluate their performance. Most of these methods were developed in the context of human disease, and only a few apply to plants or other organisms. These gene prioritisation methods vary in their application (*e.g.,* prioritising potential regulators in general or for specific diseases/traits), data sources, approaches, and assumptions about the causal genes (Moreau and Tranchevent 2012; Seyyedrazzagi and Navimipour 2017; Zolotareva and Kleine 2019). Still, no single gene prioritisation method outperforms the others in all cases (Bornigen et al. 2012). In general, the commonality of these tools is the use of prior biological knowledge of the traits as input to rank candidate genes. This prior knowledge is utilised either in the form of keywords describing the trait (*e.g.,* "seed germination" or "light response") that match the description/annotation of the causal gene or known gene-trait associations based on literature. These associations are then used to derive a particular type of scoring or selection system and suggest the most likely causal genes.

The following are the most common types of prior knowledge used by gene prioritisation tools:

Gene functional annotations: functional annotation associates the gene function in the organisms. Gene Ontology (GO) is the most well-known and comprehensive annotation source for many organisms, including Arabidopsis. GO terms form a comprehensive functional annotation categorised into biological process, molecular function, and cellular component at different levels of detail, organised in a hierarchical graph. For model organisms like Arabidopsis, genes are annotated with GO terms derived from published experiments. In addition, sequence homology and other computational methods are used to assign inferred GO terms to unannotated genes (Gene Ontology 2021).

Besides the GO, various publicly available databases contain functional annotations of genes, for example, linking them to biochemical pathways producing specific chemical compounds. One of the most comprehensive and well-maintained databases in this area is the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al. 2022). KEGG integrates sixteen databases of various biological entities, including genomic and biochemical pathway information, manually curated from published literature. Besides KEGG, AraCyc is a biochemical pathway database for

Arabidopsis. The functional annotation data in AraCyc is primarily derived computationally based on existing annotations (Rhee et al. 2006).

Protein-protein interactions: proteins interact in various ways as part of their biological functions. Identifying these so-called protein functional interactions is essential to understand the molecular basis underlying traits. Over the past years, many experiments have been performed to study protein-protein interactions, and several databases are dedicated to hosting these interaction data inferred from different evidence types. For Arabidopsis, protein-protein interactions can be retrieved from databases like STRING (Szklarczyk et al. 2021), AraNet V2 (Lee et al. 2015b) and GeneMania (Warde-Farley et al. 2010). These databases predict functional interaction based on different categories of evidence: physical interaction, genome comparison, co-expression, experiments, other databases, and literature.

Gene expression regulatory information: transcription factors (TFs) are key gene expression regulators that can act as causal genes (Albert et al. 2018). Information about TFs and their targets is essential to understand the transcriptional regulation of a specific biological process. A TF regulates its target genes by binding short DNA sequences located in promoters or enhancers of genes, called TF binding sites (TFBSs). High-throughput methods like *in vivo* chromatin immunoprecipitation (ChIP) (*e.g.,* in Chen et al. 2018), DNA affinity purification (DAP) (*e.g.,* in O'Malley et al. 2016) sequencing or SELEX-seq (Riley et al. 2014) allow the identification of TFBSs and potential targets for a certain TF. Curated TFBS sequences taken from the literature are accessible in databases like JASPAR (Castro-Mondragon et al. 2022) and TRANSFAC (Wingender et al. 2000). Furthermore, several databases (*e.g.,* PlantRegMap (Tian et al. 2020) and AtRegNet (Yilmaz et al. 2011)) present inferred links between TFs and their target genes, which can be used to construct gene regulation networks.

Sequence information: sequence properties of genes or proteins can be used to distinguish trait or disease-causing genes from others. For example, Lopez-Bigas and Ouzounis (2004) discovered that disease proteins are longer, have more homologous genes in other species, and are more conserved than the average human protein. In plants, causal QTL genes in Arabidopsis and rice tend to have higher paralog copy numbers and more SNPs resulting in premature stop codons (Lin et al. 2019). Sequence information can also be used to find potentially causal genes based on similarity to sequences of previously validated disease genes. This approach is used by Endeavour, which identifies potential new disease genes based on the similarity to protein sequences and TFBSs of known disease genes (Aerts et al. 2006).

Some methods were explicitly developed to prioritise candidate genes on QTL in plants. Bargsten et al. (2014) proposed a prioritisation method based on the assumption that the different causal genes on multiple QTLs of a trait are involved in the same biological process. They analysed the overrepresentation of Biological

Process Gene Ontology terms in the set of candidate genes on several QTL intervals of many rice traits. Next, they identify the likely candidate genes based on the annotation with these overrepresented GO terms. Removing genes not annotated with the overrepresented GO terms could drastically reduce the number of candidate genes for a QTL.

Gene functional annotation is also exploited in the QTL search tool (Warwick Vesztrocy et al. 2018), which combines it with hierarchical orthologous group data from the Orthologous Matrix project (Altenhoff et al. 2018) to prioritise candidate genes in Arabidopsis and rice. Functional annotations (e.g., from GO or ChEBI) associated with the trait under study are used as input for this tool. QTLsearch finds candidate genes on the QTL, of which homologs are associated with the trait-related annotations. It subsequently assigns a score and p-value representing the association of the candidate gene to the functional annotation. Researchers can use this score to rank the candidate genes or apply a threshold based on the p-value to reduce the number of candidate genes. Despite their capability to prioritise potential causal genes in the given examples, QTLsearch as well as the method developed by Bargsten et al. (2014) rely on functional annotations to link traits with their candidate genes. Incomplete annotation data or a lack of biological knowledge of the trait when using QTLsearch, may result in many false negatives.

Compared to these two QTL gene prioritisation tools, which only use a few data sources, two machine learning approaches called QTG-Finder and QTG-Finder2 use multiple data sources for QTL gene prioritisation in Arabidopsis and several other plant species: sequence, homology, and functional annotation (Lin et al. 2019; Lin et al. 2020). A machine learning algorithm uses these data to infer the properties of known QTL causal genes and predict new ones in the genomes. QTG-Finder and QTG-Finder2 output a probability of a gene being causal for a QTL that researchers can use to rank candidate genes. Unlike the trait-specific tools mentioned above, QTG-Finder and QTG-Finder2 prioritise causal genes in general, regardless of the trait, *i.e.,* specific information about the trait is not used. Despite the good performance in ranking genes not used in training the algorithm, the model's complexity restricts the interpretation of the gene prioritisation process, which prevents us from studying the molecular mechanism of the causal genes in trait regulation.

KnetMiner takes a different approach by integrating many kinds of prior biological knowledge in a gene prioritisation platform (Hassani-Pak et al. 2016; Hassani-Pak et al. 2021). The data sources include protein-protein interaction, phenotype-associated mutant, and genetic studies, GO annotations, GWAS data, links to relevant publications *etc*., collected from different organisms, from fungi to plants (although most species are only available in a paid version). These data were integrated using graph database management systems KnetBuilder and rdf2neo to construct a genome-scale graph-structured knowledge base (knowledge graph) connecting genes to functions, homologous genes, related phenotypes etc. Like QTLsearch, researchers must provide KnetMiner keywords associated with the trait under study, based on which it then finds related genes and scores these based on their specificity, distance, and relevance. The use of prior knowledge in knowledge graphs is a significant advantage of KnetMiner over other tools, allowing users to inspect the reasoning underlying the prioritisation. However, it is also a limitation as the prioritisation only works if knowledge is available. Combining the graph database with a prediction (*e.g.,* using machine learning) is a potential solution, allowing the identification of new causal genes while maintaining interpretability.

Gene prioritisation methods above work well to prioritise candidate genes on plant QTLs in different ways. The method developed by Bargsten et al. (2014) works by filtering likely candidates based on GO terms, resulting in a ten-fold reduction in the number of candidates in rice QTLs. QTLsearch was able to identify at least one candidate gene for 76 out of 141 published metabolite QTL (mQTL) compared to a naïve BLAST approach as a benchmark (72 mQTLs), given that the trait can be associated with ChEBI terms (Warwick Vesztrocy et al. 2018). QTG-Finder2 can recall 64% of Arabidopsis and 83% of rice-known causal genes when the top 20% of ranked genes are considered. KnetMiner can also rank candidate genes based on keywords, providing evidence supporting prioritisation.

Nevertheless, we are yet to see a gene prioritisation method that allows us to elucidate the underlying molecular mechanism leading to the trait's variation. I believe the systems genetics approach using genetical genomics data is a key to developing such a system-wide gene prioritisation method. It requires molecular QTL data (*e.g.*, eQTL and mQTL), which can be used to infer regulatory interaction on QTL but have not been fully exploited in gene prioritisation. A systems genetics approach is the central theme in this thesis, and molecular QTL data is the building block for developing a gene prioritisation method. This thesis aims to use molecular QTL data to link genes to traits in Arabidopsis.

1.5. Thesis scope and outline

This thesis describes various data-driven methods to efficiently and rationally prioritise causal genes using a systems genetics approach and molecular QTL data. Chapter 2 presents a case study of eQTL mapping in four stages of seed germination. The result complements previous phenotype and metabolite QTL data and shows hotspots of collocating different QTL types. Co-expression network analysis using different network inference methods leads to the prioritisation of candidate causal genes at QTL hotspots. Chapter 3 describes the extension of the QTG-Finder method to allow prioritisation of eQTGs, by adding gene structure, protein interaction, and gene expression features relevant to gene expression regulation. In Chapter 4, I develop a gene prioritisation method using knowledge graphs of prior biological knowledge, including gene-gene interactions inferred from various data sources, such as functional annotation, protein-protein interaction, and transcription factor binding sites. I validate this prioritisation approach using known causal genes at single eQTLs and eQTL hotspots and demonstrate its capability to identify causal genes of organism-level traits. Chapter 5 describes a computational fine-mapping method that provides a quick, low-cost alternative to traditional fine-mapping, suitable for prioritising candidate genes on multi-omics QTL hotspots. Finally, in Chapter 6, I discuss my main findings and challenges, and present an outlook for the future of plant genetic improvement using a systems genetic approach.

Chapter 2.
Network analysis prioritizes *DEWAX* and *ICE1* as the candidate genes for major eQTL hotspots in seed germination of *Arabidopsis thaliana*

Seed germination is characterized by a constant change of gene expression across different time points. These changes are related to specific processes, which eventually determine the onset of seed germination. To get a better understanding on the regulation of gene expression during seed germination, we performed a quantitative trait locus mapping of gene expression (eQTL) at four important seed germination stages (primary dormant, after-ripened, six-hour after imbibition, and radicle protrusion stage) using *Arabidopsis thaliana* Bay x Sha recombinant inbred lines (RILs). The mapping displayed the distinctness of the eQTL landscape for each stage. We found several eQTL hotspots across stages associated with the regulation of expression of a large number of genes. Interestingly, an eQTL hotspot on chromosome five collocates with hotspots for phenotypic and metabolic QTLs in the same population. Finally, we constructed a gene co-expression network to prioritize the regulatory genes for two major eQTL hotspots. The network analysis prioritizes transcription factors DEWAX and ICE1 as the most likely regulatory genes for the hotspot. Together, we have revealed that the genetic regulation of gene expression is dynamic along the course of seed germination.

This chapter is published as:

Hartanto, M., R.V.L. Joosen, B.L. Snoek, L.A.J. Willems, M.G. Sterken, D. de Ridder, H.W.M., Hilhorst, W. Ligterink, H. Nijveen.
Network Analysis Prioritizes DEWAX and ICE1 as the Candidate Genes for Major eQTL Hotspots in Seed Germination of *Arabidopsis thaliana*.
G3 Genes|Genomes|Genetics, Volume 10, Issue 11, 1 November 2020, Pages 4215–4226, https://doi.org/10.1534/g3.120.401477

2.1. Introduction

Seed germination involves a series of events starting with the transition of *quiescent* to physiologically active seeds and ends with the emergence of the embryo from its surrounding tissues. Germination is initiated when seeds become imbibed by water, leading to the activation of seed physiological activities (Bewley et al. 2013b; Nonogaki et al. 2010). Major metabolic activities occur after seeds become hydrated, for example, restoration of structural integrity, mitochondrial repair, initiation of respiration, and DNA repair (Bewley et al. 2013b; Nonogaki et al. 2010). For some species such as *Arabidopsis thaliana*, germination can be blocked by seed dormancy. Dormant seeds need to sense and respond to environmental cues to break their dormancy and complete germination. In *Arabidopsis thaliana*, seed dormancy can be alleviated by periods of dry after-ripening or moist chilling (Bewley et al. 2013b). Soon after dormancy is broken, the storage reserves are broken down, and germination-associated proteins are synthesized. Lastly, further water uptake followed by cell expansion leads to radicle protrusion through endosperm and seed coat, which marks the end of germination (Bewley et al. 2013b).

A major determinant for the completion of seed germination is the transcription and translation of mRNAs. The activity of mRNA transcription is low in dry, mature seeds (Comai and Harada 1990; Leubner-Metzger 2005), and drastically increases after seeds become rehydrated (Bewley et al. 2013a). Nevertheless, stored mRNAs of more than 12,000 genes with various functions are already present in dry seeds. These mRNAs are not only remnants from the seed developmental process, but also mRNAs for genes related to metabolism as well as protein synthesis and degradation required in early seed germination (Nakabayashi et al. 2005; Rajjou et al. 2004). Later in after-ripened seeds, only a slight change in transcript composition was detected compared to the dry seeds (Finch-Savage et al. 2007). The major shift in transcriptome takes place after water imbibition (Nakabayashi et al. 2005). Interestingly, the transcriptome at the imbibition stage depends on the status of dormancy. For non-dormant seeds, most of the transcripts are associated with protein synthesis, while for dormant seeds, the transcripts are dominated by genes associated with stress-responses (Finch-Savage et al. 2007; Buijs et al. 2019). Even the transcript composition in primary dormant seeds, which occurs when the dormancy is initiated during development, is different from that of secondary dormant seeds, which occurs when the dormancy is reinduced (Cadman et al. 2006). These findings show the occurrence of phase transitions in transcript composition along the course from dormant to germinated seed.

As omics technology becomes more widely available, several transcriptomics studies in seed germination processes have been conducted on a larger-scale. More developmental stages, i.e., stratification and seedling stage, and even spatial analyses have been included in these studies, resulting in the identification of gene co-expression patterns as well as the predicted functions of hub-genes (Narsai et al. 2011; Silva et al. 2016; Dekkers et al. 2013; Bassel et al. 2011). Through guilt-by-association, these co-expression based studies can be used for the identification of regulatory genes that are involved in controlling the expression of downstream genes. These regulatory genes can be subjected to further studies by reverse genetics to provide more insight into the molecular mechanisms of gene expression in seed germination (i.e., Silva et al. 2016). Nevertheless, this approach still has limitations. Uygun et al. (2016) argued that co-expressed genes do not always have similar biological functions. On the other hand, genes involved in the same function are not always co-expressed since gene expression regulation could be the result of post-transcriptional or other layers of regulation (Lelli et al. 2012). Further, Uygun et al. (2016) emphasized the importance of combining the expression data with multiple relevant datasets to maximize the effort in the prioritization of candidate regulatory genes.

Genetical genomics is a promising approach to study the regulation of gene expression by combining genome-wide expression data with genotypic data of a segregating population (Jansen and Nap 2001). To enable this strategy, the location of markers associated with variation in gene expression is mapped on the genome, which results in the identification of expression quantitative trait loci (eQTLs). Relative to the location of the associated gene, the eQTL can be locally or distantly mapped, known as local and distant eQTLs (Rockman and Kruglyak 2006; Brem et al. 2002). Local eQTLs mostly arise because of variations in the corresponding gene or a cis-regulatory element. In contrast, distant eQTLs typically occur due to polymorphism on trans-regulatory elements located far away from the target genes (Rockman and Kruglyak 2006). Therefore, given the positional information of distant eQTLs, one can identify the possible regulators of gene expression. However, the eQTL interval typically spans a large area of the genome and harbors hundreds of candidate regulatory genes. A large number of candidate genes would cause the experimental validation (e.g. using knockout or overexpression lines) to be costly and take a long time. Therefore, a prioritization method is needed to narrow down the list of candidate genes underlying eQTLs, particularly on distant eQTL hotspots. A distant eQTL hotspot is a genomic locus where a large number of distant eQTLs are collocated (Breitling et al. 2008). The common assumption is that the hotspot arises due to one or more polymorphic master regulatory genes affecting the expression of multiple target genes (Breitling et al. 2008). Therefore, the identification of master regulatory genes becomes the center of most genetical genomics studies as the findings might improve our understanding of the regulation of gene expression (i.e., in Keurentjes et al. 2007; Jimenez-Gomez et al. 2010; Sterken et al. 2017; Valba et al. 2015; Terpstra et al. 2010).

In this study, we carried out eQTL mapping to reveal loci controlling gene expression in seed germination. To capture whole transcriptome changes during seed germination, we included four important seed germination stages, which are primary dormant seeds (PD), after-ripened seeds (AR), six-hours imbibed seeds (IM), and seeds with radicle protrusion (RP). In total, 160 recombinant inbred lines (RILs) from a cross between genetically distant ecotypes Bay-0 and Shahdara (Bay x Sha) were used in this study (Loudet et al. 2002). Our results show that each seed germination stage has a unique eQTL landscape, confirming the stage-specificity of gene regulation, particularly for distant regulation. Based on network analysis, we identify the transcription factors ICE1 and DEWAX as prioritized candidate regulatory genes for two major eQTL hotspots in PD and RP, respectively. Finally, the resulting dataset complements the previous phenotypic QTL (Joosen et al. 2012) and metabolite QTL (Joosen et al. 2013) datasets, allowing systems genetics studies in seed germination. The identified eQTLs are available through the web-based AraQTL (http://www.bioinformatics.nl/AraQTL/) workbench (Nijveen et al. 2017).

2.2. Materials and Methods

Plant materials

In this study, we used 164 recombinant inbred lines (RILs) derived from a cross between the Bay-0 and Shahdara Arabidopsis ecotypes (Loudet et al. 2002) provided by the Versailles Biological Resource Centre for Arabidopsis (http://dbsgap.versailles.inra.fr/vnat). The plants were sown in a fully randomized setup on 4x4 cm rockwool plugs (MM40/40, Groudan B. V.) and hydrated with 1 g/l Hyponex (NPK = 7:6:19, http://www.hyponex.co.jp) in a climate chamber (20°C day, 18°C night) with 16 hours of light (35 W/m2) at 70% relative humidity. Seeds from four to seven plants per RIL were bulk harvested for the experiment (see also Joosen et al. 2012; Joosen et al. 2013). The genotypic data consisting of 1,059 markers per line was obtained from Serin et al. (2017). However, the genotypic data is available only for 160 RILs; therefore, we used this number of lines for eQTL mapping.

Experimental setup

The RIL population was grouped into four subpopulations, each one representing one of the four different seed germination stages. We used the designGG-package (Li et al. 2009) in R (version 3.6.0 Windows x64) to aid the grouping so that the distribution of Bay-0 and Sha alleles between sub-populations is optimized. The first stage is the primary dormant (PD) stage when the seeds were harvested and stored at -80°C after one week at ambient conditions. The second stage is after-ripened (AR) seeds that obtained maximum germination potential after five days of imbibition by storing at room temperature and ambient relative humidity. The third stage is the 6 hours imbibition (IM) stage. For this stage, the seeds were after-ripened and imbibed for six hours on water-saturated filter paper at 20°C and immediately transferred to a dry filter paper for 1 minute to remove the excess of water. The fourth stage is the radicle protrusion (RP) stage. To select seeds at this stage, we used a binocular to observe the presence of a protruded radicle tip.

RNA isolation

Total RNA was extracted according to the hot borate protocol modified from Wan and Wilkins (1994). For each treatment, 20 mg of seeds were homogenized and mixed with 800 µl of extraction buffer (0.2M Na boratedecahydrate (Borax), 30 mM EGTA, 1% SDS, 1% Na deoxycholate (Na-DOC)) containing 1.6 mg DTT and 48 mg PVP40 which had been heated to 80°C. Then, 1 mg proteinase K was added to this suspension and incubated for 15 min at 42°C. After adding 64 µl of 2 M KCL, the samples were incubated on ice for 30 min and subsequently centrifuged for 20 min at 12,000 g. Ice-cold 8 M LiCl was added to the supernatant in a final concentration of 2 M, and the tubes were incubated overnight on ice. After centrifugation for 20 min at 12,000 g at 4°C, the pellets were washed with 750 µl ice-cold 2 M LiCl. The samples were centrifuged for another 10 min at 10,000 g at 4°C, and the pellets were re-suspended in 100 µl DEPC treated water. The samples were phenol-chloroform extracted, DNAse treated (RQ1 DNase, Promega), and further purified with RNeasy spin columns (Qiagen) following the manufacturer's instructions. The RNA quality and concentration were assessed by agarose gel electrophoresis and UV spectrophotometry.

Microarray analysis

RNA was processed for use on the Affymetrix Arabidopsis SNPtile array (atSNPtilx520433), as described by the manufacturer. Briefly, 1 mg of total RNA was reverse transcribed using a T7-Oligo(dT) Promoter Primer in the first-strand cDNA synthesis reaction. Following RNase H-mediated second-strand cDNA synthesis, the double-stranded cDNA was purified and served as a template in the subsequent in vitro transcription reaction. The reaction was carried out in the presence of T7 RNA polymerase and a biotinylated nucleotide analog/ribonucleotide mix for complementary RNA (cRNA) amplification and biotin labeling. The biotinylated cRNA targets were then cleaned up, fragmented, and hybridized to the SNPtile array. The hybridization data were extracted using a custom R script with the help of an annotation-file based on TAIR10. Intensity data were log-transformed and normalized using the *normalizeBetweenArrays* function with the quantile method from Bioconductor package limma (Ritchie et al. 2015). Then, for each annotated gene, the log-intensities of anti-sense exon probes were averaged.

Clustering analysis

Principal component analysis for log-intensities of all parents and RIL population samples was done using the pr.comp function in R where the unscaled log intensities are shifted to be zero centered. For hierarchical clustering, we only selected genes with a minimal fold change of 2 between any pair of consecutive stages (PD to AR, AR to IM, or IM to RP). Then, the distance matrices of filtered genes and all samples were calculated using the absolute Pearson correlation. These matrices were clustered using Ward's method. We manually set the number of clusters to 8 and performed gene ontology enrichment for each of the clusters using the weight algorithm of the topGO package in R and used 29,913 genes detected by hybridization probes as the background (Alexa et al. 2006).

eQTL mapping

For eQTL mapping, we used 160 RILs separated into four subpopulations, each representing one specific seed germination stage. For each stage separately, eQTLs were mapped using a single-marker model, as in Sterken et al. (2017). The gene expression data were fitted to the linear model

$$y_{i,j} \sim x_j + e_j$$

where y is the log-intensity representing the expression of a gene i (i = 1, 2, ..., 29,913) of RIL j (j = 1, 2, ..., 160) explained by the parental allele on marker location x (x = 1, 2, ..., 1,059). The random error in the model is represented by e_j.

To account for the multiple-testing burden in this analysis, we determined the genome-wide significant threshold using a permutation approach (e.g. see Sterken et al. 2017). A permuted dataset was created by randomly distributing the log-intensities of the gene under study over the genotypes. Then, the previous eQTL mapping model was performed on this permuted dataset. This procedure was repeated 100 times for each stage. The threshold was determined using:

$$\frac{FDS}{RDS} \leq \frac{m_0}{m} q.\, log(m)\,,$$

where, at a specific significance level, the false discoveries (FDS) were the averaged permutation result, and real discoveries (RDS) were the outcome of the eQTL mapping using the unpermuted dataset. The number of true hypotheses tested (m_0) was 29,913 - $RDS,$ and the number of hypotheses (m) tested was the number of genes, which was 29,913. For the q-value, we used a threshold of 0.05. As a result, we got a threshold of 4.2 for PD and AR, 4.1 for IM, and 4.3 for RP.

The confidence interval of an eQTL was determined based on a -log$_{10}$(p-value) drop of 1.5 compared to the peak marker (as in Keurentjes et al. 2007; Cubillos et al. 2012). We determine an eQTL as local if the peak marker or the confidence interval lies within 1 Mb or less from the target gene location (as in Cubillos et al. 2012). All eQTLs that did not meet this criterion were defined as distant.

We defined a region as an eQTL hotspot if the number of distant-eQTLs mapped to a particular genomic region significantly exceeded the expectation. First, we divided the genome into bins of 2 Mb. Then, we determined the expected number of distant-eQTLs per genomic bin by dividing the total number of distant-eQTLs by the total number of bins. Based on a Poisson distribution, any bin having an actual number of distant-eQTLs larger than expected ($p < 0.0001$) was then considered as an eQTL hotspot.

Gene regulatory network inference and candidate genes prioritization of eQTL hotspot

We used a community-based approach to infer regulatory networks of genes with an eQTL on a hotspot location using expression data. In this approach, we assume the hotspot is caused by a polymorphism in or near one or more regulatory genes causing altered expression that can be detected as a local eQTL (Joosen et al. 2009; Breitling et al. 2008; Jimenez-Gomez et al. 2010; Serin et al. 2017). Based on this assumption, we labeled all genes with a local eQTL on a hotspot as candidate regulators and genes with a distant eQTL as targets. The expression of these genes was subjected to five different network inference methods to predict the interaction weight. The methods used were TIGRESS (Haury et al. 2012), Spearman correlation, CLR (Faith et al. 2007), ARACNE (Margolin et al. 2006), and GENIE3 (Huynh-Thu et al. 2010). The predictions from GENIE3 were used to establish the direction of the interaction by removing the one that has the lowest variable importance to the expression of the target genes between two pairs of genes. For instance, if the importance of gene$_i$ – gene$_j$ is smaller than gene$_j$ – gene$_i$, then the former is removed. By averaging the rank, the predictions of all inference methods were integrated to produce a robust and high performance prediction (Marbach et al. 2012). The threshold was determined as the minimum average rank where all nodes are included in the network. Finally, the network was visualized using Cytoscape (version 3.7.1) (Shannon et al. 2003), and network properties were calculated using the NetworkAnalyzer tool (Assenov et al. 2008). The candidate genes for each eQTL hotspot were prioritized based on their outdegree and closeness centrality (Pavlopoulos et al. 2011).

Data availability

The list of genetic markers, genotype, and gene expression data used in this study are given in Table S2.7, Table S2.8, and Table S2.9, respectively. Cel files of microarray data have been deposited in the ArrayExpress database at EMBL-EBI (http://www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-9080. The phenotype and metabolite measurement can be found in Table S2.10 and Table S2.11. The list of differentially expressed genes is in Table S2.12. All QTL mapping results are given in Table S2.13 (expression QTL), Table S2.14 (phenotype QTL), and Table S2.15 (metabolite QTL). The code for the analysis and visualization is available in the form of R scripts at the Wageningen University GitLab repository (https://git.wur.nl/harta003/seed-germination-qtl).

2.3. Results

Major transcriptional shifts take place after water imbibition and radicle protrusion

To visualize the transcriptional states of the parental lines and the RILs at the four seed germination stages, we performed a principal component analysis using the log-intensities of all expressed genes (Figure 2.1). The first principal component explains 55.6% of the variation and separates the samples into three groups. Germination progresses from left to right with the PD and AR seeds grouping together, indicating that the after-ripening treatment does not induce a considerable change in global transcript abundance. The large-scale transcriptome change only happens after water imbibition and radicle protrusion. This event was also observed by Finch-Savage et al. (2007) and Silva et al. (2016). The second principal component on the PCA explains 14.2% variance in the data and separates the RILs within each of the three clusters but not the parents. The source of this variation may be the genetic variation among samples and shows transgressive segregation of gene expression in RILs due to genetic reshuffling of the parental genomes during crossing and generations of selfing.

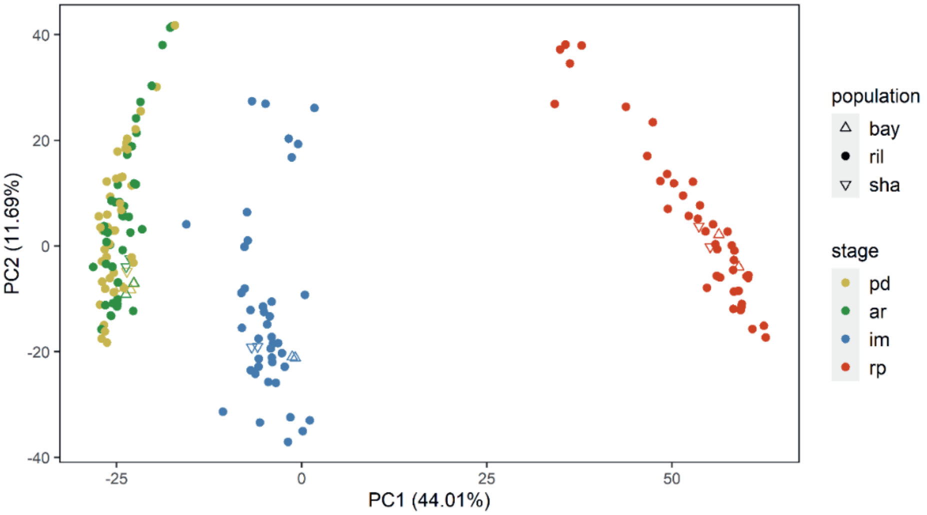


Figure 2.1. Principal component plot derived from transcriptome measurements of 164 RILs, and the Bay-0 and Sha parental lines taken at primary dormant seed (PD), after-ripened seed (AR), six-hours after imbibition (IM), and at the time when the radicle is protruded (RP).

To identify specific expression patterns among genes in the course of seed germination, we performed an additional analysis of the transcriptome data using hierarchical clustering (Figure 2.2). For this analysis, we only selected the 990 genes with a minimal fold change of two between any two consecutive stages (PD to AR, AR to IM, IM to RP). We then clustered both the genes and the seed samples. As shown in the figure, the clustering of samples shows similar grouping as in the previous PCA plot; three clusters were formed with one cluster containing both PD and AR, while IM and RP form separate clusters.

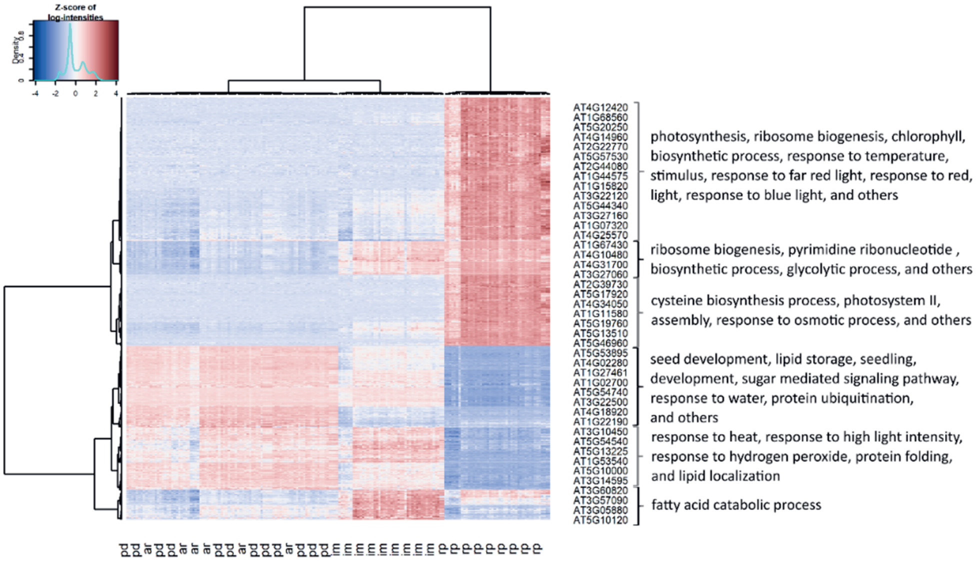


Figure 2.2. Hierarchical clustering of Bay-0, Sha, and 164 RILs transcriptome samples measured at four different seed germination stages (top) and 990 genes differentially expressed between two consecutive stages (left). Listed genes are the sample of genes for each cluster. Some enriched gene ontology terms for gene clusters are listed on the right.

The clustering of genes shows at least three distinctive gene expression patterns. In the first pattern, transcript abundance is highest in the last stage, radicle protrusion. A GO enrichment test suggests that transcripts with this expression pattern are involved in the transition from the heterotrophic seed to the autotrophic seedling stage, with enriched processes such as photosynthesis, response to various light, and response to temperature. This is in agreement with Rajjou et al. (2004), who showed that genes required for seedling growth are expressed after water imbibition. The second pattern shows an opposite trend with higher transcript abundances in the first three stages and lower expression at the end of the seed germination process. Some of these transcripts may be the remnant of seed development since the GO term related to this process is overrepresented. Moreover, transcripts involved in response to hydrogen peroxide were also overrepresented, which provides more evidence for the importance of reactive oxygen species in seed germination (for review see Wojtyla et al. 2016). The last pattern represents genes that are upregulated at the IM stage. Genes with this pattern are functionally enriched in the catabolism of fatty acids, a likely source of energy for seedling growth (Bewley et al. 2013c). Altogether, these results suggest that co-expression patterns of genes reflect particular functions during the seed germination process.

Distant eQTLs explain less variance than local eQTLs and are more specific to a seed germination stage

To map loci associated with gene expression levels, we performed eQTL mapping of 29,913 genes for each seed population representing four seed germination stages (Table 2.1). We found eQTLs, numbers ranging from 1,335 to 1,719 per stage (FDR = 0.05), spread across the genome. Among the genes with an eQTL, only a few (less than 1%) had more than one. We then categorized the eQTLs into local and distant based on the distance between the target gene and the eQTL peak marker or the confidence interval. Based on this criterion, over 72% of the eQTLs per stage were categorized as local (located within 1 Mb of the gene), while the remainder were distant. Although the total of the identified eQTLs was different between the stages, the ratio of distant to local eQTLs was relatively similar for all stages. We then calculated the fraction of the total variation that is explained by the simple linear regression model for each eQTL. By comparing the density distributions (Figure S2.1), we showed that local eQTLs generally explain a more substantial fraction of gene expression variation than distant eQTLs. Finally, we determined the number of specific and shared eQTLs across stages (Figure 2.3). Here, we show that distant eQTLs are more specific to seed germination stages. Local eQTLs, on the other hand, are commonly shared between two or more stages, which is in line with previous experiments showing overlapping local eQTLs and specific distant eQTLs across different developmental stages (Vinuela et al. 2010), environments (Snoek et al. 2012; Snoek et al. 2017; Lowry et al. 2013) and populations (Cubillos et al. 2012).

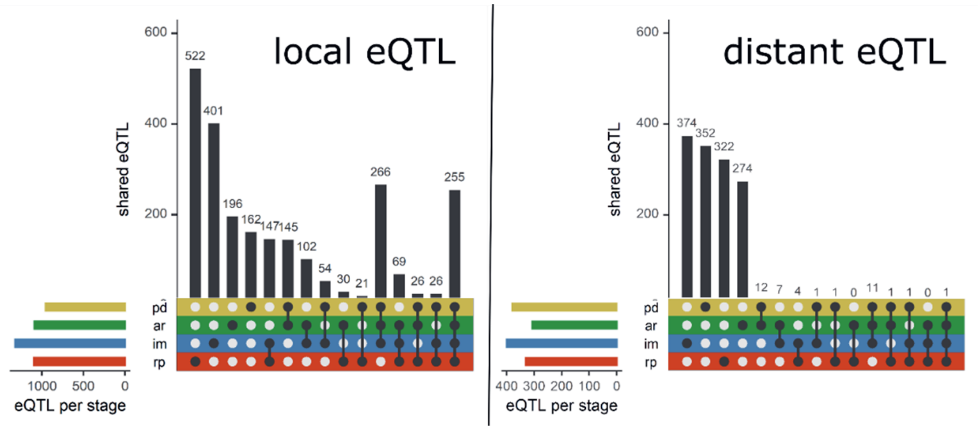


Figure 2.3. Shared local and distant eQTLs per seed germination stage.

An eQTL hotspot on chromosome 5 is associated with genes related to seed germination and collocates with multiple metabolic and phenotypic QTLs

To get an overview of how the eQTLs were mapped over the genome, we visualized the eQTL locations and their associated genes on a local/distant eQTL plot (Figure 2.4A). Here, the local eQTLs are aligned across the diagonal and spread relatively equally across the genome, while it is not the case for the distant eQTLs. Furthermore, specific loci show clustering of eQTLs, which could indicate the presence of major regulatory genes that cause genome-wide gene expression changes. We identified ten so-called (distant-) eQTL hotspots, with at least two hotspots per stage (Table 2.1). The number of distant eQTLs located within these hotspots ranges from 16 to 96. The major eQTL hotspots are PD2, IM2, and RP4, with 69, 69, and 96 distant eQTLs co-locating, respectively. Moreover, the landscape of the eQTL hotspots (Figure 2.4B) differs for every stage, including PD and AR, which is surprising since these two stages have a relatively similar transcriptome profile (Figure 2.1).

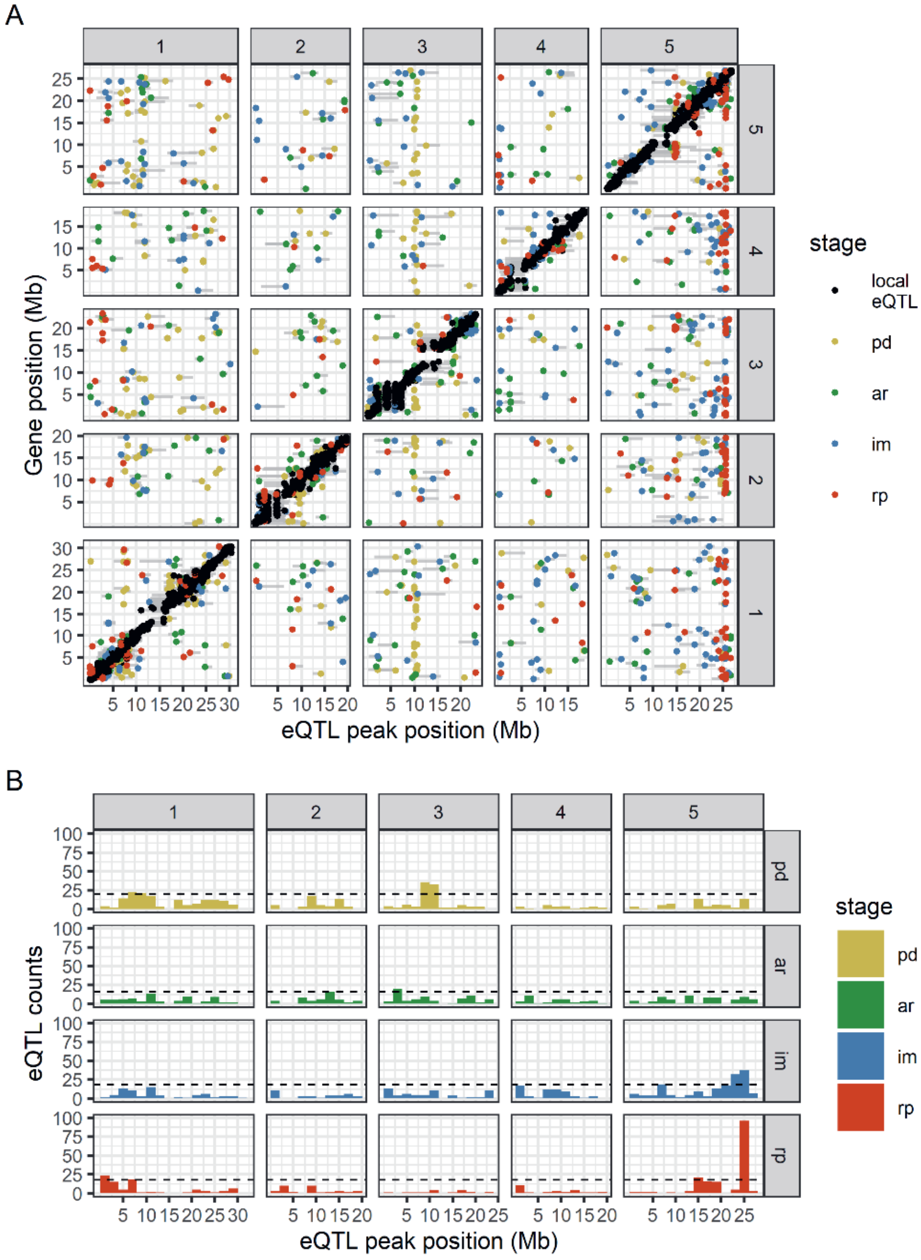


Figure 2.4. eQTL mapping from four different seed germination stages. The local-distant eQTL plot is shown on top (A). The positions of eQTLs are plotted along the five chromosomes on the x-axis and the location of the genes with an eQTL is plotted on the y-axis. The black dots (●) represent local eQTLs (located within 1 Mb of the associated gene) and the colored dots represent distant eQTLs (located far from the associated gene). The gray horizontal lines next to each dot indicate the confidence interval of the eQTL location based on a 1.5 drop in -log$_{10}$(p-value). The histogram of the number of eQTLs per genomic location is shown at the bottom (B). The horizontal dashed black lines mark the significance threshold for an eQTL hotspot.

Table 2.1. Summary of the eQTL mapping for the four different seed germination stages.

stage	eQTLs	genes with an eQTL	eQTL type	total	proportion
primary dormant	1,335	1,328	local	955	0.72
			distant	380	0.28
after-ripened	1,395	1,377	local	1,089	0.78
			distant	306	0.22
six hours after imbibition	1,719	1,702	local	1,320	0.77
			distant	399	0.23
radicle protrusion	1,426	1,418	local	1,096	0.77
			distant	330	0.23

We remapped the QTLs for previously studied seed germination phenotypes (Joosen et al. 2012) and metabolites (Joosen et al. 2013) using the RNA-seq based genetic map (Serin et al. 2017). We then visualized the resulting QTL count histograms alongside the eQTL histogram (Figure 2.5). The histogram shows that several eQTL hotspots collocate with hotspots for phenotype and metabolite QTLs (phQTLs and mQTLs, respectively). The most striking example is the collocation of QTLs on chromosome 5 around 24—25 Mb (IM2 and RP4) at the last two stages of seed germination. We performed gene ontology (GO) term enrichment analysis for genes with an eQTL mapping to these hotspots, and found ‘seed germination’ enriched among other terms (Table 2.2). These findings taken together indicate that the IM2 and RP4 hotspots harbor one or more important genes affecting gene expression during seed germination. Therefore, the identification of the regulatory gene(s) for one of these hotspots can give us more insight into the trans-regulation of gene expression during seed germination.

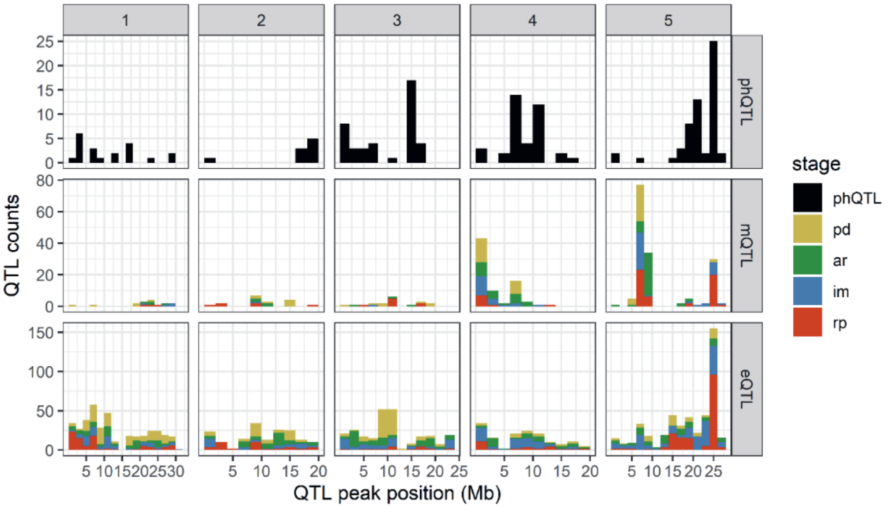


Figure 2.5. Hotspots for phQTLs, mQTLs, and eQTLs. A region of interest is located on chromosome 5 (around 24—26 Mb) where hotspots from different QTL levels collocate.

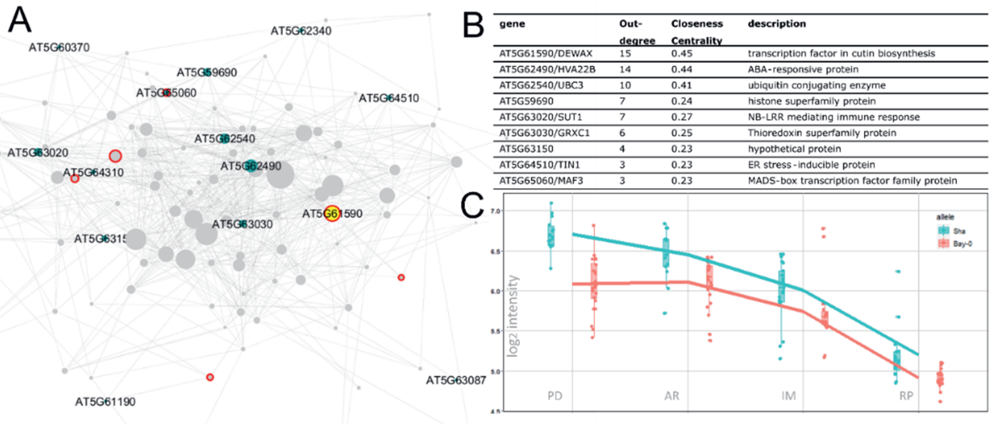
Table 2.2. Distant eQTL hotspots of the four seed germination stages. These hotspots were identified by dividing the genome into bins of 2 Mbp and performing a test to determine whether the number of distant eQTLs on a particular bin is higher than expected ($p > 0.0001$) assuming a Poisson distribution. Seed germination phenotype and metabolite data were taken from Joosen et al. (2012) and Joosen et al. (2013), respectively. Detailed information about enriched GO terms, metabolite, and phenotype can be seen on Supplemental Table S2.2 in the Supplementary Material.

hotspot ID	position	distant eQTLs	enriched GO terms	metabolite QTL	phenotype QTL
PD1	ch1:6-10 Mb	43	11	1	4
PD2	ch3:8-12 Mb	69	3	2	1
AR1	ch2:12-14 Mb	16	0	0	0
AR2	ch3:2-4 Mb	20	9	1	1
IM1	ch5:6-8 Mb	19	2	24	1
IM2	ch5:22-26 Mb	69	6	6	31
RP1	ch1:0-2 Mb	23	1	0	1
RP2	ch1:6-8 Mb	18	0	0	3
RP3	ch5:14-16 Mb	21	29	0	1
RP4	ch5:24-26Mb	96	18	20	25

Transcription factors were prioritized as the candidate genes for major eQTL hotspots

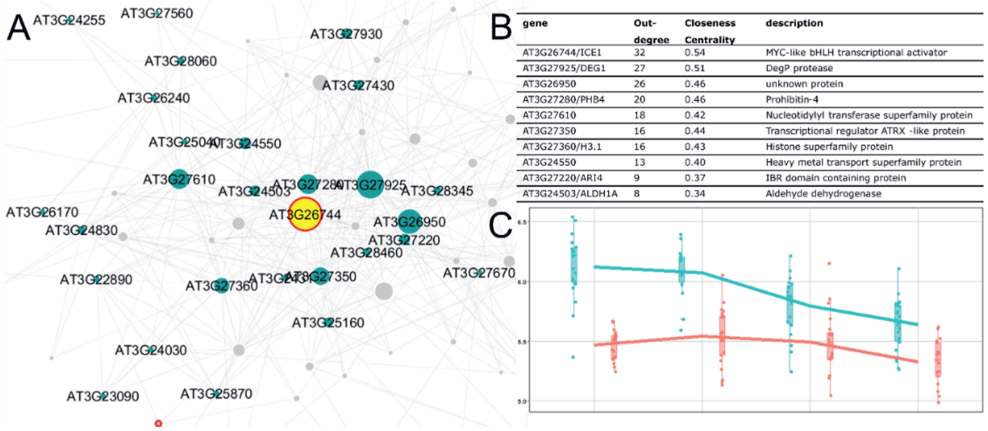
To prioritize the candidate regulatory genes underlying eQTL hotspots in this study, we constructed a network based on the expression of genes with eQTLs on the hotspot location. We built the network for two hotspots: RP4, where QTLs for expression, metabolite, and phenotype are collocated; and PD2, another major eQTL hotspot in this study. For RP4, the total number of genes used to construct the network was 116, of which 20 had a local eQTL at the hotspot, whereas for PD2, 114 genes were identified, of which 45 with a local eQTL. The genes with local eQTLs were then labeled as candidates. The networks were constructed by integrating predictions from several gene regulatory network inference methods to ensure the robustness of the result (Marbach et al. 2012). The direction of the edges in the network is predicted using the GENIE3 method (Huynh-Thu et al. 2010). For each candidate gene, we calculated the outdegree, indicating the number of outgoing edges of a gene to other genes in the network, and the closeness centrality of the candidate gene nodes, which shows the efficiency of the gene in spreading information to the rest of the genes in the network (Pavlopoulos et al. 2011). Finally, these two network properties were used to prioritize the most likely regulator of the distant eQTL hotspot.

In the resulting network, genes encoding the transcription factors *DECREASE WAX BIOSYNTHESIS/DEWAX* (*AT5G61590*), and *INDUCER OF CBP EXPRESSION 1/ICE1* (*AT3G26744*) were prioritized as the most likely candidate genes for RP4 (Figure 2.6) and PD2 (Figure 2.7), respectively. As many as 15 genes were predicted to be associated with *DEWAX* and 32 genes with *ICE1*. Note that these numbers depend on the chosen threshold; nonetheless, the current candidates are robust to changes when the parameter was changed (Table S2.3 and Table S2.4). Furthermore, these two genes also had the highest closeness centrality among the other candidates, showing that these genes have a strong influence within the network. We assessed the Bay x Sha SNP data (Genomes Consortium. Electronic address and Genomes 2016) and found several SNPs between the Bay and Sha parents in both the *DEWAX* and *ICE1* genes, including two that affect the amino acid sequence of the corresponding proteins (Table S2.5 and Table S2.6). Also, querying for *DEWAX* and *ICE1* on AraQTL showed a local eQTL for both genes in an experiment using the same RIL population on leaf tissue (West et al. 2007). This evidence supports the hypothesis that polymorphisms between the Bay and Sha alleles of *DEWAX* and *ICE1* are responsible for the steadily occurring local eQTLs at three stages (PD, IM, RP) for *DEWAX* and all four stages for *ICE1*.



gene	Out-degree	Closeness Centrality	description
AT5G61590/DEWAX	15	0.45	transcription factor in cutin biosynthesis
AT5G62490/HVA22B	14	0.44	ABA-responsive protein
AT5G62540/UBC3	10	0.41	ubiquitin conjugating enzyme
AT5G59690	7	0.24	histone superfamily protein
AT5G63020/SUT1	7	0.27	NB-LRR mediating immune response
AT5G63030/GRXC1	6	0.25	Thioredoxin superfamily protein
AT5G63150	4	0.23	hypothetical protein
AT5G64510/TIN1	3	0.23	ER stress-inducible protein
AT5G65060/MAF3	3	0.23	MADS-box transcription factor family protein

Figure 2.6. The prioritization of candidate genes for RP4 eQTL hotspot. The network of genes associated with RP4 is visualized in A. The genes in the network are represented by nodes with various sizes according to the outdegree. The unlabeled grey nodes are the targets (genes with a distant eQTL) and the labelled green nodes are the candidates (genes with a local eQTL). Nodes with a red border are transcription factors. The yellow node is *DEWAX* (*AT5G61590*). The list of top ten candidate genes for the hotspot is shown in B. The expression of *DEWAX* in 160 RILs across the four seed germination stages is visualized in C. The RILs with the Sha allele of the gene are depicted in blue, the ones with the Bay-0 allele of *DEWAX* are depicted in red.



gene	Out-degree	Closeness Centrality	description
AT3G26744/ICE1	32	0.54	MYC-like bHLH transcriptional activator
AT3G27925/DEG1	27	0.51	DegP protease
AT3G26950	26	0.46	unknown protein
AT3G27280/PHB4	20	0.46	Prohibitin-4
AT3G27610	18	0.42	Nucleotidylyl transferase superfamily protein
AT3G27350	16	0.44	Transcriptional regulator ATRX -like protein
AT3G27360/H3.1	16	0.43	Histone superfamily protein
AT3G24550	13	0.40	Heavy metal transport superfamily protein
AT3G27220/ARI4	9	0.37	IBR domain containing protein
AT3G24503/ALDH1A	8	0.34	Aldehyde dehydrogenase

Figure 2.7. The prioritization of candidate genes for the PD2 eQTL hotspot. The network of genes associated with PD2 is visualized in A. The genes in the network are represented by nodes with various sizes according to the outdegree. The unlabeled grey nodes are the targets (genes with a distant eQTL) and labelled green nodes are the candidates (genes with a local eQTL). Nodes with a red border are transcription factors. The yellow node is *ICE1* (*AT3G26744*). The list of top ten candidate genes for the hotspot is shown in B. The expression of *ICE1* in 160 RILs across the four seed germination stages is visualized in C. The RILs with the Sha allele of the gene are depicted in blue, the ones with the Bay-0 allele of *ICE1* are depicted in red.

2.4. Discussion

The function of *DEWAX* may be related to seed cuticular wax biosynthesis

In this study, we constructed a network of genes associated with the RP4 eQTL hotspot and showed that *DEWAX* was prioritized as the candidate gene for the hotspot. *DEWAX* encodes an AP2/ERF-type transcription factor that is well-known as a negative regulator of cuticular wax biosynthesis (Go et al., 2014; Suh and Go, 2014; Cui et al., 2016; Li et al., 2019) and a positive regulator of defense response against biotic stress (Froschel et al. 2019; Ju et al. 2017). This gene also seems to be involved in drought stress response (Huang et al. 2008) by inducing the expression of genes that confer drought tolerance (Sun et al. 2016), some of which (*LEA4-5*, *LTI-78*) have a distant eQTL at the RP4 hotspot. Moreover, the overexpression of *DEWAX* in Arabidopsis increases the seed germination rate (Sun et al. 2016). The role of *DEWAX* in seed germination is still unknown but may be related to cuticular wax biosynthesis.

Wax is a mixture of hydrophobic lipids, which is part of the plant cuticle together with cutin and suberin (Yeats and Rose 2013). Previous studies have demonstrated that the biosynthesis of wax in the cuticular layer of stems and leaves is negatively regulated by *DEWAX* (Go et al., 2014; Suh and Go, 2014; Cui et al., 2016; Li et al., 2019). Although the function of this gene has never been reported in seeds, the presence of a cuticular layer indeed plays a significant role in maintaining seed dormancy (Nonogaki 2019; De Giorgi et al. 2015). In Arabidopsis seeds, the thick cuticular structure covering the endosperm prevents cell expansion and testa rupture that precede radicle protrusion. Besides, this layer also reduces the diffusion of oxygen into the seed, thus preventing oxidative stress that may cause rapid seed aging and loss of dormancy (De Giorgi et al. 2015).

Besides *DEWAX*, *MUM2* is another possible regulatory gene for the RP4 hotspot based on QTL confirmation of an imbibed seed size phenotype using a heterogeneous inbred family approach (Joosen et al. 2012). In our study, we also discovered that most eQTLs on the RP4 hotspot peak at the marker located closely to the MUM2 location (Figure S2.2), which provides more evidence for this gene as the regulator for the hotspot. *MUM2* encodes a cell-wall modifying beta-galactosidase involved in seed coat mucilage biosynthesis, and the *mum2* mutant is characterized by a failure in extruding mucilage after water imbibition (Dean et al. 2007). In our analysis, *MUM2* did not have a distant eQTL on the RP4 hotspot; thus, it is not prioritized as a prominent candidate, pointing out a limitation of our approach in prioritizing candidate eQTL hotspot genes which will be discussed later. Nonetheless, we found some evidence connecting *DEWAX* to *MUM2*. First, Shi et al. (2019) found out that the mutant of *CPL2*, another gene involved in wax biosynthesis, showed a delayed secretion of the enzyme encoded by *MUM2* that disrupts seed coat mucilage extrusion. In the same study, they revealed that *CPL2* encodes a phosphatase involved in secretory protein trafficking required for the secretion of extracellular matrix materials, including wax and the cell wall-modifying enzyme *MUM2*. Although no direct connection between *DEWAX* and *CPL2* has been reported, a recent study by Xu et al. did identify *DEWAX* as a putative regulator of cell-wall-loosening *EXPANSIN* (*EXPA*) genes involved in germination (Xu et al. 2020). These findings provide a link between wax biosynthesis and cell-wall modifying enzymes, and possibly between the genes involved in these processes.

Second, the expression of *DEWAX* may be the consequence of the disruption of seed mucilage extrusion. Penfield et al. (2001) suggest that seed mucilage helps enhance water uptake to ensure efficient germination in the condition of low water potential. This is supported by the evidence that the mucilage-impaired mutant showed reduced maximum germination only on osmotic polyethylene glycol solutions (Penfield et al. 2001). Therefore, the absence of mucilage in imbibed seed under low water potential may cause osmotic stress in the seed and, in turn, induce the expression of *DEWAX*, which is known to play a role in the response of plants against osmotic stress (Sun et al. 2016). If this is the case, then a scenario could be that *DEWAX* acts downstream of *MUM2*, and the expression variation of these two genes lead to the emergence of the RP4 eQTL hotspot.

Network analysis shows the involvement of *ICE1* as a regulator of gene expression during seed germination

ICE1 is an MYC-like basic helix-loop-helix (bHLH) transcription factor that shows pleiotropic effects in plants. Earlier studies of *ICE1* mostly focus on the protein function in the acquisition of cold tolerance (Lee et al. 2005; Chinnusamy et al. 2003) and stomatal lineage development (Kanaoka et al. 2008). Recently, *ICE1* was also shown to form a heterodimer with ZOU, another bHLH transcription factor, to regulate endosperm breakdown required for embryo growth during seed development (Denay et al. 2014). At a later stage, *ICE1* negatively regulates ABA-dependent pathways to promote seed germination and seedling establishment (Liang and Yang 2015). This process involves repressing the expression of transcription factors in ABA signaling, such as *ABI3* and *ABI5*, and ABA-responsive genes, such as *EM6* and *EM1*, thus initiating seed germination and subsequent seedling establishment (Hu et al. 2019b; MacGregor et al. 2019). Loss of *ice1* has been reported to lead to reduced germination (MacGregor et al., 2019)

In this study, we performed a network analysis for genes having distant eQTLs on the PD2 hotspot and prioritized *ICE1* as the most likely regulator using network analysis. The high connectivity of *ICE1* with the other genes in the network could reflect an essential regulatory function of this gene during seed germination. However, we did not find any of the known *ICE1* target genes (i.e., *ABI3*, *ABI5*, *EM1*, and *EM6*) nor seed germination phenotype (Figure 2.5) having an eQTL at the *ICE1* locus. It could be that the *ICE1* polymorphism is not severe enough to cause considerable trait variation, especially to break a robust biological system where several buffering mechanisms exist to prevent small molecular perturbation from propagating to the phenotypic level (Signor and Nuzhdin 2018; Fu et al. 2009).

A good strategy to validate that a predicted candidate gene indeed causes a QTL hotspot would be to test one parent's allele of the gene in the genetic background of the other parent. This could be achieved by generating near-isogeneic lines, although rapid developments in site directed mutagenesis might offer a more feasible high-throughput approach for future studies. Next, being able to convert one parent's gene into the other parent's gene one SNP at a time would even allow identification of causal SNPs.

Limitations of co-expression network in identifying candidate genes of eQTL hotspots

The construction of a co-expression network is a promising approach to prioritize candidate eQTL genes (Serin et al. 2016). Despite its potential, there is a major limitation in using a co-expression network. The network is based on gene expression data; hence the identified causal genes are those that directly affect gene expression. For example, as we described above, our approach did not prioritize *MUM2* for the RP4 hotspot, possibly because the gene does not cause variation in the target gene expression but rather causes differences at another level of target gene regulation (e.g., enzyme biosynthesis) between two parental alleles in the RIL population. Other studies reported similar results where a known causal gene was not detected as a hub in the network (Jimenez-Gomez et al. 2010; Sterken et al. 2017). To overcome this, future work should focus on networks that are built upon multi-omics data by including metabolic, proteomic, and, more importantly, phenotypic measurement data (Hawe et al. 2019). Moreover, prior biological knowledge, including protein-protein interaction (Szklarczyk et al. 2017), transcription factor binding-site (Kulkarni et al. 2018), and other types of interactions (for review see Kulkarni and Vandepoele 2019) can be incorporated to construct data-driven interaction networks. Nevertheless, our approach offers a simple and straightforward way to prioritize candidate genes underlying eQTL hotspots from a limited amount of resources.

Data availability

The code and data for the analysis and visualization are available at the Wageningen University GitLab repository (https://git.wur.nl/harta003/seed-germination-qtl). Supplemental materials are available at figshare: https://doi.org/10.25387/g3.12844358.

Chapter 3. Prioritizing candidate eQTL causal genes in Arabidopsis using random forests

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 top 20% rank. 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.

This chapter is published as:

Hartanto, M., A.A. Sami, D. de Ridder, and H. Nijveen.
Prioritizing Candidate eQTL Causal Genes in Arabidopsis using Random Forests.
G3 Genes|Genomes|Genetics, Volume 12, Issue 11, November 2022,
jkac255, https://doi.org/10.1093/g3journal/jkac255

3.1. 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 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 co-expression 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 centrality (*i.e.*, the number of genes interacting with a candidate) or closeness centrality (*i.e.*, the average path length between a candidate and all other genes) (Serin et al. 2016; Hartanto et al. 2020). 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 co-expression 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. Three of 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.

3.2. 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

   We previously showed that different stages of seed germination each have a unique eQTL landscape pointing to stage-specific regulators (Hartanto et al. 2020). This indicates that variation in gene expression may help distinguish eQTL causal genes from other (non-causal) genes. We, therefore, generate seven features based on the average and standard deviation of gene expression across different tissues, accessions, and conditions (control vs. treatments):

   a. Tissues

      We downloaded RNA-seq data for nine different tissues (flower, root, male organ, seeds, female organ, stem, leaf, apical meristem, and root meristem) from CoNekT (http://www.evorepro.plant.tools/) (Julca et al. 2020). For each gene, the standard deviation is calculated and used as a feature ("SD exp. across tissues").

   b. Accessions

      We used RNA-seq data measured in seedlings of nineteen different Arabidopsis accessions (Zu-0, Wu-0, Ws-0, Wil-2, Tsu-0, Sf-2, Rsch-4, Po-0, Oy-0, No-0, Mt-0, Ler-0, Kn-0, Hi-0, Edi-0, Ct-1, Col-0, Can-0, and Bur-0). These data are obtained from the Arabidopsis RNA-seq Database (http://ipf.sustech.edu.cn/pub/athrna/) (Zhang et al. 2020). The average and standard deviation were calculated and used as features ("avg exp. across accessions" and "SD exp. across accessions").

c. Conditions
From the same database, we collected whole tissue RNA-seq data of the wild-type Col-0 accession. We divided these data into experiments with and without treatments to generate four features for average and standard deviation of treatment and control conditions ("avg exp across treatments", "avg exp. across controls", "SD exp. across treatments", and "SD exp. across controls").

We removed datasets from the Arabidopsis RNA-seq Database with a very low total read count and/or many unmapped reads. The list of samples used to generate gene expression features can be found in Table S6.

3. Structural features
The structure of causal genes and encoded proteins might differ from the other genes. Therefore, we generate four structural features: the numbers of introns, 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). Given the low number of known eQTGs, we use known QTGs and Arabidopsis orthologs of QTGs found in other species as positives and other genes as negatives, similar to QTG-Finder2. 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-operating characteristic (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 (Table S1). We ensure that these candidates are not used for hyperparameter tuning or cross-validation.

Independent validation is performed using the best combination of parameters (Table S5). 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 within 2 Mbp around it (1 Mbp upstream and 1Mbp downstream). For these genes, the average probability of being causal is calculated over 5,000 models. These average probabilities are then ranked for prioritization, and the rank is calculated as a performance measure. For example, a rank of 10% indicates that 10% of genes in the eQTL region rank higher than the candidate.

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 between the full model (with all features) and the reduced model is then calculated and used to indicate the feature importance. In addition, we calculate feature importance for clusters of correlated features. Features are clustered if they have a pairwise Pearson correlation equal to or larger than 0.6. We use the previous cross-validation framework and the best parameters to measure the model performance in this analysis.

Data analyses

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.

3.3. 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 (Figure S2.1), indicating that we add new information to the model. Figure 3.1 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 Table S3.2). 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.

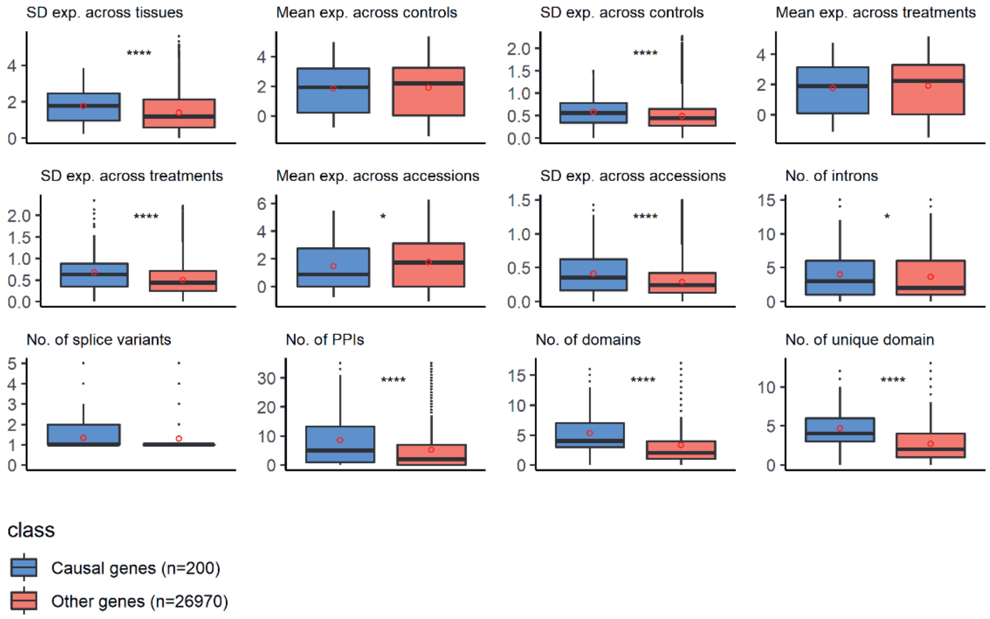


Figure 3.1. Distribution of twelve new features for known causal genes as the positive class (blue: n=200; 55 known QTGs and 145 orthologs of QTGs from other species) and the remaining genes in the genome as the negative class (red: n=26,970). Significance of differences in medians was assessed using 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 curve (AUC) of the receiver-operating characteristic (ROC) between the original QTG-Finder2 and the extended model that we labelled eQTG-Finder, and for the extended model with the class labels permutated, as a control (Figure 3.2A). 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.

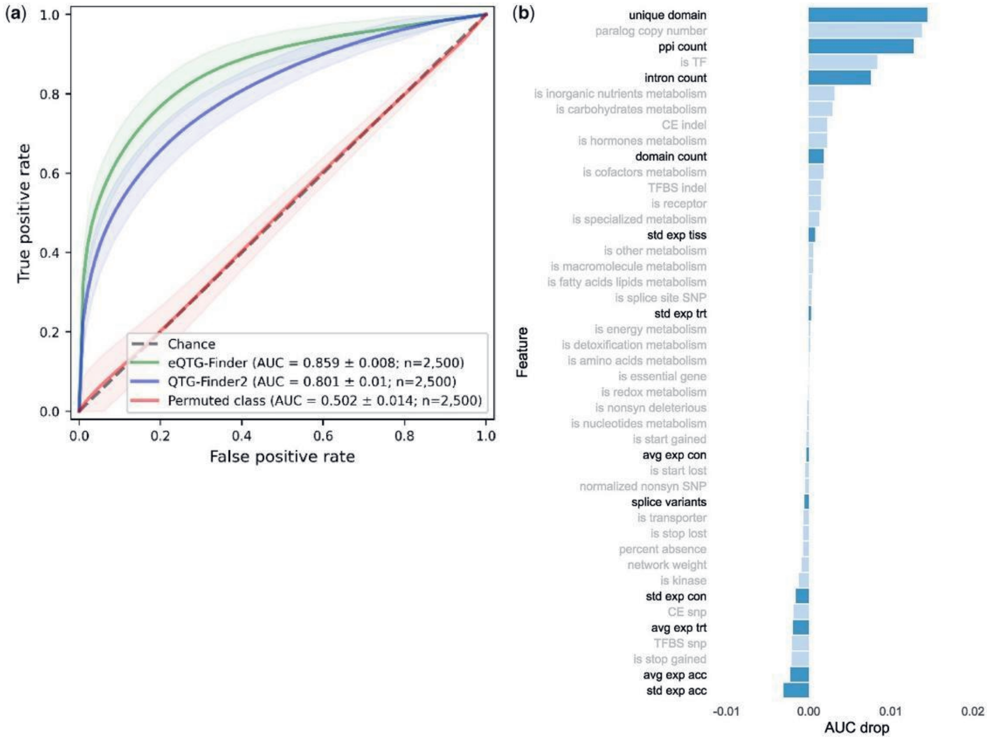


Figure 3.2. (A) Area under the curve (AUC) of the receiver-operating characteristic (ROC) of the original QTG-Finder2 model (blue) and extended eQTG-Finder model (green), and eQTG-Finder trained with randomized class labels (red) as a control. Transparent areas indicate standard deviations over 2,500 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.

To determine how the new features contribute to causal gene prediction, we calculate feature importance using a leave-one-out approach (Figure 3.2). 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.

Some features in the model are highly correlated (Figure S3.1). When one of these features is removed to calculate feature importance, the reduced model will resort to using these correlated features. As a result, the removed feature might be assigned lower importance than it should have in the model (Gregorutti et al. 2016). To avoid this, we calculated feature importance for clusters of features. The result (Figure S3.3) shows a slight change in the importance of some features, for example, "network weight" is now among the top important features since it is correlated with "ppi count".

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 (Table S1.1). 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 Materials and 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 to indicate the percentage of genes on the eQTL with higher ranks than the gene of interest (*i.e.,* a rank of 10% 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 3.3). 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 Materials and methods). Four out of six strong eQTG candidates (*AOP2, ERECTA, GIGANTEA,* and *MAM1*) rank within the top 5% by eQTG-Finder compared to only one (*ERECTA*) by QTG-Finder2. The other two strong candidates, *FRI* and *ELF3*, were ranked at 10.2% and 61.2% by eQTG-Finder. The ranks of sixteen genes are improved by eQTG-Finder, eight are worse, and one stays the same (Table S3.3). The rank of four out of six strong eQTG candidates improves, with *GIGANTEA* one of the most drastic improvements, moving from 53.7% to 4.2%. On the other hand, the rank of *ERECTA* drops (0.4% to 2.8%) but remains in the top 5%. Both models rank another strong eQTG candidate *ELF3* poorly (at 44% by QTG-Finder2 and 61.2% by eQTG-Finder). As the number of strong eQTG candidates is limited, we also show the prioritization of hypothetical and hypothetical-orthologs eQTGs. Even though the improvement was not as large as for the strong eQTG candidates, eQTG-Finder still ranks most of the hypothetical and hypothetical-ortholog eQTGs in the top 10%.

Despite the decent overall performance in candidate eQTGs 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 top 20% by eQTG-Finder, compared to nine by QTG-Finder2 (Figure S3.2).

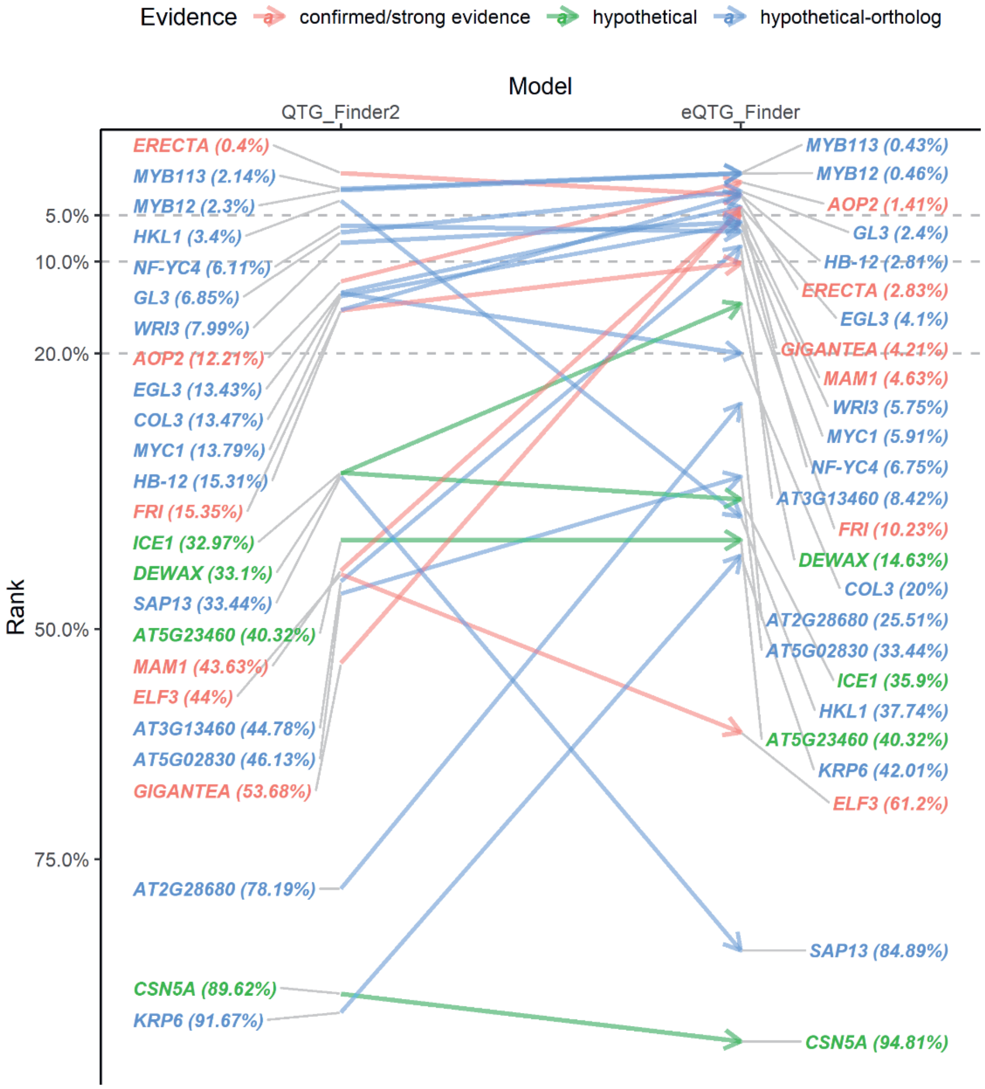


Figure 3.3. Rank comparison of sixteen candidate eQTGs using the model with new features (eQTG-Finder) and the original model (QTG-Finder2).

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 3.4). This skewed towards a low value, with a median value of 0.007 (note that the *x*-axis of Figure 3.4 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.

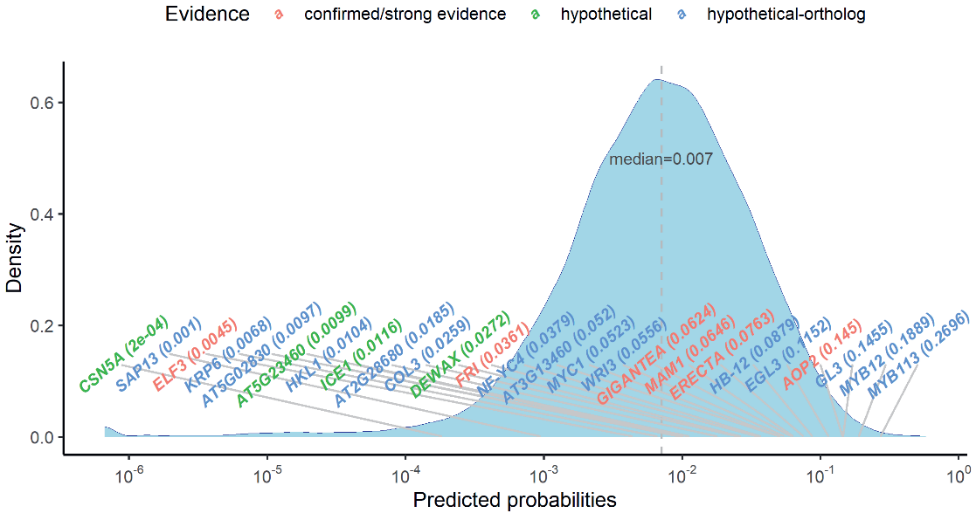
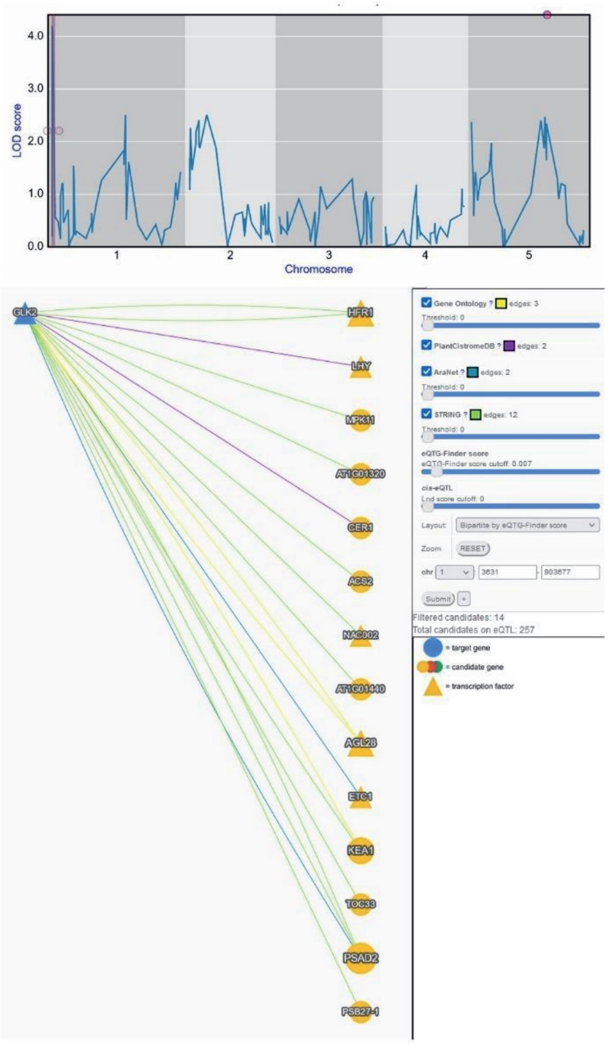


Figure 3.4. The density plot of probabilities of being causal predicted by eQTG-Finder for all Arabidopsis genes. Text labels point to the probability of the gene in the plot. The x-axis is on a $\log_{10}$ scale.

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 (Table S3.4), 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.

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 3.5, top). As many as 257 candidate regulatory genes are present in the eQTL (Figure 3.5, 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 (Chapter 4). 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 reported to regulate *GLK2*. However, this gene is a promising *GLK2* regulator candidate as the network shows that it has a transcription factor binding site(s) on the *GLK2* promoter (O'Malley et al. 2016). Moreover, *LHY* is involved in light signaling (Joo et al. 2017; Kim et al. 2003). This example suggests that integrating the eQTG-Finder score in AraQTL can help infer new regulatory interactions.



☑ Gene Ontology ? 🟨 edges: 3

Threshold: 0

☑ PlantCistromeDB ? 🟪 edges: 2

☑ AraNet ? 🟦 edges: 2

Threshold: 0

☑ STRING ? 🟩 edges: 12

Threshold: 0

eQTG-Finder score

eQTG-Finder score cutoff: 0.007

cis-eQTL

Lod score cutoff: 0

Layout: Bipartite by eQTG-Finder score

Zoom: RESET

chr: 1 3631 - 903677

Submit +

Filtered candidates: 14

Total candidates on eQTL: 257

● = target gene

● = candidate gene

▲ = transcription factor

Figure 3.5. 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.

3.4. 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 in prioritizing eQTG. eQTG-Finder outperforms QTG-Finder2 in distinguishing positive causal genes from the other genes in the genome based on a cross-validation setting (Figure 3.2). Moreover, eQTG-Finder prioritizes most eQTGs in eQTLs better than QTG-Finder2 in an independent validation test (Figure 3.3). 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 3.1). 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, tissues, and conditions. Even though the medians between causal and other genes are significantly different (Figure 3.1), features based on SD of expression have low importance in the model (Figure 3.2). A possible explanation for this could be that features based on expression are correlated (Figure S3.1) and, therefore, their importance is underestimated (Gregorutti et al. 2016). We, therefore, removed all of these correlated features and re-calculated the feature importance. The feature importance, however, remains the same. 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. A recent finding in humans showed that *cis*-eQTLs and GWAS genes are different due to the detection bias of the assays (Mostafavi et al. 2022). This detection bias could also hold for trans-eQTL and phenotype QTL genes in Arabidopsis. However, we argue that QTGs are still 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% 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 (Chapter 4). For example, eQTG-Finder prioritizes transcription factor *LHY* as the regulator of *GLK2* (Figure 3.5). 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 3.3). 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).

We observed that eQTG-Finder prioritization of candidate QTGs in independent validation was slightly worse compared to QTG-Finder2 (Figure S3.2), despite its better performance in cross-validation (Figure 3.1). The new expression-based features might bias eQTG-Finder toward prioritizing eQTGs compared to QTGs, but the complexity of the model makes it difficult to learn exactly how these features affect prioritization. Moreover, the number of eleven candidates we used for validation is too low to allow a very precise assessment of the general performance of eQTG-Finder in prioritizing QTGs.

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 are 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). Supplementary materials are available at https://academic.oup.com/g3journal/article/12/11/jkac255/6712316#supplementary-data.

Acknowledgments

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

Chapter 4. AraQTL: mining gene expression regulation in eQTLs using knowledge graphs

Genetical genomics studies are rich sources of gene regulatory information that have remained largely unexplored. These studies associate genes with genomic regions that affect their expression, so-called expression Quantitative Trait Loci or eQTLs. Identifying the actual causal gene or polymorphism in these eQTLs is complicated by the relatively large size of the regions, which often harbour dozens or even hundreds of genes. Experimental fine-mapping all the causal genes is impractical, given the thousands of eQTLs in a single study. Prior biological knowledge in the form of gene/protein interaction data can be used to rank the potential regulators and create a list of the most likely candidate genes for validation. To aid biologists in unlocking the potential of genetical genomics studies, in this case for Arabidopsis, we developed a novel knowledge graph-based method to link genes with candidate regulators using publicly available gene annotation, protein-protein interaction, and transcription factor binding site data. This method is accessible in AraQTL, a platform for systems genetics study in Arabidopsis. We provide two use cases where our method successfully recovers known regulators of a gene and prioritise the candidate master regulator of an eQTL hotspot. In addition, we extended AraQTL with more than 700 phenotype and metabolite QTL profiles and presented a use case for prioritising candidate genes for traits using QTL profile correlation. These two improvements are major enhancements of the utility of AraQTL to unravel the molecular regulation underlying variation in complex traits.

In preparation for publication.

Margi Hartanto, Marie Lefebvre, Antea Kardum, Vincent van den Berg, Basten L. Snoek, Dick de Ridder, Harm Nijveen

4.1. Introduction

One of the main objectives of genetic research is to unravel the molecular mechanism through which genotype variation affects phenotypic variation. The genetic loci that contribute to complex traits can be studied using statistical techniques such as linkage mapping or association studies (Bazakos et al. 2017). These approaches identify quantitative trait loci (QTLs), genomic regions associated with traits. However, despite its usefulness in elucidating the genetic architecture of traits, the information provided by QTL mapping only brings limited insight into the genetic regulation of traits at the molecular level. An alternative is to map QTLs for the abundance of biological molecules, for example, RNA, resulting in expression QTLs (eQTL) (Jansen and Nap 2001). The rationale for eQTL mapping is that the flow of biological information from DNA to trait occurs through a network of intermediate molecular phenotypes, such as metabolite, protein, and RNA (Civelek and Lusis 2013). eQTL mapping could therefore provide more insight into the regulation of complex traits. Several eQTL mapping studies have been carried out over the past decade, for instance, in model organisms *Arabidopsis thaliana* (Keurentjes et al. 2007; West et al. 2007; Cubillos et al. 2012; Snoek et al. 2012; Lowry et al. 2013; Hartanto et al. 2020; Imprialou et al. 2017; Kawakatsu et al. 2016; Rabanal et al. 2017). These studies yielded a vast number of eQTLs that can be used to study gene expression regulation underlying complex traits.

To give the research community better access to eQTL data sets, we developed a web-based workbench and database for eQTL studies, AraQTL (Nijveen et al. 2017) for Arabidopsis, WormQTL2 (Snoek et al. 2020) for *C. elegans,* and SolQTL for tomato (in preparation). These platforms allow researchers to analyse published eQTL datasets interactively. For example, with a simple query, one can obtain the eQTL profile of a gene of interest, pinpointing the locations of possible regulators on the genome. However, due to the low mapping resolution, the eQTL interval typically spans a large genomic interval that can still harbour tens to hundreds of candidate regulatory genes. The resolution can be improved through fine mapping, which requires the development of a new population with more recombination events on the eQTL of interest (Tuinstra et al. 1997). However, fine-mapping thousands or even just dozens of available eQTLs is impractical because it is costly and laborious. This obstacle clearly hampers the exploration of potentially valuable eQTL datasets.

To accelerate the identification of the causal genes underlying eQTLs, a computational method can help to explore the genes underlying eQTLs and prioritise the most likely regulators for further experimental validation. The method should consider different possible gene interactions since the unknown regulatory genes may interact with the target genes in many ways. Jimenez-Gomez et al. (2010) provided an example, who detected *ELF3* as the causal gene underlying many co-locating eQTLs using co-expression and Gene Ontology (GO) data. Unfortunately, although these data have been widely generated for model species, they are dispersed across several databases in different data formats with variable quality. Because of this, researchers must make considerable efforts to retrieve and evaluate these data separately before using them to prioritise the candidate genes.

We addressed the problem by developing a data integration platform and a knowledge-based network visualisation method to link genes with their likely regulator underlying eQTL regions. The regulatory evidence can be inferred from transcription factor binding sites, protein-protein interactions, and similarity in functional annotations. Our platform utilises these types of data to link genes, including transcription factor-target data from AGRIS (Yilmaz et al. 2011), PlantCistromeDB (O'Malley et al. 2016), PlantRegMap (Tian et al. 2020), iGRN (De Clercq et al. 2021), protein-protein interaction data from STRING (Szklarczyk et al. 2017) and AraNet (Lee et al. 2015b), and gene annotation data from Gene Ontology (Gene Ontology 2021) and KEGG (Kanehisa et al. 2021). These data can be retrieved and shown as an interaction network to allow visual exploration of the candidate regulators in one or more eQTL regions. Interactions can be filtered based on confidence scores to reduce the list of candidates even more. Alternatively, users can select a genomic region that affects the expression of many genes, a so-called eQTL hotspot. This method will produce and visualise an interaction network connecting the genes regulated by the region to identify potential master regulators in the region that are linked to many of the query genes. Our method successfully recovers *FRI* and *FLC* as the known regulators of *AGL20/SOC1* and prioritises *ERECTA* as the master regulator of an eQTL hotspot. Thus, the AraQTL knowledge-based data mining approach will support researchers in identifying key gene expression regulators easily. In addition, we included phenotype and metabolite QTL data in AraQTL and presented an example where QTL profile correlation helps identify candidate genes for flowering time. We believe that these new functionalities render AraQTL more useful for systems genetics studies and candidate gene prioritisation in general.

4.2. Material and methods

AraQTL stores and combines all published Arabidopsis genetical genomics data to enable easy access and visualisation of eQTLs. We extend this functionality by developing a data integration platform and network visualisation method to connect genes with their regulator on eQTLs using data from different sources.

Data

The TAIR10 Arabidopsis gene annotations were retrieved from release 53 of the Ensembl Plant Database (http://ftp.ensemblgenomes.org/pub/plants/release-53/gff3/arabidopsis_thaliana/Arabidopsis_thaliana.TAIR10.53.gff3.gz). Data to infer potential regulation between Arabidopsis genes were based on these following three categories:

1. Transcription factor binding site (TFBS)
   We use transcription factor (TF) – target gene pairs data based on TFBS experiments to identify likely transcriptional regulators on eQTLs. These data were downloaded from AGRIS (Yilmaz et al. 2011), PlantCistromeDB (O'Malley et al. 2016), PlantRegMap (Tian et al. 2020), and iGRN (De Clercq et al. 2021). For AGRIS, we downloaded the data from https://agris-knowledgebase.org/Downloads/AtRegNet.zip and filtered direct or confirmed entries, which resulted in 1,458,834 TF-target interactions. For PlantCistromeDB, peak data in the narrowPeak format were downloaded from http://neomorph.salk.edu/dap_web/pages/browse_table_aj.php, and the top 25% peaks were selected based on the signal score. The resulting data contains 2,585,567 TF-target interactions. For PlantRegMap and iGRN, we downloaded the TF-target data from http://plantregmap.gao-lab.org/download_ftp.php?filepath=08-download/Arabidopsis_thaliana/binding/regulation_merged_Ath.txt and http://bioinformatics.psb.ugent.be/webtools/iGRN/files/igrn_data/iGRN_network_full.txt, respectively. The number of interactions is 648,662 for PlantRegMap and 620,751 for iGRN. These data were downloaded in July 2022 unless indicated otherwise.
2. Protein-protein interaction (PPI)
   Functional genes or protein interactions may provide evidence for gene expression regulation. To obtain such evidence, we use interaction data from STRING (Szklarczyk et al. 2019) (https://string-db.org/cgi/download, accessed July 2022) and AraNet V2 (Lee et al. 2015b) (https://www.inetbio.org/aranet/downloadnetwork.php, accessed July 2022). We parsed the gene identifiers, interaction evidence, and interaction probability scores from these data. In total, 31,108,576 functional interactions were collected from STRING and 733,183 from AraNet V2.
3. Gene functional annotation
   Gene expression regulators and the targets may be annotated with a particular biological process or pathway. To identify likely candidate regulators with similar annotation, we use biological process annotation data from Gene Ontology (GO) (Gene Ontology 2021; Ashburner et al. 2000) and pathways from KEGG (Kanehisa et al. 2021). Arabidopsis GO annotation data were downloaded from http://current.geneontology.org/annotations/tair.gaf.gz and filtered to only contain biological process terms. We assume potential regulation based on GO term semantic similarity. We use a Python script with the Biopython module to access the KEGG online API and get the annotation for the KEGG pathway. Gene regulation is inferred if two genes have similar pathway annotations.

In addition to the data above, we also included eQTG-Finder scores (the probability of a gene being causal) based on a machine learning approach (Hartanto et al. 2022) and the *cis*-eQTL LOD score of the candidate regulators as an indication that the gene is polymorphic.

TFBS and PPI data are structured as graphs and stored in a Neo4J graph database version 4.4.9 (Robinson et al. 2015). The data structure is shown in Figure 4.1, where genes are represented as nodes connected by relationship edges. The Cypher query language is used to retrieve data from the graph database.

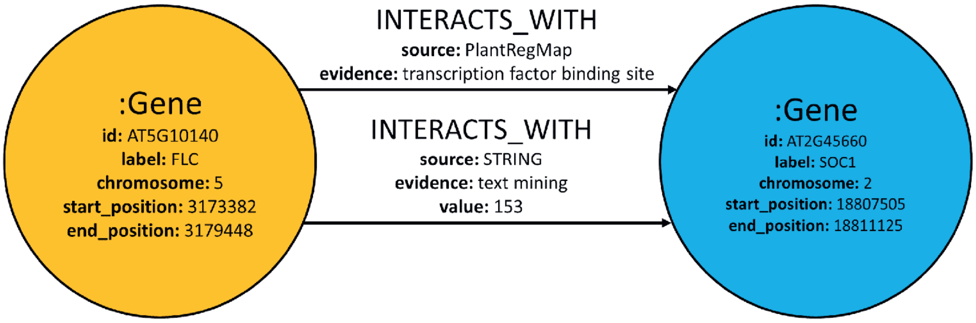


Figure 4.1. Transcription factor binding site and protein-protein interaction data structure in a Neo4J graph database. Genes are represented as nodes connected by INTERACTS_WITH relationships. Gene nodes have id, label, chromosome, start_position, and end_position properties. Relationships have source, evidence, and value properties.

GO term semantic similarity calculation

The semantic similarity score between GO terms is calculated using a custom Python script with the GOATOOLS library (Klopfenstein et al. 2018). We use the implementation of Resnik's similarity measure, defined as the information content of the most informative common parent term in the GO hierarchy (Resnik 1999). The similarity score between two genes is determined by the maximum semantic similarity of these genes' annotated GO terms.

Phenotype QTL data

We collected and remapped phenotype and metabolite QTL data and included them in AraQTL. We curated data found in the supplementary information of 45 publications, totaling 737 traits (Table S4.1) from Bay x Sha and Ler x Cvi RIL populations. Using the obtained phenotypes and the most recent genotype data (Alonso-Blanco et al. 1998; Serin et al. 2017) QTL profiles for each study were remapped using a linear single marker model:

$$y_{i,j} \sim x_j + e_j$$

where y is the phenotype value i of RIL j based on the function of marker genotype x_j and the error term e_j. Subsequently, markers were called significantly associated if the

-$\log_{10}$(p-value) exceeded 3. This analysis was performed using a custom-made script in ‘R’ (version 4.0.2, Windows x64).

Network visualisation

To visualise and explore the interaction network of one or more target genes with a set of candidate regulators, we build upon the Cytoscape.js library (v3.20) (Franz et al. 2016) and three network layout libraries: namely cytoscape-cola.js (v2.5.0) (Frans et al. 2021), cytoscape-cose-bilkent.js (v4.1.0) (Balcı et al. 2019), and cytoscape.js-cise (Mistry et al. 2013). In addition, we implemented a bipartite layout based on the default grid layout. The bipartite layout shows the nodes in separate columns based on their type (target or candidate regulator) and orders each column based on the number of interactions (node degree), eQTG-Finder score, and *cis*-eQTL LOD score to prioritise the candidate regulatory genes. The cytoscape-popper.js (v2.0.0) (Mistry et al. 2013) library is used to display tooltips. The network visualisation interface is integrated into the AraQTL multiplot and co-regulation dynamic web pages. The knowledge graph is queried using AJAX calls through a Django-based API. This API allows for retrieving interactions based on a target gene identifier and the interval of an eQTL. The result is a gene interaction network, where nodes represent genes and edges represent the interaction between genes based on various sources of evidence.

Software

AraQTL was developed using the Python Django web framework version 3.1.5. The backend runs on Debian GNU/Linux 10.9 (buster) using the Apache web server version 2.4.38 and a PostgreSQL 13.1 database. The web frontend is implemented via Django templates in HTML and Javascript, using the D3 and jQuery libraries. The software runs in Docker containers, including the PostgreSQL and Neo4J databases (Boettiger 2015).

4.3. Results

To further enhance the utility of the AraQTL web platform, we developed an eQTL causal gene prioritisation method based on a knowledge graph built from various gene-gene and protein-protein interaction information sources. Below, we present two use cases illustrating how our method allows researchers to interactively explore potential gene expression regulators. Additionally, we included over 700 phenotype QTL profiles in the database and show in a third use case that a simple correlation analysis between phenotype QTL and eQTL profiles can already help to rank potential candidate genes for the phenotype.

Use case 1: the exploration of candidate regulators underlying two eQTLs for the AGL20 gene

In Arabidopsis, thousands of eQTLs have been identified; however, only a few regulatory genes underlying these eQTLs are known. In this use case, we present a method to explore the candidate genes for single or multiple co-locating eQTL peaks. We use the eQTL profile of *AGAMOUS-LIKE 20* (*AGL20,* also known as *SOC1*) as a floral pathway integrator gene that responds to multiple environmental and endogenous signals to determine the transition to flowering (Simpson and Dean 2002). Factors affecting *AGL20* regulation, including its upstream regulator, have been well-characterised (Lee and Lee 2010), making it an excellent example of how prior-biological knowledge can help identify gene expression regulators. eQTL analysis on rosette leaf tissue of a Bayreuth x Shahdara RIL population revealed two strong eQTLs, located on chromosome 4 and chromosome 5, regulating the expression of *AGL20* (West et al. 2007). Using the interactive knowledge graph approach, we could reduce the number of candidate regulators from over 250 genes to known AGL20 regulators *FLC* and *FRI* (Figure 4.2).

We developed a simple interactive method to retrieve the interaction data in AraQTL. To explore the possible causal gene for the *AGL20* eQTL, the user can start by querying “AGL20” in the AraQTL search box and selecting the West et al. (2007) experiment to show the eQTL profile of *AGL20*. Two eQTL peaks will be shown, one on chromosome 4 and another on chromosome 5 (Figure 4.2A), indicating the presence of *AGL20* candidate regulators (harbouring 76 and 215 genes, respectively) in these two locations. By clicking “Show interaction network” and selecting one of the eQTL peaks, a new window will show a network where *AGL20* is connected to its candidate regulators based on multiple sources of evidence contained in the knowledge graph (Figure 4.2B). This action reduces the number of candidates to 9 genes for the chromosome 4 eQTL and 31 for the chromosome 5 eQTL. In the network, nodes represent genes, and edges represent the interaction connecting two genes (*e.g.,* similar GO biological process annotation).

Some interactions (*e.g.,* STRING) have scores that can be filtered by the user, thus allowing users to remove low-confidence interactions. In this case, we define an arbitrary threshold to remove low GO similarity and STRING scores in the network. The filtering narrowed the candidate for the eQTL on chromosome 4 to *FRIGIDA* (*FRI*) and the eQTL on chromosome 5 to *FLOWERING LOCUS C* (*FLC*). *FLC* is a repressor of floral transition, which works by binding to the promoter of *AGL20* and inhibiting its expression (Hepworth et al. 2002). In the chromosome 5 eQTL network, *FLC* is connected to AGL20 through TFBS evidence (the yellow nodes on the right in Figure 4.2C), which agrees with Hepworth et al. (2002). These two genes are also connected because they share similar GO annotations (“regulation of flower development” and “response to cold”). *FRI* on chromosome 4 is a positive regulator of *FLC* and an indirect inhibitor of *AGL20* (Michaels and Amasino 2001). *FRI* has two connections with *AGL20* in the chromosome 4 eQTL graph, one based on similar GO annotations and the other on co-occurrence in literature (the green nodes on the left in Figure 4.2C). *FLC* and *FRI* are also connected in the combined network, reflecting their interaction that ultimately affects *AGL20* regulation.

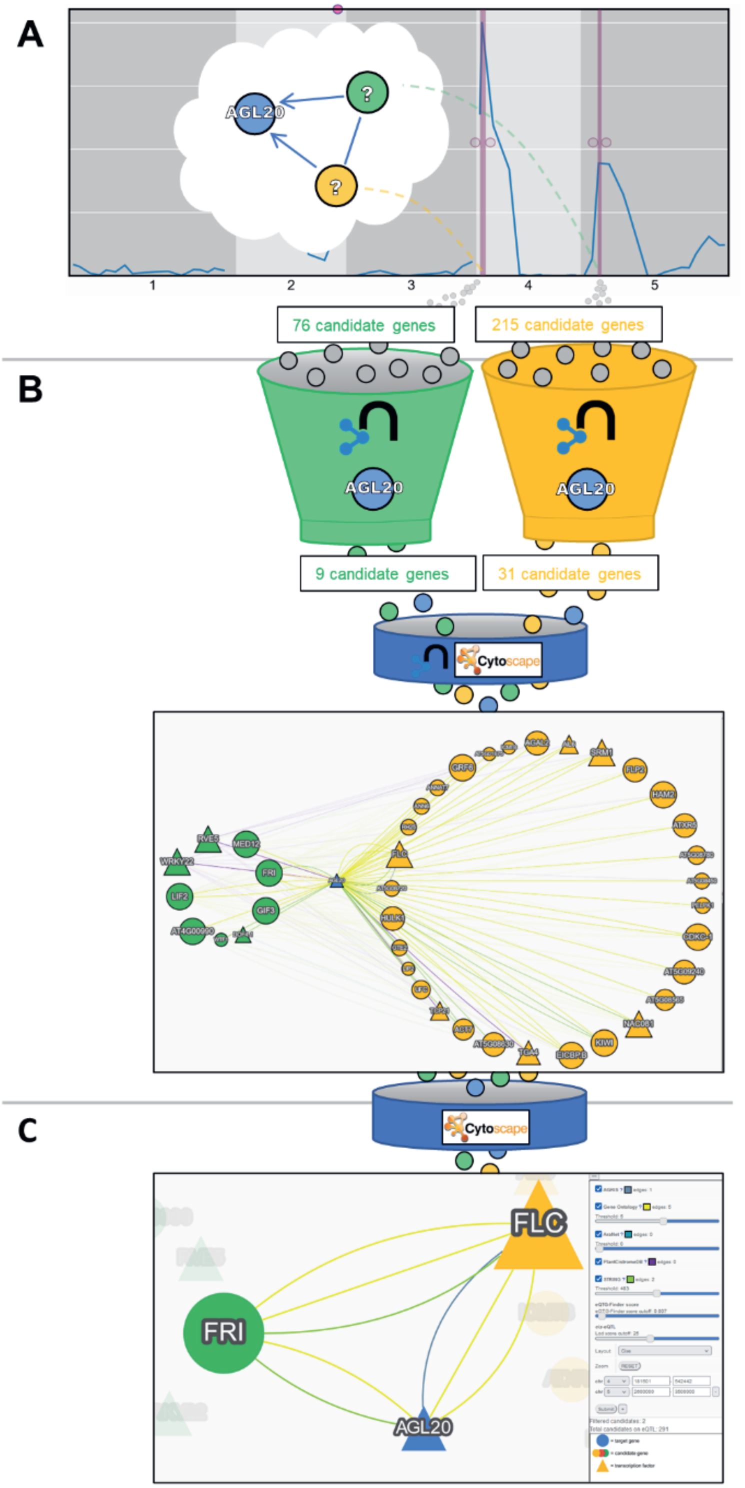


Figure 4.2. Exploring candidate regulatory genes for AGL20 using gene interaction networks in AraQTL. (A) *AGL20* has two prominent eQTLs on chromosome 4 and chromosome 5, indicating the presence of gene expression regulators in these genomic locations. Genes located in the eQTL intervals are assigned as candidate regulators. (B) The knowledge graph is used to select genes interacting with or similarly annotated as *AGL20*, drastically reducing the number of candidate genes. The result is shown as a network where nodes are genes (the target gene is the blue node), and the edge colour indicates the type of interaction evidence. (C) Further candidate selection can be done by pruning edges that are based on weak evidence (e.g., low-scored STRING interactions). By doing so, the resulting network recovers known flowering pathways where *FLC* and *FRI* regulate *AGL20*.

Use case 2: prioritisation of the master regulator underlying an eQTL hotspot

A common finding in eQTL studies is the co-location of numerous distant eQTLs in a particular genomic location. Such a genomic region is called an eQTL hotspot or trans-band and is assumed to point to one or more common master regulatory genes in the region (Breitling et al. 2008). Identifying such genes has become one of the primary efforts in eQTL studies, as it might reveal a crucial gene at the top of a regulatory hierarchy affecting many downstream genes involved in a particular biological process or pathway. Previous studies have used a co-expression network to identify the master regulator by looking for hub genes, which have the largest number of connections to other genes in the network (e.g., in Hartanto et al. 2020; Keurentjes et al. 2007; Jimenez-Gomez et al. 2010). However, using co-expression alone might not be sufficient because the polymorphism in the master regulator may not exert its effect through alteration of its expression but on another level of regulation, for example, on the protein level. This is where AraQTL comes in useful. The interactions in AraQTL's knowledge graph cover multiple levels, including gene expression and protein level, which makes it suitable to capture the potential interaction between the regulator and its target.

In this use case, we will use the prominent eQTL hotspot (chromosome 2: 11.2 Mbp) discovered in Keurentjes et al. (2007) and Snoek et al. (2012) using a Landsberg *erecta* x Cape Verde Islands RIL population (Alonso-Blanco et al. 1998). This hotspot is most likely caused by the *ERECTA* gene (Keurentjes et al. 2007; Snoek et al. 2012; Terpstra et al. 2010), which is mutated in the Landsberg *erecta* accession and known as a pleiotropic regulator of many functions in plants (van Zanten et al. 2009).

To explore the candidate genes on the eQTL hotspot, we developed a prioritisation method in AraQTL using the knowledge graph. By inspecting the eQTL *cis-trans* plot of the Keurentjes et al. (2007) experiment in AraQTL, a clear trans-eQTL hotspot on chromosome 2 can be seen as a vertical line on the plot. Clicking the histogram bar underneath the hotspot brings up the co-regulation page, showing the genes that have an eQTL at the hotspot (Figure 4.3). By selecting all of these genes in the table underneath the QTL plot and clicking the "Show interaction network" button, a network is constructed that shows the interaction between the candidate genes (located in the hotspot region) and the target genes (having an eQTL mapping to the hotspot). As before, nodes represent genes in the network, and edges represent interactions between two genes (*e.g.,* text-mining from STRING). The candidates and targets are coloured in blue and orange, respectively, and the default layout groups candidate and target genes in two columns. The genes are ordered by the number of connections, with the genes with higher degrees appearing higher in the column. Ordering the candidate regulators facilitates the prioritisation of the most likely regulator of the eQTL hotspot among the candidate genes. For the chromosome 2 eQTL hotspot network, *ERECTA* appears at the top of the prioritised candidate genes, which agrees with the literature (Keurentjes et al. 2007; Snoek et al. 2012; Terpstra et al. 2010; van Zanten et al. 2009).

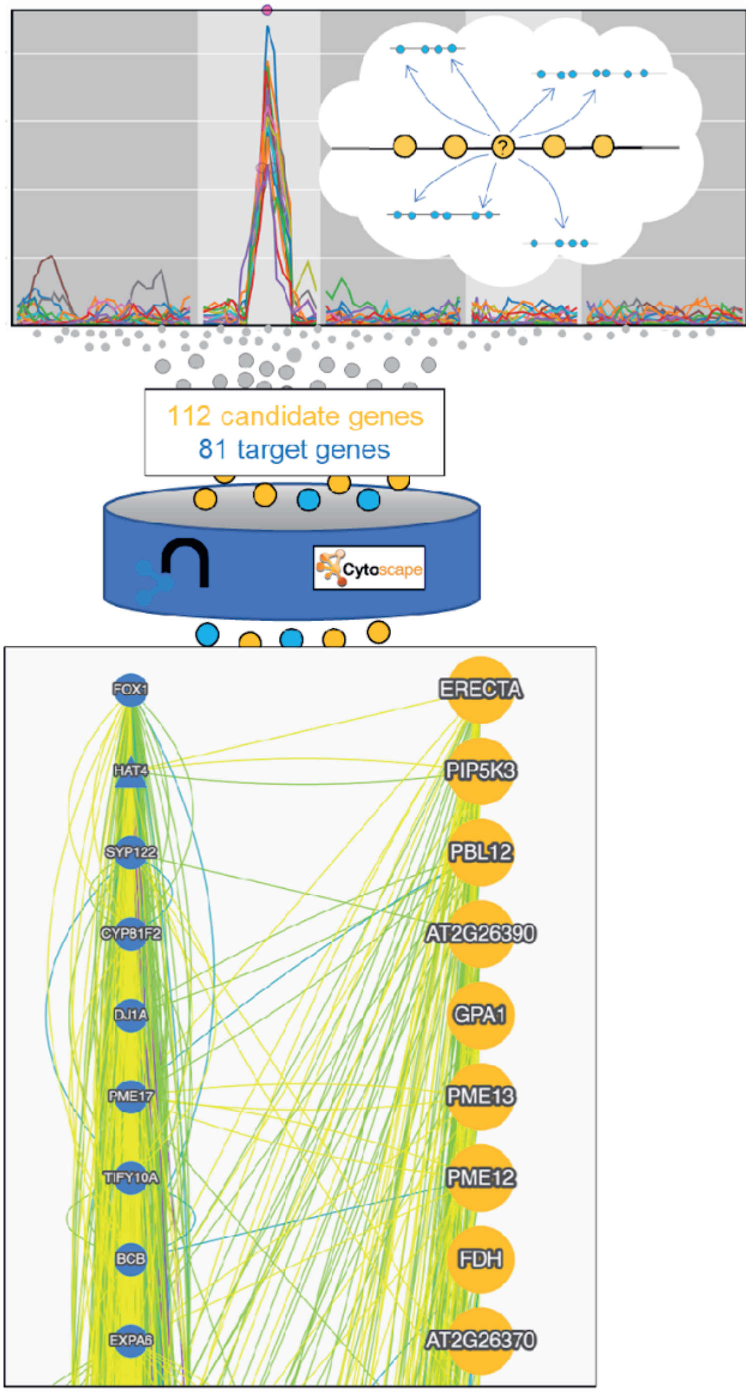


Figure 4.3. Prioritisation of candidate genes for an eQTL hotspot using a knowledge graph. eQTL mapping in seven days old Arabidopsis seedlings identified an eQTL hotspot on chromosome 2 composed of many co-locating eQTLs, potentially caused by one or more master regulatory gene(s). AraQTL can be used to identify potential master regulators on eQTL by exploring a knowledge network composed of the target genes (having an eQTL on the hotspot) and candidates (located on the hotspot). The network shows the interactions between the targets (blue) and the candidates (yellow). The network layout groups the targets and candidate regulators in two separate columns, in descending order based on the number of interactions, a measure that can be used to prioritise the master regulatory genes. In this case, the well-known master regulator *ERECTA* is the top prioritised candidate gene.

Use case 3: correlation between phenotype and gene expression QTL recovers a flowering repressor gene

To enhance the capabilities of AraQTL for systems genetics studies, we curated, remapped, and included phenotype and metabolite QTL data into AraQTL. Integrating these data will benefit researchers in studying the regulation of complex traits at the molecular level, and the simplest way to do this is through the correlation of multi-omics QTL profiles. For instance, strong correlations between phenotype or metabolite QTL with eQTL profile may reflect a shared genetic architecture or even the direct involvement of the transcript-encoding gene, *i.e.,* variation in gene expression leading to changes in phenotypes or metabolite levels. This approach can help prioritise candidate regulatory genes. In AraQTL, we extended the correlation function to also analyse the correlation between phenotype and gene expression QTL. For example, we analysed the correlation of the flowering time in a long day (FTLD) phenotype QTL of a Bay x Sha RIL population (Joosen et al. 2012) with eQTLs from experiments using the same population in (Figure 4.4). We selected the top five correlated genes and found that an *FLC* eQTL profile on dry seed (Serin 2018) has the highest Spearman correlation coefficient (0.86) with the phenotype QTL. *FLC* is highly abundant in dry seeds to regulate temperature-dependent seed germination (Chiang et al. 2009). The gene is also an essential regulator of flowering time (Michaels and Amasino 2001), explaining the high correlation between the eQTL and the flowering time QTL. This use case illustrates that QTL profile correlation analysis has the potential to identify candidate regulatory genes.

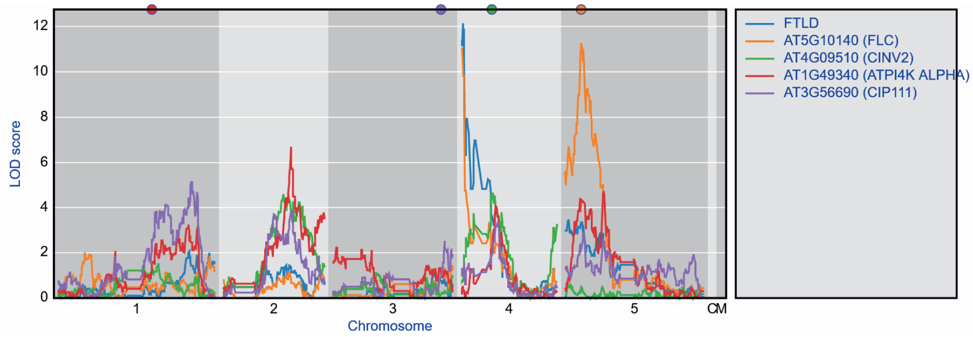


Figure 4.4. The prioritisation of candidate regulatory genes for flowering time using QTL profile correlation. Spearman correlation of flowering time between a long day (FTLD) QTL profile (blue line) and eQTLs found in all available Bay x Sha experiments is analysed. The top four most strongly correlated eQTL profiles are listed in the box on the right. The top profile is that of *FLC* (orange line), a well-known flowering time repressor.

4.4. Discussion

In this work, integrating prior biological knowledge is the key innovation to link genes to their potential regulators. This approach was inspired by similar integration platforms in plants, such as AgroLD (Venkatesan et al. 2018), pbg-ld (Singh et al.

2020), and KnetMiner (Hassani-Pak et al. 2021). These platforms provide access to knowledge mining to generate new insights in many areas within the plant science domain. Compared to these data integration platforms, our work specifically aims to explore the vast amount of underutilised eQTL data, particularly to identify regulators of gene expression. Building this prioritisation approach on top of AraQTL enables versatile and smooth exploration of eQTL data, as eQTL profile retrieval and causal gene identification are available within the same platform.

We collected gene interaction and annotation data from various sources to find a link between the query gene and potential regulators on eQTLs. Albert et al. (2018) showed that eQTLs in yeast were enriched for TF-encoding genes, pointing to an essential role for this class of genes; hence we included TF-target gene pair information based on TFBS data in our integration platform. Other classes of genes besides TFs could also act as gene expression regulators on eQTL. For example, the *ERECTA* gene of use case 2 encodes a receptor-like kinase (Terpstra et al. 2010), and *GIGANTEA* (Keurentjes et al. 2007) is not known to encode a TF. To identify non-TF eQTL regulatory genes, we include functional gene/protein interactions from STRING (Szklarczyk et al. 2019) and AraNet V2 (Lee et al. 2015b). These gene interaction data are structured as graphs in a Neo4J database, enabling simple querying and fast retrieval. Furthermore, structuring these interaction data as graphs allows us to query for more complex interactions (*e.g.,* indirect interaction) and apply various graph analyses (*e.g.,* centrality measurements and community detection). It will be interesting to explore the usefulness of these complex interactions and community detections for QTL analysis.

A limitation of using prior biological knowledge to prioritise gene expression regulators is that it can only retrieve known gene interactions, which leads to a bias towards well-studied genes (Stoeger et al. 2018; Haynes et al. 2018; Hao et al. 2010). It makes less well-characterised genes underrepresented in the knowledge graph and less often prioritised. To remedy this, we included the eQTG-Finder score as a gene property that can be used to rank the candidate regulators. This score is the probability of a gene being causal, as inferred by a machine learning approach using several important features that are not biased towards well-characterised genes, such as the number of unique domains and the number of introns (Hartanto et al. 2022).

In addition, we included the *cis*-eQTL LOD score of the candidate regulators as an alternative way to rank eQTL candidate genes. The presence of a *cis*-eQTL indicates that the gene has a polymorphism affecting its expression. This variation in expression could in turn affect the expression of the query gene *in trans* (Jimenez-Gomez et al. 2010), making genes with *cis*-eQTLs potential regulatory genes. Filtering for such genes can be done using the *cis*-eQTL LOD score. Altogether, both eQTG-Finder and *cis*-eQTL LOD scores can help prioritise gene expression regulators regardless of the availability of prior knowledge.

The identification of *ERECTA* as the most likely causal gene for a hotspot on chromosome 2 in use case 2 shows the power of prior knowledge-based prioritization of master regulators. However, it was recently shown that eQTL hotspots could arise due to variation in developmental age within the tested population (van Eijnatten et al. 2023; Francesconi and Lehner 2014), batch effect (Michaelson et al. 2009), or be attributed to the formation of specific plant tissues (Vosman et al. 2019). In such cases, it might be more fruitful to look for commonalities between the target genes since the connection to the actual regulator can involve many regulatory steps, while analysis of the target genes could point to an underlying phenomenon. The simplest way is to analyse the GO term or pathway enrichment to identify gene expression co-founder (as in (van Eijnatten et al. 2023)), which at the same time can help identify the characteristics of the causal genes based on functional annotations. This feature should be the focus of future development.

Another significant improvement is the addition of phenotype and metabolite QTL in AraQTL. Besides making these data accessible to research communities, the availability of phenotype QTL data allows us to perform integrated multi-omics QTL analyses based on the correlation between profiles. The correlated QTL profile may indicate a direct causal relationship (gene affecting phenotype) or a common regulator of both the gene and the phenotype. However, the correlations could be spurious and thus lead to false positives; therefore, results should be carefully evaluated, *e.g.* by referencing the literature. Hence, future work should focus on adding extra information to support and validate the correlation using prior knowledge. For example, phenotype can be associated with relevant GO terms (*e.g.,* flowering phenotype with GO:0009908/flower development), which can support the phenotype-expression QTL profile correlation if the transcript encoding genes are associated with the GO terms.

In the case of the flowering time phenotype and *FLC* (use case 3, Figure 4.4), the QTL correlation most likely indicates a causal relationship, as it is long known that the null mutation leads to early flowering (Michaels and Amasino 2001; Hepworth et al. 2002). Finnegan et al. (2020) showed that the activation of *FLC* expression is started during embryo development and continues until the vegetative stage. In the transcriptome analysis of dry seed of Bay x Sha RILs (Serin 2018), *FLC* is expressed and shown to be regulated by two eQTLs (Figure 4.4). eQTL on chromosome 5 is collocated with the gene location, which is a likely *cis*-eQTL (*e.g.,* variation in promoter leading to variation in gene expression). On chromosome 4, the *trans*-eQTL is potentially caused by *FRI*, a transcription activator known to promote the expression of *FLC* (Choi et al. 2011). Figure 4.2 (use case 1) shows the interaction between *FRI* and *FLC*, where these genes are connected in the network based on similarity in GO annotations and protein-protein interactions. Altogether, the transcriptional regulation leading to the variation in flowering time can be retrieved and nicely shown in AraQTL using knowledge-graph, network visualization, and QTL profile correlation.

To conclude, AraQTL now allows for exploring eQTL regions to find potential causal genes based on prior knowledge. In addition, it is extended with a large number of published phenotype QTL profiles which provides opportunities to find causal genes for traits. These two new features enhance the systems genetics capabilities of AraQTL, allowing researchers to generate novel hypotheses regarding causal genes and underlying molecular mechanisms. For example, researchers studying a specific gene or phenotype can now explore the potential causal genes in AraQTL and select the prioritized subset for further experimental validation. Future development can focus on applying this platform to agriculturally important crops for which eQTL studies are available, *e.g.,* tomatoes (Sterken et al. 2021; Wang et al. 2020) or lettuce (Zhang et al. 2017). Using homology information (Emms and Kelly 2019), interaction and annotation data can be transferred between species to increase data availability for a more reliable exploration of candidate regulatory genes.

Acknowledgements
We would like to thank Arnold Kuzniar, Weiqi Tang, and Thijn van Kempen for their help in data mining and database development and Joost Keurentjes for providing a list of phenotype data.

Supplementary Materials

Table S4.1. Phenotype QTLs in AraQTL

Study ID	Year	Trait	#traits	RIL population	Reference
Alonso-Blanco_etal_1998	1998	flowering time	12	Ler x Cvi	Alonso-Blanco et al. (1998)
Alonso-Blanco_etal_1999	1999	life history	12	Ler x Cvi	Alonso-Blanco et al. (1999)
Alonso-Blanco_etal_2003	2003	seed germination	7	Ler x Cvi	Alonso-Blanco et al. (2003)
Alonso-Blanco_etal_2005	2005	freezing tolerance	4	Ler x Cvi	Alonso-Blanco et al. (2005)
Anwer_etal_2014	2014	circadian clock	4	Bay-0 x Sha	Anwer et al. (2014)
Baxter_etal_2007	2005	ions	34	Ler x Cvi	Baxter et al. (2007)
Bentsink_etal_2000	2000	sugar and seed germination	12	Ler x Cvi	Bentsink et al. (2000)
Bentsink_etal_2003	2003	seed weight and content	5	Ler x Cvi	Bentsink et al. (2003)
Borevitz_etal_2002	2002	hypocotyl length	7	Ler x Cvi	Borevitz et al. (2002)
Botto_Coluccio_2007	2007	flowering time and leaf number	8	Ler x Cvi	Botto and Coluccio (2007)
Botto_etal_2003	2003	hypocotyl length	13	Ler x Cvi	Botto et al. (2003)
Buijs_etal_2020	2020	seed germination	9	Ler x Cvi	Buijs et al. (2020)
Darrah_etal_2006	2006	circadian clock	6	Ler x Cvi	Darrah et al. (2006)
Decroocq_etal_2006	2006	symptom development	1	Ler x Cvi	Decroocq et al. (2006)
Edwards_etal_2005	2005	leaf movement period	6	Ler x Cvi	Edwards et al. (2005)
Esch_etal_2007	2007	crossover breakpoints	1	Ler x Cvi	Esch et al. (2007)

Fulcher_etal_2015	2015	telomere length	3	Ler x Cvi	Fulcher et al. (2015)
Harada_etal_2004, 2006	2005	nitrate and potassium content	6	Ler x Cvi	Harada et al. (2004)
Hobbs_etal_2004	2004	oil and fatty acid	7	Ler x Cvi	Hobbs et al. (2004)
Hou_etal_2005	2005	alpha-subunits of 12S globulin cruciferin B (CRB)	1	Ler x Cvi	Hou et al. (2005)
Ito_etal_2007	2007	centrometric satellite repeats	1	Ler x Cvi	Ito et al. (2007)
Joosen_etal_2012	2012	seed germination	94	Bay-0 x Sha	Joosen et al. (2012)
Joosen_etal_2013	2013	seed metabolite	161	Bay-0 x Sha	Joosen et al. (2013)
Juenger_etal_2005	2005	carbon isotope	1	Ler x Cvi	Juenger et al. (2005)
Keurentjes_etal_2007	2007	flowering time	6	Ler x Cvi	Keurentjes et al. (2007)
Kliebenstein_etal_2001, 2002	2002	insect resistance	69	Ler x Cvi	Kliebenstein et al. (2002)
Kobayashi_etal_2005	2005	aluminium tolerance	1	Ler x Cvi	Kobayashi et al. (2005)
Luquez_etal_2006	2006	flowering time, leaf longevity	10	Ler x Cvi	Luquez et al. (2006)
Lustenhouwer_etal_2018	2018	silique number	3	Ler x Cvi	Lustenhouwer et al. (2018)
Payne_etal_2004	2004	Cs uptake, shoot fresh weight	3	Ler x Cvi	Payne et al. (2004)
Raschke_etal_2015	2015	hypocotyl length	5	Bay-0 x Sha	Raschke et al. (2015)
Sangster_etal_2008	2008	hypocotyl and root length	24	Ler x Cvi	Sangster et al. (2008)
Sergeeva_etal_2004	2004	PGM activity	8	Ler x Cvi	Sergeeva et al. (2004)

Sergeeva_etal_2006	2006	hypocotyl length and invertase activity	19	Ler x Cvi	Sergeeva et al. (2006)
Sicard_etal_2008	2008	symptom development and virus accumulation	2	Ler x Cvi	Sicard et al. (2008)
Staal_etal_2006	2006	fungal pathogen resistance	1	Ler x Cvi	Staal et al. (2006)
Swarup_etal_1999	1999	leaf movement	1	Ler x Cvi	Swarup et al. (1999)
Teng_etal_2005	2005	anthocyanin	1	Ler x Cvi	Teng et al. (2005)
Teng_etal_2008	2008	glucose sensitivity	1	Ler x Cvi	Teng et al. (2008)
Ungerer_etal_2002, 2003	2003	flowering time	26	Ler x Cvi	Ungerer et al. (2002)
Vallejo_etal_2010	2010	seed germination	6	Bay-0 x Sha	Vallejo et al. (2010)
Vreugdenhill_etal_2004	2004	seed cation	7	Ler x Cvi	Vreugdenhil et al. (2004)
Wentzell_etal_2007	2007	glucosinolate	126	Bay-0 x Sha	Wentzell et al. (2007)
White_etal_2003	2003	Cs uptake, shoot fresh weight	3	Ler x Cvi	White et al. (2003)
Sangster_etal_2008b	2008	hypocotyl and root length	8	Bay-0 x Sha	Sangster et al. (2008)

Chapter 5.
High-resolution QTL fine-mapping using transcriptomics data

Quantitative Trait Locus mapping using a population of recombinant inbred lines (RIL) is a powerful approach linking genomic regions to quantitative traits. These regions, called quantitative trait loci (QTL), typically contain dozens or even hundreds of genes, and linking individual genes to the trait requires expensive and laborious experimental fine-mapping. The mapping resolution is determined by the number of recombination events in the population and the marker density used for the genetic map. Here we investigate the benefits and challenges of utilizing individual transcriptome-derived SNPs to fine-map QTLs that were mapped using a binned marker-based genetic map in an *Arabidopsis thaliana* RIL population. First, we improve the SNPs' quality by addressing missing data and miscalled genotypes. Then, we perform fine-mapping by identifying SNPs with the highest association with the trait on the QTL interval. This approach can reduce the number of candidate genes of phenotype QTLs to 13% of the original number. Looking at gene expression QTLs (eQTLs), the percentages went down to 9% for distant eQTLs and 17% for local eQTLs. Half of the local eQTL fine-mapped intervals contain the transcript-encoding gene as the likely causal gene, significantly higher than size-matched random intervals. We applied the fine-mapping approach to two QTL hotspots and identified the potential causal genes, including the previously known *MUM2* gene responsible for imbibed seed size traits. The high-resolution fine-mapping presented in this study provides a quick, low-cost alternative to traditional fine-mapping in identifying causal genes underlying QTLs.

In preparation for submission.

Margi Hartanto, Eros Reij, Dick de Ridder, Basten L Snoek, Harm Nijveen

5.1. Introduction

Linkage mapping using bi-parental populations (*e.g.,* recombinant inbred lines, RILs) is a popular method to associate traits with specific genomic regions called quantitative trait loci (QTL) (Collard and Mackill 2008; Joosen et al. 2012; Collard et al. 2005; Jimenez-Gomez et al. 2010). This includes molecular traits, such as transcript levels, using expression QTL (eQTL) mapping (Jansen and Nap 2001). In general, QTL mapping is carried out by finding polymorphic genetic markers significantly associated with the trait values. However, the significant markers typically do not cause the traits themselves but are in linkage disequilibrium (LD) and thus highly correlated with the actual causal variants. Consequently, a QTL is defined as a confidence interval based on the location of the significant markers. The QTL confidence interval is usually defined by determining the markers neighbouring the most significant marker with a decrease of several units of the logarithm of odds (LOD) scores (Collard et al. 2005). Depending on the marker density and the number of recombination effects in the population, a QTL interval may span several million base pairs, containing dozens to hundreds of candidate causal genes. The large QTL interval makes it hard to identify the actual causal gene(s) or variant(s) for further investigation.

The number of candidate genes is traditionally reduced by fine-mapping, *i.e.,* increasing the resolution, using substitution mapping (Paterson et al. 1990). This approach involves crossing two lines with distinct genotypes at the QTL of interest to obtain new recombinant lines for that region. Subsequently, the recombinant individuals are screened by identifying more markers on the QTL region. In an Arabidopsis Bay-0 x Sha RIL population, a fine-mapping experiment has been done using near-isogenic lines (plants with identical genetic backgrounds except for a few loci) to narrow down a zinc-tolerance QTL on chromosome 3 to an interval spanning seven genes, including toxin efflux transporter *FERRIC REDUCTASE DEFECTIVE3* (*FRD3*) as the potential causal gene (Pineau et al. 2012). Another experiment using a heterogeneous-inbred family (genetically identical plants segregating for a genomic region of interest) identifies *NAD-DEPENDENT MALIC ENZYME 1* (*NAD-ME1*) as the most likely gene controlling plant metabolism and circadian clock outputs from 14 candidates on the fine-mapped interval (Francisco et al. 2021). However, while traditional fine-mapping effectively narrows the QTL interval, it is time-consuming and costly.

Besides experimental fine-mapping to introduce additional recombination events in the QTL region, a high-density genetic map based on next-generation sequencing data can also increase QTL mapping resolution (Zhou et al. 2016; Gonda et al. 2019; Huang et al. 2009; Snoek et al. 2021). In the early days of QTL mapping, genetic maps of RIL populations typically had no more than one hundred markers spanning the whole genome (Loudet et al. 2002; Alonso-Blanco et al. 1998), while current approaches using next-generation sequencing (NGS) can identify up to hundreds of thousands of single nucleotide polymorphisms (SNPs) in the population (Zhou et al. 2016; Gonda et al. 2019; Huang et al. 2009). However, sequencing errors preclude using each SNP as a marker; instead, a sliding window or binning approach is used for genotype calling. This approach also reduces the number of markers, thus lowering the multiple testing correction required in QTL mapping. An example is provided by Serin et al. (2017), who constructed a high-density genetic map for an Arabidopsis Bay-0 x Sha RIL population using RNA-seq-based SNPs. SNPs were grouped into 100 kbp bins, and genotypes were assigned based on the consensus, resulting in a set of 1,059 bin markers – which was a substantial improvement over previous Bay-0 x Sha RIL genetic maps (69 markers from (Loudet et al. 2002) and 497 markers from (Zych et al. 2017)). This larger number of markers allowed for the discovery of more recombination breakpoints and increased mapping resolution, thus reducing QTL confidence intervals compared to the other two maps. However, the reduced QTL intervals (a median length of 6 Mbps) were still relatively large, containing many candidate genes.

We hypothesised that the individual SNPs of the Serin et al. study could be exploited to increase the mapping resolution further by discovering novel recombination events and more precise localisation of known breakpoints. We perform high-resolution QTL fine-mapping to investigate this using individual SNPs found in RNA-seq data as markers. We first identified and corrected miscalled genotypes and imputed missing data. Next, we confirm that in local eQTL data, our fine-mapping approach narrows down QTL regions to the location of the causal gene (i.e., the expression of which is used as a molecular trait). Finally, we present several cases where our fine-mapping approach identifies potential causal genes in QTL hotspots.

5.2. Materials and methods

Phenotype and gene expression QTLs

We used seed germination phenotypes and gene expression data of dry seeds matured in four different maternal environments of 161 Arabidopsis RILs from Serin (2018) to test our fine-mapping approach. Seed germination phenotype QTLs (phQTLs) were mapped using the single marker method as in Hartanto et al. (2020). For gene expression QTL (eQTL) analysis, the gene expression data from RILs matured in all four different maternal environments were combined to identify the genetic main effect eQTLs, *i.e.,* the QTL effect without interaction with the environment (Serin 2018). We identify the original location of phQTLs and eQTLs using the 1,059 bin markers found in (Serin et al. 2017). QTL signals and peak markers were detected using the "signal.find_peaks" function from the Scipy Python library (version 1.7.3). A LOD threshold of 4.1 was used to determine significant QTLs, as in Serin (2018). QTL intervals were defined as the region between the upstream and downstream markers where a 1.5 LOD drop from the QTL peak marker was reached. We defined an eQTL as local if the gene encoding the transcript is located within 1 Mbp from the peak marker and the eQTL interval; otherwise, it is defined as distant (Cubillos et al. 2012). The physical locations of the Arabidopsis genes were retrieved from the TAIR 10 genome release (https://www.arabidopsis.org/download_files/Genes/TAIR10_genome_release/TAIR10_gff3/TAIR10_GFF3_genes.gff).

RNA-seq-based SNPs

A merged variant call format (VCF) file of 12 Bay-0, 12 Sha, and 160 Bay-0 x Sha Arabidopsis RIL samples was obtained from Serin et al. (2017). A Python script was written to generate a genotype file and ordering them based on physical location. SNPs close to another SNP (< 100 bp) were considered redundant and the downstream one are removed as they are potentially in a high linkage disequilibrium. We observed that some homozygous genotypes were misassigned as heterozygous; we corrected these based on the Phred-scaled likelihoods of the genotype. We loaded the map data into the R environment as a cross object using the "read.cross" function of the qtl library (version 1.52). We set the "crosstype" to "riself" and the "F.gen" parameter to 6 to indicate the F6 generation of selfed RILs. The residual heterozygosity is 3-4%; however, most heterozygous SNPs are either out-of-phase with the surrounding markers or only present in a few lines. Therefore, we treated these as likely errors and replaced them with NA. Further genotyping error detection and correction were performed using the "calc.errorlod" function of the qtl library and a custom R function. The errors were identified based on double-crossovers spanning only one or two SNPs (Broman and Sen 2009) and fixed based on the genotype of the surrounding markers. For example, "AAAABBAAAA" is replaced with "AAAAAAAAAA". This approach does not work for missing genotypes in recombinant intervals (e.g., AAA-----BBB); therefore, we left them unchanged. Missing genotypes in non-recombinant intervals (*e.g.,* AAA-----AAA) were replaced by the genotype of surrounding markers (*e.g.,* AAAAAAAAAAA).

High-resolution fine-mapping

High-resolution fine-mapping was performed in phQTL and eQTL intervals by associating the RNA-seq-based SNPs with the trait values using the single marker analysis method. For comparison, we also performed fine-mapping using the binned genetic map of (Serin et al. 2017) and a genetic map where SNPs were binned per gene. The fine-mapped QTL interval was defined as the region where the LOD score reached its maximum, including a flanking marker, each upstream and downstream. The number of candidate genes was compared between the original QTL and the fine-mapped QTL.

Simulation

To obtain a background probability distribution of the occurrence of causal genes by chance, we simulated randomly positioned intervals on each local eQTL with sizes based on those of the observed fine-mapped local eQTLs. Local eQTLs include (or are near to) the gene encoding the transcript and can arise due to the variation in *cis*-regulatory elements or gene itself (known as local- or *cis*-eQTL) (Rockman and Kruglyak 2006; Jansen and Nap 2001). Because the transcript-encoding gene can be assumed causal (Jimenez-Gomez et al. 2010; Snoek et al. 2012; Cubillos et al. 2012), local eQTLs provide a means to validate our approach. We noticed that crossover events are absent in around 11% of local eQTLs, resulting in the fine-mapped region spanning almost the whole interval; we excluded these eQTLs from our simulation. The probability of finding a causal gene by chance was calculated as the average percentage of random intervals containing a causal gene over 1,000 repetitions.

Prioritization of candidate genes at QTL hotspots

We used our previously developed knowledge graph (described in chapter 4) to prioritize candidate genes at fine-mapped QTL hotspots. In brief, the database contains gene-gene functional interactions predicted based on gene annotation, protein-protein interaction, and regulatory information. We used this information to create gene interaction networks where nodes represent genes and edges represent the interaction between these genes. The network was created using genes with an eQTL at the hotspot as target genes and genes in the hotspot interval as candidate regulatory genes. Each candidate was scored and prioritized based on node eigenvector centrality as one of the most accurate centrality measures to determine hub genes (Ozgur et al. 2008). A high eigenvector centrality score means the gene is connected to many genes with high scores and vice versa. We calculated eigenvector centrality using the "eigenvector_centrality" function implemented in the NetworkX Python package (version 2.6.3).

Gene ontology enrichment

Gene ontology enrichment analyses were performed using the TopGO library in R (Alexa et al. 2006) with the algorithm's default 'weight01' parameter, which is a mixture of the 'elim' and 'weight' methods. Fisher's exact test is used to calculate the statistical significance of the GO term enrichment of a gene set with an eQTL at the hotspot with respect to a set of background genes: all genes having expression levels significantly higher than 0, determined using single sample t-tests (Bonferroni adjusted $p < 0.05$).

5.3. Results

Serin et al. (2017) showed that a high-density genetic map consisting of 1,059 bin-markers for the Bay-0 x Sha RIL Arabidopsis population increased the resolution of QTL mapping for 510 published seed germination phenotypes compared to previous marker sets (Loudet et al. 2002; Alonso-Blanco et al. 1998). However, the number of candidate genes at each QTL remains relatively large. In this study, we used the individual SNPs found in the RNA-seq data Serin et al. (2017) used to further fine-map the QTL intervals. We first identify and correct likely genotyping errors and impute missing alleles. Second, we test our fine-mapping approach on local eQTLs to verify that the fine-mapped intervals indeed contain the likely causal genes, based on the assumption that most local eQTLs are *cis*-eQTLs, i.e., that the causal variant is in the tested gene. Finally, we present a case in which our fine-mapping approach successfully narrows down a hotspot QTL region to the location of potential causal genes.

Missing data imputation and genotype correction of the high-resolution genetic map

The quality of QTL mapping relies on the quality of the genetic map. In our case, the genetic map constructed from 17,148 RNA-seq-based SNPs still contains many missing genotypes and potential genotyping errors (Figure 5.1A). To fix this, we imputed the missing data and corrected possibly miscalled genotypes. Missing genotypes in non-recombinant intervals were imputed by the genotype of surrounding markers. Imputation drastically reduces the percentage of missing genotypes from 5.5% to 0.35% over all RILs. Additionally, unusually close double crossovers spanning only one to two SNPs were identified as potential genotyping errors and replaced by the parental allele of the surrounding markers. This method significantly decreases the number of short genotype blocks (two markers or less) from 1,658 to 138. The raw genetic map before and after imputation and correction and the binned version for the first 2 Mbp of chromosome 1 are shown in Figure 5.1.

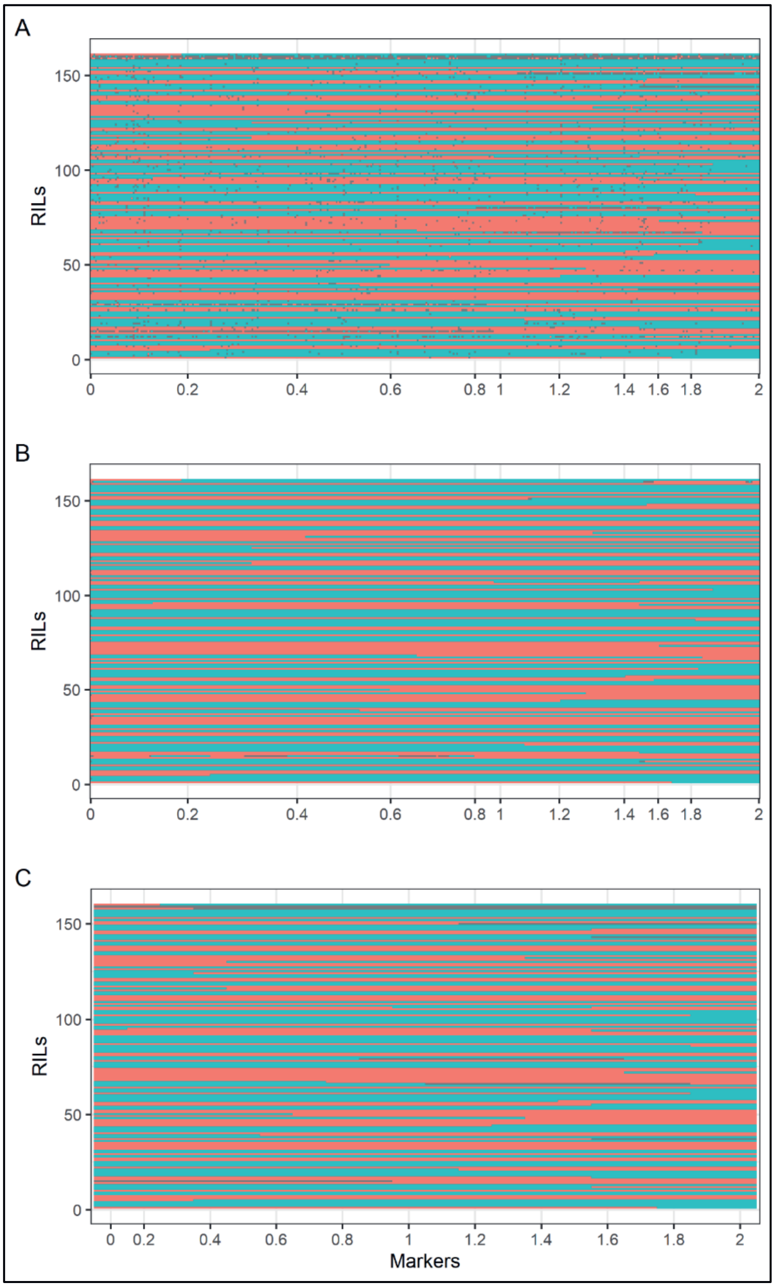


Figure 5.1. The high-resolution genetic map of the 160 RILs for the first 2 Mb of Chromosome 1 before (A) and after imputation and correction (B), and the binned version of the genetic map (C). Each row correspond to a RIL. Columns represent the genetic marker physically anchored on the chromosome. The genotypes of Bay-0 and Sha are indicated by red and green bar, respectively. The missing genotypes is indicated by grey bars.

High-resolution fine-mapping significantly reduces the number of candidate genes

The high-resolution fine-mapping was done by analysing the QTL region using RNA-seq-based SNP markers that are much denser than the binned markers (Serin et al. 2017). Here we present a fine-mapping example of the *AT4G36760* local eQTL at the tail of chromosome 4, containing 55 candidate genes (Figure 5.2A). We analysed the association of 45 SNP markers on the eQTL with the transcript level of *AT4G36760*. As we hypothesized, denser markers uncover more recombination breakpoints, indicated by an increase in mapping resolution (Figure 5.2B). SNP markers in the fine-mapped interval have much higher LOD scores than the binned marker, allowing for more precise localisation of the QTL region. We define the fine-mapped interval using the locations of markers with the highest LOD scores, ranging from 17.31 Mbp to 17.34 Mbp, reducing the number of candidate genes from 55 to 15. Furthermore, we identified RILs where recombination occurs upstream (RIL65, RIL131, and RIL164) and downstream (RIL21 and RIL95) of the fine-mapped interval. For the traditional fine-mapping approach, this information will help identify individuals with desired recombination patterns to produce the desired offspring with a small QTL interval. Plotting the gene expression values of the RILs shows that the effect of parental alleles on the *AT4G36760* transcript level is the highest in the fine-mapped interval compared to the upstream and downstream intervals (Figure 5.2C).

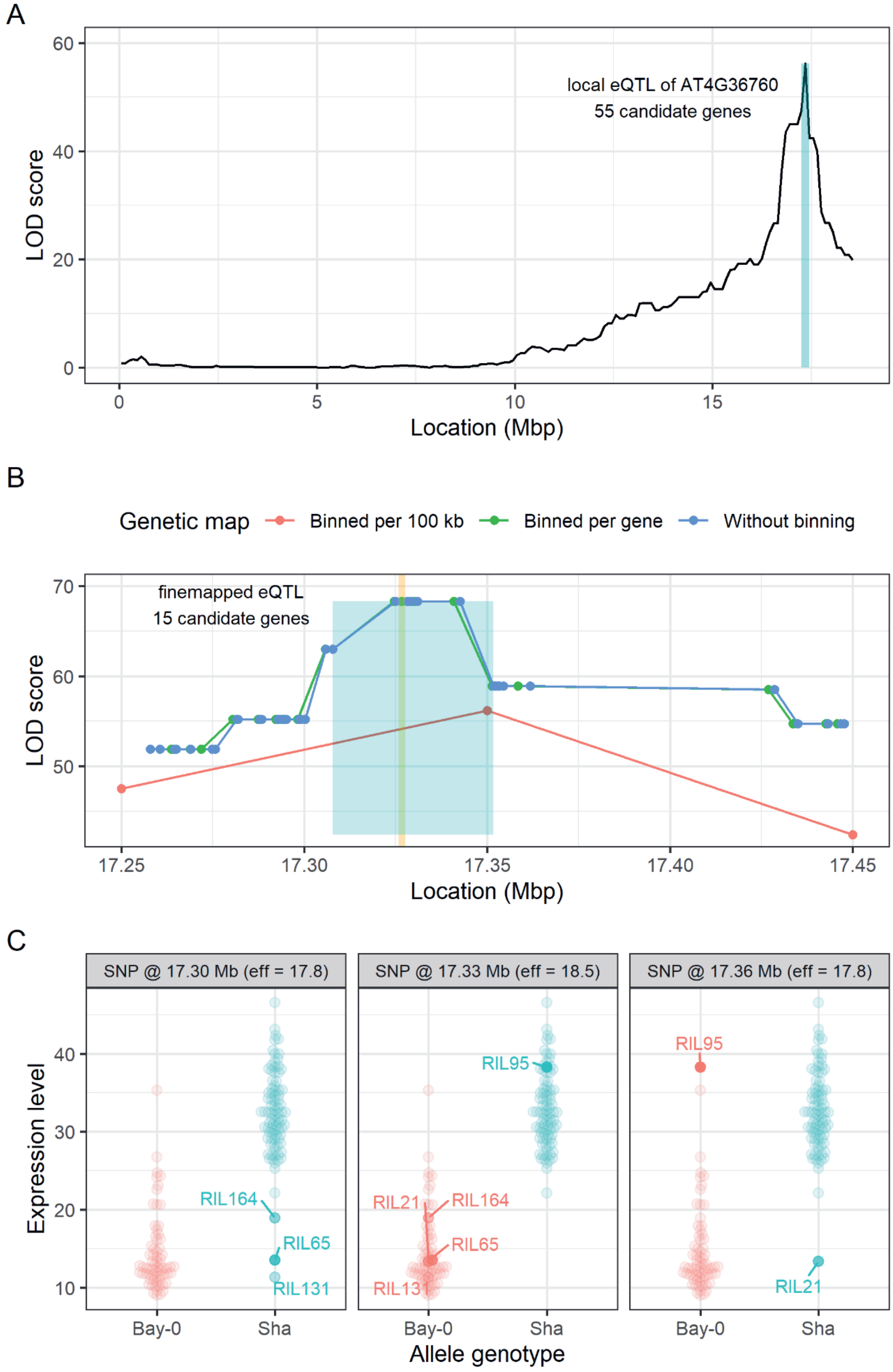

A
LOD score
60
40
20
0
0
5
10
15
Location (Mbp)
local eQTL of AT4G36760
55 candidate genes
B
Genetic map
Binned per 100 kb
Binned per gene
Without binning
LOD score
70
60
50
17.25
17.30
17.35
17.40
17.45
Location (Mbp)
finemapped eQTL
15 candidate genes
C
SNP @ 17.30 Mb (eff = 17.8)
SNP @ 17.33 Mb (eff = 18.5)
SNP @ 17.36 Mb (eff = 17.8)
Expression level
40
30
20
10
RIL164
RIL65
RIL131
RIL95
RIL21
RIL164
RIL65
RIL131
RIL95
RIL21
Bay-0
Sha
Bay-0
Sha
Bay-0
Sha
Allele genotype

Figure 5.2. High-resolution fine-mapping of a local eQTL for the AT4G36760 gene. The local eQTL was first identified using binned markers (A). Fine-mapping the eQTL using all individual SNP markers narrowed the interval to a smaller region including the gene (B). The fine-mapped region is shown in blue, the location of the gene is indicated by a yellow vertical bar. Fine-mapping uncovered additional recombination breakpoints, with recombination events in RIL65, RIL131, and RIL164 upstream of the fine-mapped region and downstream for RIL21 and RIL95. The dot plots (C) show the expression levels of the gene separated by the parental allele, upstream (17.30 Mbp), within (17.33 Mbp), and downstream (17.35 Mbp) of the fine-mapped region. eff = eQTL effect on the gene expression.

To validate our fine-mapping approach more broadly, we applied it to 577 seed germination phenotype QTLs (phQTLs), 7,044 dry seed local eQTLs, and 15,076 distant ones eQTLs (Serin 2018). Generally, the number of candidate genes is reduced for all types of QTL data. The median number of candidates within the original phQTLs, local eQTLs, and distant eQTLs are 438, 207, and 413 genes, respectively; after fine-mapping, these numbers are reduced to 32, 40, and 35, respectively, corresponding to 13%, 9%, and 17% of the total number of original candidate genes (Figure 5.3). Even though the final numbers of candidate genes are similar for the three QTL types, fine-mapping yields a larger reduction for distant eQTLs and phQTLs than for local eQTLs. It is because local eQTLs usually already have small intervals, rendering the fine-mapping approach less effective in reducing the number of candidate genes. Looking further, we learned that our fine-mapping approach is only effective in the presence of crossovers. For example, increasing the resolution of the *AT5G09240* local eQTL does not reduce the region's size due to the absence of a crossover (Figure S5.1).

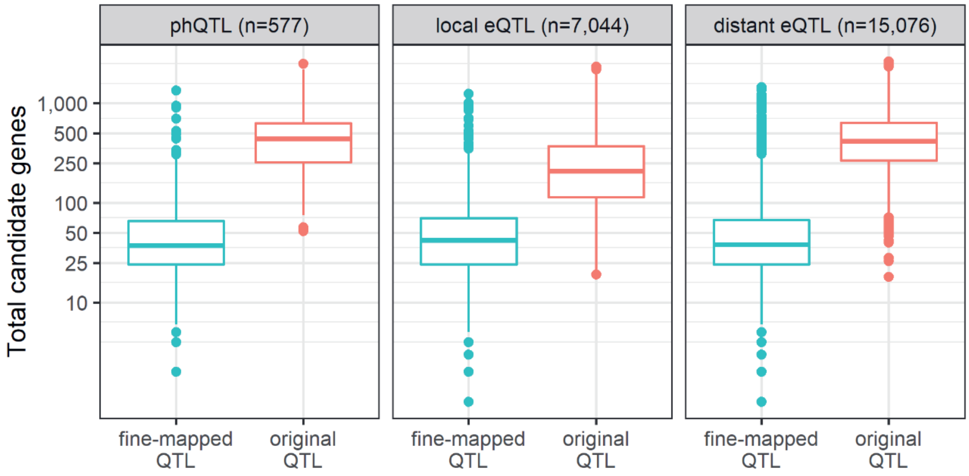


Figure 5.3. Boxplots comparing the number of QTL candidate genes before and after fine-mapping for different QTL types. Fine-mapping strongly reduces the number of candidate genes per QTL. The y-axis is $\log_{10}$ scaled.

We next set out to analyse whether our fine-mapping approach indeed narrows down the QTL interval to the location of the likely causal gene. It is not trivial because for only very few QTLs, the actual causal genes are known (Chapter 3). As an alternative, we used local eQTLs, mapped near or on the location of the gene whose expression is the molecular trait. It is generally assumed that most of these are *cis*-eQTL, where the causal variant lies in the gene itself (either in the coding sequence or in cis-regulatory elements) (Rockman and Kruglyak 2006). We analysed the fine-mapped local eQTLs and determined that 51% contained the transcript-encoding gene. This percentage is significantly higher ($p < 0.001$) than that in randomly generated size-matched fine-mapped intervals on local eQTLs (26±0.4% on average) (Figure 5.4A). For the remaining 49% of local eQTLs, where the fine-mapped region does not contain the gene, the distance between the fine-mapped interval and the gene is smaller than that of randomly generated size-matched intervals (Figure 5.4B). Furthermore, the original intervals are larger (Figure 5.4C), and the average distance between the peak marker and the transcript-encoding gene (Figure 5.4D) is higher than in the local eQTLs where the fine-mapped interval contains the gene. The larger distance between the peak marker and the gene may indicate that these are local *trans*-eQTLs rather than *cis*-eQTLs.

High-resolution fine-mapping of multi-omics QTL hotspots supports multiple causal genes causing the hotspots

We applied the method to prioritize potential causal genes in seed germination phQTL hotspots. phQTL hotspots are locations on the genome where multiple phQTLs collocate. These are assumed to be due to a polymorphism in one or a few master regulators on the loci (Breitling et al. 2008). In our data set, two of the most prominent phQTL hotspots are located on chromosome 5 at 20.5-21.5 Mbp (referred to as hotspot_21, with 290 candidate genes) and 25-26 Mbp (hotspot_25; 302 candidate genes) based on the colocation of 46 and 57 phQTLs (Figure 5.5). Both phQTL hotspots are colocated with many distant eQTLs: 222 eQTLs at hotspot_21 and 301 distant eQTLs at hotspot_25. Identifying master regulators causing these hotspots would help understand the molecular mechanisms underlying the seed germination phenotypes. We address this problem by fine-mapping each of the phQTLs and eQTLs located on the hotspots to pinpoint the potential location of the regulators.

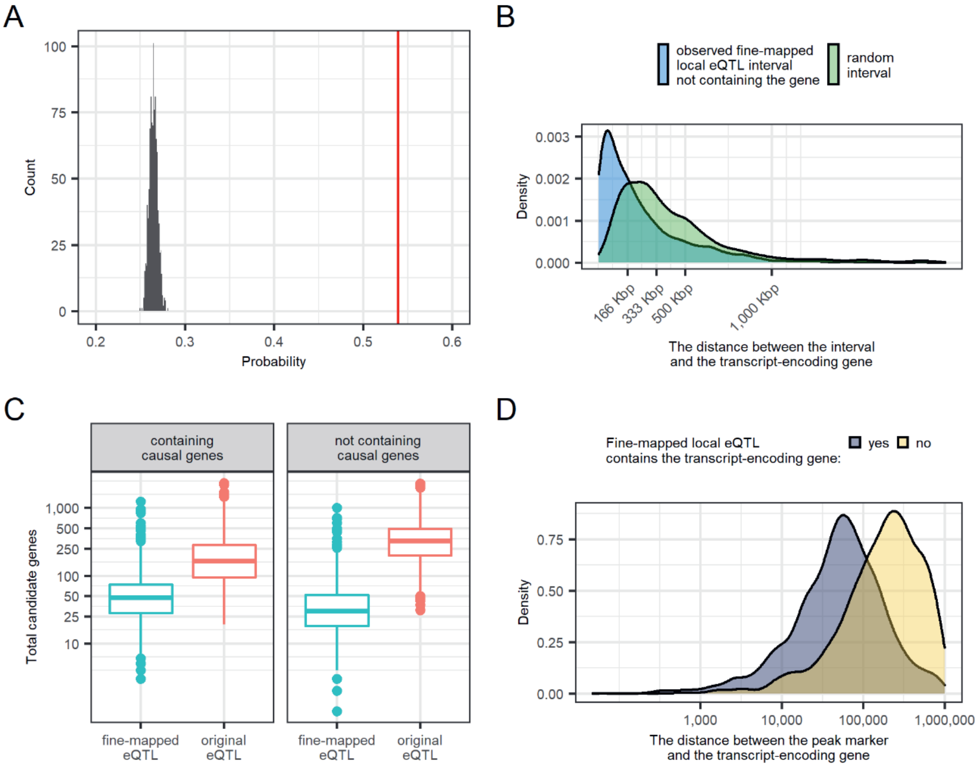


Figure 5.4. For local eQTLs, the location of the gene with respect to the fine-mapped region. (A) The observed proportion of fine-mapped local eQTLs containing the transcript-encoding gene (red line) and the probability distribution of selecting a random region in the original local QTL that contains the gene. The distribution is calculated using 1,000 randomly positioned size-matched intervals (black). (B) Distance in kilo base pairs to the actual gene location for fine-mapped local eQTLs that do not contain the gene (blue), compared to randomly positioned size-matched intervals (green). (C) Distribution ($\log_{10}$-scaled) of candidate gene counts in the original local eQTLs and the fine-mapped eQTLs that do (left) or do not (right) contain the transcript-encoding gene. (D) Distribution ($\log_{10}$-scaled) of the distance (in base pairs) between the peak marker and the transcript-encoding gene in cases where the fine-mapped intervals do or do not contain the gene.

Fine-mapping phQTLs and eQTLs at the hotspot narrows the intervals into several overlapping ones (Figure 5.5). This overlap supports the robustness of our fine-mapping approach. The presence of multiple fine-mapped intervals may indicate that the hotspot contains multiple causal genes. If this is true, we can decompose the hotspot based on the smaller fine-mapped intervals and analyse these separately to find the causal genes. The result of these analyses is presented in Table 5.1. At hotspot_21, most phQTLs associated with seed germination under osmotic and cold stress were fine-mapped to 21,017,143-21,121,588 bp, with 31 candidate genes (referred to as hotspot_21_1). This region is also supported by 38 fine-mapped distant eQTLs where the genes are enriched for several GO terms, including 'response to water deprivation' (3-fold enrichment; p-value = 0.032) which is related to seed germination in osmotic stress. A small list of candidate genes (31) makes it feasible to examine them manually, for example, using the Gene Ontology (GO) annotations of the genes based on the trait under study. Accordingly, we found *ABA-HYPERSENSITIVE GERMINATION 1* (*AHG1*; *AT5G51760*) in the region, a gene annotated with 'response to abscisic acid' GO term. In addition, using a gene interaction network*, C-REPEAT-BINDING FACTOR 4* (*CBF4*; *AT5G51990*) is the top prioritized gene at hotspot_21. There is no record of the involvement of this gene in seed germination; however, recent studies suggest that *CBF4* is involved in osmotic stress response (Vonapartis et al. 2022) and cold response (Vyse et al. 2022).

At hotspot_25, the phQTLs were fine-mapped to at least two smaller intervals, with the majority at 25,433,663-25,670,432 bp (referred to as hotspot_25_1). phQTLs mapped to hotspot_25_1 are associated with imbibed seed size and germination under ABA treatment. By examining the GO terms of the candidate genes, we found *MUCILAGE-MODIFIED 2* (*MUM2*; *AT5G63980*), in which the accession Sha contains a polymorphism affecting the imbibed seed size (Macquet et al. 2007). *MUM2* was also proposed as the potential causal gene for imbibed seed size traits in another study using the Bay-0 x Sha RIL population (Joosen et al. 2012). The fine-mapping of many seed-size phQTLs to the location of *MUM2* suggests that our approach effectively pinpoints the causal gene's location. Besides *MUM2*, we also found *SAL1* as a potential causal gene for seed germination under ABA treatment based on prioritization using an interaction network of genes with eQTLs on hotspot_25_1. This finding is supported by an experiment showing that *SAL1* mutants were more sensitive to ABA, resulting in low seed germination (Pornsiriwong et al. 2017).

Besides those on hotspot_21_1 and hotspot_25_1, the remaining phQTLs were fine-mapped to other smaller intervals, some overlapping. Considering that only a few phQTLs support these fine-mapped intervals, we doubt whether they are accurate or may be due to genotyping errors. Prioritizing the genes in one interval, 25,649,503-25,737,045 bp (hotspot_25_2), led to a promising top prioritized candidate gene *EXORDIUM LIKE 2* (*EXL2*; *AT5G64260*). Based on inference from electronic annotation, *EXL2 is* involved in response to ABA and water deprivation (Depuydt and Vandepoele 2021), which is related to phQTLs for seed germination in osmotic stress and after ABA treatment that was fine-mapped to hotspot_25_2. Despite the evidence we provide here, experimental validation is needed to confirm the association of these genes to the traits. Altogether, we provided an example of fine-mapping a large number of QTLs to identify causal genes at QTL hotspots, which is unfeasible using the traditional fine-mapping approach.

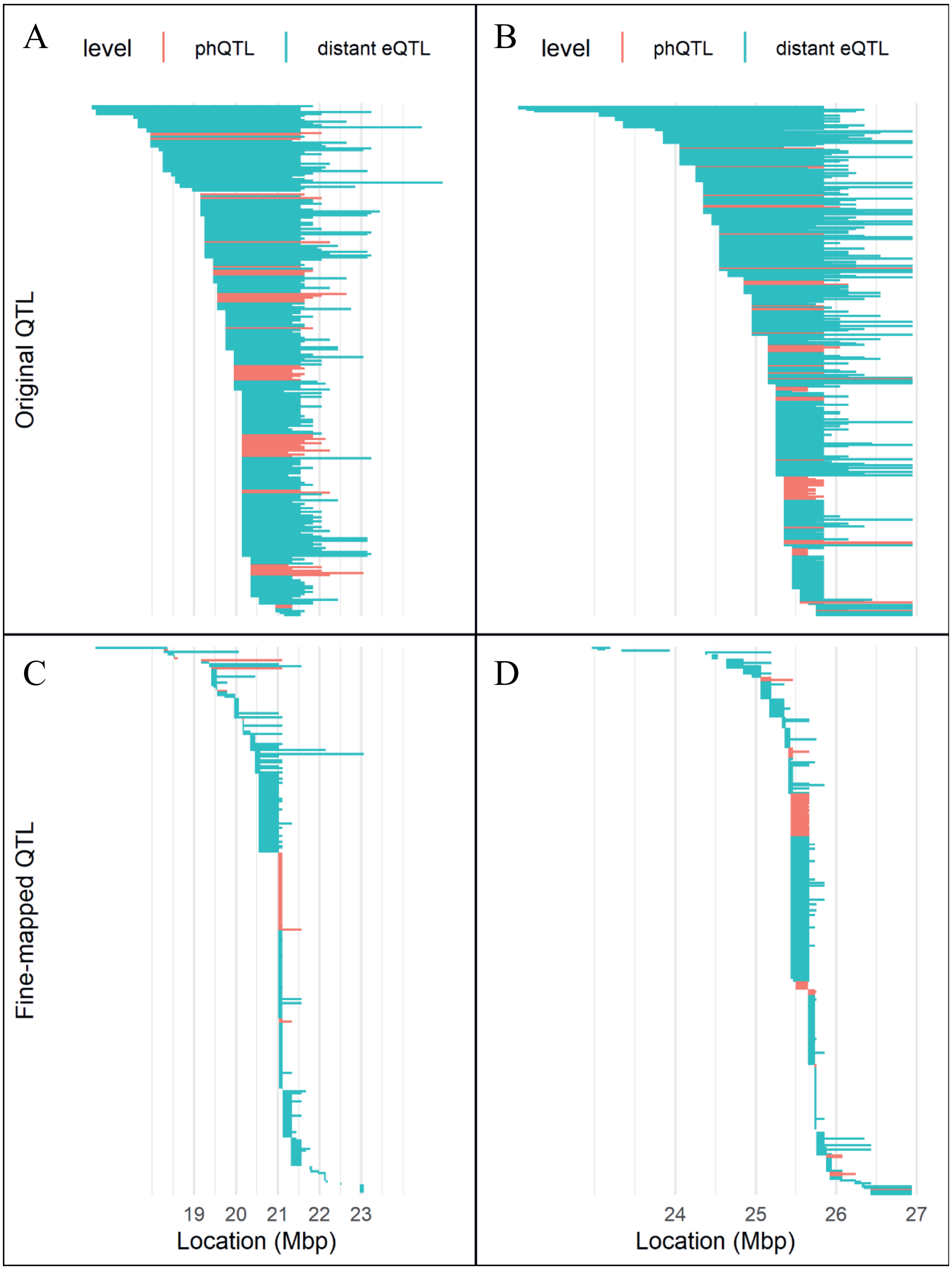


Figure 5.5. Comparing the original QTL regions with their fine-mapped counterparts on two hotspots, for phQTLs (red) and distant eQTLs (blue) on chromosome 5. The original phQTL and eQTL intervals colocated at around 21 Mbp (A) and 25 Mbp (B), sorted by start position. Fine-mapped intervals of the same QTLs colocated on a hotspot at 21 Mbp (C) and 25 Mbp (D).

Table 5.1. The two most prominent phQTL hotspots with a summary of the fine-mapping and gene prioritization analyses.

Hotspot ID (interval range in Mb)	Total candidate genes	Total phQTLs	Total eQTLs	Fine-mapped hotspot ID	Interval range	Total candidate genes	Fine-mapped phQTLs	Traits	Fine-mapped eQTLs	Enriched GO terms	Prioritized genes based on GO (relevant GO term)	Top 3 prioritized genes based on the interaction network
hotspot_21 (20.5—21.5)	290	46	222	hotspot_21_1	21,017,143 — 21,121,588	31	36	-Seed germination in osmotic -Seed germination in cold stress	38	-Response to water deprivation -Purine nucleotide metabolic process, etc.	-*AHG1* (response to abscisic acid), - *SBT1.3* (seed development) -*PGMP* (response to cold)	*CBF4* *TCP19* *PGMP*
hotspot_25 (25—26)	302	57	301	hotspot_25_1	25,433,663 — 25,670,432	66	33	-Imbibed seed size -Seed germination in ABA treatment	81	- Regulation of transcription , auxin-activated signaling pathway, -Response to temperature stimulus, etc.	-*MUM2* (mucilage biosynthetic process involved in seed coat development) -*FCLY* (ABA-activated signaling pathway), *SAL1* (ABA-activated signaling pathway)	*NAC103* *SAL1* *FLS2*

				hotspot_25_2	25,649,503 — 25,737,045	28	3	-Seed germination in osmotic stress of the high-light maternal environment -Seed germination in ABA treatment of the high-temperature maternal environment	42	-Abscisic acid-activated signaling pathway, -Cellular response to DNA damage stimulus, -mRNA processing, etc.	- *AT5G64230* (response to water deprivation, response to ABA), -*EXL2* (response to salt stress, response to ABA, response to temperature stimulus)	*CAMTA2* *EXL2* *AT5G64270*

5.4. Discussion

In this study, we investigated the hypothesis that fine-mapping using high-density SNP markers effectively increases the QTL mapping resolution and reduces the number of candidate genes. We developed an approach to fine-map QTLs using RNA-seq-based SNP markers. The fine-mapping of 544 phQTLs, 7,044 local eQTLs, and 15,076 distant eQTLs results in drastic reductions in the numbers of candidate genes of these three different QTL types, leaving only 13%, 9%, and 17% of the original median numbers of candidate genes, respectively. As a validation of the approach, we found that half of the fine-mapped local eQTLs contain the genes encoding the transcript as the most likely causal ones, which suggests that our fine-mapping approach is practical for a large number of cases. We also presented the application of high-resolution fine-mapping on two QTL hotspots and showed that most QTLs were fine-mapped to a smaller interval, containing the potential causal genes. One gene is *MUM2*, a previously described causal gene that affects imbibed seed size (Macquet et al. 2007; Joosen et al. 2012).

A limitation of high-resolution fine-mapping is its heavy reliance on existing recombination breakpoints in the population, as the approach is ineffective without it (Figure S5.1). Lack of recombination occurs for around 11% of local eQTLs, where the fine-mapped intervals are almost as large as the original ones. The QTL interval can only be narrowed for these cases by introducing more recombinations using the traditional fine-mapping approach. Even in the presence of recombination, the traditional fine-mapping approach can add more recombination and reduce the number of candidates down to a few genes, as in Pineau et al. (2012). However, fine-mapping using substitution mapping is much more expensive and time-consuming than our fully computational approach.

Another important limitation is SNP genotype accuracy. SNP genotyping using next-generation sequencing technology is prone to errors, especially for low read depths. For example, it is estimated that the genotyping error of a high-density genetic map of rice RILs based on whole-genome resequencing can be as high as 0.71-4.12% (Huang et al. 2009), which is affected by the quality of the parental reference genome sequence as well. Moreover, in genotype calling using RNA-seq, the variation in gene expression level may affect read depths and, thereby, the quality of genotypes. For example, for a gene with very low expression, there is a high probability that the reads are mapped to only one of the two genotype variants of diploid individuals (Nielsen et al. 2011; Zhou et al. 2016). We took several measures to identify and correct miscalled genotypes. However, these are still susceptible to error. Our naive approach of assigning 1-2 out-of-phase markers as miscalled genotypes might nullify the occurrence of gene conversions that are potentially causal for QTLs (Wijnker et al. 2013). Replacing missing This and other remaining mislabelled genotypes can alter the genetic map, resulting in incorrectly fine-mapped intervals (Zych et al. 2017). Future work should focus on comparing different strategies of variant and genotype calling, missing data imputation, and genotype error correction, including using DNA resequencing data to improve the quality of the genetic map and fine-mapping results.

Next to possible errors in the genetic map, there may also be biological reasons for a fine-mapped local eQTL interval being located elsewhere than the gene encoding the transcript. First, local eQTLs may be due to variations in *cis*-regulatory elements located far away from the gene, such as enhancers. Zhu et al. (2015) analysed 10,044 putative intergenic distant enhancers in Arabidopsis and found that an enhancer can be located as far as 200 Kbp from the nearest gene, which is within the distance between a gene and its local eQTL as defined in this study (1 Mbp). Second, the causal gene may not be the transcript-encoding gene but can still be located nearby (local *trans*-eQTL). To evaluate the presence of *trans*-eQTLs in a local eQTL data set, we modelled the distribution of the distance between the gene and its *cis* and *trans* eQTL on the same chromosome using local and distant eQTL, respectively, as in Fauman and Hyde (2022). According to the model (Figure S5.2), an estimated small proportion of *trans*-eQTLs is present in the pool of local eQTLs. The presence of *trans*-eQTLs may explain why some fine-mapped regions do not contain the gene encoding the transcript. For 57% of these assumed local *trans*-eQTLs, we found potential causal genes, including some transcription factors, based on the gene interaction database.

Despite these limitations, our work clearly demonstrates the feasibility of high-resolution fine-mapping, which works for many cases based on validation using local eQTL data using individual SNPs of NGS-based genetic maps. The main advantage of this method is that it can be performed immediately after QTL mapping without the need to develop a new plant population and perform genotype screening for recombinants. Nevertheless, the fine-mapping method presented in this study will not replace the traditional fine-mapping needed to obtain experimentally validated results. The main application of our effective method is to identify causal genes, particularly in cases where many QTLs are to be analysed (*e.g.,* colocated on a hotspot) and as an extension to and refinement of the traditional fine-mapping approach.

Supplementary materials

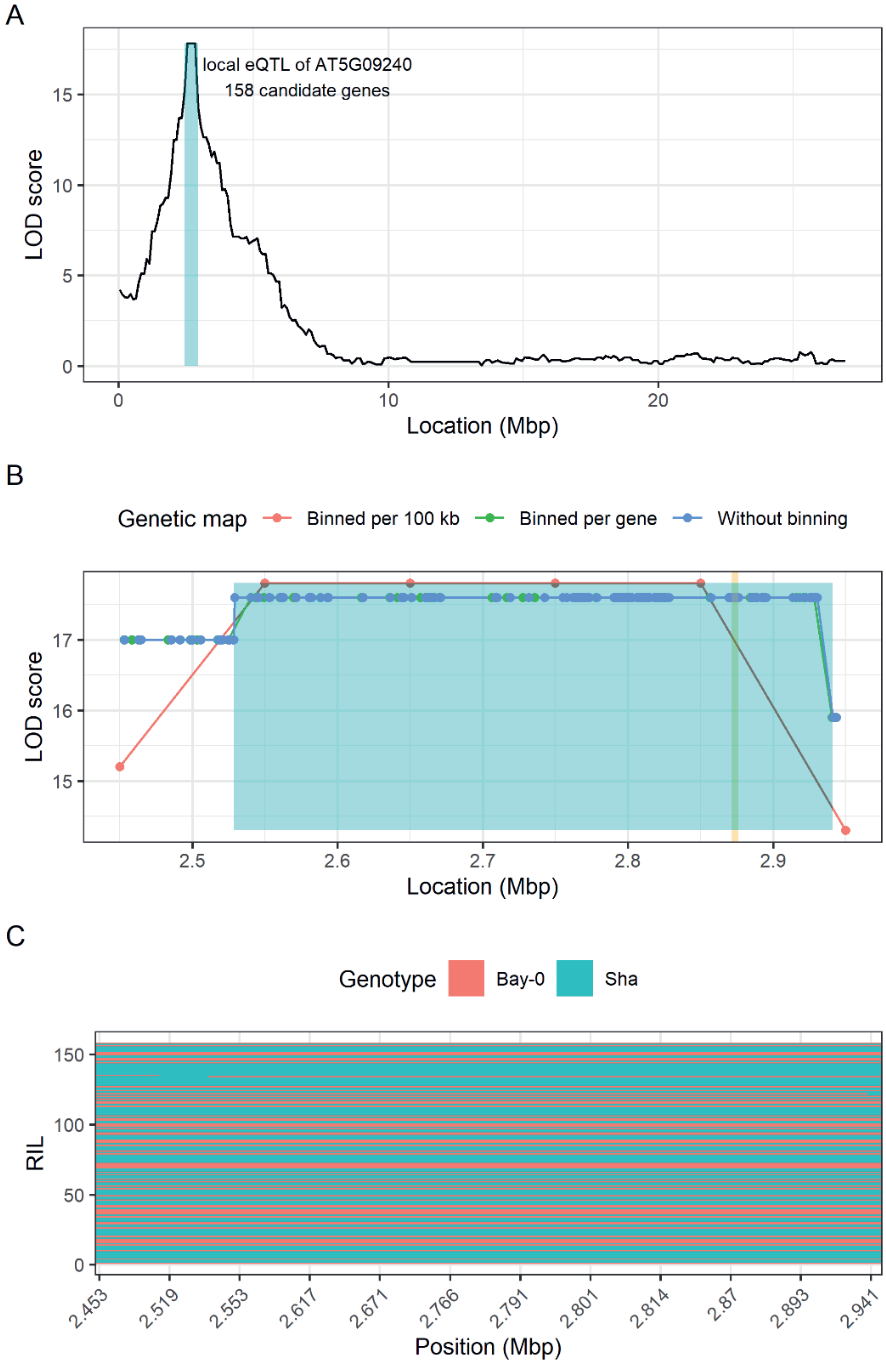


Figure S5.1. An example where fine-mapping is ineffective in reducing the eQTL interval. The local eQTL of AT5G09240 is located at the beginning of chromosome 5 (A). Fine-mapping does not reduce the interval (B). The genetic map for the eQTL region shows a lack of crossover events, explaining the ineffectiveness of the fine-mapping approach (C).

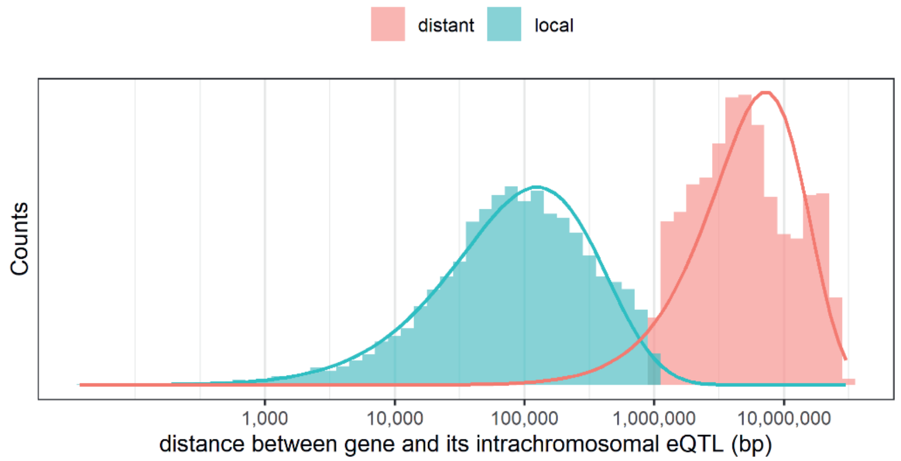


Figure S5.2. The Weibull distribution of the distance between genes and their eQTLs mapped on the same chromosome. The two distinct distributions are assumed to be the distribution of intrachromosomal *cis*- (blue line) and *trans*-eQTL (red line), which are fitted using local (peak marker-gene distance <=1 Mbp; blue bars) and distant eQTL (peak marker-gene > 1 Mbp; red bars). The x-axis is $\log_{10}$ scaled.

Chapter 6. Discussion

Systems genetics is an attractive approach to analyse the molecular machinery that underpins plant traits by inferring gene regulatory networks through genetical genomics (*i.e.,* eQTL mapping) to link genes with their regulators. However, the often low mapping resolution complicates the identification of such regulators. In this thesis, I addressed this problem by developing gene prioritization methods to explore potential regulators using prior knowledge. In Chapter 2, I provided a systems genetics use case and demonstrated a gene prioritization method for QTL hotspots using co-expression networks. In Chapter 3 and Chapter 4, I continued developing methods to identify gene expression regulators using prior knowledge, based on machine learning and using prior knowledge. Finally, I directly addressed the low mapping resolution by utilizing SNPs derived from RNA-seq to fine-map QTLs in Chapter 5. In the following sections, I will describe some considerations when developing gene prioritization methods using the systems genetics approach and offer my perspective on addressing them. After that, I will provide some recommendations to continue this work, particularly to apply the approach to crops. Finally, I will present my point of view regarding the implementation of systems genetics in plant breeding and the challenges involved.

6.1. Experimental considerations in systems genetics studies

Low mapping resolution

The low resolution of QTL mapping, typically yielding tens to hundreds of candidate causal genes, was the motivation for developing the gene prioritisation methods described in this thesis. Obviously, increasing the resolution of QTL mapping would reduce the number of candidates needed to be evaluated in gene prioritisation or even make it less necessary if only a few genes remain. The low resolution is both a detection issue as well as an inherent property of the used populations. The number of markers used to genotype a population is usually too low to map all recombination events precisely. Previously, approaches using next-generation sequencing data have been proposed to increase the mapping resolution, for example, by developing a high-density genetic map (Huang et al. 2009; Serin et al. 2017; Gonda et al. 2019) or fine-mapping using individual SNPs (Chapter 5). In an Arabidopsis biparental population, a high-density map could reduce the QTL size to a median of ~1 Mb containing ± 200 genes (Serin et al. 2017), while fine-mapping using individual SNPs further reduced it to ~300 kb containing ± 60 genes (Chapter 5).

While dense genetic markers address the detection constraints on mapping resolution, the limited recombination frequency in biparental recombinant inbred line (RIL) populations still causes many genes to be inherited together. A much higher mapping resolution can be achieved in genome-wide association studies (GWAS) using a natural population. However, detecting low-effect or rare variants in such a population is often difficult because of a lack of statistical power, requiring a large number (hundreds to, preferably, thousands) of individuals (Korte and Farlow 2013). The same issue prevents most GWAS experiments from detecting *trans*-eQTLs (see Chapter 1.3) (Liu et al. 2019), while they are important for inferring gene expression regulation underlying traits (Chapters 2, Chapter 3, Chapter 4). Although individual *trans*-eQTLs usually have small effect sizes, their combination has been found to explain a more significant fraction of gene expression variation than *cis*-eQTLs (Albert et al. 2018; Liu et al. 2019). The lack of statistical power for *trans*-eQTL GWAS analysis prevents the identification of gene expression regulators underpinning plant traits, prohibiting systems genetics study.

The choice to use natural or biparental populations is thus a trade-off between statistical power and mapping resolution. These two aspects can be balanced using multi-parent advanced generation inter-cross (MAGIC) populations (Kover et al. 2009; Huang et al. 2011; Pascual et al. 2016; Scott et al. 2020). A MAGIC population is constructed by crossing several founder lines multiple times, followed by selfing to generate inbred lines. This complex breeding scheme results in higher recombination rates and, thus, a higher mapping resolution than in biparental populations but still provides high statistical power as the crossing scheme prevents the allele frequencies from becoming too low (Kover et al. 2009; Pascual et al. 2016; Scott et al. 2020). A study in tomato showed that QTL mapping using a MAGIC population results in smaller QTL intervals than a biparental RIL population, as it provided almost twice the number of recombination breakpoints but also found approximately twice the number of QTLs compared to GWAS (Pascual et al. 2016). Future studies should thus consider using MAGIC populations to increase mapping resolution without losing statistical power to detect low-effect variants.

Technical limitations leading to false positives

eQTL mapping results must be interpreted carefully because of false positives due to limitations in the method used to measure mRNA levels and in the experimental design. For example, when using microarrays, sequence polymorphisms in the mRNA region targeted by the probes may cause mRNA variants to hybridize differently. The difference in the affinity of the mRNA to the probes can result in the identification of false *cis*-eQTL (Alberts et al. 2007). Transcriptome analysis using RNA-seq can also produce false positives (Saha and Battle 2018) when reads from a gene are incorrectly mapped to another gene with a similar sequence. If a *cis*-eQTL regulates the expression of the former gene, a false *trans*-eQTL on the exact location will be falsely detected, targeting the latter gene. Besides limitations in the method, false positives can also be caused by the experimental design. For example, eQTL hotspots could arise due to confounding factors such as batch effects (Michaelson et al. 2009) or variation in developmental age within the tested population (van Eijnatten et al. 2023). Given the prevalence of false positives, it is clear that strong claims regarding eQTL results (and other types of statistical association studies) should be avoided before experimental validation (*e.g.,* using the reverse genetics approach). In this regard, the eQTG-Finder score (Chapter 3) and knowledge graphs (Chapter 4) in AraQTL serve as alternative ways of validation by checking whether a potential or known gene expression regulator is present in the eQTL region.

Integrating different types of functional interaction data

In Chapter 4, I described the prioritisation of eQTL causal genes using a knowledge graph based on different types of functional interaction (i.e. protein-protein interaction, transcription factor-target pairs, functional similarity). The prioritisation now is simply based on node degree, *i.e.,* candidate genes are ranked by the number of functional interactions (edges) connecting them to the target gene. However, the knowledge graph is heterogeneous, *i.e.*, edges represent different types of interaction, and data sources are not equally reliable. For example, interactions inferred via homology are less reliable than those with direct experimental evidence, yet these two data sources are currently considered equal in ranking the candidate genes. Additionally, some of the interaction types already come with weights or scores. However, their calculation and interpretation are different, *e.g.,* protein-protein interaction scores (from STRING and AraNet) are based on inferred probabilities, where GO semantic similarity scores are based on the shared information content of the associated GO terms. A form of score integration is needed to get a more confident ranking of the candidates.

Several approaches can be used to integrate different types of evidence in a heterogeneous network. The simplest method is to normalise the interaction scores (*e.g.,* using z-score) and use the average to evaluate the reliability of the candidate regulator-target gene interaction. In the case of gene-disease association in humans, this naïve approach of averaging different functional interaction scores works better than using only one type of interaction in disease gene prioritisation (Valentini et al. 2014). Additionally, differential weighing of the contributions of each data source (i.e., less reliable data contributes less to the final score) can lead to more reliable regulator-target interaction inference. A network where edge weights were calculated using a supervised learning algorithm indeed improved the predictive performance in gene-disease association (Valentini et al. 2014). Supervised learning has been used for evidence integration in other cases as well, for example, in predicting genetic interactions in yeast (Wong et al. 2004; Pandey et al. 2010) and transcription factor-target gene regulation in Arabidopsis (De Clercq et al. 2021).

Despite good prediction performance in the above cases, integrating data by supervised learning for the inference of causal eQTL-target gene interactions is complicated by the limited number of reliably labelled gene-gene interaction instances, *i.e.* known eQTL causal genes with their target (positive instances) and non-target genes (negative instances). The sample size can be increased through fine-mapping of eQTL intervals, but this is costly and laborious. A more feasible option to increase the number of positive instances would be by mining the literature manually or using artificial intelligence (*e.g.,* PlantConnectome (Kevin et al. 2023)) for experimentally validated interactions between candidate genes in the eQTL interval and the target genes. Several studies use literature mining to generate positive instances for machine learning algorithms, resulting in a decent prediction performance (Chapter 3¸ Lin et al. 2019; Lin et al. 2020; Fu et al. 2020). To gather negative instances, a naïve approach would select genes located outside of eQTL regions as potential non-regulators for a target gene. However, regulatory interactions are specific to the tissue, environment, and developmental stage (Chapter 2, Vinuela et al. 2010, Snoek et al. 2012; Snoek et al. 2017; Lowry et al. 2013, Cubillos et al. 2012). The absence of an eQTL does not necessarily mean that the region does not harbour a gene expression regulator; it could just be inactive in that particular experimental condition or simply lack variants affecting its activity. Therefore, negative instances should be inferred from as many experiments as possible under different conditions, using different populations, that report the absence of an eQTL. It is clear that currently, available eQTL studies do not exhaustively cover all possible tissues, environments, developmental stages, etc.

With limited numbers of positive instances and the absence of a reliable set of negative instances between eQTL causal genes and their target, a suitable approach for data integration and predicting unknown interactions could be positive-unlabelled (PU) learning. PU learning can train classifiers using a limited number of positive samples and many unlabelled instances. As obtaining labelled instances is often expensive, it has become a popular approach in bioinformatics. There are several approaches to PU learning (Li et al. 2022). Most commonly, reliable negative instances are identified by negative expansion. This process involves iteratively adding negative samples using a classifier trained, assuming unlabelled samples are putatively negative. Another method, called label propagation, identifies negative samples based on their distance to positive samples. A third approach adapts the base classifiers in an ensemble approach (*e.g.,* a support vector machine), as in Yang et al. (2016). This idea is similar to the work described in Chapter 3, where an ensemble of models is trained using unlabelled data as negative instances, and the class probability estimates (multiplied by a correction factor to adjust for class imbalance) from these models are averaged to obtain the final prediction result. Future research should evaluate the feasibility of PU learning in predicting eQTL causal-target genes.

6.2. Recommendation for future research

Knowledge graphs for crops

Systems genetics studies using eQTL mapping have been performed in agriculturally important crops. Most of these studies identify causal genes by constructing a co-expression network similar to the one in Chapter 2, for example, in potatoes (van Muijen et al. 2016), rape seed (Basnet et al. 2016), maize (Wang et al. 2018), and melon (Galpaz et al. 2018). This approach effectively identifies potential candidate regulatory genes; however, it does not work so well in detecting genes that regulate the activity of other genes beyond the transcriptional level (i.e., at the level of epigenetics, post-transcription, translation, and post-translation) (Harrison and Shanahan 2014). As described before, gene prioritisation using a knowledge graph can provide an alternative based on more heterogeneous sources of information.

In Chapter 4, the knowledge graph was constructed from public data widely available for Arabidopsis. For some major crops, these kinds of data are also available in several databases. For example, functional interaction predictions between genes/proteins are available for soybean in SoyNet (Kim et al. 2017a), for tomato in TomatoNet (Kim et al. 2017b), for wheat in WheatNet (Lee et al. 2017), and for rice in RiceNet (Lee et al. 2015a). STRING stores predicted protein-protein interactions for a broad range of crop species (Szklarczyk et al. 2021). Transcription factor-target gene relations can be retrieved in PlantRegMap for multiple non-model species (Tian et al. 2020). Similarly, gene ontology (GO) and KEGG pathway annotations are available for many crops (Gene Ontology 2021; Kanehisa et al. 2021). Breeding companies might have their own in-house data, but this is often not freely available. Generally, the coverage, however, is not as complete as for Arabidopsis. Some databases address this limitation by transferring information using homology, *e.g.,* in STRING from version 9 onward using hierarchically arranged orthologous groups (Szklarczyk et al. 2019) and phylogenetically-inferred Gene Ontology annotations (Gaudet et al. 2011). Taken together, also for crops, sufficient data should be available to construct a knowledge graph.

A further development would be to generate a pan-plant knowledge graph containing information from and for all members of plant species. Such a database would support easy knowledge transfer between species by linking orthologous genes. Regulatory information for genes in a well-studied species can then be used to infer interactions in less-studied species, *e.g.,* neglected and underutilized crops. Moreover, knowledge graphs can help identify conserved causal genes in different species. This topic will be discussed further in the next section.

Orthology analysis in QTL intervals to identify conserved causal genes

Orthologous genes may be causal for a QTL for a particular trait in different species. For example, Yoshihara et al. (2022) identified seven one-to-one orthologs in maize and Arabidopsis gravitropism-related trait QTLs, four demonstrating a specific role in gravitropism. The same concept could be applied to molecular phenotypes, *i.e.* for finding orthologous genes causing variation in metabolite or gene expression regulators. For example, I analysed eQTLs from dry seeds of Arabidopsis (Serin 2018), rape seed (Basnet et al. 2016), soybean (Bolon et al. 2014), and tomato (Sterken et al. 2021) and found several candidate-target orthologs in the eQTL intervals of two or more species (data not shown). Even though experimental validation is needed, the orthologs of candidate-target genes can be used to support the presence of gene regulatory interactions in different species.

Integration of multi-omics data

The overarching goal of systems genetics is to reveal the flow of information from variation in the DNA sequence to a phenotype. Combining multi-omics QTLs and looking for hotspots where QTLs of different omics levels collocate (Chapter 2) is a promising approach to discover interactions between molecular phenotypes (*e.g.,* transcript and metabolite levels), leading to variation in the phenotype. With the increasing availability of multi-omics data, methods to infer interactions between different molecule types have been developed, ranging from simple correlation to machine learning algorithms (Hawe et al. 2019). Typically, analyses involve calculating pair-wise associations between raw data of different omics from multiple samples. However, the heterogeneity (*e.g.,* in the data types, source of noise, and distribution) and the complexity of the interactions make inference of interactions based on multi-omics data hard.

The genetic correlation calculated on LOD scores between different molecular phenotypes could offer a more standardized approach for multi-omics data integration. In Chapter 4, the LOD score profile of the *FLC* transcript, of which the protein product is known to repress flowering, is highly correlated to that of a flowering time phenotype (FTLD). The same is observed for the *AOP* gene transcript and the insect resistance phenotype. These findings indicate that similarity in genetic architectures reflected in correlated LOD scores can help infer functional interactions between molecular phenotypes of different omics levels. These interactions may point to regulatory interactions or the presence of a common regulator. Spurious correlations may occur, so a specific filtering method should be applied, for example, by permuting LOD scores multiple times and determining the correlation threshold based on the 95th percentile of the null distribution. Given the availability of large amounts of QTL profiles from different omics levels (*e.g.* in AraQTL), a large-scale inference of multi-omics interaction can be performed to screen for potential interactions. Highly correlated (intermediate) phenotypes can then be validated using a knowledge graph (see Chapter 4) to explore the presence of functional interactions (*e.g.,* GO annotations for gene-trait associations or KEGG for gene-metabolite associations).

6.3. Potential application of a systems genetics approach in plant breeding

The identification of targets for genome editing

Most recent progress in plant breeding was achieved using molecular marker technology, *i.e.*, marker-assisted breeding. The marker is selected based on statistical association (*i.e.,* linkage) with the trait without any prior functional information or molecular mechanism regarding its effect (*e.g.,* on the causal gene); this can be considered a “black box” method. While the genetic marker is sufficient for selection, the information on a causal gene can be useful for genome editing, a powerful technique for modifying the DNA sequence at the nucleotide level (Aglawe et al. 2018). Simple traits regulated by a single locus or a few loci with large effect sizes (*e.g.,* biotic stress resistance (Louthan and Kay 2011)) are particularly suitable for genome editing. Despite strict regulation in some countries, genome editing was already shown to speed up plant breeding programs by targeting the function of a major effect gene, for example, in the domestication of a wild species of groundcherry (*Physalis pruinosa*) (Lemmon et al. 2018) or in the production of waxy corn by a deletion of the *Waxy* gene (Gao et al. 2020).

Target genes for genome editing can be identified using a systems genetics approach by correlating trait and gene expression QTLs, as in Chapter 4. Combined with a knowledge graph, we can unravel the underlying gene regulatory networks, the biochemical pathways, and the essential genes that regulate the downstream physiological response. The key components (*e.g.,* the hub genes) of the regulatory network are the potential targets for genome editing to develop novel plant varieties, as proposed by Tripathi and Wilkins (2021).

The identification of intermediate phenotypes to improve genomic prediction

Many important traits in plant breeding (*e.g.,* yield) are controlled by a large number of genes, and, therefore, genome editing targeting a single or few causal genes is ineffective for trait improvement. Genomic selection is a widely applied technique to improve such complex traits, where genome-wide SNP markers are used in the genomic prediction of traits and breeding values (Meuwissen et al. 2001; de Koning 2016). It is also a black box technique, assuming that at least some markers are linked to the causal genes. Many studies show that including prior biological knowledge (*e.g.,* causal variants (Meuwissen et al. 2022; MacLeod et al. 2016), QTL (Teng et al. 2020), or Gene Ontology annotation (Farooq et al. 2020)) in genomic prediction models improves their prediction accuracy. The main challenge, however, is identifying which biological data are relevant to the target trait, as including non-relevant information will likely not improve and might even worsen performance (Farooq et al. 2020; Teng et al. 2020).

Intermediate phenotypes identified using a systems genetic approach (*e.g.,* transcripts or metabolites through eQTL or mQTL mapping, respectively) are potentially beneficial to be included as prior knowledge in genomic prediction. In particular, transcripts or metabolites with a QTL colocating with the target trait QTL (*e.g.,* on a hotspot) may be intermediates in the molecular pathways connecting variation in genome sequence to changes in the trait. For example, in Chapter 2, 96 eQTLs and 20 mQTLs were mapped to the same region on chromosome 5 as a number of seed germination QTLs. The corresponding genes were enriched for GO terms related to “Seed Development”, “Lipid Storage”, and “Seed Germination”.

Previously, gene expression information has been evaluated in genomic prediction to improve prediction accuracy. Transcriptomics data were used as additional variables, including whole-genome transcript levels in addition to SNPs (Guo et al. 2016; Hu et al. 2019a; Li et al. 2019; Morgante et al. 2020; de Las Heras-Saldana et al. 2020; Perez et al. 2022; Wade et al. 2022). However, measuring whole-genome transcript levels (*e.g.,* using RNA-seq) for hundreds to thousands of plants is not cost-efficient in practice, given that only a few transcripts are likely relevant for trait regulation. The necessary transcripts can be identified using the systems genetics approach mentioned above (eQTL and mQTL mapping) and measured using a more targeted method, such as RT-qPCR, which is less expensive for a few transcripts, thus allowing for measurements of many replicates to get better estimates of transcript levels (Nonis et al. 2014). The benefit of this approach in the genomic prediction model should be evaluated in future studies.

Challenges

The implementation of systems genetics in plant breeding is expensive and complicated. Transcriptomic analyses of a whole plant population are already far from routine in breeding programs, and developing a specific population to improve the power of detection (*e.g.*, RILs and MAGIC populations) is even less common. Moreover, the expression of genes, proteins and metabolites is very specific to the environment, tissue, and developmental stage. Consequently, measurements should be taken under conditions relevant to the trait. The identification of causal genes or the increase in genomic prediction accuracy of a trait may not be worth such investment of money, time and labour unless the trait is extremely important (*e.g.,* a new disease resistance trait during an outbreak). Nevertheless, as sequencing technologies are becoming cheaper over time in combination with high-throughput phenotyping systems, routine adoption of systems genetics in plant breeding programmes may become possible in the future.

6.4. Closing remarks

Systems genetics offers a holistic approach to untangling the genetic complexity of plant traits. By exploiting advances in multi-omics technology, systems genetics addresses the interconnecting molecules from different omics levels to understand the flow of information from variations in the DNA to organism-level phenotypes. The underlying processes are complex, and capturing the full detail of trait regulation is fundamentally difficult (and unnecessary in most cases). Nevertheless, these processes can be modelled or otherwise simplified (*e.g.,* using a gene regulatory network or gene prioritization method), allowing us to learn general principles of trait regulation. Implementing systems genetics still faces many challenges, and I cannot see it being applied in plant breeding programs in the near future, at least on a large scale. Nevertheless, systems genetics provides a new perspective for fundamental plant science research by adding a new dimension (*i.e.,* molecular phenotype) to link genotype to phenotype.

References

1001 Genomes Consortium, 2016 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* 166 (2):481-491.

Acquaah, G., 2012 *Principles of Plant Genetics and Breeding*: John Wiley & Sons, Ltd.

Aerts, S., D. Lambrechts, S. Maity, P. Van Loo, B. Coessens *et al.*, 2006 Gene prioritization through genomic data fusion. *Nat Biotechnol* 24 (5):537-544.

Aglawe, S.B., K.M. Barbadikar, S.K. Mangrauthia, and M.S. Madhav, 2018 New breeding technique "genome editing" for crop improvement: applications, potentials and challenges. *3 Biotech* 8 (8):336.

Albert, F.W., J.S. Bloom, J. Siegel, L. Day, and L. Kruglyak, 2018 Genetics of trans-regulatory variation in gene expression. *Elife* 7.

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

Alberts, R., P. Terpstra, Y. Li, R. Breitling, J.P. Nap *et al.*, 2007 Sequence polymorphisms cause many false cis eQTLs. *PLoS One* 2 (7):e622.

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.

Alonso-Blanco, C., L. Bentsink, C.J. Hanhart, H. Blankestijn-de Vries, and M. Koornneef, 2003 Analysis of natural allelic variation at seed dormancy loci of Arabidopsis thaliana. Genetics 164 (2):711-729.

Alonso-Blanco, C., H. Blankestijn-de Vries, C.J. Hanhart, and M. Koornneef, 1999 Natural allelic variation at seed size loci in relation to other life history traits of Arabidopsis thaliana. Proc Natl Acad Sci U S A 96 (8):4710-4717.

Alonso-Blanco, C., C. Gomez-Mena, F. Llorente, M. Koornneef, J. Salinas et al., 2005 Genetic and molecular analyses of natural variation indicate CBF2 as a candidate gene for underlying a freezing tolerance quantitative trait locus in Arabidopsis. Plant Physiol 139 (3):1304-1312.

Alonso-Blanco, C., A.J. Peeters, M. Koornneef, C. Lister, C. Dean et al., 1998 Development of an AFLP based linkage map of Ler, Col and Cvi Arabidopsis thaliana ecotypes and construction of a Ler/Cvi recombinant inbred line population. Plant J 14 (2):259-271.Altenhoff, A.M., N.M. Glover, C.M. Train, K. Kaleb, A. Warwick Vesztrocy *et al.*, 2018 The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Res* 46 (D1):D477-D485.

Anwer, M.U., E. Boikoglou, E. Herrero, M. Hallstein, A.M. Davis *et al.*, 2014 Natural variation reveals that intracellular distribution of ELF3 protein is associated with function in the circadian clock. *Elife* 3.

Arabidopsis Genome, I., 2000 Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408 (6814):796-815.

Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler *et al.*, 2000 Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25 (1):25-29.

Baxter, I., M. Ouzzani, S. Orcun, B. Kennedy, S.S. Jandhyala *et al.*, 2007 Purdue ionomics information management system. An integrated functional genomics platform. *Plant Physiol* 143 (2):600-611.

Assenov, Y., F. Ramirez, S.E. Schelhorn, T. Lengauer, and M. Albrecht, 2008 Computing topological parameters of biological networks. *Bioinformatics* 24 (2):282-284.

Balcı, H., metincansiper, M. Franx, D. Fong, M. Cheung *et al.*, 2019 cytoscape/cytoscape.js-cose-bilkent: 4.1.0 (v4.1.0), Zenodo.

Bargsten, J.W., J.P. Nap, G.F. Sanchez-Perez, and A.D. van Dijk, 2014 Prioritization of candidate genes in QTL regions based on associations between traits and biological processes. *BMC Plant Biol* 14:330.

Basnet, R.K., D.P. Del Carpio, D. Xiao, J. Bucher, M. Jin *et al.*, 2016 A Systems Genetics Approach Identifies Gene Regulatory Networks Associated with Fatty Acid Composition in Brassica rapa Seed. *Plant Physiol* 170 (1):568-585.

Bassel, G.W., H. Lan, E. Glaab, D.J. Gibbs, T. Gerjets *et al.*, 2011 Genome-wide network model capturing seed germination reveals coordinated regulation of plant cellular phase transitions. *Proc Natl Acad Sci U S A* 108 (23):9709-9714.

Bazakos, C., M. Hanemian, C. Trontin, J.M. Jimenez-Gomez, and O. Loudet, 2017 New Strategies and Tools in Quantitative Genetics: How to Go from the Phenotype to the Genotype. *Annu Rev Plant Biol* 68:435-455.

Bebber, D.P., T. Holmes, and S.J. Gurr, 2014 The global spread of crop pests and pathogens. *Global Ecology and Biogeography* 23 (12):1398-1407.

Bebber, D.P., M.A.T. Ramotowski, and S.J. Gurr, 2013 Crop pests and pathogens move polewards in a warming world. *Nature Climate Change* 3 (11):985-988.

Bewley, J.D., K.J. Bradford, H.W.M. Hilhorst, and H. Nonogaki, 2013a Dormancy and the Control of Germination, pp. 247-297 in *Seeds: Physiology of Development, Germination and Dormancy, 3rd Edition*, edited by J.D. Bewley, K.J. Bradford, H.W.M. Hilhorst and H. Nonogaki. Springer New York, New York, NY.

Bewley, J.D., K.J. Bradford, H.W.M. Hilhorst, and H. Nonogaki, 2013b Germination, pp. 133-181 in *Seeds: Physiology of Development, Germination and Dormancy, 3rd Edition*, edited by J.D. Bewley, K.J. Bradford, H.W.M. Hilhorst and H. Nonogaki. Springer New York, New York, NY.

Bewley, J.D., K.J. Bradford, H.W.M. Hilhorst, and H. Nonogaki, 2013c Synthesis of Storage Reserves, pp. 85-131 in *Seeds: Physiology of Development, Germination and Dormancy, 3rd Edition*, edited by J.D. Bewley, K.J. Bradford, H.W.M. Hilhorst and H. Nonogaki. Springer New York, New York, NY.

Boettiger, C., 2015 An introduction to Docker for reproducible research. *ACM SIGOPS Operating Systems Review* 49 (1):71-79.

Bolon, Y.T., D.L. Hyten, J.H. Orf, C.P. Vance, and G.J. Muehlbauer, 2014 eQTL Networks Reveal Complex Genetic Architecture in the Immature Soybean Seed. *The Plant Genome* 7 (1).

Bornigen, D., L.C. Tranchevent, F. Bonachela-Capdevila, K. Devriendt, B. De Moor *et al.*, 2012 An unbiased evaluation of gene prioritization tools. *Bioinformatics* 28 (23):3081-3088.

Bentsink, L., C. Alonso-Blanco, D. Vreugdenhil, K. Tesnier, S.P. Groot *et al.*, 2000 Genetic analysis of seed-soluble oligosaccharides in relation to seed storability of Arabidopsis. *Plant Physiol* 124 (4):1595-1604.

Bentsink, L., K. Yuan, M. Koornneef, and D. Vreugdenhil, 2003 The genetics of phytate and phosphate accumulation in seeds and leaves of Arabidopsis thaliana, using natural variation. *Theor Appl Genet* 106 (7):1234-1243.

Borevitz, J.O., J.N. Maloof, J. Lutes, T. Dabi, J.L. Redfern *et al.*, 2002 Quantitative trait loci controlling light and hormone response in two accessions of Arabidopsis thaliana. *Genetics* 160 (2):683-696.

Botto, J.F., C. Alonso-Blanco, I. Garzaron, R.A. Sanchez, and J.J. Casal, 2003 The Cape Verde Islands allele of cryptochrome 2 enhances cotyledon unfolding in the absence of blue light in Arabidopsis. *Plant Physiol* 133 (4):1547-1556.

Botto, J.F., and M.P. Coluccio, 2007 Seasonal and plant-density dependency for quantitative trait loci affecting flowering time in multiple populations of Arabidopsis thaliana. *Plant Cell Environ* 30 (11):1465-1479.

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.

Broman, K.W., 2015 R/qtlcharts: interactive graphics for quantitative trait locus mapping. *Genetics* 199 (2):359-361.

Broman, K.W., and S. Sen, 2009 *A Guide to QTL Mapping with R/qtl*.

Buijs, G., A. Vogelzang, H. Nijveen, and L. Bentsink, 2019 Dormancy cycling: Translation related transcripts are the main difference between dormant and non-dormant seeds in the field. *Plant J.*

Buijs, G., L.A.J. Willems, J. Kodde, S.P.C. Groot, and L. Bentsink, 2020 Evaluating the EPPO method for seed longevity analyses in Arabidopsis. *Plant Sci* 301:110644

Cadman, C.S., P.E. Toorop, H.W. Hilhorst, and W.E. Finch-Savage, 2006 Gene expression profiles of Arabidopsis Cvi seeds during dormancy cycling indicate a common underlying dormancy control mechanism. *Plant J* 46 (5):805-822.

Castro-Mondragon, J.A., R. Riudavets-Puig, I. Rauluseviciute, R.B. Lemma, L. Turchi *et al.*, 2022 JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 50 (D1):D165-D173.

Chen, D., W. Yan, L.Y. Fu, and K. Kaufmann, 2018 Architecture of gene regulatory networks controlling flower development in Arabidopsis thaliana. *Nat Commun* 9 (1):4534.

Chiang, G.C., D. Barua, E.M. Kramer, R.M. Amasino, and K. Donohue, 2009 Major flowering time gene, flowering locus C, regulates seed germination in Arabidopsis thaliana. *Proc Natl Acad Sci U S A* 106 (28):11661-11666.

Chinnusamy, V., M. Ohta, S. Kanrar, B.H. Lee, X. Hong *et al.*, 2003 ICE1: a regulator of cold-induced transcriptome and freezing tolerance in Arabidopsis. *Genes Dev* 17 (8):1043-1054.

Choi, K., J. Kim, H.J. Hwang, S. Kim, C. Park *et al.*, 2011 The FRIGIDA complex activates transcription of FLC, a strong flowering repressor in Arabidopsis, by recruiting chromatin modification factors. *Plant Cell* 23 (1):289-303.

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

Collard, B.C., and D.J. Mackill, 2008 Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philos Trans R Soc Lond B Biol Sci* 363 (1491):557-572.

Collard, B.C.Y., M.Z.Z. Jahufer, J.B. Brouwer, and E.C.K. Pang, 2005 An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica* 142 (1-2):169-196.

Comai, L., and J.J. Harada, 1990 Transcriptional activities in dry seed nuclei indicate the timing of the transition from embryogeny to germination. *Proc Natl Acad Sci U S A* 87 (7):2671-2674.

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.

Darrah, C., B.L. Taylor, K.D. Edwards, P.E. Brown, A. Hall *et al.*, 2006 Analysis of phase of LUCIFERASE expression reveals novel circadian quantitative trait loci in Arabidopsis. *Plant Physiol* 140 (4):1464-1474.

Decroocq, V., O. Sicard, J.M. Alamillo, M. Lansac, J.P. Eyquard *et al.*, 2006 Multiple resistance traits control Plum pox virus infection in Arabidopsis thaliana. *Mol Plant Microbe Interact* 19 (5):541-549.

De Clercq, I., J. Van de Velde, X. Luo, L. Liu, V. Storme *et al.*, 2021 Integrative inference of transcriptional networks in Arabidopsis yields novel ROS signalling regulators. *Nat Plants* 7 (4):500-513.

De Giorgi, J., U. Piskurewicz, S. Loubery, A. Utz-Pugin, C. Bailly *et al.*, 2015 An Endosperm-Associated Cuticle Is Required for Arabidopsis Seed Viability, Dormancy and Early Control of Germination. *PLoS Genet* 11 (12):e1005708.

de Koning, D.J., 2016 Meuwissen et al. on Genomic Selection. *Genetics* 203 (1):5-7.

de Las Heras-Saldana, S., B.I. Lopez, N. Moghaddar, W. Park, J.E. Park *et al.*, 2020 Use of gene expression and whole-genome sequence information to improve the accuracy of genomic prediction for carcass traits in Hanwoo cattle. *Genet Sel Evol* 52 (1):54.

Dean, G.H., H. Zheng, J. Tewari, J. Huang, D.S. Young *et al.*, 2007 The Arabidopsis MUM2 gene encodes a beta-galactosidase required for the production of seed coat mucilage with correct hydration properties. *Plant Cell* 19 (12):4007-4021.

Dekkers, B.J., S. Pearce, R.P. van Bolderen-Veldkamp, A. Marshall, P. Widera *et al.*, 2013 Transcriptional dynamics of two seed compartments with opposing roles in Arabidopsis seed germination. *Plant Physiol* 163 (1):205-215.

Denay, G., A. Creff, S. Moussu, P. Wagnon, J. Thevenin *et al.*, 2014 Endosperm breakdown in Arabidopsis requires heterodimers of the basic helix-loop-helix proteins ZHOUPI and INDUCER OF CBP EXPRESSION 1. *Development* 141 (6):1222-1227.

Depuydt, T., and K. Vandepoele, 2021 Multi-omics network-based functional annotation of unknown Arabidopsis genes. *Plant J* 108 (4):1193-1212.

Edwards, K.D., J.R. Lynn, P. Gyula, F. Nagy, and A.J. Millar, 2005 Natural allelic variation in the temperature-compensation mechanisms of the Arabidopsis thaliana circadian clock. *Genetics* 170 (1):387-400.

Emms, D.M., and S. Kelly, 2019 OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* 20 (1):238.

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.

Esch, E., J.M. Szymaniak, H. Yates, W.P. Pawlowski, and E.S. Buckler, 2007 Using crossover breakpoints in recombinant inbred lines to identify quantitative trait loci controlling the global recombination frequency. *Genetics* 177 (3):1851-1858.

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.

Evans, K.S., M.H. van Wijk, P.T. McGrath, E.C. Andersen, and M.G. Sterken, 2021 From QTL to gene: C. elegans facilitates discoveries of the genetic mechanisms underlying natural variation. *Trends Genet* 37 (10):933-947.

Faith, J.J., B. Hayete, J.T. Thaden, I. Mogno, J. Wierzbowski *et al.*, 2007 Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol* 5 (1):e8.

FAO, 2011 Save and grow: A policymaker’s guide to the sustainable intensification of smallholder crop production., Rome.

FAO, 2016 The State of Food and Agriculture 2016. Climate change, agriculture and food security, Rome.

Farooq, M., A.D.J. van Dijk, H. Nijveen, M.G.M. Aarts, W. Kruijer *et al.*, 2020 Prior Biological Knowledge Improves Genomic Prediction of Growth-Related Traits in Arabidopsis thaliana. *Front Genet* 11:609117.

Finch-Savage, W.E., C.S. Cadman, P.E. Toorop, J.R. Lynn, and H.W. Hilhorst, 2007 Seed dormancy release in Arabidopsis Cvi by dry after-ripening, low temperature, nitrate and light shows common quantitative patterns of gene expression directed by environmentally specific sensing. *Plant J* 51 (1):60-78.

Finnegan, E.J., M. Robertson, and C.A. Helliwell, 2020 Resetting FLOWERING LOCUS C Expression After Vernalization Is Just Activation in the Early Embryo by a Different Name. *Front Plant Sci* 11:620155.

Francesconi, M., and B. Lehner, 2014 The effects of genetic variation on gene expression dynamics during development. *Nature* 505 (7482):208-211.

Francisco, M., D.J. Kliebenstein, V.M. Rodriguez, P. Soengas, R. Abilleira *et al.*, 2021 Fine mapping identifies NAD-ME1 as a candidate underlying a major locus controlling temporal variation in primary and specialized metabolism in Arabidopsis. *Plant J* 106 (2):454-467.

Frans, M., M. Cheung, H. Balci, D. Fong, A. Li *et al.*, 2021 cytoscape/cytoscape.js-cola: (v2.5.1), Zenodo.

Franz, M., C.T. Lopes, G. Huck, Y. Dong, O. Sumer *et al.*, 2016 Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics* 32 (2):309-311.

Froschel, C., T. Iven, E. Walper, V. Bachmann, C. Weiste *et al.*, 2019 A Gain-of-Function Screen Reveals Redundant ERF Transcription Factors Providing Opportunities for Resistance Breeding Toward the Vascular Fungal Pathogen Verticillium longisporum. *Mol Plant Microbe Interact* 32 (9):1095-1109.

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.

Fu, Y., J. Xu, Z. Tang, L. Wang, D. Yin *et al.*, 2020 A gene prioritization method based on a swine multi-omics knowledgebase and a deep learning model. *Commun Biol* 3 (1):502.

Fulcher, N., A. Teubenbacher, E. Kerdaffrec, A. Farlow, M. Nordborg *et al.*, 2015 Genetic architecture of natural variation of telomere length in Arabidopsis thaliana. *Genetics* 199 (2):625-635.

Gao, H., M.J. Gadlage, H.R. Lafitte, B. Lenderts, M. Yang *et al.*, 2020 Superior field performance of waxy corn engineered using CRISPR-Cas9. *Nat Biotechnol* 38 (5):579-581.

Gaudet, P., M.S. Livstone, S.E. Lewis, and P.D. Thomas, 2011 Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Brief Bioinform* 12 (5):449-462.

Gene Ontology, C., 2021 The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res* 49 (D1):D325-D334.

Godfray, H.C., J.R. Beddington, I.R. Crute, L. Haddad, D. Lawrence *et al.*, 2010 Food security: the challenge of feeding 9 billion people. *Science* 327 (5967):812-818.

Gonda, I., H. Ashrafi, D.A. Lyon, S.R. Strickler, A.M. Hulse-Kemp *et al.*, 2019 Sequencing-Based Bin Map Construction of a Tomato Mapping Population, Facilitating High-Resolution Quantitative Trait Loci Detection. *Plant Genome* 12 (1).

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

Guo, Z., M.M. Magwire, C.J. Basten, Z. Xu, and D. Wang, 2016 Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. *Theor Appl Genet* 129 (12):2413-2427.

Hao, L., X. Ge, H. Wan, S. Hu, M.J. Lercher *et al.*, 2010 Human functional genetic studies are biased against the medically most relevant primate-specific genes. *BMC Evol Biol* 10:316.

Harada, H., T. Kuromori, T. Hirayama, K. Shinozaki, and R.A. Leigh, 2004 Quantitative trait loci analysis of nitrate storage in Arabidopsis leading to an investigation of the contribution of the anion channel gene, AtCLC-c, to variation in nitrate levels. *J Exp Bot* 55 (405):2005-2014.

Harrison, A., and H. Shanahan, 2014 An Overview of Gene Regulation, pp. 21-69 in *Approaches in Integrative Bioinformatics*.

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.

Hartanto, M., A.A. Sami, D. de Ridder, and H. Nijveen, 2022 Prioritizing Candidate eQTL Causal Genes in Arabidopsis using Random Forests. *G3 (Bethesda)*.

Hassani-Pak, K., M. Castellote, M. Esch, M. Hindle, A. Lysenko *et al.*, 2016 Developing integrated crop knowledge networks to advance candidate gene discovery. *Appl Transl Genom* 11:18-26.

Hassani-Pak, K., A. Singh, M. Brandizi, J. Hearnshaw, J.D. Parsons *et al.*, 2021 KnetMiner: a comprehensive approach for supporting evidence-based gene discovery and complex trait analysis across species. *Plant Biotechnol J* 19 (8):1670-1678.

Haury, A.C., F. Mordelet, P. Vera-Licona, and J.P. Vert, 2012 TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. *BMC Syst Biol* 6:145.

Hawe, J.S., F.J. Theis, and M. Heinig, 2019 Inferring Interaction Networks From Multi-Omics Data. *Front Genet* 10:535.

Haynes, W.A., A. Tomczak, and P. Khatri, 2018 Gene annotation bias impedes biomedical research. *Sci Rep* 8 (1):1362.

Hepworth, S.R., F. Valverde, D. Ravenscroft, A. Mouradov, and G. Coupland, 2002 Antagonistic regulation of flowering-time gene SOC1 by CONSTANS and FLC via separate promoter motifs. *EMBO J* 21 (16):4327-4337.

Hill, W.G., 2010 Understanding and using quantitative genetic variation. *Philos Trans R Soc Lond B Biol Sci* 365 (1537):73-85.

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

Hobbs, D.H., J.E. Flintham, and M.J. Hills, 2004 Genetic control of storage oil synthesis in seeds of Arabidopsis. *Plant Physiol* 136 (2):3341-3349.

Hou, A., K. Liu, N. Catawatcharakul, X. Tang, V. Nguyen *et al.*, 2005 Two naturally occurring deletion mutants of 12S seed storage proteins in Arabidopsis thaliana. *Planta* 222 (3):512-520.

Hu, X., W. Xie, C. Wu, and S. Xu, 2019a A directed learning strategy integrating multiple omic data improves genomic prediction. *Plant Biotechnol J* 17 (10):2011-2020.

Hu, Y., X. Han, M. Yang, M. Zhang, J. Pan *et al.*, 2019b The Transcription Factor INDUCER OF CBF EXPRESSION1 Interacts with ABSCISIC ACID INSENSITIVE5 and DELLA Proteins to Fine-Tune Abscisic Acid Signaling during Seed Germination in Arabidopsis. *Plant Cell* 31 (7):1520-1538.

Huang, D., W. Wu, S.R. Abrams, and A.J. Cutler, 2008 The relationship of drought-related gene expression in Arabidopsis thaliana to hormonal and environmental factors. *J Exp Bot* 59 (11):2991-3007.

Huang, X., Q. Feng, Q. Qian, Q. Zhao, L. Wang *et al.*, 2009 High-throughput genotyping by whole-genome resequencing. *Genome Res* 19 (6):1068-1076.

Huang, X., M.J. Paulo, M. Boer, S. Effgen, P. Keizer *et al.*, 2011 Analysis of natural allelic variation in Arabidopsis using a multiparent recombinant inbred line population. *Proc Natl Acad Sci U S A* 108 (11):4488-4493.

Huynh-Thu, V.A., A. Irrthum, L. Wehenkel, and P. Geurts, 2010 Inferring regulatory networks from expression data using tree-based methods. *PLoS One* 5 (9).

Imprialou, M., A. Kahles, J.G. Steffen, E.J. Osborne, X. Gan *et al.*, 2017 Genomic Rearrangements in Arabidopsis Considered as Quantitative Traits. *Genetics* 205 (4):1425-1441.

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.

Joosen, R.V., D. Arends, Y. Li, L.A. Willems, J.J. Keurentjes *et al.*, 2013 Identifying genotype-by-environment interactions in the metabolism of germinating arabidopsis seeds using generalized genetical genomics. *Plant Physiol* 162 (2):553-566.

Joosen, R.V., D. Arends, L.A. Willems, W. Ligterink, R.C. Jansen *et al.*, 2012 Visualizing the genetic landscape of Arabidopsis seed performance. *Plant Physiol* 158 (2):570-589.

Joosen, R.V., W. Ligterink, H.W. Hilhorst, and J.J. Keurentjes, 2009 Advances in genetical genomics of plants. *Curr Genomics* 10 (8):540-549.

Ju, S., Y.S. Go, H.J. Choi, J.M. Park, and M.C. Suh, 2017 DEWAX Transcription Factor Is Involved in Resistance to Botrytis cinerea in Arabidopsis thaliana and Camelina sativa. *Front Plant Sci* 8:1210.

Juenger, T., J.M. Perez-Perez, S. Bernal, and J.L. Micol, 2005 Quantitative trait loci mapping of floral and leaf morphology traits in Arabidopsis thaliana: evidence for modular genetic architecture. *Evol Dev* 7 (3):259-271.

Julca I, Ferrari C, Flores-Tornero M, Proost S, Lindner A-C, Hackenberg D, Steinbachová L, Michaelidis C, Pereira SG, Misra CS, et al. Comparative transcriptomic analysis reveals conserved transcriptional programs underpinning organogenesis and reproduction in land plants. bioRxiv. 2020.

Kanaoka, M.M., L.J. Pillitteri, H. Fujii, Y. Yoshida, N.L. Bogenschutz *et al.*, 2008 SCREAM/ICE1 and SCREAM2 specify three cell-state transitional steps leading to arabidopsis stomatal differentiation. *Plant Cell* 20 (7):1775-1785.

Kanehisa, M., M. Furumichi, Y. Sato, M. Ishiguro-Watanabe, and M. Tanabe, 2021 KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res* 49 (D1):D545-D551.

Kanehisa, M., M. Furumichi, Y. Sato, M. Kawashima, and M. Ishiguro-Watanabe, 2022 KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res*.

Kawakatsu, T., S.C. Huang, F. Jupe, E. Sasaki, R.J. Schmitz *et al.*, 2016 Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. *Cell* 166 (2):492-505.

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.

Keurentjes, J.J., L. Bentsink, C. Alonso-Blanco, C.J. Hanhart, H. Blankestijn-De Vries *et al.*, 2007 Development of a near-isogenic line population of Arabidopsis thaliana and comparison of mapping power with a recombinant inbred line population. *Genetics* 175 (2):891-905.

Kevin, F., C. Yu Song, F. Herman, D. Emilia Emmanuelle, F. Melissa *et al.*, 2023 PlantConnectome: knowledge networks encompassing > 100,000 plant article abstracts. *bioRxiv*:2023.2007.2011.548541.

Khush, G.S., 2001 Green revolution: the way forward. *Nat Rev Genet* 2 (10):815-822.

Kim, E., S. Hwang, and I. Lee, 2017a SoyNet: a database of co-functional networks for soybean Glycine max. *Nucleic Acids Res* 45 (D1):D1082-D1089.

Kim, H., B.S. Kim, J.E. Shim, S. Hwang, S. Yang *et al.*, 2017b TomatoNet: A Genome-wide Co-functional Network for Unveiling Complex Traits of Tomato, a Model Crop for Fleshy Fruits. *Mol Plant* 10 (4):652-655.

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.

Klopfenstein, D.V., L. Zhang, B.S. Pedersen, F. Ramirez, A. Warwick Vesztrocy *et al.*, 2018 GOATOOLS: A Python library for Gene Ontology analyses. *Sci Rep* 8 (1):10872.

Kliebenstein, D., D. Pedersen, B. Barker, and T. Mitchell-Olds, 2002 Comparative analysis of quantitative trait loci controlling glucosinolates, myrosinase and insect resistance in Arabidopsis thaliana. *Genetics* 161 (1):325-332.

Kobayashi, Y., Y. Furuta, T. Ohno, T. Hara, and H. Koyama, 2005 Quantitative trait loci controlling aluminium tolerance in two accessions of Arabidopsis thaliana (Landsberg erecta and Cape Verde Islands). *Plant, Cell and Environment* 28 (12):1516-1524.

Kooke, R., E. Wijnker, and J.J. Keurentjes, 2012 Backcross populations and near isogenic lines. *Methods Mol Biol* 871:3-16.

Koornneef, M., and D. Meinke, 2010 The development of Arabidopsis as a model plant. *Plant J* 61 (6):909-921.

Korte, A., and A. Farlow, 2013 The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 9:29.

Kover, P.X., W. Valdar, J. Trakalo, N. Scarcelli, I.M. Ehrenreich *et al.*, 2009 A Multiparent Advanced Generation Inter-Cross to fine-map quantitative traits in Arabidopsis thaliana. *PLoS Genet* 5 (7):e1000551.

Kulkarni, S.R., and K. Vandepoele, 2019 Inference of plant gene regulatory networks using data-driven methods: A practical overview. *Biochim Biophys Acta Gene Regul Mech*:194447.

Kulkarni, S.R., D. Vaneechoutte, J. Van de Velde, and K. Vandepoele, 2018 TF2Network: predicting transcription factor regulators and gene regulatory networks in Arabidopsis using publicly available binding site information. *Nucleic Acids Res* 46 (6):e31.

Lamesch, P., T.Z. Berardini, D. Li, D. Swarbreck, C. Wilks *et al.*, 2012 The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res* 40 (Database issue):D1202-1210.

Lavarenne, J., S. Guyomarc'h, C. Sallaud, P. Gantet, and M. Lucas, 2018 The Spring of Systems Biology-Driven Breeding. *Trends Plant Sci* 23 (8):706-720.

Lee, B.H., D.A. Henderson, and J.K. Zhu, 2005 The Arabidopsis cold-responsive transcriptome and its regulation by ICE1. *Plant Cell* 17 (11):3155-3175.

Lee, J., and I. Lee, 2010 Regulation and function of SOC1, a flowering pathway integrator. *J Exp Bot* 61 (9):2247-2254.

Lee, T., S. Hwang, C.Y. Kim, H. Shim, H. Kim *et al.*, 2017 WheatNet: a Genome-Scale Functional Network for Hexaploid Bread Wheat, Triticum aestivum. *Mol Plant* 10 (8):1133-1136.

Lee, T., T. Oh, S. Yang, J. Shin, S. Hwang *et al.*, 2015a RiceNet v2: an improved network prioritization server for rice genes. *Nucleic Acids Res* 43 (W1):W122-127.

Lee, T., S. Yang, E. Kim, Y. Ko, S. Hwang *et al.*, 2015b AraNet v2: an improved database of co-functional gene networks for the study of Arabidopsis thaliana and 27 other nonmodel plant species. *Nucleic Acids Res* 43 (Database issue):D996-1002.

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

Lemmon, Z.H., N.T. Reem, J. Dalrymple, S. Soyk, K.E. Swartwood *et al.*, 2018 Rapid improvement of domestication traits in an orphan crop by genome editing. *Nat Plants* 4 (10):766-770.

Leubner-Metzger, G., 2005 beta-1,3-Glucanase gene expression in low-hydrated seeds as a mechanism for dormancy release during tobacco after-ripening. *Plant J* 41 (1):133-145.

Li, F., S. Dong, A. Leier, M. Han, X. Guo *et al.*, 2022 Positive-unlabeled learning in bioinformatics and computational biology: a brief review. *Brief Bioinform* 23 (1).

Li, Y., M.A. Swertz, G. Vera, J. Fu, R. Breitling *et al.*, 2009 designGG: an R-package and web tool for the optimal design of genetical genomics experiments. *BMC Bioinformatics* 10 (1):188.

Li, Z., N. Gao, J.W.R. Martini, and H. Simianer, 2019 Integrating Gene Expression Data Into Genomic Prediction. *Front Genet* 10:126.

Liang, C.H., and C.C. Yang, 2015 Identification of ICE1 as a negative regulator of ABA-dependent pathways in seeds and seedlings of Arabidopsis. *Plant Mol Biol* 88 (4-5):459-470.

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.

Liu, X., Y.I. Li, and J.K. Pritchard, 2019 Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell* 177 (4):1022-1034 e1026.

Lopez-Bigas, N., and C.A. Ouzounis, 2004 Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res* 32 (10):3108-3114.

Loudet, O., S. Chaillou, C. Camilleri, D. Bouchez, and F. Daniel-Vedele, 2002 Bay-0 × Shahdara recombinant inbred line population: a powerful tool for the genetic dissection of complex traits in Arabidopsis. *Theoretical and Applied Genetics* 104 (6):1173-1184.

Louthan, A.M., and K.M. Kay, 2011 Comparing the adaptive landscape across trait types: larger QTL effect size in traits under biotic selection. *BMC Evol Biol* 11:60.

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.

Luquez, V.M., Y. Sasal, M. Medrano, M.I. Martin, M. Mujica *et al.*, 2006 Quantitative trait loci analysis of leaf and plant longevity in Arabidopsis thaliana. *J Exp Bot* 57 (6):1363-1372.

Lustenhouwer, N., J.L. Williams, J.M. Levine, and J. Cahill, 2018 Evolution during population spread affects plant performance in stressful environments. *Journal of Ecology* 107 (1):396-406.

MacGregor, D.R., N. Zhang, M. Iwasaki, M. Chen, A. Dave *et al.*, 2019 ICE1 and ZOU determine the depth of primary seed dormancy in Arabidopsis independently of their role in endosperm development. *Plant J* 98 (2):277-290.

MacLeod, I.M., P.J. Bowman, C.J. Vander Jagt, M. Haile-Mariam, K.E. Kemper *et al.*, 2016 Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genomics* 17:144.

Macquet, A., M.C. Ralet, O. Loudet, J. Kronenberger, G. Mouille *et al.*, 2007 A naturally occurring mutation in an Arabidopsis accession affects a beta-D-galactosidase that increases the hydrophilic potential of rhamnogalacturonan I in seed mucilage. *Plant Cell* 19 (12):3990-4006.

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.

Margolin, A.A., I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky *et al.*, 2006 ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7 Suppl 1:S7.

Meuwissen, T., B. Hayes, I. MacLeod, and M. Goddard, 2022 Identification of Genomic Variants Causing Variation in Quantitative Traits: A Review. *Agriculture* 12 (10).

Meuwissen, T.H., B.J. Hayes, and M.E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157 (4):1819-1829.

Michaels, S.D., and R.M. Amasino, 2001 Loss of FLOWERING LOCUS C activity eliminates the late-flowering phenotype of FRIGIDA and autonomous pathway mutations but not responsiveness to vernalization. *Plant Cell* 13 (4):935-941.

Michaelson, J.J., S. Loguercio, and A. Beyer, 2009 Detection and interpretation of expression quantitative trait loci (eQTL). *Methods* 48 (3):265-276.

Mistry, H., M. Franz, lcparsons, D. Fong, D. Sabsay *et al.*, 2013 CiSE: a circular spring embedder layout algorithm. *IEEE Trans Vis Comput Graph* 19 (6):953-966.

Moose, S.P., and R.H. Mumm, 2008 Molecular plant breeding as the foundation for 21st century crop improvement. *Plant Physiol* 147 (3):969-977.

Moreau, Y., and L.C. Tranchevent, 2012 Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat Rev Genet* 13 (8):523-536.

Morgante, F., W. Huang, P. Sorensen, C. Maltecca, and T.F.C. Mackay, 2020 Leveraging Multiple Layers of Data To Predict Drosophila Complex Traits. *G3 (Bethesda)* 10 (12):4599-4613.

Mostafavi, H., J.P. Spence, S. Naqvi, and J.K. Pritchard, 2022 Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery. *bioRxiv*.

Nakabayashi, K., M. Okamoto, T. Koshiba, Y. Kamiya, and E. Nambara, 2005 Genome-wide profiling of stored mRNA in Arabidopsis thaliana seed germination: epigenetic and genetic regulation of transcription in seed. *Plant J* 41 (5):697-709.

Narsai, R., S.R. Law, C. Carrie, L. Xu, and J. Whelan, 2011 In-depth temporal transcriptome profiling reveals a crucial developmental switch with roles for RNA processing and organelle metabolism that are essential for germination in Arabidopsis. *Plant Physiol* 157 (3):1342-1362.

Nielsen, R., J.S. Paul, A. Albrechtsen, and Y.S. Song, 2011 Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* 12 (6):443-451.

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.

Noman, A., M. Aqeel, and S. He, 2016 CRISPR-Cas9: Tool for Qualitative and Quantitative Plant Genome Editing. *Front Plant Sci* 7:1740.

Nonis, A., B. De Nardi, and A. Nonis, 2014 Choosing between RT-qPCR and RNA-seq: a back-of-the-envelope estimate towards the definition of the break-even-point. *Anal Bioanal Chem* 406 (15):3533-3536.

Nonogaki, H., 2019 Seed germination and dormancy: The classic story, new puzzles, and evolution. *J Integr Plant Biol* 61 (5):541-563.

Nonogaki, H., G.W. Bassel, and J.D. Bewley, 2010 Germination—Still a mystery. *Plant Science* 179 (6):574-581.

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.

Ozgur, A., T. Vu, G. Erkan, and D.R. Radev, 2008 Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics* 24 (13):i277-285.

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).

Pascual, L., E. Albert, C. Sauvage, J. Duangjit, J.P. Bouchet *et al.*, 2016 Dissecting quantitative trait variation in the resequencing era: complementarity of bi-parental, multi-parental and association panels. *Plant Sci* 242:120-130.

Paterson, A.H., J.W. DeVerna, B. Lanini, and S.D. Tanksley, 1990 Fine mapping of quantitative trait loci using selected overlapping recombinant chromosomes, in an interspecies cross of tomato. *Genetics* 124 (3):735-742.

Pavlopoulos, G.A., M. Secrier, C.N. Moschopoulos, T.G. Soldatos, S. Kossida *et al.*, 2011 Using graph theory to analyze biological networks. *BioData Min* 4:10.

Payne, K.A., H.C. Bowen, J.P. Hammond, C.R. Hampton, J.R. Lynn *et al.*, 2004 Natural genetic variation in caesium (Cs) accumulation byArabidopsis thaliana. *New Phytologist* 162 (2):535-548.

Pazhamala, L.T., H. Kudapa, W. Weckwerth, A.H. Millar, and R.K. Varshney, 2021 Systems biology for crop improvement. *Plant Genome* 14 (2):e20098.

Penfield, S., R.C. Meissner, D.A. Shoue, N.C. Carpita, and M.W. Bevan, 2001 MYB61 Is Required for Mucilage Deposition and Extrusion in the Arabidopsis Seed Coat. *The Plant Cell* 13 (12):2777-2791.

Perez, B.C., M. Bink, K.L. Svenson, G.A. Churchill, and M.P.L. Calus, 2022 Adding gene transcripts into genomic prediction improves accuracy and reveals sampling time dependence. *G3 (Bethesda)* 12 (11).

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.

Pineau, C., S. Loubet, C. Lefoulon, C. Chalies, C. Fizames *et al.*, 2012 Natural variation at the FRD3 MATE transporter locus reveals cross-talk between Fe homeostasis and Zn tolerance in Arabidopsis thaliana. *PLoS Genet* 8 (12):e1003120.

Pornsiriwong, W., G.M. Estavillo, K.X. Chan, E.E. Tee, D. Ganguly *et al.*, 2017 A chloroplast retrograde signal, 3'-phosphoadenosine 5'-phosphate, acts as a secondary messenger in abscisic acid signaling in stomatal closure and germination. *Elife* 6.

Rabanal, F.A., V. Nizhynska, T. Mandakova, P.Y. Novikova, M.A. Lysak *et al.*, 2017 Unstable Inheritance of 45S rRNA Genes in Arabidopsis thaliana. *G3 (Bethesda)* 7 (4):1201-1209.

Rajjou, L., K. Gallardo, I. Debeaujon, J. Vandekerckhove, C. Job *et al.*, 2004 The effect of alpha-amanitin on the Arabidopsis seed proteome highlights the distinct roles of stored and neosynthesized mRNAs during germination. *Plant Physiol* 134 (4):1598-1613.

Raschke, A., C. Ibanez, K.K. Ullrich, M.U. Anwer, S. Becker *et al.*, 2015 Natural variants of ELF3 affect thermomorphogenesis by transcriptionally modulating PIF4-dependent auxin response genes. *BMC Plant Biol* 15:197.

Resnik, P., 1999 Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research* 11:95-130.

Rhee, S.Y., P. Zhang, H. Foerster, and C. Tissier, 2006 AraCyc: Overview of an Arabidopsis Metabolism Database and its Applications for Plant Research, pp. 141-154 in *Plant Metabolomics*.

Riley, T.R., M. Slattery, N. Abe, C. Rastogi, D. Liu *et al.*, 2014 SELEX-seq: a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes. *Methods Mol Biol* 1196:255-278.

Ritchie, M.E., B. Phipson, D. Wu, Y. Hu, C.W. Law *et al.*, 2015 limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* 43 (7):e47-e47.

Robinson, I., J. Webber, and E. Eifrem, 2015 *Graph Databases*. Beijing: O'Reilly.

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

Royal Society of London, 2009 Reaping the Benefits: Science and the Sustainable Intensification of Global Agriculture, London.

Saha, A., and A. Battle, 2018 False positives in trans-eQTL and co-expression analyses arising from RNA-sequencing alignment errors. *F1000Res* 7:1860.

Sangster, T.A., N. Salathia, S. Undurraga, R. Milo, K. Schellenberg *et al.*, 2008 HSP90 affects the expression of genetic variation and developmental stability in quantitative traits. *Proc Natl Acad Sci U S A* 105 (8):2963-2968.

Scott, M.F., O. Ladejobi, S. Amer, A.R. Bentley, J. Biernaskie *et al.*, 2020 Multi-parent populations in crops: a toolbox integrating genomics and genetic mapping with breeding. *Heredity (Edinb)* 125 (6):396-416.

Sergeeva, L.I., J.J. Keurentjes, L. Bentsink, J. Vonk, L.H. van der Plas *et al.*, 2006 Vacuolar invertase regulates elongation of Arabidopsis thaliana roots as revealed by QTL and mutant analysis. *Proc Natl Acad Sci U S A* 103 (8):2994-2999.

Sergeeva, L.I., J. Vonk, J.J. Keurentjes, L.H. van der Plas, M. Koornneef *et al.*, 2004 Histochemical analysis reveals organ-specific quantitative trait loci for enzyme activities in Arabidopsis. *Plant Physiol* 134 (1):237-245.

Serin, E.A., 2018 Environmental tuning of the genetic control of seed performance : a systems genetics approach in *Laboratory of Plant Physiology*. Wageningen University, Wageningen.

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.

Serin, E.A.R., L.B. Snoek, H. Nijveen, L.A.J. Willems, J.M. Jimenez-Gomez *et al.*, 2017 Construction of a High-Density Genetic Map from RNA-Seq Data for an Arabidopsis Bay-0 x Shahdara RIL Population. *Front Genet* 8:201.

Seyyedrazzagi, E., and N.J. Navimipour, 2017 Disease genes prioritizing mechanisms: a comprehensive and systematic literature review. *Network Modeling Analysis in Health Informatics and Bioinformatics* 6 (1).

Shannon, P., A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang *et al.*, 2003 Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13 (11):2498-2504.

Shi, J., R. Li, D. Qiu, C. Jiang, Y. Long *et al.*, 2009 Unraveling the complex trait of crop yield with quantitative trait loci mapping in Brassica napus. *Genetics* 182 (3):851-861.

Shi, L., G.H. Dean, H. Zheng, M.J. Meents, T.M. Haslam *et al.*, 2019 ECERIFERUM11/C-TERMINAL DOMAIN PHOSPHATASE-LIKE2 Affects Secretory Trafficking. *Plant Physiol* 181 (3):901-915.

Sicard, O., O. Loudet, J.J. Keurentjes, T. Candresse, O. Le Gall *et al.*, 2008 Identification of quantitative trait loci controlling symptom development during viral infection in Arabidopsis thaliana. *Mol Plant Microbe Interact* 21 (2):198-207.

Signor, S.A., and S.V. Nuzhdin, 2018 The Evolution of Gene Expression in cis and trans. *Trends Genet* 34 (7):532-544.

Silva, A.T., P.A. Ribone, R.L. Chan, W. Ligterink, and H.W. Hilhorst, 2016 A Predictive Coexpression Network Identifies Novel Genes Controlling the Seed-to-Seedling Phase Transition in Arabidopsis thaliana. *Plant Physiol* 170 (4):2218-2231.

Simpson, G.G., and C. Dean, 2002 Arabidopsis, the Rosetta stone of flowering time? *Science* 296 (5566):285-289.

Singh, G., A. Kuzniar, M. Brouwer, C. Martinez-Ortiz, C.W.B. Bachem *et al.*, 2020 Linked Data Platform for Solanaceae Species. *Applied Sciences* 10 (19).

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

Snoek, B.L., M.G. Sterken, R.P.J. Bevers, R.J.M. Volkers, A. Van't Hof *et al.*, 2017 Contribution of trans regulatory eQTL to cryptic genetic variation in C. elegans. *BMC Genomics* 18 (1):500.

Snoek, B.L., M.G. Sterken, M. Hartanto, A.J. van Zuilichem, J.E. Kammenga *et al.*, 2020 WormQTL2: an interactive platform for systems genetics in Caenorhabditis elegans. *Database (Oxford)* 2020.

Snoek, B.L., M.G. Sterken, H. Nijveen, R.J.M. Volkers, J. Riksen *et al.*, 2021 The genetics of gene expression in a Caenorhabditis elegans multiparental recombinant inbred line population. *G3 (Bethesda)* 11 (10).

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.

Sterken, M.G., H. Nijveen, M. van Zanten, J.M. Jiménez-Gómez, N. Geshnizjani *et al.*, 2021 Plasticity of maternal environment dependent expression-QTLs of tomato seeds. *bioRxiv*.

Sterken, M.G., L. van Bemmelen van der Plaat, J.A.G. Riksen, M. Rodriguez, T. Schmid *et al.*, 2017 Ras/MAPK Modifier Loci Revealed by eQTL in Caenorhabditis elegans. *G3 (Bethesda)* 7 (9):3185-3193.

Stoeger, T., M. Gerlach, R.I. Morimoto, and L.A. Nunes Amaral, 2018 Large-scale investigation of the reasons why potentially important genes are ignored. *PLoS Biol* 16 (9):e2006643.

Staal, J., M. Kaliff, S. Bohman, and C. Dixelius, 2006 Transgressive segregation reveals two Arabidopsis TIR-NB-LRR resistance genes effective against Leptosphaeria maculans, causal agent of blackleg disease. *Plant J* 46 (2):218-230.

Sun, Z.M., M.L. Zhou, W. Dan, Y.X. Tang, M. Lin *et al.*, 2016 Overexpression of the Lotus corniculatus Soloist Gene LcAP2/ERF107 Enhances Tolerance to Salt Stress. *Protein Pept Lett* 23 (5):442-449.

Swarup, K., C. Alonso-Blanco, J.R. Lynn, S.D. Michaels, R.M. Amasino *et al.*, 1999 Natural allelic variation identifies new genes in the Arabidopsis circadian system. *Plant J* 20 (1):67-77.

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.

Szklarczyk, D., A.L. Gable, K.C. Nastou, D. Lyon, R. Kirsch *et al.*, 2021 The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res* 49 (D1):D605-D612.

Szklarczyk, D., J.H. Morris, H. Cook, M. Kuhn, S. Wyder *et al.*, 2017 The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res* 45 (D1):D362-D368.

Teng, J., S. Huang, Z. Chen, N. Gao, S. Ye *et al.*, 2020 Optimizing genomic prediction model given causal genes in a dairy cattle population. *J Dairy Sci* 103 (11):10299-10310.

Teng, S., J. Keurentjes, L. Bentsink, M. Koornneef, and S. Smeekens, 2005 Sucrose-specific induction of anthocyanin biosynthesis in Arabidopsis requires the MYB75/PAP1 gene. *Plant Physiol* 139 (4):1840-1852.

Teng, S., S. Rognoni, L. Bentsink, and S. Smeekens, 2008 The Arabidopsis GSQ5/DOG1 Cvi allele is induced by the ABA-mediated sugar signalling pathway, and enhances sugar sensitivity by stimulating ABI4 expression. *Plant J* 55 (3):372-381.

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.

Tripathi, R.K., and O. Wilkins, 2021 Single cell gene regulatory networks in plants: Opportunities for enhancing climate change stress resilience. *Plant Cell Environ* 44 (7):2006-2017.

Tuinstra, M.R., G. Ejeta, and P.B. Goldsbrough, 1997 Heterogeneous inbred family (HIF) analysis: a method for developing near-isogenic lines that differ at quantitative trait loci. *Theoretical and Applied Genetics* 95 (5-6):1005-1011.

Ungerer, M.C., S.S. Halldorsdottir, J.L. Modliszewski, T.F. Mackay, and M.D. Purugganan, 2002 Quantitative trait loci for inflorescence development in Arabidopsis thaliana. *Genetics* 160 (3):1133-1151.

Uygun, S., C. Peng, M.D. Lehti-Shiu, R.L. Last, and S.H. Shiu, 2016 Utility and Limitations of Using Gene Expression Data to Identify Functional Associations. *PLoS Comput Biol* 12 (12):e1005244.

Valba, O.V., S.K. Nechaev, M.G. Sterken, L.B. Snoek, J.E. Kammenga *et al.*, 2015 On predicting regulatory genes by analysis of functional networks in C. elegans. *BioData Min* 8:33.

Vallejo, A.J., M.J. Yanovsky, and J.F. Botto, 2010 Germination variation in Arabidopsis thaliana accessions under moderate osmotic and salt stresses. *Ann Bot* 106 (5):833-842.

Valentini, G., A. Paccanaro, H. Caniza, A.E. Romero, and M. Re, 2014 An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods. *Artif Intell Med* 61 (2):63-78.

van Eijnatten, B., M. Sterken, J. Kammenga, H. Nijveen, and S. B.L., 2023 The effect of developmental variation on expression QTLs in a multi parental C. elegans population. *bioRxiv*.

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.

van Zanten, M., L.B. Snoek, M.C. Proveniers, and A.J. Peeters, 2009 The many functions of ERECTA. *Trends Plant Sci* 14 (4):214-218.

Varshney, R.K., A. Bohra, J. Yu, A. Graner, Q. Zhang *et al.*, 2021 Designing Future Crops: Genomics-Assisted Breeding Comes of Age. *Trends Plant Sci* 26 (6):631-649.

Venkatesan, A., G. Tagny Ngompe, N.E. Hassouni, I. Chentli, V. Guignon *et al.*, 2018 Agronomic Linked Data (AgroLD): A knowledge-based system to enable integrative biology in agronomy. *PLoS One* 13 (11):e0198270.

Vinuela, A., L.B. Snoek, J.A. Riksen, and J.E. Kammenga, 2010 Genome-wide gene expression regulation as a function of genotype and age in C. elegans. *Genome Res* 20 (7):929-937.

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.

Vonapartis, E., D. Mohamed, J. Li, W. Pan, J. Wu *et al.*, 2022 CBF4/DREB1D represses XERICO to attenuate ABA, osmotic and drought stress responses in Arabidopsis. *Plant J* 110 (4):961-977.

Vosman, B., A. Kashaninia, W. Van't Westende, F. Meijer-Dekens, H. van Eekelen *et al.*, 2019 QTL mapping of insect resistance components of Solanum galapagense. *Theor Appl Genet* 132 (2):531-541.

Vreugdenhil, D., M.G.M. Aarts, M. Koornneef, H. Nelissen, and W.H.O. Ernst, 2004 Natural variation and QTL analysis for cationic mineral content in seeds of Arabidopsis thaliana. *Plant, Cell and Environment* 27 (7):828-839.

Vyse, K., S. Schaarschmidt, A. Erban, J. Kopka, and E. Zuther, 2022 Specific CBF transcription factors and cold-responsive genes fine-tune the early triggering response after acquisition of cold priming and memory. *Physiol Plant* 174 (4):e13740.

Wade, A.R., H. Durufle, L. Sanchez, and V. Segura, 2022 eQTLs are key players in the integration of genomic and transcriptomic data for phenotype prediction. *BMC Genomics* 23 (1):476.

Wan, C.Y., and T.A. Wilkins, 1994 A Modified Hot Borate Method Significantly Enhances the Yield of High-Quality RNA from Cotton (Gossypium hirsutum L.). *Analytical Biochemistry* 223 (1):7-12.

Wang, X., L. Gao, C. Jiao, S. Stravoravdis, P.S. Hosmani *et al.*, 2020 Genome of Solanum pimpinellifolium provides insights into structural variants during tomato breeding. *Nat Commun* 11 (1):5817.

Warde-Farley, D., S.L. Donaldson, O. Comes, K. Zuberi, R. Badrawi *et al.*, 2010 The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res* 38 (Web Server issue):W214-220.

Warwick Vesztrocy, A., C. Dessimoz, and H. Redestig, 2018 Prioritising candidate genes causing QTL using hierarchical orthologous groups. *Bioinformatics* 34 (17):i612-i619.

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.

White, P.J., K. Swarup, A.J. Escobar-Gutiérrez, H.C. Bowen, N.J. Willey *et al.*, 2003 *Plant and Soil* 249 (1):177-186.

Wijnker, E., G. Velikkakam James, J. Ding, F. Becker, J.R. Klasen *et al.*, 2013 The genomic landscape of meiotic crossovers and gene conversions in Arabidopsis thaliana. *Elife* 2:e01426.

Wingender, E., X. Chen, R. Hehl, H. Karas, I. Liebich *et al.*, 2000 TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res* 28 (1):316-319.

Wojtyla, L., K. Lechowska, S. Kubala, and M. Garnczarska, 2016 Different Modes of Hydrogen Peroxide Action During Seed Germination. *Front Plant Sci* 7:66.

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.

Xu, H., O. Lantzouni, T. Bruggink, R. Benjamins, F. Lanfermeijer *et al.*, 2020 A Molecular Signal Integration Network Underpinning Arabidopsis Seed Germination. *Curr Biol*.

Yang, P., S.J. Humphrey, D.E. James, Y.H. Yang, and R. Jothi, 2016 Positive-unlabeled ensemble learning for kinase substrate prediction from dynamic phosphoproteomics data. *Bioinformatics* 32 (2):252-259.

Yeats, T.H., and J.K. Rose, 2013 The formation and function of plant cuticles. *Plant Physiol* 163 (1):5-20.

Yilmaz, A., M.K. Mejia-Guerra, K. Kurz, X. Liang, L. Welch *et al.*, 2011 AGRIS: the Arabidopsis Gene Regulatory Information Server, an update. *Nucleic Acids Res* 39 (Database issue):D1118-1122.

Yoshihara, T., N.D. Miller, F.A. Rabanal, H. Myles, I.Y. Kwak *et al.*, 2022 Leveraging orthology within maize and Arabidopsis QTL to identify genes affecting natural variation in gravitropism. *Proc Natl Acad Sci U S A* 119 (40):e2212199119.

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.

Zhang, L., W. Su, R. Tao, W. Zhang, J. Chen *et al.*, 2017 RNA sequencing provides insights into the evolution of lettuce and the regulation of flavonoid biosynthesis. *Nat Commun* 8 (1):2264.

Zhou, Z., C. Zhang, Y. Zhou, Z. Hao, Z. Wang *et al.*, 2016 Genetic dissection of maize plant architecture with an ultra-high density bin map based on recombinant inbred lines. *BMC Genomics* 17:178.

Zolotareva, O., and M. Kleine, 2019 A Survey of Gene Prioritization Tools for Mendelian and Complex Human Diseases. *J Integr Bioinform* 16 (4).

Zych, K., B.L. Snoek, M. Elvin, M. Rodriguez, K.J. Van der Velde *et al.*, 2017 reGenotyper: Detecting mislabeled samples in genetic data. *PLoS One* 12 (2):e0171324.

Summary

One of the main objectives of genetics is to understand how genetic variation shapes traits. Genetic research has revolutionised along with the advances in other scientific fields, and the recent developments in omics technologies gave rise to a new branch of genetics: systems genetics. Systems genetics revolves around molecular phenotypes (e.g., phenotypes at the level of transcripts, metabolites, and proteins), which mediate the flow of information from genotype to phenotype. These so-called "intermediate" phenotypes form complex interaction networks, shaping the molecular machinery underlying organism-level phenotypes (e.g., yield, disease resistance). A common approach to combine genetics and genomics, known as genetical genomics, can be used to associate variants in the genome with variations in transcript levels. This approach finds so-called gene expression QTLs (eQTLs), genomic regions associated with gene expression variation. An eQTL can be interpreted as mapping the location of a gene expression regulator. The resolution of eQTL mapping can be quite low, necessitating the prioritisation of dozens, if not hundreds, of genes to find the most likely regulator. Linking genes to their regulators allows us to infer gene regulatory networks, an important step towards understanding traits at the molecular level.

In this thesis, I describe the development of methods to prioritise candidate trait-related genes in model plant *Arabidopsis thaliana* using a systems genetics approach. I first introduce the field, describe the current state-of-the-art gene prioritisation methods and argue that systems genetics is a promising approach to improve these methods. I then present a case study applying systems genetics to find novel regulators for Arabidopsis seed maturation. By combining eQTL mapping with metabolite QTLs and seed germination phenotype QTLs, I discovered a QTL hotspot on chromosome 5 where QTLs of different omics are colocated. A QTL hotspot is assumed to emerge due to the presence of a master regulatory gene, so identifying it could improve our understanding of seed germination. To identify this potential master regulator, I use a community-based approach, combining the results of five co-expression network inference methods. While this expression-based prioritisation approach is simple and requires a minimal amount of data, it comes with a limitation: gene expression regulation in eukaryotes is complex, and a co-expression network does not capture regulation that occurs beyond the transcriptional level (e.g., at the level of translation or protein activity). To address this, I take two different approaches. First, I use a machine-learning approach that extends the QTG-Finder2 method for prioritising candidate causal genes on phenotype QTLs. By adding features related to gene structure, protein interactions, and gene expression, I train a new model called eQTG-Finder. With this model, I can calculate a predicted probability of causality (the eQTG-Finder score) for all Arabidopsis genes that can be used for eQTL causal gene prioritisation. Second, I use prior biological knowledge to prioritise candidate

regulatory genes by integrating gene-gene interaction evidence from gene annotations, protein-protein interactions, and transcription factor binding sites into a knowledge graph. This graph can be queried to identify potential regulators for the gene of interest, which is visualised in an interaction network. Both the knowledge graph and eQTG-Finder scores are available in our systems genetics platform for Arabidopsis called AraQTL, allowing researchers to interactively explore the predictions.

Next, I address the typical low QTL mapping resolution by using RNA-seq-derived SNPs to computationally fine-map QTLs. I demonstrate the effectiveness of this method on QTL hotspots, which results in smaller QTL intervals and fewer candidate genes. This fine-mapping strategy can be used as a quick and low-cost alternative to traditional fine-mapping to explore potential causal genes underlying QTLs.

Finally, I discuss the challenges of the systems genetics approach for causal gene prioritisation. I conclude this thesis by discussing the future of systems genetics research and its potential applications in plant breeding.

Acknowledgments

الْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِينَ

I would like to express my gratitude and appreciation to all those who have supported me throughout this challenging journey of completing my PhD thesis. This accomplishment would not have been possible without their unwavering encouragement, guidance, and assistance.

First and foremost, I extend my heartfelt thanks to my supervisors, Harm Nijveen and Dick de Ridder, whose guidance and expertise have been invaluable in shaping the direction of my research and also my personal and professional growth. Harm, you were always present from beginning to end of this journey. I learned many things from you about science and life in general. I will miss our weekly meeting, which kept motivating me to finish this thesis. Dick, your thought-provoking feedback helped me maintain the quality of my work. I always appreciated your quick response whenever I needed support from you. It is a privilege to have you both as my supervisors.

I also wish to express my gratitude to the people I collaborated with in writing this thesis: Mark, Basten, Marie, Wilco, Ronny, Leo, and Henk. Working with you was a valuable learning experience for me. Thank you for your significant contribution to this thesis.

I am also grateful to have Wilco as my external supervisor. Our discussion about research and careers in the industry was very insightful to me. Too bad we only met a couple of times during my PhD period.

I would like to acknowledge the members of my thesis committee, Leónie Bentsink, Geo Velikkakam James, Jan Kammenga, and Luisa Trindade, for their willingness to spare their precious time to review this thesis and attend my defence as an opponent.

Thank you, Iqbal and Dirk-Jan, for your willingness to become my paranymphs and assisting me in preparing the defence.

Many thanks to all the people of the Bioinformatics Group (fellow PhD students, postdocs and the academic staff) who provided a conducive academic environment, resources, and support that enabled me to thrive and succeed in my doctoral studies. I will miss all the fun during lunches, retreats, and meetings. Special thanks to Mas Satria, my first friend in the group who gave me a gentle introduction into bioinformatics and life as PhD student. I want to acknowledge the BSc and MSc students who helped me develop the chapters of this thesis, especially Asif, Thijn, Weiqi, Antea, Vincent, and Eros. I also wish to thank Marie-José and Maria for helping me with administration and Susan and Juliane for your help with the training and courses.

Kepada teman-teman Indonesia di Wageningen, mulai dari saya MSc sampai sekarang, terima kasih atas pertemanannya dan bantuannya selama ini. Untuk geng Keluarga Newbie, semoga kita masih bisa ngumpul-ngumpul/makan-makan/ngerumpi lagi di Indonesia.

Ibu, bapak, terima kasih atas segalanya. Kalian menjadi sumber inspirasi untuk memulai dan menyelesaikan perjalanan ini. Teruntuk keluarga besar saya di Indonesia, terima kasih atas dukungan moral dan doanya.

Finally, I would like to dedicate this work to Nindya and Asa. Tanpa kalian berdua, perjalanan ini bakal hambar rasanya. Terima kasih telah menuangkan warna di sepanjang perjalanan ini.

This thesis is a culmination of the collective efforts of many, and I am truly fortunate to have had such a wonderful support system. Thank you all for being a part of this journey.

Thanks to you for reading my thesis.

Margi Hartanto

Education Statement of the Graduate School

Experimental Plant Sciences

Issued to:	**Margi Hartanto**
Date:	**12 December 2023**
Group:	**Bioinformatics**
University:	**Wageningen University & Research**

1) Start-Up Phase	*date*	*cp*
► **First presentation of your project**		
Finding genes using system genetics	21 Mar 2019	1.5
► **Writing or rewriting a project proposal**		
► **MSc courses**		
SSB-30306 Molecular Systems Biology	04-25 Jan 2020	6.0
Subtotal Start-Up Phase		7.5

2) Scientific Exposure	*date*	*cp*
► **EPS PhD days**		
EPS PhD days 'Get2Gether', Soest, The Netherlands	10-11 Feb 2020	0.6
EPS PhD days 'Get2Gether', Soest, The Netherlands	3-4 May 2022	0.6
► **EPS theme symposia**		
EPS Theme 4 Symposium 'Genome Biology', Wageningen, The Netherlands	13 Dec 2019	0.3
EPS Theme 4 Symposium 'Genome Biology', online	11 Dec 2020	0.2
► **National platform meetings**		
Annual Meeting Experimental Plant Sciences, Lunteren, The Netherlands	8-9 Apr 2019	0.6
Annual Meeting Experimental Plant Sciences, online	12-13 Apr 2021	0.5
The Dutch Bioinformatics & Systems Biology (BioSB) 2019 conference, Lunteren, The Netherlands	2-3 Apr 2019	0.6
The Dutch Bioinformatics & Systems Biology (BioSB) 2020 conference, online	27-28 Oct 2020	0.5
The Dutch Bioinformatics & Systems Biology (BioSB) 2021 conference, Lunteren, The Netherlands	15-16 Jun 2021	0.6

	The Dutch Bioinformatics & Systems Biology (BioSB) 2022 conference, Lunteren, The Netherlands	28-29 Jun 2022	0.6
►	**Seminars (series), workshops and symposia**		
	B-Wise Seminar: Gerben Hermes & Pariya Behrouzi, Wageningen, The Netherlands	5 Feb 2019	0.2
	B-Wise Seminar: Martijn Huijnen & Mark Sterken, Wageningen, The Netherlands	5 Mar 2019	0.2
	KeyGene's webinar Data Science for Genome Insights: Andrew Carrol & Marcel van Verk, online	9 Jul 2020	0.2
	KeyGene's webinar PanGenomes: Rod Wing & Allen Sessions, online	17 Feb & 18 Mar 2021	0.2
	EPSO 15th Plant Science Seminar – "Data Science for Understanding Plants": Franziska Turck, Alisandra Denton & Thomas Hartwig, online	15 Sep 2021	0.2
	Plant-RX symposium online event: Artificial intelligence in plant science and breeding, online	24 Feb 2021	0.2
	EPS CRISPR/Cas Workshop, Wageningen, The Netherlands	13-14 Sep 2021	0.6
►	**Seminar plus**		
►	**International symposia and congresses**		
	Wageningen Indonesia Scientific Exposure 2019, Wageningen, The Netherlands	12 Mar 2019	0.3
	Workshop 'Modern Statistics for Interdisciplinary Omics and Big Data', online	28-30 Jun 2021	0.9
	9th Plant Genomics & Gene Editing Congress Europe 2022, Den Haag, The Netherlands	11-12 Apr 2022	0.6
►	**Presentations**		
	Poster: "Finding genes for traits using system genetics" at Wageningen Indonesian Scientific Exposure 2019, BioSB conference 2019 & Annual meeting EPS 2019	12 Mar 2019	1.0
	Poster: "The prediction of eQTL causal genes: an initial approach using classification model" at BioSB conference 2020	27-28 Oct 2020	1.0
	Poster: "Exploring Gene Regulatory Interaction in eQTLs using Prior Knowledge Networks" at annual meeting EPS	12-13 Apr 2021	1.0
	Poster: "Mining Gene Expression Regulation in eQTLs using Knowledge Graphs" at Transcription Factors & Transcriptional Regulation course	13-15 Dec 2021	1.0
	Talk: "Network Analysis Prioritizes DEWAX and ICE1 as the Candidate Genes for Major	11 Dec 2020	1.0

eQTL Hotspots in Seed Germination of Arabidopsis thaliana" at EPS Theme 4 Symposium		
Talk: "Linking genes to phenotypes by combining QTL and prior knowledge" at 9th Plant Genomics & Gene Editing Congress Europe 2022	11 Apr 2022	1.0
► **Interviews**		
► **Excursions**		
EPS company visit to Bejo Zaden, online	14 Dec 2020	0.2
EPS company visit Genetwister Technologies, online	22 Mar 2022	0.2
EPS company visit Enza Zaden & Seed meets Technology	29 Sep 2022	0.3

Subtotal Scientific Exposure 15.4

3) In-Depth Studies	*date*	*cp*
► **Advanced scientific courses & workshops**		
PE&RC/WIMEK course: R and Big Data, Wageningen, The Netherlands	26-27 Sep 2019	0.6
BioSB Core Course: Machine Learning for bioinformatics and systems biology, online	5-9 Oct 2020	3.0
BioSB Core Course: Algorithms for Biological Networks, online	1-5 Feb 2021	1.5
EPS course: Transcription Factors & Transcriptional Regulation, Wageningen, The Netherlands	13-15 Dec 2021	1.0
WIAS course: Genomic Prediction in Animal & Plant Breeding, Wageningen, The Netherlands	17-21 Oct 2022	1.5
► **Journal club**		
Bioinformatics literature discussion sessions	2018-2022	3.0
► **Individual research training**		

Subtotal In-Depth Studies 10.6

4) Personal Development	*date*	*cp*
► **General skill training courses**		
WGS PhD Competence Assessment, Wageningen, The Netherlands	18 Jun 2019	0.3
EPS Introduction Course, Wageningen, The Netherlands	28 Sep 2020	0.3
WGS course: Brain friendly working and writing, online	18 March 2020	0.3
WGS course: Project and Time Management, online	13 May - 24 Jun 2020	1.5

WGS course: Career Perspectives, Wageningen, The Netherlands	3 Jun - 5 Jul 2022	1.6
► **Organisation of scientific meetings, PhD courses or outreach activities**		
► **Membership of EPS PhD Council**		

Subtotal Personal Development 4.0

5) Teaching & Supervision Duties	*date*	*cp*
► **Courses**		
BIF-30806 Advanced Bioinformatics	2019/2020/2021	3.0
► **Supervision of BSc/MSc projects**		
BSc thesis "The phenotype quantitative trait locus in Arabidopsis thaliana integrated into AraQTL"	2020	1.0
MSc thesis "Comparative QTL analysis in Arabidopsis and Tomato"	2020	1.0
MSc thesis "The prioritization of eQTL causal genes using machine learning algorithm"	2021	1.0

Subtotal Teaching & Supervision Duties 6.0

TOTAL NUMBER OF CREDIT POINTS*	**43.5**

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS with a minimum total of 30 ECTS credits.

** A credit represents a normative study load of 28 hours of study.*