Using Genome-wide Approaches to Understand Natural Variation in Plant Fitness- and Defense- Related Traits

by

Hao Ji

Bachelor of Science, Shanghai Jiao Tong University, 2007

Submitted to the Graduate Faculty of

The Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH

The Dietrich School of Arts and Sciences

This dissertation was presented

by

Hao Ji

It was defended on

August 27th 2013

and approved by

Carina Barth, Ph.D.

Susan Kalisz, Ph.D., Professor

Mark Rebeiz, Ph.D., Assistant Professor

Stephen Tonsor, Ph.D., Associate Professor

Dissertation Advisor: Brian Traw, Ph.D., Assistant Professor

Using Genome-wide Approaches to Understand Natural Variation in Plant Fitness- and Defense- Related Traits

Hao Ji, Ph.D.

University of Pittsburgh, 2013

Plants have developed sophisticated defense networks to defend against their enemies and increase their fitness. However, there is often significant natural variation in plant fitness- and defense- related traits. If high levels of defense and reproduction are important, then why do plants not all have the highest possible values for these traits? One possible explanation is that these traits are costly to produce and allocation to one trait limits the resources available for another. Such tradeoffs are likely to be very important in understanding why phenotypic variation is maintained in natural populations. However, the genetic mechanisms that underlie them remain poorly understood. Coding regions of known defense- and seed- related genes are typically highly conservative in their sequences, suggesting that sequence differences in those genes are not responsible for the observed phenotypic variation. How then do plants achieve large phenotypic differences in these traits?

In my dissertation studies, I have dissected the genetic architecture underlying tradeoffs in defense and reproductive traits in the model plant, *Arabidopsis thaliana* and identified tradeoffs in a worldwide collection of wild *Arabidopsis* populations between a defense trait, leaf trichome number, and a reproductive trait, mass per seed. I found that two traits were negatively correlated at a major trichome-related locus, ETC2, which encodes a transcription factor. I also performed a Genome-wide Association (GWA) study searching novel candidate genes accounting for the observed variation in resistance to a bacterial pathogen, *Pseudomonas*

syringae DC3000 and found that one of the candidate genes, AtABCG16, appears to assist plant resistance through its effect on abscisic acid-signaling in leaves.

Collectively, my work has taken advantage of reverse genetic mapping, traditional genetic techniques ,and bioinformatical analysis to improve our knowledge of plant natural variation. Moreover, my work suggests that the genetic basis of tradeoffs may involve both transcription factors and hormonal signaling. These findings are novel and are likely to explain why phenotypic variation persists in important defense traits. These results will therefore be important for predicting how these traits will respond to environmental change and how they can be manipulated in agricultural crops.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I am in debt to my committee of Brian Traw, Susan Kalisz, Steve Tonsor, Mark Rebeiz, and Carina Barth for their guidance of my dissertation research. I would like to specifically thank my academic advisor, Brian Traw, for his tutoring and financial support of this research. I also want to thank Joe Martens for his advice and suggestions on my research when he was on my committee. I am grateful for the attention I received from the senior students, Maya Groner, Tarek Elnaccash, Ping Zhang and Lin Hao, particularly during my first year when the environment was totally new to me. The classmate in my cohort, Aaron Stoler, Marnin Wolf, Alison Hale and Sara Hainer, gave me a lot of help during my first year.

I was heavily influenced by George Tseng's microarray analysis course and Daniel Weeks's quantitative genetics course during my third year at University of Pittsburgh and thank to my advisor, Brian Traw, for his encouragement to decide pick genome-wide association study as my dissertation topic. I also owe Lun-ching Chang a debt for his mentorship in helping me improved my computer programming skills.

My comprehensive proposal and oral presentation was improved by the comments of Maya Groner, Tarek Elnaccash and Alison Hale. Especially Maya provided very helpful feedback.

Andy Lariviere and Muhammad Saleem were all great lab mates. They gave me a lot of suggestions on my research and I really enjoyed discussing science with them. Seth Reighard,

Juhyun Kim, Justin Seaman and David Punihaole were very good undergraduate researchers in the lab. They helped me in collecting many of the phenotypic traits that I mapped. I also thank my good friends, Yihong Kang, Desheng Li, Xing Chao, Hailu Jiang, and Yue Zhang, for their kindly help in my life.

I thank my parents and parents in law, Lifang Wang, Zhigang Ji, Xiuqing De and Wei Liu, for their strong encouragement of my interest in statistical genetics and the financial support they have given to me. I owe my wife, Qing, the greatest debt for coming into my life. She is not just the wife who made the food ready when I finished a whole day work in the lab but also used what she learnt from her graduate training to provide me great academic assistance.

TABLE OF ABBREVIATIONS

ABA: **Ab**scisic **A**cid

ABC: **A**TP-**B**inding **C**assette

ABCG: **A**TP-**B**inding **C**assette, **G** subfamily

AFLP: **A**mplified **F**ragments **L**ength **P**olymorphisms

ETC2: **E**nhancer of **T**RY and **C**PC **2**

EMMA: **E**fficient **M**ixed **M**odel **A**ssociation

GO: **G**ene **O**ntology

GWA: **G**enome-**W**ide **A**ssociation

IAA: **I**ndole-3-**A**cetic **A**cid

JA: **J**asmonic **A**cid

LD: **L**inkage **D**isequilibrium

MYB-TF: **My**elo**b**lastosis transcription factor

QTL: **Q**uantitative **T**rait **L**oci

RFLP: **R**estriction **F**ragment **L**ength **P**olymorphisms

RIL: **R**ecombinant **I**nbred **L**ines

RTN: **R**educe **T**richome **N**umber

SNP: **S**ingle **N**ucleotide **P**olymorphisms

SA: **S**alicylic **A**cid

1.0 INTRODUCTION

Understanding the causes and consequences of natural variation is a long-term and major task of ecology (Bolnick *et al.*, 2011) of which one ultimate goal is to discover natural alleles that can be used to improve crops and another is to better understand how plants interact with their biotic and abiotic environments. The natural phenotypic differences, which could be either within or across species, are generally considered as complex outcomes of the interactions among environmental factors and genotypic variations. Plants, particularly, differ significantly in almost every aspect of their life cycles such as how much defense they produced against variable enemies, how many offspring and their qualities, how quick they can be and how big they want, etc. This is likely because plants are stationary and must be equipped for a wide range of environmental conditions. These complex traits are likely constrained by energy economics and evolutionary history. The interaction between genetic and environmental variation makes understanding and manipulating genetic materials involving in such interesting traits a big challenge. To study the genotypic variation and its effects on phenotypic differences, researchers first used genetic markers as the guide to search either in recombinant inbred lines (RILs) or a collection of wild accessions. After evaluating the trait-marker association, genes close to the significant markers were considered for possible causal effects on the phenotypic variation. The first generation of genetic markers such as microsatellites, Amplified Fragments Length Polymorphisms (AFLPs), and Restriction Fragment Length

Polymorphisms (RFLPs) were not easy to acquire and had sparse coverage of the chromosomes (Alonso-Blanco and Koornneef, 2000). With the development of sequencing technology, Single Nucleotide Polymorphisms (SNPs) became popular and led to great breakthroughs in dissecting complex traits. However, even with these recent advances, and an increasing number of genes implicated, few of their functions have been revealed. New methods to increase the accuracy of genetic mapping to evaluate candidates are both urgently needed. In the following paragraphs, I will first talk about the first important genetic mapping method: "QTLs mapping". Secondly, I will introduce a recently developed method: "GWAs mapping". Then, I will compare these two methods and suggest some future directions of this area.

1.1 QTL MAPPING

Genetic mapping is a method that evaluates the association between genotypic markers and phenotypic traits using multiple statistic models. Such types of methods are usually grouped into reverse genetics and are different from forward genetics, which basically acquires the candidate genes from extensive genetic mutant screens. The candidates from traditional mutagen screens are important for understanding how organisms work, but are less useful for understanding how one organism differs from another. So in finding alleles that cause different phenotypic output, reverse genetics is more common and useful. In the model plant, *Arabidopsis thaliana*, the first tool for analyzing genetic architectures of quantitative traits and searching candidate loci accounting for them is Quantitative Trait Loci (QTLs) analysis, which

is also known as traditional linkage mapping (Nam *et al.*, 1989; Z., B., Zeng, 1994). It uses the offspring that usually are the eighth generation (F8) of the inbreeding cross of well-known pedigrees, such as Cvi-0, Ler-0, Ws-0 and Col-0 (Alonso Blanco *et al.*, 1999). The molecular markers used are typically AFLPs and RFLPs but not limited to these two types (Alonso-Blanco *et al.*, 1998; Alonso-Blanco and Koornneef, 2000). By testing the significance of the correlations between the genetic architectures and phenotypic variations, candidate QTLs may then be found for further forward genetic analysis (Larkin *et al.*, 1996; Symonds *et al.*, 2005).

1.2 GWA MAPPING

Genome-wide association mapping (GWAs), a high throughput method, has been developed recently following the great breakthrough in sequencing technologies (Nordborg *et al.*, 2005; Clark *et al.*, 2007; Weigel and Mott, 2009; Platt *et al.*, 2010; Cao *et al.*, 2011). In 2005, Aranzana et al., for the first time, analyzed and reported the genetic information of 96 natural *A.thaliana* accessions, which contained around 3,500 single nucleotide polymorphisms (SNPs). The quick decay of linkage disequilibrium (LD) in *A.thaliana* allows more spontaneous recombination happening, which make the genetic mapping possible (Nordborg *et al.*, 2005). Then in the following seven years, due to the decreasing of cost in genotyping interested populations or species, about 7,000 wild *Arabidopsis* accessions have been genotyped and from the whole genomic sequences of this set of large number of accessions, the SNPs are becoming increasingly dense (Weigel, 2012). By evaluating the association between SNPs and mapped wild accessions, genes or other genetic elements such as non-coding RNA that contain significant SNPs will be considered to be involved in causing the phenotypic variations among

wild populations (Aranzana *et al.*, 2005; Ehrenreich *et al.*, 2009; Chan *et al.*, 2010; Li *et al.*, 2010; Atwell *et al.*, 2010; Brachi *et al.*, 2010; Chan *et al.*, 2011; Fournier-Level *et al.*, 2011; Chao *et al.*, 2012; Filiault and Maloof, 2012).

Given that we have seen many successful cases that GWAs successfully found good candidate genes, some parts of the big picture are still missing (Table 1). First, how should we select candidate genes from a region for which none of the genes has been previously reported? Till present, there are ten major publications of GWAs mapping in *Arabidopsis*, among which four used an empirical significance to select candidate SNPs and five of the rest used arbitrary p-values. For using arbitrary p-values as the threshold, there would be no difference if one or several obvious or major peaks can be found on the map (Atwell *et al.*, 2010; Chao *et al.*, 2012). However, when the mapped quantitative traits are additive and each locus plays a relatively weak role on the overall outcome, empirical threshold could help us to keep the loci on the candidate list. Second, only one paper mapped fitness traits (pod number, Fournier-Level *et al.*, 2011). Understanding the genetic basis underlying the variation of plant fitness traits is important for both agriculture and ecology, thus there is an urgent demand to provide GWAs study on fitness-related traits such as seed mass, seed number and plant bio mass, etc. Third, just half of the publications evaluated their candidate genes with other genome-wide methods and only one paper tested their candidate genes experimentally (Chan *et al.*, 2011). Indeed, lacking of enough experimental evidence has led to many arguments about the accuracy of GWAs study. Providing more genetic evidence of candidate genes will significantly improve the development of GWAs. Last but not least, how to take advantage of using this Arabidopsis accession and SNP dataset mapping system to understand other biological questions? These accessions are collected around the word and the environmental information they experienced is

encrypted on their genome. Dissecting such kind of genetic architectures could significantly improve our knowledge of genome-environment interaction.

Table 1. Summary of the major publications that have ever used GWAs mapping in *Arabidopsis thaliana*.

GWAs-mapping Study	Mapped traits	Rules of picking candidates	Dissecting alleles and loci	Evaluation of the map
Aranzana et al. 2005	Flowering Time Pathogen Resistance	N/A	N/A	N/A
Ehrenreich et al. 2009	Flowering Time	Empirical top 5%	N/A	N/A
Chan et al. 2010	Glucosinolates	Empirical top 0.1%	Yes	QTL analysis
Li et al. 2010	Flowering Time	Arbitrary p-value	N/A	QTL analysis
Atwell et al. 2010*	Flowering Time Isonomics Development Resistance	Arbitrary p-value	N/A	N/A
Brachi et al. 2010	Flowering Time	Arbitrary p-value	N/A	QTL analysis
Chan et al. 2011	Glucosinolates	Empirical top 0.1%	N/A	Proteomic/transcriptomic data T-DNA insertion
Fournier-Level et al. 2011	Survival and Silique	Empirical top 0.05%	Yes	No
Chao et al.2012	Metal Tolerance	Arbitrary p-value	Yes	QTL analysis Complementation analysis qRT-PCR
Filiault et al. 2012	Shade Avoidance	Arbitrary p-value	N/A	GO analysis Microarray

1.3 COMPARING TWO MAPPING METHODS

Indeed, QTLs mapping and GWAs mapping do not differ from each other fundamentally. They both use statistic models to evaluate trait-marker associations. Meanwhile, due to the design of two mapping methods, combining these two methods may lead to better understanding of mapping phenotypic variation (Table 2). GWAs mapping is a high throughput method, which can scan genetic loci accounting for variation in hundreds of populations at a time. It also provides better resolution since all of its genetic markers are single nucleotide. However, it has its own considerations, of which the most important is population structure. Since GWAs mapping mostly relays on the natural recombination across the entire genome, population structure could cause significant bias and false positive thus become a strong confounder in association studies, which was estimated up to 40% in one extreme case (Bergelson and Roux, 2010). Statistically, the Wilcoxon rank-sum test that is usually used in evaluating trait-mark accession requires variables to be independent. Population structure, which here means some accessions are more closed to each other, clearly violates this assumption. Biologically, if the importance of a particular allele in explaining an interested phenotypic variation among a global collection of wild accessions is always determined or influenced by a certain rare group of populations, GWAs mapping could miss such important loci due to either too many groups or too few accessions with a group. On the other side, QTLs mapping, because of its mapping populations being created by artificial crossing of two different parental lines, has no issues with population structure. The considerations of QTLs mapping are lower coverage on the chromosome and small number of accessions to search with. However, these are what GWAs can complement. So there are suggestions and experimental evidence showing that these two methods can be used together to search candidate QTLs

(Nordborg and Weigel, 2008; Brachi *et al.*, 2010; Weigel, 2012). For example, flowering time is a quantitative trait that has been studied for a long time. What Brachi et al. reported is that some of the significant SNPs identified from GWAs mapping are within 20kb of previously reported QTL regions (Brachi *et al.*, 2010). Similarly, Chao et al. used GWAs and QTL to find a heavy metal ATPase transporter, HMA3, involved in plant leaf cadmium response (Chao *et al.*, 2012). My unpublished data also suggests that the beginning of first chromosome of *Arabidopsis* is important for regulating seed size, which is supported by both GWAs and QTL mapping (Alonso Blanco *et al.*, 1999). Mean while, it is not uncommon that some loci are not seen under QTL mapping but are detected by GWAs and vice verse since they have different mechanisms. But if the goal is to find major loci contributing to phonotypic variation, this should not be considered as a problem.

Table 2. Comparison of artificial-population based (QTL) and natural-accession based (GWA) genetic mapping methods to dissect the complex traits in *Arabidopsis thaliana*.

	QTL Mapping	GWA Mapping
Disadvantages	▪ Mapping genetic diversity form two parental pedigrees at a time. ▪ Crossing and purifying genotype are time consuming and sometimes are hard to acquire.	▪ Confounding bias caused by population structure. ▪ Needing genome-wide sequences of a large number of natural accessions. ▪ Missing rare and weak effect alleles
Advantages	▪ No issues about population structures. ▪ Identifying less frequent alleles. ▪ No need of large amount of sequencing.	▪ Mapping a large set of natural accessions at a time. ▪ Allowing fine map because of high density of genetic markers.
Benefits of combining two methods	▪ Using GWA mapping to first scan thousands populations and using QTL to study specific population with rare or weak alleles ▪ Using QTL mapping to check and control potential bias in GWA mapping caused by population structure.	

1.4 FUTURE DIRECTION

Now, let's go back to the very basic of these genome-wide approaches. GWAs mapping finds the candidate loci causing the phenotypic variation. So there should be at least two requirements for such candidates. First, the primary statement is that they need to participate in regulating or causing the interested phenotype. Then, second, there are polymorphisms existed in the candidates that generate natural differences. Based on these criteria, other genome-wide approaches could be able to help us evaluate the candidates of GWAs mapping. Trancriptomics such as microarrays and RNA-seq are great tools. Since the mid 1990s, DNA/cDNA based microarrays have been the primary techniques for genome-wide search of gene expression levels in plants. The ability of these arrays to measure thousands of transcripts at a time has led to important advances in searching novel candidate genes for interested phenotypic changes such as hormonal response and pathogen resistance (Schenk *et al.*, 2000; Thibaud-Nissen *et al.*, 2006). The fact of its ability to scan the whole genome makes it a good complementary tool of GWAs mapping in evaluating the mapping results. Nonetheless, microarrays have their limitations. The major one is the final results is pre-determined by the probes and chips. In other words, if there is no probe coverage on the chromosomes, then nothing can be detected from that region. For example, there are three tandem repeats of MYB-like transcription factor at the Reduce Trichome Number (RTN) locus. However, microarray chip ATH-121501 only contains the probe to detect one of the three genes, ENHANCER of TRY and CPC 2 (ETC2). Detecting the expression of the other two trichome related genes, Trichome Less 1 and Trichome Less 2 (TCL1 and TCL2), are not available if using this chip. Recently, high-throughput RNA sequencing, also known as RNA-seq, has shown its power and advantages in measuring gene expression. In other species, several technologies, including those developed

by 454 Life Sciences (Margulies *et al.*, 2005) and Illumina (Bennett *et al.*, 2010), has successfully been used to explore genetic variation (Korbel *et al.*, 2007; Mortazavi *et al.*, 2008; Pickrell *et al.*, 2010). Although there has been no similar report so far in *A.thaliana,* I suggest that RNA-seq might be a great tool for detecting natural differences among wild *Arabidopsis* populations for the following reasons (Marioni *et al.*, 2008; R., Lister *et al.*, 2009). First, RNA-seq is sequence-based. It provides the counts of detected fragments, which would allow us to see the variation of sequences and their effect on gene expressions among different accessions, which is also known as allelic specific expression (Sun, 2011; Sun and Hu, 2013). It may provide the information about alternative spliced fragments (Sun and Hu, 2013). Second, RNA-seq allows us to measure and compare gene expressions under different SNPs more accurately. Microarray, on the other hand, has a strong issue with its probes' different hybridization properties (Marioni *et al.*, 2008). Since the *Arabidopsis* genome is smaller and simpler, RNA-seq can focus on searching SNPs across mapping populations without considering heterozygosity, which will make it even more efficient compared with the usage in Human genetics.

1.5 GOALS OF MY THESIS

In my dissertation study, I focused on the following questions: 1) Can we combine GWA and QTL to search for candidate loci in explaining interested phenotypic variation and how? 2) Can forward genetic techniques be used to evaluate the candidate genes found by GWA mapping successfully, particularly in explaining variations in wild accessions? 3) Why do some of the wild accessions maintain low levels of beneficial traits?

To better introduce how I combined bioinformatics, biostatistics and genetic experimental procedures to address the three main goals of my dissertation research, I created a loop diagram listing six key steps of understanding natural variation in plant fitness- and defense- related traits (Figure 1). 1) First, I scored the phenotypic values of interested traits from wild *Arabidopsis* accessions. If significant variation was found, 2) I then performed genetic mapping analysis searching candidate genes associated with observed phenotypic variation. 3) Next, I assess the potential function candidate genes in effecting mapped phenotypes by analyzing other genome-wide data such as microarray. 4) Then I experimentally test the hypothesis I made in step3 about how the candidate genes affecting studied traits. 5) After the genes' function had been revealed, I then tried to dissect the alleles at candidate gene loci and then 6) test the allelic effect on the studied phenotypes to see if these alleles could cause the observed phenotypic variation. These six steps, together, formed a loop beginning with observing phenotypic variation and ending up with finding allelic effects causing such variation.

My thesis will contain three data chapters in response to answer the three questions listed above: 1) Using GWA and QTL maps together to find candidate loci associated with seed mass of wild *Arabidopsis* accessions. 2) Genome-wide association mapping to locate an ABC-transporter, ABCG16, involved in plant resistance to bacterial pathogen, *Pseudomonase syringae pv. tomato* DC3000 (*Pst.* DC3000). 3) Genome-wide approaches revealing a major fitness trade-off at a particular defense-related locus, ENHANCER of TRY and CPC 2 (ETC2).

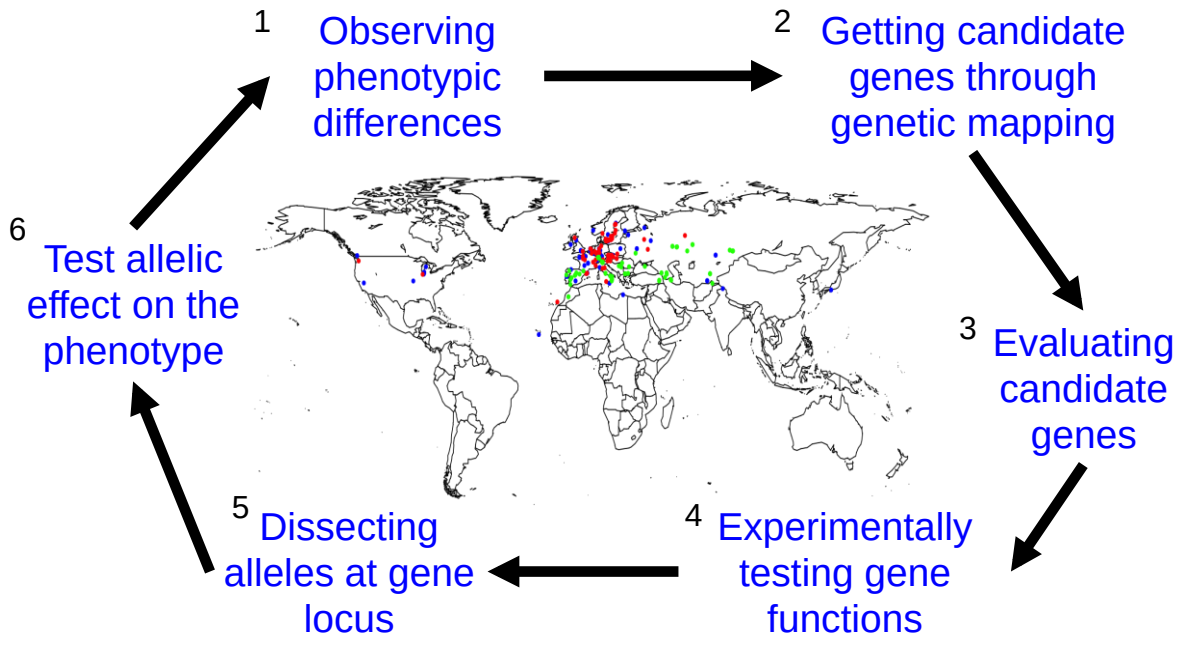


Figure 1. A "loop" containing six steps to study natural variation in plant traits.

2.0 GLOBAL MAP OF NATURAL ALLELIC VARATION IN SEED MASS OF *ARABIDOPSIS THALIANA* WILD ACCESSIONS

Seed mass is a complex plant fitness trait, and it plays a major role in plant growth, development, migration, colonization and ecological succession. It is a trait of both ecological and agricultural importance. Presently, little is known about the genetic architecture of this trait, and the underlying allelic variations in wild populations remain to be identified. Here, I show the outcome of a broad genome-wide association study based on genotyping 204,902 SNP variants across 178 RegMap global collection of *Arabidopsis thaliana* that identifies candidate genes likely to determine the phenotypic variation in mass per seed. I report five loci and forty genes that could be responsible for natural variation in seed mass. Using published QTL analysis, microarray data and GO terms, I find that my candidates are strongly enriched with seed-related functions and provide an important resource for future work. Also, I observed a significant geographic pattern with alleles for heavier seeds being more represented in high latitude populations. My findings will be helpful in future studies aiming at investigating seed mass in plants with importance for both agriculture and ecology.

2.1 INTRODUCTION

Seed mass is an important trait of both agricultural and ecological importance and provides us food, fiber and fuel. As a fundamentally important but inherently complex trait, it determines plant growth, development, migration, colonization and ecological succession (Oerke, 2006; Westoby *et al.*, 2002; Martínez-Andújar *et al.*, 2012). The variation in seed mass across the plant species is significant (Todesco *et al.*, 2010; Moles *et al.*, 2005) and having small or large seeds both have certain advantages and disadvantages for plants. For instance, relatively small seed producing species are thought to be superior colonizers while heavy seeds are considered to produce larger seedlings, which are capable to better survive in the presence of different environmental stressors such as light or nutrients stress, dry spells, partial damage, shade and other biotic stressors (Jones and Dangl, 2006; Westoby *et al.*, 2002). Though these is voluminous amount of literature which deals with see production and seed mass in different plant species from agricultural to ecological perspectives. However, a little is known about natural allelic variations in seed mass at the genotypic level in wild populations. Such kind of lack of genetic understanding of this important trait in wild populations limits our intellectual and scientific vision to address the key issues such as enhanced and quality food production (seed production) and species conservation in nature. Interestingly, there are certain plant species, which cover a broad geographical niche from east to west, and thus could be used as model plant systems to investigate the genetic bases of key functional traits such as seed mass in nature. However, no latitudinal assessment of seed size in *Arabidopsis* has been done to my knowledge.

Arabidopsis thaliana is one of these unique plant model systems that offer wonderful genetic resources, ecological and geographic breadth to analyze natural allelic variation in seed mass. There have been several reports that showed the importance of particular genes required for making a seed in *A. thaliana* (Vlot *et al.*, 2009; Jako *et al.*, 2001; Ohto *et al.*, 2005; Fang *et al.*, 2012), but whether these same genes also control differences in mass per seed among wild populations remains largely unknown today. Meanwhile, previous assessments of the genetic architecture underlying variation in seed mass have focused on QTL mapping of some lines (Vlot *et al.*, 2009; Alonso Blanco *et al.*, 1999; Herridge *et al.*, 2011; Van Daele *et al.*, 2012). These approaches have identified only a handful of likely to influence seed mass in those accessions of *A. thaliana* (de Torres Zabala *et al.*, 2007; Mizukami and Fischer, 2000; Flint-Garcia *et al.*, 2003; Jofuku *et al.*, 2005; Luo *et al.*, 2005; Schruff *et al.*, 2006; Herridge *et al.*, 2011; Van Daele *et al.*, 2012). Very recently, Van Daele and colleagues (2012) suggested that the gene(s) responsible for seed traits are required to be identified through either further fine mapping of the QTL region that is a time-consuming process or through genome-wide association (GWA) mapping by extensively genotyping of wild lines. Presently, there is a pressing demand to dissect the genetic architecture of this major trait to answer unresolved questions in seed biology of 21st century to address the forthcoming challenges in agricultural and ecological research (Vlot *et al.*, 2009; Nambara and Nonogaki, 2012; Sreenivasulu and Wobus, 2013).

In fact, at present, there is no study on global mapping of seed mass in wild populations. My study therefore provides a first integrated view of the genetic basis of seed mass by combing GWA and QTL analysis in 178 wild populations of *A. thaliana*. Such combinations of taking advances of both mapping methods have been suggested to be more efficient and

precisely in locating candidate genetic material previously (de Torres Zabala *et al.*, 2007; Nordborg and Weigel, 2008; Weigel, 2012) but have only been done twice previously (Brachi *et al.*, 2010, Chao *et al.*, 2012). Here, I extensively map and show the genetic bases of these traits by identifying key loci on different chromosomes, and unveil the amount of potential genes that may determine variation in seed mass in wild populations.

2.2 MATERIALS AND METHODS

2.2.1 Mass per seed measurement

Seeds were obtained from the Regmap collections (Atwell *et al.*, 2010, Horton *et al.*, 2012). For each accession, three sets of seeds were measured phenotypically using a Mettler Toledo MX5 microbalance at the University of Pittsburgh and the seed number was counted. Each of the three sets contained about ten to thirty seeds. Mass per seed was then calculated by dividing the observed mass by the number of seeds. The value used for the GWA map was the averaged value of the three measurements.

2.2.2 Genome-wide association mapping

Wilcoxon rank-sum tests were used to calculate the significance of association between each SNP and the phenotypic values using standard methods (Atwell *et al.*, 2010). To handle confounding caused by population structure, I used the standard approach of mixed model analysis known as Efficient Mixed-Model Association (EMMA) (H., M., Kang *et al.*, 2008).

All programs were coded in R (http://cran.r-project.org/). I presented log transformed P-values following the standard approach and called this type of values "GWAs score".

2.2.3 Geographic analysis

The geographic information (region, country, latitude, and longitude) I used in this study was acquired from the *Arabidopsis* 1001 program (http://www.1001genomes.org/). The polynomial regression, contour plot, and 3D surface plot were performed by R package *rsm* (http://cran.r-project.org/web/packages/rsm/index.html). The map of Europe was made by R package "*maps"* (http://cran.r-project.org/web/packages/maps/).

2.2.4 Keyword enrichment

Keyword analyses were performed based on TAIR through searching genes with the keyword webpage (http://www.arabidopsis.org/servlets/). The "keyword term" menu was set as "contains". The keywords for searching genes were: "growth", "metabolism", "transport", "transcription", "protein synth", "signaling", "cell cycle", "embryo", "seed", "endosperm", "leaf", "root", "flower". The gene lists returned from the website were then used as the gene pool and I counted how many of the forty candidate genes were included in each of the gene pools. To test the significance statistically, I performed two methods. First, I performed a Chi-sq test using the total gene number, number of genes with each keyword and numbers of candidate genes included in each gene pool. I also performed a permutation analyses by asking given a certain number of genes with each keyword, what is the probability that I would get the observed number in the forty candidate gene list. For each keyword, I permuted 20,000 times to

create the null distribution. The total number of genes in *Arabidopsis thaliana*, the number and the locus information of genes with each keyword search were all collected in between February and March 2013. Since the GO terms are continuously being updated, those numbers will change in the future.

2.3 RESULTS

2.3.1 Genome-wide association analysis of mass per seed in *Arabidopsis thaliana*

To rigorously assess the distribution of natural variation in seed size, I measured 194 wild *A.thaliana* accessions from the stock center, collected worldwide (Figure 2A, *Arabidopsis* biological Resource Center, http://abrc.osu.edu/), and I found that these accessions differed significantly in their seed masses, ranging from 16.4 to 35.8 μg (Figure 2B, Table 3). To dissect the genetic bases underlying the differences in seed mass, I used 204,902 genome-wide SNP markers that were derived from the 250K SNP data version 3.06 (Atwell *et al.*, 2010). All SNPs used in the analysis were diallelic and had the minor nucleotide represented in more than 5% of the accessions. The Wilcoxon rank-sum test, a non-parametric approach, was used to test the significance of association between each SNP and the observed seed mass. Moreover, to control the confounding caused by population structure, mixed model analysis commonly known as EMMA (Efficient Mixed-Model Association, (H., M., Kang *et al.*, 2008)) was used. During GWAs mapping, seventeen wild accessions were excluded since their SNP information was not available, thus making the total number of mapped accessions to 178 (Table 3). The obtained values were then plotted against each SNP position on five chromosomes (Figure 3A).

SNPs with higher significance in explaining large portions of variations of mass per seed were then arranged in peaks. To decrease the noise and control the side effect of having single SNPs with a higher GWA score in a certain region of the chromosome, I then smoothed the map by averaging the GWA score of every ten consecutive SNPs. Some distinct peaks were then visible under maps using both the Wilcoxon and the EMMA methods (Figure 3B-C). Compared with dominant and obvious peaks in other GWA studies of trichome density and heavy metal tolerance (Atwell *et al.*, 2010; Chao *et al.*, 2012), the moderate and multiple peaks indicate that the variation of mass per seed here is more polygenic. Interestingly, the peaks identified by the Wilcoxon method showed relatively high significance (i.e., *p*-values) as compared to those peaks observed in the EMMA scan. The difference in *p*-values is consistent with the previous reports that EMMA would cause a decrease or a loss of genetic patterns on the map (Filiault and Maloof, 2012; Atwell *et al.*, 2010). Therefore, I decided to use both Wilcoxon and EMMA results since they all had their advantages/disadvantages. To select the candidate loci and genes, I averaged the GWAs score of every ten adjacent SNPs and performed this smoothing across the entire genome. I did this type of smoothing for both the Wilcoxon and EMMA maps. This smoothing method helped me to decrease the noise on the map and focus on major areas with multiple significant SNPs instead of looking in a region with just a single SNP with high GWAs scores. I focused on peaks that contained SNPs with top 0.1% empirical smoothed GWAs scores. Only peaks present in both smoothed Wilcoxon and EMMA map were processed for later evaluation. For each selected peak, the SNP with the highest GWAs score in the region was set as the center and the linkage disequilibrium between every SNP and the central SNP was calculated. Finally, I plotted the LD values against the chromosome positions and fitted the LD curve using the *smooth.spline* method in R package.

Only the region where the curve was above 0.3 was considered as the candidate locus and the genes with this locus were considered as candidate genes (Figure 4). This methodology is substantially more rigorous than most previous GWAs studies. However, since the LD decays quickly and variably across the *Arabidopsis* genome (Nordborg *et al.*, 2005), deciding the width of candidate loci by calculating local LD is most appropriate. Genes presented within my LD blocks were then considered as candidate genes (Table 4, Figure 5). Overall, following the described selection criteria of candidate SNPs, GWA scanning pinpointed five major peaks (Figure 3, Figure 5), which contained 40 candidate genes (Table 4). Interestingly, focal SNPs that were significantly associated with phenotypic variation were also tightly linked during recombination (Figure 7).

Table 3. Information of all *Arabidopsis* accessions used in this study.

Orange blocks represent the accessions of which the information was not available in the database. Yellow, blue and gray colors in the five loci represent the heavy, light and mixed alleles, respectively.

Name	ecotype_id	MPS Class	Mass per Seed (ug)	±SD	Latitude	Longitude	Region	Country	Euro.Line	Alleles@Locus1	Alleles@Locus2	Alleles@Locus3	Alleles@Locus4	Alleles@Locus5
RRS-7	7514	24	27.2	1.1	41.56	-86.43	Midwest	USA	No	Light Allele	Heavy Allele	Heavy Allele	Mixed Allele	Heavy Allele
RRS-10	7515	24	26.9	1.2	41.56	-86.43	Midwest	USA	No	Heavy Allele	Mixed Allele	Heavy Allele	Heavy Allele	Heavy Allele
KNO-10	6927	24	27.9	0.8	41.28	-86.62	Midwest	USA	No	Heavy Allele	Mixed Allele	Heavy Allele	Heavy Allele	Heavy Allele
KNO-18	6928	24	27.2	1.0	41.28	-86.62	Midwest	USA	No	Heavy Allele	Mixed Allele	Heavy Allele	Mixed Allele	Light Allele
RMX-A02	7524	28	28.1	0.5	42.04	-86.51	Midwest	USA	No	Heavy Allele	Mixed Allele	Heavy Allele	Heavy Allele	Heavy Allele
RMX-A180	7525	20	21.6	0.2	42.04	-86.51	Midwest	USA	No	Light Allele	Mixed Allele	Light Allele	Light Allele	Heavy Allele
PNA-17	7523	24	25.4	0.6	42.09	-86.33	Midwest	USA	No	Mixed Allele	Light Allele	Heavy Allele	Light Allele	Heavy Allele
PNA-10	7526	24	27.1	1.6	42.09	-86.33	Midwest	USA	No	Heavy Allele	Mixed Allele	Heavy Allele	Heavy Allele	Heavy Allele
Eden-1	6009	28	28.5	0.6	62.88	18.18	N Sweden	SWE	Yes	Heavy Allele	Heavy Allele	Mixed Allele	Heavy Allele	Heavy Allele
Eden-2	6913	20	23.6	0.8	62.88	18.18	N Sweden	SWE	Yes	Heavy Allele	Light Allele	Heavy Allele	Heavy Allele	Heavy Allele
Lov-1	6043	24	27.8	1.7	62.80	18.08	N Sweden	SWE	Yes	Heavy Allele	Heavy Allele	Heavy Allele	Heavy Allele	Heavy Allele
Lov-5	6046	28	28.8	0.7	62.80	18.08	N Sweden	SWE	Yes	Heavy Allele	Heavy Allele	Heavy Allele	Mixed Allele	Heavy Allele
Fab-2	6917	20	23.5	1.5	63.02	18.32	N Sweden	SWE	Yes	Heavy Allele	Heavy Allele	Heavy Allele	Heavy Allele	Light Allele
Fab-4	6918	28	29.6	1.8	63.02	18.32	N Sweden	SWE	Yes	Heavy Allele	Heavy Allele	Heavy Allele	Heavy Allele	Heavy Allele
Bil-5	6900	24	24.1	1.2	63.32	18.48	N Sweden	SWE	Yes	Heavy Allele	Heavy Allele	Mixed Allele	Light Allele	Light Allele
Bil-7	6901	20	23.8	1.0	63.32	18.48	N Sweden	SWE	Yes	Heavy Allele	Heavy Allele	Heavy Allele	Light Allele	Light Allele
Var-2-1	7516	28	29.4	1.1	55.58	14.33	S Sweden	SWE	Yes	Light Allele	Heavy Allele	Mixed Allele	Mixed Allele	Heavy Allele
Var-2-6	7517	32	34.3	0.7	55.58	14.33	S Sweden	SWE	Yes	Light Allele	Heavy Allele	Heavy Allele	Mixed Allele	Heavy Allele
Spr-1-2	6964	24	25.1	0.5	56.30	16.00	C Sweden	SWE	Yes	Light Allele	Mixed Allele	Light Allele	Mixed Allele	Light Allele
Spr-1-6	6965	20	20.4	0.6	58.42	14.16	C Sweden	SWE	Yes	Light Allele	Heavy Allele	Light Allele	Light Allele	Light Allele
Omo-2-1	7518	20	20.7	0.9	56.15	15.77	S Sweden	SWE	Yes	Light Allele	Light Allele	Light Allele	Mixed Allele	Light Allele
Omo-2-3	7519	16	19.4	0.4	56.15	15.77	S Sweden	SWE	Yes	Light Allele	Mixed Allele	Light Allele	Light Allele	Heavy Allele
Ull-2-5	6974	16	19.3	1.3	56.06	13.97	S Sweden	SWE	Yes	Light Allele	Light Allele	Light Allele	Light Allele	Light Allele
Ull-2-3	6973	20	22.7	1.9	56.06	13.97	S Sweden	SWE	Yes	Heavy Allele	Heavy Allele	Light Allele	Mixed Allele	Light Allele
Zdr-1	6984	24	26.7	0.4	49.39	16.25	Moravia	CZE	Yes	Light Allele	Heavy Allele	Light Allele	Mixed Allele	Light Allele
Zdr-6	6985	20	23.4	1.3	49.39	16.25	Moravia	CZE	Yes	Light Allele	Mixed Allele	Mixed Allele	Light Allele	Light Allele
Bor-1	5837	24	24.7	0.7	49.40	16.23	Moravia	CZE	Yes	Light Allele	Light Allele	Mixed Allele	Mixed Allele	Light Allele
Bor-4	6903	20	22.9	0.5	49.40	16.23	Moravia	CZE	Yes	Light Allele	Mixed Allele	Mixed Allele	Light Allele	Light Allele
Pu2-7	6956	20	21.9	0.3	49.42	16.36	Eastern Europe	CZE	Yes	Light Allele	Heavy Allele	Light Allele	Light Allele	Light Allele
Pu2-23	6951	20	23.5	0.6	49.42	16.36	Eastern Europe	CZE	Yes	Light Allele	Heavy Allele	Mixed Allele	Mixed Allele	Light Allele
LP2-2	7520	20	22.6	0.4	49.38	16.81	Moravia	CZE	Yes	Light Allele	Light Allele	Light Allele	Light Allele	Mixed Allele
LP2-6	7521	20	23.0	0.4	49.38	16.81	Moravia	CZE	Yes	Light Allele	Light Allele	Light Allele	Mixed Allele	Light Allele
HR-5	6924	16	19.8	0.5	51.41	-0.64	Northern Europe	UK	Yes	Light Allele	Light Allele	Light Allele	Mixed Allele	Light Allele
HR-10	6923	32	35.5	3.0	51.41	-0.64	Northern Europe	UK	Yes	Light Allele	Mixed Allele	Heavy Allele	Mixed Allele	Heavy Allele
NFA-8	6944	20	22.8	1.2	51.41	-0.64	Northern Europe	UK	Yes	Light Allele	Light Allele	Heavy Allele	Light Allele	Light Allele
NFA-10	6943	20	22.3	1.0	51.41	-0.64	Northern Europe	UK	Yes	Light Allele	Mixed Allele	Light Allele	Mixed Allele	Light Allele
SQ-1	6966	16	18.1	1.1	51.41	-0.64	Northern Europe	UK	Yes	Light Allele	Light Allele	Light Allele	Mixed Allele	Light Allele
SQ-8	6967	20	20.3	0.2	51.41	-0.64	Northern Europe	UK	Yes	Light Allele	Light Allele	Heavy Allele	Light Allele	Heavy Allele
CIBC-5	6730	24	24.4	0.7	51.41	-0.64	Northern Europe	UK	Yes	Light Allele	Mixed Allele	Heavy Allele	Mixed Allele	Light Allele
CIBC-17	6907	24	25.7	2.4	51.41	-0.64	Northern Europe	UK	Yes	Light Allele	Mixed Allele	Heavy Allele	Light Allele	Heavy Allele
TAMM-2	6968	20	21.4	0.9	60.00	23.50	Northern Europe	FIN	Yes	Heavy Allele	Heavy Allele	Mixed Allele	Heavy Allele	Light Allele
TAMM-27	6969	20	23.0	0.8	60.00	23.50	Northern Europe	FIN	Yes	Heavy Allele	Heavy Allele	Mixed Allele	Heavy Allele	Light Allele
KZ-1	6930	16	18.5	0.9	49.50	73.10	Central Asia	KAZ	No	Light Allele	Light Allele	Light Allele	Mixed Allele	Mixed Allele
KZ-9	6931	16	17.0	0.8	49.50	73.10	Central Asia	KAZ	No	Light Allele	Heavy Allele	Light Allele	Light Allele	Mixed Allele
GOT-7	6921	32	32.7	0.1	51.53	9.94	Western Europe	GER	Yes	Heavy Allele	Heavy Allele	Light Allele	Mixed Allele	Heavy Allele
GOT-22	6920	32	33.3	0.5	51.53	9.94	Western Europe	GER	Yes	Heavy Allele	Heavy Allele	Light Allele	Mixed Allele	Heavy Allele
REN-1	6959	24	26.0	1.2	48.50	-1.41	Western Europe	FRA	Yes	Light Allele	Light Allele	Heavy Allele	Light Allele	Heavy Allele
REN-11	6960	20	22.4	1.0	48.50	-1.41	Western Europe	FRA	Yes	Light Allele	Light Allele	Mixed Allele	Light Allele	Light Allele
Uod-1	6975	20	20.2	1.0	48.30	14.45	Western Europe	AUT	Yes	Light Allele	Light Allele	Light Allele	Light Allele	Light Allele
Uod-7	6976	20	23.5	0.7	48.30	14.45	Western Europe	AUT	Yes	Light Allele	Mixed Allele	Light Allele	Light Allele	Light Allele
Cvi-0	6911	32	32.7	1.3	15.11	-23.62	Macaronesia	CPV	No	Light Allele	Heavy Allele	Mixed Allele	Mixed Allele	Heavy Allele
Lz-0	6936	28	28.3	0.8	46.00	3.30	Western Europe	FRA	Yes	Light Allele	Light Allele	Mixed Allele	Mixed Allele	Mixed Allele
Ei-2	6915	20	22.2	0.5	50.30	6.30	Western Europe	GER	Yes	Light Allele	Light Allele	Light Allele	Light Allele	Heavy Allele
Gu-0	6922	20	22.3	0.1	50.30	8.00	Western Europe	GER	Yes	Light Allele	Heavy Allele	Mixed Allele	Light Allele	Light Allele
Ler-1	6932	24	24.0	0.3	47.98	10.87	Western Europe	GER	Yes	Light Allele	Light Allele	Light Allele	Light Allele	Light Allele
Nd-1	6942	24	24.6	1.6	50.00	10.00		SUI	Yes	Heavy Allele	Light Allele	Light Allele	Mixed Allele	Heavy Allele
C24	6906	28	28.1	0.3	40.21	-8.43	Southern Europe	POR	Yes	Light Allele	Heavy Allele	Light Allele	Mixed Allele	Light Allele
CS22491	7438	16	17.0	0.6	61.36	34.15		RUS	Yes	Light Allele	Light Allele	Light Allele	Light Allele	Light Allele
Wei-0	6979	24	26.5	0.3	47.25	8.26	Western Europe	SUI	Yes	Light Allele	Heavy Allele	Light Allele	Mixed Allele	Heavy Allele
Ws-0	6980	20	20.4	0.1	52.30	30.00	Eastern Europe	RUS	Yes	Light Allele	Light Allele	Light Allele	Light Allele	Light Allele
Yo-0	6983	24	24.4	1.6	37.45	-119.35	Northern California	USA	No	Light Allele	Light Allele	Heavy Allele	Light Allele	Light Allele
Col-0	6909	20	22.6	0.7	38.30	-92.30	Midwest	USA	No	Light Allele	Light Allele	Light Allele	Mixed Allele	Light Allele
An-1	6898	20	23.0	0.4	51.22	4.40	Western Europe	BEL	Yes	Light Allele	Mixed Allele	Heavy Allele	Mixed Allele	Light Allele
Van-0	6977	20	22.4	0.7	49.30	-123.00	Pacific Northwest	CAN	No	Light Allele	Light Allele	Light Allele	Mixed Allele	Heavy Allele
Br-0	6904	24	24.7	1.1	49.20	16.62	Moravia	CZE	Yes	Light Allele	Mixed Allele	Light Allele	Light Allele	Heavy Allele
Est-1	6916	24	24.8	0.3	58.30	25.30		RUS	Yes	Light Allele	Heavy Allele	Light Allele	Mixed Allele	Heavy Allele
Ag-0	6897	28	29.1	0.8	45.00	1.30	Western Europe	FRA	Yes	Light Allele	Light Allele	Mixed Allele	Light Allele	Heavy Allele
Gy-0	8214	24	25.1	2.2	49.00	2.00	Western Europe	FRA	Yes	Light Allele	Mixed Allele	Heavy Allele	Light Allele	Heavy Allele
Ra-0	6958	24	26.2	0.3	46.00	3.30	Western Europe	FRA	Yes	Light Allele	Light Allele	Light Allele	Mixed Allele	Heavy Allele
Bay-0	6899	24	26.3	1.2	49.00	11.00	Western Europe	GER	Yes	Light Allele	Light Allele	Light Allele	Mixed Allele	Light Allele
Ga-0	6919	20	23.4	0.4	50.30	8.00	Western Europe	GER	Yes	Light Allele	Mixed Allele	Light Allele	Light Allele	Light Allele
Mrk-0	6937	32	32.3	0.9	49.00	9.30	Western Europe	GER	Yes	Light Allele	Heavy Allele	Light Allele	Light Allele	Heavy Allele
Mz-0	6940	20	23.5	1.2	50.30	8.30	Western Europe	GER	Yes	Light Allele	Light Allele	Light Allele	Mixed Allele	Heavy Allele
Wt-5	6982	20	23.3	1.4	52.30	9.30	Western Europe	GER	Yes	Heavy Allele	Light Allele	Light Allele	Mixed Allele	Heavy Allele
Kas-1	8424	28	31.8	0.3	35.00	77.00	South Asia	IND	No	Light Allele	Heavy Allele	Light Allele	Mixed Allele	Mixed Allele
Ct-1	6910	20	23.9	0.9	37.30	15.00	Southern Europe	ITA	Yes	Light Allele	Light Allele	Light Allele	Light Allele	Light Allele
Mr-0	7522	32	32.9	0.7	44.15	9.65	Southern Europe	ITA	Yes	Light Allele	Heavy Allele	Light Allele	Mixed Allele	Heavy Allele
Tsu-1	6972	24	24.3	0.9	34.43	136.31	Eastern Asia	JPN	No	Light Allele	Mixed Allele	Light Allele	Mixed Allele	Mixed Allele
Mt-0	6939	20	21.0	0.1	32.34	22.46	Cyrenaica	LIB	No	Light Allele	Light Allele	Light Allele	Mixed Allele	Light Allele
Nok-3	6945	24	24.7	0.8	52.24	4.45		NED	Yes	Light Allele	Light Allele	Light Allele	Light Allele	Light Allele
Wa-1	6978	28	30.8	2.4	52.30	21.00	Eastern Europe	POL	Yes	Light Allele	Light Allele	Light Allele	Mixed Allele	Light Allele
Fei-0	8215	20	21.1	0.6	40.50	-8.32	Southern Europe	POR	Yes	Light Allele	Light Allele	Light Allele	Light Allele	Heavy Allele
Se-0	6961	24	25.4	1.5	38.33	-3.53	Southern Europe	ESP	Yes	Light Allele	Light Allele	Light Allele	Light Allele	Heavy Allele
Ts-1	6970	24	24.6	0.7	41.72	2.93	Southern Europe	ESP	Yes	Light Allele	Heavy Allele	Heavy Allele	Light Allele	Heavy Allele
Ts-5	6971	28	28.9	0.9	41.72	2.93	Southern Europe	ESP	Yes	Light Allele	Mixed Allele	Heavy Allele	Light Allele	Heavy Allele
Pro-0	8213	20	23.9	1.0	43.25	-6.00	Southern Europe	ESP	Yes	Light Allele	Mixed Allele	Mixed Allele	Light Allele	Heavy Allele
LL-0	6933	16	19.2	0.6	41.59	2.49	Southern Europe	ESP	Yes	Light Allele	Mixed Allele	Light Allele	Light Allele	Heavy Allele
Kondara	6929	20	22.9	0.4	38.48	68.49	Central Asia	TJK	No	Light Allele	Heavy Allele	Light Allele	Mixed Allele	Mixed Allele
Shahdara	6962	20	22.8	0.4	38.35	68.48	Central Asia	TJK	No	Light Allele	Heavy Allele	Light Allele	Mixed Allele	Mixed Allele
Sorbo	6963	24	27.1	1.1	38.35	68.48	Central Asia	TJK	No	Light Allele	Heavy Allele	Light Allele	Mixed Allele	Mixed Allele
Kin-0	6926	20	23.3	0.7	44.46	-85.37	Midwest	USA	No	Light Allele	Mixed Allele	Light Allele	Light Allele	Heavy Allele
Ms-0	6938	20	21.8	0.8	55.75	37.63		RUS	Yes	Light Allele	Heavy Allele	Light Allele	Mixed Allele	Light Allele
Bur-0	6905	32	35.8	1.7	54.10	-6.20	Northern Europe	IRL	Yes	Heavy Allele	Heavy Allele	Light Allele	Mixed Allele	Mixed Allele
Edi-0	6914	24	26.8	0.5	55.95	-3.16	Northern Europe	UK	Yes	Light Allele	Heavy Allele	Heavy Allele	Light Allele	Heavy Allele
Oy-0	6946	24	24.5	1.4	60.23	6.13	Northern Europe	NOR	Yes	Light Allele	Heavy Allele	Light Allele	Light Allele	Light Allele
Ws-2	6981	16	16.9	0.6	52.30	30.00	Eastern Europe	RUS	Yes	Light Allele	Light Allele	Light Allele	Mixed Allele	Light Allele
Duk	6008	24	24.0	0.6	49.10	16.20	Moravia	CZE	Yes	Light Allele	Heavy Allele	Light Allele	Mixed Allele	Light Allele
Eds-1	6016	32	33.8	0.7	62.90	18.40	N Sweden	SWE	Yes	Heavy Allele	Heavy Allele	Heavy Allele	Heavy Allele	Heavy Allele
Hovdala-2	6039	20	22.4	0.8	56.10	13.74	S Sweden	SWE	Yes	Light Allele	Heavy Allele	Light Allele	Light Allele	Heavy Allele
Kni-1	6040	20	21.9	0.2	55.66	13.40	S Sweden	SWE	Yes	Light Allele	Mixed Allele	Mixed Allele	Light Allele	Light Allele
Lom 1-1	6042	20	20.2	0.4	56.09	13.90	S Sweden	SWE	Yes	Light Allele	Light Allele	Light Allele	Mixed Allele	Heavy Allele
Nyl-2	6064	28	30.9	0.4	62.95	18.28	N Sweden	SWE	Yes	Heavy Allele	Heavy Allele	Heavy Allele	Heavy Allele	Heavy Allele
Or-1	6074	20	21.6	0.6	56.46	16.14	S Sweden	SWE	Yes	Light Allele	Heavy Allele	Mixed Allele	Light Allele	Light Allele
Tottarp-2	6243	16	18.0	1.1	55.95	13.85	S Sweden	SWE	Yes	Light Allele	Light Allele	Light Allele	Mixed Allele	Light Allele
Lis-2	8222	16	19.3	2.5	56.03	14.78	S Sweden	SWE	Yes	Light Allele	Light Allele	Light Allele	Mixed Allele	Mixed Allele
Algutsrum	8230	20	23.6	0.5	56.68	16.50	S Sweden	SWE	Yes	Light Allele	Mixed Allele	Mixed Allele	Mixed Allele	Light Allele

Table 3. Continued

Name	ecotype_id	MPS Class	Mass per Seed (ug)	±SD	Latitude	Longitude	Region	Country	Euro.Line	Alleles@Locus1	Alleles@Locus2	Alleles@Locus3	Alleles@Locus4	Alleles@Locus5
Bro-1-6	8231	20	22.4	1.1	56.30	16.00	S Sweden	SWE	Yes	Light Allele	Light Allele	Light Allele	Mixed Allele	Light Allele
Dem-4	8233	24	27.2	0.6	41.19	-87.19	Midwest	USA	No	Heavy Allele	Mixed Allele	Heavy Allele	Heavy Allele	Heavy Allele
Gul1-2	8234	20	22.5	0.5										
Hod	8235	16	16.6	0.2	48.80	17.10	Moravia	CZE	Yes	Light Allele	Light Allele	Light Allele	Light Allele	Light Allele
HSm	8236	20	22.3	0.9	49.33	15.76	Moravia	CZE	Yes	Light Allele	Mixed Allele	Mixed Allele	Mixed Allele	Mixed Allele
Kavlinge-1	8237	20	20.0	0.6	55.80	13.10	S Sweden	SWE	Yes	Light Allele	Light Allele	Light Allele	Light Allele	Heavy Allele
Kent	8238	20	20.0	1.2	51.15	0.40		UK	Yes	Light Allele	Light Allele	Light Allele	Mixed Allele	Light Allele
Koln	8239	16	19.6	2.9	51.00	7.00		GER	Yes	Light Allele	Light Allele	Mixed Allele	Mixed Allele	Heavy Allele
Kulturen-1	8240	24	24.8	0.3	55.71	13.20	S Sweden	SWE	Yes	Light Allele	Mixed Allele	Mixed Allele	Light Allele	Heavy Allele
Liarum	8241	16	18.2	0.7	55.95	13.82	S Sweden	SWE	Yes	Light Allele	Light Allele	Light Allele	Light Allele	Light Allele
Lillo-1	8242	24	25.1	1.7	56.15	15.79	S Sweden	SWE	Yes	Light Allele	Heavy Allele	Light Allele	Mixed Allele	Light Allele
PHW-2	8243	20	23.7	0.2	43.77	11.25		ITA	Yes	Light Allele	Mixed Allele	Heavy Allele	Mixed Allele	Heavy Allele
PHW-34	8244	24	24.2	0.5										
Seattle-0	8245	16	18.8	0.2	47.00	-122.20	Pacific Northwest	USA	No	Light Allele	Mixed Allele	Light Allele	Light Allele	Heavy Allele
NC-6	8246	20	22.4	0.8										
San-2	8247	24	24.8	0.2	56.07	13.74	S Sweden	SWE	Yes	Light Allele	Mixed Allele	Mixed Allele	Light Allele	Heavy Allele
Vimmerby	8249	24	24.1	1.3	57.70	15.80	S Sweden	SWE	Yes	Light Allele	Heavy Allele	Light Allele	Mixed Allele	Light Allele
Alc-0	8252	16	18.5	0.3										
Ang-0	8254	16	17.9	0.4	50.30	5.30	Western Europe	BEL	Yes	Light Allele	Light Allele	Light Allele	Light Allele	Light Allele
Ba-1-2	8256	16	19.6	0.7	56.40	12.90	S Sweden	SWE	Yes	Light Allele	Light Allele	Mixed Allele	Light Allele	Light Allele
Ba3-3	8257	16	18.3	1.0										
Ba-4-1	8258	20	21.7	0.4	56.40	12.90	S Sweden	SWE	Yes	Mixed Allele	Heavy Allele	Mixed Allele	Light Allele	Heavy Allele
Ba-5-1	8259	20	20.5	0.7	56.40	12.90	S Sweden	SWE	Yes	Mixed Allele	Heavy Allele	Mixed Allele	Light Allele	Heavy Allele
Bg-2	8261	28	28.0	0.9										
Bla-1	8264	24	26.4	0.9	41.68	2.80	Southern Europe	ESP	Yes	Light Allele	Mixed Allele	Mixed Allele	Mixed Allele	Mixed Allele
Blh-1	8265	16	19.6	0.4	48.00	19.00	Eastern Europe	CZE	Yes	Light Allele	Mixed Allele	Light Allele	Light Allele	Light Allele
Bs-1	8270	20	21.8	0.2	47.50	7.50	Western Europe	SUI	Yes	Light Allele	Mixed Allele	Light Allele	Light Allele	Light Allele
Bu-0	8271	32	33.2	0.0	50.50	9.50	Western Europe	GER	Yes	Light Allele	Light Allele	Mixed Allele	Mixed Allele	Light Allele
Can-0	8274	16	17.7	1.0	29.21	-13.48	Southern Europe	ESP	Yes	Light Allele	Mixed Allele	Light Allele	Light Allele	Light Allele
Cen-0	8275	20	20.2	0.6	49.00	0.50	Western Europe	FRA	Yes	Light Allele	Light Allele	Light Allele	Light Allele	Heavy Allele
Co	8278	32	37.9	0.4										
Dra-3-1	8283	24	24.5	0.6	55.76	14.12	S Sweden	SWE	Yes	Light Allele	Light Allele	Light Allele	Light Allele	Light Allele
DraII-1	8284	20	23.5	0.6	49.41	16.28	Moravia	CZE	Yes	Light Allele	Heavy Allele	Light Allele	Light Allele	Light Allele
DraIII-1	8285	16	19.7	0.4	49.41	16.28	Moravia	CZE	Yes	Light Allele	Mixed Allele	Mixed Allele	Light Allele	Light Allele
En-1	8290	20	23.6	1.0	50.00	8.50	Western Europe	GER	Yes	Light Allele	Light Allele	Light Allele	Light Allele	Light Allele
Gd-1	8296	20	23.0	0.7	53.50	10.50	Western Europe	GER	Yes	Light Allele	Light Allele	Mixed Allele	Mixed Allele	Mixed Allele
Ge-0	8297	28	30.1	1.4	46.50	6.08	Western Europe	SUI	Yes	Light Allele	Heavy Allele	Mixed Allele	Light Allele	Heavy Allele
Gr-1	8300	16	18.1	0.6	47.00	15.50		AUT	Yes	Light Allele	Light Allele	Light Allele	Light Allele	Light Allele
H55	8303	16	19.8	0.2										
Hi-0	8304	24	24.3	0.2	52.00	5.00	Western Europe	NED	Yes	Light Allele	Heavy Allele	Light Allele	Light Allele	Light Allele
Hov-4-1	8306	20	21.9	1.0	56.10	13.74	S Sweden	SWE	Yes	Light Allele	Heavy Allele	Mixed Allele	Light Allele	Heavy Allele
Hs-0	8310	24	25.5	0.3	52.24	9.44	Western Europe	GER	Yes	Heavy Allele	Heavy Allele	Mixed Allele	Mixed Allele	Mixed Allele
In-0	8311	20	22.5	0.6	47.50	11.50	Western Europe	AUT	Yes	Light Allele	Mixed Allele	Light Allele	Mixed Allele	Light Allele
Is-0	8312	24	24.9	0.5	50.50	7.50	Western Europe	GER	Yes	Light Allele	Heavy Allele	Mixed Allele	Light Allele	Light Allele
Jm-0	8313	16	19.7	0.3	49.00	15.00	Eastern Europe	CZE	Yes	Light Allele	Heavy Allele	Light Allele	Light Allele	Light Allele
Ka-0	8314	20	22.7	0.3	47.00	14.00	Western Europe	AUT	Yes	Light Allele	Light Allele	Light Allele	Mixed Allele	Light Allele
Kz-13	8321	20	22.7	0.5										
Lc-0	8323	16	16.4	0.3	57.00	-4.00		UK	Yes	Light Allele	Mixed Allele	Light Allele	Mixed Allele	Light Allele
Lip-0	8325	20	23.4	0.2	50.00	19.30		POL	Yes	Light Allele	Light Allele	Light Allele	Mixed Allele	Light Allele
Lis-1	8326	20	20.2	0.6	56.03	14.78	S Sweden	SWE	Yes	Light Allele	Light Allele	Light Allele	Light Allele	Light Allele
Lm-2	8329	20	23.2	0.6	48.00	0.50	Western Europe	FRA	Yes	Light Allele	Mixed Allele	Light Allele	Mixed Allele	Heavy Allele
Lu-1	8334	20	20.1	0.6	55.71	13.20	S Sweden	SWE	Yes	Light Allele	Heavy Allele	Light Allele	Light Allele	Mixed Allele
Lund	8335	20	23.5	0.9	55.71	13.20	S Sweden	SWE	Yes	Light Allele	Mixed Allele	Mixed Allele	Light Allele	Heavy Allele
Mir-0	8337	16	19.8	0.4	44.00	12.37		ITA	Yes	Light Allele	Light Allele	Light Allele	Mixed Allele	Heavy Allele
Na-1	8343	24	24.4	0.9	47.50	1.50	Western Europe	FRA	Yes	Light Allele	Heavy Allele	Light Allele	Light Allele	Light Allele
NW-0	8348	20	22.7	0.7	50.50	8.50		GER	Yes	Light Allele	Mixed Allele	Light Allele	Light Allele	Heavy Allele
Ost-0	8351	24	26.1	1.1	60.25	18.37	C Sweden	SWE	Yes	Light Allele	Heavy Allele	Mixed Allele	Heavy Allele	Heavy Allele
Pa-1	8353	16	17.2	0.8	38.07	13.22	Sicily	ITA	Yes	Light Allele	Light Allele	Light Allele	Mixed Allele	Light Allele
Per-1	8354	20	23.1	1.4	58.00	56.32		RUS	No	Light Allele	Light Allele	Light Allele	Light Allele	Mixed Allele
Petergof	8355	20	20.6	1.0										
Pi-0	8356	28	28.6	0.5										
Pla-0	8357	16	19.7	0.2	41.50	2.25	Southern Europe	ESP	Yes	Light Allele	Light Allele	Light Allele	Light Allele	Heavy Allele
Pu2-8	8363	20	22.4	0.7										
Rak-2	8365	20	20.8	0.6	49.00	16.00		CZE	Yes	Light Allele	Light Allele	Light Allele	Light Allele	Light Allele
Rd-0	8366	24	25.3	0.7	50.50	8.50		GER	Yes	Light Allele	Heavy Allele	Light Allele	Mixed Allele	Heavy Allele
Rev-1	8369	16	19.8	0.7	55.69	13.45	S Sweden	SWE	Yes	Light Allele	Mixed Allele	Mixed Allele	Light Allele	Light Allele
Rsch-4	8374	16	17.8	0.8	56.30	34.00		RUS	Yes	Light Allele	Light Allele	Mixed Allele	Mixed Allele	Light Allele
Rubnoe-1	8375	16	17.8	0.5										
Sanna-2	8376	28	29.5	0.7	62.69	18.00	N Sweden	SWE	Yes	Heavy Allele	Heavy Allele	Mixed Allele	Heavy Allele	Heavy Allele
SantaClara	8377	20	20.4	0.3										
Sap-0	8378	16	19.2	0.5	49.49	14.24	Eastern Europe	CZE	Yes	Light Allele	Heavy Allele	Light Allele	Light Allele	Light Allele
Sf-1	8380	24	24.6	0.4										
Sr-5	8386	16	19.2	0.3	58.90	11.20	C Sweden	SWE	Yes	Light Allele	Light Allele	Light Allele	Mixed Allele	Light Allele
St-0	8387	20	20.2	0.4	59.00	18.00	C Sweden	SWE	Yes	Light Allele	Heavy Allele	Mixed Allele	Light Allele	Heavy Allele
Stw-0	8388	20	20.6	0.7	52.00	36.00	Eastern Europe	RUS	Yes	Light Allele	Light Allele	Light Allele	Light Allele	Light Allele
Ta-0	8389	20	23.8	0.5	49.50	14.50	Eastern Europe	CZE	Yes	Light Allele	Heavy Allele	Mixed Allele	Light Allele	Light Allele
Tu-0	8395	24	26.6	0.8	45.00	7.50		ITA	Yes	Light Allele	Mixed Allele	Light Allele	Mixed Allele	Mixed Allele
Rd-0	8411	20	22.3	0.7	50.50	8.50		GER	Yes	Light Allele	Heavy Allele	Light Allele	Mixed Allele	Heavy Allele
Sav-0	8412	20	23.9	1.0	49.18	15.88		CZE	Yes	Light Allele	Light Allele	Mixed Allele	Mixed Allele	Light Allele
Wil-1	8419	16	16.8	0.3										
Kelsterbach-4	8420	16	19.8	0.3	50.07	8.53	Western Europe	GER	Yes	Light Allele	Light Allele	Light Allele	Mixed Allele	Heavy Allele
Fja-1-1	8422	24	24.1	0.1	56.06	14.29	S Sweden	SWE	Yes	Heavy Allele	Mixed Allele	Mixed Allele	Light Allele	Heavy Allele
Hov-2-1	8423	20	20.0	0.4	56.10	13.74	S Sweden	SWE	Yes	Light Allele	Light Allele	Light Allele	Light Allele	Light Allele
Ull-1-1	8426	20	22.5	0.9	56.06	13.97	S Sweden	SWE	Yes	Light Allele	Heavy Allele	Light Allele	Mixed Allele	Light Allele
Uod-2	8428	16	19.5	0.7	48.30	14.45	Western Europe	AUT	Yes	Light Allele	Light Allele	Light Allele	Mixed Allele	Light Allele
Lisse	8430	28	28.5	0.6	52.25	4.57		NED	Yes	Heavy Allele	Mixed Allele	Light Allele	Light Allele	Heavy Allele
Vinslov	9057	20	20.0	0.2	56.10	13.92	S Sweden	SWE	Yes	Light Allele	Light Allele	Light Allele	Mixed Allele	Heavy Allele
Vastervik	9058	20	23.0	0.4	57.75	16.63		SWE	Yes	Heavy Allele	Light Allele	Light Allele	Light Allele	Light Allele

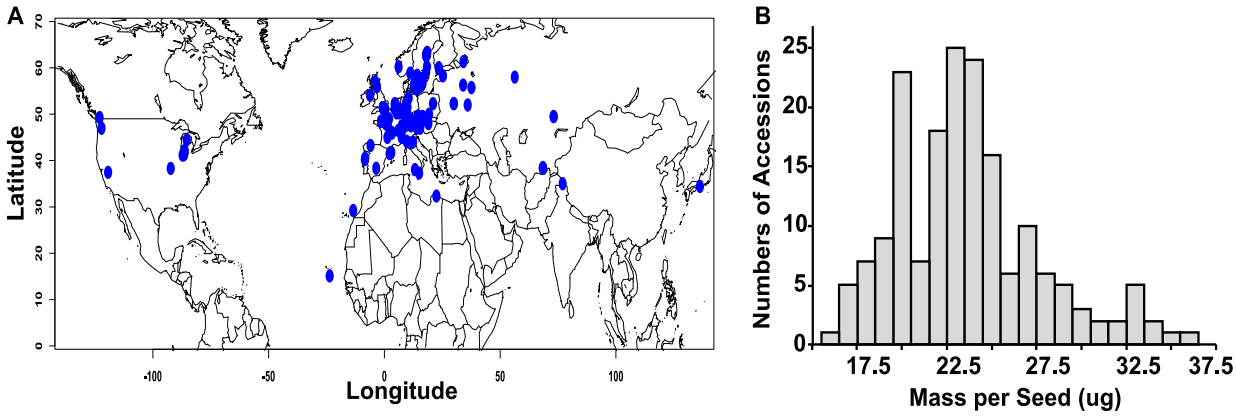


Figure 2. Seed mass pattern of globally collected *A.thaliana* wild

(A) Geographic map showing the collected sites of the mapped accessions. **(B)** Histograms of the mass per seed.

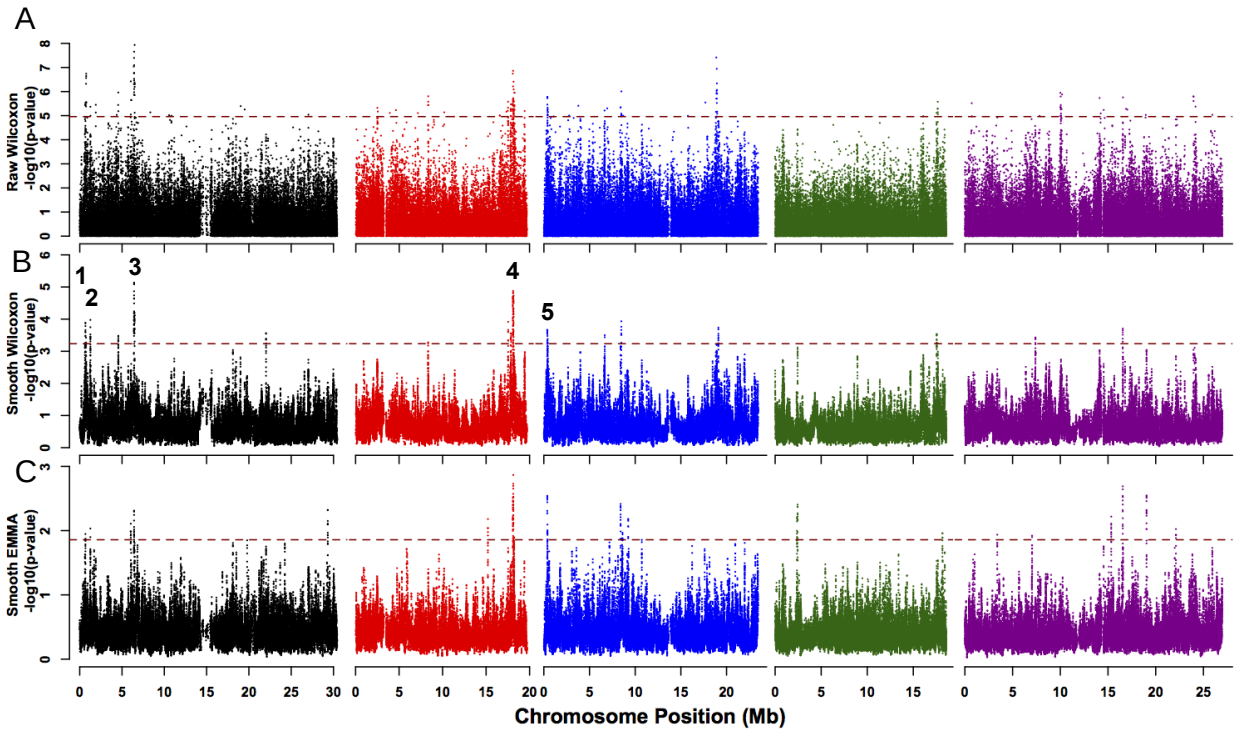


Figure 3. Genome-wide association map on mass per seed

(A) Raw Wilcoxon map. (**B)** Smoothed Wilcoxon map. **(C)** Smoothed EMMA map. Number 1-5 represent the positions of the five loci that survived after EMMA correction.

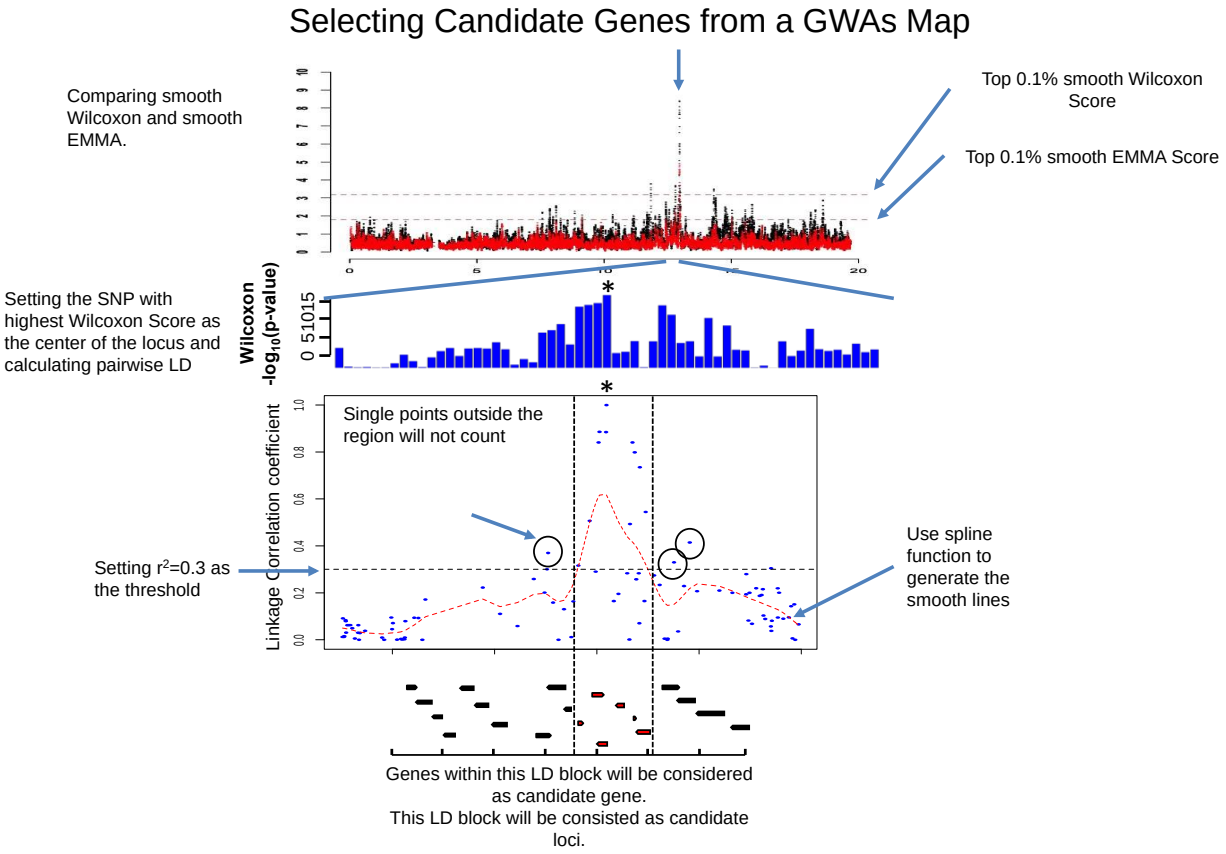


Figure 4. Diagram of the procedure to select candidate genes from GWAs analysis.

The asterisk represents the central SNP.

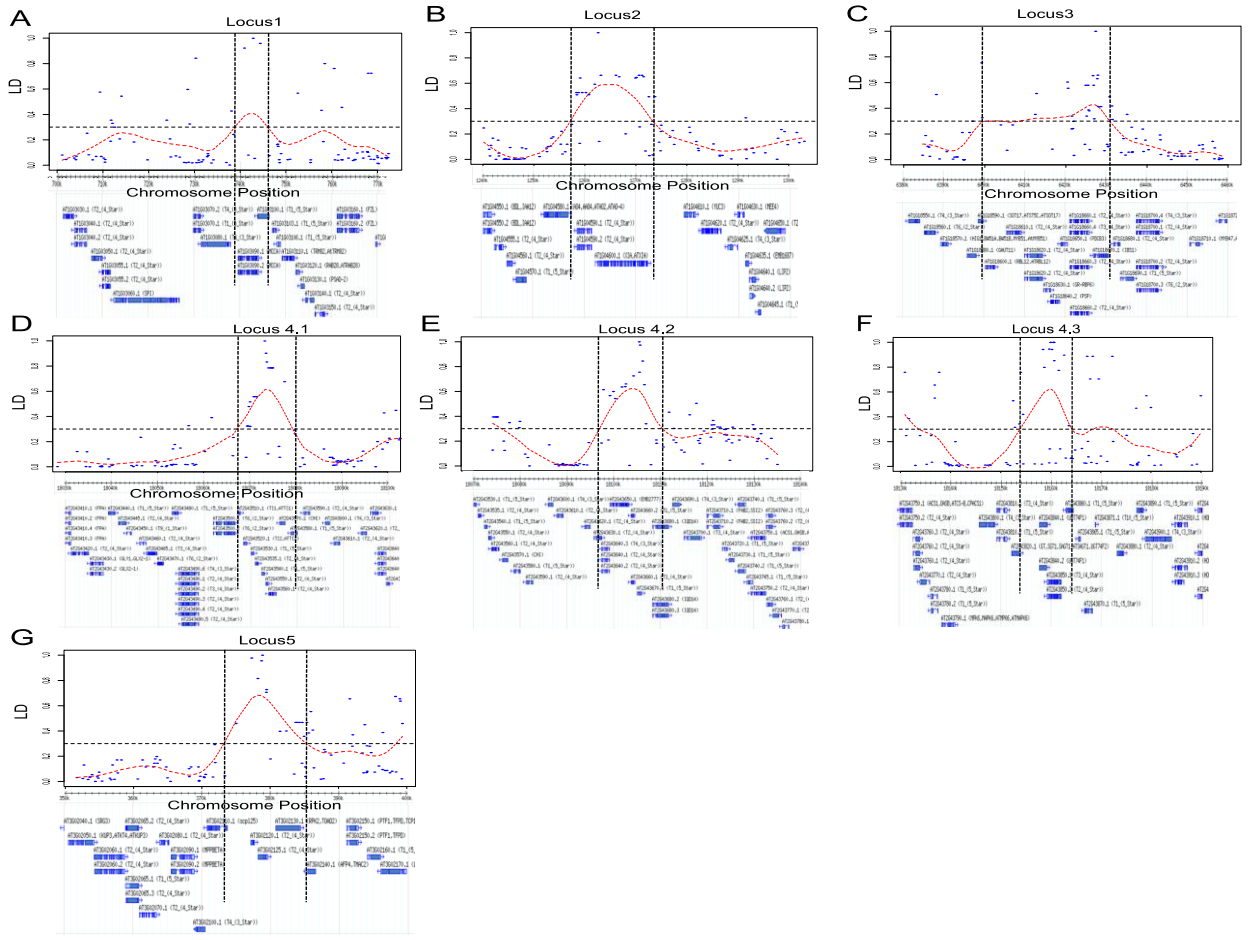


Figure 5. Local linkage disequilibrium plot showing the width of each locus and the genes within it.

(A)-(G) The LD values of +/- 50 SNPs around the central SNP are plotted according to their base pair positions. Red dash line represents the fitted LD curve. Black dash line shows the threshold (0.3) and the borders of each locus. Gene information is screen-printed from TAIR10 GBrowse (http://gbrowse.arabidopsis.org/cgi-bin/gbrowse/arabidopsis/).

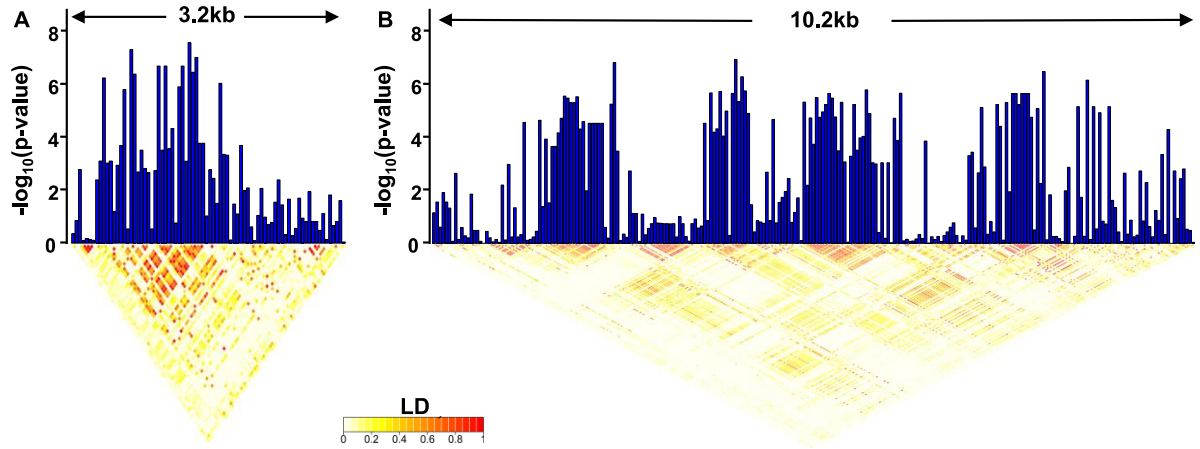


Figure 6. Fine map of two mass per seed candidate loci.

(A) Locus #3. **(B)** Locus #4. Blue bars represent Wilcoxon GWAs score. Triangle heatmap show linkage disequilibrium in two loci. Red color represents tightly linked and light yellow represents no linkage. The size of two loci on chromosomes are annotated at the top.

Table 4. Description of the forty candidate genes for mass per seed GWAs map.

Annotation based on TAIR10.

Locus	Gene	Name	Description
1	AT1G03090	MCCA	Biotinylated subunit of the dimer MCCase, which is involved in leucine degradation.
1	AT1G03100	Unk	Pentatricopeptide repeat (PPR) superfamily protein.
2	AT1G04580	AAO4	Aldehyde Oxidase
2	AT1G04590	Unk	Pentatricopeptide repeat (PPR) superfamily protein.
2	AT1G04600	XIA	MYOSIN XI A. Member of Myosin-like proteins.
3	AT1G18590	SOT17	SULFOTRANSFERASE 17. desulfoglucosinolate sulfotransferase
3	AT1G18600	RBL12	Proteolysis / endopeptidase
3	AT1G18610	Unk	Galactose oxidase/kelch repeat superfamily protein
3	AT1G18620	TRM3	Unknown
3	AT1G18630	RBP6	glycine-rich RNA binding protein.
3	AT1G18640	PSP	3-phosphoserine phosphatase in serine biosynthesis
3	AT1G18650	PDCB3	PLASMODESMATA CALLOSE-BINDING PROTEIN3
3	AT1G18660	Unk	Zinc finger (C3HC4-type RING finger) family protein.
3	AT1G18670	IBS1	IMPAIRED IN BABA-INDUCED STERILITY 1
4	AT2G43500	RWP-RK	positive regulation of transcription
4	AT2G43510	TI1	trypsin inhibitor
4	AT2G43520	TI2	trypsin inhibitor
4	AT2G43530	Unk	starch biosynthesis
4	AT2G43535	Unk	starch biosynthesis
4	AT2G43540	Unk	myristoylation
4	AT2G43550	Unk	ion channel inhibitor
4	AT2G43560	Unk	FKBP-like peptidyl-prolyl cis-trans isomerase family protein.
4	AT2G43570	CHI	Chitinase
4	AT2G43580	Unk	Chitinase
5	AT2G43630	Unk	actin pollen tube development
5	AT2G43640	Unk	Signal recognition particle, SRP9/SRP14 subunit.
5	AT2G43650	EMB2777	EMBRYO DEFECTIVE 2777.
5	AT2G43660	Unk	Carbohydrate-binding X8 domain superfamily protein.
5	AT2G43670	Unk	Carbohydrate-binding X8 domain superfamily protein.
5	AT2G43680	IQD14	IQ-DOMAIN 14.
6	AT2G43820	SAGT1	Induced by Salicylic acid. tryptophan synthesis
6	AT2G43830	Unk	Pseudogene, NPK1-related protein kinase 3
6	AT2G43840	UGT74F1	UDP-GLYCOSYLTRANSFERASE 74
6	AT2G43850	Unk	Ankyrin repeat
6	AT2G43860	Unk	Pectin lyase-like superfamily protein
7	AT3G02110	SCPL25	serine carboxypeptidase-like 25
7	AT3G02120	Unk	hydroxyproline-rich glycoprotein family protein.
7	AT3G02125	Unk	Unknown
7	AT3G02130	RPK2	RECEPTOR-LIKE PROTEIN KINASE 2. A regulator of meristem maintenance
7	AT3G02140	AFP4	ABI FIVE BINDING PROTEIN 4 Negative regulator of ABA

2.3.2 Grouping alleles at candidate loci

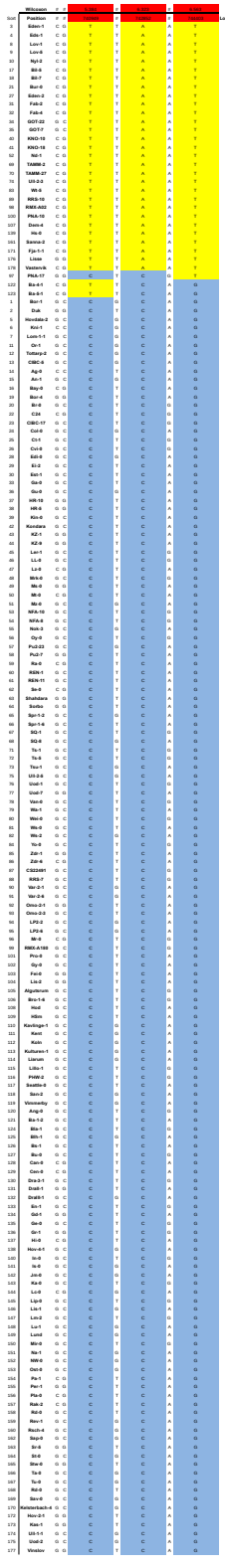
One of the major advantages of GWA mapping is that it allows researchers to identify alleles at candidate loci for future characterization of allelic effects (Todesco *et al.*, 2010). To identify alleles associated with seed mass, I chose SNPs with the top 0.1% GWAs score calculated by Wilcoxon method and coded the SNPs associated with heavy or light seeds with yellow or blue color, respectively. I then sorted the whole SNPs set by colors in the order of their GWAs score. SNPs at all sorted positions with yellow color were then grouped and considered as alleles associated with heavy seeds ("Heavy allele"). At the same time, SNPs with all blue color were grouped and considered as alleles associated with light seeds ("Light allele"). The SNPs at these positions with mixed colors were grouped and called as "Mixed alleles" (Table 5). My GWA mapping pinpointed five candidate loci that might have genetic variants for explaining the observed differences in mass per seed (Table 1, Figure 3). These results clearly support the idea that seed mass is a very complex trait like other plant fitness-related traits, which are genetically governed by different loci (Bergelson and Roux, 2010).

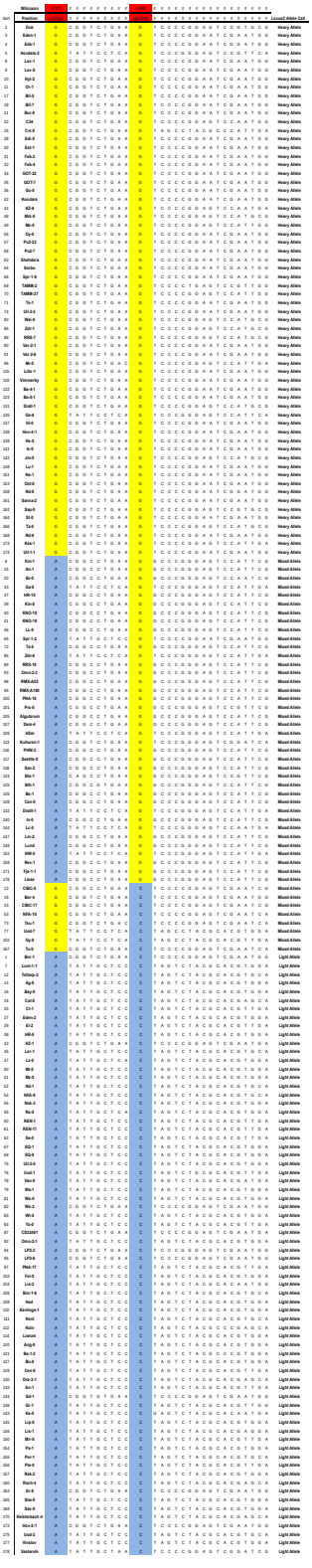
After grouping alleles at all five different loci, I found that all of the heavy alleles were significantly associated with heavy mass per seed at each of the five loci (Figure 6). Surprisingly, although the locus 4 covered over 10 kb on chromosome 2 that was about three to five times wider than other loci, there were still 16 accessions sharing identical SNPs at all 31 positions across this entire locus (Figure 6B, Table 5). This tight linkage disequilibrium might indicate the importance of the genetic material in this locus for plant fitness. All five loci, together, could explain 43.28 % observed variation in mass per seed. Among the 178 wild accessions, only four maintained heavy alleles at all five loci (Eds-1, Fab-4, Lov-1, and Nyl-2)

and ten maintained heavy alleles at four out of five loci. On the other hand, 17 accessions maintained light alleles at all five loci and 38 had light alleles at four out of five loci (Table 5). This contrast might indicate the limitation of producing heavy seeds and suggest the existence of cost-benefit trade-offs at these genetic loci. Indeed, the trade-off between seed mass and seed number has already been reported previously (Alonso Blanco *et al.*, 1999).

Table 5. SNPs and alleles at the five candidate loci for mass per seed GWAs map.

Locus1 Locus2 Locus3





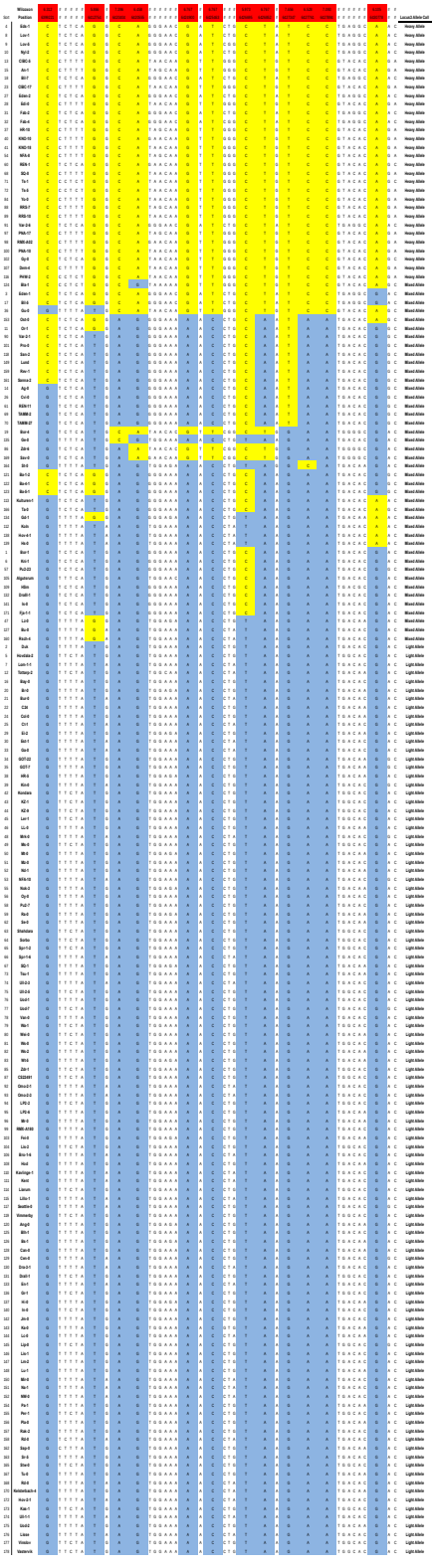
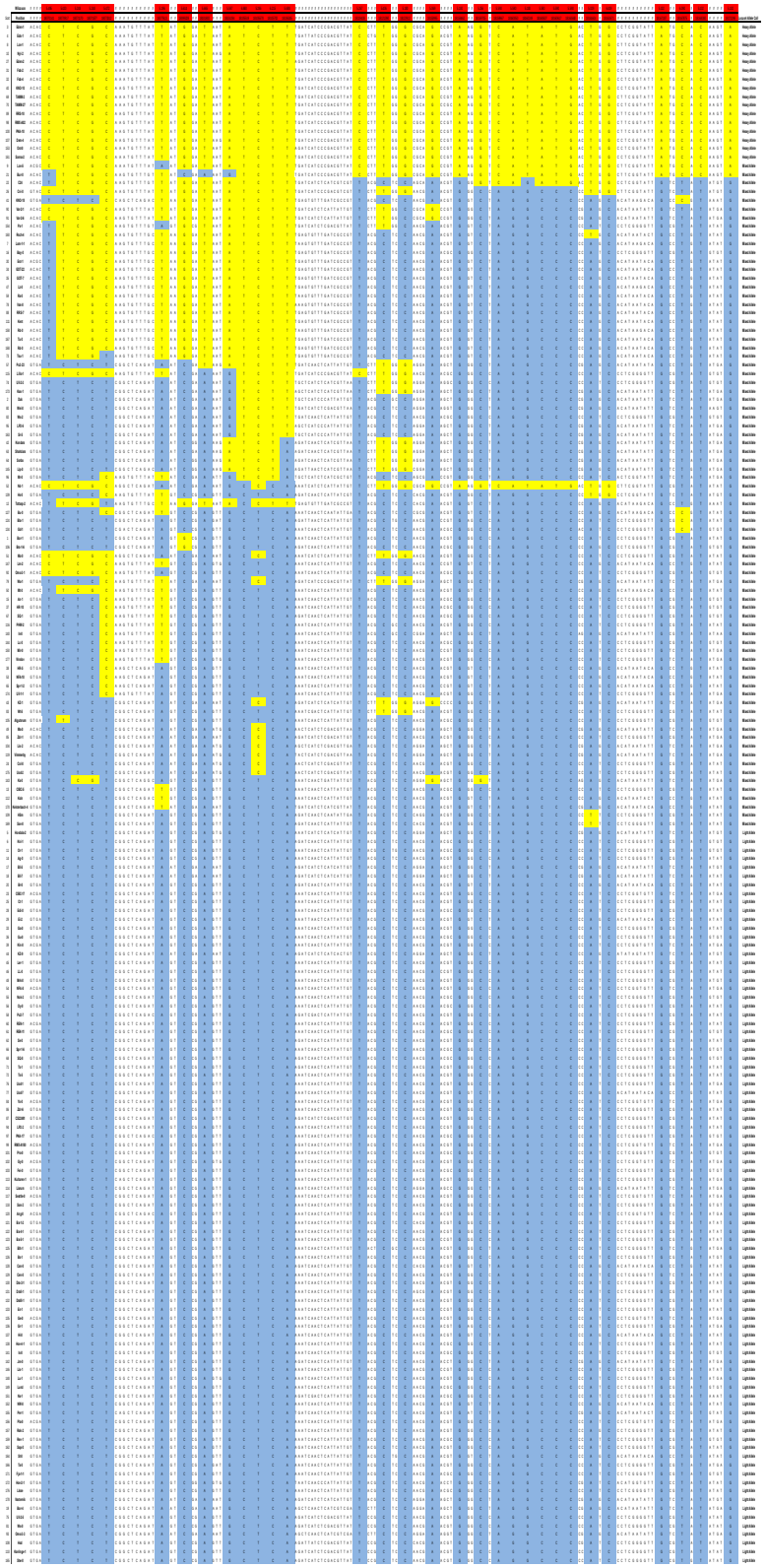
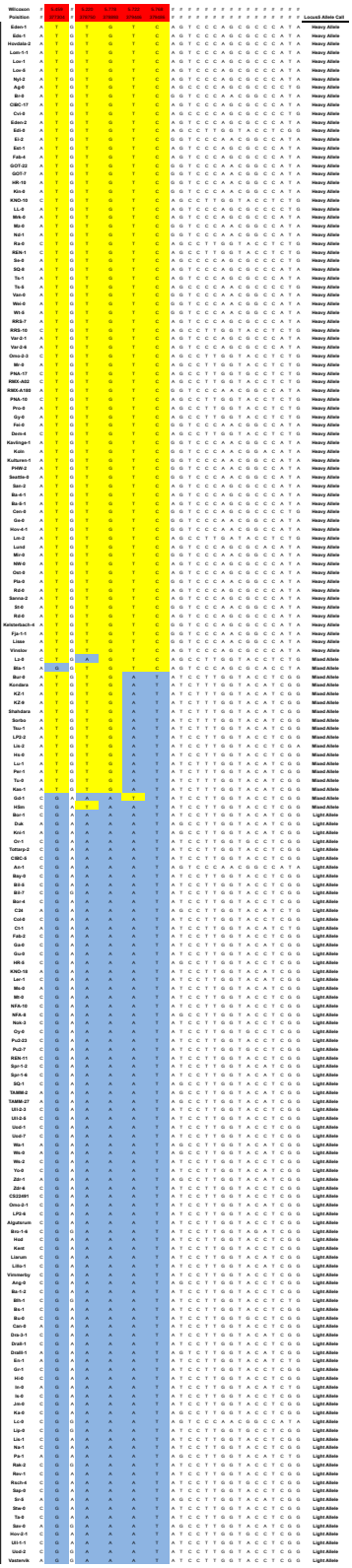


Table 5. Continued

Locus4



Locus5



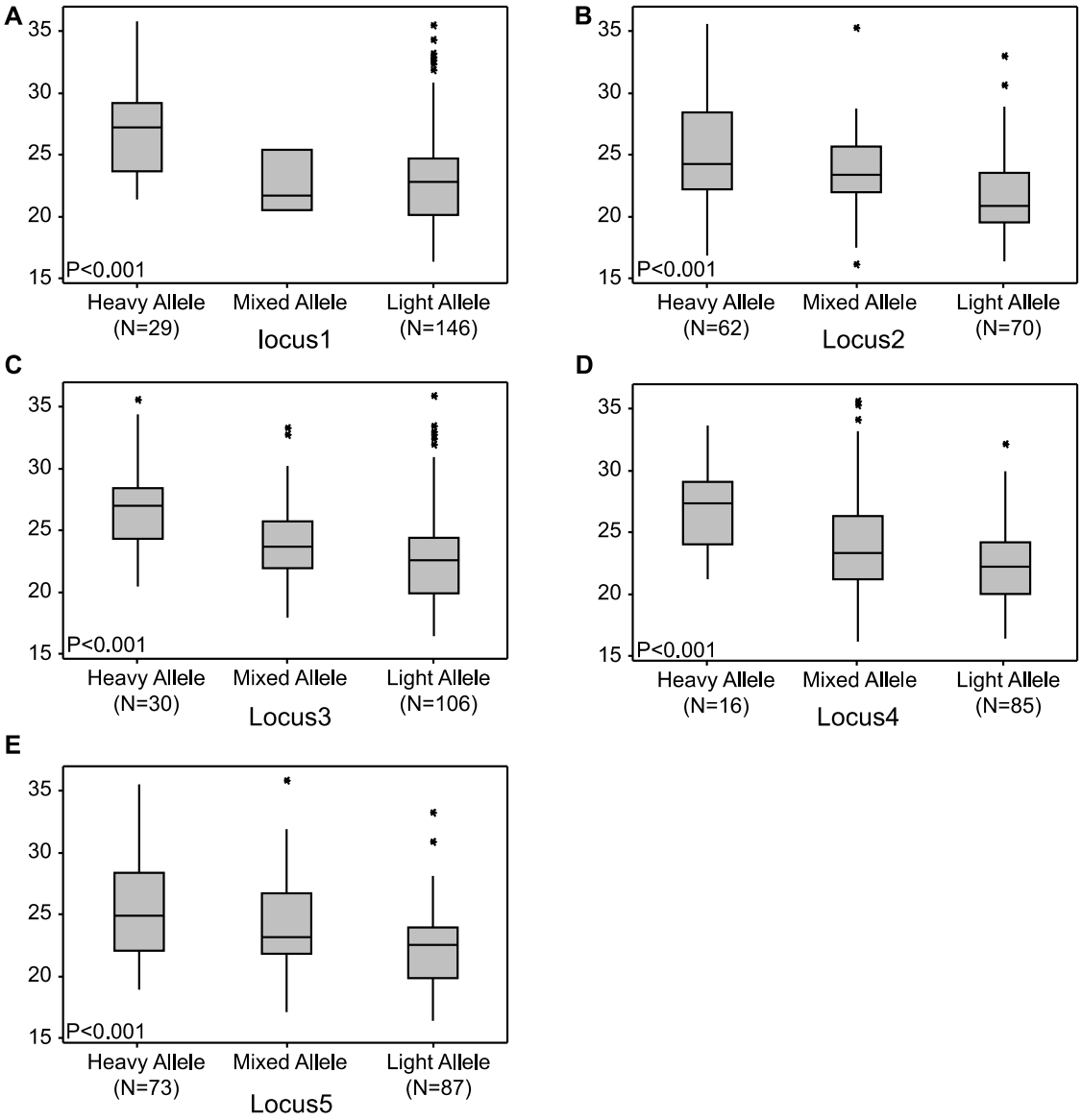


Figure 7. Boxplots of the mass per seed for each GWA locus genotype.

(A)-(E). The y-axis is mass per seed (μg). For each of the five loci, alleles were assigned to 178 wild *A.thaliana* accessions based on their SNPs information (Table3 and Table 5). The numbers of accessions with alleles at each of the five loci were listed at the bottom of each plot. p-value on each plot was acquired by ANOVA.

2.3.3 Geographic analysis of seed mass alleles

Environment has a strong effect on plants, particularly on fitness-related traits. However, the genetic basis is not well understood (Atwell *et al.*, 2010; Fournier-Level *et al.*, 2011). To address geographic patterns, I focused on 154 natural *Arabidopsis* accessions from Europe and analyzed the effect of geographic distribution on mass per seed and the five loci identified from the GWA map. Phenotypically, the fitted 3D surface based on polynomial regression showed that the mass per seed increased when latitude went high and longitude went small (Figure 8A). When I projected this surface to the geographic map, the contour map showed that the higher predicted values displaced at the top left part of the map. Indeed, the 3D plot of raw mass per seed against longitude and latitude showed that the accessions with higher mass per seed are more frequent in the north and northwest part in Europe (Figure 8B). Interestingly, I found no significant correlation between mass per seed and longitude or latitude, individually (Figure 10A-B, p=0.054 & 0.471, respectively).

However, when having longitude and latitude together in the model, I saw both terms had significant correlation with mass per seed (Figure 9, MPS = $0.13^{*} \times$ Latitude - $0.12^{**} \times$ Longitude + 18.34^{***}. *Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05*). I also tested the interaction and saw no significant pattern. Genetically, at the allelic level, accessions with heavy alleles at four out of five loci had a higher latitude compared with those with the light alleles (Figure 10C-G). For example, at locus two, 55 accessions had a heavy allele, of which the mean of latitude was 54.65 (Figure 10D). Meanwhile, the mean of latitude in 63 accessions with light alleles was 50.91, which was significantly lower ($p<0.001$). Interestingly, heavy alleles at locus 2 tended to cluster at the northwest and the light alleles are more frequently in

the southeast of Europe (Figure 8C). Only locus 5 from the top of chromosome 3 had no difference of latitude between two types of alleles (Figure 10G).

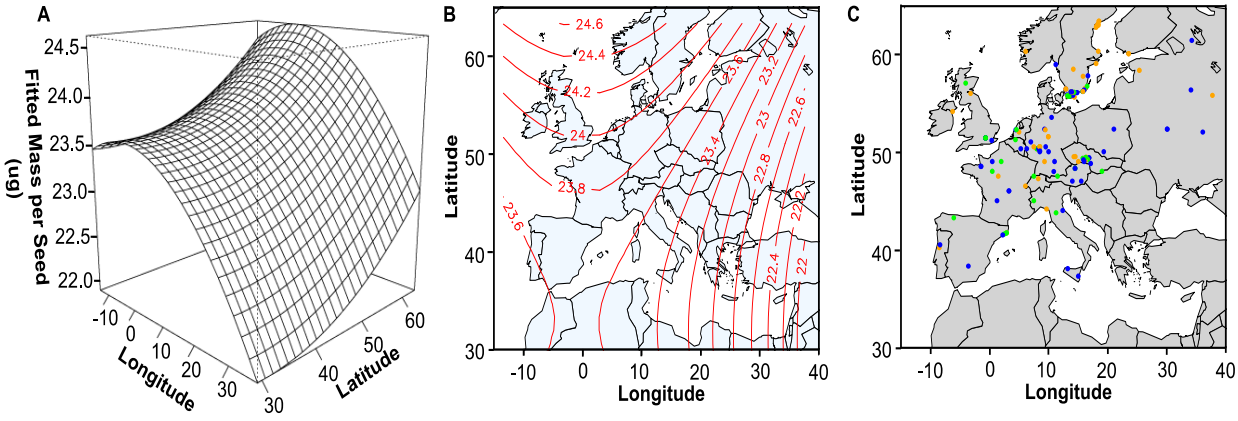


Figure 8. Geographic pattern of mass per seed in 178 wild *A. thaliana* accessions.

(A) 3D plot of fitted surface of mass per seed (z-axis) against longitude (x-axis) and latitude (y-axis) in Europe. **(B)** Contour plot of predicted mass per seed based on polynomial regression. **(C)** Geographic distribution of alleles at locus 2. Orange, green and blue colors represent heavy, mixed and light alleles, respectively.

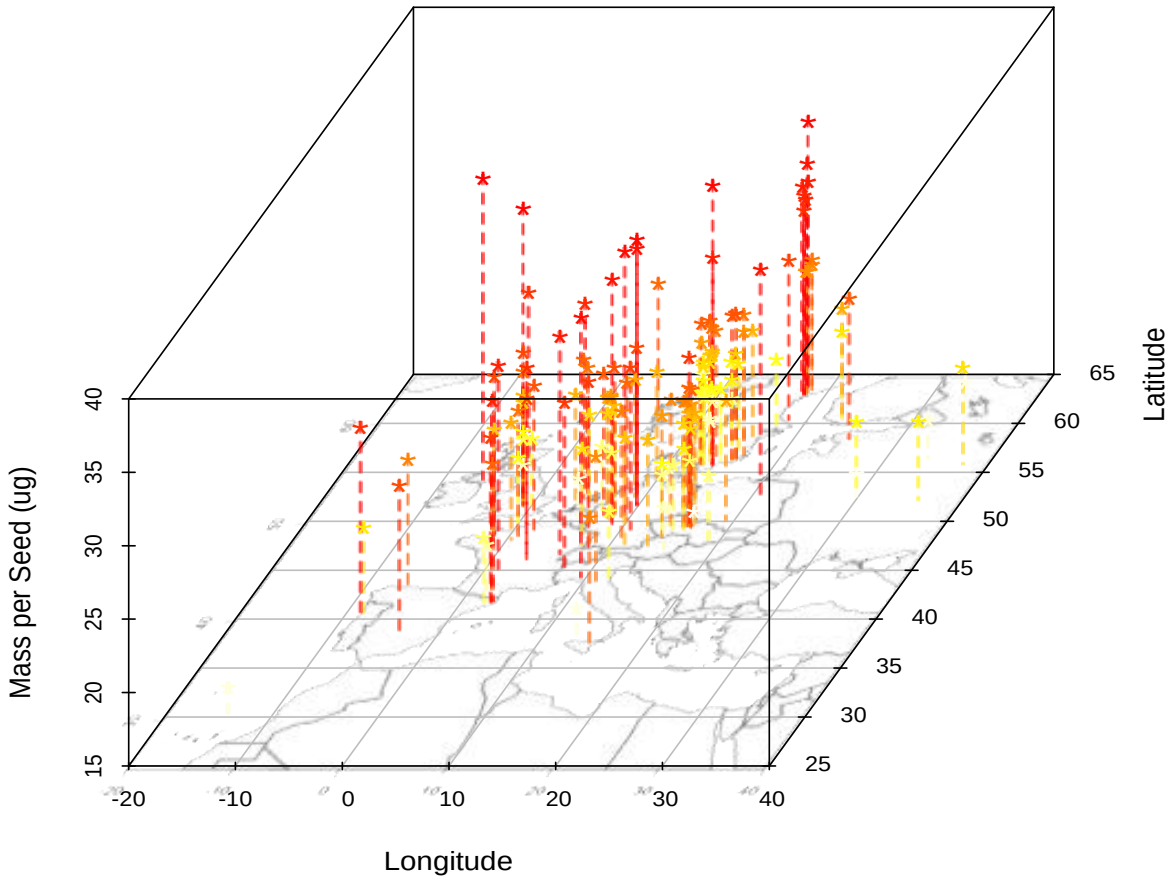


Figure 9. 3D plot of mass per seed against longitude and latitude.

The height of the dashed line represents the mass per seed of each accession. Asterisks represent the locations of each line, which are based on its geographic coordinates. Lines and asterisks are also color-coded by their value. Yellow represents lower mass per seed. Red represents higher mass per seed. I observed a significant linear regression model: MPS = $0.13^{*} \times$ Latitude - $0.12^{**} \times$ Longitude + 18.34^{***}

(*Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05*).

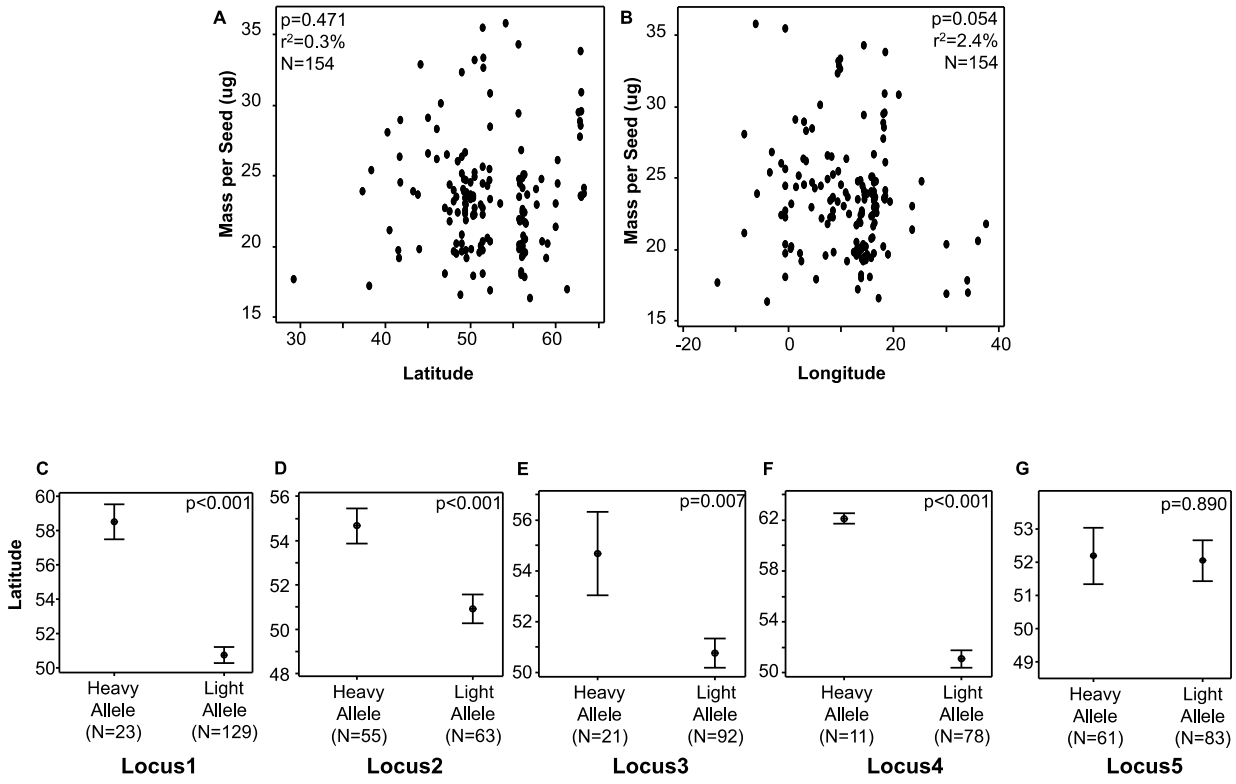


Figure 10. Geographic analysis of MPS alleles.

(A)-(B) Scatter plot of mass per seed against latitude and longitude among the 154 wild *A.thaliana* accessions in Europe. **(C)-(G)** Interval plot of latitude against each of the five loci. p-values are calculated from ANOVA. Numbers of accessions with each allele are labeled at the bottom.

2.3.4 Candidate genes from GWA map on 178 wild Arabidopsis accessions

In my GWA analysis, forty genes from five candidate loci were identified (Table 4, Table 6), including 14 genes from the three loci at the top of chromosome 1 and 21 gene from the wide locus at the bottom of chromosome 2. I took advantages of metadata to evaluate the current candidate genes. First, I analyzed the GO annotation for the candidate gene list (Table 4, Table 6). In the keywords related to general functions, the forty candidate genes had significant enrichment in "growth", "metabolism" and "signaling" (Figure 10A. Chi-sq=15.06, 5.43 and 5.00, respectively, D.f.=1). I didn't see significant enrichment in other general functions including "transport", "transcription", "protein synthesis" and "cell wall". For the keywords related to specific tissues, I found significant enrichment in "embryo" and "seed" categories (Figure 11B. Chi-sq=16.33 and 11.77, respectively, D.f.=1). These results supported the quality of my map and the candidate gene list I provided. Interestingly, I also noticed that my genes were significantly enriched in "leaf", "root" and "flower" (Figure 11B, Chi-sq= 13.83, 19.51 and 23.93, respectively). This might indicate that the seed mass is an important reproductive outcome related to plant growth condition in general. Thus the tissues being important for plant life activity were found by my map. Particularly, the keyword "flower" has the most hits with the highest statistical significance, which suggested the direct effect of flowering on seed production. I also searched the expression data of the forty genes in previously published microarray data (Le *et al.*, 2010). I focused on the mature green stage since it was the primary step of storage reserve accumulation (Le *et al.*, 2010). I searched the expression data in seed coat and embryo since they are the two major parts of the seed. My analysis revealed that 12 and 15 out of 40 genes had at least a 2-fold up-regulation in the embryo and seed coat, respectively (Figure 12. Average fold change of the 12 and 15 genes =

64.01 and 14.62, p-value=0.014 and 0.152, respectively). Among them, the candidate gene with highest up-regulation in embryos is AT2G43520 (*ATTI2*), which had a 661.8-fold change, ranking 32nd out of 22,591 genes of the microarray. It is predicted to be a trypsin inhibitor and might be involved in defense against herbivory. However, there is no genetic study of this gene. The candidate gene with the highest up-regulation in seed coat was AT1G04580 (*AAO4*), which had a 148.07-fold change (rank=67th). It encodes an aldehyde oxidase that converts benzaldhyde to benzoic acid (Ibdah *et al.*, 2009), and regulates the endogenous ABA level (Pandey *et al.*, 2010; Seo *et al.*, 2004) during seed development. Again, its function in regulating seed weight is unknown.

Table 6. Keyword enrichment for mass per seed map candidates.

Locus	Gene ID	Gene Name[1]	GWAs Wilcoxon[2]	Microarray (fold-Change[3])		General Function Keyword							Specific Tissue Keyword					
				Seed Coat	Embryo	Growth	Metabolism	Transport	Transcription	Protein Synth	Signaling	Cell Cycle	Embryo	Seed	Endosperm	Leaf	Root	Flower
1	AT1G03090	MCCA	6.3	0.5	2.4	✔	✔						✔	✔		✔	✔	✔
1	AT1G03100	Unk	6.5	1.3	1.0	✔							✔	✔		✔	✔	✔
2	AT1G04580	AAO4	5.3	148.1	0.0	✔	✔	✔										✔
2	AT1G04590	Unk	4.2	1.2	0.6	✔	✔					✔	✔	✔		✔	✔	✔
2	AT1G04600	XIA	5.0	0.2	1.0	✔							✔	✔		✔	✔	✔
3	AT1G18590	SOT17	6.3	4.0	2.0	✔	✔						✔	✔		✔	✔	✔
3	AT1G18600	RBL12		0.8	0.9	✔	✔						✔	✔		✔	✔	✔
3	AT1G18610	Unk	3.0	0.6	0.4	✔							✔	✔		✔	✔	✔
3	AT1G18620	TRM3	5.9	1.7	1.6	✔					✔		✔	✔		✔	✔	✔
3	AT1G18630	RBP6		1.1	0.5	✔							✔	✔		✔	✔	✔
3	AT1G18640	PSP	0.5	0.4	1.1	✔	✔						✔	✔		✔	✔	✔
3	AT1G18650	PDCB3	7.4	0.9	7.5	✔							✔	✔		✔	✔	✔
3	AT1G18660	Unk	6.8	1.3	2.0	✔	✔	✔					✔	✔		✔	✔	✔
3	AT1G18670	IBS1	7.7	0.8	2.3	✔	✔			✔			✔	✔		✔	✔	✔
4	AT2G43500	RWP-RK	3.9	3.2	5.9	✔	✔		✔				✔	✔		✔		✔
4	AT2G43510	TI1		6.1	0.3	✔							✔	✔		✔	✔	✔
4	AT2G43520	TI2		11.0	661.8	✔							✔	✔		✔		✔
4	AT2G43530	Unk	5.5	0.7	7.3	✔	✔						✔	✔		✔	✔	✔
4	AT2G43535	Unk	5.2	2.3	21.6	✔							✔	✔		✔	✔	✔
4	AT2G43540	Unk		1.7	1.0	✔	✔			✔			✔	✔		✔	✔	✔
4	AT2G43550	Unk	5.5	3.4	0.2	✔							✔	✔		✔	✔	✔
4	AT2G43560	Unk	4.5	1.1	1.5	✔	✔						✔	✔		✔	✔	✔
4	AT2G43570	CHI	0.6	3.2	0.9	✔	✔						✔	✔		✔	✔	✔
4	AT2G43580	Unk	6.7	5.6	2.4	✔	✔				✔							
5	AT2G43630	Unk	0.6	0.5	1.1	✔	✔						✔	✔		✔	✔	✔
5	AT2G43640	Unk	4.5	0.4	0.9	✔	✔	✔	✔	✔			✔	✔		✔	✔	✔
5	AT2G43650	EMB2777	5.7	0.6	0.2	✔	✔						✔	✔		✔	✔	✔
5	AT2G43660	Unk	5.6	0.4	0.1	✔											✔	✔
5	AT2G43670	Unk	6.9	2.6	0.1	✔					✔						✔	✔
5	AT2G43680	IQD14	0.8	5.2	3.3	✔							✔	✔		✔	✔	✔
6	AT2G43820	SAGT1	3.4	12.6	48.3	✔	✔	✔						✔		✔	✔	✔
6	AT2G43830	Unk																
6	AT2G43840	UGT74F1	5.2	0.8	1.3	✔	✔									✔		✔
6	AT2G43850	Unk	5.6	1.4	0.6	✔	✔				✔		✔	✔		✔	✔	✔
6	AT2G43860	Unk	1.8	2.8	0.5	✔	✔											
7	AT3G02110	SCPL25	2.8	1.8	2.6	✔	✔						✔	✔		✔	✔	✔
7	AT3G02120	Unk	5.5	0.9	2.1	✔						✔	✔	✔		✔	✔	✔
7	AT3G02125	Unk	5.8	0.4	1.9								✔					✔
7	AT3G02130	RPK2	3.2	5.1	0.8	✔	✔				✔		✔	✔		✔	✔	✔
7	AT3G02140	AFP4	4.4	3.9	0.7	✔	✔		✔		✔		✔	✔		✔	✔	✔
					Total Count	38	24	4	3	3	6	2	32	32	0	33	32	37

1. Gene names are based on TAIR10 (http://www.arabidopsis.org).
2. Wilcoxon values are the maxed GWAs score of each gene. Gray blocks represent no SNP inside the gene.
3. Fold-changes are colored based on the efp browser (http://bar.utoronto.ca/efp/cgi-bin/efpWeb.cgi?dataSource=Seed).

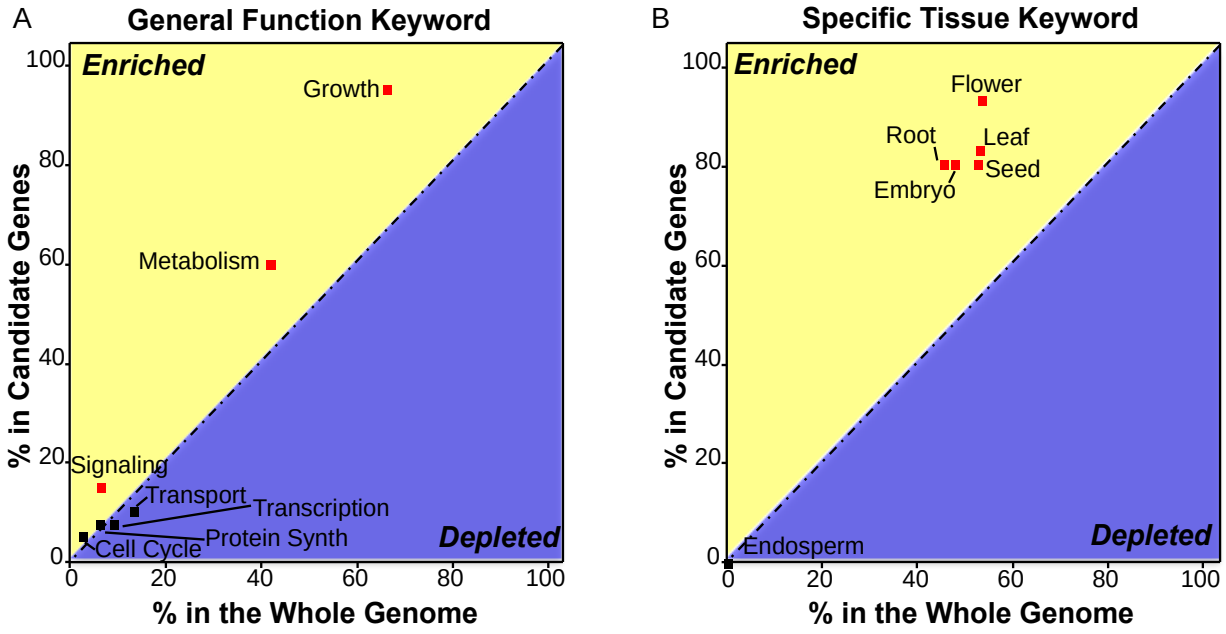


Figure 11. Keyword analysis of 40 candidate genes.

(A) Enrichment in keyword related to general function. **(B)** Enrichment in keyword related to specific tissue. The y-axis represents the percentage of 40 candidate genes hits in one of the keywords. The x-axis represents the percentage of the rest of the genome hits in one of the keywords. Percentage that is higher in a candidate gene than in the rest of the genome is considered as enriched and the area on the plot is colored by yellow. Percentage that is lower in the candidate gene than in the rest of the genome is considered as depleted and the area on the plot is colored by blue. Dash line represents the diagonal. Red boxes represent the enrichment is statistically significant using both Chi-sq test and permutation analysis. Black boxes mean no significance observed. Keywords are listed next to the boxes.

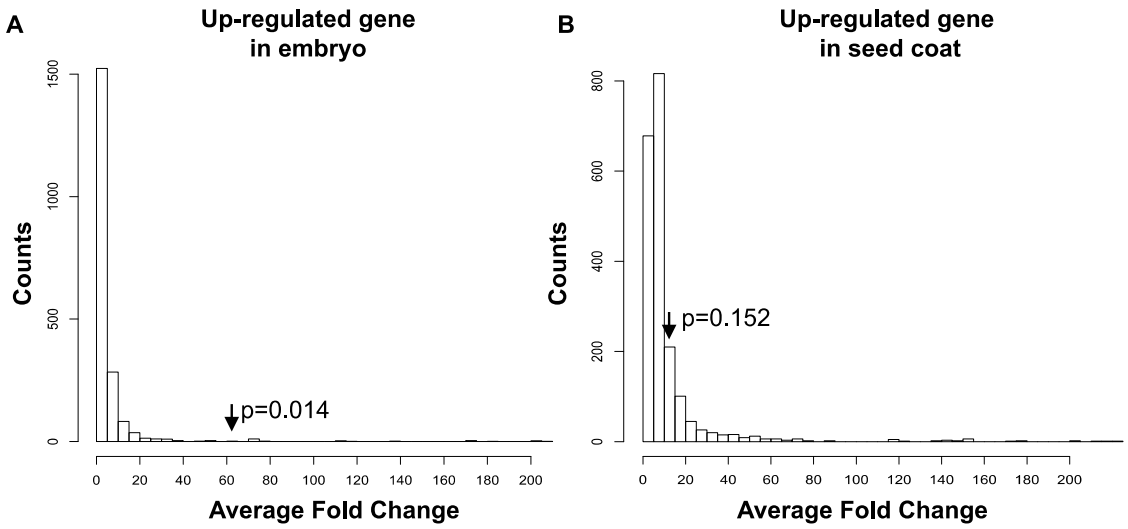


Figure 12. Permutation of microarray data.

(A) Histogram of the average fold change of 12 genes randomly sampled from 5031 genes which have at-least 2-fold upregulation. **(B)** Histogram of the average fold change of 15 genes randomly sampled from 7716 genes which have at-least 2-fold upregulation. The arrows point out the value observed from these candidate gene. p-values are calculated by deviding the number higher than the observed value by the total permutation times (n=2000).

2.4 DISCUSSION

Seed mass is a critical fitness trait for plants because it provides the nutrients for initial germination and carries the maternal information for early development. Understanding the genetic basis underlying the regulation of seed mass is important for both ecology and agriculture. Quantitative genetists have put deliberate efforts in searching QTLs causing natural variations in seed mass. However, only a few cases have been reported in using traditional QTL mapping method (Alonso Blanco *et al.*, 1999; Van Daele *et al.*, 2012) and few genes have been found in regulating seed development (Mizukami and Fischer, 2000; Ohto *et al.*, 2005; Jofuku *et al.*, 2005; Luo *et al.*, 2005; Schruff *et al.*, 2006; Herridge *et al.*, 2011; Van Daele *et al.*, 2012). This current map is novel and important for the following reasons. First, my study is the first that uses GWAs mapping methods to search candidate SNPs and genes associated with seed mass in *any plant*. It is also one of the first GWAs maps that try to understand the QTLs related to plant fitness traits (Atwell *et al.*, 2010; Fournier-Level *et al.*, 2011). I found five candidate loci distributed at the top of chromosome 1 and the bottom of chromosome 2. These locations have been suggested by a QTL map before (Alonso Blanco *et al.*, 1999). Because it uses SNPs as the genetic markers, my map is likely to provide better resolution and more accurate identification of candidate genes. Second, I also tried to evaluate my candidate genes using the available metadata, including keyword GO annotations and microarray results. I found great enrichment of seed-related keywords in my candidate gene list and several genes that are highly upregulated during seed development. Third, I tried to understand the geographic effect on plant fitness. Similar effort has only been seen in few reports previously (Fournier-Level *et al.*, 2011). I found that alleles associated with heavier seed mass tended to be present in the north

part of Europe. This is consistent with findings from other plant species showing larger seed sizes at high latitudes (Moles *et al.*, 2006) .

GWAs mapping has several advantages over previous approaches such as high throughput, fine resolution and less effort in creating mapping populations. However, GWAs is also vulnerable to the effects of population structure. Since the GWAs map relies on natural recombination events, genetic similarity could make high false positive ratios and thus blur the real candidates. While theory suggests that EMMA may correct the potential false positive of GWA scan due to population structure (Kang *et al.*, 2008), the empirical effect is unknown. QTL analysis, based on the experimental design, has no issue about genetics distance, thus is a great complementary tool for evaluating the result of the GWA analysis. Indeed, the combination of using GWA and QTL mapping methods together has been suggested and successfully reported several times (e.g., (Atwell *et al.*, 2010; Nordborg and Weigel, 2008; Brachi *et al.*, 2010; Chao *et al.*, 2012; Weigel, 2012)). As described before, the current GWA map identified three loci at the top of chromosome 1. While the SNPs with the highest GWA score were present at 0.74 Mb, 1.26 Mb and 6.42 Mb positions (Table 5). Particularly, the SNP at 6,427,347 bp is the most significant SNP across the entire genome and the alleles at this locus could alone explain 15.16 % of observed variations in seed mass by itself. The data suggested that the genetic regions around 0.74 to 0.75 Mb, 1.25 to 1.28 Mb and 6.40 to 6.43 Mb on chromosome 1 contain an important predictor of natural variation in seed mass of the tested wild accessions (Figure 3). Previous QTL analysis of LerxCvi offspring by Alonso-Branco et al. (1999) supported my results in the sense that they also found genetic material around the markers at the beginning of chromosome 1 that explained a significant portion of variations in twelve different fitness traits including the seed weight per 100 seed. Similarly some

researchers using BurxCol, CvixCol and LerxCvi RILs reported the top of chromosome 1 as the candidate region generating variations in seed size as well (Weigel, 2012; Moore *et al.*, 2013; Van Daele *et al.*, 2012; Herridge *et al.*, 2011). Similarly, an area around the second half of chromosome 2 showed up in several maps (Atwell *et al.*, 2010; Alonso Blanco *et al.*, 1999; Van Daele *et al.*, 2012; Herridge *et al.*, 2011). I also identified a big locus (locus #4) close to this region, which covered from 18.06 Mb to 18.16 Mb, including three sub loci. Similarly, GWA analysis identified the fifth locus at the top of chromosome 3, ranging from 3.7 to 3.9 Mb, which has also been suggested by one of the previous QTL maps (C., Lister and C., Dean, 1993; Alonso *et al.*, 1999). Although none of the reported QTL studies has seen my five candidate loci together, it is not surprising because none of the available parents differ at all five loci. Indeed, only Eds-1, Fab-4, Lov-1 and Nyl-2 have the heavy allele at all five loci. The precise locations I suggested here were not revealed by any of the previous QTL maps. This was most likely because previous QTL analysis only explored variations within Cvi, Ler, Bur and Col lines, and the genetic markers they used were AFLPs and RFLPs, which are much bigger than SNPs. Every identified QTL locus only explained the observed variation between the two parental lines, while on the other hand, my GWA analysis investigated candidate loci across the 178 lines. The high throughput and the pinpointed accuracy are the major advantages of GWA map as compared to traditional QTL analysis. However, QTL maps did not have the population structure issue, and thus it may serve as a great tool to re-check the loci identified by GWA analysis since EMMA could only correct the potential bias theoretically. Overall, these results suggest that using GWA and QTL mapping together could lead to successful identification of candidate loci.

2.5 FUTURE DIRECTION

The next step is to experimentally evaluate the candidate genes. I will order SALK lines corresponding to each gene and grow them under the same common garden conditions. I will first use PCR and qrt-PCR to check the effect of T-DNA insertions. Then I will measure the seed mass and compare the values to the Col-0 background. If I find the knockout lines have seed mass differing from the background significantly, I will then try to understand the gene's function during seed production.

2.6 ACKNOWLEDGEMENTS

I thank Dr. Muhammad Saleem for comments on the writing. I thank Seth Reighard and Juhyun Kim for assistance with data collection. I thank the Salk Institute Genomic Analysis Laboratory for the T-DNA insertion mutants distributed by ABRC. This research was supported by US National Science Foundation Grant #1050138 (M.B.T.).

3.0 *ARABIDOPSIS* ABC TRANSPORTER ATABCG16 INCREASES PLANT TOLERANCE TO ABSCISIC ACID AND ASSISTS IN BASAL RESISTANCE AGAINST *PSEUDOMONAS SYRINGAE* DC3000

Plants close their stomata following perception of the virulent bacterial pathogen, *Pseudomonas syringae* pv. *tomato* DC3000 (*Pst* DC3000) and the hormone triggering this stomatal closure is abscisic acid (ABA). These bacteria, on the other hand, secrete coronatine which blocks ABA signaling in guard cells and forces stomata to reopen. Once inside the leaf, *Pst* DC3000 again alters ABA signaling and suppresses SA-dependent resistance. Some wild plants exhibit resistance to *Pst* DC3000, but the mechanisms by which they achieve this resistance remain unknown. Here, I used genome-wide association mapping to identify an ATP-dependent binding cassette transporter gene, *AtABCG16*, in *Arabidopsis thaliana* that contributes to wild plant resistance to *Pst* DC3000. Through microarray analysis and GUS reporter lines, I show that the gene is upregulated by ABA, bacterial infection, and coronatine. Our collaborator at the University of Tennessee provided images using a GFP-fusion protein to show that the transporter localizes to the plasma membrane. T-DNA insertion and RNAi knockdown lines also exhibited consistent defective tolerance of exogenous ABA, impaired stomatal aperture regulation, and reduced resistance to infection by *Pst* DC3000. My conclusion is that AtABCG16 is involved in ABA tolerance and contributes to plant resistance against *Pst* DC3000. This is one of the first examples of ABC transporter involvement in plant resistance

to infection by a bacterial pathogen. It also suggests a possible mechanism by which plants prevent the negative effects of ABA accumulation during pathogen attack. Collectively, these results improve our understanding of basal resistance in *Arabidopsis* and offer novel ABA-related targets for improving the innate resistance of plants to bacterial infection.

3.1 INTRODUCTION

Bacterial pathogens of plants disperse through the global hydrological cycle (Morris *et al.*, 2008) and are likely to exert selection favoring defenses of wild plants at large spatial scales. In agriculture, these pathogens cause significant losses and necessitate extensive spraying of antibiotics (Oerke, 2006). While there is great potential to learn the mechanisms that underlie this resistance (Todesco *et al.*, 2010), the genetic underpinnings of resistance in wild plants remain essentially unknown. Identification of natural mechanisms by which resistant wild plants prevent the growth of bacterial pathogens has the potential to improve crop yields and reduce exogenous antibiotic use in agriculture.

The immune system of plants is sophisticated and involves coordinated actions of hundreds of genes (Jones and Dangl, 2006) and requires the presence of a single major hormone, salicylic acid (SA) (Vlot *et al.*, 2009). This pathway is cross-regulated by other signaling pathways in the plant, including those controlled by jasmonic acid (JA) (Endo *et al.*, 2008; Vlot *et al.*, 2009) and abscisic acid (ABA) (de Torres Zabala *et al.*, 2007). However, how and to what extent plants can regulate these hormonal-dependent pathways and crosstalk is still not clear. SA-dependent defenses are effective against many bacterial pathogens, including the

causal agents of bacterial spot disease, *Pseudomonas syringae* (Vlot *et al.*, 2009). However, at least one particularly virulent strain, *Pseudomonas syringae pv. tomato* (*Pst* DC3000), has evolved a novel virulence mechanism in which it secretes coronatine to "hijack" plant abscisic acid (ABA) biosynthesis (de Torres Zabala *et al.*, 2007), which elevates ABA levels and suppresses *ICS1* expression, leading to down-regulation of SA-dependent defenses (de Torres Zabala *et al.*, 2009). What has remained unknown is whether plants can counteract this suppression and, if so, how. ABA is mainly produced in vascular cells but is likely to act in opposite directions in mesophyll and guard cells during pathogen infection (Kuromori and Shinozaki, 2010; Zheng *et al.*, 2012). Intercellular ABA is favored in guard cells during bacteria attack since it keeps stomata closed (Munemasa *et al.*, 2007). In contrast, high ABA in the mesophyll cells is likely to cause lower SA-dependent resistance (de Torres Zabala *et al.*, 2009). Re-distribution of ABA among plant cells is therefore likely to be critical for plant defense, but the underlying genetics have not been identified previously.

ATP-dependent binding cassette (ABC) transporters have received particular attention recently for their importance in moving hormones, metal ions and other compounds across membranes (Rea, 2007). The largest subfamily of ABC transporters in *Arabidopsis thaliana*, the "AtABCG" group, contains 43 genes that share a common nucleotide-binding domain (NBD) and transmembrane domain (TBD) orientation (Verrier *et al.*, 2008). To date, three members of this AtABCG subfamily have been associated with ABA transport directly. AtABCG22 and AtABCG40 are found in the guard cells specifically and AtACBG25 is found in the vascular bundles (J., Kang *et al.*, 2010; Kuromori *et al.*, 2010; Kuromori *et al.*, 2011), transporting ABA and improving drought tolerance. AtABCG36 (PEN3/PDR8) is the only gene in AtABCG subfamily that has been previously reported for involvement in plant

resistance to bacterial and fungal pathogens (Kobae *et al.*, 2006; Stein *et al.*, 2006). The functions of other AtABCGs and particularly their possible roles in ABA transport or plant-pathogen interaction remain largely unknown. Here, I present evidence that one of the AtABCG transporters, AtABCG16, identified by genome-wide association (GWAs) mapping, is involved in both ABA response and plant resistance against *Pst* DC3000.

3.2 MATERIALS AND METHODS

3.2.1 Plant materials and growth conditions

Plants used for infiltration inoculation, including wild *Arabidopsis* accessions, nontransgenic Col-0 (CS60000), *atabcg16* mutants with T-DNA insertions (SALK_087501 and SALK_119868C) were acquired from the *Arabidopsis* Biological Resource Center (ABRC, http://www.arabidopsis.org/servlets/Order?state=catalog). Seeds were put on soil and cold-treated for one week before putting into the growth room at 22 ℃ under a 12h light/12h dark photoperiod. Seeds of plants used for flood inoculation, including Col-0, SALK knockouts, overexpressors and RNAi knockdowns, were sterilized using bleach. In brief, the seeds were first incubated with 70% ethanol for 1min and then incubated in 50 % (v/v) bleach with 0.05 % (v/v) tween-20 (FisherBiotech BP337-100) for 10 min. After surface sterilization, seeds were washed with sterile distilled water for four times and germinated on petri dishes containing 0.8 % agar with 0.5X strength MS medium (Sigma-Aldrich M5524). The plates were cold-treated for one week and then transported to the growth chamber at 22 ℃ under a 12h light/12h dark photoperiod with 55 % humidity.

3.2.2 Genome-wide association mapping

To identify candidates in plant resistance to *Pst* DC3000, I created a GWA map for 96 accessions from the RegMap panel using data published previously (Atwell *et al.*, 2010) of *Arabidopsis thaliana*, collected across the worldwide range. The dataset contained 205,803 SNPs derived from the 250K SNP data version 3.06. (https://cynin.gmi.oeaw.ac.at/home/resources/atpolydb/250k-snp-data). All SNPs used in the analysis were diallelic and had the minor nucleotide represented in more than 5% of the accessions. Wilcoxon rank-sum tests were used to calculate the significance of association between each SNP and the phenotypic values using standard methods (Atwell *et al.*, 2010). To handle confounding caused by population structure, we used the standard approach of mixed model analysis known as EMMA (Efficient Mixed-Model Association, (H., M., Kang *et al.*, 2008)). I presented log transformed P-values following the standard approach (Atwell *et al.*, 2010) and called the value as GWAs score in the paper. To control the bias of rare single extreme GWAs score, I smoothed the map by averaging the GWAs score of every 10 adjacent SNPs. The 13 loci identified were the ones that were common to both the Wilcoxon and EMMA maps.

3.2.3 Microarray data

The microarray data I used in this study are list as follows:

1. DC3000. AtGenExpress: Response to virulent, avirulent, typeIII-secretion system deficient and nonhost bacteria (NASCARRAYS-120, Numberger Lab).

2. ABA. AtGenExpress: ABA time course in wildtype seedlings (NASCARRAYS-176, Shimada Lab).

3. MeJA. AtGenExpress: Methyl Jasmonate time course in wildtype (NASCARRAYS-174, Shimada Lab).

4. IAA. AtGenExpress: IAA time course in wildtype seedlings (NASCARRAYS-175, Shimada Lab).

5. SA. AtGenExpress: Effect of ibuprofen, salicylic acid and daminozide on seedlings (NASCARRAYS-192, Shimada Lab).

6. ABA leaf. Boolean modeling of transcriptome data reveals novel modes of heterotrimeric G-protein action. (Pandey *et al.*, 2010)

7. ABA seed. AtGenExpress: Effect of ABA during seed imbibition (NASCARRAYS-183, Kamiya Lab).

8. ABA guard cells. Boolean modeling of transcriptome data reveals novel modes of heterotrimeric G-protein action. (Pandey *et al.*, 2010)

9. ABA mesophyll cells. Isolation of a strong *Arabidopsis* guard cell promoter and its potential as a research tool. (Yang *et al.*, 2008)

3.2.4 Bacterial assay

The bacterial strains I used were a lab strain, *Pst* DC3000 and a mutant strain, *Pst* DC3661, created by Dr. Cuppels and provided by Dr. Tambong (Agriculture and Agri-Food, Canada). For the GWAs mapping accessions, I used plant averages from an experiment initially described elsewhere (Atwell *et al.*, 2010). Briefly, those plants were infected by blunt syringe with a 1 x 10^4 cfu/ml solution of DC3000 at four days of growth. To perform the infiltration inoculation, I followed Aranzana et al. (Aranzana *et al.*, 2005). For each plant, I inoculated two leaves with 0.1 ml of 10^5 cfu/ml bacteria in 10 mM $MgSO_4$ buffer using a blunt-tipped syringe.

For the SALK knockout lines, I sampled six independent plants for each time point. For each sampled leaf, I used paper hole puncher to collect leaf disk ($0.28cm^2$) and then ground the leaf disk in 200 µl 10 mM $MgSO_4$ buffer. Forty microliters from diluted mixture were then plated on the agar plates with 50 µg/ml rifampicin (Sigma-Aldrich R3501-5G) and incubated at 28 ℃ for three days before counting the number of colonies. To perform the flood inoculation, I followed Ishiga et al. (Ishiga *et al.*, 2011). Nine seeds of each tested line were planted on agar plates with 0.5X MS salts. Inoculation was taken when plants were 4-weeks old. Bacteria was bulked up overnight and resuspended at $5x10^6$ cfu/ml with 0.025% Silwet L-77. Each plate received 40 ml of inoculation mixture for three minutes. On day three, for overexpressing and RNAi knockdown lines, three leaves from each plate (three plates per line per treatment) were randomly subsampled following the method mentioned in the infiltration-inoculation part. For SALK knockout lines, four leaves from each plate (six plates per line per treatment) were randomly subsampled.

3.2.5 Hormone plate assay

Seeds were sterilized and washed using the same method mentioned in the plant growth part. 25 seeds were placed on each plate of 0.5X strength MS medium containing different concentrations of ABA, IAA and SA. The plates were then incubated in the growth chamber at 22 ℃ under a 12 h light/12 h dark photoperiod with 55% humidity for measuring germination (on day 4), root length (on day 10) and mortality (on day 30). The means and standard errors were calculated based on at least four independent replications. For measuring root length, roots on day 10 were traced using a sharpie marker on transparent slides along with a 1 cm scale bar. The slides were then scanned and the pixels were measure by ImageJ. The root length was

calculated by dividing the pixels of root tracing by the pixels of 1 cm scale bars. The experiment was repeated twice with six replications each time.

3.2.6 Expression studies using (q)RT-PCR

(This part provided by Dr. Yanhui Peng at the University of Tennessee)

Total RNA was isolated from leaves, stem, flowers and root tissues of 6-week-old nontransgenic *Arabidopsis* plants; whereas total RNA was isolated only from leaves tissues of T-DNA knockout, overexpression and amiRNAi knockdown lines, respectively, utilizing TriReagent according to manufacturer's protocol (MRC, Cincinnati, OH). The residual genomic DNA in the extract was removed by treatments with RNase-free DNase I (Invitrogen, USA). First strand cDNA was synthesized using: ~2 μg of total RNA, 0.5 μg oligo$(dT)_{18}$ and SuperScript® III reverse transcriptase, according to the manufacturer's instructions (Invitrogen, USA). RT-PCR was carried out in an Eppendorf mastercycler (Hamburg, Germany) using GoTaq® Green Master Mix (Promega USA), which was programmed as follows: 2 min at 95 ℃ for pre-denature; 25-40 cycles of 15 s at 94 ℃, 15 s at 55 ℃, 20 s at 72 ℃ for each gene. qRT-PCR was carried out in an ABI-7900 thermal cycling system using a Real-Time PCR Power Mix Kit (ABI, USA) and was programmed at the same condition. For control reactions, either no sample was added or RNA alone was added without reverse transcription to test if the RNA sample was contaminated with genomic DNA. *AtActin2*, was selected as reference the gene. The oligonucleotide primers (Table 9) were designed with the Primer Express 2.0 software (Applied Biosystems-Perkin-Elmer, Foster City, Calif.). To test the suitability of these primer sets, the specificity and identity of the RT-PCR products were monitored by a melting curve analysis (65–99 ℃, 5 ℃ s^{-1}) of the reaction products, which can distinguish the gene-

specific PCR products from the nonspecific PCR products. All primers were synthesized by Integrated DNA Technologies (IDT, USA). Data analysis was performed as described by Yuan et al. 2006.

3.2.7 Generation of *Arabidopsis* transgenic lines

(This part provided by Dr. Yanhui Peng at the University of Tennessee)

To generate overexpression lines, the ORF of ABCG16 was first amplified using PCR and cloned into pENTR/D-TOPO vector (Invitrogen, USA). DNA sequencing confirmed ABCG16 was then cloned into Gateway pMDC32 vector via LR reaction to make the overexpression construct (Curtis and Grossniklaus, 2003). *Arabidopsis* plants (ecotype Columbia-0) were transformed with this transgene using the floral-dip method (Clough and Bent, 1998). Hygromycin-resistant lines (T1) were screened and PCR confirmed. Experiments were conducted with homozygote T3 plants.

Sequences of WBC type ABC transporters are highly conserved, and the hairpinRNAi knockdown strategy does not apply to generate individual mutants for members of this kind gene family. Then, the artificial microRNA method was chosen for its high efficiency and target specific knockdown. AmiRNA precursor sequences have been designed using the Web MicroRNA Designer (http://wmd3.weigelworld.org). The most appropriate amiRNA candidates were chosen by considered relative and absolute hybridization energy, GC content and the target sites (target: TACTAAGCTGTGTCACTCCTT; and AmiRNA :TACTAAGCTGTGTCACTGCTT). Two pair specific primers and one pair universal primers were designed; site-directed mutagenesis on endogenous miRNA precursor (MIR319a) was performed using overlapping PCR to generate the amiRNA sequences (Schwab *et al.,* 2006).

The modified miRNA precursor was cloned into pENTR vector. The sequencing confirmed amiRNA fragment was then cloned into gateway pMDC32 vector via LR reaction to make the RNAi construct. Hygromycin-resistant lines (T1) were screened and PCR confirmed. Experiments were conducted with homozygote T3 plants.

3.2.8 Promoter activity analysis and GUS staining

(This part provided by Dr. Yanhui Peng at the University of Tennessee)

For AtABCG16 promoter-driven GUS expression lines, a 1388 bp AtABCG16 promoter region was amplified by using EX Taq (Takara, Japan) with the primers listed in Table S9. Then the PCR product was cloned into the PCR8/GW/TOPO vector (Invitrogen, USA); after sequencing confirmed and integrated into the promoter analysis vector pMDC162 via LR reaction (Curtis and Grossniklaus, 2003). Transgenic plants were made as described above. T2 lines were used for the gus expression analysis. To test the expression profile of ABCG16 and its response to ABA, one-week-old seedling were soaked in 10 μM $MgCl_2$ control or 3 μM ABA for 24 h. For other treatments, leaves of 3-week-old plants were soaked in 10 μM $MgCl_2$ control, 10^8 cfu/ml DC3000 (COR+), 10^8 cfu/ml DC3661 (COR-), 3 μM ABA and 1 μM coronatine for 24 h, respectively. GUS staining was performed following standard procedure (Jefferson *et al.*, 1987).

3.2.9 Subcellular localization

(This part provided by Dr. Yanhui Peng at the University of Tennessee)

The 2211 bp AtABCG16 cDNA was clone as described above. The sequence of this clone (pENTR-AtABCG16) was integrated into the GFP-fusion protein vector pMDC43 using LR clonase (Invitrogen,USA). The construct of 2X35S::GFP- AtABCG16 was introduced into *Arabidopsis* by using an Agrobacterium-mediated transformation system. One-week-old seedling of T2 plants were used for study the Subcellular localization of GFP-fusion protein. Microscopic images were taken under a FITC filtered epifluorescence microscopy (Olympus BX51 model) with blue light at 400× magnification. For plasmolysis, root sections were treated with 20% sucrose for 10 min.

3.2.10 Stomatal response following treatments

I followed a standard procedure to measure stomatal aperture on the whole *Arabidopsis* leaves (Melotto *et al.*, 2006). To make most of the stomata at the open stage, I kept the plants under light for three hours. Fully expanded leaves detached from 4-week-old plants were immersed in water (with 0.1 % (v/v) methanol), 10 mM $MgSO_4$, 10 μM ABA, 0.5 ng/μl coronatine or $1x10^8$ cfu/ml bacterial (DC3000/DC3661) suspension. At 1hr and 3hr time points, intact leaves were observed and photographed under a compound microscope at 400X (Carl-Zeiss). The width of the stomatal aperture was measured using the software ImageJ and averaged by at least 30 stomata per line per treatment.

3.3 RESULTS

3.3.1 Genome-wide association mapping and microarray analysis suggest the involvement of AtABCG16 in plant resistance to bacterial pathogen *Pst* DC3000

To identify candidate genes that may explain differences in defense against bacterial infection in wild plants, I assessed a GWAs map (Figure 13A, Table 7) constructed from 96 genotypes of the RegMap panel of *A. thaliana*, consisting of a worldwide collection of plants for which over 200,000 single nucleotide polymorphisms (SNPs) have been identified (Atwell *et al.*, 2010). GWAs mapping has emerged as a particularly powerful method for identifying novel genes, because it utilizes genetic polymorphisms that are already present in wild populations and, therefore, is likely to pinpoint genetic variation that is important in nature (Weigel, 2012). To assess resistance to infection, those plants were challenged by blunt syringe infiltration with a 10^4 cfu/ml solution of *Pst* DC3000 as described previously (Atwell *et al.*, 2010). My assessment of the GWAs map revealed 13 major loci (Figure 13A) and 54 candidate genes (Table 8), of which one of the top candidates was an ABCG transporter, *AtABCG16* at Locus#10 on chromosome 3 (Figure 13B, Table 8).

In order to better understand the possible defense-related functions of *AtABCG16*, I first mined published microarray data for the widely-used Col-0 accession and found that *AtABCG16* had a 2.76-fold upregulation two hours after exposure to *Pst* DC3000 diminishing to 0.72-fold at 24 hr after exposure (Figure 13C). The early response to bacteria is consistent with the functional role in the initial stages of plant responses to infection. Given the key roles of SA, ABA, IAA and MeJA in plant signaling and known transport roles of ABCG transporters in moving these compounds (J., V., Dean and Mills, 2004; Kuromori *et al.*, 2010; Vlot *et al.*,

2009), I then asked whether *AtABCG16* expression was altered following exogenous application of these hormones to plants. I found that *AtABCG16* was induced 3.8-fold and 3.5-fold by 10 μM exogenous SA and IAA respectively, and not at all by MeJA (Figure 13D). In contrast, *AtABCG16* was upregulated nearly 40-fold by ABA. Indeed, this was the 34th highest response out of 22,591 *A. thaliana* genes, and this gene was the only one within the 250 Kb neighborhood to exhibit a response (Figure 13D). Furthermore, tissue-specific microarrays showed that *AtABCG16* was upregulated 65-fold in mesophyll cells and nearly 20-fold in guard cells treated with exogenous ABA (Figure 13E). Collectively, these data strongly showed that this transporter responds to *Pst* DC3000 and ABA.

To assess experimentally whether AtABCG16 is required for plant tolerance of exogenous ABA, SA, or IAA, we used a standard hormone assay on MS-media plates (Kuromori *et al.,* 2010) comparing the tolerance of two T-DNA insertion knockout lines relative to the background (Col-0). To characterize the knockouts, we first ran qRT-PCR (Figure 14A, Table 9) and found that they reduced the *AtABCG16* gene expression by 92% (SALK_087501, hereafter "*abcg16-1*") and 84% (SALK_119868C, hereafter "*abcg16-2*"). Because SA and IAA likely operate at different concentrations from ABA *in planta,* we included both a low range test (1 and 3 μM) and a high range test (100 and 300 μM) of ABA, SA, and IAA (Figure 15, Table 10). Through the plate assay, we found that both *abcg16-1* and *abcg16-2* had significantly reduced germination in the presence of exogenous ABA relative to the Col-0 background ($F_{4,45} = 18.49$, $P < 0.0001$, Figure 14B, Table 10), whereas no such response was detected in the presence of IAA (Figure 14C, Table 10) or SA (Figure 14D, Table 10). The loss of ABA tolerance was dramatic. Specifically, in the presence of 3 μM exogenous ABA (Figure 14B, Table 10), the *abcg16-1* line had a 94.8% reduction in germination (Dunnett T = -8.7, P <

0.001) and the *abcg16-2* line had a 43.5% reduction in germination (Dunnett T = -4.0, P = 0.002), relative to the background. Our conclusion from these assays was that *AtABCG16* was necessary for successful germination in the presence of exogenous ABA, but not SA or IAA, at the concentrations tested.

To further assess the tolerance of *abcg16-1* and *abcg16-2* to exogenous ABA, I also tested these lines across a smaller range of 0 to 1.5 μM exogenous ABA following an established protocol (Kuromori *et al.*, 2010) and again found highly significant defects in the tolerance of both knockout lines relative to the Col-0 background in terms of reduced germination ($F_{6,108}$ = 3.71, P = 0.002, Figure 14F, Table 19), reduced root growth ($F_{6,108}$ = 7.03, P < 0.001, Figure 14G, Table 11), and elevated mortality of the knockouts ($F_{6,103}$ = 3.92, P = 0.001, Figure 14H, Table 11). Together, these experiments demonstrated that knockouts of *AtABCG16* were defective in ABA tolerance and that the response was present at multiple stages of plant development including germination and subsequent growth and survival of germinated seeds.

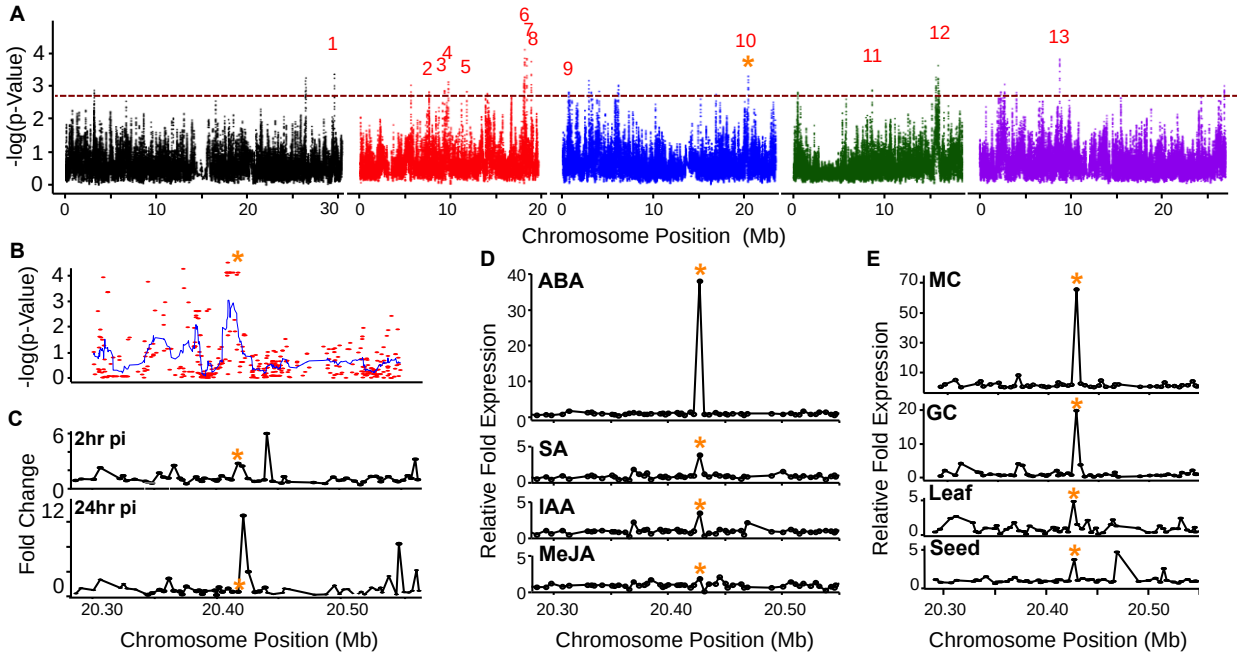


Figure 13. Genome-wide approaches suggest ABCG16 as a candidate gene of plant resistance through response to ABA.

(A) Genome-wide association map (GWAs) showing thirteen loci, including Locus#10 which contains the ABCG transporter, AtABCG16. **(B)** Fine GWAs map of Locus #10 showing interval containing exactly 30 genes upstream and downstream of AtABCG16 covering roughly 250Kb of Chromosome 3. **(C)** Fine map of Locus #10 showing gene expression 2 hr (top panel) and 24 hr (bottom panel) after *Pst* DC3000 induction. **(D)** Exceptional expression response of AtABCG16 to exogenous abscisic acid (ABA) application (34th highest out of 21,233 genes in the microarray) and moderate responses to salicylic acid (SA), indole-3-acetic acid (IAA), methyl jasmonate (MeJA) at 3hrs. **(E)** Tissue/location specific expression pattern in response to ABA. "MC" and "GC" stand for mesophyll cell and guard cell, respectively. All microarray data in this figure are acquired from *Arabidopsis* eFP browser (http://bar.utoronto.ca/efp/cgi-bin/efpWeb.cgi,(Winter *et al.*, 2007)). Orange asterisks indicate the position of AtABCG16.

Table 7. Average log(cfu) of *Pst* DC3000 used for GWAs map.

RegMap panel lines at 0, 1, 4, and 7 days post inoculation from three independent experiments described previously (Atwell *et al.*, 2010) sorted by average titer on day 4.

Titer Day 4			First Experiment				Second Experiment				Third Experiment			
Average log(cfu)	Name	ID	Day0	Day1	Day4	Day7	Day0	Day1	Day4	Day7	Day0	Day1	Day4	Day7
3.70	Fab-2	6917	1.30	3.00	4.30	3.30	1.90	1.60	3.20	3.30	1.30	-	3.60	3.30
3.89	Lov-5	6046	1.30	3.00	4.30	3.30	1.30	4.08	4.08	3.78	2.00	-	3.30	3.30
3.98	Zdr-6	6985	1.30	2.30	4.30	3.30	2.15	4.20	4.34	6.25	2.15	-	3.30	4.78
3.99	Eden-1	6009	2.08	2.30	4.78	3.30	2.34	2.00	3.90	3.30	1.30	-	3.30	3.30
4.40	Spr-1-2	6964	1.30	2.30	4.30	3.30	2.73	1.60	5.60	3.30	1.30	-	3.30	3.30
4.40	Omo-2-1	7518	2.38	3.30	4.30	7.15	1.78	4.15	5.60	4.15	1.30	-	3.30	5.65
4.44	Ei-2	6915	2.08	3.78	4.30	4.30	1.30	4.00	4.30	5.92	2.38	-	4.72	3.30
4.60	TAMM-2	6968	1.78	3.00	4.30	3.30	1.78	3.78	5.90	5.02	2.20	-	3.60	3.30
4.60	Wa-1	6978	1.78	3.30	4.30	4.15	1.30	4.85	6.20	5.89	2.53	-	3.30	7.99
4.62	GOT-22	6920	1.78	3.30	4.30	3.30	1.60	3.78	5.78	5.37	2.08	-	3.78	3.90
4.72	Lov-1	6043	1.30	3.30	4.30	3.30	2.00	3.60	6.56	4.30	1.90	-	3.30	3.30
4.78	Bay-0	6899	2.45	4.08	4.60	4.30	2.62	4.34	5.30	3.90	2.15	-	4.45	3.30
4.87	C24	6906	1.30	3.30	4.90	3.30	1.60	1.30	4.75	3.30	1.78	-	4.95	3.30
4.89	Fab-4	6918	2.45	3.00	4.30	3.60	1.30	3.90	7.06	6.10	2.92	-	3.30	3.30
4.91	RRS-10	7515	1.60	3.90	5.34	3.30	1.30	4.30	6.08	4.70	1.90	-	3.30	3.30
4.91	Pu2-23	6951	1.78	3.30	5.34	3.30	1.30	4.20	5.60	4.26	1.30	-	3.78	3.30
4.99	Shahdara	6962	2.26	3.60	4.30	3.30	2.64	4.66	7.37	5.48	2.30	-	3.30	3.90
5.01	Bor-4	6903	1.30	3.60	5.00	8.09	1.30	4.20	6.72	5.38	2.83	-	3.30	5.65
5.01	Ts-1	6970	1.30	3.78	4.30	7.85	3.36	4.41	7.43	3.60	2.89	-	3.30	5.43
5.01	NFA-8	6944	1.60	3.60	5.86	3.30	1.30	4.66	4.51	5.48	2.98	-	4.68	5.12
5.03	An-1	6898	1.90	4.20	4.30	7.15	2.00	4.38	7.48	3.30	1.30	-	3.30	4.58
5.09	GOT-7	6921	1.30	3.30	4.90	4.73	1.30	3.60	5.60	4.15	2.15	-	4.78	5.35
5.13	TAMM-27	6969	2.00	3.00	5.89	6.78	2.87	4.08	5.90	3.30	1.30	-	3.60	4.73
5.20	Ms-0	6938	2.08	2.30	4.30	3.30	1.78	2.78	5.30	5.68	2.08	-	6.00	5.51
5.27	Eden-2	6913	1.30	3.30	4.30	3.30	2.20	2.58	8.20	3.30	1.30	-	3.30	3.30
5.30	Yo-0	6983	2.08	3.30	5.20	7.81	1.30	4.72	7.38	5.07	2.30	-	3.30	7.41
5.38	Est-1	6916	2.15	3.60	4.78	5.23	2.30	4.34	7.58	5.58	1.90	-	3.78	3.60
5.40	PNA-10	7526	1.90	3.78	6.56	7.30	1.60	4.64	6.34	4.45	1.60	-	3.30	6.60
5.48	Bil-7	6901	1.30	3.30	4.30	3.30	2.26	3.30	7.15	6.35	2.20	-	5.00	7.88
5.50	Omo-2-3	7519	2.20	3.90	6.34	3.60	1.30	4.76	4.83	5.03	1.30	-	5.33	7.62
5.52	Mz-0	6940	2.00	2.30	5.91	6.60	3.13	3.60	4.64	3.30	1.30	-	6.00	3.30
5.53	KZ-1	6930	1.90	3.30	5.98	7.08	2.15	3.30	6.99	3.30	2.53	-	3.60	3.30
5.56	Edi-0	6914	2.20	3.30	4.30	3.30	1.30	4.88	6.87	-	2.34	-	5.52	5.36
5.63	Wei-0	6979	1.30	3.30	4.30	6.90	1.30	3.30	7.74	5.66	1.78	-	4.86	7.66
5.64	PNA-17	7523	1.78	3.30	5.38	7.20	2.20	4.41	8.25	6.20	2.20	-	3.30	7.26
5.77	Mr-0	7522	1.30	3.00	4.30	-	1.30	3.30	6.15	6.68	1.30	-	6.86	5.41
5.80	HR-10	6923	1.90	2.30	4.30	6.30	2.45	4.20	6.81	6.51	1.30	-	6.30	7.38
5.81	KNO-10	6927	1.30	3.30	5.60	7.64	1.30	4.34	6.83	4.08	1.90	-	4.99	7.60
5.93	Ws-0	6980	2.20	3.60	4.30	6.30	1.30	3.90	6.11	5.89	2.38	-	7.38	5.12
5.94	Nd-1	6942	1.78	3.60	6.43	7.62	1.90	4.38	8.07	5.76	2.87	-	3.30	7.41
5.97	RMX-A180	7525	1.90	2.30	6.49	7.64	1.30	4.08	6.41	4.68	2.08	-	5.02	7.51
5.99	Kas-1	8424	2.41	3.30	4.30	7.20	2.20	3.78	5.78	3.30	2.00	-	7.89	3.30
6.03	Zdr-1	6984	1.78	3.30	4.78	6.60	1.60	4.51	6.70	4.93	1.78	-	6.60	5.63
6.05	CIBC-5	6730	1.60	3.30	4.30	6.60	1.30	3.30	6.91	6.81	1.90	-	6.94	3.30
6.07	CS22491	7438	1.30	3.00	6.15	7.76	2.51	4.00	6.56	6.41	2.85	-	5.51	8.38
6.12	RMX-A02	7524	2.60	3.00	5.76	5.42	2.30	4.48	8.20	5.51	2.38	-	4.38	6.30
6.12	Van-0	6977	1.90	3.90	6.90	5.47	2.64	4.34	7.15	4.38	2.45	-	4.30	3.30
6.13	Uod-1	6975	1.78	2.30	6.58	7.15	2.08	4.64	6.48	5.77	1.78	-	5.35	5.61
6.16	Pu2-7	6956	1.78	2.30	6.43	7.64	2.08	4.56	7.46	6.46	2.76	-	4.58	3.30
6.18	Kondara	6929	2.08	3.30	6.45	6.90	2.34	4.20	7.54	5.21	2.15	-	4.53	6.30
6.21	LP2-2	7520	2.00	3.30	4.30	3.30	1.60	4.15	7.60	5.06	1.60	-	6.72	7.15
6.28	Ts-5	6971	1.90	3.60	5.26	7.53	1.30	4.15	6.72	6.32	2.48	-	6.86	4.95
6.31	Lz-0	6936	1.30	3.30	5.00	4.62	1.30	4.30	6.91	4.34	2.30	-	7.02	5.14
6.34	Fei-0	8215	2.08	3.30	6.19	7.00	3.02	4.70	7.40	5.06	1.30	-	5.43	5.55
6.36	Bil-5	6900	2.26	2.30	4.30	3.30	1.60	4.00	7.43	5.89	2.64	-	7.34	3.30
6.39	Ler-1	6932	2.00	3.60	4.30	7.87	2.53	4.81	7.95	6.11	2.79	-	6.92	4.86
6.43	Ws-2	6981	1.30	3.90	4.60	8.29	2.08	2.41	7.11	6.90	1.30	-	7.58	7.08
6.45	RRS-7	7514	1.78	3.78	6.97	7.91	2.34	4.30	8.01	5.95	1.90	-	4.38	7.62
6.53	Bur-0	6905	2.00	4.08	6.55	7.45	2.15	4.00	7.04	5.70	1.90	-	6.00	5.19
6.61	Ull-2-3	6973	1.90	4.15	6.62	4.56	1.60	4.00	8.06	6.60	1.30	-	5.15	5.27
6.67	Cvi-0	6911	2.58	3.78	6.59	7.79	1.30	4.00	7.45	6.68	2.41	-	5.97	7.73
6.70	SQ-1	6966	2.30	3.30	6.31	7.41	3.29	3.60	7.39	6.51	1.78	-	6.38	7.58
6.70	KZ-9	6931	2.56	3.00	5.92	6.90	2.79	4.48	7.27	3.30	2.38	-	6.90	3.30
6.72	Spr-1-6	6965	1.90	3.00	6.63	7.73	1.30	4.38	5.39	6.90	1.30	-	8.15	7.15
6.77	Se-0	6961	2.00	3.30	6.97	7.91	2.68	4.53	7.45	5.46	2.68	-	5.90	3.30
6.78	Gu-0	6922	2.15	3.00	5.45	7.08	2.58	3.90	8.11	4.85	2.08	-	6.78	6.20
6.79	Uod-7	6976	2.41	3.60	7.00	6.78	2.30	4.00	6.26	5.37	2.00	-	7.11	6.78
6.79	LP2-6	7521	1.30	4.15	6.81	7.68	1.78	4.72	7.23	5.92	2.60	-	6.34	5.54
6.80	NFA-10	6943	2.15	3.78	6.94	8.24	3.16	3.30	6.20	3.30	2.81	-	7.26	6.60
6.81	KNO-18	6928	2.00	3.30	6.50	7.00	2.76	4.60	7.06	4.45	1.78	-	6.87	7.91
6.82	Bor-1	5837	2.20	4.15	6.32	6.30	2.26	4.15	6.85	6.27	1.30	-	7.28	7.41
6.83	Tsu-1	6972	2.34	2.30	6.16	7.15	1.60	4.38	7.07	6.40	2.15	-	7.25	7.48
6.85	Sorbo	6963	2.48	3.30	6.65	7.56	2.15	3.90	7.21	6.40	2.30	-	6.70	6.30
6.86	Oy-0	6946	1.90	3.30	5.82	7.86	2.30	4.00	7.94	6.32	2.20	-	6.82	5.20
6.87	Wt-5	6982	2.53	3.60	5.75	6.90	2.15	4.15	7.97	6.68	1.90	-	6.89	7.20
6.91	Var-2-6	7517	1.30	3.30	5.91	-	1.30	4.00	7.26	4.81	2.15	-	7.55	3.30
6.92	Gy-0	8214	2.08	2.30	7.08	7.72	2.45	4.30	7.97	6.40	1.60	-	5.71	6.90
6.97	LL-0	6933	1.30	3.78	6.43	7.48	2.51	4.15	7.46	3.30	2.53	-	7.01	7.53
7.03	Ga-0	6919	2.00	3.30	6.28	7.73	1.30	3.30	8.10	5.55	2.15	-	6.72	5.68
7.05	Ag-0	6897	1.78	3.00	6.71	7.00	1.30	4.48	7.15	6.81	2.00	-	7.30	3.30
7.10	Col-0	6909	2.45	4.00	6.25	7.83	2.20	4.30	8.11	6.90	2.30	-	6.95	7.75
7.13	Pro-0	8213	2.00	3.60	6.45	7.48	1.30	4.86	7.53	6.68	2.45	-	7.42	7.92
7.18	Mt-0	6939	2.30	3.30	6.00	7.53	2.56	4.15	7.70	6.68	1.30	-	7.85	7.30
7.19	REN-1	6959	2.00	3.30	6.13	4.66	3.03	4.93	7.78	4.88	3.11	-	7.65	7.56
7.24	CIBC-17	6907	2.38	3.78	6.54	8.11	1.90	4.51	7.64	6.08	2.00	-	7.55	4.81
7.26	HR-5	6924	2.45	3.30	7.17	8.38	1.30	4.75	6.95	5.45	2.15	-	7.67	4.90
7.32	REN-11	6960	1.30	4.38	7.26	8.26	2.08	4.00	7.65	6.98	2.72	-	7.06	5.32
7.34	Ct-1	6910	1.30	3.60	6.54	6.60	1.78	4.38	7.38	3.30	1.30	-	8.08	7.15
7.36	Br-0	6904	1.60	3.78	6.36	6.78	2.30	4.08	8.18	6.81	2.15	-	7.55	7.15
7.44	Mrk-0	6937	2.15	3.60	7.39	8.63	1.30	4.20	7.92	7.11	1.30	-	7.01	8.87
7.44	Nok-3	6945	2.08	3.00	6.80	5.18	1.30	4.20	8.35	7.11	1.60	-	7.19	8.60
7.48	Ra-0	6958	1.90	3.30	6.20	7.56	1.30	4.08	8.37	5.67	1.60	-	7.87	8.60
7.50	Kin-0	6926	2.15	3.60	7.65	8.09	2.48	4.34	7.81	4.96	3.33	-	7.03	7.53
7.56	SQ-8	6967	2.41	3.30	6.44	7.62	1.60	4.89	8.00	6.81	2.30	-	8.24	8.07
7.69	Ull-2-5	6974	1.30	4.20	7.35	7.60	2.66	5.08	7.32	6.31	1.30	-	8.41	7.76
7.99	Var-2-1	7516	2.76	3.78	7.15	4.20	1.30	4.56	8.55	5.56	2.00	-	8.25	7.53

Table 8. Candidate genes from a genome-wide association map for resistance to *Pst* DC3000.

Sorted by Wilcoxon score. N/As represent that genes are inside the candidate loci but no SNP is found inside the gene coding regions.

Rank	Wilcoxon.DC3000	Locus	Gene	Name	Description
1	5.42	3	AT2G21910	CYP96A5	CYTOCHROME P450
2	4.97	7	AT2G44581	Unknown	RING/U-box superfamily protein
3	4.93	8	AT2G45850	Unknown	AT hook motif DNA-binding family protein
3	4.93	8	AT2G45890	RHS11	Kinase partner family protein
3	4.93	8	AT2G45900	Unknown	Phosphatidylinositol N-acetyglucosaminlytransferase subunit
6	4.91	7	AT2G44570	GH9B12	Glycosyl hydrolase 9B12,hydrolase activity
6	4.91	7	AT2G44590	DL1D	DYNAMIN-like 1D,GTPase activity
8	4.89	8	AT2G45870	Unknown	Bestrophin-like protein
9	4.77	12	AT4G32970	Unknown	Unknown
10	4.74	8	AT2G45860	Unknown	Unknown
11	4.58	4	AT2G22950	Unknown	Cation transporter/ E1-E2 ATPase family protein
12	4.54	10	AT3G55070	Unknown	LisH/CRA/RING-U-box domains-containing protein
13	4.53	12	AT4G32960	Unknown	Unknown
14	4.42	12	AT4G32980	ATH1	Transcription factor involved in photomorphogenesis
15	4.29	2	AT2G17580	Unknown	Undecaprenyl pyrophosphate synthetase family protein
16	4.25	6	AT2G43850	Unknown	Integrin-linked protein kinase family
17	4.17	8	AT2G45830	DTA2	Downstream target of AGL15, post-germinative development
18	4.16	1	AT1G78660	GGH1	Gamma-glutamyl hydrolase cleaving pentaglutamates
19	4.14	10	AT3G55080	Unknown	SET domain-containing protein
19	4.14	10	AT3G55090	ABCG16	ABC-2 type transporter family protein
21	3.86	2	AT2G17560	Unknown	HMGB (high mobility group B) protein
22	3.82	2	AT2G17570	Unknown	Undecaprenyl pyrophosphate synthetase family protein
23	3.81	6	AT2G43820	SAGT1	Induced by Salicylic acid, virus, fungus and bacteria
24	3.79	6	AT2G43840	UGT74F1	Transfers UDP:glucose to salicylic acid (forming a glucoside)
25	3.57	12	AT4G32950	Unknown	Protein phosphatase 2C family protein
26	3.55	9	AT3G03060	Unknown	P-loop containing nucleoside triphosphate hydrolase
27	3.51	4	AT2G22970	SCPL11	Serine-type carboxypeptidase activity
28	3.48	11	AT4G15180	SDG2	SET domain protein 2
29	3.37	5	AT2G27600	SKD1	Suppressor of K+ Transport Growth Defect1
29	3.37	5	AT2G27610	Unknown	Tetratricopeptide repeat (TPR)-like superfamily protein
29	3.37	5	AT2G27630	Unknown	Ubiquitin carboxyl-terminal hydrolase-related protein
32	3.34	1	AT1G78650	POLD3	Similar to DNA polymerase delta
33	2.95	9	AT3G03050	RHD7	Root hair development
34	2.90	9	AT3G03070	Unknown	NADH-ubiquinone oxidoreductase-related
35	2.68	4	AT2G22960	Unknown	Alpha/beta-Hydrolases superfamily protein
N/A	N/A	2	AT2G17556	Unknown	Unknown
N/A	N/A	2	AT2G17590	Unknown	Cysteine/Histidine-rich C1 domain family protein
N/A	N/A	3	AT2G21905	Unknown	Pseudogene
N/A	N/A	3	AT2G21920	Unknown	F-box associated ubiquitination effector family protein
N/A	N/A	4	AT2G22942	Unknown	Growth factors
N/A	N/A	4	AT2G22955	Unknown	Potential natural antisense gene
N/A	N/A	4	AT2G22980	SCPL13	Serine carboxypeptidase-like 13
N/A	N/A	5	AT2G27650	Unknown	Ubiquitin carboxyl-terminal hydrolase-related protein
N/A	N/A	6	AT2G43830	Unknown	Pseudogene
N/A	N/A	6	AT2G43860	Unknown	Pectin lyase-like superfamily protein
N/A	N/A	7	AT2G44578	Unknown	RING/U-box superfamily protein
N/A	N/A	7	AT2G44580	Unknown	Zinc ion binding
N/A	N/A	8	AT2G45840	Unknown	InterPro DOMAIN/s
N/A	N/A	8	AT2G45880	BAM7	Encodes a beta-amylase-like protein
N/A	N/A	10	AT3G55060	Unknown	Unknown
N/A	N/A	11	AT4G15165	Unknown	N-terminal nucleophile aminohydrolases
N/A	N/A	13	AT5G25310	Unknown	Exostosin family protein
N/A	N/A	13	AT5G25320	Unknown	ACT-like superfamily protein
N/A	N/A	13	AT5G25330	Unknown	Acetylglucosaminyltransferase family protein

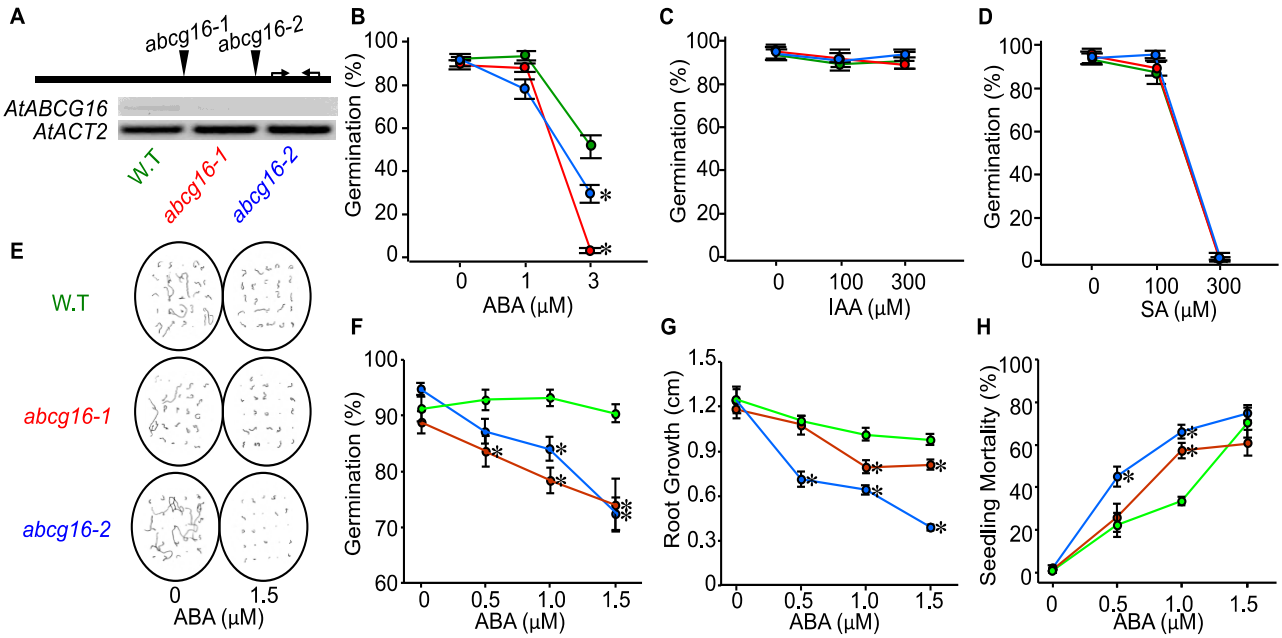


Figure 14. T-DNA insertion knockouts of *AtABCG16* are less tolerant to exogenous ABA.

(A) *AtABCG16* gene structure and T-DNA insertion sites of two *abcg16* mutants. *AtABCG16* contains a single exon, represented by the black square box. Transposon insertions in *abcg16-1* and *abcg16-2* are shown as triangles. Arrows show the positions of primers for detecting gene expression. The constitutive gene expression level of *AtABCG16* in Col-0 (W.T) and two knockouts (*abcg16-1*, *abcg16-2*) are checked by semi-qPCR, compared with the housekeeping gene (*AtACT2*). Gel pic provided by Dr. Yanhui Peng. **(B)-(D)** Plate assays comparing germination responses of the two T-DNA insertion knockouts of *AtABCG16* (*abcg16-1* and *abcg16-2*) relative to the background (CS60000) in the presence of exogenous B) 0-3 μM ABA, C) 0-300 μM SA, and D) 0-300 μM IAA. Each mean represents the average of five replicate plates, each containing 25 seeds. Asterisk (*) indicates significant Dunnett contrasts at $p<0.05$. **(E)** Representative tracings of root lengths for 0 and 1.5 μM ABA treatments on day 10. In this and all other figures, results are shown as mean ±1SE unless otherwise noted. **(F)-(H)** Plate assays testing the effect of 0, 0.5, 1.0, or 1.5 μM exogenous ABA on F) germination (%), G) root length (cm), H) mortality (%). Each mean represents ten replicate plates, each containing 25 seeds. Germination, root length, and mortality were measured on Day 4, 10, and 30, respectively. Asterisk (*) indicates significant Dunnett contrasts at $p < 0.05$.

Table 9. RT-PCR measuring AtABCG16 expression in T-DNA knockouts, RNAi knockdowns and overexpressors.

SALK Knockouts

	WT	***abcg16-1***	***abcg16-2***
Related to ACT2 rpp1	0.00044	0.00004	0.00010
Related to ACT2 rpp2	0.00051	0.00005	0.00012
Related to ACT2 rpp3	0.00085	0.00005	0.00007
Average	0.00060	0.00005	0.00010
Ratio	1	0.07749	0.16058
STDEV		0.01577	0.03776

Transgenic Lines

	WT	**amiR.3**	**amiR.5**	**amiR.7**	**amiR-14**	**OE 2**	**OE 3**	**OE 9**	**Empty vector 1**
Related to ACT2 rpp1	0.00443	0.00087	0.00074	0.00033	0.00087	0.10629	0.13921	0.18921	0.00625
Related to ACT2 rpp2	0.00478	0.00095	0.00135	0.00023	0.00095	0.07567	0.10702	0.16702	0.00373
Related to ACT2 rpp3	0.00444	0.00094	0.00078	0.00019	0.00094	0.04422	0.10479	0.16479	0.00289
Average	0.00455	0.00092	0.00096	0.00025	0.00092	0.07539	0.11701	0.17368	0.00429
Ratio	1	0.20186	0.21036	0.05529	0.20545	16.56608	25.71065	38.16213	0.94255
STDEV		0.00004	0.00034	0.00007	0.00004	0.03104	0.01926	0.01350	0.00175

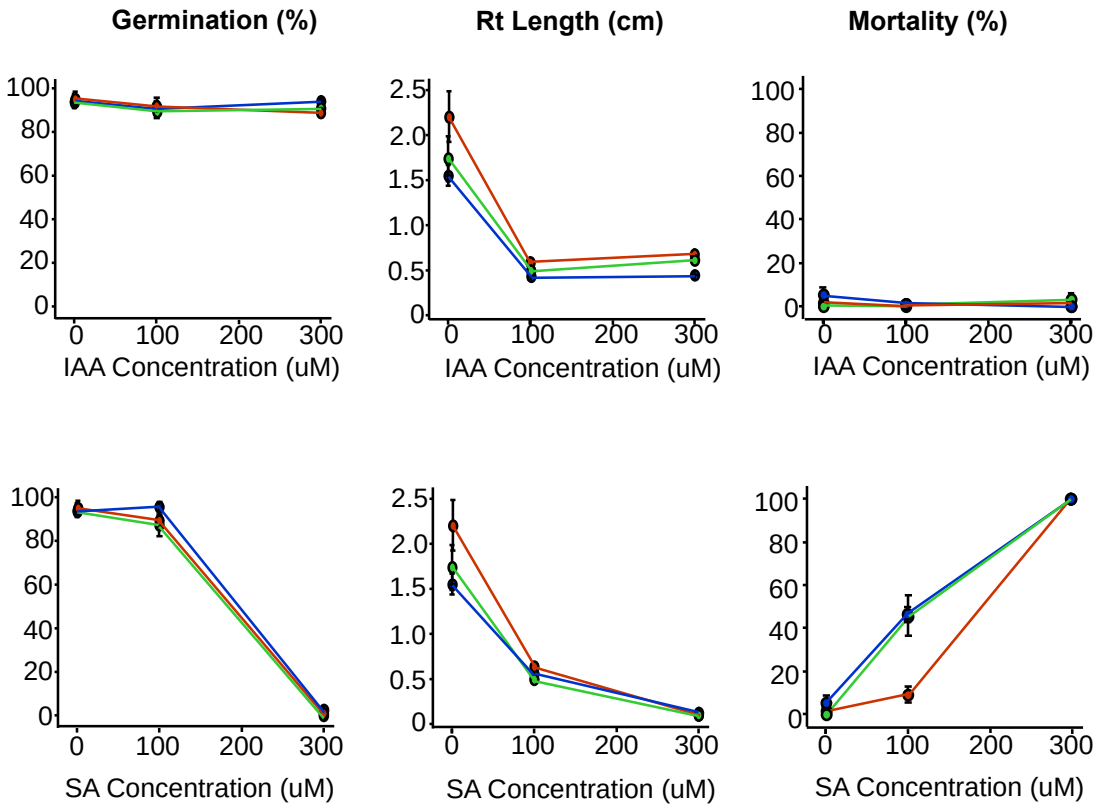


Figure 15. High range treatment of plants with IAA and SA, respectively.
Green, red and blue represent background (Col-0), *abcg16-1* and *abcg16-2*, respectively.

Table 10. Large-scale hormonal plate assay.

Plate assay results for germination (%) of *abcg16-1* and *abcg16-2* knockouts relative to the CS60000 background in response to low (0-3 μM) or high (0-300 μM) range abscisic acid (ABA), indole-3-acetic acid (IAA), or salicylic acid (SA).

Experiment	Trt	Line	No. plates	No. Seeds per plate	Germination Average (%)	+/- SE
Low Range	Control	Background	6	25	92.7	0.7
		abcg16-1	6	25	89.3	2.2
		abcg16-2	6	25	91.3	3.0
	ABA 1uM	Background	6	25	94.0	2.3
		abcg16-1	6	25	88.0	2.1
		abcg16-2	6	25	78.0	4.5
	ABA 3uM	Background	6	25	52.0	5.4
		abcg16-1	6	25	2.7	1.3
		abcg16-2	6	25	29.3	4.2
	IAA 1uM	Background	6	25	92.7	1.6
		abcg16-1	6	25	88.7	3.5
		abcg16-2	6	25	78.7	2.0
	IAA 3uM	Background	6	25	93.3	2.2
		abcg16-1	6	25	88.7	2.8
		abcg16-2	6	25	82.0	4.6
	SA 1uM	Background	6	25	94.0	3.2
		abcg16-1	6	25	88.0	4.5
		abcg16-2	6	25	86.7	3.2
	SA 3uM	Background	6	25	93.3	1.3
		abcg16-1	6	25	92.0	2.7
		abcg16-2	6	25	85.3	4.3
High Range	Control	Background	5	25	93.6	2.4
		abcg16-1	5	25	95.2	3.2
		abcg16-2	5	25	94.4	3.0
	ABA 100uM	Background	5	25	0.0	0.0
		abcg16-1	5	25	0.0	0.0
		abcg16-2	5	25	0.0	0.0
	ABA 300uM	Background	5	25	0.0	0.0
		abcg16-1	5	25	0.0	0.0
		abcg16-2	5	25	0.0	0.0
	IAA 100uM	Background	5	25	89.6	1.6
		abcg16-1	5	25	92.0	2.2
		abcg16-2	5	25	91.2	4.8
	IAA 300uM	Background	5	25	91.2	3.9
		abcg16-1	5	25	88.8	2.0
		abcg16-2	5	25	94.4	1.6
	SA 100uM	Background	5	25	87.2	5.0
		abcg16-1	5	25	89.6	3.5
		abcg16-2	5	25	96.0	1.8
	SA 300uM	Background	5	25	0.0	0.0
		abcg16-1	5	25	0.8	0.8
		abcg16-2	5	25	2.4	1.6

Table 11. Small-scale hormonal plate assay.

Plate assay results for germination (%) on day 4, root length (cm) on day 10, and mortality (%) on day 30 of *abcg16-1* and *abcg16-2* knockouts relative to the CS60000 background in response to four concentrations of abscisic acid (ABA).

				Germination (%)		Root Length (cm)		Mortality (%)	
Treatment	Line	No. Plates	No. Seeds / Plate	Average	+/- SE	Average	+/- SE	Average	+/- SE
Control	Background	10	25	91.2	2.2	1.25	0.07	0.9	0.9
	abcg16-1	10	25	88.8	2.0	1.19	0.07	1.3	0.7
	abcg16-2	10	25	94.8	1.0	1.25	0.09	2.1	1.3
0.5uM ABA	Background	10	25	92.8	1.8	1.11	0.04	22.3	5.6
	abcg16-1	10	25	83.6	2.7	1.08	0.06	25.9	6.6
	abcg16-2	10	25	87.2	2.3	0.72	0.05	45.1	4.5
1.0uM ABA	Background	10	25	93.2	1.5	1.02	0.04	33.5	2.0
	abcg16-1	10	25	78.4	2.2	0.80	0.04	57.4	3.8
	abcg16-2	10	25	84.0	2.1	0.65	0.03	66.2	3.3
1.5uM ABA	Background	10	25	90.4	1.6	0.99	0.04	70.5	7.1
	abcg16-1	10	25	74.0	4.7	0.82	0.03	61.0	5.9
	abcg16-2	10	25	72.4	2.9	0.40	0.02	75.1	3.5

3.3.2 *abcg16*-RNAi knockdowns and *AtABCG16*-overexpressor display altered tolerance to ABA

To further test the function of AtABCG16 in ABA tolerance, I used transgenic lines with altered expression of *AtABCG16* (Figure 16A). The two independent small RNA interference (RNAi) lines, amiR-3 and amiR-14, had 5% and 20% of normal *AtABCG16* transcript levels, respectively (Figure 16A, Table 9). Both RNAi lines exhibited lower germination ratio and shorter root length compared with wild type plants (Figure 16C-D). Indeed, only 3% of amiR-3 seeds could germinate on ABA plates while 80% germinated on the control plates (Dunnett $T = -5.3$, $p < 0.001$, Table 12). Additionally, the seedlings had roots on ABA plates which were only 7% of the length on the control plates, but this difference was marginally nonsignificant (Dunnett $T = -2.8$, $p = 0.08$, Figure 16D, Table 12). I also used the transgenic lines with overexpressed *AtABCG16* in wild type Col-0 (Figure 16A) and found that the overexpressor exhibited elevated germination and root length in the presence of ABA, relative to the RNAi lines (Figure 16C-E). I didn't observe any difference between empty vector lines and wild type in these analyses, indicating no effect *per se* of pMDC32, the vector used to create the transgenic lines. Collectively, these data from my RNAi and overexpressor lines indicated a role of AtABCG16 in tolerance of exogenous ABA, confirming my previous finding from the T-DNA knockout lines.

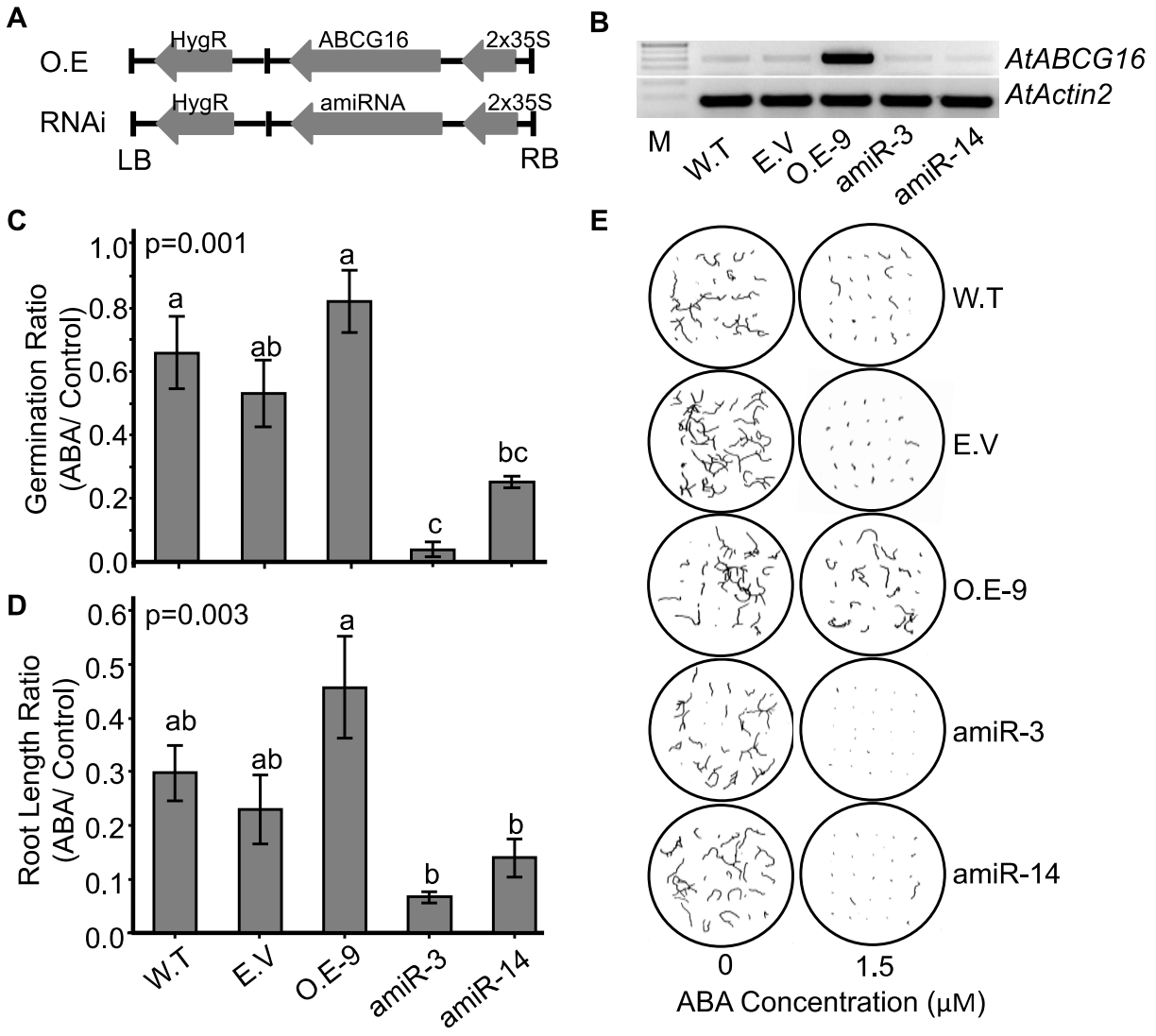


Figure 16. Response of ABCG16 overexpression and amiRNAi knockdown mutants to ABA.

(A) Sketch of overexpression and amiRNAi constructions. **(B)** Examination of AtABCG16 expression in leaves of different transgenic lines using semi-qPCR. Actin2 is amplified as a loading control. Gel pic provided by Dr. Yanhui Peng. **(C)-(D)** Plate assays testing the effect of 1.5 μM exogenous ABA on C) germination and D) root length controlled by the performance of respective lines on control plates. Each mean represents four replicate plates, each containing 25 seeds. P-values are calculated by general linear model and the means that do not share a letter are significantly different. **(E)** Representative tracings of root lengths for 0 and 1.5 μM ABA treatments on day 10. In this and all other figures, W.T, E.V, O.E, amiR stand for nontransgenic *Arabidopsis* (Col-0), wild type with empty pMDC32 plasmid, overexpressing lines and RNAi lines, respectively.

Table 12. Plate assay results for germination percentage (GP) and root length (RL) of wild type, empty vector, overexpressor, amiR-3 and amiR-14 response to 1.5 μM ABA.

Lines	No.Plates	Aver.GP % (Control)	Ave.GP % (ABA)	Relative GP (ABA/Control)	Ave.RL cm (Control)	Ave.RL cm (ABA)	Relative RL (ABA/Control)
Wild Type	4	99	65	0.66 ± 0.12	1.28	0.37	0.30 ± 0.05
E.V	4	100	53	0.53 ± 0.11	2.19	0.39	0.23 ± 0.06
O.E-9	4	89	73	0.82 ± 0.10	1.68	0.75	0.46 ± 0.10
amiR-3	4	80	3	0.04 ± 0.02	1.42	0.09	0.07 ± 0.01
amiR-14	4	100	25	0.25 ± 0.02	1.91	0.21	0.14 ± 0.04

3.3.3 AtABCG16 localizes to the plasma membrane

To localize AtABCG16, we use a construct consisting of the cauliflower mosaic virus (CaMV) 35S promoter driving the green fluorescent protein (GFP) fused to AtABCG16 protein (2X35S::GFP-AtABCG16). Subcellular localization of the fusion protein was observed using the filtered epifluorescence microscopy imaging of the green fluorescence signals in cells of the 2X35S::GFP-AtABCG16 transformed plants (Figure 17). In root cells, fluorescence was highest on the cell surface, indicating that GFP-AtABCG16 was located in the plasma membrane, but not in the tonoplast or cytoplasm (Figure 17). To further exclude the possibility that the protein was on the cell wall, we treated the root tip cells with 20% sucrose to create a high osmotic environment following a published protocol (Kuromori and Shinozaki, 2010). In the osmotically-treated cells, the fluorescence signal was observed on the plasma membrane rather than on the cell wall (Figure 17C-D). In leaf cells, we also observed fluorescence signals in both pavement cells and guard cells (Figure 17E-F). Collectively, these results clearly established the localization of AtABCG16 on the plasma membrane.

3.3.4 *AtABCG16* gene expression is induced by hormones and bacteria

To further investigate how this gene is regulated, we created transgenic lines containing the GUS reporter gene driven by a native *AtABCG16* promoter (Figure 18A). This native promoter of *AtABCG16* (Figure 18B) contains important binding motifs for the ABA response factor, ABRE, at -370, -250, and -80, as well as a binding site at -470 for WRKY38, a transcription factor well established to be directly linked to infection by *Pst* DC3000 (Kim *et al.*, 2008). In *AtABCG16 pro*::GUS transgenic plants, in the absence of exogenous ABA, the GUS signals were observed in the center of the roots and hypocotyl (Figure 18C-D), and the veins of

leaves (Figure 18E-F), consistent with association with vascular bundles. After treatment with 3 μM ABA for 24hr, GUS signals were observed ubiquitously in both roots and leaves (Figure 18C-F), indicating significant upregulation in all types of cells. To see if *AtABCG16* was upregulated by bacterial pathogens, we tested the transgenic plants with the coronatine secreting strain (COR+), *Pst* DC3000 and a coronatine knockout strain (COR-), *Pst* DC3661, along with two hormones, ABA and coronatine. We found that coronatine upregulated *AtABCG16* even more strongly than ABA did (Figure 18G). However, COR+ and COR- bacteria both induced *AtABCG16* expression, indicating that coronatine alone does not determine the induction. Also, we observed uniform GUS levels between guard cells and pavement cells under ABA induction (Figure 18H), which confirmed that AtABCG16 was responding in all types of leaf cells as indicated by the microarray data (Figure 18E). Interestingly, under coronatine treatment, *AtABCG16* was turned on significantly higher in surrounding tissues than in the guard cells (Figure 18H). Given that guard cells do not produce ABA (Endo *et al.*, 2008), our result is consistent with *AtABCG16* responding to induced ABA biosynthesis in the pavement cells. Collectively, our GUS assay data showed that AtABCG16 responded to bacterial pathogen, ABA, and coronatine but also that the genetic mechanisms of the responses to ABA and coronatine might be different.

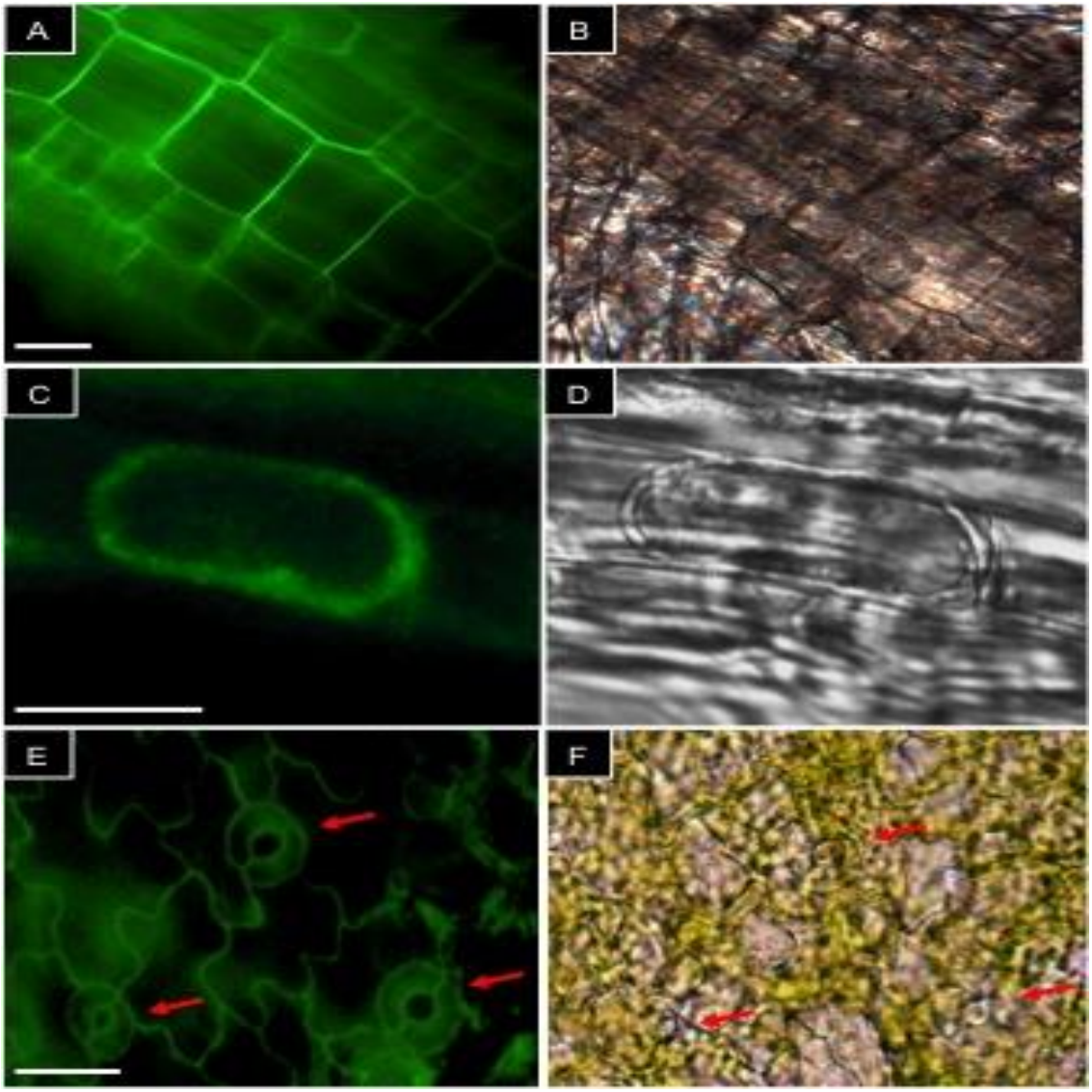


Figure 17. Plasma membrane localization of 2X35S::GFP-AtABCG16 fusion protein.

(A), (C) and **(E)** show GFP expression. **(B), (D)** and **(F)** are taken under white light. A-D are root cells. C and D show plasmolyzed root cells after treatment with 20% sucrose. E and F are leaf epidermal cells. White bars = 10 μm. Red arrows indicate stomata. Image provided by Dr. Yanghui Peng.

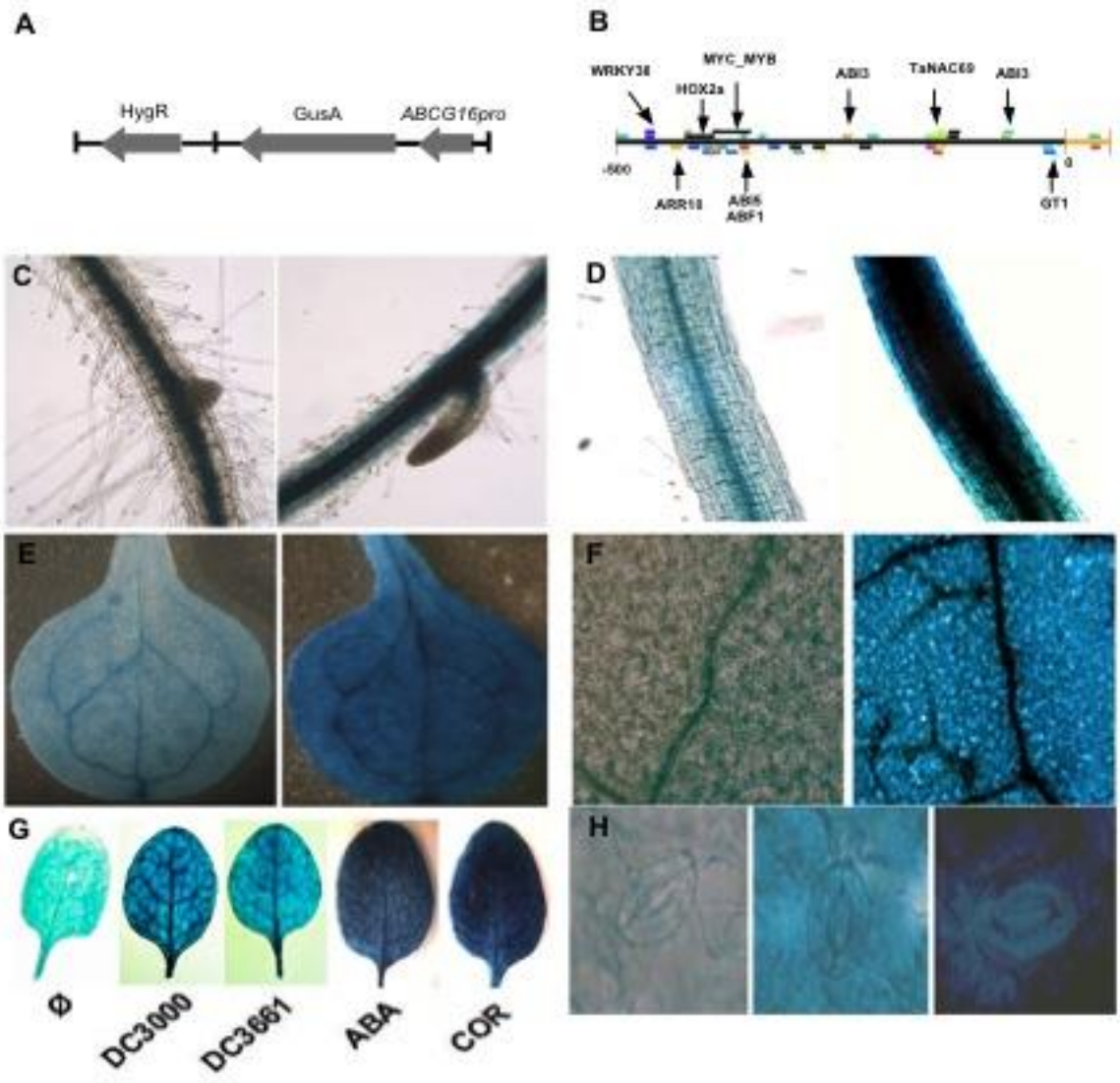


Figure 18. *AtABCG16* expression in plant tissues.

(A) Diagram of *AtABCG16pro::GUS* construct. **(B)** Diagram of the *ABCG16* promoter from Athamap, showing elements in the 500 bp upstream region, including three binding sites for ABI3, a known ABA-related transcription factor, and WRKY38, a known bacterial-related transcription factor. **(C)–(F)** GUS staining without (left) and with (right) ABA in root (C), hypocotyl (D), cotyledon (E) and 3-weeks leaf (F). **(G)** GUS staining of leaves in response to10 μM $MgCl_2$ control, 10^8 cfu/ml DC3000 (COR+), 10^8 cfu/ml DC3661 (COR-), 3μM ABA and 1μM coronatine (left through right). **(H)** GUS staining around stomata cells. From left to right are: control, 3 μM ABA and 1 μM coronatine. All plants were treated for 24 hours. Image provided by Dr. Yanghui Peng.

3.3.5 AtABCG16 is involved in stomatal closure and induced by hormones and bacteria

Given that ABA and coronatine have been reported to regulate stomatal aperture (Figure 19A) in response to biotic and abiotic stresses (Ren *et al.*, 2010; W., Zeng and He, 2010; Zheng *et al.*, 2012), I tested stomatal responses to both hormones and bacterial pathogens using the overexpression lines of *AtABCG16*. For hormonal response, ABA treatment strongly induced all lines to close the stomata and coronatine kept the stomata opened (Figure 19B). When ABA and coronatine were applied together, the average stomatal aperture was 64% lower in the overexpressor line than in the nontransgenic control (Figure 19B, Dunnett $T = -6.77$, $p < 0.001$). Meanwhile, two RNAi lines (amiR-3 and amiR-14) had significantly larger stomatal aperture (mean = 3.12 and 2.58, SE = 0.31 and 0.15, Dunnett $T = 4.84$ and 3.09, $p < 0.001$ and $p = 0.01$, respectively, Table 21). When treated with water, I did not observe any significant differences among nontransgenic controls, overexpressor and RNAi lines ($F = 0.81$, $p = 0.492$, Figure 19B). To test the effect of bacterial application on stomata aperture, I applied either DC3000 (COR+) or DC3661 (COR-) to the Col-0 background and overexpressor line (Figure 19C, Table 13). At 1 hr, all plants had closed stomata in response to the bacterial application. Three hours post infection, for the Col-0 background, plants treated with DC3000 reopened their stomata (mean = 2.68, SE = 0.19, Table 13), whereas plants treated by DC3661, still kept their stomata closed to a significant degree (mean = 1.35, SE = 0.13). However, DC3661 and coronatine together, caused stomata to reopen (mean = 3.10, SE = 0.11). In contrast, the overexpression line retained closed stomata under all of the bacterial treatment conditions (DC3000, DC3661, or DC3661 plus coronatine, Table 13), which was significantly different from the Col-0 background plants (Dunnett $T = -10.9$, -2.75, and -16.45, respectively).

To assess whether stomatal closure was associated with the function of AtABCG16 during infection by bacterial pathogens, I challenged the Col-0 background plants, two AtABCG16-overexpressing lines, and three RNAi lines with *Pst* DC3000 or *Pst* Dc3661 using flood-inoculation (Figure 19D). At day three following infection, one overexpression line (O.E-3) had 8% lower *Pst* DC3000 compared to nontransgenic control (Dunnett T = -5.339, $p < 0.001$). The other overexpression line, O.E-2, had a lower mean titer, but the difference was marginally non-significant (Dunnett T = -2.475, p = 0.075). Meanwhile, all three *abcg16*-RNAi knockdown lines (amiR-3, amiR-5, amiR-7) had significantly higher bacterial titers (Figure 19D, Dunnett T = 4.891, 5.057 and 5.958, respectively). Titers of DC3661 were lower on all lines, but were not affected by the presence of *AtABCG16* (F = 1.25, p = 0.296, Figure 19D). In separate tests, I found that the knockout lines (*abcg16-1* and *abcg16-2*) were more susceptible to DC3000 but not to DC3661 relative to the Col-0 background line (Figure 19E-F). Because coronatine and ABA are known to not only affect plant resistance in guard cells (Zheng *et al.*, 2012) but also in the mesophyll cells as well (de Torres Zabala *et al.*, 2009), I infiltration-inoculated *abcg16-1* and *abcg16-2* to bypass the guard cell protection (Figure 20B). The bacterial titer of DC3000 in both T-DNA insertion lines were significantly higher than in the wild type, as with the flood inoculation (Dunnett T = 3.968 and 3.422, p = 0.024 and 0.071) and again there were no significant differences among DC3661 treated plants (Dunnett T = 0.276 and 0.580, p = 0.946 and 0.789). Together, these measurements indicated that overexpressing *AtABCG16* could maintain stomata closure under infection conditions that caused stomata to reopen in the Col-0 background plants (Figure 19G). However, AtABCG16 was clearly also providing a role in the defense of mesophyll cells as well, given that I also observed

compromised resistance when the guard cells were bypassed through blunt syringe (Figure 20B).

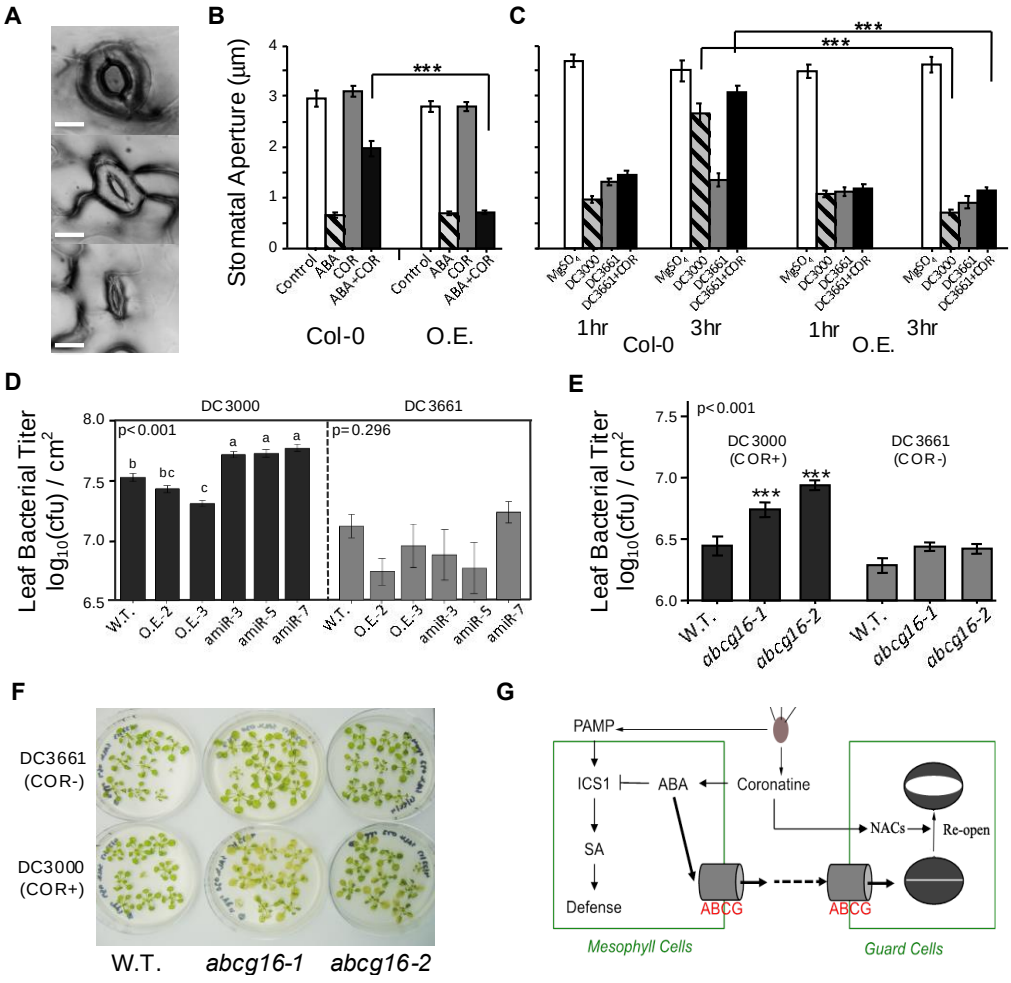


Figure 19. Response of ABCG16 mutants to hormone and bacterial treatment.

(A) Leaf stomata fully open (top), intermediate (middle), and closed (bottom). Scale bars represent 3µM **(B)** Stomatal aperture in intact leaves of Col-0 and overexpression line exposed to water with 0.1 % methanol (white), 10 µM ABA (diagonal), 0.5 ng/µl coronatine (gray) or 10 µM ABA+0.5 ng/µl coronatine (black bar). **(C)** Stomatal aperture in intact leaves of Col-0 and overexpression line exposed to $MgSO_4$ buffer (white), DC3000 (diagonal), DC3661 (gray) or DC3661+0.5ng/µl coronatine (black bar) at 1 hr and 3 hrs after treatment. *** represents p<0.001. Bacterial flood inoculation response to $5x10^6$ cfu showing **(D)** stronger disease symptoms at 3 dpi and **(E)** higher bacterial titers for both T-DNA insertion knockouts challenged with virulent DC3000, whereas all plant lines respond similarly to the cor- line (DC3661). (**F)** Plants leaf pattern 3 days after bacteria infection.

(**G)** Model of proposed ABCG16 function in which transporter moves ABA out of mesophyll cells to relive suppression of the SA-dependent pathway and into guard cells to keep stomata closed.

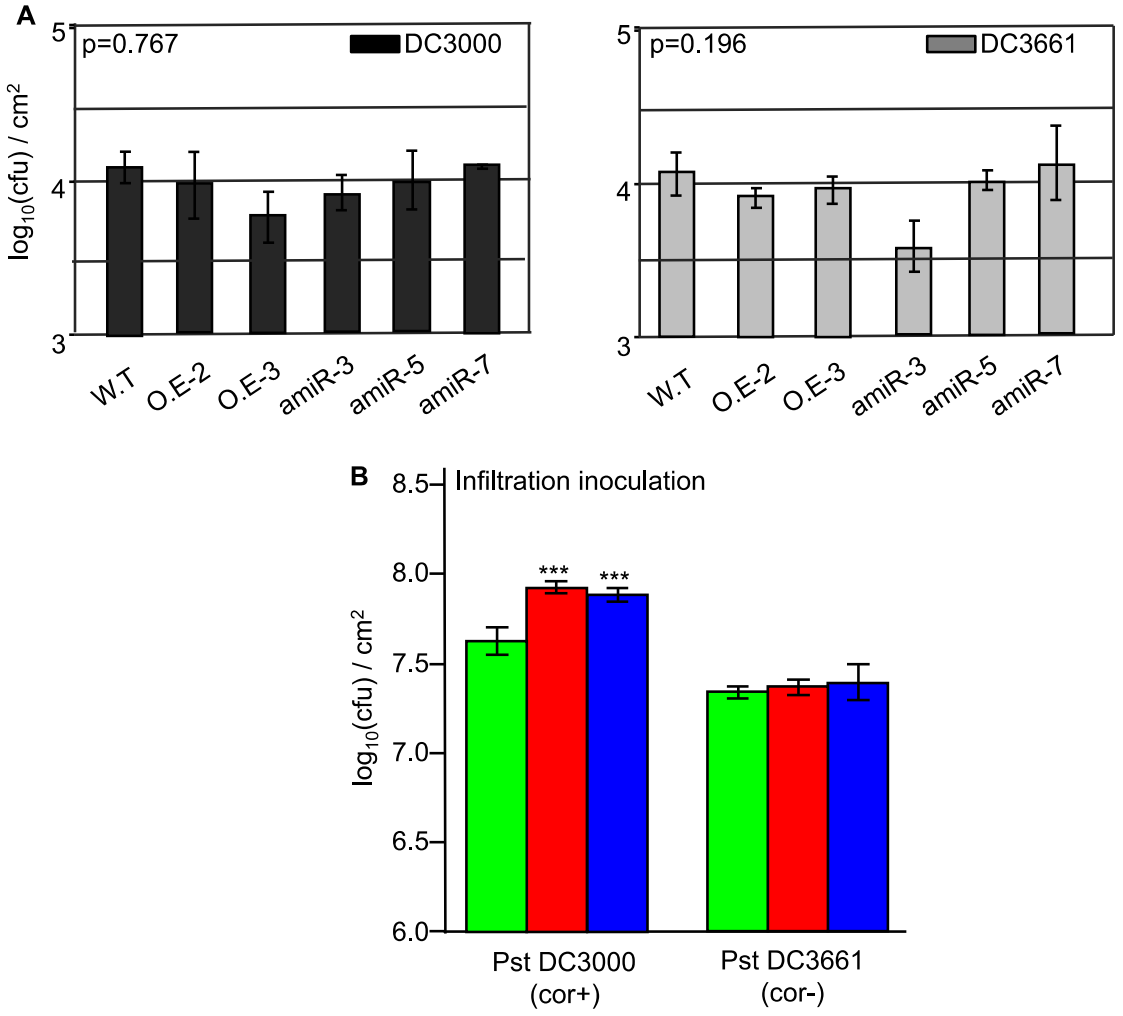


Figure 20. Additional bacterial growth measurement

(A) Bacterial titer at 5hr post flood-inoculation. **(B)** Bacterial titer on 4day post infiltration-inoculation. *** represents p<0.001. Green, red and blue represent wild type (Col-0), *atabcg16-1* and *atabcg16-2*, respectively.

Table 13. Stomatal aperture measurement of hormonal and bacterial response.

Stomatal Aperture (μm)
(mean±s.e)

Hormonal Response (3hr)

Lines	Water*	ABA	COR	ABA+COR
Wild Type	2.96±0.16	0.66±0.05	3.09±0.11	1.97±0.16
O.E-9	2.79±0.11	0.70±0.09	2.79±0.10	0.71±0.04
amiR-3	2.73±0.09	0.97±0.09	2.61±0.09	3.12±0.31
amiR-14	2.74±0.10	0.77±0.07	2.98±0.16	2.58±0.15

Bacterial Response (1hr)

Lines	MgSO4	DC3000	DC3661	DC3661+COR
Wild Type	3.71±0.12	0.97±0.07	1.31±0.06	1.14±0.10
O.E-9	3.48±0.14	1.08±0.06	1.12±0.09	1.18±0.09
amiR-3	2.78±0.08	1.05±0.11	1.18±0.08	1.93±0.16
amiR-14	2.57±0.08	0.92±0.09	0.84±0.06	1.73±0.10

Bacterial Response (3hr)

Lines	MgSO4	DC3000	DC3661	DC3661+COR
Wild Type	3.52±0.20	2.68±0.19	1.35±0.13	3.10±0.11
O.E-9	3.63±0.15	0.71±0.05	0.91±0.11	1.15±0.06
amiR-3	2.89±0.14	2.53±0.14	0.94±0.07	2.87±0.11
amiR-14	2.78±0.13	2.64±0.13	1.11±0.09	2.57±0.11

* Water contains 0.1% methanol (v/v).

Table 14. Bacteria growth on knockout, knockdown and overexpressing lines treated through flood or infiltration inoculation.

Bacteria Titer (log_{10}(cfu)/cm^2) (mean±s.e)

SALK lines blunt syringe inoculation

Lines	Time	DC3000	DC3661
Wild Type	Day4	7.63±0.08	7.34±0.03
abcg16-1	Day4	7.93±0.03	7.37±0.05
abcg16-2	Day4	7.89±0.04	7.40±0.10

SALK lines blunt syringe inoculation

Lines	Time	DC3000	DC3661
Wild Type	Day4	7.63±0.08	7.34±0.03
abcg16-1	Day4	7.93±0.03	7.37±0.05
abcg16-2	Day4	7.89±0.04	7.40±0.10

Transgenic lines flood inoculation

Lines	Time	DC3000	DC3661
Wild Type	5hr	4.09±0.10	4.05±0.13
O.E-2	5hr	3.95±0.22	3.90±0.06
O.E-3	5hr	3.75±0.16	3.95±0.09
amiR-3	5hr	3.89±0.11	3.56±0.16
amiR-5	5hr	3.96±0.19	3.99±0.04
amiR-7	5hr	4.08±0.01	4.09±0.24
Wild Type	Day3	7.66±0.04	7.11±0.10
O.E-2	Day3	7.54±0.03	6.72 ±0.11
O.E-3	Day3	7.41±0.03	6.94±0.18
amiR-3	Day3	7.88±0.03	6.87±0.20
amiR-5	Day3	7.89±0.03	6.76±0.21
amiR-7	Day3	7.93±0.03	7.22 ±0.09

3.3.6 Wild accessions with high tolerance to ABA were more resistant to bacterial infection.

A candidate gene from a GWAs map needs to meet two criteria. First, it should be able to be involved in causing, regulating or affecting the mapped phenotype. Second, it should be able to be associated with the observed natural difference of mapped phenotype. I found AtABCG16 from a map searching candidate genes that cause 4-fold difference of how much defense that plant produced against bacterial pathogen *Pst* DC3000. I provided strong evidence that AtABCG16 assisted in plant basal resistance against *Pst* DC3000 through the ABA response. To further investigate the relation between ABA tolerance and bacterial resistance in wild *Arabidopsis* accessions, I picked eleven of the most resistant wild accessions and eleven of the most susceptible ones. Eden-2, which is one of the resistant lines, didn't germinate. This made the total number of tested lines down to 21. I found that the accessions that had high resistance to *Pst* DC3000 were dramatically more tolerant to exogenous ABA (Figure 21B-C). Indeed, at 6 μM ABA, 70% of the high resistance lines were able to germinate in the presence of 6 μM ABA, whereas only 10% of the low resistance lines germinated in ABA-media (Figure 21B, Table 15). I found a similar trend when applying 1.5 μM ABA (Table 15). This pattern was consistent across all of the three replicates plates tested. Meanwhile, I analyzed the promoter and coding sequence of AtABCG16 using Clustal X. Interestingly, I found the two alleles at AtABCG16 were both old and co-existed for a long time (Figure 21A).

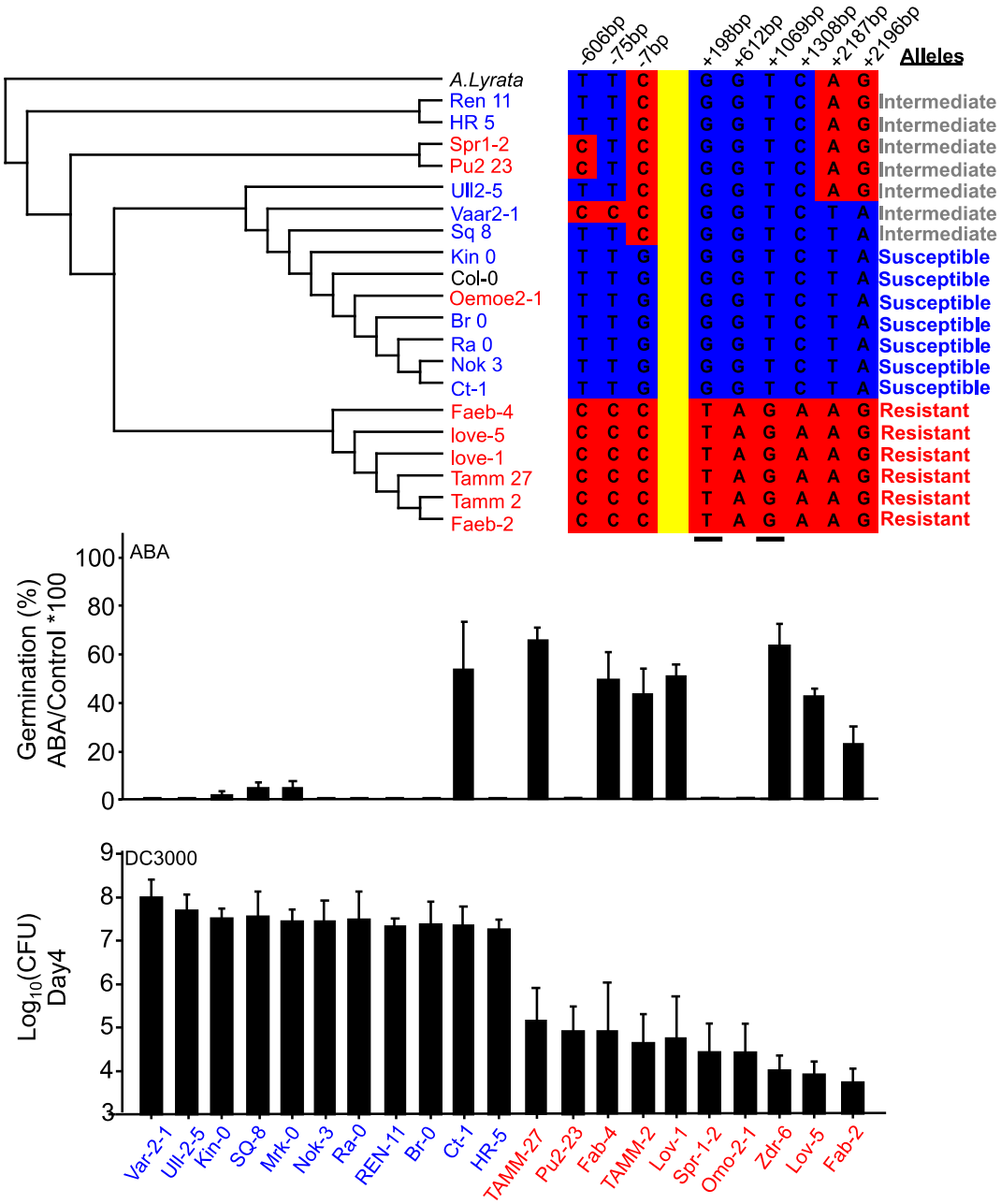


Figure 21. ABA tolerance and bacterial resistance of wild *Arabidopsis* accessions
(A) AtABCG16 gene tree and corresponding SNPs. The tree was made by Clustal X. The two black bars at the bottom represent the two SNPs used in the GWAs map. The other SNPs are collected from the SALK1001 database. Accession names with red and blue colors represent that accessions are either resistant or susceptible, respectively. Alleles with gray, blue and red colors represent what type of alleles the accessions have, intermediate, susceptible and resistant, repectively. **(B)** ABA tolerance of extreme wild accessions. **(C)** Bacterial resistance of extreme wild accessions.

Table 15. ABA tolerance of wild Arabidopsis accessions.

Type	Rank	Line	Accession	Rpp	DC3000log(cfu)	GP 0um	GP 1.5um	GP 6um
S	1	03A	Var-2-1	1	7.15	88.00	0.00	0.0
S	1	03A	Var-2-1	2	8.55	100.00	0.00	0.0
S	1	03A	Var-2-1	3	8.25	100.00	0.00	0.0
S	2	03G	Ull-2-5	1	7.35	88.00	0.00	0.0
S	2	03G	Ull-2-5	2	7.32	92.00	0.00	0.0
S	2	03G	Ull-2-5	3	8.41	100.00	0.00	0.0
S	3	12C	Kin-0	1	7.65	84.00	0.00	0.0
S	3	12C	Kin-0	2	7.81	76.00	0.00	5.0
S	3	12C	Kin-0	3	7.03	92.00	0.00	0.0
S	4	05F	SQ-8	1	6.44	88.00	45.45	8.7
S	4	05F	SQ-8	2	8.00	84.00	33.33	4.5
S	4	05F	SQ-8	3	8.24	88.00	63.64	0.0
S	5	09H	Mrk-0	1	7.39	64.00	62.50	4.5
S	5	09H	Mrk-0	2	7.92	52.00	7.69	9.5
S	5	09H	Mrk-0	3	7.01	72.00	22.22	0.0
S	6	10H	Nok-3	1	6.80	96.00	20.83	0.0
S	6	10H	Nok-3	2	8.35	96.00	0.00	0.0
S	6	10H	Nok-3	3	7.19	84.00	23.81	0.0
S	7	09E	Ra-0	1	6.20	100.00	0.00	0.0
S	7	09E	Ra-0	2	8.37	96.00	8.33	0.0
S	7	09E	Ra-0	3	7.87	100.00	0.00	0.0
S	8	06H	REN-11	1	7.26	60.00	20.00	0.0
S	8	06H	REN-11	2	7.65	60.00	40.00	0.0
S	8	06H	REN-11	3	7.06	68.00	5.88	0.0
S	9	09A	Br-0	1	6.36	100.00	0.00	0.0
S	9	09A	Br-0	2	8.18	100.00	0.00	0.0
S	9	09A	Br-0	3	7.55	100.00	0.00	0.0
S	10	10D	Ct-1	1	6.54	100.00	92.00	60.0
S	10	10D	Ct-1	2	7.38	100.00	96.00	16.0
S	10	10D	Ct-1	3	8.08	100.00	76.00	84.0
S	11	05A	HR-5	1	7.17	96.00	0.00	0.0
S	11	05A	HR-5	2	6.95	88.00	0.00	0.0
S	11	05A	HR-5	3	7.67	84.00	0.00	0.0
R	13	06B	TAMM-27	1	5.89	0.00	0.00	76.2
R	13	06B	TAMM-27	2	5.90	0.00	0.00	60.0
R	13	06B	TAMM-27	3	3.60	0.00	0.00	60.0
R	14	04F	Pu2-23	1	5.34	88.00	45.45	0.0
R	14	04F	Pu2-23	2	5.60	64.00	50.00	0.0
R	14	04F	Pu2-23	3	3.78	60.00	26.67	0.0
R	15	02F	Fab-4	1	4.30	96.00	100.00	36.0
R	15	02F	Fab-4	2	7.06	100.00	84.00	40.0

Table 15. continued.

Type	Rank	Line	Accession	Rpp	DC3000log(cfu)	GP 0um	GP 1.5um	GP 6um
R	15	02F	Fab-4	3	3.30	88.00	100.00	72.0
R	16	06A	TAMM-2	1	4.30	100.00	76.00	50.0
R	16	06A	TAMM-2	2	5.90	100.00	84.00	22.2
R	16	06A	TAMM-2	3	3.60	100.00	72.00	57.1
R	17	02C	Lov-1	1	4.30	100.00	76.00	60.0
R	17	02C	Lov-1	2	6.56	96.00	75.00	45.8
R	17	02C	Lov-1	3	3.30	100.00	88.00	45.8
R	18	03C	Spr-1-2	1	4.30	96.00	4.17	0.0
R	18	03C	Spr-1-2	2	5.60	96.00	4.17	0.0
R	18	03C	Spr-1-2	3	3.30	100.00	4.00	0.0
R	19	03E	Omo-2-1	1	4.30	100.00	68.00	0.0
R	19	03E	Omo-2-1	2	5.60	96.00	29.17	0.0
R	19	03E	Omo-2-1	3	3.30	100.00	32.00	0.0
R	20	04B	Zdr-6	1	4.30	0.00	0.00	79.2
R	20	04B	Zdr-6	2	4.34	0.00	0.00	62.5
R	20	04B	Zdr-6	3	3.30	0.00	0.00	47.8
R	21	02D	Lov-5	1	4.30	96.00	29.17	36.4
R	21	02D	Lov-5	2	4.08	96.00	45.83	47.6
R	21	02D	Lov-5	3	3.30	96.00	58.33	42.9
R	22	02E	Fab-2	1	4.30	88.00	95.45	32.0
R	22	02E	Fab-2	2	3.20	100.00	88.00	8.0
R	22	02E	Fab-2	3	3.60	96.00	75.00	28.0

3.4 DISCUSSION

The first defense of plants following perception of the virulent bacterial pathogen, *Pseudomonas syringae* pv. *tomato* DC3000 (*Pst* DC3000) is to close stomata (Zheng *et al.*, 2012; W., Zeng and He, 2010), and the hormone triggering this stomatal closure is ABA (de Torres Zabala *et al.*, 2007; de Torres Zabala *et al.*, 2009). While ABC transporters have been shown previously to transport ABA (J., Kang *et al.*, 2010; Kuromori *et al.*, 2010; Kuromori *et al.*, 2011), this function has not been previously linked to resistance to bacterial infection. Here, I have shown that an ABCG transporter, AtABCG16, is important for ABA tolerance and also assists in plant basal resistance against *Pst* DC3000. Overexpression of the gene results in improved tolerance of plants to exogenous ABA and improved resistance to bacterial infection, while silencing the gene has the reverse effects. As such, the study provides one of the first demonstrations of a link between an ABC transporter and plant resistance to bacterial pathogens. Previously, only one other ABCG transporter, AtABCG36 (PEN3/PDR8), has been found to be involved in plant defense against pathogens (Kobae *et al.*, 2006), but in that case auxin transport is implicated (Strader and Bartel, 2009). GWAs mapping has attracted a great deal of interest in the literature (Atwell *et al.*, 2010), but examples of successful use of this methodology remain very rare in plants. My study is one of the first to identify a candidate gene directly from a genome-wide association map and find supporting experimental evidence of relevant gene function.

I propose that AtABCG16 assists plant basal resistance to *Pst* DC3000 as a byproduct of transporting ABA (Figure 19G). In this scenario, attack by *Pst* DC3000 begins with the production of PAMPs (Jones and Dangl, 2006) by which the plant initiates systemic acquired

resistance in mesophyll cells (Vlot *et al.*, 2009) as well as stomatal closure (W., Zeng and He, 2010). In turn, the pathogen secretes coronatine which upregulates ABA concentrations and suppresses the production of SA-dependent defenses (Zheng *et al.*, 2012). In resistant plants, I hypothesize that AtABCG16 and possibly other ABC transporters move ABA out of the mesophyll cells, which then relieves suppression of the SA-dependent pathway. Coronatine also has the function through the NACs of forcing stomata to reopen (W., Zeng and He, 2010; Zheng *et al.*, 2012) and it is also possible that resistant plants may move the relocated ABA into the guard cells and thereby keep stomata closed robustly against the coronatine-forced stomatal reopening. I provide three lines of evidence that support this model. First, my analyses of stomata response showed that AtABCG16 helped the plants to keep stomata closed against *Pst* DC3000 and such closure was robust to coronatine when we overexpressed the gene. Second, the GUS assays indicated strong AtABCG16 expression in response to both the bacterial pathogen and hormone treatments. Indeed, AtABCG16 expression is induced by flg22 alone (C. Danna, pers. correspondence). This likely explains why the gene was induced by both *Pst* DC3000 and *Pst* DC3661 in the experiments. Interestingly, when we applied ABA to *AtABCG16pro*::GUS plants, the signal was uniformly detected across the leaf surface, including guard cells and non-guard cells. However, when I applied coronatine, I observed a dramatic response in non-guard cells but only moderate upregulation in guard cells. Given the fact that most ABA is produced in xylem (Boursiac *et al.*, 2013), these results when combined with my stomatal aperture assay suggested that AtABCG16 is likely to respond to ABA directly and coronatine indirectly. Third, overexpression of AtABCG16 showed lower bacteria growth and turning down the gene resulted in higher bacteria growth. Collectively, these results suggest

that AtABCG16 has a primary relationship to ABA regulation, and contributes to resistance to *Pst* DC3000 as a result of this pathogen's ability to manipulate ABA concentrations in cells.

While it is likely that AtABCG16 transports ABA, it is important to note that membrane transport assays (Kuromori and Shinozaki, 2010) have not yet been conducted with this protein to our knowledge. Such assays use labeled ABA as substrate and measure the accumulation within vesicles in the presence and absence of ATP. These assays will be an important step in future work with this gene. For these reasons, it is appropriate to currently conclude that AtABCG16 is involved in ABA tolerance and may indeed be an ABA transporter.

How the AtABCG16 protein interacts with other ABC transporters is a question of some importance. AtABCG16 and the other three ABC transporters implicated in ABA transport (AtABCG22, AtABCG25 and AtABCG40) are all half transporters, meaning that their products must form either homodimers or heterodimers to become functional transporters. Heterodimerisation has been shown previously for ABC transporters (McFarlane et al. 2010) and provides a mechanism by which a large number of transporter structures could be assembled from a relatively small number of genes. It is possible that the transport function of AtABCG16 will involve the proteins from these or other half ABCG transporters. Interestingly, the expression pattern of AtABCG16 is different from AtABCG25, AtABCG22, and AtABCG40 (J., Kang *et al.*, 2010; Kuromori and Shinozaki, 2010; Kuromori *et al.*, 2011). In leaves, AtABCG25 is located on xylem while AtABCG22 and AtABCG40 are expressed mostly on the membrane of guard cells. This would seem to suggest that AtABCG16 may not form heterodimers with these particular proteins and is consistent with the finding that AtABCG22 and AtABCG40 do not appear to interact *in planta* (Kuromori *et al.*, 2011). My analysis of ABCG family genes (Figure 22) indicates that AtABCG16 is most similar to

AtABCG1 and my data suggest that the gene is most likely to have generalized expression across cell types, which may only be turned on under stress, while AtABCG25, AtABCG22 and AtABCG40 may have evolved more specialized functions with restricted expression that is cell-type specific (Ukitsu *et al.*, 2007; Kuromori *et al.*, 2011).

It is worth noting that AtABCG16 exists on the third chromosome as a part of a tandem repeat of three ABC transporters that includes AtABCG17 and AtABCG18. There is no indication however that these genes have overlapping function and indeed, the differential responses of these genes to exogenous ABA, SA, and IAA are striking as shown in the expression neighborhood plot (Figure 13D). Notably, my GWAs map only identifies significant nucleotide polymorphisms in AtABCG16 and not in the other two genes (Table 8). To understand whether particular coding and promoter region differences among wild lines affect the function of this gene, it will be important to test the relative activities of alleles from susceptible and resistant wild lines in a common null background.

In summary, I have provided evidence that AtABCG16 is involved in ABA tolerance and contributes to plant resistance against *Pst* DC3000. We have localized AtABCG16 to the plasma membrane and confirmed microarray data that indicate very low levels of constitutive expression but dramatic induction of this gene by ABA and coronatine and expression in both mesophyll and guard cells. Collectively, these results improve our understanding of basal resistance in *Arabidopsis* and offer novel ABA-related targets for improving the innate resistance of plants to bacterial infection.

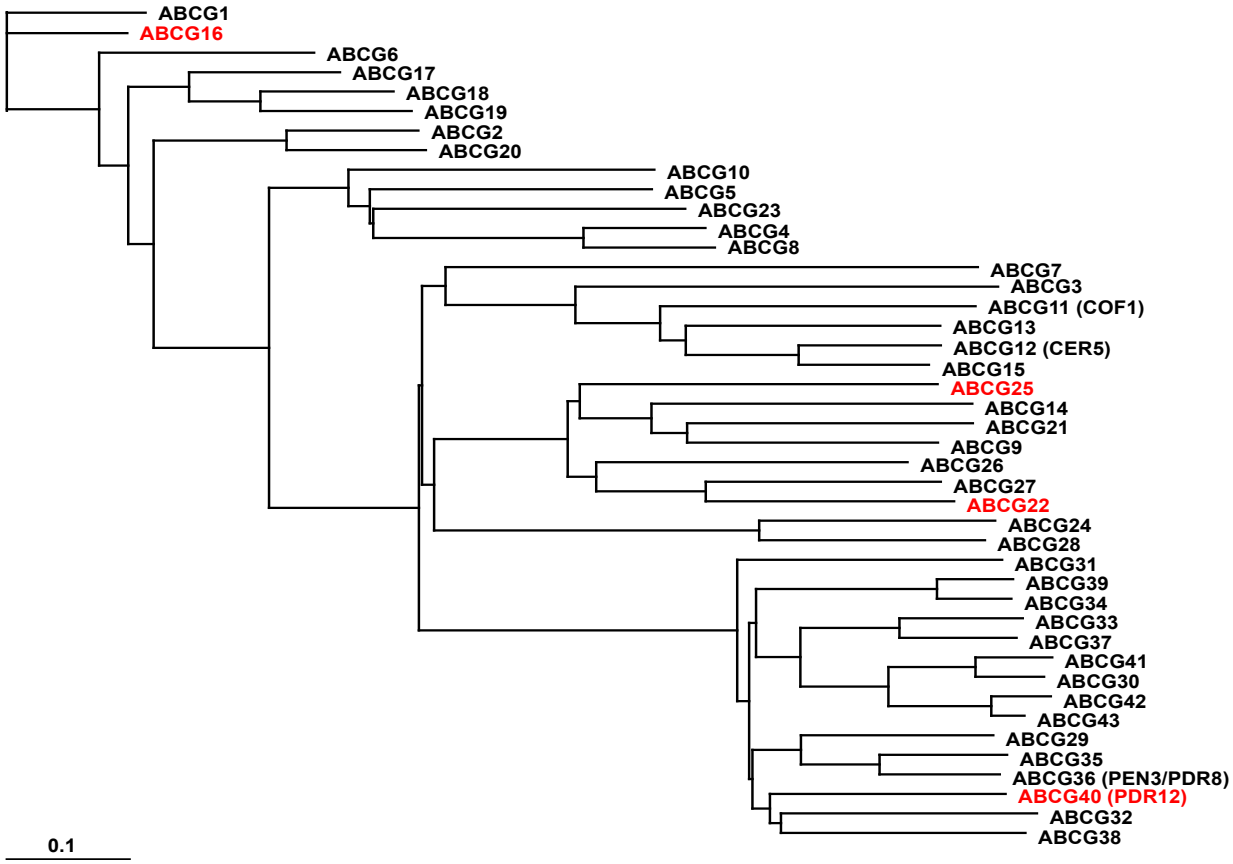


Figure 22. ABCG subfamily of ABC transporters in *Arabidopsis*.

Phylogenetic analysis is taken after previous publications (Ukitsu *et al.*, 2007; Kuromori *et al.*, 2011). Nucleotide sequences were acquired from TAIR with the latest updates in March 2013. (http://www.arabidopsis.org/servlets/Search). Sequence Alignment and phylogeny tree construction were performed by Clustal X. Red color indicates the proteins transporting ABA.

3.5 FUTURE DIRECTION

The next step is to find the allelic effect at the AtABCG16 locus. Why do two alleles persist at this locus? Why don't all accessions maintain high level of resistance to *Pst* DC3000? Our collaborators at the University of Tennessee, Dr. Yanhui Peng and Dr. Neal Stewart are putting the two versions of AtABCG16 into a null background and I can then assess the allelic effect of this gene. Few reports ever tried this kind of effort (Hilscher *et al.*, 2009; Todesco *et al.*, 2010). It will be a great improvement of our understanding of natural variation in plant-bacterial interaction.

3.6 ACKNOWLEDGEMENTS

I thank Dr. Yanhui Peng and Dr. Neal Stewart for creating the transgenic lines and carrying GFP and GUS experiment, which is a big support of the whole work. I thank Steve McCalley, Kaiyuan Pang, Darve Robinson, Justin Seaman, Andy Lariviere, Dr. Muhammad Saleem, and Rongjian Ye for assistance with data collection. I thank the Numberger, Shimada, and Kamiya Labs for providing the microarray data to Nottingham Arabidopsis Stock Centre's microarray database. I thank Dr. DrDiane Cuppels and Dr. James Tambong for providing the COR- bacteria strains. I thank the Salk Institute Genomic Analysis Laboratory for providing the sequence-indexed *Arabidopsis* T-DNA insertion mutants. This research was supported by US National Science Foundation Grants #1051581 (CNS) and #1050138 (M.B.T.).

4.0 GENOME-WIDE ASSOCIATION MAPPING REVEALS A MAJOR FITNESS TRADE-OFF AT A TRICHOME SUPPRESSOR GENE, ETC2, OF *ARABIDOPSIS THALIANA*

A long-standing question in ecology is why variation in defense strategies persists in natural plant populations. While environmental heterogeneity is often implicated in explaining this variation, pleiotropic roles of particular genes are increasingly realized for their potential importance. Using a combination of genome-wide association co-mapping, reverse genetics, and transgenic approaches in *Arabidopsis thaliana*, I show that allelic variation in ENHANCER OF TRIPTYCHON AND CAPRICE 2 (ETC2), a gene known to suppress trichome production in natural accessions, is also associated with heavier seeds, an important trait for reproductive success. This trade-off may explain why individuals with low trichome numbers persist in wild *A. thaliana* and provides evidence for the importance of allelic variation in the maintenance of phenotypic variation in natural populations.

4.1 INTRODUCTION

Given their sessile lifestyle, plants are particularly vulnerable to damage by abiotic and biotic stressors, and possible loss of fitness (Marquis, 1984). In response, plants produce a remarkable diversity of toxic secondary compounds and physical defenses that can reduce this

damage (Bednarek *et al.*, 2010). There is however, abundant variation in natural populations (Weigel, 2012) with individuals often representing a range of investments in defense (Steets *et al.*, 2010; Kawagoe *et al.*, 2011). Why is it that all plants are not heavily defended? High costs of producing these defenses are one possible explanation (Bergelson *et al.*, 1996; Todesco *et al.*, 2010). Such costs can occur when production of defense diverts resources from growth and fecundity (Bazzaz *et al.*, 1987). Identification of the genetic underpinnings of these negative phenotypic correlations between defenses and other traits is of great importance for prediction of the evolutionary trajectories of these populations and, more generally, for understanding the maintenance of biodiversity (Bednarek *et al.*, 2010; Weigel, 2012). However, identification of such defense trade-offs has been a significant challenge, precisely because they must by nature be small (otherwise a plant could not afford them) and because they are typically observable only when those stressors are absent (Kawagoe *et al.*, 2011). To what extent then, is there allelic variation in the genes that influence plant defense and do these genes have pleiotropic roles with respect to other traits, such as those related to reproduction?

4.2 MATERIALS AND METHODS

4.2.1 Genome-wide association mapping

To map the genetic basis of variation in trichome number and mass per seed, I used 168 accessions from the RegMap panel (Horton *et al.*, 2012) of *Arabidopsis thaliana*, collected across the worldwide range (Atwell *et al.*, 2010). The dataset contained 202,967 SNPs derived from the 250K SNP data version3.06.

(https://cynin.gmi.oeaw.ac.at/home/resources/atpolydb/250k-snp-data). All SNPs used in the analysis were diallelic and had the minor nucleotide represented in more than 5% of the accessions. Wilcoxon rank-sum tests were used to calculate the significance of association between each SNP and the phenotypic values using standard methods (Atwell *et al.*, 2010). To handle confounding caused by population structure, I used the standard approach of mixed model analysis known as EMMA (H., M., Kang *et al.*, 2008)(Efficient Mixed-Model Association). All programs are coded in R (http://cran.r-project.org/). I presented log transformed P-values following the standard approach (Atwell *et al.*, 2010; Filiault and Maloof, 2012).

4.2.2 Phenotypic measurements on the RegMap Collection

To assess plant allocation to trichomes, I grew replicate plants of all lines simultaneously under common conditions in growth chambers with unlimited access to water. I assessed the plants every four days for damage by insects or pathogens and documented that there was no evidence of damage or disease. Seedlings of the 168 RegMap *Arabidopsis* accessions were germinated in soil (Premier Pro-Mix BX) in 36-cell flats in a plant growth room at the University of Pittsburgh. Flats were vernalized first at 4^{o}C for seven days, then placed in a growth room at 20^{o}C with short day lighting (12 hr of 350 umol*m^{2}*s^{-1}, 1:1 mixture of sodium: metal halide bulbs). Each plant received a low amount of fertilizer, consisting of 14 ml of full strength Peter's 20:20:20 NPK fertilizer on day 10 of growth. Plants received fertilization only once every two months. To control for the effect of allometric correlations, I selected a single developmental stage (the ninth leaf), which is possessed by all individuals. On day 27, the ninth leaf was identified, removed from the plant, and traced. Leaf area was

determined from the tracing using ImageJ. A leaf disk (Area = 0.29 cm^2) was then removed by hole punch from the center of the leaf blade and counted for adaxial trichome number under a dissecting microscope (Zeiss, Germany). Trichome density was calculated as the trichome number per disk divided by the disk area. Because trichomes are distributed uniformly across the upper surface of mature leaves of *A. thaliana* (Larkin *et al.*, 1996), total number of trichomes on the adaxial surface of the leaf was then calculated by multiplying the trichome number per disk by the ratio of leaf size to disk size. I chose total trichome number per leaf rather than trichome density because area of mature leaves changes through increased cell expansion. I included three independent plants per wild line. To measure mass per seed, I weighed at least three batches, each containing five seeds, for all genotypes using a Mettler Toledo MX5 microbalance at the University of Pittsburgh. Average mass per seed was calculated by dividing the weighed mass by five.

4.2.3 Assessment of Col-0 x Ler recombinant inbred lines

Trichome numbers for the RILs were provided by J. Larkin from measurements on the first and second leaf of plants in a common garden experiment (Larkin *et al.*, 1996). To measure mass per seed on these same lines, I weighed three batches each containing five seeds, as described above. Average mass per seed was calculated by dividing the total mass by three.

4.2.4 T-DNA Insertion Lines and complementation tests

Mass per seed was measured using seeds provided by ABRC. For each T-DNA insertion knockout I weighed five batches each containing five seeds as described above. Average mass per seed was calculated by dividing the total mass by five. The chimeric ETC2

constructs and the transformation effects on trichome numbers are described elsewhere (Hilscher *et al.*, 2009; Alonso *et al.*, 2003). For each of the 10 independent transformants per construct, I weighed a total of 25 T2 seeds.

4.3 RESULT

4.3.1 GWAs mapping found ETC2 locus accounting for a large portion of natural variation of leaf trichome number and suggested a role in differing seed mass in *Arabidopsis thaliana* wild populations

To assess the natural variations in plant defense at a genetic level, I mapped variation in trichome (a defense trait) production in the RegMap global collection of *Arabidopsis thaliana* (Horton *et al.*, 2012). I focused on trichomes because they are specialized epidermal cells that protect plants from enemies and abiotic stressors (Mauricio, 1998; Agrawal *et al.*, 2009) and because of the dramatic variation in trichome numbers that a plant can produce per leaf (Table 16). My resulting genome-wide association (GWAs) map for leaf trichome number (Figure 23A) confirmed the importance of a trichome–related gene, *ETC2*, in which the SNP with highest GWAs score out of about 215,000 SNPs was found (Figure 23A). By calculating the linkage disequilibrium between the adjacent SNPs with that focal SNP, I identified a linkage block containing twenty SNPs presenting six genes (Figure 23B), which in my dataset explained nearly one-fifth of all trichome number variation raging from zero to over 350, in this worldwide population collection. Two major alleles were identified in this region (Table 17, Table 18), hereafter called the "Col-0-like" allele and "Ler-1-like" allele, which are shared by 40 and 32 accessions, respectively. Plants with the Col-0-like allele possessed 176.1 trichomes per leaf on average, which is more than twice the number for plants with the Ler-1-like allele (Figure 23E).

To test the hypothesis whether trichome production has trade off with reproduction traits, I measured the seed mass of those 168 accessions. Seed mass largely determines plant

future performance (Westoby *et al.*, 2002). If trichome production results in a cost to seed allocation, then loci that are associated with elevated trichome production would be expected to also associate with a lower mass per seed. Indeed, I found a dramatic negative association between the nucleotide difference in trichome production and the corresponding difference in mass per seed at this *ETC2* locus (Figure 23C). In other words, SNPs that were associated with trichome numbers were also associated with a low mass per seed. This association was highly significant by Chi-square test ($\chi^2 = 16.36$, df = 1, N = 20, $P < 0.001$). To control for the non-independence of these SNPs, I also assessed the Chi-square value against the empirical distribution of all Chi-square tests for phenotypic comparisons of all possible groups of 20 adjacent SNPs across the genome. The Chi-square value at the *ETC2* locus was in the extreme 0.3% of the most negative associations across the entire genome (Figure 23D). These results suggest that genetic variation at the *ETC2* locus has a unique importance in conferring both elevated trichome numbers and a reduction in mass per seed. Given my results from the Chi-square tests, I predicted that plants with the Col-0-like allele would have lighter seeds than plants with the Ler-1-like allele. Using mixed-model ANOVA corrected by population structure, I found that the mass per seed of plants with the Col-0-like allele was 2.21 μg lower than those of plants having Ler-1-like allele (Figure 23F), which amounted to an 8.5 % net cost of possessing the high trichome allele (T = -2.37, $P = 0.0192$, Table 17). More important, since both trichome number and seed mass data were collected from the same set of plants growing in the same common garden condition with deliberately limited resource, the observed negative correlation suggested the existence of trade-off at the particular genetic locus.

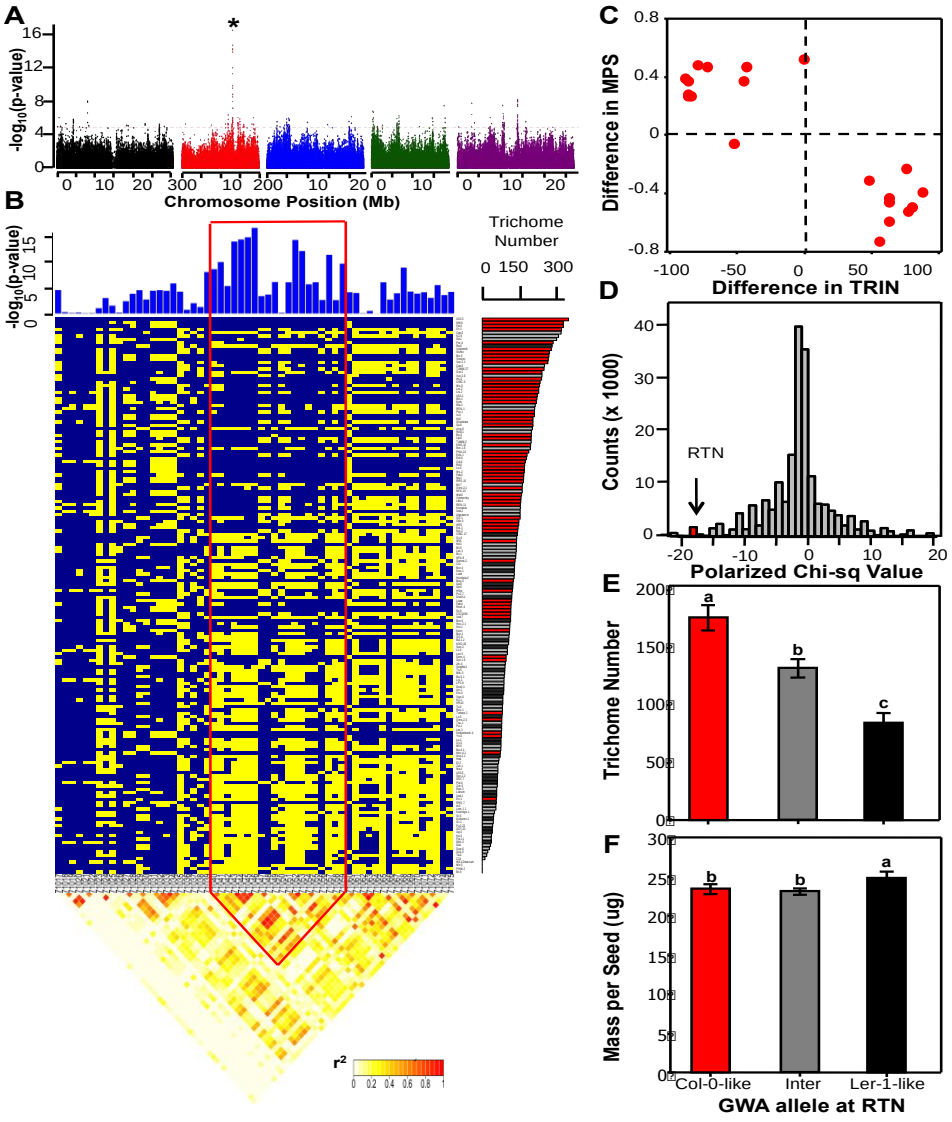


Figure 23. GWAs analysis of leaf trichome number and mass per seed.

(A) The strongest association of SNPs with trichome number was on chromosome 2 near 12.96 Mb and was robust to correction for population structure. (**B)** High correspondence of SNPs and phenotypic values of trichome number at the *ETC2* locus and low correlation with adjacent SNPs. **(C)** Chi-Square association between difference in mass per seed and trichome number with the closest 20 SNPs is significantly negative (χ^2 = 16.36, df = 1, N = 20, $P < 0.001$). (**D)** Empirical distribution of possible Chi-Square values across the full genome. Means (+/- SE) of **(E)** trichome numbers on the ninth leaf and **(C)** population structure controlled mass per seed (ug) for Ler-1-like (N=32), intermediate (N=96), and Col-0-like (N=40) genotypes (Table S3). Different letters indicate significant differences $P < 0.05$.

Table 16. Summary of trhicome number and seed mass for 168 accessions of RegMap Panel.

Accession	Ecotype_id	RTN Allele	Leaf Trichome Number	Mass per Seed(ug)	Structure Group
Ag.0	6897	Ler-1-like	40	29.1	1
Algutsrum	8230	Inter	153	23.7	7
An.1	6898	Inter	87	23.0	1
Ang.0	8254	Inter	202	17.9	8
Ba.1.2	8256	Ler-1-like	107	19.6	4
Ba.4.1	8258	Inter	75	21.7	4
Ba.5.1	8259	Inter	89	20.5	4
Bay.0	6899	Col-0-like	125	26.3	7
Bil.5	6900	Inter	137	24.1	6
Bil.7	6901	Inter	162	23.8	6
Bla.1	8264	Inter	214	26.4	1
Blh.1	8265	Col-0-like	220	19.6	1
Bor.1	5837	Inter	109	24.7	7
Bor.4	6903	Inter	130	22.9	7
Br.0	6904	Ler-1-like	0	24.8	1
Bro.1.6	8231	Inter	194	22.4	7
Bs.1	8270	Inter	136	21.8	1
Bu.0	8271	Inter	198	33.2	1
Bur.0	6905	Col-0-like	114	35.8	4
C24	6906	Inter	11	28.1	1
Can.0	8274	Inter	251	17.7	1
Cen.0	8275	Inter	318	20.2	1
CIBC.17	6907	Ler-1-like	141	25.7	1
CIBC.5	6730	Col-0-like	233	24.4	1
Col.0	6909	Col-0-like	173	22.6	2
CS22491	7438	Inter	115	17.0	5
Ct.1	6910	Ler-1-like	131	23.9	7
Dem.4	8233	Col-0-like	96	27.2	3
Dra.3.1	8283	Ler-1-like	72	24.5	7
DraII.1	8284	Ler-1-like	87	23.5	7
DraIII.1	8285	Ler-1-like	119	19.7	7
Duk	6008	Inter	33	24.0	7
Edi.0	6914	Col-0-like	178	26.8	1
Eds.1	6016	Inter	181	33.8	6
Ei.2	6915	Inter	66	22.2	1
En.1	8290	Ler-1-like	57	23.7	7
Est.1	6916	Col-0-like	145	24.8	2
Fab.2	6917	Inter	168	23.6	6
Fab.4	6918	Inter	117	29.6	6
Fei.0	8215	Col-0-like	327	21.1	8
Fja.1.1	8422	Ler-1-like	36	24.1	7
Ga.0	6919	Ler-1-like	124	23.4	7
Gd.1	8296	Inter	300	23.0	1
Ge.0	8297	Ler-1-like	123	30.1	1
GOT.22	6920	Ler-1-like	40	33.3	4
GOT.7	6921	Ler-1-like	62	32.7	4
Gr.1	8300	Inter	44	18.1	7
Gu.0	6922	Col-0-like	205	22.3	1

Table 16. Continued

Accession	Ecotype_id	RTN Allele	Leaf Trichome Number	Mass per Seed(ug)	Structure Group
Gy.0	8214	Inter	139	25.1	8
Hi.0	8304	Col-0-like	205	24.3	1
Hod.	8235	Inter	67	16.6	7
Hov.2.1	8423	Inter	112	20.0	7
Hov.4.1	8306	Col-0-like	73	21.9	7
Hovdala.2	6039	Inter	128	22.4	7
HR.10	6923	Ler-1-like	83	35.5	8
HR.5	6924	Col-0-like	147	19.8	8
Hs.0	8310	Col-0-like	239	25.5	1
HSm	8236	Inter	122	22.3	7
In.0	8311	Inter	55	22.5	7
Is.0	8312	Col-0-like	210	24.9	1
Jm.0	8313	Inter	93	19.7	7
Ka.0	8314	Inter	37	22.7	7
Kas.1	8424	Inter	129	31.9	5
Kavlinge.1	8237	Inter	49	20.0	7
Kelsterbach.4	8420	Col-0-like	77	19.8	1
Kent	8238	Col-0-like	110	20.0	1
Kin.0	6926	Ler-1-like	86	23.3	8
Kni.1	6040	Inter	111	21.9	4
KNO.10	6927	Col-0-like	195	27.9	3
KNO.18	6928	Inter	106	27.2	1
Koln	8239	Col-0-like	218	19.6	1
Kondara	6929	Inter	154	22.9	5
Kulturen.1	8240	Ler-1-like	45	24.8	7
KZ.1	6930	Inter	83	18.5	5
KZ.9	6931	Inter	307	17.0	5
Lc.0	8323	Col-0-like	170	16.4	1
Ler.1	6932	Ler-1-like	78	24.0	7
Liarum	8241	Ler-1-like	58	18.2	7
Lillo.1	8242	Inter	155	25.1	7
Lip.0	8325	Inter	197	23.4	7
Lis.1	8326	Inter	88	20.2	7
Lis.2	8222	Inter	228	19.3	7
Lisse	8430	Col-0-like	118	28.5	1
LL.0	6933	Inter	101	19.2	1
Lm.2	8329	Inter	228	23.2	1
Lom.1.1	6042	Inter	55	20.2	7
Lov.1	6043	Inter	137	27.8	6
Lov.5	6046	Inter	101	28.8	6
LP2.6	7521	Inter	88	23.0	7
Lu.1	8334	Inter	76	20.2	7
Lund	8335	Ler-1-like	129	23.5	4
Lz.0	6936	Ler-1-like	80	28.3	1
Mir.0	8337	Ler-1-like	0	19.8	1
Mrk.0	6937	Inter	90	32.3	1
Mt.0	6939	Ler-1-like	76	21.0	1
Mz.0	6940	Col-0-like	269	23.6	1

Table 16. Continued

Accession	Ecotype_id	RTN Allele	Leaf Trichome Number	Mass per Seed(ug)	Structure Group
Na.1	8343	Col-0-like	167	24.4	1
Nd.1	6942	Col-0-like	138	24.6	1
NFA.10	6943	Col-0-like	161	22.3	8
NFA.8	6944	Inter	135	22.8	8
Nok.3	6945	Inter	36	24.7	1
NW.0	8348	Col-0-like	327	22.7	1
Nyl.2	6064	Inter	143	31.0	6
Omo.2.1	7518	Inter	161	20.7	7
Omo.2.3	7519	Inter	79	19.4	7
Or.1	6074	Inter	320	21.6	4
Ost.0	8351	Inter	241	26.1	6
Oy.0	6946	Ler-1-like	76	24.5	7
Pa.1	8353	Inter	78	17.2	1
Per.1	8354	Inter	210	23.1	5
PHW.2	8243	Inter	0	23.7	1
Pla.0	8357	Inter	61	19.7	1
PNA.10	7526	Col-0-like	187	27.1	3
Pro.0	8213	Inter	286	23.9	1
Pu2.23	6951	Inter	42	23.5	7
Pu2.7	6956	Inter	120	21.9	7
Ra.0	6958	Ler-1-like	282	26.2	1
Rak.2	8365	Ler-1-like	58	20.8	7
Rd.0	8366	Col-0-like	171	25.3	2
Rd.0.1	8411	Col-0-like	198	22.3	2
REN.1	6959	Inter	211	26.0	8
REN.11	6960	Inter	154	22.4	8
Rev.1	8369	Inter	81	19.8	4
RRS.10	7515	Col-0-like	166	27.0	3
RRS.7	7514	Col-0-like	57	27.2	1
Rsch.4	8374	Col-0-like	116	17.8	1
San.2	8247	Ler-1-like	102	24.8	4
Sanna.2.	8376	Inter	132	29.5	6
Sap.0	8378	Inter	29	19.2	7
Sav.0	8412	Inter	22	23.9	7
Seattle.0	8245	Inter	92	18.9	8
Shahdara	6962	Inter	205	22.8	5
Sorbo	6963	Inter	270	27.1	5
Spr.1.2	6964	Ler-1-like	62	25.1	7
Spr.1.6	6965	Inter	96	20.4	7
SQ.1	6966	Inter	153	18.1	8
SQ.8	6967	Inter	108	20.3	1
Sr.5	8386	Inter	46	19.2	7
St.0	8387	Col-0-like	115	20.2	1
Stw.0	8388	Inter	148	20.6	7
Ta.0	8389	Inter	18	23.8	7
TAMM.2	6968	Inter	197	21.4	6
TAMM.27	6969	Inter	249	23.0	6
Tottarp.2	6243	Inter	81	18.0	7

Table 16. Continued

Accession	Ecotype_id	RTN Allele	Leaf Trichome Number	Mass per Seed(ug)	Structure Group
Ts.5	6971	Ler-1-like	91	28.9	1
Tsu.1	6972	Inter	79	24.3	1
Tu.0	8395	Inter	81	26.6	1
Ull.1.1	8426	Col-0-like	222	22.5	7
Ull.2.3	6973	Col-0-like	348	22.7	1
Ull.2.5	6974	Inter	63	19.3	7
Uod.1	6975	Inter	58	20.2	7
Uod.2	8428	Col-0-like	153	19.5	2
Uod.7	6976	Ler-1-like	114	23.5	7
Van.0	6977	Ler-1-like	83	22.5	1
Var.2.1	7516	Inter	265	29.4	4
Var.2.6	7517	Inter	241	34.3	4
Vastervik	9058	Inter	279	23.0	4
Vimmerby	8249	Inter	156	24.1	7
Vinslov	9057	Col-0-like	266	20.0	4
Wa.1	6978	Inter	63	30.8	7
Wei.0	6979	Col-0-like	159	26.5	1
Wil.1.Dean	100000	Inter	0	20.7	7
Ws.0	6980	Inter	231	20.4	7
Ws.2	6981	Col-0-like	169	16.9	7
Wt.5	6982	Inter	138	23.3	1
Yo.0	6983	Col-0-like	77	24.4	3
Zdr.1	6984	Inter	63	26.7	7
Zdr.6	6985	Ler-1-like	61	23.4	7

Table 17. Classification of two major alleles at the *ETC2* locus.
Orange means SNPs are in *ETC2*. Yellow means SNPs are not in *ETC2* but they are candidate SNPs. Black means non-significant SNPs.

Accession	RTN Allele	71040	71041	71042	71043	71044	71045	71046	71047	71048	71049	71050	71051	71052	71053	71054	71055	71056	71057	71058	71059
Bay.0	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Blh.1	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Bur.0	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
CIBC.5	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Col.0	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Dem.4	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Edi.0	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Est.1	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Fei.0	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Gu.0	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Hi.0	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Hov.4.1	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
HR.5	Col-0-like	T	G	C	A	T	G	G	C	T	A	T	G	G	T	G	G	C	A	G	A
Hs.0	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Is.0	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Kelsterbach.4	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Kent	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
KNO.10	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Koln	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Lc.0	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Lisse	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Mz.0	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Na.1	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Nd.1	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
NFA.10	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
NW.0	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
PNA.10	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Rd.0	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Rd.0.1	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
RRS.10	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
RRS.7	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Rsch.4	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
St.0	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Ull.1.1	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Ull.2.3	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Uod.2	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Vinslov	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Wei.0	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Ws.2	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Yo.0	Col-0-like	T	G	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Algutsrum	Inter	T	G	C	A	T	G	G	T	C	A	T	T	G	T	T	G	T	A	C	A
An.1	Inter	C	C	C	G	C	A	A	C	T	A	A	T	C	G	T	T	C	G	C	T
Ang.0	Inter	C	C	T	G	C	A	G	T	T	C	A	T	C	G	T	T	C	G	C	T
Ba.4.1	Inter	T	C	C	G	C	A	A	C	T	A	A	T	C	G	G	G	C	G	C	T
Ba.5.1	Inter	T	C	C	G	C	A	A	C	T	A	A	T	C	G	G	G	C	G	C	T
Bil.5	Inter	T	C	C	G	C	A	A	C	T	A	T	T	C	G	G	G	C	G	C	T
Bil.7	Inter	T	C	C	G	C	A	A	C	T	A	T	T	C	G	G	G	C	G	C	T
Bla.1	Inter	C	C	C	A	T	G	G	T	C	A	T	G	G	T	T	G	C	A	C	T
Bor.1	Inter	C	C	C	G	C	A	A	C	T	A	T	T	C	G	G	G	C	G	G	T
Bor.4	Inter	C	C	C	G	C	A	A	C	T	A	T	T	C	G	T	G	C	G	C	T
Bro.1.6	Inter	C	C	C	G	C	A	A	C	T	A	A	T	C	G	G	G	C	G	G	T
Bs.1	Inter	T	C	C	G	C	A	A	C	T	A	A	T	C	G	G	G	C	G	C	T
Bu.0	Inter	T	G	C	A	T	G	G	C	T	C	A	G	G	T	G	G	C	A	G	A
C24	Inter	C	C	T	G	C	A	A	C	T	A	A	T	C	G	G	G	C	G	C	A
Can.0	Inter	C	C	C	A	T	G	G	C	T	A	T	G	G	G	G	G	C	A	G	A
Cen.0	Inter	C	C	T	G	C	A	A	C	T	A	A	T	C	G	T	T	C	G	C	T
CS22491	Inter	T	G	C	A	T	G	G	T	C	A	T	T	G	T	G	G	T	A	C	A
Duk	Inter	C	C	C	G	C	A	A	C	T	C	A	T	C	G	T	T	C	G	C	T
Eds.1	Inter	T	G	C	A	T	G	G	T	C	A	T	T	G	T	G	G	T	A	C	A
Ei.2	Inter	C	C	C	G	C	A	A	C	T	C	T	T	C	G	T	T	C	G	G	A
Fab.2	Inter	C	C	C	A	T	G	G	T	C	A	T	G	G	T	G	G	C	A	G	A
Fab.4	Inter	C	C	C	A	T	G	G	T	C	A	T	G	G	T	G	G	C	A	G	A
Gd.1	Inter	T	G	C	A	C	G	G	C	T	A	A	G	G	T	G	G	C	G	G	A
Gr.1	Inter	C	C	C	G	C	A	A	C	T	A	T	T	C	G	G	G	C	G	G	T
Gy.0	Inter	C	C	T	G	C	G	A	T	T	C	A	T	C	G	T	T	C	G	C	T
Hod.	Inter	C	C	C	A	C	A	A	C	T	A	A	T	C	G	G	G	C	G	G	T
Hov.2.1	Inter	T	C	C	G	T	G	G	C	T	A	A	G	C	T	G	T	C	A	G	A
Hovdala.2	Inter	C	C	C	G	C	A	A	C	T	A	A	T	C	G	G	G	T	A	C	A
HSm	Inter	C	G	C	A	T	G	G	T	C	A	T	G	G	T	G	G	C	A	C	A
In.0	Inter	C	C	C	G	C	A	A	C	T	A	T	T	C	G	T	T	C	G	C	T
Jm.0	Inter	C	C	C	A	T	G	G	T	C	A	T	G	G	T	G	G	T	A	C	T
Ka.0	Inter	C	C	C	A	C	G	A	T	T	A	T	T	C	G	T	G	C	G	G	T
Kas.1	Inter	C	C	C	A	T	G	G	T	C	A	T	G	G	T	G	G	C	A	C	T
Kavlinge.1	Inter	T	G	C	G	C	A	A	C	T	A	T	T	C	G	G	G	T	A	C	A
Kni.1	Inter	T	G	C	G	C	A	A	C	T	A	T	T	C	G	G	G	C	G	C	T
KNO.18	Inter	T	G	C	G	C	A	A	C	T	A	A	T	C	G	G	G	C	G	C	T
Kondara	Inter	T	G	C	A	T	G	G	T	C	A	T	T	G	T	G	G	T	A	C	A
KZ.1	Inter	C	C	C	A	T	A	A	C	C	A	T	G	G	T	G	G	C	A	C	T
KZ.9	Inter	T	G	C	A	T	G	G	T	C	A	T	T	G	G	G	G	T	A	C	A
Lillo.1	Inter	T	G	C	A	T	G	G	T	C	A	T	T	G	T	G	G	T	A	C	A
Lip.0	Inter	T	G	C	A	T	G	G	T	C	A	T	T	G	G	G	G	T	G	C	A
Lis.1	Inter	C	C	C	G	C	A	A	C	T	A	T	T	C	G	G	G	C	G	G	T
Lis.2	Inter	T	G	C	A	T	G	G	T	C	A	T	T	G	T	G	G	T	A	C	A
LL.0	Inter	C	C	C	G	C	A	A	C	T	C	A	T	C	G	T	G	C	G	C	T
Lm.2	Inter	C	C	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Lom.1.1	Inter	C	C	C	G	C	A	A	C	T	C	A	T	C	G	T	T	C	G	C	T
Lov.1	Inter	T	C	C	G	C	A	A	C	T	A	A	T	C	G	G	G	C	G	C	T
Lov.5	Inter	T	C	C	G	C	A	A	C	T	A	T	T	C	G	G	G	C	G	C	T

Table17. continued

Accession	RTN Allele	71040	71041	71042	71043	71044	71045	71046	71047	71048	71049	71050	71051	71052	71053	71054	71055	71056	71057	71058	71059
LP2.6	Inter	C	G	C	A	T	G	G	T	C	A	T	G	G	T	G	G	C	A	C	T
Lu.1	Inter	C	C	C	G	C	A	A	C	T	A	T	T	C	G	G	G	C	G	C	A
Mrk.0	Inter	T	C	C	G	C	A	A	C	T	A	A	T	C	G	G	G	C	G	C	T
NFA.8	Inter	C	C	T	G	C	G	G	C	T	C	A	T	C	G	T	T	C	G	C	T
Nok.3	Inter	C	C	T	G	C	A	A	C	T	C	T	T	C	G	T	G	C	G	C	T
Nyl.2	Inter	T	G	C	A	T	G	G	T	C	A	T	T	G	T	G	G	T	A	C	A
Omo.2.1	Inter	T	G	C	A	T	A	G	T	C	A	T	T	G	T	G	G	T	A	C	A
Omo.2.3	Inter	T	C	C	G	C	A	A	C	T	A	T	T	C	G	G	G	C	G	G	T
Or.1	Inter	T	G	C	A	T	G	G	T	C	A	T	T	G	T	G	G	T	A	C	A
Ost.0	Inter	C	C	C	G	C	A	A	C	T	A	A	T	C	G	G	G	C	G	G	T
Pa.1	Inter	C	C	T	G	C	A	A	C	T	A	A	T	C	G	G	G	C	G	C	A
Per.1	Inter	C	G	C	A	T	G	G	T	C	A	T	T	G	T	G	G	T	A	C	A
PHW.2	Inter	C	C	T	G	C	A	A	T	C	A	T	G	G	T	G	G	T	A	C	T
Pla.0	Inter	C	C	C	A	T	G	G	T	C	A	T	G	G	T	T	G	C	A	C	T
Pro.0	Inter	C	C	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Pu2.23	Inter	C	C	C	G	C	A	A	C	T	A	A	T	C	G	G	G	C	G	G	T
Pu2.7	Inter	C	C	C	A	T	G	G	T	C	A	T	G	G	T	T	G	C	A	C	T
REN.1	Inter	T	G	C	A	T	G	G	C	T	A	T	G	G	T	T	G	C	A	C	T
REN.11	Inter	T	G	C	A	T	G	G	C	T	A	T	G	G	T	T	G	C	A	C	T
Rev.1	Inter	T	G	C	G	C	A	A	C	T	A	A	T	C	G	G	G	T	A	C	A
Sanna.2.	Inter	T	G	C	A	T	G	G	T	C	A	T	T	G	T	G	G	T	A	C	A
Sap.0	Inter	C	C	C	G	C	A	A	C	T	C	A	T	C	G	T	T	C	G	C	T
Sav.0	Inter	C	C	C	G	C	A	A	C	T	A	A	T	C	G	G	G	C	G	C	A
Seattle.0	Inter	C	C	T	G	C	G	A	T	T	C	T	T	C	G	T	T	C	G	C	T
Shahdara	Inter	T	G	C	A	T	G	G	T	C	A	T	T	G	T	G	G	T	A	C	A
Sorbo	Inter	T	G	C	A	T	G	G	T	C	A	T	T	G	T	G	G	T	A	C	A
Spr.1.6	Inter	C	G	C	A	T	G	G	T	C	A	T	G	G	T	G	G	C	A	C	A
SQ.1	Inter	C	C	T	G	C	G	G	T	T	C	T	T	C	G	T	T	C	G	C	T
SQ.8	Inter	C	C	T	G	C	A	A	C	T	A	A	G	G	T	G	G	C	A	G	A
Sr.5	Inter	C	C	C	G	C	A	A	C	T	C	T	T	C	G	T	T	C	G	C	T
Stw.0	Inter	T	C	C	A	T	G	G	T	C	A	T	T	G	T	G	G	T	A	C	A
Ta.0	Inter	C	C	C	G	C	A	A	C	T	A	A	T	C	G	G	G	C	G	G	T
TAMM.2	Inter	T	G	C	A	T	G	G	T	C	A	T	T	G	T	G	G	T	A	C	A
TAMM.27	Inter	T	G	C	A	T	G	G	T	C	A	T	T	G	T	G	G	T	A	C	A
Tottarp.2	Inter	C	C	C	G	C	A	A	C	T	C	A	T	C	G	T	T	C	G	C	T
Tsu.1	Inter	C	C	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Tu.0	Inter	C	C	C	A	T	G	G	C	T	A	A	G	G	T	G	G	C	A	G	A
Ull.2.5	Inter	T	G	C	G	C	A	A	C	T	A	A	T	C	G	G	G	T	A	C	A
Uod.1	Inter	C	C	C	G	C	A	A	C	T	A	T	T	C	G	G	G	C	G	G	T
Var.2.1	Inter	C	G	C	A	T	G	G	T	C	A	T	T	G	T	G	G	T	A	C	A
Var.2.6	Inter	C	C	C	A	T	G	G	T	C	A	T	G	G	T	G	G	C	A	G	A
Vastervik	Inter	T	G	C	A	T	G	G	T	C	A	T	T	G	T	G	G	T	A	C	A
Vimmerby	Inter	T	G	C	A	T	G	G	T	C	A	T	T	G	T	G	G	T	A	C	A
Wa.1	Inter	C	C	C	G	C	A	A	C	T	A	A	T	C	G	G	G	C	G	G	T
Wil.1.Dean	Inter	T	G	T	G	C	A	A	C	T	A	A	T	C	G	G	G	C	G	G	A
Ws.0	Inter	C	C	C	A	T	G	G	T	C	A	T	G	G	T	G	G	C	A	C	A
Wt.5	Inter	C	C	T	G	C	A	A	C	T	C	T	T	C	G	T	G	C	G	C	T
Zdr.1	Inter	C	C	C	G	C	A	A	C	T	A	A	T	C	G	G	G	C	G	G	T
Ag.0	Ler-1-like	C	C	T	G	C	A	A	C	T	C	A	T	C	G	T	T	C	G	C	T
Ba.1.2	Ler-1-like	C	C	T	G	C	A	A	C	T	C	T	T	C	G	T	T	C	G	C	T
Br.0	Ler-1-like	C	C	T	G	C	A	A	C	T	C	T	T	C	G	T	T	C	G	C	T
CIBC.17	Ler-1-like	C	C	T	G	C	A	A	C	T	C	A	T	C	G	T	T	C	G	C	T
Ct.1	Ler-1-like	C	C	T	G	C	A	A	C	T	C	T	T	C	G	T	T	C	G	C	T
Dra.3.1	Ler-1-like	C	C	T	G	C	A	A	C	T	C	A	T	C	G	T	T	C	G	C	T
DraII.1	Ler-1-like	C	C	T	G	C	A	A	C	T	C	T	T	C	G	T	T	C	G	C	T
DraIII.1	Ler-1-like	C	C	T	G	C	A	A	C	T	C	A	T	C	G	T	T	C	G	C	T
En.1	Ler-1-like	C	C	T	G	C	A	A	C	T	C	A	T	C	G	T	T	C	G	C	T
Fja.1.1	Ler-1-like	C	C	T	G	C	A	A	C	T	C	A	T	C	G	T	T	C	G	C	T
Ga.0	Ler-1-like	C	C	T	G	C	A	A	C	T	C	T	T	C	G	T	T	C	G	C	T
Ge.0	Ler-1-like	C	C	T	G	C	A	A	C	T	C	A	T	C	G	T	T	C	G	C	T
GOT.22	Ler-1-like	C	C	T	G	C	A	A	C	T	C	T	T	C	G	T	T	C	G	C	T
GOT.7	Ler-1-like	C	C	T	G	C	A	A	C	T	C	A	T	C	G	T	T	C	G	C	T
HR.10	Ler-1-like	C	C	T	G	C	A	A	T	T	C	T	T	C	G	T	T	C	G	C	T
Kin.0	Ler-1-like	C	C	T	G	C	A	A	C	T	C	A	T	C	G	T	T	C	G	C	T
Kulturen.1	Ler-1-like	C	C	T	G	C	A	A	C	T	C	T	T	C	G	T	T	C	G	C	T
Ler.1	Ler-1-like	C	C	T	G	C	A	A	C	T	C	T	T	C	G	T	T	C	G	C	T
Liarum	Ler-1-like	C	C	T	G	C	A	A	C	T	C	T	T	C	G	T	T	C	G	C	T
Lund	Ler-1-like	C	C	T	G	C	A	A	C	T	C	A	T	C	G	T	T	C	G	C	T
Lz.0	Ler-1-like	C	C	T	G	C	A	A	C	T	C	A	T	C	G	T	T	C	G	C	T
Mir.0	Ler-1-like	C	C	T	G	C	A	A	C	T	C	T	T	C	G	T	T	C	G	C	T
Mt.0	Ler-1-like	C	C	T	G	C	A	A	C	T	C	A	T	C	G	T	T	C	G	C	T
Oy.0	Ler-1-like	C	C	T	G	C	A	A	C	T	C	T	T	C	G	T	T	C	G	C	T
Ra.0	Ler-1-like	C	C	T	G	C	A	A	C	T	C	T	T	C	G	T	T	C	G	C	T
Rak.2	Ler-1-like	C	C	T	G	C	A	A	C	T	C	T	T	C	G	T	T	C	G	C	T
San.2	Ler-1-like	C	C	T	G	C	A	A	C	T	C	A	T	C	G	T	T	C	G	C	T
Spr.1.2	Ler-1-like	C	C	T	G	C	A	A	C	T	C	A	T	C	G	T	T	C	G	C	T
Ts.5	Ler-1-like	C	C	T	G	C	A	A	C	T	C	T	T	C	G	T	T	C	G	C	T
Uod.7	Ler-1-like	C	C	T	G	C	A	A	C	T	C	T	T	C	G	T	T	C	G	C	T
Van.0	Ler-1-like	C	C	T	G	C	A	A	C	T	C	T	T	C	G	T	T	C	G	C	T
Zdr.6	Ler-1-like	C	C	T	G	C	A	A	C	T	C	T	T	C	G	T	T	C	G	C	T

Table 18. GWAs result at the *ETC2* locus.

SNPs are sorted by their Wilcoxon score in a descending order. Orange means SNPs are in *ETC2*. Yellow means SNPs are not in *ETC2* but they are candidate SNPs.

Gene	Chr	SNP Position	SNP Number	SNP1	SNP2	Wilcoxon.TRIN	EMMA.TRIN	Diff in TRIN	Diff in MPS
AT2G30420	2	12961873	71046	A	G	16.361	12.282	-89	0.4
AT2G30420	2	12961795	71045	A	G	14.558	9.820	-87	0.38
AT2G30410	2	12960512	71044	C	T	14.142	9.519	-87	0.29
NA	2	12966970	71052	C	G	14.050	9.390	-87	0.28
AT2G30410	2	12960334	71043	A	G	13.753	9.964	87	-0.38
NA	2	12967408	71053	G	T	11.930	6.326	-85	0.28
AT2G30430	2	12968411	71057	A	G	11.201	6.308	79	-0.485
NA	2	12958593	71041	C	G	9.822	6.145	-80	0.49
AT2G30432	2	12969555	71059	A	T	9.506	6.656	76.5	-0.515
NA	2	12956368	71040	C	T	8.452	4.649	-73	0.48
NA	2	12965938	71049	A	C	5.994	4.146	62	-0.58
AT2G30430	2	12968196	71055	G	T	5.955	3.393	62	-0.45
NA	2	12966482	71051	G	T	5.946	3.088	75	-0.22
NA	2	12967708	71054	G	T	5.564	3.215	62	-0.42
AT2G30410	2	12959770	71042	C	T	5.228	3.738	55	-0.72
NA	2	12964216	71048	C	T	3.591	2.539	47	-0.3
NA	2	12963312	71047	C	T	3.300	2.026	-46	0.38
AT2G30430	2	12968316	71056	C	T	2.549	2.245	-44	0.48
AT2G30432	2	12969304	71058	C	G	2.529	1.542	-53	-0.05
NA	2	12966368	71050	A	T	0.112	0.046	-1	0.53

4.3.2 Phenotypic and genotypic data from RILs support the trade-off between trichome number and seed mass at ETC2 locus

To control further for genetic background, I assessed the effects of the alleles at *ETC2* using a set of recombinant inbred lines (RILs) from the cross of Ler x Col-0 (C., Lister and C., Dean, 1993). RILs are superior for this purpose, owing to the fact that each RIL is a unique combination of alleles from the two parents, allowing the effects of allelic differences at *ETC2* to be tested independently of the rest of the genome. The closest marker to *ETC2*, m283C showed a significant positive relation between the Col-0 allele and high trichome numbers (Figure 24A, T = -9.543, $P < 0.001$, trichome data (provided by Dr. John Larkin). At the same marker, a highly significant reduced weight of 1.37 μg per seed (5.6 %) was present for RILs that possessed the Col-0 allele (Figure 24B, T = 5.436, $P < 0.001$, Table 19). This effect was unique to the *ETC2* region in my GWA analysis (Figure 24C). In fact, the most significant loci on the other chromosomes, all were in the opposite direction, accounting for the overall lower seed weight for the Ler parent relative to the Col-0 parent (Figure 24D) (17.1 versus 21.3 ug, $F_{1,58} = 124.7$, $P<0.001$).

4.3.3 T-DNA lines at ETC2 locus showed decrease of mass per seed

Finally, to experimentally assess the relationship between mass per seed and trichome number at the *ETC2* locus, I measured mass per seed of all available T-DNA insertion knockouts (Alonso *et al.*, 2003) of the focal gene *ETC2* since it contains the most significant SNPs of the entire genome. The knockouts of *ETC2* included SALK and GABI-KAT (German Plant Genomics Program 'Genome Analysis in Biological Systems-Knockout *Arabidopsis thaliana*') lines (Alonso *et al.*, 2003; Kleinboelting *et al.*, 2012), generated by two independent

institutes. Given the known role of *ETC2* in trichome regulation (Hilscher *et al.*, 2009; Wang *et al.*, 2007), I hypothesized that the loss-of-function mutants of *ETC2* gene would exhibit low mass per seed. Supporting this prediction, SALK knockout (040390C) had a reduction of 3.33 ug (16.3 %) in seed mass relative to the background line (CS70000). Interestingly, all four individual lines of a GABI-KAT knockout (GK-105E10) had an average of 2.48 ug lower seed mass (12.1%) compared to CS70000 as well, where three of them exhibited significant lower values (Figure 25, Table 20).

4.3.4 Complementing knockout line with a functional copy of ETC2 decreased trichome number and increased seed mass

To further test the gene function of ETC2 on both trichome numbers and seed mass, I analyzed the seed mass of the transgenic plants and found that all *etc2*- lines with a functional copy of *ETC2* have significantly higher seed mass as compared to their null background and the mutants with empty vectors, which were created by Dr. Hilscher and provided by Dr. Hauser (Hilscher *et al.*, 2009). In Hilscher et al. 2009, they introduced chimeric copy of natural ETC2 alleles back into null background (SALK_040390C). What they found was that all of their transgenic lines developed fewer trichomes (Figure 26, Table 21). Since ETC2 was a trichome suppressor, this was strong evidence to support the function of ETC2 in trichome development. Through analyzing the seed mass of these transgenic lines, I found they all showed higher seed mass compered with the null background. Although the construction of transgenic lines were not ideal for this experiment, this still was strong evidence to support the gene function of ETC2 in both trichome and seed mass.

Collectively, my three independent approaches (mixed-model ANOVA, analyzing RILs and weighing T-DNA knockouts) each revealed significant costs of trichome production (8.6 %, 5.6 %, 16.3 % for SALK lines and 12.1 % for GABI-KAT lines, respectively) in terms of reductions of seed mass. The data support earlier findings of costs of trichomes in *A. thaliana* (Mauricio, 2001; Züst *et al.*, 2008) and other plants (Ågren and Schemske, 1993; Hare *et al.*, 2003; Kärkkäinen *et al.*, 2004) and pinpoint for the first time that a well-known trichome related gene (*ETC2*) have been found to be involved in regulating fitness related trait (seed mass) in the opposite direction. It clearly suggests the reason why the high- and low- trichome related alleles could persist in the wild Arabidopsis accessions at this locus. Researchers have recently documented exceptional variation in leaf trichome numbers among individuals within and among populations (Steets *et al.*, 2010; Kawagoe *et al.*, 2011). My results are likely to explain why such low trichome phenotypes are common in wild populations.

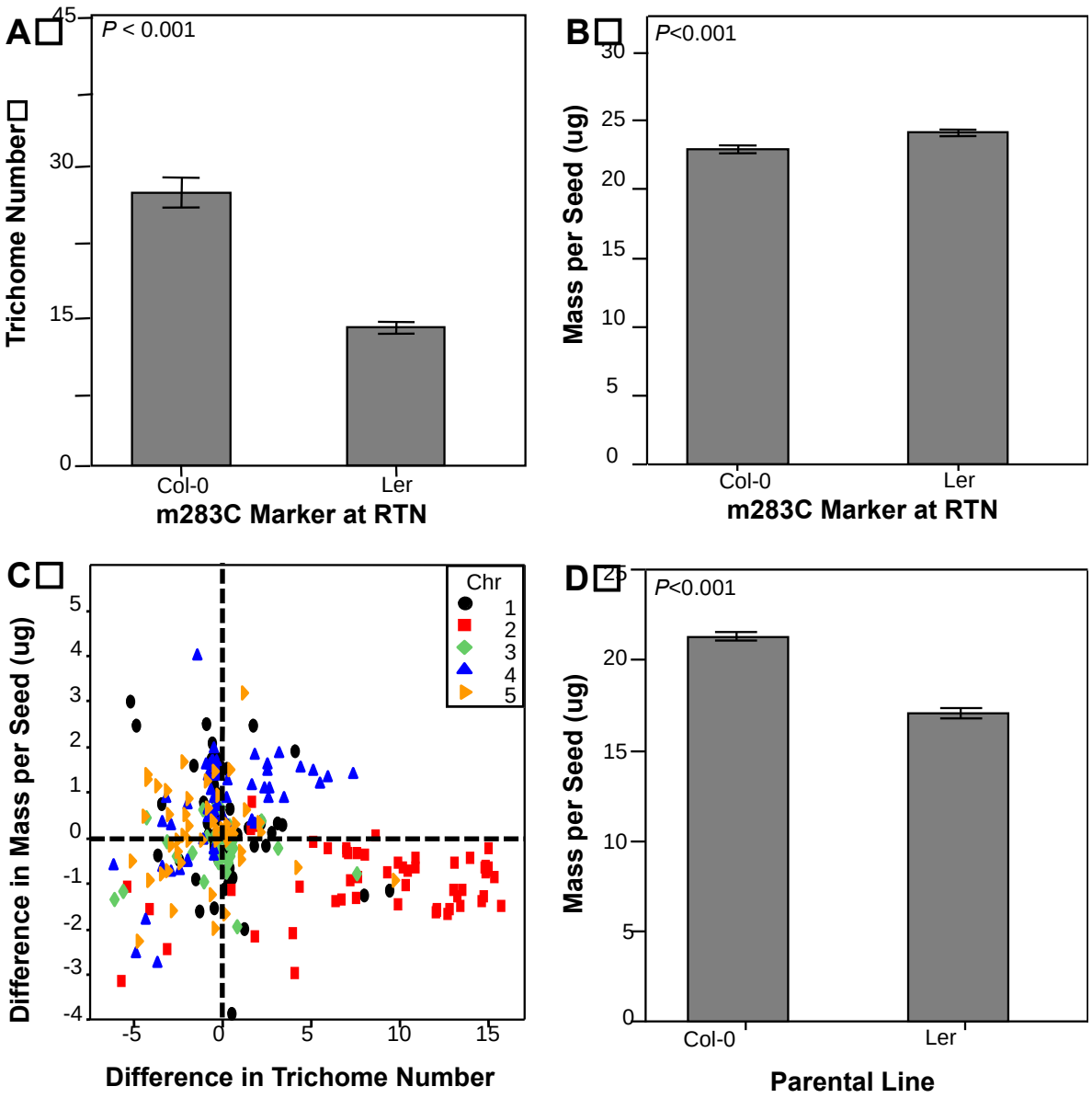


Figure 24. The effect of the *ETC2* locus on seed weight in recombinant inbred lines (RILs) between Ler x Col-0.

(A) Average **trichome n**umber (+/- SE) of the first and second leaf (data provided by J. Larkin) and **(B)** average mass per seed (+/- SE) of RILs homozygous for the m283C marker, which is the closest marker to the *ETC2* locus. **(C)** Scatterplot of allelic difference in mass per seed (µg) and trichome number for the genome-wide RIL QTL markers with physical map position. **(D)** Average mass per seed (+/- SE) of parental lines.

Table 19. Mass per seed and trichome numbers for the recombinant inbred lines from the cross of Ler x Col-0.

QTL Cross	RIL#	Marker m283C on Chr2	Avg Mass per Seed (ug)	Avg Trichome Number
Ler x Col-0	1900	-	23.5	11.3
Ler x Col-0	1901	-	22.1	22.8
Ler x Col-0	1903	-	21.6	10.4
Ler x Col-0	1904	-	22.9	32.2
Ler x Col-0	1905	Col-0	23.8	27.6
Ler x Col-0	1906	Col-0	26.9	25.1
Ler x Col-0	1907	Ler	25.1	12.7
Ler x Col-0	1908	Ler	27.3	10.5
Ler x Col-0	1909	Col-0	21.3	30.4
Ler x Col-0	1910	Ler	23.2	9.7
Ler x Col-0	1911	-	21.4	24.5
Ler x Col-0	1912	Ler	25.0	10.9
Ler x Col-0	1913	-	22.5	24.6
Ler x Col-0	1914	Col-0	23.2	22.7
Ler x Col-0	1915	-	20.5	13.1
Ler x Col-0	1916	Col-0	21.5	26.8
Ler x Col-0	1917	Ler	24.3	14.2
Ler x Col-0	1918	Ler	27.7	11
Ler x Col-0	1919	Col-0	23.1	11.9
Ler x Col-0	1920	Col-0	26.3	33.5
Ler x Col-0	1921	Ler	24.9	14.4
Ler x Col-0	1922	Col-0	20.6	15.2
Ler x Col-0	1923	Ler	21.4	12.4
Ler x Col-0	1924	-	22.1	25
Ler x Col-0	1925	Col-0	16.5	25.8
Ler x Col-0	1926	Col-0	27.6	29.8
Ler x Col-0	1927	-	20.9	11.3
Ler x Col-0	1928	Col-0	23.3	18.6
Ler x Col-0	1929	Col-0	18.3	30.7
Ler x Col-0	1930	Ler	26.2	8.7
Ler x Col-0	1931	Ler	22.6	8.3
Ler x Col-0	1932	Col-0	18.1	31.6
Ler x Col-0	1933	Ler	23.1	17.3
Ler x Col-0	1934	Ler	20.8	9.1
Ler x Col-0	1935	Ler	21.5	11.9
Ler x Col-0	1936	Ler	24.6	11.8
Ler x Col-0	1937	Ler	25.1	16.7
Ler x Col-0	1938	Ler	22.2	15.1
Ler x Col-0	1939	Ler	21.8	16.8
Ler x Col-0	1940	Ler	27.0	14.1
Ler x Col-0	1941	Col-0	18.9	29.9
Ler x Col-0	1942	-	21.1	21.6
Ler x Col-0	1943	Ler	25.6	9.4
Ler x Col-0	1944	-	21.0	22.1
Ler x Col-0	1945	-	25.9	31.7
Ler x Col-0	1946	-	18.3	10.4
Ler x Col-0	1947	Ler	29.1	14.9
Ler x Col-0	1948	Col-0	26.1	29.3
Ler x Col-0	1949	-	20.8	29.3
Ler x Col-0	1950	Col-0	22.7	35.2
Ler x Col-0	1951	Ler	22.1	14.2
Ler x Col-0	1952	Ler	27.0	16.9

Table 19. continued

QTL Cross	RIL#	Marker m283C on Chr2	Avg Mass per Seed (ug)	Avg Trichome Number
Ler x Col-0	1953	-	27.7	17.4
Ler x Col-0	1954	Col-0	24.3	29
Ler x Col-0	1955	Col-0	24.3	33.3
Ler x Col-0	1956	-	22.5	13.3
Ler x Col-0	1957	Ler	26.8	17.3
Ler x Col-0	1958	Ler	21.9	17
Ler x Col-0	1959	Col-0	21.9	29.1
Ler x Col-0	1960	Ler	23.3	12.2
Ler x Col-0	1961	Ler	25.7	16.8
Ler x Col-0	1962	Col-0	23.5	33.7
Ler x Col-0	1963	Ler	24.7	24.9
Ler x Col-0	1964	Ler	24.0	13.9
Ler x Col-0	1965	-	22.1	15.4
Ler x Col-0	1966	-	19.4	18
Ler x Col-0	1967	-	22.9	10.3
Ler x Col-0	1968	-	23.1	32.8
Ler x Col-0	1969	Ler	25.3	19.7
Ler x Col-0	1970	-	23.5	14.1
Ler x Col-0	1971	Col-0	24.5	22.7
Ler x Col-0	1972	Ler	25.1	17.3
Ler x Col-0	1973	Col-0	24.4	45.7
Ler x Col-0	1974	Ler	25.6	21.4
Ler x Col-0	1975	Ler	20.1	17.1
Ler x Col-0	1976	Ler	22.6	16
Ler x Col-0	1977	Ler	21.5	17.7
Ler x Col-0	1978	Col-0	26.2	36.8
Ler x Col-0	1979	Col-0	21.7	48.4
Ler x Col-0	1980	Ler	23.7	17.2
Ler x Col-0	1981	Col-0	23.4	28.9
Ler x Col-0	1982	Ler	24.3	10.5
Ler x Col-0	1983	Col-0	24.7	17.3
Ler x Col-0	1984	Ler	21.5	12
Ler x Col-0	1985	Col-0	21.1	16.9
Ler x Col-0	1986	Ler	20.9	9.6
Ler x Col-0	1987	Ler	24.1	11.5
Ler x Col-0	1988	Col-0	22.3	19.5
Ler x Col-0	1989	Col-0	34.3	30.1
Ler x Col-0	1990	Ler	24.8	16.9
Ler x Col-0	1991	Ler	30.7	16.2
Ler x Col-0	1992	-	22.5	10.9
Ler x Col-0	1993	Col-0	21.1	39
Ler x Col-0	1994	-	21.1	8.9
Ler x Col-0	1995	Ler	23.3	11.8
Ler x Col-0	1996	Ler	24.7	11.1
Ler x Col-0	1997	Col-0	24.9	28.9
Ler x Col-0	1998	Col-0	24.3	16.5
Ler x Col-0	1999	-	18.4	37.2
Ler x Col-0	4686	Col-0	20.9	
Ler Parent	20		17.1	10.2
Col-0 Parent	933		21.3	33.2

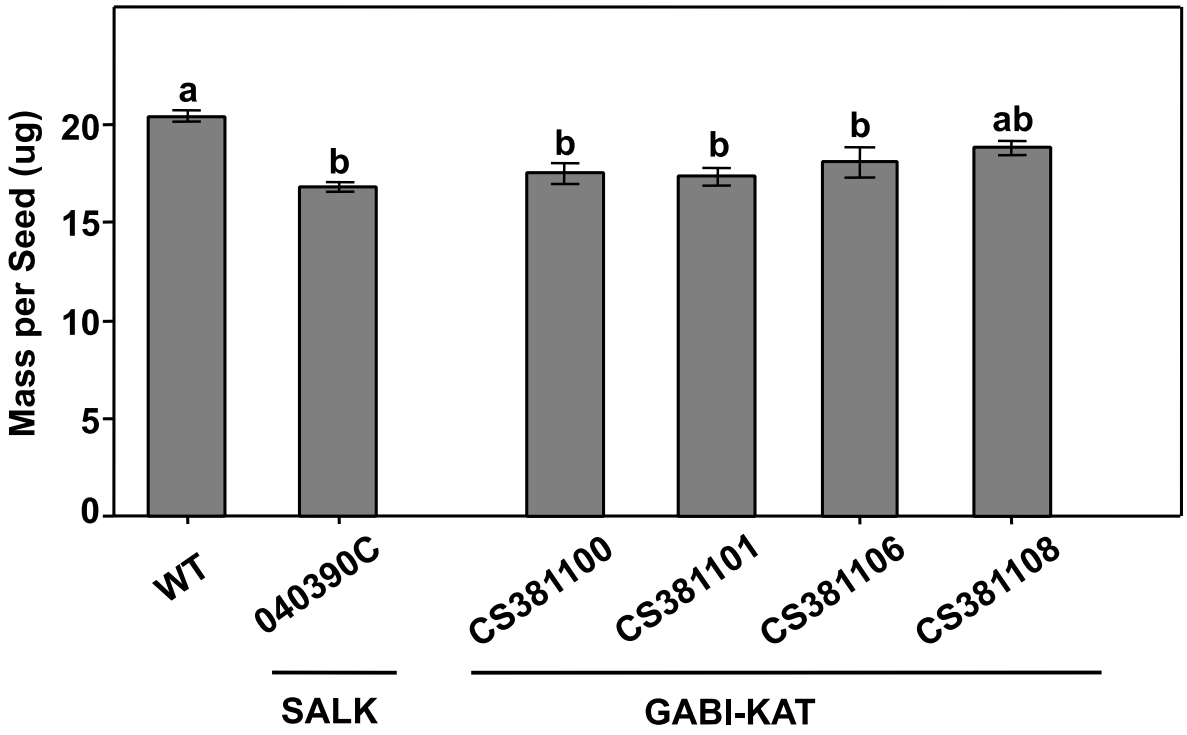


Figure 25. Average seed mass (+/- SE) for wild type Columbia-0 CS70000, SALK T-DNA mutant 040390C, GABI-KAT mutants: CS381100, CS381101, CS381106 and CS381108.

Table 20. Mass per seed (ug) of T-DNA insertion lines at ETC2.

Gene	Description	ABRC Stock	AVG Mass per Seed
Background	Wildtype Col-0	CS70000	20.4
ETC2	SALK-etc2-	SALK_040390C	16.8
ETC2	GB-etc2-	CS381097	18
ETC2	GB-etc2-	CS381100	17.5
ETC2	GB-etc2-	CS381101	17.3
ETC2	GB-etc2-	CS381106	18.1
ETC2	GB-etc2-	CS381108	18.8

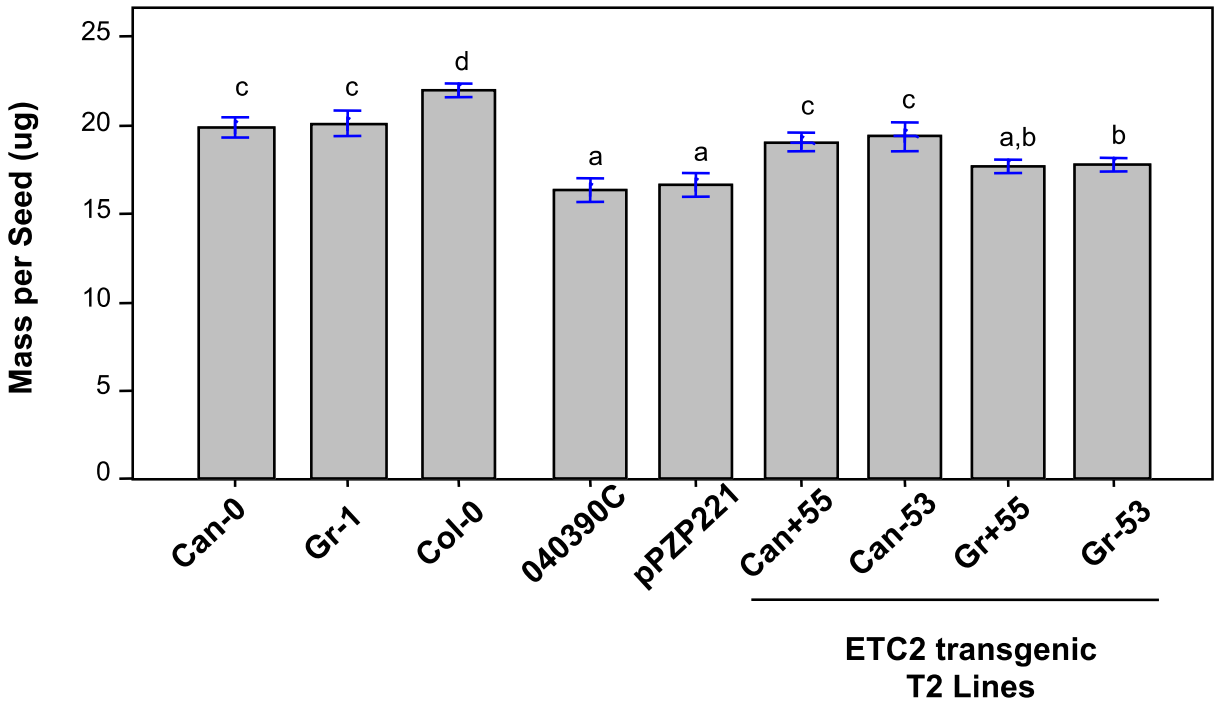


Figure 26. Complementation of the *etc2*-2 T-DNA insertion mutant with genomic *ETC2* chimeras.

The seeds were provided by Dr. Hauser and created by Dr. Hilscher. Averages (+/- SE) of mass per seed (ug) of Can-0, Gr-1, Col-0 accessions (N = 30 each) in comparison to Col-0 mutant *etc2-2* allele (SALK_040390C) (N=30) and T2 seeds of the *etc2-2* line transformed with the empty vector pPZP221 or four different chimeric versions of the *ETC2* gene as follows: Can+55 (Can-0 allele mutated to Gr-1 at position+55, Can-53 (Can-0 allele mutated to Gr-1 at position-53), Gr+55 (Gr-0 allele mutated to Can-0 at position +55), and Gr-53 (Gr-0 allele mutated to Can-0 at position -53). Each bar represents an average of five values of batches of five seeds each for 10 independent transformants of each chimeric sequence for a total of 50 seed mass measurements per transformation. Means with significant difference by Tukey's test at $P = 0.05$ are indicated by a difference in letters.

Table 21. Mass per seed (ug) of *ETC2* SNP Constructs Transformed into the *etc2*- background.

Lines		Total Batches Weighed	Total Seeds	Average Mass per Seed (ug)
T2 generation of transformed lines:				
Empty pPZP221 vector (control)	A	5	25	15.8
transformed into SALK_040390C	B	5	25	12.5
	C	5	25	19.4
(10 independent transformations)	D	5	25	14.2
	E	5	25	17.6
	F	5	25	18.5
	G	5	25	16.3
	H	5	25	18.6
	J	5	25	17.4
	K	5	25	16.0
$ETC2^{Can-0}$ coding sequence driven by	P	5	25	17.1
chimeric Can-0 & Gr-1 promoter	Q	5	25	17.6
which has Gr-1 nucleotide at position -53	R	5	25	20.1
construct transformed into SALK_040390C	S	5	25	19.0
therefore named "Can -53"	T	5	25	19.5
	U	5	25	17.9
(10 independent transformations)	V	5	25	18.2
	W	5	25	25.3
	X	5	25	17.0
	Z	5	25	22.3
$ETC2^{Gr-1}$ coding sequence driven by	A	5	25	15.4
chimeric Gr-1 & Can-0 promoter	B	5	25	16.6
which has Can-0 nucleotide at position -53	C	5	25	19.0
construct transformed into SALK_040390C	D	5	25	18.8
therefore named "Gr -53"	E	5	25	18.4
	F	5	25	17.6
(10 independent transformations)	G	5	25	18.0
	H	5	25	18.8
	I	5	25	17.0
	K	5	25	18.4
$ETC2^{Can-0Gr-1}$ Chimeric Coding Sequence	AA	5	25	22.6
which has Gr-1 nucleotide at position +55	AB	5	25	19.0
driven by Can-0 promoter	AC	5	25	17.8
construct transformed into SALK_040390C	AE	5	25	17.9
therefore named "Can+55"	AG	5	25	20.5
	AH	5	25	18.0
(10 independent transformations)	AJ	5	25	20.1
	AK	5	25	18.9
	AL	5	25	18.7
	AM	5	25	17.3
$ETC2^{Gr-1Can-0}$ Chimeric Coding Sequence	E	5	25	17.8
which has Can-0 nucleotide at position +55	F	5	25	18.7
driven by Gr-1 promoter	G	5	25	18.5
construct transformed into SALK_040390C	H	5	25	20.6
therefore named "Gr+55"	J	5	25	17.0
	K	5	25	17.2
(10 independent transformations)	L	5	25	16.8
	M	5	25	17.2
	O	5	25	16.3
	P	5	25	17.0
Other Lines:				
Can-0		30	150	19.9
Gr-1		30	150	20.1
Col-0		30	150	22.0
SALK_040390C		30	150	16.4

4.4 DISCUSSION

While GWA studies can be especially vulnerable to confounding influences of population structure (H., M., Kang *et al.*, 2008; Platt *et al.*, 2010), my interpretations of trade-off costs at the RTN locus are robust for three reasons. First, the seed mass reduction that I see associated with the high trichome number allele in wild accessions was more significant when population structure (Kärkkäinen *et al.*, 2004) was incorporated. Second, there was no obvious geographic pattern in the distribution of the two alleles at RTN (Figure 27), suggesting that this allelic diversity is both widespread and relatively ancient (Platt *et al.*, 2010). Third, the independent tests with the Ler x Col-0 RIL lines and the complementation of the T-DNA insertion mutant at ETC2 both provide independent evidence supporting this allelic effect at RTN on trichome number and seed weight (Figure 24-26).

Heavy seeds are favored for survival in deep shade and in the presence of environmental stresses, such as drought and competition with other plants (Kirik *et al.*, 2004). Given these important fitness consequences of seed provisioning, it is remarkable how little is known about the genetic basis of its natural variation. One possible concern about experimental measurement of mass per seed is its stability over time in storage. Under my conditions, I saw no evidence of a storage effect; mass per seed had not changed significantly for any lines that I measured initially in 2009 and retested in 2012 (Figure 28, Table 22). Moreover, I have also found that parent and offspring values for mass per seed are highly correlated (Figure 28, Table 23), suggesting high heritability for this trait in *A. thaliana*. Considering the plants grew in a common garden condition with deliberately limited resource, this observed heritability was highly likely to be a genetic outcome, which suggested that plants might have a sophisticated

regulatory system in controlling energy into offspring production or into developing other important traits.

In summary, my results suggest an important role of trade-offs in maintaining allelic variation at ETC2. My results are likely to explain why an allele that confers low trichome numbers persists in natural populations of *A. thaliana*. More generally, my results document the potential of using GWA mapping to co-map traits that are critical for understanding resistance to pests in agriculture (Hancock *et al.*, 2011). In present scenarios of climate change and emerging plant enemies, such examination of allelic variation beneath these fitness and defense tradeoffs will be critical to predicting the responses of wild populations.

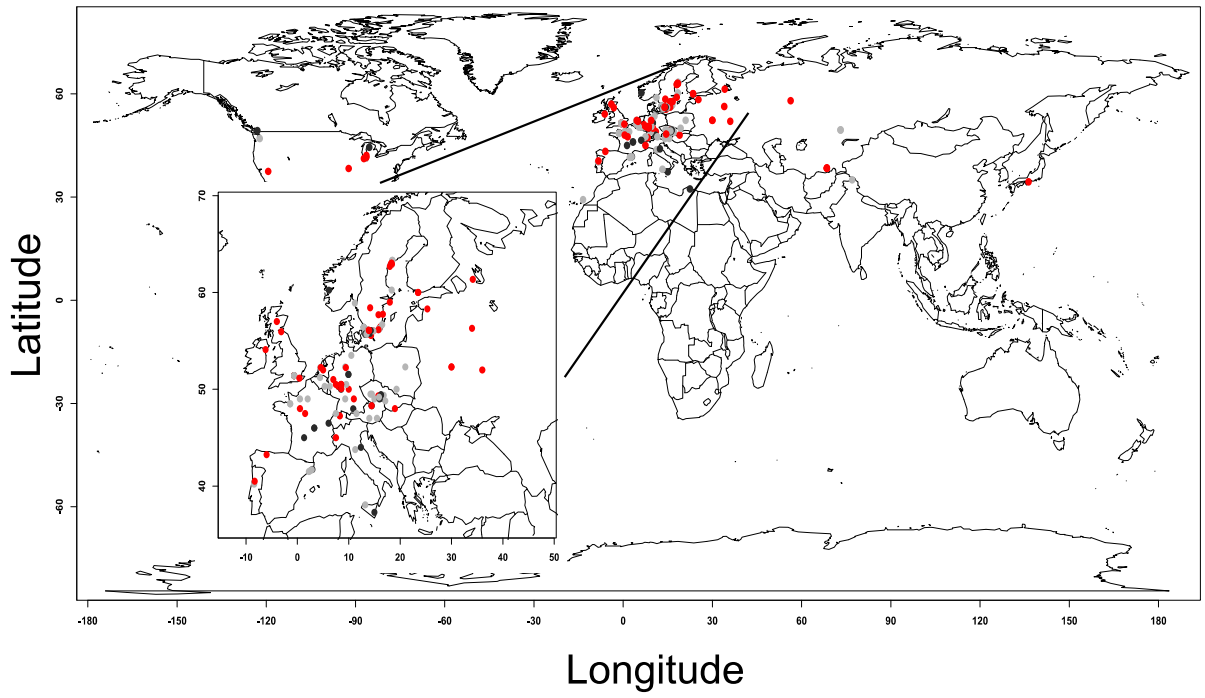


Figure 27. Geographic distribution of the 168 accessions from the RegMap collection showing no significant correlation between latitude and allele at *ETC2*.

Red, black, and gray represent the Col-0-like, Ler-1-like, and intermediate alleles, respectively, at *ETC2* (as in Figure 12 E-F).

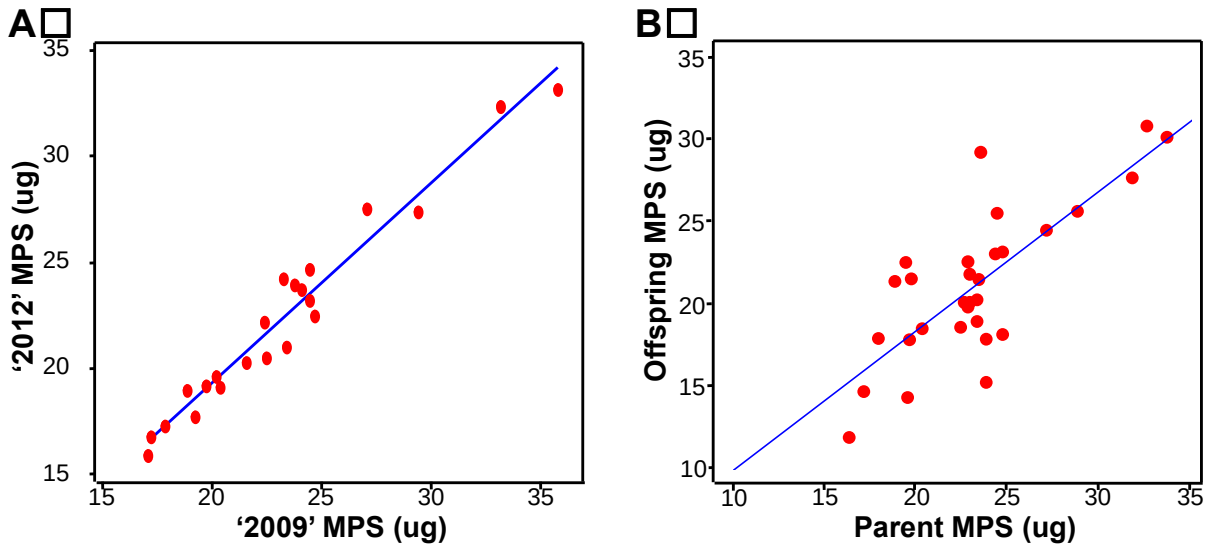


Figure 28. Seed mass is a stable and heritable trait.

(A) Regression of two measurements of mass per seed showing small change of overall mass per seed under my storage conditions. Data represent the averages of three replicate aliquots of five seeds each for each accession weighed in 2009 and reweighed in 2012 (Table 22). **(B)** Parent - offspring regression of mass per seed values for a random picked subset of 30 genotypes from the RegMap Panel (Table 23). Data represent the averages of five replicate aliquots of five seeds each for each accession where the parents were grown together in one environment and the offspring were grown together again in a common garden environment. These data indicate a high broad-sense significant heritability of mass per seed of around 0.85 in this group of the RegMap Panel lines.

Table 22. Mass per seed (μg) of a subset of accessions weighed 2009 and reweighed 2012.

Accession	Line	2009 Mass per seed (ug)	2012 Mass per seed (ug)
Bil-7	6901	23.8	23.9
Bro 1-6	8231	22.4	22.2
Bu-0	8271	33.2	32.3
Bur-0	6905	35.8	33.1
CIBC-5	6730	24.4	24.7
CS6187/Seattle-0	8245	18.8	18.9
Dra-3-1	8283	24.5	23.2
Kin-0	6926	23.3	24.2
KZ-9	6931	17.0	15.9
LL-0	6933	19.2	17.7
Lom 1-1	6042	20.2	19.6
Nok-3	6945	24.7	22.5
Pa-1	8353	17.2	16.7
Pla-0	8357	19.7	19.2
RMX-A180	7525	21.6	20.3
Rsch-4	8374	17.8	17.3
Sorbo	6963	27.1	27.6
Spr-1-6	6965	20.4	19.1
Ull 1-1	8426	22.5	20.5
Var-2-1	7516	29.4	27.4
Vimmerby	8249	24.1	23.7
Zdr-6	6985	23.4	21.0

Table 23. Comparison of parent and offspring values for a subset of 30 RegMap Panel lines.

Accession	Ecotype_id	RTN.Allele	Leaf.Trichome.Number	Parent MPS (ug)	Offspring MPS (ug)
Blh.1	8265	Col-0-like	220	19.6	14.28
Bor.4	6903	Inter	130	22.9	22.56
Br.0	6904	Ler-1-like	0	24.8	23.16
Dem.4	8233	Col-0-like	96	27.2	24.48
Dra.3.1	8283	Ler-1-like	72	24.5	25.52
DraIII.1	8285	Ler-1-like	119	19.7	17.8
Eds.1	6016	Inter	181	33.8	30.16
Est.1	6916	Col-0-like	145	24.8	18.12
Fab.2	6917	Inter	168	23.6	29.24
Ga.0	6919	Ler-1-like	124	23.4	20.24
GOT.7	6921	Ler-1-like	62	32.7	30.84
Kas.1	8424	Inter	129	31.9	27.68
Kelsterbach.4	8420	Col-0-like	77	19.8	21.52
Kondara	6929	Inter	154	22.9	19.8
Lc.0	8323	Col-0-like	170	16.4	11.84
Lip.0	8325	Inter	197	23.4	18.92
LP2.6	7521	Inter	88	23	20.08
Na.1	8343	Col-0-like	167	24.4	23.04
Pa.1	8353	Inter	78	17.2	14.64
Pro.0	8213	Inter	286	23.9	15.2
Pu2.23	6951	Inter	42	23.5	21.48
Sav.0	8412	Inter	22	23.9	17.84
Seattle.0	8245	Inter	92	18.9	21.36
Spr.1.6	6965	Inter	96	20.4	18.48
TAMM.27	6969	Inter	249	23	21.8
Tottarp.2	6243	Inter	81	18	17.88
Ts.5	6971	Ler-1-like	91	28.9	25.64
Ull.1.1	8426	Col-0-like	222	22.5	18.56
Ull.2.3	6973	Col-0-like	348	22.7	20.08
Uod.2	8428	Col-0-like	153	19.5	22.52

4.5 FUTURE DIRECTION

The primary experiment that need to be done is putting the two types of ETC2 alleles (Col-0-allele and Ler-1-allele) back into the null etc2- background, which is being carried on by our collaborators at the University of Tennessee. When I have the transgenic plants, I will first grow them under the common garden condition and measure their trichome number. When plants are done, I will then measure the seed mass and see if the pattern of trade-off still exists.

4.6 ACKNOWLEDGEMENTS

I thank Dr. Julia Hilscher and Dr. Marie-Theres Hauser for providing their chimeric complementary transgenic lines. I thank Dr. Daniel Weeks, Dr. John Wilson, and Qing Liu for helpful feedback regarding statistical analyses, Dr. Bjarni Vilhjalmsson, Dr. Christopher Toomajian, and Lunching Chang for help with R programming, and Dr. Susan Kalisz, Dr. Mark Rebeiz, and Dr. Jeff Brodsky for comments on the manuscript. I thank Seth Reighard and Juhyun Kim for assistance with data collection, and Dr. John Larkin for providing published trichome data for the Ler x Col-0 mapping lines. I thank the Salk Institute Genomic Analysis Laboratory for the T-DNA insertion mutants distributed by ABRC. This research was supported by US National Science Foundation Grant #1050138 (M.B.T.).

5.0 CONCLUSION

In my dissertation work, my major study system has been the RegMap worldwide collection of 195 wild Arabidopsis accessions collected and the SNP data from their fully sequenced genomes. Using this powerful system, I performed GWAs mapping and compared the results with traditional QTL mapping. For one of my GWAs map, I provided new criteria for selecting and evaluating candidate genes. I also performed multiple experimental analyses to explore the function of a candidate gene, *AtABCG16,* from one of the map (*Pst* DC3000 resistance) and provided the evidence that this gene was likely to explain a portion of variation I observed. Moreover, I tried to extensively develop the usage of this system. I created a co-mapping method to search genetic loci account for phenotypic trade-off.

5.1 BOTTOM OF CHROMOSOME 2 AND TOP OF CHROMOSOME 5

In my dissertation study, I have mapped twelve phenotypic traits, including salicylic acid production, vitamin C production, abscisic acid tolerance, leaf trichome number, *Pst* DC3000 resistance, seedling max extension, seedling root length, plant rosette mass, mass per seed, number of seed per fruit, number of fruit per plant, and total number of seeds. Interestingly, when I summarized all of the candidate loci from the twelve maps, I found that

some area on the genome was hotter than the other places (Figure 29). The bottom of chromosome 2 and the top of chromosome 5, particularly, when I looked at the entire genome at a 5Mb-large scale, each had ten out of twelve maps hitting for at least once. One possible argument was that some of the traits I mapped were correlated, such as seed mass, seed numbers and pod numbers. However, the candidate loci in these two regions were actually not entirely overlapped, which indicated that the causal genetic elements were close on the chromosome but did not totally overlap with each other. Interestingly, meanwhile, some parts on the chromosome were relatively quiet. I would like to argue that some of the genetic loci accounting for the variation of phenotypic traits meanwhile some other genetic loci account for determining the phenotypic traits at the same time. Causing or determining a phenotype was different from cause the variation of this phenotype. For many important genes, like resistant genes, seed genes, they were too important to allow any room for natural variation. Under this scenario, GWAs cannot detect these areas since there was no genetic polymorphism present. However, since plants were stationary, they kept a portion of variation on their genome to allow phenotypic plasticity in respond to the biotic and abiotic changes in the environment. For this kind of genetic material, GWAs could easily pick up signals. From what I found, the bottom of chromosome 2 and the top of chromosome 5 were likely to be a genetic pool that accounted for many phenotypic variations.

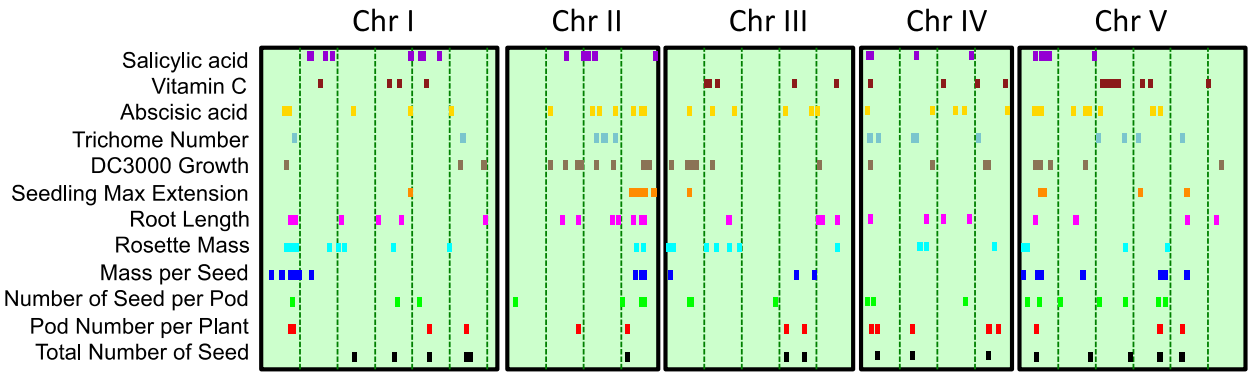


Figure 29. Summary of 12 GWAs maps.

Each dot represents a genetic locus which is among top 0.1% of the GWAs map. The five boxes represent five chromosomes. Green dash lines show the position of every 5Mb.

5.2 QTL MAPPING AND GWA MAPPING

In chapter #2, I used seed mass GWAs map as an example to describe how to create a GWAs map, to select the candidate genes, and to control population structure. The primary goal of this chapter was to show that traditional QTL mapping and modern GWAs mapping can be combined to search qualified candidate genes. Three candidate loci were found on the top of chromosome 1 and one big locus was found at the bottom of chromosome 2. These locations, at the chromosome level, matched up with previously published studies using QTLs mapping method. QTL and GWA mapping, indeed, are very similar. They all provide candidates based on the statistical results of trait-marker association. The essential difference is that QTL uses artificial created mapping population relying on artificial crossing while GWA uses wild accessions relying on natural recombination. GWA mapping can easily acquire a large set of mapping population. However, QTL don't have the concern about population structure. To archive the most accurate genetic map, combining the two mapping methods would be a better way than doing just one of them.

5.3 TRADE-OFF MODEL

In chapter #4, I tried to answer the question: " why do plants persist in having such big variation even when a trait is beneficial? " Indeed, this question should be asked after every map. Because if there is no variation, GWAs mapping never works. Only genomes with genetic differences that cause phenotypic variation can lead to significant GWAs scores. However, if the mapped trait is a benefit to plants, why could I still observe significant

variations? The most likely answer is the "trade-off" theory. Nutrients and resources are limiting and extensively developing beneficial traits is costly. Some of the wild accessions decide to maintain a normal or lower amount of phenotypes in order to gain advantages at other aspects. In chapter #4, I showed a negative correlation between a defense trait, leaf trichome number, and a fitness trait, mass per seed. I found the ETC2 locus, which accounted for over 60% natural variation of trichome number, was also involved in seed mass determination. And the effects on the two traits were in opposite directions at this locus. This finding supported the "trade-off" theory. More importantly, it was one of the first publications that attempted to dissect the effect of multiple alleles at a genetic locus on two important phenotypes which, phenotypically, are not associated.

5.4 FUTURE DIRECTIONS

In *Arabidopsis*, GWAs has been broadly used in searching for genetic loci to account for quantitative traits and has lead to many successful findings. However, this area is still developing and many big challenge needs to be answered.

First, what can we do if the interested genetic region is not a small part of chromosome? Figure 28 indicated two parts of the chromosome that are likely to be important for multiple traits. However, traditional genetic methods such as T-DNA knockout, RNA interference, and 35S-driven overexpression cannot fit in this scenario since the target locus is too big. The SNPs associated with important phenotypic variations sometimes are tightly linked and cover over 2-5Mb of a chromosome. Scientists need to figure out ways that can understand effects of genetic variation on phenotypic difference at the chromosome level.

Second, what else can we use to evaluate candidates of a GWAs map besides using microarray and GO annotations? Since GWAs is genome-wide, the evaluation needs to be genome-wide as well. At this point, RNA-seq is a great candidate. It covers the entire genome and most important, it can be used to analyze the wild accessions, thus make linking expression level and nucleotide polymorphisms become possible.

Last but not least, how can we improve the current study system? In other words, do we want more accessions or do we want more genetic markers? Increasing the number of accessions can increase the phenotype density, then provide a better gradient of phenotypic variation. On the other hand, increasing the number of genetic markers (SNPs) could increase the resolution of the map, which could help us to locate tiny but important genetic factors such as small RNAs.

Together, my dissertation work provides experimental and statistical evidence that shows that GWAs can successfully find novel candidate genes related to mapped natural variation. It also explores the new area of using the wild *Arabidopsis* accessions and SNP dataset. Such extension can contribute to our knowledge of the interaction between environmental factors and genetic information. Moreover, this experimental system is a good tool to dissect the genetic basis of observed trade-offs, which has been long existed and is a interested topic in ecology and evolution of broad interest.

APPENDIX A

CONSIDERATIONS OF USING WILD ARABIDOPSIS ACCESSIONS FOR QUANTITATIVE ANALYSES

The *Arabidopsis* mapping system has been becoming more and more complex since the number collected and sequenced wild accession kept increasing during the past six years. In 2005, when the first set of accessions had been collected and used for scientific study, there were only 96 accessions and about 3,000 high-quality SNP markers. In 2010, two big papers were published in GWAs mapping area. Atwell et al. (Atwell *et al.*, 2010) used about 200 accessions, although in a fair number of their maps the phenotypic data were collected from the first 96 accessions. Cao et al. (Cao *et al.*, 2011) used a set of wild accessions which was totally different from Atwell's paper. The total available accessions at that time were around 500. Now, on the 1001 project website, SNP data of over 1,000 accessions is available. The break through of next-gen sequence has significantly improved this field. Indeed, in the big data era, when raw materials are so easy to get, the limitation for quantitative genetics in this area has been changed from creating/collecting mapping populations to how to sample, calculate, analyze, understand and interpret the outcome of the maps. These steps are critical because when sequencing techniques being introduced into other species like maize, strawberry,

tobacco, rice and so on, GWAs studies will be performed in all of these long- and well- studied systems. At that moment, the big time for ecology and evolution will come and every effort we have been putting into this field using Arabidopsis as the study system will be followed and modified for scientists who are working on other plant species. Here, I would like to put some of my thoughts of the two most important and needed considerations of using wild Arabidopsis accessions: population structure and chromosome differences.

A.1 HOW TO UNDERSTAND THE POPULATION STRUCTURE

One big challenge when doing statistics using the wild *Arabidopsis* accessions is population structure. What is population structure and why does it need to be handled properly during the genome-wide studies? Population structure is also called genetic distance. As it is literally written, it is the distance between the genomes of those wild accessions. In other words, for a give accession, its genome is more similar to some accessions than it is to others. So, genetically, they are more "closer" to each other. There are many reasons that can cause this. The primary one is kinship, where parents and offspring and siblings share genetic similarity due to common descent. This causes plants close to each other to share more of their genome in common than plants farther apart. However, a second reason that plants that are close together can share common alleles is directly because of the environment. This is because geographic elements such as elevation, latitude, humidity, temperature and sunshine all favor particular alleles, thus plants close together may be unrelated but have the favored allele for that location. Indeed, previous publications have found that wild accessions which were geographically closed were also genetically closed as well (Horton *et al.*, 2012) and these wild

accessions perform better in the environment similar to their habitats showing local adaptation (Fournier-Level *et al.*, 2011). All of this information tells us one thing, these wild accessions are not independent from each other. They belong to genetic clusters and the overall pattern of these clusters is "population structure".

How does population structure affect quantitative genetics? It's very simple. The statistic methods we used for understanding the association between genetic markers and phenotypic traits, which usually are ANOVA and Wilcoxon rank sum test, no matter whether they are parametric or non-parametric, all require one assumption that is all tested variables need to be independent against each other. Since some of the wild accessions are similar or close to each other, directly throwing them into any of these statistical models is not correct. The results are supposed to return higher false-positive ratio.

To control the effect caused by population structure, for GWAs mapping particularly, researchers have introduced a mix-model method called "EMMA" (H., M., Kang *et al.*, 2008). This method efficiently reduces the false-positive ratio but, at the same time, decreases the overall signal intensity (Atwell *et al.*, 2010). This mixed-model method has been used in many statistic cases when tested variables are not independent to each other. The important factor for this method is to provide the most accurate cluster information, which is still a challenge for quantitative genetists in plant science. Here, I propose a method which is based on principle component analysis and borrows the idea from machine learning method. First, we use PCA to decrease the dimensions and plot PC1 against PC2. Then I use K-means clustering method to let the computer tell me which accessions are more close to each other (Figure 30, Table16). The number of clusters can be changed upon the experimental needs. This method is unbiased by researchers' knowledge of accessions. Horton et al. 2012 had a similar idea but the

difference is that they color-coded the accessions on the plot and found a fair number of accessions which were geographically closed were also closed on the plot. But my method is indeed different from theirs since I never introduced the geographic information into the model. My method let the computer cluster these accessions only based on their genetic similarity. So what Horton et al. had actually supports my method.

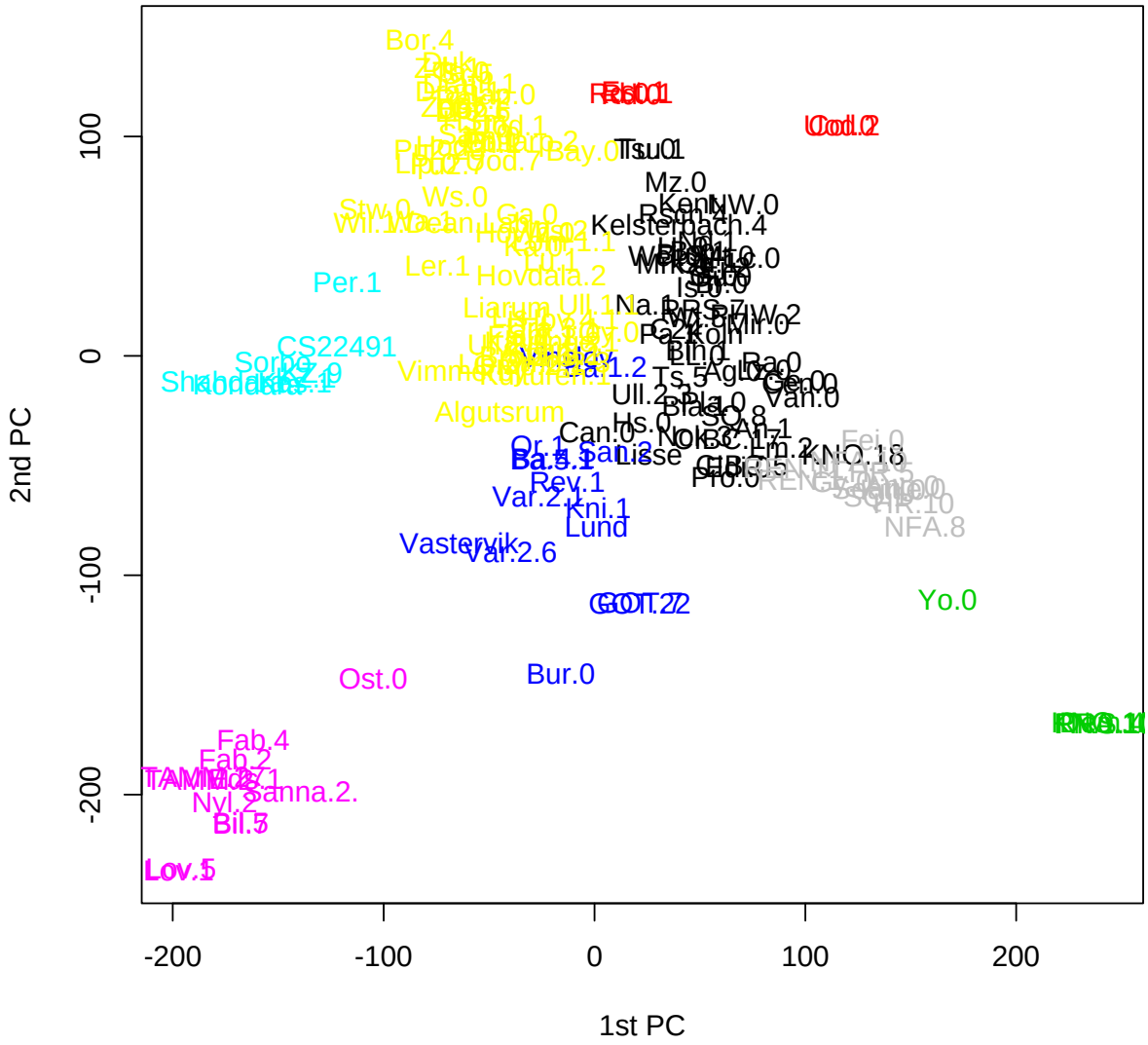


Figure 30. Population structure from PCA/K-Means method using 168 *Arabidopsis* accessions

A.2 HOW TO MAP MULTIPLE TRAITS AT A TIME?

Here, I take two traits, mass per seed and Number of seed per pod, for an example. I will provide three independent methods that try to map these two traits simultaneously.

A.2.1 Sliding regression method (Figure 31A)

I used the SNP effects on mass per seed and seed number per pod (Table calculated in polarized overlay map and plot the values of every 100 constitutive SNPs on five chromosomes. For each 100-SNPs window, we run linear regression and collect the t-value to evaluate the correlation between delta MPS and delta SPF of this window. The t-values were then assigned to the 51st SNP of each 100-SNPs window and plotted against their chromosome positions. To handle potential bias of the genetic structure, we then performed a permutation analysis. The MPS and SPF phenotypic values of 164 accessions were shuffled and randomly assigned back to the wild accessions for 100 times. For each round of assignments, we recalculated the SNP effects on MPS and SPF based on the random generated permuted phenotypic values. Then we performed regressions, extracted t-values as described above. For the 100 permuted t-values of each SNP, we set the 5% and 95% as the lower and upper boarder values. These values were then plotted against their chromosome positions. Genome-widely, the lowest lower boarder value was -19.27, on chromosome 5. We used this value as the threshold and considered the genes containing the SNPs of which the t-values were smaller than this value were then considered as candidate genes.

A.2.2 Polarized overlay method (Figure 31B)

I first extracted the p-values of MPS-Wilcoxon and SPF-Wilcoxon maps. For each SNP, I multiplied the p-value from MPS map by the p-value from SPF map. The product was then inverse- and logarithm- transformed, and was called as "overlay score". For each SNP, I also calculated the difference between averaged MPS or SPF of wild accessions with one of the two alleles. This "delta MPS" or "delta SPF" was then called the "SNP effect". If the SNP effect on MPS and SPF was in the same direction, we then assigned a positive symbol to the corresponded overlay score. Negative symbols were assigned to those SNPs of which the effects on MPS and SPF were in opposite directions. These overlay scores combined with positive or negative signs were called as "polarized overlay" values and plotted against their chromosome positions. SNPs with lowest 0.2% GWAs score of the map were then screened out and the genes containing these SNPs were considered as candidate genes.

A.2.3 First principle component method (Figure 31C)

I first performed a principle component analysis of MPS and SPF together. The values of the first principle component (PC1), which could explain 68.5% variation of observed trade-off, were then used for GWAs mapping following the procedure described above. SNPs with highest 0.2% GWAs score of each map were then screened out and the genes containing these SNPs were considered as candidate genes.

In all, my identification of a major locus on chromosome 2 of *A. thaliana* is novel and provides an important target for genes to improve agricultural crop yields.

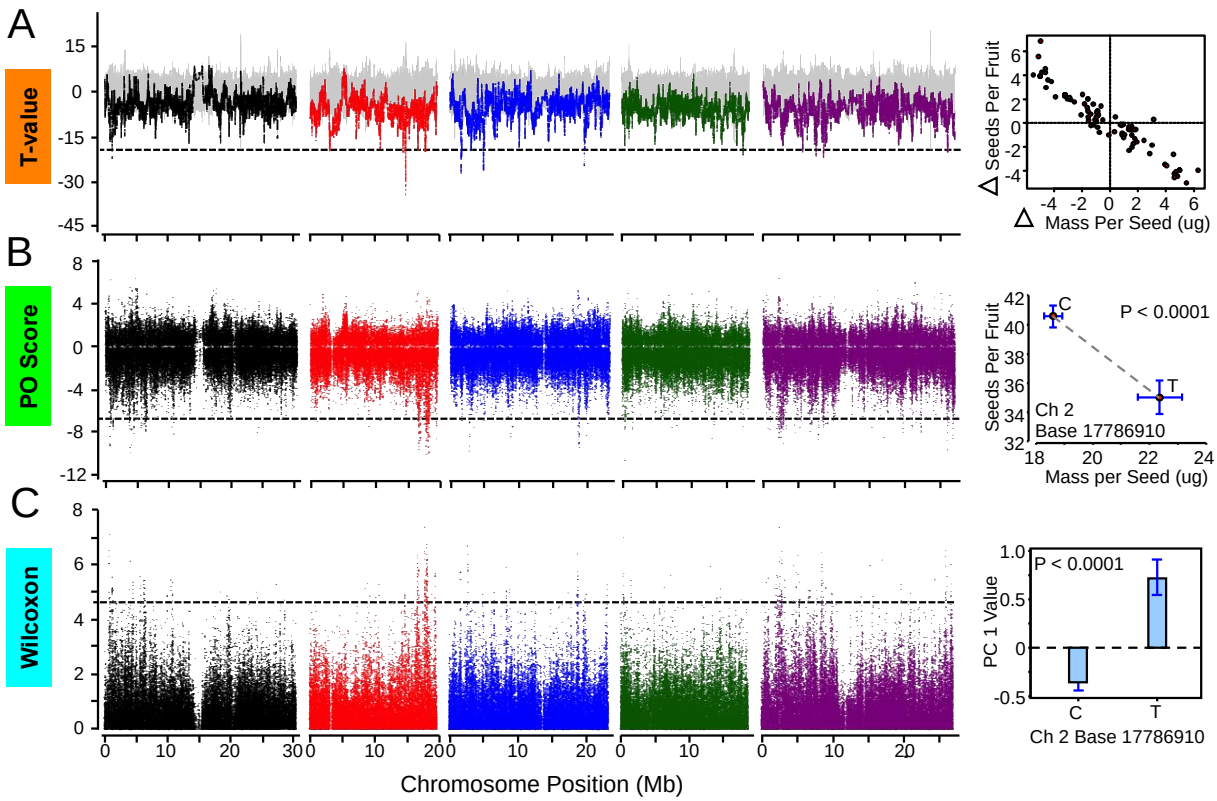


Figure 31. Three methods of genome-wide co-mapping
(A) Sliding regression method and the most significant SNP out of the map. **(B)** Polarized overlay method and the most significant SNP out of the map. **(C)** PCA-GWAs method and the most significant SNP out of the map.

Table 24. Phenotype values for trade-off co-map in 164 wild Arabidopsis accessions.

Name	Line	Seed Numebr per Pod	Mass per Seed
Ag-0	09C	45.3	21.7
Algutsrum	8230	37.8	19.4
An-1	08G	37.1	15.9
Ang-0	8254	42.1	14.1
Ba-1-2	8256	39.3	21.3
Ba-4-1	8258	53.7	21.2
Ba-5-1	8259	51.7	13.7
Bay-0	09F	33.7	21.8
Bil-5	02G	33.6	21.8
Bil-7	02H	34.4	21.8
Bla-1	8264	35.5	21
Blh-1	8265	42.3	12.3
Bor-1	04C	37.3	18.6
Bor-4	04D	60.3	20
Br-0	09A	32.5	19
Bro-1-6	8231	48.4	17.9
Bs-1	8270	31.1	21.4
Bu-0	8271	15.8	32.7
Bur-0	12E	22.8	26.2
C24	08A	40.6	19
Can-0	8274	45.5	23.4
Cen-0	8275	42	15
CIBC-17	05H	47	20.2
CIBC-5	05G	26.7	20.6
Col-0	08F	49.8	19.4
CS22491	08B	32.4	15.1
Ct-1	10D	58.2	13.8
Dem-4	8233	33	24.1
DraII-1	8284	39.7	19.5
DraIII-1	8285	40.1	16.3
Duk	6008	28.3	18.9
Eden-2	02B	33.4	27.6
Edi-0	12F	44.3	23.2
Eds-1	6016	48.3	29.2
Ei-2	07E	44.2	20.8
En-1	8290	45.5	17.2
Est-1	09B	31.3	16.2
Fab-2	02E	31.5	26.5
Fab-4	02F	32.8	29.4
Fei-0	11B	43.3	15.8
Ga-0	09G	28.4	18.6
Gd-1	8296	44	18.6
Ge-0	8297	42.3	23.1
GOT-22	06F	35.7	28
GOT-7	06E	44.6	23.8
Gr-1	8300	37.5	15.2
Gu-0	07F	37	19.7
Gy-0	09D	40.7	19.9
Hi-0	8304	39.9	16.6
Hod	8235	40.6	15.3
Hov-2-1	8423	41	20.6
Hov-4-1	8306	44.1	23.4
Hovdala-2	6039	34.8	22.7
HR-5	05A	55.4	15.2
Hs-0	8310	43.4	16.8
HSm	8236	33.5	18
In-0	8311	36.2	19.4
Is-0	8312	46.8	18.4
Jm-0	8313	41	22.5
Ka-0	8314	48.3	16.5
Kas-1	10C	34	23.5
Kavlinge-1	8237	51.9	16.6
Kelst-4	8420	57.4	18.6
Kent	8238	38.7	16.5
Kin-0	12C	48.7	16
Kni-1	6040	34.3	19.3
KNO-10	01C	33	24.4
KNO-18	01D	37.8	24.3
Koln	8239	36.5	16.4
Kondara	11H	34.3	19.1
Kulturen-1	8240	40.1	18
KZ-1	06C	32.3	15.2
KZ-9	06D	NA	11.8
Lc-0	8323	NA	10.1
Ler-1	07G	33.3	16.2
Liarum	8241	61.1	14.5
Lillo-1	8242	36.8	21.1
Lip-0	8325	20	19.3
Lis-1	8326	54.3	17
Lis-2	8222	32.7	18.4
Lisse	8430	41.1	20.2
LL-0	11G	NA	11.6
Lm-2	8329	36.7	16.9
Lom 1-1	6042	39.1	16.8
Lov-1	02C	37.5	30.2
Lov-5	02D	30	22.6
LP2-6	04H	41.6	16.2
Lu-1	8334	59.7	24.8
Lund	8335	39.7	26.4
Lz-0	07D	34	18.8
Mir-0	8337	49.8	14
Mrk-0	09H	17.2	27.4
Mt-0	10G	44.8	15.9
Mz-0	10A	40.9	16.6
Na-1	8343	29	20.6
Nd-1	07H	42.5	15
NFA-10	05D	42.8	17.5
NFA-8	05C	31.3	18.5
Nok-3	10H	47.3	19.2
NW-0	8348	40.5	16.5
Nyl-2	6064	38.4	29.8
Omo-2-1	03E	49	19.2
Omo-2-3	03F	43.6	14.5
Or-1	6074	32.8	22.8
Ost-0	8351	43.6	24
Oy-0	12G	29.6	22.3
Pa-1	8353	54.2	14.9
Per-1	8354	51.2	15
PHW-2	8243	42.3	22
Pla-0	8357	32.6	17.4
PNA-10	01H	36.9	23.5
Pro-0	11F	48.8	15.4
Pu2-23	04F	33.9	16.3
Pu2-7	04E	38.9	22.5
Ra-0	09E	37.4	16.4
Rak-2	8365	39.2	13
Rd-0	8366	43.8	18.5
Rd-0	8411	31.2	18.8
REN-1	06G	24.6	15
REN-11	06H	51.9	12.8
Rev-1	8369	41.3	18.5
RRS-10	01B	34.5	23.2
RRS-7	01A	48.6	21.2
Rsch-4	8374	34.5	16.6
San-2	8247	34.6	22
Sap-0	8378	42.8	16.2
Sav-0	8412	36.1	19.9
Seattle-0	8245	35.6	20.2
Shahdara	12A	30.7	18.2
Sorbo	12B	19.4	20.8
Spr-1-2	03C	44.9	15.4
Spr-1-6	03D	44.1	17.4
SQ-1	05E	40.7	15.8
SQ-8	05F	46.8	17.2
Sr-5	8386	31.1	19.1
St-0	8387	31.1	20.1
Stw-0	8388	39.4	19.5
Ta-0	8389	33	23.1
TAMM-2	06A	42.6	21.6
TAMM-27	06B	38.6	20.5
Tottarp-2	6243	35.2	18.5
Ts-5	11E	38	19.4
Tsu-1	10F	29.9	18.2
Tu-0	8395	39.8	18.1
Ull-1-1	8426	34.9	16.2
Ull-2-3	03H	37.3	15.3
Ull-2-5	03G	42.2	23.3
Uod-1	07A	38.1	16.7
Uod-2	8428	40.2	19
Uod-7	07B	32.4	14.9
Van-0	08H	27.7	16.3
Var-2-1	03A	27.3	35.7
Var-2-6	03B	22.7	41.8
Vastervik	9058	29.6	24.2
Vimmerby	8249	38.6	20.1
Vinslov	9057	40.1	16
Wa-1	11A	24.9	22.6
Wei-0	08C	38.6	20.9
Wil-1Dean	100000	43.8	16.6
Ws-2	12H	44.7	12.5
Wt-5	10B	37.1	18.1
Yo-0	08E	33	21.4
Zdr-1	04A	35.4	21.1
Zdr-6	04B	45.9	19.1

BIBLIOGRAPHY

Agrawal, A.A., Fishbein, M., Jetter, R., Salminen, J.-P., Goldstein, J.B., Freitag, A.E. and Sparks, J.P. (2009) Phylogenetic ecology of leaf surface traits in the milkweeds (Asclepias spp.): chemistry, ecophysiology, and insect behavior. *New Phytol.*, **183**, 848–867.

Alonso Blanco, C., Blankestijn-de Vries, H., Hanhart, C.J. and Koornneef, M. (1999) Natural allelic variation at seed size loci in relation to other life history traits of Arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 4710–4717.

Alonso, J.M., Hirayama, T., Roman, G., Nourizadeh, S. and Ecker, J.R. (1999) EIN2, a bifunctional transducer of ethylene and stress responses in Arabidopsis. *Science*, **284**, 2148–2152.

Alonso, J.M., Stepanova, A.N., Leisse, T.J., et al. (2003) Genome-wide insertional mutagenesis of Arabidopsis thaliana. *Science*, **301**, 653–657.

Alonso-Blanco, C. and Koornneef, M. (2000) Naturally occurring variation in Arabidopsis: an underexploited resource for plant genetics. *Trends in Plant Science*, **5**, 22–29.

Alonso-Blanco, C., Peeters, A.J., Koornneef, M., Lister, C., Dean, C., van den Bosch, N., Pot, J. and Kuiper, M.T. (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. *The Plant Journal*, **14**, 259–271.

Aranzana, M.J., Kim, S., Zhao, K., et al. (2005) Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. *PLoS Genet*, **1**, e60.

Atwell, S., Huang, Y.S., Vilhjálmsson, B.J., et al. (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, **465**, 627–631.

Ågren, J. and Schemske, D.W. (1993) The cost of defense against herbivores: an experimental study of trichome production in Brassica rapa. *Am. Nat.*, **141**, 338–350.

Bazzaz, F.A., Chiariello, N.R., Coley, P.D. and Pitelka, L.F. (1987) Allocating resources to reproduction and defense. *BioScience*, **37**, 58–67.

Bednarek, P., Kwon, C. and Schulze-Lefert, P. (2010) Not a peripheral issue: secretion in plant-microbe interactions. *Current Opinion in Plant Biology*, **13**, 378–387.

Bennett, B.J., Farber, C.R., Orozco, L., et al. (2010) A high-resolution association mapping panel for the dissection of complex traits in mice. *Genome Research*, **20**, 281–290.

Bergelson, J. and Roux, F. (2010) Towards identifying genes underlying ecologically relevant traits in Arabidopsis thaliana. *Nature Reviews Genetics*, **11**, 867–879.

Bergelson, J., Purrington, C.B., Palm, C.J. and López-Gutiérrez, J.C. (1996) Costs of resistance: a test using transgenic Arabidopsis thaliana. *Proc. Biol. Sci.*, **263**, 1659–1663.

Bolnick, D.I., Amarasekare, P., Araújo, M.S., et al. (2011) Why intraspecific trait variation matters in community ecology. *Trends in Ecology & Evolution*, **26**, 183–192.

Boursiac, Y., Léran, S., Corratgé-Faillie, C., Gojon, A., Krouk, G. and Lacombe, B. (2013) ABA transport and transporters. *Trends in Plant Science*, **18**, 325–333.

Brachi, B., Faure, N., Horton, M., Flahauw, E., Vazquez, A., Nordborg, M., Bergelson, J., Cuguen, J. and Roux, F. (2010) Linkage and Association Mapping of Arabidopsis thaliana Flowering Time in Nature. *PLoS Genet*, **6**, e1000940.

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

Chan, E.K.F., Rowe, H.C. and Kliebenstein, D.J. (2010) Understanding the Evolution of Defense Metabolites in Arabidopsis thaliana Using Genome-wide Association Mapping. *Genetics*, **185**, 991–1007.

Chan, E.K.F., Rowe, H.C., Corwin, J.A., Joseph, B. and Kliebenstein, D.J. (2011) Combining genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in Arabidopsis thaliana. *PLoS Biol*, **9**, e1001125.

Chao, D.-Y., Silva, A., Baxter, I., Huang, Y.S., Nordborg, M., Danku, J., Lahner, B., Yakubova, E. and Salt, D.E. (2012) Genome-wide association studies identify heavy metal ATPase3 as the primary determinant of natural variation in leaf cadmium in Arabidopsis thaliana. *PLoS Genet*, **8**, e1002923.

Clark, R.M., Schweikert, G., Toomajian, C., et al. (2007) Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. *Science*, **317**, 338–342.

Clough, S.J. and Bent, A.F. (1998) Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. *Plant J.*, **16**, 735–743.

Curtis, M.D. and Grossniklaus, U. (2003) A gateway cloning vector set for high-throughput functional analysis of genes in planta. *Plant Physiol*, **133**, 462–469.

de Torres Zabala, M., Bennett, M.H., Truman, W.H. and Grant, M.R. (2009) Antagonism between salicylic and abscisic acid reflects early host-pathogen conflict and moulds plant defence responses. *The Plant Journal*, **59**, 375–386.

de Torres Zabala, M., Truman, W., Bennett, M.H., Lafforgue, G., Mansfield, J.W., Rodriguez Egea, P., Bögre, L. and Grant, M. (2007) Pseudomonas syringae pv. tomato hijacks the Arabidopsis abscisic acid signalling pathway to cause disease. *EMBO J*, **26**, 1434–1443.

Dean, J.V. and Mills, J.D. (2004) Uptake of salicylic acid 2-O-beta-D-glucose into soybean tonoplast vesicles by an ATP-binding cassette transporter-type mechanism. *Physiol Plant*, **120**, 603–612.

Ehrenreich, I.M., Hanzawa, Y., Chou, L., Roe, J.L., Kover, P.X. and Purugganan, M.D. (2009) Candidate gene association mapping of Arabidopsis flowering time. *Genetics*, **183**, 325–335.

Endo, A., Sawada, Y., Takahashi, H., et al. (2008) Drought induction of Arabidopsis 9-cis-epoxycarotenoid dioxygenase occurs in vascular parenchyma cells. *Plant Physiol*, **147**, 1984–1993.

Fang, W., Wang, Z., Cui, R., Li, J. and Li, Y. (2012) Maternal control of seed size by EOD3/CYP78A6 in Arabidopsis thaliana. *The Plant Journal*, **70**, 929–939.

Filiault, D.L. and Maloof, J.N. (2012) A Genome-Wide Association Study Identifies Variants Underlying the Arabidopsis thaliana Shade Avoidance Response R. Mauricio, ed. *PLoS Genet*, **8**, e1002589.

Flint-Garcia, S., Thornsberry, J., S, E. and IV, B. (2003) S TRUCTURE OF L INKAGE D ISEQUILIBRIUM IN P LANTS*. *Annu. Rev. Plant Biol.*, **54**, 357–374.

Fournier-Level, A., Korte, A., Cooper, M. and Nordborg, M. (2011) A Map of Local Adaptation in Arabidopsis thaliana. *Science*.

Hancock, A.M., Brachi, B., Faure, N., Horton, M.W., Jarymowycz, L.B., Sperone, F.G., Toomajian, C., Roux, F. and Bergelson, J. (2011) Adaptation to climate across the Arabidopsis thaliana genome. *Science*, **334**, 83–86.

Hare, J.D., Elle, E. and Van Dam, N.M. (2003) Costs of glandular trichomes in Datura wrightii: a three-year study. *Evolution*, **57**, 793–805.

Herridge, R.P., Day, R.C., Baldwin, S. and Macknight, R.C. (2011) Rapid analysis of seed size in Arabidopsis for mutant and QTL discovery. **7**, 3.

Hilscher, J., Schlötterer, C. and Hauser, M.-T. (2009) A single amino acid replacement in ETC2 shapes trichome patterning in natural Arabidopsis populations. *Curr. Biol.*, **19**, 1747–1751.

Horton, M.W., Hancock, A.M., Huang, Y.S., et al. (2012) Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. *Nat Genet*, **44**, 212–216.

Ibdah, M., Chen, Y.-T., Wilkerson, C.G. and Pichersky, E. (2009) An aldehyde oxidase in developing seeds of Arabidopsis converts benzaldehyde to benzoic Acid. *Plant Physiol*, **150**, 416–423.

Ishiga, Y., Ishiga, T., Uppalapati, S.R. and Mysore, K.S. (2011) Arabidopsis seedling flood-inoculation technique: a rapid and reliable assay for studying plant-bacterial interactions. *Plant Methods*, **7**, 32.

Jako, C., Kumar, A., Wei, Y., Zou, J., Barton, D.L., Giblin, E.M., Covello, P.S. and Taylor, D.C. (2001) Seed-specific over-expression of an Arabidopsis cDNA encoding a diacylglycerol acyltransferase enhances seed oil content and seed weight. *Plant Physiol*, **126**, 861–874.

Jefferson, R.A., Kavanagh, T.A. and Bevan, M.W. (1987) GUS fusions: beta-glucuronidase as a sensitive and versatile gene fusion marker in higher plants. *EMBO J*, **6**, 3901.

Jetz, W., Sekercioglu, C.H. and Böhning-Gaese, K. (2008) The worldwide variation in avian clutch size across species and space. *PLoS Biol*, **6**, 2650–2657.

Jofuku, K.D., Omidyar, P.K., Gee, Z. and Okamuro, J.K. (2005) Control of seed mass and seed yield by the floral homeotic gene APETALA2. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 3117–3122.

Jones, J.D.G. and Dangl, J.L. (2006) The plant immune system. *Nature*, **444**, 323–329.

Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J. and Eskin, E. (2008) Efficient control of population structure in model organism association mapping. *Genetics*, **178**, 1709–1723.

Kang, J., Hwang, J.-U., Lee, M., Kim, Y.-Y., Assmann, S.M., Martinoia, E. and Lee, Y. (2010) PDR-type ABC transporter mediates cellular uptake of the phytohormone abscisic acid. *Proceedings of the National Academy of Sciences*, **107**, 2355–2360.

Kawagoe, T., Shimizu, K.K., Kakutani, T. and Kudoh, H. (2011) Coexistence of trichome variation in a natural plant population: a combined study using ecological and candidate gene approaches. *PLoS One*, **6**, e22184.

Kärkkäinen, K., Løe, G. and Ågren, J. (2004) Population structure in Arabidopsis lyrata: evidence for divergent selection on trichome production. *Evolution*, **58**, 2831–2836.

Kim, K.-C., Lai, Z., Fan, B. and Chen, Z. (2008) Arabidopsis WRKY38 and WRKY62 transcription factors interact with histone deacetylase 19 in basal defense. *Plant Cell*, **20**, 2357–2371.

Kirik, V., Simon, M., Wester, K., Schiefelbein, J. and Hulskamp, M. (2004) ENHANCER of TRYand CPC 2 (ETC2) reveals redundancy in the region-specific control of trichome development of Arabidopsis. *Plant Mol Biol*, **55**, 389–398.

Kleinboelting, N., Huep, G., Kloetgen, A., Viehoever, P. and Weisshaar, B. (2012) GABI-Kat SimpleSearch: new features of the Arabidopsis thaliana T-DNA mutant database. *Nucleic acids research*, **40**, D1211–5.

Kobae, Y., Sekino, T., Yoshioka, H., Nakagawa, T., Martinoia, E. and Maeshima, M. (2006) Loss of AtPDR8, a plasma membrane ABC transporter of Arabidopsis thaliana, causes hypersensitive cell death upon pathogen infection. *Plant and cell physiology*, **47**, 309–318.

Korbel, J.O., Urban, A.E., Affourtit, J.P., et al. (2007) Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. *Science*, **318**, 420–426.

Kuromori, T. and Shinozaki, K. (2010) ABA transport factors found in Arabidopsis ABC transporters. *psb*, **5**, 1124–1126.

Kuromori, T., Miyaji, T., Yabuuchi, H., Shimizu, H., Sugimoto, E., Kamiya, A., Moriyama, Y. and Shinozaki, K. (2010) ABC transporter AtABCG25 is involved in abscisic acid transport and responses. *Proceedings of the National Academy of Sciences*, **107**, 2361–2366.

Kuromori, T., Sugimoto, E. and Shinozaki, K. (2011) Arabidopsis mutants of AtABCG22, an ABC transporter gene, increase water transpiration and drought susceptibility. *The Plant Journal*, **67**, 885–894.

Larkin, J., Young, N., Prigge, M. and Marks, M. (1996) The control of trichome spacing and number in Arabidopsis. *Development*, **122**, 997.

Le, B.H., Cheng, C., Bui, A.Q., et al. (2010) Global analysis of gene activity during Arabidopsis seed development and identification of seed-specific transcription factors. *Proceedings of the National Academy of Sciences*, **107**, 8063–8070.

Li, Y., Huang, Y., Bergelson, J., Nordborg, M. and Borevitz, J.O. (2010) Association mapping of local climate-sensitive quantitative trait loci in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 21199–21204.

Lister, C. and Dean, C. (1993) Recombinant inbred lines for mapping RFLP and phenotypic markers in Arabidopsis thaliana. *The Plant Journal*, **4**, 745–750.

Lister, R., Gregory, B.D. and Ecker, J.R. (2009) Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. *Current Opinion in Plant Biology*, **12**, 107–118.

Luo, M., Dennis, E., Berger, F., Peacock, W. and Chaudhury, A. (2005) MINISEED3 (MINI3), a WRKY family gene, and HAIKU2 (IKU2), a leucine-rich repeat (LRR)

KINASE gene, are regulators of seed size in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 17531.

Margulies, M., Egholm, M., Altman, W.E., et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Marioni, J.C., Mason, C.E., Mane, S.M. and Stephens, M. (2008) RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome …*.

Marquis, R.J. (1984) Leaf herbivores decrease fitness of a tropical plant. *Science*, **226**, 537–539.

Martínez-Andújar, C., Martin, R.C. and Nonogaki, H. (2012) Seed traits and genes important for translational biology--highlights from recent discoveries. *Plant and cell physiology*, **53**, 5–15.

Mauricio, R. (1998) Costs of resistance to natural enemies in field populations of the annual plant Arabidopsis thaliana. *Am. Nat.*, **151**, 20–28.

Mauricio, R. (2001) Mapping quantitative trait loci in plants: uses and caveats for evolutionary biology. *Nature Reviews Genetics*, **2**, 370–381.

Melotto, M., Underwood, W., Koczan, J., Nomura, K. and He, S.Y. (2006) Plant stomata function in innate immunity against bacterial invasion. *Cell*, **126**, 969–980.

Mizukami, Y. and Fischer, R.L. (2000) Plant organ size control: AINTEGUMENTA regulates growth and cell numbers during organogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 942–947.

Moles, A.T., Ackerly, D.D., Webb, C.O., Tweddle, J.C., Dickie, J.B. and Westoby, M. (2005) A brief history of seed size. *Science*, **307**, 576–580.

Moles, A.T., Ackerly, D.D., Tweddle, J.C., et al. (2007) Global patterns in seed size. *Global Ecology and Biogeography*, **16**, 109–116.

Moore, C.R., Gronwall, D.S., Miller, N.D. and Spalding, E.P. (2013) Mapping quantitative trait loci affecting Arabidopsis thaliana seed morphology features extracted computationally from images. *G3 (Bethesda)*, **3**, 109–118.

Morris, C.E., Sands, D.C., Vinatzer, B.A., Glaux, C., Guilbaud, C., Buffière, A., Yan, S., Dominguez, H. and Thompson, B.M. (2008) The life history of the plant pathogen Pseudomonas syringae is linked to the water cycle. *ISME J*, **2**, 321–334.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, **5**, 621–628.

Munemasa, S., Oda, K., Watanabe-Sugimoto, M., Nakamura, Y., Shimoishi, Y. and Murata, Y. (2007) The coronatine-insensitive 1 mutation reveals the hormonal signaling

interaction between abscisic acid and methyl jasmonate in Arabidopsis guard cells. Specific impairment of ion channel activation and second messenger production. *Plant Physiol*, **143**, 1398–1407.

Nam, H.G., Giraudat, J., Boer, Den, B., Moonan, F., Loos, W., Hauge, B.M. and Goodman, H.M. (1989) Restriction Fragment Length Polymorphism Linkage Map of Arabidopsis thaliana. *THE PLANT CELL ONLINE*, **1**, 699–705.

Nambara, E. and Nonogaki, H. (2012) Seed biology in the 21st century: perspectives and new directions. *Plant and cell physiology*, **53**, 1–4.

Nordborg, M. and Weigel, D. (2008) Next-generation genetics in plants. *Nature*, **456**, 720–723.

Nordborg, M., Hu Tina, T., Ishino, Y., et al. (2005) The pattern of polymorphism in Arabidopsis thaliana. *PLoS Biol*, **3**, e196.

Oerke, E.-C. (2006) Crop losses to pests. *JOURNAL OF AGRICULTURAL SCIENCE-CAMBRIDGE-*, **144**, 31.

Ohto, M.-A., Fischer, R.L., Goldberg, R.B., Nakamura, K. and Harada, J.J. (2005) Control of seed mass by APETALA2. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 3123–3128.

Pandey, S., Wang, R.-S., Wilson, L., Li, S., Zhao, Z., Gookin, T.E., Assmann, S.M. and Albert, R. (2010) Boolean modeling of transcriptome data reveals novel modes of heterotrimeric G-protein action. *Mol. Syst. Biol.*, **6**, 372.

Pickrell, J.K., Marioni, J.C., Pai, A.A., et al. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, **464**, 768–772.

Platt, A., Horton, M., Huang, Y.S., et al. (2010) The scale of population structure in Arabidopsis thaliana. *PLoS Genet*, **6**, e1000843.

Rea, P.A. (2007) Plant ATP-binding cassette transporters. *Annu. Rev. Plant Biol.*, **58**, 347–375.

Ren, X., Chen, Z., Liu, Y., Zhang, H., Zhang, M., Liu, Q., Hong, X., Zhu, J.-K. and Gong, Z. (2010) ABO3, a WRKY transcription factor, mediates plant responses to abscisic acid and drought tolerance in Arabidopsis. *The Plant Journal*.

Schenk, P.M., Kazan, K., Wilson, I., Anderson, J.P., Richmond, T., Somerville, S.C. and Manners, J.M. (2000) Coordinated plant defense responses in Arabidopsis revealed by microarray analysis. *Proceedings of the National Academy of Sciences*, **97**, 11655–11660. Available at: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=11027363&retmode=ref&cmd=prlinks.

Schruff, M.C., Spielman, M., Tiwari, S., Adams, S., Fenby, N. and Scott, R.J. (2006) The AUXIN RESPONSE FACTOR 2 gene of Arabidopsis links auxin signalling, cell division, and the size of seeds and other organs. *Development*, **133**, 251–261.

Schwab, R., Ossowski, S., Riester, M., Warthmann, N. and Weigel, D. (2006) Highly specific gene silencing by artificial microRNAs in Arabidopsis. *Plant Cell*, **18**, 1121–1133.

Seo, M., Aoki, H., Koiwai, H., Kamiya, Y., Nambara, E. and Koshiba, T. (2004) Comparative studies on the Arabidopsis aldehyde oxidase (AAO) gene family revealed a major role of AAO3 in ABA biosynthesis in seeds. *Plant and cell physiology*, **45**, 1694–1703.

Sreenivasulu, N. and Wobus, U. (2013) Seed-development programs: a systems biology-based comparison between dicots and monocots. *Annu. Rev. Plant Biol.*, **64**, 189–217.

Steets, J.A., Takebayashi, N., Byrnes, J.M. and Wolf, D.E. (2010) Heterogeneous selection on trichome production in Alaskan Arabidopsis kamchatica (Brassicaceae). *Am. J. Bot.*, **97**, 1098–1108.

Stein, M., Dittgen, J., Sánchez-Rodríguez, C., Hou, B.-H., Molina, A., Schulze-Lefert, P., Lipka, V. and Somerville, S. (2006) Arabidopsis PEN3/PDR8, an ATP binding cassette transporter, contributes to nonhost resistance to inappropriate pathogens that enter by direct penetration. *Plant Cell*, **18**, 731–746.

Strader, L.C. and Bartel, B. (2009) The Arabidopsis PLEIOTROPIC DRUG RESISTANCE8/ABCG36 ATP binding cassette transporter modulates sensitivity to the auxin precursor indole-3-butyric acid. *Plant Cell*, **21**, 1992–2007.

Sun, W. (2011) A Statistical Framework for eQTL Mapping Using RNA-seq Data. *Biometrics*, **68**, 1–11.

Sun, W. and Hu, Y. (2013) eQTL Mapping Using RNA-seq Data. *Stat Biosci*, **5**, 198–219.

Symonds, V., Godoy, A., Alconada, T., Botto, J., Juenger, T., Casal, J. and Lloyd, A. (2005) Mapping quantitative trait loci in multiple populations of Arabidopsis thaliana identifies natural allelic variation for trichome density. *Genetics*, **169**, 1649.

Thibaud-Nissen, F., Wu, H., Richmond, T., Redman, J.C., Johnson, C., Green, R., Arias, J. and Town, C.D. (2006) Development of Arabidopsis whole-genome microarrays and their application to the discovery of binding sites for the TGA2 transcription factor in salicylic acid-treated plants. *Plant J.*, **47**, 152–162.

Todesco, M., Balasubramanian, S., Hu Tina, T., et al. (2010) Natural allelic variation underlying a major fitness trade-off in Arabidopsis thaliana. *Nature*, **465**, 632–636.

Ukitsu, H., Kuromori, T., Toyooka, K., et al. (2007) Cytological and biochemical analysis of COF1, an Arabidopsis mutant of an ABC transporter gene. *Plant and cell physiology*, **48**, 1524–1533.

Van Daele, I., Gonzalez, N., Vercauteren, I., De Smet, L., Inzé, D., Roldán-Ruiz, I. and Vuylsteke, M. (2012) A comparative study of seed yield parameters in Arabidopsis thaliana mutants and transgenics. *Plant Biotechnology Journal*, **10**, 488–500.

Verrier, P.J., Bird, D., Burla, B., et al. (2008) Plant ABC proteins--a unified nomenclature and updated inventory. *Trends in Plant Science*, **13**, 151–159.

Vlot, A.C., Dempsey, D.A. and Klessig, D.F. (2009) Salicylic Acid, a multifaceted hormone to combat disease. *Annu. Rev. Phytopathol.*, **47**, 177–206.

Wang, S., Kwak, S.-H., Zeng, Q., Ellis, B.E., Chen, X.-Y., Schiefelbein, J. and Chen, J.-G. (2007) TRICHOMELESS1 regulates trichome patterning by suppressing GLABRA1 in Arabidopsis. *Development*, **134**, 3873–3882.

Weigel, D. (2012) Natural variation in Arabidopsis: from molecular genetics to ecological genomics. *Plant Physiol*, **158**, 2–22.

Weigel, D. and Mott, R. (2009) The 1001 Genomes Project for Arabidopsis thaliana. *Genome Biol*, **10**, 107.

Westoby, M., Falster, D.S., Moles, A.T., Vesk, P.A. and Wright, I.J. (2002) Plant ecological strategies: some leading dimensions of variation between species| Macquarie University ResearchOnline.

Winter, D., Vinegar, B., Nahal, H., Ammar, R., Wilson, G.V. and Provart, N.J. (2007) An "electronic fluorescent pictograph" browser for exploring and analyzing large-scale biological data sets. *PLoS One*, **2**, e718.

Yang, Y., Costa, A., Leonhardt, N., Siegel, R.S. and Schroeder, J.I. (2008) Isolation of a strong Arabidopsis guard cell promoter and its potential as a research tool. *Plant Methods*, **4**, 6.

Zeng, W. and He, S.Y. (2010) A prominent role of the flagellin receptor FLAGELLIN-SENSING2 in mediating stomatal response to Pseudomonas syringae pv tomato DC3000 in Arabidopsis. *Plant Physiol*, **153**, 1188–1198.

Zeng, Z.B. (1994) Precision mapping of quantitative trait loci. *Genetics*, **136**, 1457–1468.

Zheng, X.-Y., Spivey, N.W., Zeng, W., Liu, P.-P., Fu, Z.Q., Klessig, D.F., He, S.Y. and Dong, X. (2012) Coronatine promotes Pseudomonas syringae virulence in plants by activating a signaling cascade that inhibits salicylic acid accumulation. *Cell Host Microbe*, **11**, 587–596.

Züst, T., Joseph, B., Shimizu, K.K., Kliebenstein, D.J. and Turnbull, L.A. (2008) Identification of indole glucosinolate breakdown products with antifeedant effects on Myzus persicae (green peach aphid). *The Plant Journal*, **278**, 1015–1026.