Imaging genetics of seed performance

Ronny Viktor Louis Joosen

Thesis committee

Promotor

Prof. dr. L.H.W. van der Plas
Professor of Plant Physiology
Wageningen University

Co-promotors

Dr. H.W.M. Hilhorst
Associate Professor, Laboratory of Plant Physiology
Wageningen University

Dr. W. Ligterink
Researcher, Laboratory of Plant Physiology
Wageningen University

Other members

Prof. dr. M. Koornneef, Wageningen University / Max Planck Institut, Germany
Prof. dr. F.A. van Eeuwijk, Wageningen University
Dr. J.M. Jiménez-Gómez, Max Planck Institut, Germany
Dr. J. Buitink, Institut National de la Recherche Agronomique, France

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

Imaging genetics of seed performance

Ronny Viktor Louis Joosen

Thesis

Submitted in fulfillment of the requirements for the degree of doctor

at Wageningen University

by the authority of the Rector Magnificus

Prof. dr. M.J. Kropff,

in the presence of the

Thesis Committee appointed by the Academic Board

to be defended in public

on Friday 24 May 2013

at 1.30 p.m. in the Aula.

Ronny Joosen

Imaging genetics of seed performance

196 pages

Thesis, Wageningen University, Wageningen, NL (2013)

With references, with summaries in English and Dutch

ISBN 978-94-6173-497-6

In loving memory of my mother

Liza Joosen – Huijgens

Forever you will flower
because I've planted you within my soul

CONTENTS

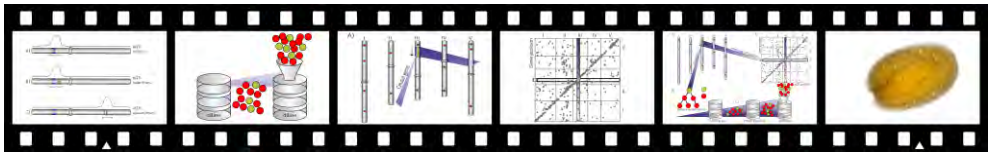
“Don't stop sowing just because the birds ate a few seeds.”

1 GENERAL INTRODUCTION

Adapted from two published reviews:

Joosen RVL, Ligterink W, Hilhorst HWM, Keurentjes JJ (2009) Advances in genetical genomics of plants. Current Genomics. Vol. 10: 540-549.

Ligterink W, Joosen RVL, Hilhorst HWM (2012) Unravelling the complex trait of seed quality: using natural variation through a combination of physiology, genetics and -omics technologies. Seed Science Research. Vol. 22: S45–S52.



Abstract

Seed quality is a complex trait that is the result of a large variety of developmental processes. The molecular-genetic dissection of these seed processes and their relationship with seed and seedling phenotypes will allow the identification of the regulatory genes and signaling pathways involved and thus, provide the means to predict and enhance seed quality. Natural variation for seed quality aspects found in recombinant inbred line (RIL) populations is a great resource to help unraveling the complex networks involved in the acquisition of seed quality. Besides extensive phenotyping, RILs can also be profiled by 'omics' technologies like transcriptomics, proteomics and metabolomics in a sophisticated so-called generalized genetical genomics approach. This combined use of physiology, genetics and several 'omics' technologies, followed by advanced data analysis allows the construction of regulatory networks involved in the different aspects of seed and seedling quality. This type of analysis of the genetic variation in RIL populations in combination with genome wide association (GWA) studies will allow a relatively quick identification of genes that are responsible for quality related traits of seeds and seedlings. New developments in several 'omics' technologies, especially the fast evolving next generation sequencing techniques, will make a similar system wide approach more applicable to non-model species in the near future and this will be a huge boost for the possibilities to breed for seed quality.

Seed Quality

Seed quality is a complex trait and comprises many different attributes describing the condition of a seed batch. These attributes include germination characteristics, dormancy, seed and seedling vigor, uniformity in seed size, normal embryo- and seedling morphology, storability, absence of mechanical damage, as well as the ability to develop into a normal plant (Dickson 1980; Hilhorst and Toorop 1997). Seed quality is largely established during seed development and maturation, as a result of, often complex, interactions between the genome and the environment. This mechanism is part of the normal adaptation of plants to a varying environment and is aimed at maximizing the probability of successful offspring (Huang et al. 2010).

The practical definition of seed quality is determined by the end user and therefore, will differ substantially, depending on the use of seeds as propagule or commodity. For farmers or plant growers high quality seeds are those seeds that germinate and produce seedlings to a high percentage under a wide range of field conditions. On the other hand, high quality seeds for use in the food industry may be seeds with a high starch or oil content or oil seeds with a specific fatty acid composition (Nesi *et al.* 2008). As a result of the complexity of seed quality, testing for seed quality in order to predict subsequent behavior in the field is troublesome and at best an 'educated guess' (Powell 2006). Therefore, seed producers have included additional attributes to the term seed quality such as usable plants and seedling and crop establishment. The trait 'usable plants' is one of the main characteristics of seed quality used by seed producers and plant growers.

Seed companies may enhance seed quality at all the different steps of the production process. At present seed companies try to obtain the best possible seeds mainly by varying the time and method of harvest, but especially by post-harvest treatments such as cleaning, sorting, coating and priming and controlling the storage conditions. Besides these methods, seed quality can also be improved by controlling the production environment. It is known that seed quality is largely acquired during seed development and particularly during the maturation phase by the successive acquisition of seed quality parameters such as germinability, desiccation tolerance, dormancy, vigor and longevity (Harada 1997) and that the environmental conditions during development have a huge impact on these different seed quality aspects. As a result, quality of different seed lots that are produced in different seasons and locations will vary. Nevertheless, influencing production environments is difficult, even under greenhouse conditions. Furthermore, since there is a complex interaction between the genome and the environment during development, the final effect of the environment on seed quality is difficult to determine and still largely unknown. Finally, the genetic component of the interaction between the genome and the environment can be investigated and this variation in genetic adaptation provides great opportunities for seed companies to breed for seed quality.

Natural variation for seed quality

Although abundant natural variation for seed quality exists, genetic components of seed quality have hardly been used in breeding programs. Exploiting natural variation is a powerful way to find the genes influencing important physiological processes. There are several ways to exploit natural variation, but in plants QTL (Quantitative Trait Loci) analysis of recombinant inbred line (RIL) populations have been widely used. In this type of analysis, linkage is sought between the genetic variety and the variation of phenotypic traits in the different RILs (Alonso-Blanco and Koornneef 2000) whereby the QTL represent the genomic regions explaining the phenotypic variation that is identified in this way. QTL analysis in plants has revealed a long list of genomic regions with variation for a broad variety of phenotypes and several of the genes underlying these QTLs have been cloned (reviewed in (Salvi and Tuberosa 2005; Gupta *et al.* 2009)). The complex nature of the trait seed quality makes it a perfect trait to decipher with a QTL approach, particularly because different aspects of seed quality have been proven to have sufficient natural variation to tackle this subject. In *Arabidopsis thaliana* different QTLs were found for dormancy (Bentsink *et al.* 2010) and several germination characteristics (Clerkx *et al.* 2004; Galpaz and Reymond 2010; Joosen *et al.* 2010). In tomato, different QTLs for germination characteristics under stress (Foolad *et al.* 2003; Foolad *et al.* 2007) and for seed size (Doganlar *et al.* 2000) have been identified. In *Medicago truncatula* several QTLs were identified for germination at extreme temperatures (Dias *et al.* 2011) and germination and seedling growth under osmotic stress (Vandecasteele *et al.* 2011). Zeng *et al.* (2006) have identified QTLs for seed storability in rice and in lettuce QTLs have been identified for several germination characteristics including thermo inhibition (Argyris *et al.* 2008).

In spite of these and other studies on specific aspects of seed quality, a systematic study of the genetics of seed quality is still lacking. A more systematic approach studying genetic populations differing in seed and seedling quality parameters will provide valuable insight in the involvement of genes, and the processes they control, in the acquisition of seed quality. Until now, only a few QTL positions have been cloned and characterized in detail, but if genes or gene sets associated with seed quality parameters become available, they may be used as diagnostic tools to assess seed quality, in marker-assisted breeding, or in genetic modification to enhance seed quality.

High throughput phenotyping

With the developments in sequencing technologies that enable fast and relative inexpensive genotyping and expression analysis, accurate phenotyping is becoming the limiting step in studying large genetic populations. To overcome this problem several initiatives have been taken to enhance phenotyping, mainly by implementing high-throughput phenotyping platforms for analyzing plant morphology as in the Australian 'High Resolution Plant Phenomics Centre' (HRPPC) (www.plantphenomics.org/hrppc), and

the Lemnatec systems (www.lemnatec.de) that perform fully automated imaging and subsequent data extraction of growing plants. For the systemic analysis of the different aspects of seed quality several (semi-)automatic phenotyping systems can be used. One of the most important aspects is the (semi-)automatic scoring of germination. Several methods to achieve this have been reported by Dell'Aquila (2009) and, more recently, by Joosen *et al.* (2010), who introduced the GERMINATOR package. Furthermore analysis of seedling shape and growth with systems like that of the previously mentioned HRPPC and Lemnatec and analysis of the root architecture of the seedlings with programs such as EZ-Rhizo (Armengaud *et al.* 2009) and Roottrace (French *et al.* 2009) will become important for the in depth analysis of seed quality.

Genetical genomics: omics QTL analysis

Fine mapping of QTL is a crucial step for plant breeding as genetic drag should be minimized in every step during the breeding process. Furthermore cloning of genes responsible for the QTL can provide great insight in the molecular mechanism underlying the adaptation. Although the causal genes for several seed-quality QTLs have been cloned and more are underway (Salvi and Tuberosa 2005), fine-mapping and ultimate cloning of these genes is very labor-intensive and time-consuming. Therefore classical QTL analysis can be considered as a low throughput technique.

Like for many physiological traits, variation in gene expression often shows a quantitative distribution, hence, all the classical statistical tools and concepts for QTL mapping can be applied for its genetic dissection. Thus, subjecting expression variation to linkage analysis identifies genetic regulatory loci, and ideally genes, explaining the observed variation. Knowing the position of genes and their corresponding expression QTLs (eQTLs) renders great opportunities for dissecting quantitative traits. This was first recognized by Jansen and Nap (2001) who outlined a concept, coined 'genetical genomics', in which the combination of a genotyped segregating population (*i.e.* genetics) and genome-wide expression profiling (*i.e.* genomics) is used to formulate hypothetic regulatory pathways and unravel complex traits in a more high-throughput manner. Analogously, similar approaches can be followed for data derived from other 'omic' technologies such as proteomics (pQTLs) and metabolomics (mQTLs) (Keurentjes *et al.* 2008).

The first study reporting a proof of principle of genetical genomics was performed in *Saccharomyces cerevisiae* (Brem *et al.* 2002). In a relatively small population of 40 haploid segregants from a cross between a laboratory and a wild type strain, it was shown that parental differences in gene expression were highly heritable and amenable to genetic mapping. This first report was quickly followed by more comprehensive eQTL studies in higher eukaryotes (Schadt *et al.* 2003) and has now been applied in a broad range of taxonomic groups including yeast (Brem *et al.* 2002; Yvert *et al.* 2003; Bing and Hoeschele 2005; Leach *et al.* 2007), nematodes (Li *et al.* 2006), insects (Wittkopp *et al.* 2004; Hsieh *et al.* 2007), plants (DeCook *et al.* 2006; Keurentjes *et al.* 2007; West *et al.* 2007; Potokina *et*

al. 2008), rodents (Bystrykh *et al.* 2005; Chesler *et al.* 2005; Hubner *et al.* 2005) and humans (Monks *et al.* 2004; Dixon *et al.* 2007; Göring *et al.* 2007; Myers *et al.* 2007; Stranger *et al.* 2007; Emilsson *et al.* 2008). All studies demonstrated the power of combining gene expression and genetic analyses to refine molecular pathways involved in complex phenotypes and to identify key driver genes thereof. Moreover, they have shown general and conserved mechanisms of expression regulation which improved our understanding of adaptive strategies and evolutionary concepts (Wittkopp *et al.* 2004; Mitchell-Olds and Schmitt 2006).

Genetic architecture of gene expression variation

The detection of eQTLs depends on a number of factors, which together determine the proportion of genetically regulated genes that can be observed. First, biological factors such as the assayed tissue, developmental stage or environmental conditions and the genotypic diversity present in the mapping population determine which genes are expressed and exhibit allelic variants, respectively. Second, statistical issues like population type and size, genetic map quality, measurement accuracy and the number of genes analyzed determine mapping power and detection thresholds. Because all these aspects vary between different experiments, reported fractions of regulated genes range from only a handful to over 50% of the total gene content.

Regulation in cis

Given the prerequisite of allelic variation, there can be many reasons why genes are differentially expressed in genotypically diverse individuals of a species. Well-known phenomena are allelic variants of transcription factors and other regulators, *cis*-elemental variation in promoter sequences, differences in mRNA stability, copy number variation and genomic rearrangements such as translocations, insertions and deletions. The latter include gene loss and duplication, resulting in neo- and sub-functionalization. Most of these variations in DNA structure will result in eQTLs but depending on the position of the causal polymorphism, an important dissection is made in local and distant eQTLs (Figure 1.1) (Rockman and Kruglyak 2006). Local eQTLs can be the result of closely linked *trans*-acting factors but in the majority of cases result from *cis*-regulatory variation in the genes under study. By definition eQTLs acting *in cis* affect transcription initiation, rate and/or transcript stability in an allele-specific manner. In addition, *cis*-regulated genes might encode regulators affecting the expression of downstream target genes *in trans*. Although the exact proportion varies between studies the occurrence of *cis*-acting eQTLs is substantial, ranging from one-third to half of the total number of eQTLs (Gibson and Weir 2005).

However, because of limitations in mapping resolution, eQTL support intervals may still contain multiple genes and as a result the classification of *cis*-eQTLs should be used with care. To discriminate true *cis*-regulatory polymorphisms from local *trans*-regulation, allele specific expression (ASE) assays can be performed (Cowles *et al.* 2002).

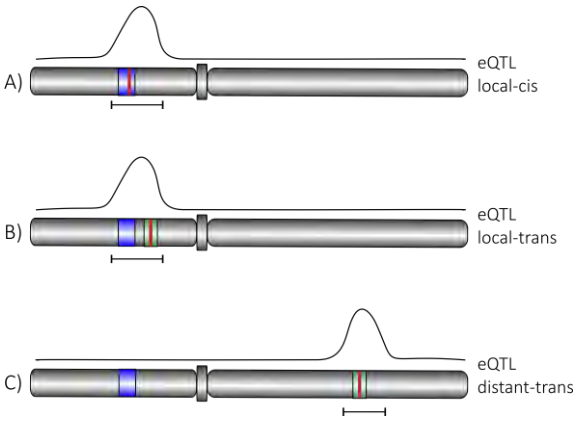


Figure 1.1: Classification of eQTLs (solid line) based on the expression of the gene under study (blue box) and location of the causal polymorphism (red bar). A) local cis-eQTL, result from allelic variation of the gene under study. B) local trans-eQTL, causal polymorphism within the eQTL confidence interval but not inside the gene under study (no allele specific expression) C) distant trans-eQTL, gene under study is located outside the confidence interval of its eQTL.

In such assays a transcribed polymorphism is used to enable discrimination between the parental transcripts and test for allele specific expression in an F_1 hybrid. Because both parental alleles share the same genetic background in F_1 hybrids, and therefore are equally exposed to *trans*-acting factors, any difference in expression can only be explained by true *cis*-acting variation. Usually, ASE-assays are performed by single gene qRT-PCR approaches but the recent development of whole genome SNP-tile microarrays (*e.g.* in Arabidopsis) enables the simultaneous testing of genome-wide ASE (Zhang *et al.* 2007).

Although expression differences are treated as quantitative traits in mapping approaches, qualitative differences, characterized by a total lack of expression for one of the allelic variants, can also be observed. The variation in a measurable detection signal can be due to differences in hybridization efficiency, which can be confirmed with genomic DNA hybridization, or genuine loss of transcription. Hybridization efficiency differences are often caused by polymorphisms in the complementary sequences of the microarray probes or mRNA splice variation and are not necessarily accompanied by transcription differences. True transcription variation however, can be caused by strong polymorphisms in promoter regions, premature stop mutations and even the complete absence of genes in one of the parental lines (Gilad and Borevitz 2006). Both hybridization and true transcription variation will lead to strong *cis*-eQTLs which can subsequently be used as molecular markers, allowing the construction of high-resolution maps (Borevitz and Chory 2004; West *et al.* 2006).

Regulation in trans

The majority of differentially expressed genes will show a quantitative expression profile with complex inheritance patterns. This is because in general genes are regulated by many independent factors which can show up as *trans*-eQTLs. Because of the multiplicity of regulators and the often-observed epistasis between them, each *trans*-eQTL can have a relatively small effect. In addition, compared to the direct regulation of *cis*-eQTLs, the accumulation of stochastic variation in the expression of *trans*-regulated genes is indirectly

also determined by the expression variation of one or more regulators. As a result the detected number of *trans*-eQTLs relative to the number of *cis*-eQTLs drops when the stringency for detection is increased (Doss *et al.* 2005).

Whereas *cis*-eQTLs are inherently associated with the gene in which they reside, a single gene can be responsible for the appearance of multiple *trans*-eQTLs throughout the genome. As a consequence the genome-wide distribution of *cis*-eQTLs is dependent on local gene density, although variation in chromatin structure can have an impact on the exposure of eQTLs. The distribution of *trans*-eQTLs however, can deviate substantially from what can be expected based on gene density. The identification of so-called hot spots, genomic regions with a high density of *trans*-eQTLs, can be explained by major regulators, *e.g.* transcription factors, which influence the expression of many downstream genes. In Arabidopsis this was illustrated by the large number of genes mapping to the *ERECTA* locus, a gene well-known for its pleiotropic effects on many morphological and developmental traits (Keurentjes *et al.* 2007). These findings suggest that the effects of key-regulators in gene expression are progressed to the phenotypic level. This was recently confirmed in a QTL study comparing transcript, protein and metabolite data with phenotypic traits (Fu *et al.* 2009). Here, only a limited number of QTL hot spots with major, system-wide effects were detected, indicating that most of the genotypic variation is phenotypically buffered. These findings support the theory of biological robustness where hotspots indicate fragilities in this genetic buffering system (Kitano 2004). Until now only a few reported hotspots have been verified and the number of detected hotspots is far from consistent between different genetical genomics studies. The latter reflects differences in the analyzed populations, species and conditions used and additionally might be the consequence of different statistical procedures used to identify eQTLs (Breitling *et al.* 2008). Because of the difficulties in cloning QTLs and the large biological relevance of hotspots, additional sources of information are often used to reduce the number of candidate genes or even predict the causal regulator. Such methods use information on gene ontology, (co-)expression, transcription factor binding sites and targets, ChIP-Seq and protein-protein interaction (Zhu *et al.* 2008). Together with computational methods such as regulatory modeling this can severely reduce the number of candidate genes and prioritize remaining candidates for further experimentation (Figure 1.2).

Genetical genomics in plants

As discussed above many principles of genetic regulation are shared among different phylogenetic taxa. Not all species however are equally suited for large-scale experimentation. Sometimes evolutionary distances withhold translation of biological relevant findings in less conserved mechanisms, *e.g.* in yeast and Drosophila, or long generation times, inbreeding depression and moral and ethical issues hinder experimentation, *e.g.* in humans and other mammals. Plants, representing one of the largest kingdoms, are therefore often used to test concepts in genetic studies.

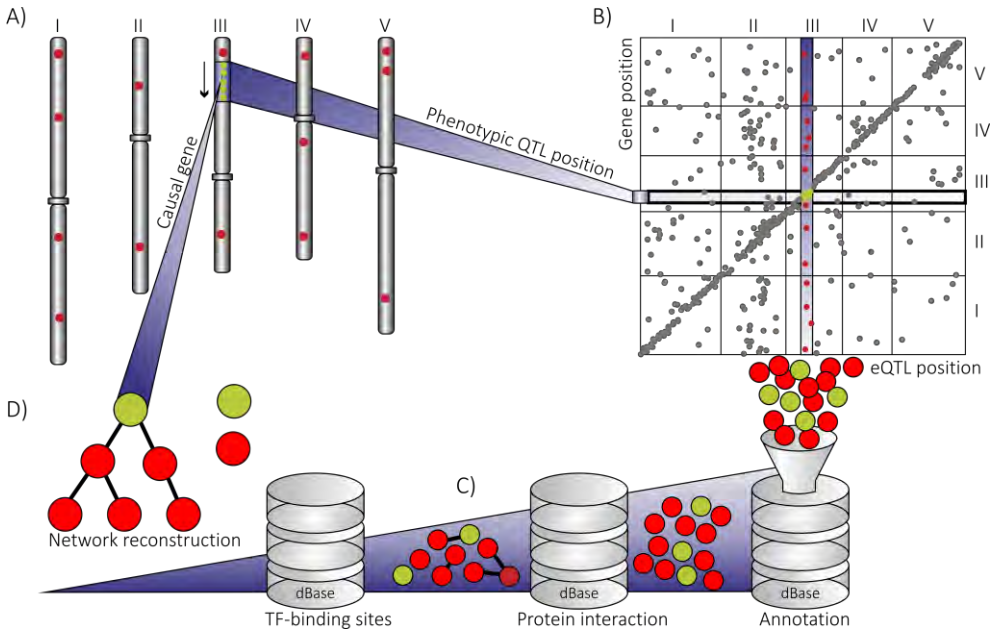


Figure 1.2: Schematic representation of candidate gene selection with a genetical genomics approach. A) Five chromosomes of Arabidopsis thaliana (I-V) with a QTL support interval for a phenotypic trait indicated on chromosome III. B) eQTL plot showing the position of genes and their corresponding eQTLs. Genes within the support interval can be causal for the observed QTL, of which cis-regulated genes, indicated in green, represent the strongest candidates. Genes outside the QTL support interval but regulated in trans by the same locus, indicated in red, might be involved in the biological process under study and represent downstream effects of the QTL. C) Available prior information of the selected genes such as gene ontology and biological interaction data can assist in limiting the number of genes to those most likely involved in the trait under study. D) Connectivity between the remaining genes is then used to construct maximum likelihood hypothetical regulatory networks which will suggest the strongest candidate regulator gene causal for the observed phenotypic QTL.

The ease to generate large families from experimental crosses and the ability to store genotypes in the form of seeds or clonal propagation make plants ideal subjects to study the mechanistic basis of genetic regulation of traits.

Arabidopsis as a reference plant

The comprehensive resources which are available for *Arabidopsis thaliana*, such as a whole genome sequence, a large collection of natural variants and an ever-increasing number of molecular tools, made it the favorable model for genetical genomics research. As a non-obligate selfing species Arabidopsis combines the ability to cross-pollinate with high tolerance to inbreeding. Together with its short generation time and high reproductive success rate this enables the fast generation of large experimental populations such as Recombinant Inbred (RI) and Introgression Line (IL) populations. The availability and immortal character of such populations enable the accurate estimation of phenotypic

values through replicated measurements and allows the testing of traits in different environments (Paran and Zamir 2003).

Traditionally QTL studies of 'classical' physiological traits in RIL populations are followed by mendelizing detected QTLs in near isogenic lines (NILs) for detailed analyses. By isolating QTLs from their genetic background it becomes much simpler to study their genetic effect and relate resulting phenotypes to other processes. Because it is expected that much of the phenotypic variation is the resultant of differences in gene expression and phenotypic perturbation in turn leads to transcriptional reprogramming, data mining for relationships between trait values and expression levels has become a common tool (Weckwerth *et al.* 2004). Very often mutants, knockouts or over-expression lines are used for these purposes, in which the effect of a single gene perturbation is tested on both the phenotypic and the expression level. For complex traits however, the causal genes leading to altered phenotypes are often not known and QTL analyses only identify genomic regions containing such genes. Nevertheless, using RIL populations to identify QTLs for a phenotypic trait and subsequently analyzing NILs for expression differences can be a powerful alternative to explore the functional relationship between genotype and phenotype (e.g. Juenger *et al.* 2005; Juenger *et al.* 2006). Although the regions spanned by NILs can still contain hundreds of genes, of which many may display allelic variation between accessions, the *cis*-regulated genes are strong candidates explaining phenotypic diversity. High detection stringency can limit the number of differentially expressed genes to a reasonable number of candidate genes with strong local eQTLs (Juenger *et al.* 2006).

The availability of a whole genome sequence in Arabidopsis provides unique opportunities, especially when multiple (epistatic) phenotypic QTLs are detected. Knowing the position of genes allows the identification of strong *cis*-regulated genes co-locating with phenotypic QTLs. An early eQTL study in Arabidopsis analyzed genome-wide gene expression in a limited population of only 30 individuals, mimicking shoot regeneration conditions (DeCook *et al.* 2006). Two of the eQTL hotspots found coincided with shoot regeneration QTLs. The most significant eQTLs within these hotspot regions showed local chromosal linkage with their corresponding genes but the majority acted distantly. These results suggest that heritable *cis*-regulated expression changes of key-regulators determine *in trans* the expression of many genes related to differences in shoot regeneration efficiency between accessions. It also indicates that a long signaling cascade may exist between the causal genotypic polymorphism and the eventual phenotype.

In contrast to the former study it is not always necessary to combine phenotypic measurements with expression analysis. Often, many genes are known to play a role in the exposure of certain traits without knowledge about the genetic regulation of these genes. Specific analysis of such genes can help to identify common regulators. In the first genome-wide eQTL study in Arabidopsis, using a complete RIL population (162 lines), this concept was used to predict possible key-regulators of flowering time and circadian rhythms (Keurentjes *et al.* 2007). The benefits of using large populations for eQTL studies became also apparent in another study where expression analyses were performed in a RIL

population of 211 individuals (West *et al.* 2007). Whereas in the majority of cases only a single QTL could be detected per differentially expressed gene in the aforementioned studies, here the expression of many genes was controlled by multiple eQTLs. Moreover, a much larger fraction of genetically regulated genes was identified with a higher proportion of *trans*-regulated genes of which the vast majority exhibited small effects.

The studies performed in Arabidopsis show that the statistical power to detect eQTLs depends largely on population size. Nonetheless, it cannot be excluded that differences in the analyzed tissues, developmental stages and populations used, such as parental variation, linkage distortion and recombination frequency, are responsible for part of the observed differences. All studies however, clearly demonstrated that variation in gene expression is for a large part genetically controlled, with much stronger effects of *cis*-eQTLs compared to *trans*-eQTLs. In general, *cis*-eQTLs also exhibit much higher heritability values and are obvious candidates to act as causal regulators of genes showing *trans*-eQTLs in the hotspots that could be detected in each of the discussed studies. The detection of regulatory loci for gene expression and the elucidation of their interaction networks might therefore provide the research community with a powerful tool to unravel the complex nature of natural variation in quantitative traits.

A good example of the power of genetical genomics is described by Jimenez-Gomez *et al.* (2010) who identified *EARLY FLOWERING 3* (*ELF3*) as the most likely candidate gene affecting the shade avoidance response of Arabidopsis in a Bayreuth-0 x Shahdara population. For narrowing down to *ELF3* as the only candidate causal gene for a shade avoidance QTL identified in this population, they combined publicly available datasets to perform network analysis with eQTL data (West *et al.* 2007) co-expression analysis (Winter *et al.* 2007) and functional classification (Ashburner *et al.* 2000). Drastically narrowing down on the number of candidate genes with this kind of approach is feasible for all QTLs where the causal alleles result in differential gene expression of the causal gene. However, this approach will not be applicable to the cases where the alleles causal for a QTL do not have an effect on gene-expression, but on activity or stability of the encoded protein. In these cases, other levels like pQTL or mQTL and other data types, including protein-protein interactions and metabolic pathways can help to narrow down on the causal genes.

Applications in crop species

Genetical genomics studies in Arabidopsis and other model species have shown the enormous benefits of the availability of an annotated genome sequence. However, until now full annotated genome sequence information for crop species is only available for *Zea Mays* (Schnable *et al.* 2009), *Solanum lycopersicum* (Tomato Genome Consortium 2012), *Sorghum bicolor* (Paterson *et al.* 2009), *Oryza sativa* (Goff *et al.* 2002; Yu *et al.* 2002), *Populus trichocarpa* (Tuskan *et al.* 2006),*Vitis vinifera* (Jaillon *et al.* 2007) and *Carica papaya* (Ming *et al.* 2008). This relatively low number of sequenced crop species can be explained by their often immense (polyploid) genome sizes and the highly repetitive nature

of many crop genomes (Burke et al. 2007). Nevertheless, sequence efforts for many more species, are ongoing and the increasing power of next-generation sequencing will soon lead to an almost unrestricted availability of genomic sequence information. Although an annotated genome is a valuable resource for the comparison of the genomic position of genes and their respective eQTLs, for most crop species this is not feasible yet. Nonetheless, several studies in crops for which genetic maps are available have shown that comprehensive genetical genomics approaches are possible without the need for annotated genome sequences (Kirst *et al.* 2004; Street *et al.* 2006; An *et al.* 2007; Poormohammad Kiani *et al.* 2007; Shi *et al.* 2007; Venu *et al.* 2007).

Illustratively, one of the first large genetical genomics experiments was performed in an economically important species, *viz. Eucalyptus* (Kirst *et al.* 2004). QTL analysis of transcript levels of lignin-related genes showed that their mRNA abundance is regulated by two genetic loci coinciding with QTLs for stem diameter growth. Genetic mapping of some of the candidate genes showed that most of the lignin genes are under control of a *trans* eQTL hotspot which suggests that transcription of many of the genes in this pathway are under a higher level of coordinated control. A strong *cis*-regulated gene encoding S-adenosylmethionine synthase, co-locating with the growth and transcription QTLs, was presented as the possible rate limiting step in lignin biosynthesis and as such a strong candidate for the observed QTLs (Kirst *et al.* 2004).

In some crops the required availability of genomic sequence data for large-scale classification of *cis/trans* eQTLs can be circumvented by making use of synteny with other species. In wheat, synteny with rice was used to assist the physical mapping of wheat genes (Jordan *et al.* 2007). A genetical genomics approach was conducted in a segregating population of 41 doubled haploid (DH) lines to study agronomic important seed quality parameters. Assuming that the most significantly different expressed genes were *cis*-regulated, a selection of genes was subjected to synteny analyses. This enabled the positioning of genes with biologically relevant linkage to phenotypic traits in a species for which full genome sequence is not available yet.

In the absence of genome-wide micro-arrays, expressed sequence tag (EST) libraries allow the construction of species specific sub genome-scale microarrays. In maize, cell-wall digestibility, which is the major target for improving the feeding value of forage maize, was analyzed in a RIL population (Shi *et al.* 2007). In addition forty extreme RIL lines were hybridized on a small microarray with 439 preselected candidate ESTs for cell-wall digestibility genes for which 89 eQTLs could be mapped. One eQTL hotspot co-located with a cell-wall digestibility related QTL (Shi *et al.* 2007). The application of genetical genomics approaches can be of special interest here when the detection of eQTLs is combined with ASE assays. The thus identified *cis*-regulated genes can then be positioned on the genetic map where they may serve as candidate genes underpinning phenotypic QTLs.

An interesting alternative for species for which no (EST) sequence information is available at all, and hence no microarrays can be produced, is a gel-based cDNA-AFLP approach (Vuylsteke *et al.* 2006). Here AFLP band intensities, reflecting expression

differences, are profiled for a large proportion of the transcribed gene pool enabling standard eQTL analyses procedures. AFLP bands showing significant eQTLs can subsequently be sequenced to obtain the identity of the gene from which the fragment was derived. Additionally, the cDNA-AFLPs can be used to construct a genetic map.

The examples given above show that genetical genomics is not necessarily restricted to model species but can be applied to any species in which experimental crosses are possible even in the absence of genomic sequence or genetic map information. The potential of combining phenotypic QTL analysis with gene expression traits is shown in a number of economically important species, e.g. *Populus* (Street *et al.* 2006), cotton (An *et al.* 2007), rice (Venu *et al.* 2007) and sunflower (Poormohammad Kiani *et al.* 2007). The application of genetical genomics is particulary promising in breeding programs of crops that take advantage of hybrid vigor. The eQTLs involved in heterosis will segregate consistently in a F_1 backcross population thereby identifying valuable targets for marker assisted breeding for the best combination of alleles in the parents of the hybrid (Kirst *et al.* 2005).

Network reconstruction

Genetical genomics harbors the potential to dissect the genetic regulation of a specific biological process. Therefore, methods to reconstruct regulatory networks from eQTL data have obtained much attention. Prioritizing on *cis*-eQTLs that co-locate with a phenotypic QTL is a valuable approach for causal gene discovery, but in many cases little is known about the global regulation, interaction and function of genes that control a biological process. Identification of a set of genes with a *trans*-eQTL at an identical position can help to dissect genetic variation that is influencing an entire pathway and can lead to the identification of initiating polymorphisms upstream in a network (Hansen *et al.* 2008). Questions about the regulatory level at which *trans* polymorphisms act in the global gene expression network and what their effect is on phenotypic variation and heritability can only be addressed when eQTLs are further dissected.

With a genetical genomics approach one can use the natural genetic variation as a source of perturbations to elucidate the structure of networks. In a summation approach eQTLs for all genes in the analysis are simply superimposed to identify common regions which control many genes (Schadt *et al.* 2003). Such an approach does not require any *a priori* network information but applies subsequent Gene Set Enrichment Analysis (GSEA) using gene ontology (GO) annotation or other descriptors to test whether selected genes share a common biological function (Subramanian *et al.* 2005). If the network under study is largely known or at least predicted, an *a priori* analysis can be performed. Here, the expression levels of individual genes in the network are converted into a common measure for the expression level of the entire network which is then used as the trait for QTL analysis. This strategy was tested in an Arabidopsis RIL population for 20 gene expression networks and resulted in statistically significant network variation for eighteen of the 20

predefined networks (Kliebenstein *et al.* 2006). Combining summation, GSEA and *a priori* network analyses allows the generation of a more specific hypothesis about phenotypic effects of network eQTLs. In a study using 175 genes, selected to be involved in regulation of flowering and circadian rhythms, 83 genes showed an eQTL (Keurentjes *et al.* 2007). By combining co-expression analysis, which becomes feasible for microarray compendia of large populations, and positional information of genes and their eQTLs, it was possible to construct regulatory networks of key-regulators and their target genes, predicting unknown relationships and confirming common knowledge.

Pre-selection of known pathways can obviously hinder the elucidation of novel networks in a species, for which much effort is made to develop methods to translate eQTL data into network information using an *a posteriori* approach. As the precise balance of active components within a tightly controlled biological pathway is in part maintained by coordinately regulated gene expression, this creates possibilities to model networks by exploring co-expression of untargeted genes. To validate this hypothesis, gene expression in liver from a population of 60 mice with variation in diabetes susceptibility was analyzed (Lan *et al.* 2006). The combination of correlation analysis across a genetic dimension and linkage mapping enabled the identification of regulatory networks, functional predictions for uncharacterized genes and characterization of novel members of known pathways. A similar approach in *Drosophila*, complemented with information about gene ontology, tissue specific expression and transcription factor binding sites, led to the construction of multiple interconnected networks with biological relevance for phenotypic traits (Ayroles *et al.* 2009).

Understanding the mechanisms underlying trait regulation requires the identification of specific causal polymorphisms. For this purpose sophisticated self-learning algorithms have been developed which make use of conservation, type and position of a particular SNP to prioritize causal regulators by estimating the likeliness that it plays a causal role in gene expression variation (Lee *et al.* 2009). Extending such approaches might also provide the means to distinguish whether variation in gene expression or a regulatory network is the cause or a consequence of an altered phenotype, resulting in the construction of probabilistic directional networks (Rockman 2008). Defining such causal networks is also known as reverse engineering, because it aims at understanding how the system works as an integrated whole instead of only defining the functionally related components.

Next level networks: integration of other 'omics' data

Although phenotypic variation can be partly explained by genetic variation in gene expression, this alone does not fully cover the possible differences in the regulatory mechanisms of an organism. Similar transcript levels of allelic gene variants can still result in varying protein levels because of variance in translational activity, protein degradation and post-translational modifications (Stylianou *et al.* 2008). Furthermore, variation in

coding sequences can alter protein function resulting in a flexible metabolome in terms of chemical structure and function (Keurentjes and Sulpice 2009). Integrating 'omics' data such as gene expression, SNPs, metabolomics and proteomics in genetic studies can therefore reduce the number of candidate genes for a given QTL from hundreds to a manageable list without excluding regulatory mechanisms *a priori*. Because of the analytical complexity in analyzing large numbers of protein samples, genetical proteomics studies are limited (e.g. Chevalier *et al.* 2004) but advances made in biochemical detection have already enabled the large-scale untargeted genetic analysis of metabolic content (Keurentjes *et al.* 2006; Schauer *et al.* 2006; Lisec *et al.* 2008).

The complex relationship between different levels of regulation was illustrated in a study integrating parallel QTL analyses of the expression of genes, activity of encoded enzymes and metabolites involved in primary carbohydrate metabolism (Keurentjes *et al.* 2008). It could be shown that regulation acted on each of the intermediary levels of the path from genotype to phenotype. Although seemingly specific independent regulation could be observed for each analyzed trait, a strong interconnectivity existed between them resulting in coherent systematic differences between populations of individuals.

The importance of the tight regulation of such an essential component in plant development as primary metabolism was also demonstrated in an Arabidopsis RIL population where plant biomass was related to the metabolic profile (Meyer *et al.* 2007). Again, no relationship could be observed between individual metabolites and plant growth but a strong canonical correlation was observed between biomass and a specific combination of metabolites in central metabolism. The power of large-scale metabolomic profiling combined with detailed morphological analysis was also shown in tomato (Schauer *et al.* 2006). Significant QTLs could be detected for the accumulation of a large number of primary metabolites together with loci that modify yield-associated traits. With this information a correlation network revealing associations between phenotype, metabolic content and nutritional value could be generated. These studies show that analyzing phenotypic traits and metabolic profiles in a genetic mapping population has great potential for the generation of biomarkers in breeding programs.

Whereas primary metabolites are essential in central metabolism governing growth and development, plants also accumulate large amounts of secondary metabolites. These are believed to be less essential but may play an important role in the adaptation of plants to local environments. Since Arabidopsis can be found in a wide variety of habitats, variation in secondary metabolism might explain much of the evolutionary success of the species. A large untargeted screen of variation in secondary metabolic composition indeed revealed a high proportion of genetically controlled compounds (Keurentjes *et al.* 2006). The highly flexible nature of the metabolome was clearly shown by the fact that more than one-third of the compounds present in the RILs were not detected in either parent but were the result of recombination in biosynthesis pathways. The genetic information obtained from such studies is of great value for the construction of molecular biosynthesis networks, especially if they can be combined with expression data.

This strategy was applied in the genetic analysis of glucosinolate biosynthetic networks which were studied at both the transcriptional and metabolic level (Wentzell *et al.* 2007). In all cases, variation in gene expression also affected the accumulation of metabolites but epistasis was detected more frequently for metabolic traits as compared to transcript traits. Within such an *a priori* defined framework it was possible to identify and unravel complex regulatory mechanisms like metabolic feed-back loops in which metabolic content regulated gene expression and vice versa.

The examples discussed here highlight the technological advances made in high-throughput characterization of the transcriptome, the proteome and metabolome which enables an integrated multidisciplinary approach to unravel the regulatory mechanisms involved in natural variation of complex traits.

Future challenges

Although much progress is being made in understanding the influence of genetic factors on a biological system we still have limited understanding of the interplay between environment and genetic factors. The discovery of molecular networks with genetical genomics approaches is often limited to a single experimental condition. An interesting concept, called generalized genetical genomics, uses controlled environmental perturbations combined with genetical genomics (Li *et al.* 2008). This generalization of genetical genomics will detect how the response to environmental changes is influenced by the genotype (*i.e.* genotype x environment interactions). Here, spatial and temporal variation can also be regarded as different environments since specific tissues and developmental stages often determine the biological context in which regulatory networks function.

The advances in next generation sequence technology will continue to produce huge amounts of sequence data. Good examples are the human 1000 (Siva 2008) and the 1001 Arabidopsis (Ossowski *et al.* 2008) genome projects which aim at resequencing over 1000 different humans and accessions respectively. However, *de novo* sequencing of economically or phylogenetically chosen species is of equal importance. The accumulation of genomic information, in combination with genetical genomics approaches, will enable the precise definition of functional important polymorphisms and their role in adaptation to changing environments and species formation. Having access to complete genome sequences also enables the generation of full genome tiling arrays for different (crop) species, which have been proven to be very useful for expression profiling (Laubinger *et al.* 2008; Matsui *et al.* 2008). When used within a genetical genomics approach this offers unique features to elucidate the genetics behind the mechanistic basis of transcriptional differences. For Arabidopsis for instance, a SNPtile microarray was developed harboring tiling probes covering both strands of the genome and in addition probes for genome-wide detection of SNPs and CpG methylation (Zhang *et al.* 2007). A properly designed genetical genomics study using such arrays might reveal genetic variation for gene expression,

alternative splicing, regulation of *cis*-natural antisense transcripts, allele specific expression and epigenetic regulation.

As a result of developments in SNP-discovery and platforms for genotyping large collections of individuals, the application of Linkage Disequilibrium (LD) mapping for complex traits has become within reach. LD or association mapping detects the non-random inheritance of alleles at separate loci located on the same chromosome. In an experimental F_2 or RIL population the genetic variation is limited to the extent of natural variation present in the parental lines and resolution depends on the recombination frequency within and size of the population. In contrast LD mapping makes use of large collections of natural (wild) accessions or elite breeding lines, sampling a much larger fraction of the natural variation present within a species. Moreover, it benefits from the much higher frequency of recombination events accumulated during the evolutionary history of a species allowing higher resolution mapping (Buckler and Gore 2007). The extent of LD varies between species and traits analyzed but the gain in resolution relative to experimental populations lies in the order of magnitudes, equally increasing the need for dense marker spacing to enable genome wide scans (Sorkheh *et al.* 2008). This high number of necessary markers has always been a big limitation for LD mapping but next generation sequencing will tremendously increase the available number of markers. Therefore, we see great potential for phenotyping and expression profiling of LD populations to detect causal genes for natural variation and enable marker-assisted selection in breeding programs.

Concluding remarks

Since its introduction the concept of genetical genomics has proven to be a powerful approach to dissect genetic variation. Studies in crop species revealed major *cis*-eQTLs which co-located with important phenotypic traits and therefore will facilitate faster crop improvement. The genetical genomics studies in model species help to understand the extent of genetic variation and much effort is spent to develop statistical tools for building and elucidating causal networks. Recent developments of inexpensive high-throughput sequencing techniques and next generation tiling microarrays will soon create opportunities to extend genetical genomics to unravel the genetic variation of gene expression, alternative splicing, allele specific expression and epigenetic polymorphisms. Similarly, continuing technological developments have increased the power of both proteomic and metabolomic approaches. Integration of phenotypic, genetic, transcriptomic, proteomic and metabolomic data will enable accurate and detailed network reconstruction for traits such as seed quality. Ultimately, this increased knowledge about the factors influencing seed quality will open new possibilities for the breeding industry to understand and control the effects of the maternal environment on seed quality and above all allow breeding for high quality seeds.

Outline of this thesis

The objective of the research presented in this thesis was to explore the possibilities of using genetic mapping populations to study the molecular mechanisms which are important for seed performance related traits.

Chapter 1 introduces the definition of seed quality and how natural variation has been explored to detect QTLs for seed related traits. It emphasizes on the importance to invest on high throughput phenotyping methods. The concept of genetical genomics is introduced and examples of studies are provided both for the model plant *Arabidopsis thaliana* and for several crop species. The potential to use genetical genomics experiments to deduce regulatory networks is discussed including the possibilities to integrate several types of 'omics' data.
Chapter 2 describes the development of the Germinator package. To allow genetic mapping of seed related traits it is important to be able to perform high-throughput detailed phenotyping of the germination process. We show that automatic scoring of seed germination is possible with the use of image analysis. This allows, in combination with a curve fitting procedure, a detailed analysis of the whole germination process. An example is provided showing the possibilities and advantages to use such an automatic scoring system to study natural variation for salt tolerance.
Chapter 3 presents the results of using the germinator package to perform a detailed analysis of natural variation for a large range of seed performance traits. In total this analysis resulted in 327 trait scores over different harvests of the Bay-0 x Sha RIL population. This demanded new methods to allow high-throughput QTL analysis. Therefore a user friendly script was developed in the statistical programming language R which performs automatic multiple QTL mapping (MQM), reporting and visualization. Multitrait visualizations are used to detect co-locating QTLs and QTLxQTL interactions are described. An alternative approach was used to detect QTL x Environment interactions. A range of QTLs are confirmed using the heterogeneous inbred family strategy. Together, this resulted in a large dataset describing natural variation for seed germination in the Bay-0 x Sha RIL population which provides a solid resource for further dissection of the detected QTLs.
Chapter 4 is focused on efficient data visualization which is a prerequisite for large scale 'omics' data analysis. A selection of seed specific transcriptome studies from publicly available microarray resources for *Arabidopsis thaliana* was used to identify functional categories which are influenced during seed dormancy and germination. The MapMan tool allows visualizing transcript, metabolite and protein levels on custom diagrams. We created two of such diagrams tailored to use for seed related research. The first diagram provides an overview of all enriched functional categories during seed dormancy and germination while the other diagram allows a focused view of cell wall changes. Four examples are provided to show the power of using the new diagrams to study molecular processes related to dormancy and seed germination.

Chapter 5 describes a metabolomics study of germinating seeds using a generalized genetical genomics approach. Genotype x environment interaction (GEI) requires experimentation in multiple conditions and is often ignored due to the expensive nature of genetical genomics studies. An alternative concept to reduce the experimental load while allowing GEI is coined generalized genetical genomics. This concept is used to study natural variation for metabolic changes during four developmental stages of seed germination. New statistical procedures are developed to allow analysis of this type of data. Two identified metabolite QTL hotspots are confirmed by using the heterogeneous inbred family approach.

Chapter 6 is continuing on the generalized genetical genomics approach and is focused on gene expression differences. Gene expression was profiled in four developmental stages of seed germination using a whole genome tiling microarray. An overview of the influenced molecular processes is provided using the Mapman diagrams that were developed in Chapter 4. Expression QTLs could be mapped using the statistical procedures that were developed in chapter 5 and are shown using a cytoscape marker-trait network which was introduced in chapter 3. Several examples of co-locating eQTLs are shown which allow building hypotheses on molecular regulation. The presented results are regarded as a 'sneak preview' because the used microarrays also enable identification of genetic variation for alternatively spliced exons and anti-sense transcripts when analyzed to its full potential.

Chapter 7 is exploring the possibilities of genome wide association (GWA) to detect natural variation for seed germination in *Arabidopsis thaliana.* We used a well-defined selection of 360 worldwide collected natural accessions (HapMap population) which are genotyped with 214051 single nucleotide polymorphism markers. Genome wide association has become a promising tool to dissect natural variation with much higher resolution compared to traditional linkage mapping. GWA results are compared to mapping results obtained in the Bay-0 x Sha RIL population (Chapter 3).

Chapter 8 summarizes and discusses the most important results from this thesis. Two examples are provided to show the possibilities and the complexity of data integration for molecular network reconstruction. The chapter is finalized with future considerations to study genetics and molecular mechanisms of seed performance.

Acknowledgements

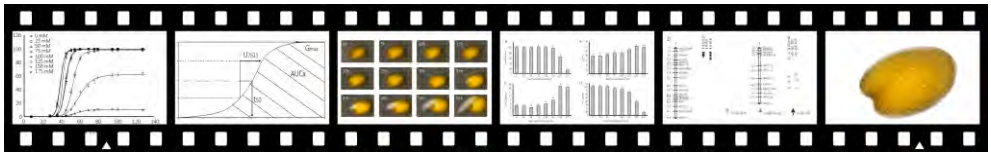
This work was supported by grants from the Netherlands Organization for Scientific Research (STW 10027), VENI scheme (863.08.019) and the Centre for Biosystems Genomics (CBSG, Netherlands Genomics Initiative).

"Whoever sows alongside the road, tires the oxen and wastes seed."

2 GERMINATOR
A software package for high throughput scoring and curve fitting of *Arabidopsis* seed germination

Joosen RVL, Kodde J, Willems LAJ, Ligterink W, van der Plas LHW, Hilhorst HWM (2010).

Plant Journal. Vol. 62: 148-159



Abstract

Over the past few decades seed physiology research has contributed to many important scientific discoveries and has provided valuable tools for the production of high quality seeds. An important instrument for this type of research is the accurate quantification of germination; however gathering cumulative germination data is a very laborious task which is often prohibitive to the execution of large experiments. In this paper we present the Germinator package: a simple, highly cost efficient and flexible procedure for high-throughput automatic scoring and evaluation of germination that can be implemented without the use of complex robotics. The Germinator package contains three modules; 1) design of experimental setup with various options to replicate and randomize samples; 2) automatic scoring of germination based on the color contrast between the protruding radicle and seed coat on a single image; 3) curve fitting of cumulative germination data and the extraction, recap and visualization of the various germination parameters. The curve fitting module enables analysis of general cumulative germination data and can be used for all plant species. We show that the automatic scoring system works for *Arabidopsis thaliana* and *Brassica spp.* seeds, but is likely to be applicable to other species, as well. In this paper we show the accuracy, reproducibility and flexibility of the Germinator package. We have successfully applied it to evaluate natural variation for salt tolerance in a large population of Recombinant Inbred Lines (RIL) and were able to identify several QTL for salt tolerance. Germinator is a low-cost package that allows the monitoring of several thousands of germination tests, several times a day by a single person.

Introduction

Seeds are the most sophisticated means of propagation created by plant evolution. They are indispensible for human society as food source and as starting material for new crops. Seed physiology and technology have provided valuable tools for the production of high quality seeds, various seed treatments and optimal storage conditions. In fundamental research seeds are studied exhaustively and systems biology approaches are undertaken to fully explore dormancy and germination (Penfield and King 2009). An important instrument to indicate the performance of a seed lot is the accurate quantification of germination by gathering cumulative germination data.

Completion of germination is defined as the protrusion of the radicle through the endosperm and seed coat (Bewley 1997). The uptake of water of the dry seed during imbibition is triphasic and consists of a rapid initial uptake (phase I) followed by a plateau phase (phase II) and a further increase in water uptake (phase III). During phase III the embryo axis elongates and breaks through the testa. In Arabidopsis the testa is dead tissue whereas the endosperm layer is living tissue. The action of several cell-wall-modifying proteins is required to enable a break through the endosperm. For accurate scoring of seed germination a careful discrimination should be made between the testa and endosperm rupture because this lag phase may vary among germination conditions and treatments (Liu *et al.* 2005; Finch-Savage and Leubner-Metzger 2006; Müller *et al.* 2006).

Although very often used, the total percent germination after a nominated period of time is not very explanatory. It lacks information about start, rate and uniformity of germination, which are essential parameters of a normally distributed seed population, for many traits such as dormancy, stress tolerance and seed aging. Information about germination at various time intervals is required to calculate a cumulative germination curve, but the number of samples that can be handled with manual counting is usually the limiting factor. Moreover, *Arabidopsis* seeds are small, requiring the use of a binocular or magnifying glass. Therefore, a fast and reliable automated procedure would enable high-throughput screens and unlock the full potential of seed science research.

Arabidopsis thaliana is a popular model plant for seed science and provides insight in common physiological processes which can be translated to economically important crops (Lin *et al.* 1999). The availability of mutants, ecotypes, inbred populations and sequence information enables the molecular-genetic analysis of many seed germination related traits. For example, mutant analysis has identified seeds with reduced dormancy and altered flavonoid biosynthesis, as well as altered germination tolerances to stresses like salt and osmotic potential, desiccation, heat and cold (Shirley *et al.* 1995; Leon-Kloosterziel *et al.* 1996; Espinosa-Ruiz *et al.* 1999; Hong and Vierling 2000; Wehmeyer and Vierling 2000; Kim *et al.* 2005). Screens for natural variation in the various available inbred populations revealed, among other traits, loci involved in dormancy, storability, glucosinolate production, salt tolerance, storage oil production and mineral content (van Der Schaar *et al.* 1997; Bentsink *et al.* 2000; Kliebenstein *et al.* 2001; Quesada *et al.* 2002;

Hobbs *et al.* 2004; Vreugdenhil *et al.* 2004). Mutant analysis and QTL localization is complemented with the exhausting inventory provided by transcriptomics, proteomics and metabolomics approaches (Gallardo *et al.* 2001; Cadman *et al.* 2006; Routaboul *et al.* 2006; Goda *et al.* 2008). Taken together it is evident that *Arabidopsis* has become an important and valuable tool for seed scientists but that high-throughput detailed phenotyping for effects on seed germination could extend its prospects.

When studying natural variation for germination performance in an inbred population or performing mutant screens, the number of germination assays required is tremendous. In this type of large experiments it is difficult to manually score germination at multiple time points per day for a number of days or weeks. Although major progress for semi-automated scoring of seed germination of *Lactuca sativa,* using a flat-bed scanner, was made by Teixeira *et al.* (2007), the setup suggested by these authors only allows a limited number of samples. Also the system with a camera above a Jacobsen table for *Helianthus annuus* seeds as described by Ducournau *et al.* (2005) does not accomodate high-throughput screens without expensive robotics, as it requires proper alignment between two consecutive images. The setup that we developed enables large screens without the need for expensive robotics. We are making use of germination trays which are kept in climatized cabinets. Digital photographs are made from these trays at flexible time intervals and automatically analyzed by our germinator scripts. The power of this procedure is that it does not score germination based on the difference between two consecutive pictures but instead uses the information from two different color threshold analyses on a single picture, which circumvents alignment problems.

Interpretation of germination performance can be accomplished by extracting the relevant parameters from the germination-time curve. We have used a method described by El-Kassaby *et al.* (2008) to mathematically fit the germination curve using the four-parameter Hill function (4PHF). This function allows extraction of biologically relevant parameters such as maximum percentage of germination (G_{max}), time to reach 50% germination (t_{50}), $t_{(x)}$=time to reach a user defined percentage of germination and uniformity of germination (like U_{7525}: time interval between 25% and 75% of viable seeds to germinate). Integration of the area under the curve (AUC) provides a value that enumerates these parameters and often shows a high discriminative power between samples. To enable the quick analysis of many cumulative germination curves in large experiments we developed a curve fitting module which results in a clearly formatted output that summarizes the biological relevant parameters, describing germination behavior. The curve fitting module enables analysis of any type of cumulative germination data and is not restricted to any plant species.

Various experiments were performed to test and validate the procedure. We used a one-hour interval measurement to quantify the germination of *Arabidopsis* accessions Landsberg erecta (Ler) and Columbia (Col). We compared manual with automatic counting and assessed the accuracy of the curve fitting at different time intervals. Furthermore, we tested the procedure for germination of *Arabidopsis* Col. at different concentrations of

NaCl. The application of salt stress results in different levels of maximal germination percentage, germination rate and uniformity of germination which provides an ideal test for the flexibility and accuracy of our automatic germination scoring procedure. Next, an *Arabidopsis* recombinant inbred population consisting of 165 lines was used to show the power of high-throughput germination phenotyping. Plant salt tolerance is a complex trait, which is polygenic and hence difficult to dissect and manipulate. We used the Germinator package to score and analyze germination in control versus salt conditions (which equals ~2000 germination assays), and tested the genetic variation for salt tolerance. Finally, we analyzed germination of seeds from a *Brassica spp.* recombinant inbred line with a huge variation in seed color to show that the germinator package might be applicable to many more species. Recently it was shown in maize that measuring the rate of germination time is a good indicator for relative vigor and field performance, which underlines the importance of high-throughput methods for scoring germination in commercial crop testing as well (Khajeh-Hosseini *et al.* 2009).

Results

We have divided the process of analyzing germination in three basic steps: experimental setup, image analysis and data analysis. These 3 steps are represented by different modules of the Germinator package and can be carried out independently. However, especially in large scale experiments a solid administration is crucial. Therefore we complemented the package with a Microsoft Office Excel visual basic script that creates an overview of the experiments performed with active links to the generated output (Germinator_menu1.0.xls).

Experimental setup

To enable as much automation as possible we have standardized the whole experimental setup. In the first module of the Germinator package (Germinator_table1.0.xls) the user can define the number of samples, treatments, and repetitions and whether a randomized setup is desired. These choices result in an 'Experiment Setup' (ES) table which can be used to set up the experiment. The exact starting times of the individual tests can be added to the ES tables. Multiple time ranges within one experiment are allowed and can be handled by both the automatic scoring and curve fitting scripts. We use transparent germination trays which can be stacked in an incubator with light from the sides (Figure 2.1A, materials and methods for details). The content of these trays, consisting of a blue filter paper with six samples of seeds, are manually photographed at different time intervals. The blue filter paper is used to obtain optimal contrast between seed, radicle and filter paper. All images are automatically named with tray number, date and time. This data is used to automatically match the pictures to the correct tray, different treatments and samples and extract information about the time intervals as mentioned in the ES tables.

Image analysis

In large scale experiments the number of images can become prohibitive; therefore an automated procedure for image analysis is required. First, the images are batch preprocessed in Adobe Photoshop CS3 with the action 'crop.atn' which divides each image into 6 individual pictures and saves them under a unique name (Figure 2.1B). Subsequent image analysis is performed with ImageJ and is based on segmentation by color-thresholding (Figure 2.1C).

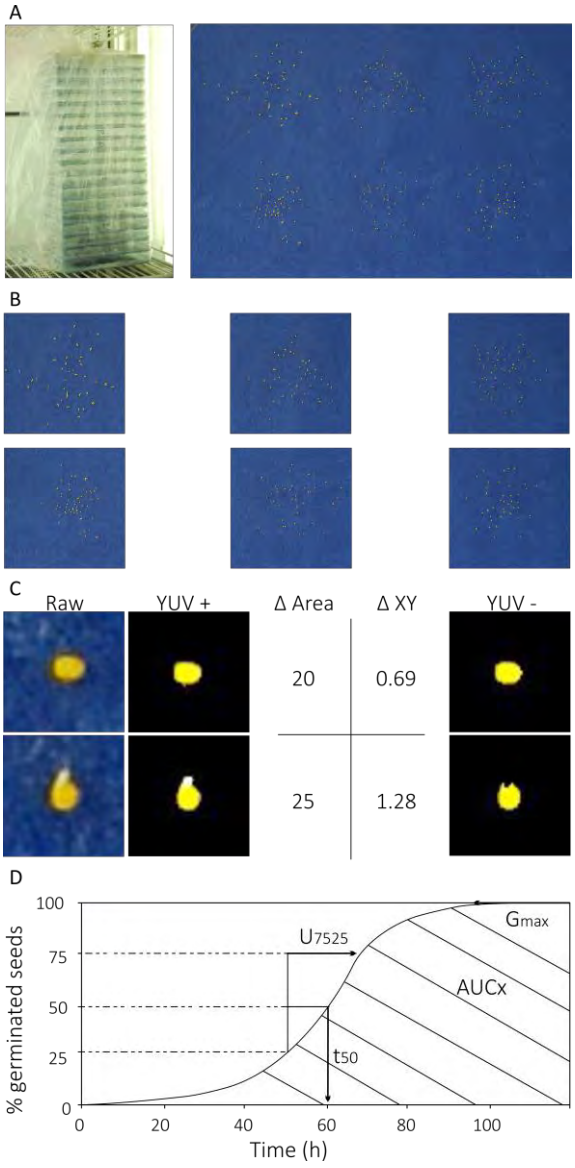


Figure 2.1: The workflow of the automated scoring of Arabidopsis germination. A) a pile of plastic germination trays in a climatized cabinet and the raw image from each tray; B) an Adobe Photoshop action crops the picture in 6 individual pictures; C) Scoring of germination is based on the double color thresholding of a single image; indicated are a magnification of the image, the Δarea and the ΔXY (both in pixels) between the color threshold that selects seedcoat only (YUV-) and the color threshold that selects both seedcoat and radicle (YUV+); D) Cumulative germination data is used as input for the curve fitting module. Multiple germination parameters are automatically extracted. G_{max} indicates the maximum germination capacity of a seedlot. The t_{50} is the time required for 50% of viable seeds to germinate (t_{50}). Uniformity (U_{7525}) of germination is the time interval between 75% and 25% of viable seeds to germinate. The Area Under the Curve (AUC_x) is the integration of the fitted curve between t = 0 and a user-defined endpoint (x=120h).

Using visual scripting for ImageJ (Baecker and Travo 2006) two batch scripts were developed which perform contrast enhancement, color-threshold, invert image, particle analysis and reporting of the results. Every image is analyzed twice; first with a color threshold that only selects the yellow/brown seed ($Y_{100\text{-}255}$;$U_{0\text{-}80}V_{130\text{-}255}$) and second with a color-threshold that selects everything but the background ($Y_{100\text{-}255}U_{0\text{-}130}V_{80\text{-}255}$). For this reason, fungal contamination during the experiment should be prevented as much as possible because this may cause false positive scoring of germination. The YUV model defines a color space in terms of one luma (Y) and two chrominance (UV) components. By using this color space we obtained the best separation between seed coat and protruding radicle. In both analyses the XY position (average of X and Y coordinates of all the pixels in the selection) and size (area + perimeter in pixels) for each individual seed are extracted to output tables which will be saved in tab delimited format. The output tables are analyzed with the help of a Microsoft Excel visual basic script (Germinator_table1.0.xls) that compares the XY position and the size of each individual seed. Seeds are scored as not germinated when both the difference between XY position and size of the two color-thresholds are within a user defined limit. To prevent artifacts caused by clustered seeds, a size restriction is added. The total number of seeds is extracted from the first image; this number is used in the later time points to calculate the number of germinated seeds based on the detection of non-germinated seeds. To set accurate thresholds for both XY-position and size differences we developed a 'parameter screen' function as part of the Germinator table script that empirically compares manual versus automatic counts and determines the most optimal settings. The germination data and time intervals are transported to the initial ES tables. These final cumulative germination tables can automatically be loaded into the third Germinator module (Germinator_curve-fitting1.0.xls), which performs curve fitting and parameter extraction (Figure 2.1D).

Data analysis

Using the visual basic module from the Microsoft Excel package we developed a script which performs automated curve fitting on cumulative germination data using the Solver add-in (Germinator_curve-fitting1.0.xls). The Solver is used in combination with the least sum of squares method to find the right parameters to fit the curves to the 4-parameter Hill function (El-Kassaby *et al.* 2008):

$$y = y_0 + \frac{ax^b}{c^b + x^b}$$

where y is the cumulative germination percentage at time x (hours), y_0 is the intercept on the y axis (≥ 0), a is the maximum cumulative germination percentage (≤ 100), b is controlling the shape and steepness of the curve and c is the time required for 50% of viable seeds to germinate (t_{50}). Initial values for the parameters a and c are extracted from the cumulative germination count and b is set to 20. With these initial values the solver performs an iterative process (max 10,000) until the sum of squares between the measured

cumulative germination and the calculated curve does not decrease any further. Because in rare cases the first iteration does not result in optimal parameters a second iteration is performed using the results of this first iteration as starting values. The iteration resulting in the lowest sum of squares is taken as the final result. The user can define a threshold for the minimum number of germinated seeds, since curve fitting on very small amounts of germinated seeds won't be very informative. In these situations the script returns a 'False' and the data is not used in the statistical analysis. Uniformity (U_{b-a}) of germination is the time interval between a% and b% of viable seeds to germinate. Users can define values used for a and b. The Area Under the Curve (AUC) is the integration of the fitted curve between t = 0 and a user-defined endpoint, which results in a parameter that combines information on maximum germination, t_{50} and uniformity. As described by El-Kassaby *et al.* (2008) the AUC can also be used to calculate a dormancy index (DI), by subtracting the area under the curve after dormancy release (e.g by cold stratification) with the area under the curve of dormant seeds. By the same analogy the AUC can be used to measure the effect of any stress treatment and calculate a stress index (SI). The Germinator curve fitting script will summarize the results by calculating averages and standard errors for repeated samples, performing a student-t test, and provides a clearly formatted output including graphs for the different germination parameters.

Accuracy and flexibility

The completion of seed germination of Arabidopsis is a two-step process: first rupture of the testa, followed by the protrusion of the radicle through the micropylar endosperm (Liu *et al.* 2005). The germination of a single seed was followed in time with high resolution imaging (see www.germinator.wageningenseedlab.nl for a time-laps movie). The two steps of Arabidopsis germination are clearly distinguishable on these images: testa rupture after 35 hours followed by endosperm rupture after 40 hours (Figure 2.2A). The difference in the threshold area and threshold XY position in time are shown in Figure 2.2B. This figure clearly shows that both the increase in area and the shift in XY position can serve as accurate indicators for germination sensu stricto.

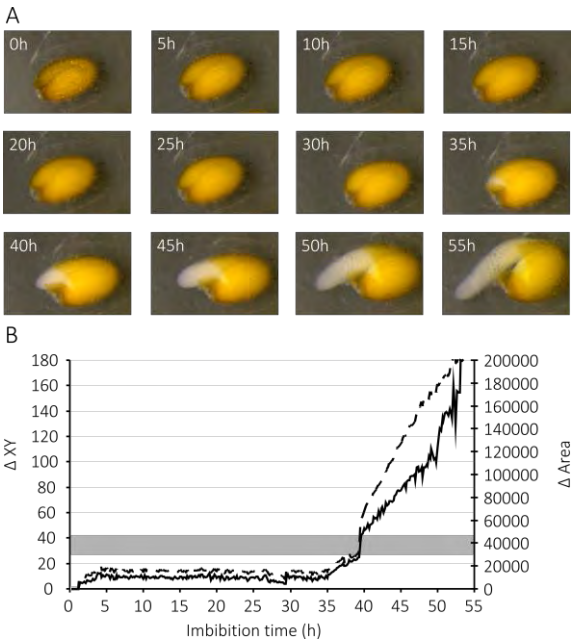


Figure 2.2: Analysis of germination by comparing the difference in either the area or XY position from a double color-threshold approach using a 10 minutes timelaps high resolution imaging series of a single seed. Dashed line = delta XY, solid line = delta area. The gray bar indicates the range for both area and XY position in which germination sensu stricto can be determined.

To test the accuracy of the automatic germination scoring with lower resolution images that can be used to study seed batches we performed an interval experiment measuring the progression of germination of seed lots from two Arabidopsis thaliana accessions (Ler; 147 seeds and Col. (Col-0); 172 seeds) every hour. The automatic counts were verified at 9 time points by manual counting (Figure 2.3 and Supporting Information, S2.1).

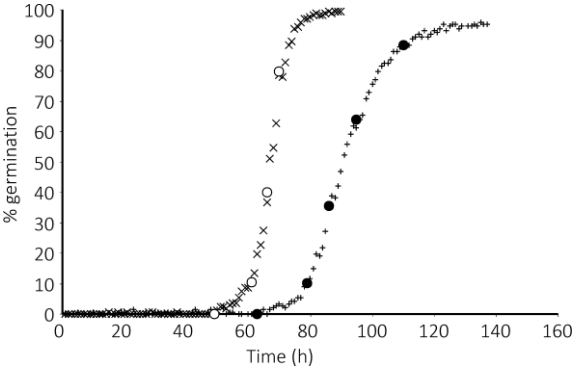


Figure 2.3: Comparison of manual (open circles; Col-0, filled circles; Ler) and automated scoring of Arabidopsis thaliana Col-0 (x) and Ler (+) germination.

Measuring germination at one-hour intervals provides very accurate data, which enables precise curve-fitting. However, it is impossible to apply this without expensive robotics in large scale experiments. Therefore, we wanted to assess the effect of the number of measurements on the accuracy of the fitted curve using the data from the

experiment depicted in Figure 2.3. Different time intervals were artificially created by removing data points from the one hour interval dataset and curve fitting was tested to assay the minimum number of required data points during germination (Table 2.1, Supporting Information S2.2).

Table 2.1: Comparison of germination curve fitting parameters t_{50}, uniformity (U_{7525}) and area under the curve until 120h (AUC_{120}) with measurements at different time intervals (hours).

	Col-0				Ler			
Interval (h)	t_{50}	U_{7525}	AUC_{120}	r^2 fit	t_{50}	U_{7525}	AUC_{120}	r^2 fit
1	67.06	6.19	81.62	0.999	90.78	14.72	56.21	0.999
4	67.08	6.06	81.64	0.999	90.75	14.78	56.26	0.999
8	67.36	6.50	81.36	0.999	90.85	15.41	56.47	0.999
12	66.64	6.11	81.94	1.000	91.55	15.55	55.64	1.000
16	66.83	4.35	81.79	1.000	91.81	15.15	55.39	1.000

Abbreviations: r^2 fit - determination coefficient; t_{50} - time to obtain 50% of germinated seeds; U_{7525} - time between 25% and 75% of germinated seeds; $AUC_{(120)}$ - Area under the curve until 120h.

From the example in Table 2.1 it is clear that our curve fitting module is able to accurately predict the various parameters. The desired interval will be dependent on the required accuracy and the level of difference between samples. Often, it is more convenient for practical reasons to use flexible intervals. Therefore, we tested with 5 replicates of an Arabidopsis thaliana Col-0 seed lot for which only six time points were acquired; care had been taken to obtain at least two measurements during the exponential phase of the curve (Figure 2.4, Table 2.2).

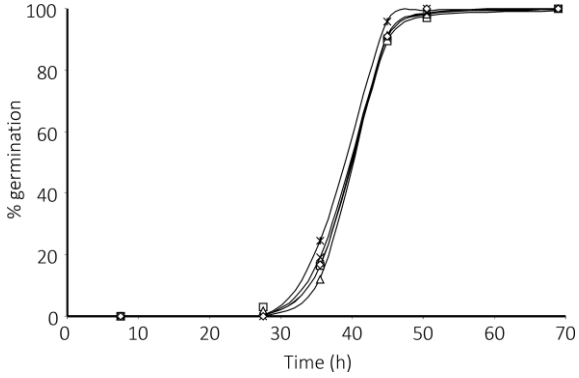


Figure 2.4: Germination curves of 5 replications (indicated by the various symbols) of Arabidopsis thaliana Col-0.

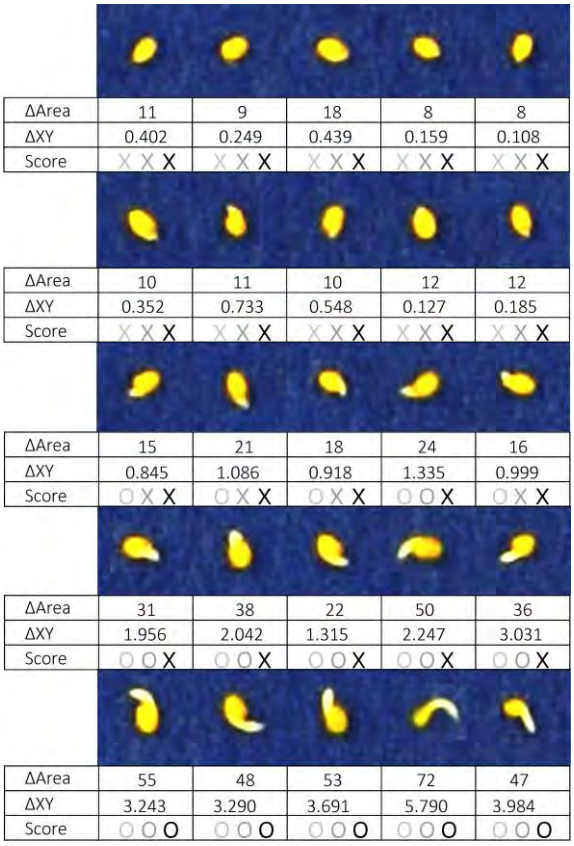
Table 2.2: Parameters characterizing seed germination curves of 5 replications of Arabidopsis thaliana Col-0 (Figure 2.4)

Replicate #	1	2	3	4	5
Number of seeds	78	67	58	63	49
r^2 fit	1.00	0.99	1.00	1.00	1.00
t_{50}	39.1	39.3	39.5	38.9	37.8
U_{7525}	5.1	5.4	4.7	5.5	4.6
$AUC_{(60)}$	20.7	20.4	20.3	20.8	22.1

Abbreviations: r^2 fit - determination coefficient; t_{50} - time to obtain 50% of germinated seeds; U_{7525} - time between 25% and 75% of germinated seeds; $AUC_{(60)}$ - Area under the curve until 60h.

This experiment shows that an accurate prediction for the various germination parameters can be obtained by as less as 6 data points with two data points in the exponential phase of the curve.

Salt tolerance during germination is an important but complex trait. Salt stress consists of an ionic and an osmotic component which is influencing homeostasis signaling pathways, detoxification response pathways, and pathways for growth regulation (Zhu 2002). Therefore, salt may influence the lag phase between testa and endosperm rupture, radicle growth and seedling establishment and germination on salt was used to test the accuracy of the measurement of germination sensu stricto. Figure 2.5 shows 25 seeds at different stages of germination during imbibition in 125 mM NaCl. Careful optimalization of the threshold for both the area and XY difference between the double color-threshold enables scoring of germination which resembles manual scoring as close as possible. The most accurate threshold settings can be determined with the 'parameter screen' script that we included in the Germinator_table file.



ΔArea	11	9	18	8	8
ΔXY	0.402	0.249	0.439	0.159	0.108
Score	X X X	X X X	X X X	X X X	X X X
ΔArea	10	11	10	12	12
ΔXY	0.352	0.733	0.548	0.127	0.185
Score	X X X	X X X	X X X	X X X	X X X
ΔArea	15	21	18	24	16
ΔXY	0.845	1.086	0.918	1.335	0.999
Score	O X X	O X X	O X X	O O X	O X X
ΔArea	31	38	22	50	36
ΔXY	1.956	2.042	1.315	2.247	3.031
Score	O O X	O O X	O O X	O O X	O O X
ΔArea	55	48	53	72	47
ΔXY	3.243	3.290	3.691	5.790	3.984
Score	O O O	O O O	O O O	O O O	O O O

Figure 2.5: Arabidopsis germination on 125 mM NaCl on blue filter paper. A compilation of images of different stages of germination derived from the original images used for automatic scoring is shown. Indicated are the differences in XY position and area between two color-thresholds and scoring of germination based on different threshold settings (x=not germinated, 0=germinated at thresholds: light grey = 20 area/0.8 XY, grey = 30 area/1.2 XY, black = 40 area/3.0 XY)

Both the germination rate and the maximum germination capacity are inhibited by sodium chloride (NaCl). We used a concentration range of NaCl to test the accuracy and

flexibility of the automatic scoring on six replicates of an Arabidopsis thaliana Col-0 seed lot, here we have set the limit for Δarea to 30 pixels and ΔXY to 1.2 (Figure 2.6, Supporting Information S2.4).

Separation of the germination behavior in specific parameters can help to describe and compare many lines. Figure 2.7 shows a comparison of the four different parameters and their discriminative power for germination under salt stress based on the germination characteristics shown in Figure 2.6.

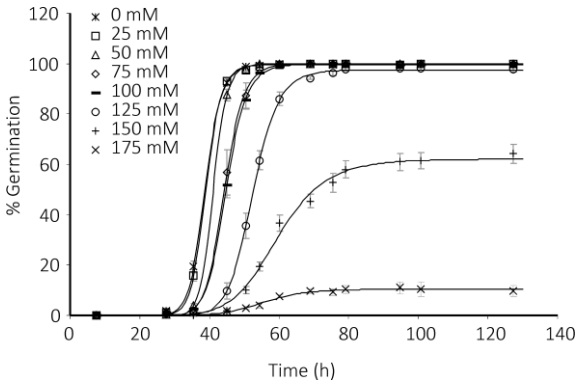


Figure 2.6: Germination of Arabidopsis thaliana (Col-0) seeds on different concentrations of NaCl, analyzed with the Germinator curve fitting module. Error bars represent SEM (n=6).

As shown in Figure 2.6 and Figure 2.7 the cumulative germination is inhibited by NaCl. At lower concentrations only the t_{50} is reduced where at higher concentrations the maximum germination (G_{max}) is affected as well.

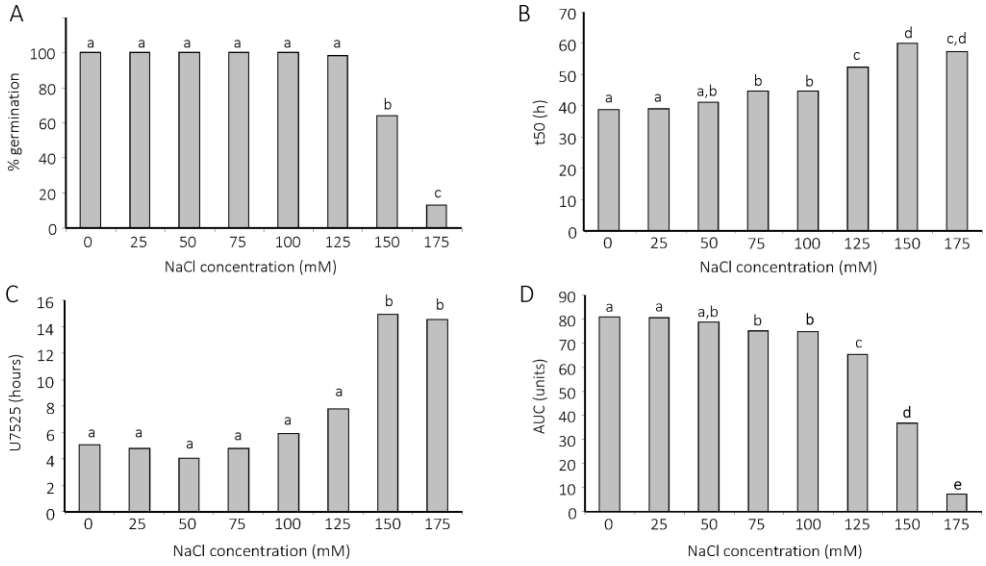


Figure 2.7: Germination of Arabidopsis thaliana (Col-0) seeds on different concentrations of NaCl. A) Maximum germination after 5d, B) time to reach 50% germination, t_{50}, C) uniformity (U_{7525}), D) Area under the curve until 120h (AUC_{120}). Letters (a,b,c,d,e) represent statistical different subsets (Tukey HSD, p=0.05)

Natural variation for salt tolerance

To fully exploit the power of high throughput cumulative germination data we used the Germinator scripts to screen the core set (165 lines) of a Bay-0 x Sha recombinant inbred population (Loudet *et al.* 2002) for salt tolerance. After-ripened seeds were germinated on water and 100 mM NaCl without prior stratification. We performed duplicate measurements of three different harvests resulting in a total of 1980 individual germination assays. Values obtained for germination on 100 mM salt were subtracted from values derived from germination on water (Supporting Information, S2.5). Figure 2.8A-C shows the frequency distribution of non-normalized data for G_{max}, t_{50} and AUC in this population. Both parental lines are indicated with an arrow showing the large extent of transgression. After normalization (see Experimental procedures for details) of these trait data we detected multiple QTLs for salt tolerance in 6 regions (Figure 2.8D). The QTLs for both maximum germination and area under the curve could explain 49% of the total variance. The QTLs for t_{50} could explain 39% of the total variance. No QTL for uniformity (U_{7525}) was detected. The QTL on top of chromosome 1 is affecting germination capacity (G_{max}) but not t_{50}. By contrast, we see QTLs on chromosome 5 that affect rate of germination without affecting germination capacity.

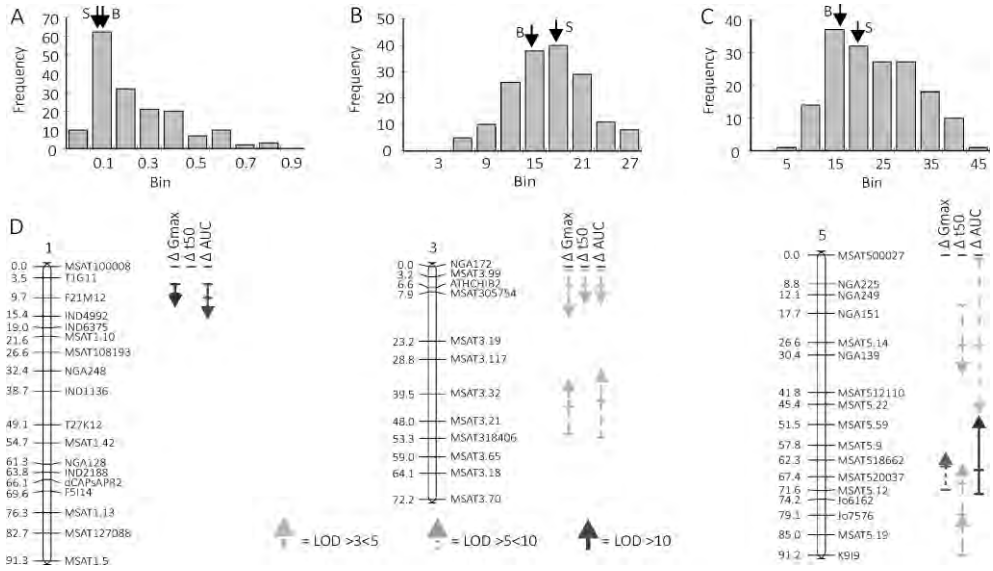
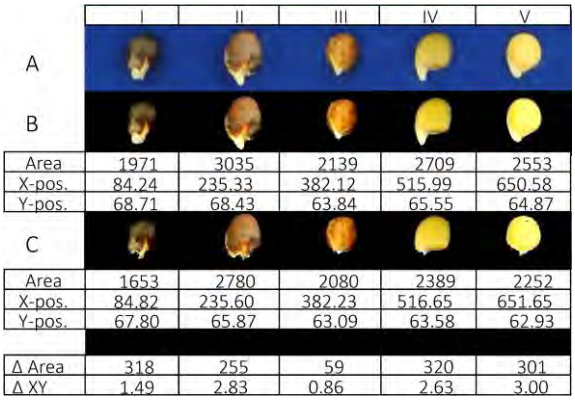


Figure 2.8: QTL analysis of salt tolerance of germination. A) Frequency distribution of non-normalized data for G_{max}, B) t_{50} and C) AUC in the Bay-0xSha RIL population for germination on 100 mM salt and corrected for germination on water: $\Delta G_{max} = G_{max}(water) - G_{max}(salt)$, $\Delta t_{50} = t_{50}(salt) - t_{50}(water)$, $\Delta AUC = AUC(water) - AUC(salt)$. D) The Bay-0xSha linkage map showing the genetic locations affecting germination on 100 mM salt. Mapped traits are indicated above each lane. Grayscales of the arrows indicate the LOD-score (darker = higher LOD scores). Arrows indicate the direction of the phenotypic effect; up: Sha increasing, Bay-0 decreasing; down: Bay-0 increasing, Sha decreasing. The length of the arrow depicts the 2-LOD support interval determined with restricted MQM mapping.

Scoring Brassica germination

Currently the whole Germinator procedure is optimized for use with *Arabidopsis* but it is probably suitable to handle many other species as well. To test whether the same script is also applicable to other species we tested some lines from a *Brassica* doubled haploid population from which the seeds strongly varied in color. Although some seeds are almost black while others are pale yellow it was possible to define two color thresholds which can distinguish between the protruding radicle and the rest of the seed (Figure 2.9). The differences between the two color thresholds can be used to automatically score germination, here we have set the limit for Δarea to 100 pixels and ΔXY to 1.4 (Figure 2.10, Supporting Information S2.6). Every seed with values below one of both limits will be scored as not germinated (e.g. Figure 2.9, seed III).



	I	II	III	IV	V
Area	1971	3035	2139	2709	2553
X-pos.	84.24	235.33	382.12	515.99	650.58
Y-pos.	68.71	68.43	63.84	65.55	64.87
Area	1653	2780	2080	2389	2252
X-pos.	84.82	235.60	382.23	516.65	651.65
Y-pos.	67.80	65.87	63.09	63.58	62.93
Δ Area	318	255	59	320	301
Δ XY	1.49	2.83	0.86	2.63	3.00

Figure 2.9: Five Brassica seeds which strongly vary in seed coat color were analyzed with the Germinator scripts. A) original image; B) image after color threshold with settings: + $Y_{0\text{-}255}U_{0\text{-}125}V_{135\text{-}255}$; C) image after color threshold with settings: – $Y_{0\text{-}255}U_{0\text{-}125}V_{120\text{-}255}$. The difference in Δ Area and Δ XY enables automatic scoring of germination.

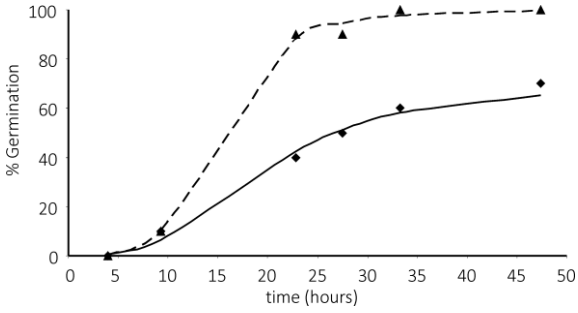


Figure 2.10: Germinator scripts were used for automatic scoring of Brassica seed germination with strong variation in seed coat color. Two different mixtures of 24 lines that strongly vary in seed color were used.

Discussion

Features of Germinator

The procedure presented here represents a novel and efficient analysis tool for high-throughput monitoring of seed germination. A few other studies have used image analysis to score seed germination for other plant species such as cabbage, broccoli, cauliflower, sunflower, lentil, pepper, radish and tomato (Dell'Aquila *et al.* 2000; Dell'Aquila 2004; Dell'Aquila 2005; Ducournau *et al.* 2005; Dell'Aquila 2007; Teixeira *et al.* 2007). Basically all systems use a fixed imaging system which allows full automated scoring of germination. The advantage from such a setup is the possibility to precisely follow phenotypic properties of individual seeds and germination can be scored based on e.g. increase in seed size or changes in roundness of the individual seeds. Since all these systems require a proper alignment between consecutive images they do not allow high-throughput analysis without huge investments in robotics. Therefore, we have chosen for a semi-automatic approach and have developed a system that can handle many samples which may be germinated at different environmental conditions. The power of the presented procedure is that it does not score germination based on the difference between two consecutive images but instead uses the information from two different color threshold analyses on a single image. This allows a much more flexible setup for screening large populations, but requires a good level of contrast between the radicle and seedcoat which may limit the usability for several species.

The high level of automation during experimental setup, image analysis and curve fitting provides a solid, reproducible and yet flexible system which can be implemented at very low costs. We have shown the accuracy of the automatic scoring and showed that only a limited number of measurements are needed for an accurate prediction of the germination curves. To achieve best accuracy care should be taken to obtain at least two measurements in the exponential phase of germination, although this can be difficult when screening large populations with different times and rates of germination. Therefore, a critical assessment of the calculated germination curves cannot be skipped. Also, fungal contamination during the experiment should be prevented as much as possible because this may cause false positive scoring of germination. The flexibility of the curve fitting is shown by germinating *Arabidopsis* on different concentrations of NaCl which caused reduced germination percentage, rate and uniformity. The curve fitting module was able to efficiently and accurately describe those curves.

High-throughput phenotyping

The availability of large genetic and mutant populations offers great potential for seed science research but also demands a high-throughput procedure to score seed germination. Until now, scoring of *Arabidopsis* germination is a time consuming task,

requiring manual inspection using binoculars. Especially in large scale experiments these human observations can easily lead to misjudgment and the number of experiments which can be handled is restricted by the desired interval between two inspections and the time it takes to count the individual experiments. This time is reduced dramatically in the procedure suggested here. Scoring six individual experiments (each 50-200 seeds) is reduced to the time it takes to take one photograph (approx. 5 seconds). Registration of time intervals is automated and can therefore be corrected for each individual experiment. The short measuring time also contributes to the accuracy and reliability. Together, this enables the researcher to follow the cumulative germination in time and determine the germination curves in large scale experiments. Nevertheless, it should also be clear that automatic scoring of germination cannot be left unsupervised. If screening environmental perturbations like germination in the presence of NaCl one might expect effects on the lag time between testa and radicle protrusion and effects on radicle growth and seedling establishment. This requires accurate parameter thresholding which can only be achieved by manual counting of a small subset.

Curve fitting for cumulative germination data

As described by Brown and Mayer (1988) fitted curves allow germination to be summarized in terms of a few curve coefficients, which offers a much better description of the time-course of germination than single value indices, such as the widely used maximum germination percentage. To optimize our analysis pipeline the Germinator curve fitting script was developed. It provides a fast and easy tool for fitting germination curves from cumulative germination data. On a standard desktop computer (dual core 2.3, 4 Gb, Windows XP) the curve fitting and parameter extraction for 5000 germination tests was calculated in less than 15 minutes. The overall quality of the fit is of course strongly dependent on the quality and amount of data points but overall the coefficient of determination was close to 1.00. Here it should be noted that less data points automatically results in a higher coefficient of determination but that this might not always reflect the true germination curve. Therefore, the value attached to the coefficient of determination in experiments with only a few datapoints should be considered with care. The output from the Germinator curve fitting script contains all the parameters from the 4PHF function, total germination (G_{max}), time to reach 50% germination (t_{50}), uniformity (U_{b-a}) and Area under the Curve (AUC). It offers the possibility to depict both the individual data points and the fitted curves and it can summarize the data by calculating averages and performing student t-tests. The Germinator curve fitting script enables analysis of general cumulative germination data and can be used for all plant species.

Gathering detailed germination data in an experiment on salt tolerance (Figure 2.6) clearly shows the added value of the cumulative germination curve compared to e.g. the total germination after 5 days. The latter is not discriminative until a concentration of 125 mM NaCl. The t_{50} is not discriminative until 75 mM NaCl and not between 150 and 175

mM. The uniformity (U_{7525}) only shows a significant difference between the 125 mM and 150 mM points. The combined interpretation of the parameters G_{max}, t_{50} and U_{7525} can accurately describe the cumulative germination curve. The AUC is summarizing these three parameters effectively and shows optimal discrimination among the different treatments (Figure 2.7). Calculating the difference of the area under the curve (AUC) between germination on water and germination on a specific concentration of NaCl generates a value (Stress Index, SI) which can summarize the effect of the salt treatment based on maximum germination, t_{50} and U_{7525}. This approach can be used for any type of stress as well as for the release of dormancy by e.g. cold stratification. This parameter was introduced by El-Kassaby, *et al.* (2008) as a useful index to describe dormancy (Dormancy Index, DI). We suggest to use this parameter as well for normalized values of stress-treatments (Stress Index, SI).

Natural variation for salt tolerance

The ability to handle large scale experiments was shown in a screen for allelic variation for salt tolerance in the *Arabidopsis* Bay-0 x Sha recombinant inbred population (165 lines). Repetitions and water control experiments raised the number of individual germination experiments to 1980. It would have been impossible for one person to manually count this large number of experiments multiple times a day. The same experiment also clearly shows the large benefit of acquiring detailed germination curves (Figure 2.8D). The QTL on top of chromosome 1 is affecting germination capacity (G_{max}) but not rate of germination (t_{50}). On the contrary, on chromosome 5 QTLs are observed that affect rate of germination without affecting germination capacity. The area under the curve (AUC) is summarizing both parameters and show QTLs that are affected either in germination capacity or rate. Multiple QTLs for germination on NaCl were identified in 6 regions. Distinct loci where either Bay-0 or Sha alleles improved germination were found, which could explain the observed transgression (Figure 2.8A-C). Comparing QTLs for maximum germination capacity found in the *Arabidopsis* Ler x Sha population (Clerkx *et al.* 2004) revealed that the QTLs on chromosome 1, 3 and lower arm of chromosome 5 could be in the same regions and show similar directions of the Shakdara allelic effects. One of the apparent advantages of using cumulative germination data over endpoint germination is the ability to measure genetic variation for stress tolerance at lower concentrations (Figure 2.7). Furthermore, it is known that salt tolerance is realized via distinct pathways for high and low salt concentrations (Munnik *et al.* 1999). Cumulative germination data might allow separate analysis of these pathways, whereas endpoint germination might be restricted to pathways for higher concentrations.

Prospects

Although we optimized the Germinator scripts for *Arabidopsis thaliana* we were able to show that the same basic setup can also be employed for other seeds which have a

good contrast between the seed coat and protruding radicle. The *Brassica* seeds we used for this test displayed considerable variation in seed color but, nevertheless, it was possible to define a color threshold setting that efficiently distinguished between seed coat and the protruding radicle.

We show that we have developed a package for high-throughput seed germination phenotyping. The Germinator pipeline offers a well-defined and robust experimental setup but is very flexible in terms of numbers and treatments. The improved efficiency and absence of subjectivity are great advantages of computer aided assessment. The procedure presented in this paper offers great potential to perform high-throughput germination tests in large mutant or genetic populations. Automatic germination scoring is optimized for use with *Arabidopsis* and will most likely work for many other species as well. The curve fitting script enables analysis of general cumulative germination data and can be used for all plant species. Although we tried to optimize the package it is of crucial importance to set accurate thresholds by comparing the automated scoring with manual scoring. In conclusion, Germinator is a low-cost package that allows the monitoring of several thousands of germination tests, several times a day by a single person.

Experimental procedures

Plant material and growth conditions

Arabidopsis thaliana plants from accessions Columbia and Landsberg erecta were grown on soil in a climate chamber (20°C day, 18°C night) with 16 hours of light (35W/m^2) at a relative humidity of 70%. Seeds were bulk harvested and stored at 20°C under ambient relative humidity (around 40%) for 5 months. Seeds from the core population (165 lines) of the *Arabidopsis* Bayreuth-0 x Shakdara recombinant inbred population (Loudet *et al.* 2002) were obtained from the Versailles Biological Resource Centre for Arabidopsis (http://dbsgap.versailles.inra.fr/vnat/) and were grown in triplicate of 5 plants each in a fully randomized setup. Plants were grown on 4x4 cm rockwool plugs (MM40/40, Grodan B.V.) and watered with 1 g/l Hyponex (NPK=7:6:19 http://www.hyponex.co.jp) fertilizer in a climate chamber (20°C day, 18°C night) with 16 hours of light (35W/m2) at a relative humidity of 70%. Seeds were bulk harvested and after-ripened until they reached their maximum germination after 5 d of imbibition. Subsequently, the seeds were dried for 1 week at a relative humidity of 20% and stored at -80°C until further experimentation. To prevent fungus contamination during the experiment we surface sterilized the seeds by placing 50 mg of seeds per seed lot for 2 hours in a desiccator jar above a solution of 100 ml 4% sodium hypochlorite + 3 ml concentrated HCL. *Brassica rapa* plants from a combined Double Haploid (DH) population containing plants from the DH38 population of Ping Lou et al. (2008) and plants from similar but reciprocal crossing were grown in the greenhouse. Seeds were harvested and stored at 20°C until use. Seeds from 24 lines representing the different classes of seed coat colors were used to test the Germinator.

Germination assay

Germination experiments were performed in plastic (15x21 cm) trays (ref 109, DBP Plastics, Belgium; www.dbp.be) containing 42 ml water or NaCl solution and two layers of blue filter paper (5.6' X 8' Blue Blotter Paper, Anchor Paper Company, St Paul, MN; www.seedpaper.com). Six samples of approximately 50-200 Arabidopsis seeds were dispersed on the filter paper using a mask to ensure an accurate and reproducible spacing. Clustering of seeds was prevented as much as possible. A maximum of 20 trays were piled with, on both the top and the bottom of the stack, two empty trays with 42 ml water and 2 layers of blue filter paper to prevent unequal evaporation and ensure equal distribution of light. The whole pile was wrapped in a closed transparent plastic bag and placed in an incubator. The incubator (type 5042, Seed Processing Holland, Enkhuizen, The Netherlands; www.seedprocessing.nl) provides light from 3 sides and was set to a temperature of 20°C. For the interval experiment, the lower filter paper was used as a wedge inserted in a tray filled with water to prevent drying of the seeds and enable automatic hourly measurements. The experiment was carried out in an air-conditioned room (20°C). Experiment set up, automatic scoring and curvefitting was performed with the germinator package.

Imaging

A digital camera (Nikon D80 with Nikkor AF-S 60mm f/2.8 G Micro ED) was fixed to a repro stand and connected to a computer, using Nikon camera control pro software version 2.0. Two vertically placed fluorescent tl-tubes (150 cm), 1.5 meter left and right from the camera, were used as indirect light source; great care was taken to prevent any reflection. The camera was set to full manual control (ISO400, F/18, 1/3 sec, manual focus). Image files are named following a strict convention: mmddyy-hhmm#seq, whereby the seq is an automatic sequential number indicating the tray number. A position mask is used to make sure that the trays are placed at the correct position under the camera.

QTL analysis

For QTL analysis a genetic map consisting of 69 markers (provided by dbsgap.versaiiles.inra.fr/vnat/) with an average distance between the markers of 6.1 cM was used. To test and correct normality of the trait values we used the software package Distribution analyzer v1.2 (www.variation.com). Multiple-QTL model mapping (MQM) was carried out by using the software package MapQTL (version 5.0, Kyazma B.V. Wageningen, The Netherlands). Cofactors were selected according to the program's reference manual and the 2LOD interval was determined with restricted MQM mapping. MapChart v2.2 (Plant Research International, Wageningen, The Netherlands) was used to construct the linkage map shown in Figure 2.8.

Downloading Germinator

The full Germinator package (Windows operating systems) is freely available for the scientific community. It can be downloaded from www.wageningenseedlab.nl. This website also contains a full manual and video demonstrations about the use of the various modules.

Acknowledgements

This work was supported by the Technology Foundation STW, the Applied Science Division of NWO and the Technology Program of the Ministry of Economic Affairs. We would like to thank Marie Retiere for testing the Brassica seeds.

Supporting Information

Supporting information can be downloaded from either the online version of this article (Joosen *et al.* 2010) or from:
www.wageningenseedlab.nl/thesis/rvljoosen/SI/chapter2

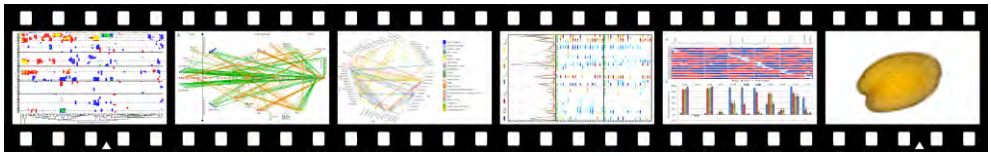
Table S2.1: The Germinator curve fitting module containing the raw cumulative germination data, fitted curves and extracted parameters from the interval experiment shown in Figure 2.2.
Table S2.2: The Germinator curve fitting module containing the raw cumulative germination data, fitted curves and extracted parameters from the reduced interval experiment shown in Table 2.1.
Table S2.3: The Germinator curve fitting module containing the raw cumulative germination data, fitted curves and extracted parameters from 5 replicates of Col-0 as shown in Figure 2.3 and Table 2.2.
Table S2.4: The Germinator curve fitting module containing the raw cumulative germination data, fitted curves and extracted parameters from germination on various concentrations of NaCl as shown in Figure 2.4 and Figure 2.5.
Table S2.5: Values for germination on NaCl for the *Arabidopsis* Bay-0 x Sha recombinant inbred core population (Figure 2.6). All values are corrected for germination on water by subtraction.
Table S2.6: The Germinator curve fitting module containing the raw cumulative germination data, fitted curves and extracted parameters from scoring *Brassica* germination as shown in Figure 2.8.

"It is from a small seed that the giant Iroko tree has its beginning."

3 Visualizing the genetic landscape of *Arabidopsis* seed performance

Joosen RVL, Arends D, Willems LAJ, Ligterink W, Jansen RC, Hilhorst HW (2012).

Plant Physiology. Vol. 158(2): 570-589



Abstract

Perfect timing of germination is required to encounter optimal conditions for plant survival and it is the result of a complex interaction between molecular processes, seed characteristics and environmental cues. To disentangle these processes we made use of natural genetic variation present in an *Arabidopsis thaliana* Bayreuth x Shahdara RIL population. For a detailed analysis of the germination response we characterized rate, uniformity and maximum germination and discussed the added value of such precise measurements. The effects of after-ripening, stratification and controlled deterioration as well as the effect of salt (NaCl), mannitol, heat, cold and ABA with and without cold stratification were analyzed for these germination characteristics. Seed morphology (size, length) of both dry and imbibed seeds was quantified by using image analysis. For the overwhelming amount of data produced in this study we developed new approaches to perform and visualize high throughput QTL analysis. We show correlation of trait data, (shared) QTL positions and epistatic interactions. The detection of similar loci for different stresses indicate that often the molecular processes regulating environmental responses converge into similar pathways. Seven major QTL hotspots were confirmed using a HIF approach. QTLs co-locating with previously reported QTLs and well characterized mutants are discussed. A new connection between dormancy, ABA and a cripple mucilage formation due to a natural occurring mutation in the *MUM2* gene is proposed and this is an interesting lead for further research on the regulatory role of ABA in mucilage production and its multiple effects on germination parameters.

Introduction

Seed Germination

Colonizing plants are subject to a wide variety of environmental conditions. For successful adaptation to new habitats the timing of developmental transitions is especially important. Seed germination is one of these important transitions as it determines the seasonal environment experienced in further plant life (Huang *et al.* 2010). Natural populations that develop under distinct environmental conditions may reveal genetic adaptation, which can be used to disentangle the signaling routes that are involved. Seed germination is described by three phases of water uptake. In phase I the seed imbibes and reinitiates metabolic processes followed by a lag phase (phase II). Further water uptake results in protrusion of the radicle through the testa and endosperm (phase III). The moment of radicle protrusion through the endosperm is considered to be the moment of germination sensu stricto (Finch-Savage and Leubner-Metzger 2006). To characterize the genetic variation of germination related traits we focused on the effect of the environment that a seed perceives during germination rather than the effect of the environment during maternal plant growth, which has been the subject of other studies (Gutterman 2000; Dechaine *et al.* 2009; Elwell *et al.* 2011). Seed content (e.g. oil) is often used as commodity and modifications to the content can therefore be regarded as seed quality parameters as well. To prevent confusion we will use the term seed performance to indicate that the focus of our study was restricted to seed germination characteristics.

The production of high quality crop seed not only entails knowledge about maternal plant growth, harvesting and storage of seeds, but also of germination conditions (Rivero-Lepinckas *et al.* 2006). To obtain better germination and field performance, many seed companies rely on enhancement methods, such as seed priming and coating and/or pelleting, but these methods are reaching their limits. Dissecting the molecular mechanisms underlying seed germination and its tolerance to the environment may unlock the full genetic potential and enable targeted breeding for seed performance.

In this study we used a recombinant inbred line (RIL) population derived from two *Arabidopsis thaliana* ecotypes: Bayreuth (Bay-0) which originates from a fallow land habitat in Germany and Shahdara (Sha) which grows at high altitude in the Pamiro-Alay mountains in Tadjikistan (Loudet *et al.* 2002). The Bay-0 x Sha RIL population has been used in many previous studies to map QTL positions for root morphology (Loudet *et al.* 2005; Reymond *et al.* 2006), anion content (Loudet *et al.* 2003), nitrogen use efficiency (Loudet *et al.* 2003), cell wall digestibility (Barriere *et al.* 2005), carbohydrate content (Calenge *et al.* 2006), sulfate content (Loudet *et al.* 2007), leaf senescence (Diaz *et al.* 2006), morning-specific growth (Loudet *et al.* 2008) and cold-dark germination (Meng *et al.* 2008). We have used the natural variation present in this RIL population to map the response of germination characteristics to environmental conditions to which a seed is exposed.

Freshly harvested viable *Arabidopsis* seeds often don't germinate even when placed under conditions favorable for germination. This event, called primary dormancy, is shown to be subject to natural variation (Bentsink *et al.* 2010). In many Arabidopsis ecotypes, this primary dormancy is released after a period of dry storage at room temperature. Another dormancy breaking treatment is cold stratification where seeds are imbibed in water and stored at 4°C in the dark for four days before putting them into optimal conditions for germination (Finch-Savage and Leubner-Metzger 2006). Unfavorable conditions during seed germination may result in a changed rate or even failure of germination. In Arabidopsis, it has been shown that the responsiveness to temperature is closely related to the level of after-ripening (Tamura *et al.* 2006). High salt concentrations induce osmotic stress and ion toxicity resulting in both a delay and reduction of maximum germination (Galpaz and Reymond 2010). Often, these different environmental stresses are interconnected and will cause osmotic and associated oxidative stress (Zhu 2002; Chinnusamy *et al.* 2004). The plant hormone Abscisic Acid (ABA) plays a predominant role in plant responses to different environmental stresses and can activate various signal transduction pathways leading to a global change in transcription (Finkelstein *et al.* 2002; Xiong *et al.* 2002). Exogenous application of ABA during germination results in a distinction between testa and endosperm rupture. At certain concentrations the testa will rupture but germination sensu stricto (radicle protrusion through the endosperm) will be inhibited. This phenomenon, caused by reduced weakening of the endosperm cap, is the consequence of a complex interplay between ABA, GA and ethylene signals (Linkies *et al.* 2009). In this report, we determined germination sensu stricto for primary dormancy in freshly harvested seeds, germination of fully after-ripened seeds with and without a preceding cold stratification period (see material and methods for conditions), and germination under various stress conditions (low/high temperature, salt/osmotic stress and ABA) to assess natural variation in the Bay-0xSha RIL population. Additionally, seed morphology (size and length) and flowering time were phenotyped as they have been shown to be strong determinants of plant trait variation (Chiang *et al.* 2009; Orsi and Tanksley 2009; Elwell *et al.* 2011). We correlated these traits to our germination related traits to evaluate possible causality. In total this analysis resulted in 327 trait scores over different harvests. Evaluation of these high numbers of phenotypes demanded methods of QTL analysis that extended beyond mapping of individual traits and that allowed comprehensive and comprehensible visualization.

Analysis of natural variation that is captured in well-defined recombinant inbred populations has shown to be a powerful tool to detect important loci that influence the traits under study (Alonso-Blanco *et al.* 2009). To uncover the loci with genetic variation a statistical framework is needed. For this, any programming language can be used which supports statistics. In the life sciences the statistical language R is often the prime candidate. R is open source, contains the latest in statistical analysis methods and has a large community for help and support (http://www.r-project.org/). Furthermore, it has the R/qtl package (Broman *et al.* 2003), which contains an array of different QTL mapping

methods, including Single Marker Mapping, Interval Mapping and Multiple QTL Mapping (MQM) (Arends *et al.* 2010). Although all possibilities to perform a detailed QTL analysis including data preprocessing and output formatting are present in R, it requires extensive knowledge of the R-syntax to combine all necessary steps in a single analysis protocol that can loop through hundreds or thousands of traits. In this paper we present a script that can perform these tasks. This type of automated analysis combined with efficient data visualization is a necessary step to keep up with the increasing rate of biological data production. For using single trait mapping the effect of a certain treatment, e.g. germination at high temperature, must be corrected by the germination characteristics under control conditions. Here, we subtracted the observed germination under stress conditions from values for germination under control conditions. This correction can lead to complicated interpretation, especially when the environment under study affects loci with already strong effects under control conditions. Further, it can reduce statistical power due to summation of the error components. Therefore we performed an additional analysis using a QTL by environment (QTLxE) approach (Jansen *et al.* 1994; Malosetti *et al.* 2004; Moreau *et al.* 2004; Eeuwijk *et al.* 2006). Instead of considering individual responses, one can then treat the stress conditions as a set of environmental perturbations and evaluate a single trait (such as germination percentage). Because several environments are taken into account simultaneously, the statistical power to detect loci that are affected by several environments increases and interpretation becomes more intuitive as the need for correcting the stress response by the control response is eliminated (Boer *et al.* 2007; Payne *et al.* 2011).

The Bay-0 x Sha RIL population consists of 420 lines that were genotyped in the F6. This relatively low degree of inbreeding provoked residual heterozygosity present at almost all genome positions. This residual heterozygosity can be used to confirm QTL positions, as it provides a possibility to study both parental alleles at the locus of interest in an elsewhere homozygous background (Tuinstra *et al.* 1997). In contrast to conventional near isogenic lines (NILs) the genetic background of heterogeneous inbred lines (HIFs) consist of a mix of the two parental genomes. The availability of a genome wide set of HIF lines for the Bay-0xSha RIL population provides a fast and accurate mean to confirm QTL loci.

Results

Phenotyping Seed Germination

To map the genetic architecture of seed germination traits we have used the core-population (165 lines) of the Bay-0 x Sha RIL population (Loudet *et al.* 2002). Flowering time was recorded during maternal plant growth and showed good correlation (Pearson r^2 = 0.75) with previous published data of this population (Loudet *et al.* 2002). We used freshly harvested seeds (2 weeks of after-ripening) and tested germination with and without stratification (see material en methods for conditions). Further, we tested germination of fully after-ripened seeds and assayed germination under several stress conditions, with and without stratification (Table 3.1). Germination was measured with the automated scoring system; the Germinator (Joosen *et al.* 2010, Chapter 2). Using this package we were able to describe the cumulative germination curve under all conditions tested. The germination curve can accurately be described by extracting five parameters: G_{max} = maximum germination, U_{8416} = uniformity of germination, time between 16 and 84% of germination, t_{10} = initiation of germination, time to reach 10% of germination, t_{50} = rate of germination, time to reach 50% of germination and AUC = area under the germination curve until 100h.

To reduce environmental variation we took great care in the growth and harvest of the maternal plants. We used a fully randomized setup and grew the population twice in a climate chamber. In the first growth we separated the harvest in 3 blocks (ABC), each with 3-5 plants/RIL. In the second growth we pooled the harvest of 4-7 plants/RIL (D). The overall Pearson correlation between block A-B, A-C and B-C was higher compared to A-D, B-D and C-D (Supporting Information, S3.1). Broad sense heritability's were calculated with the QTL data analysis tools in Genstat 14, using the preliminary single environment analysis and adding the block as an additional fixed term (Table 3.2). Heritability values can range between 0 (no heritability) and 1 (maximum heritability). Overall, heritability was high, indicating a large genetic variance and small effect of the different harvests. The lower heritability for maximum germination at low temperature (AR.NS.Cold) and after cold stratification (AR.WS) is the result of low genetic variance for these traits, as many of the lines germinated to 100% under these conditions. However, we were able to capture the genetic variation for these traits in the other parameters (AUC, t_{50}, t_{10} and U_{8416}). Although this breadth of phenotypic screens is common nowadays, easy tools for performing high throughput QTL mapping and generating clear overview figures were not available. It is important to detect and correct data errors, enable selection of traits that should be analyzed in more detail, detect possible epistatic interactions or find strong correlations between phenotypes. Because these are crucial steps in determining the biological relevance we invested in the development of an automated analytical protocol that allowed large scale single trait QTL analysis.

Table 3.1: Overview of traits in this study and the harvest(s) used for the measurement. The indicated color-code is used in all figures throughout this paper. For each mentioned experiment G_{max}, AUC, t_{50}, t_{10} and U_{8416} were determined. Abbreviations in codes: AR: after-ripened, NS: no stratification, WS: with stratification, CD: Controlled deterioration.

Trait group		Harvest	Description	Codes
Germination		ABCD	Germination of after-ripened seeds	AR.NS
After-ripening		ABCD	Delta between germination of freshly harvested seeds and germination of after-ripened seeds without stratification	AR.NS - Fresh.NS
Fresh + stratification		ABCD	Delta between germination of freshly harvested seeds without stratification and germination of freshly harvested seeds with stratification	Fresh.WS -Fresh.NS
AR + stratification		ABCD	Delta between germination of after-ripened seeds without stratification and germination of after-ripened seeds with stratification	AR.WS - AR.NS
NaCl		ABCD	Delta between germination of after-ripened seeds on 100 mM NaCl and germination of after-ripened seeds on water, without stratification (Joosen et al., 2010)	AR.NS - NaCl.NS
NaCl + stratification		ABCD	Delta between germination of after-ripened seeds on 125 mM NaCl and germination of after-ripened seeds on water, with stratification	AR.WS - NaCl.WS
Mannitol		AD	Delta between germination of after-ripened seeds on -0.5 mP Mannitol and germination of after-ripened seeds on water, without stratification	AR.NS - AR.Mann.NS
Mannitol + stratification		AD	Delta between after-ripened seed germination on -0.5 mP Mannitol and after-ripened seed germination on water, with stratification	AR.WS - AR.Mann.WS
Cold Fresh		D	Delta between germination of freshly harvested seeds at 10°C and germination of freshly harvested seeds at 20°C, without stratification	Fresh.NS - Fresh.Cold.NS
Cold		AD	Delta between germination of after-ripened seeds at 10 °C and germination of after-ripened seeds at 20°C, without stratification	AR.NS - AR.Cold.NS
Cold + stratification		D	Delta between after-ripened seed germination at 10 °C and germination of after-ripened seeds at 20°C, with stratification	AR.WS - AR.Cold.WS
Heat Fresh		D	Delta between germination of freshly harvested seeds at 30 °C and germination of after-ripened seeds at 20°C, without stratification	Fresh.NS - Fresh.Heat.NS
Heat		D	Delta between germination of after-ripened seeds at 30 °C and after-ripened seed germination at 20°C, without stratification	AR.NS - AR.Heat.NS
Heat + stratification		D	Delta between germination of after-ripened seeds at 30 °C and after-ripened seed germination at 20°C, with stratification	AR.WS - AR.Heat.WS
Controlled deterioration		D	Delta between germination of after-ripened seeds after controlled deterioration and germination of after-ripened seeds on water, without stratification	AR.NS - AR.CD.NS
Controlled deterioration + stratification		D	Delta between germination of after-ripened seeds after controlled deterioration and germination of after-ripened seeds on water, with stratification	AR.WS - AR.CD.WS
ABA		D	Delta between germination of after-ripened seeds with 0.5 µM ABA and germination of after-ripened seeds on water, without stratification	AR.NS - AR.ABA.NS
ABA + stratification		D	Delta between germination of after-ripened seeds with 0.5 µM ABA and germination of after-ripened seeds on water, with stratification	AR.WS - AR.ABA.WS
Seed size		ABD	Seed size and length of dry seeds	Size.Area
				Size.length
Seed size, imbibed		ABD	Seed size of imbibed seeds	Size.imbibed
Flowering time		ABC	Time till first open flower under long day (16D/8N) conditions	FTLD

Table 3.2: Overview of the broad sense heritability scores. Included are those traits for which different blocks were tested (Trait code descriptions can be found in Table 3.1)

Trait	G_{max}	AUC	t_{50}	t_{10}	U_{8416}
AR.NS	0.82	0.87	0.86	0.79	0.82
AR.NS.Cold	0.51	0.77	0.73	0.66	0.48
AR.NS.Mannitol	0.61	0.79	0.70	0.55	0.62
AR.NS.NaCl	0.9	0.94	0.80	0.76	0.43
AR.WS	0.63	0.79	0.78	0.72	0.72
AR.WS.NaCl	0.91	0.93	0.86	0.78	0.70
Fresh.NS	0.92	0.94	0.81	0.70	0.76
Fresh.WS	0.40	0.81	0.87	0.84	0.76

Single trait QTL mapping

To evaluate the response of germination to a certain treatment, we first subtracted the observed germination at test conditions from germination at the proper control conditions. For example, the effect of NaCl on germination after cold stratification is determined by subtracting G_{max} on NaCl from G_{max} on water. This subtraction was reversed for the rate and uniformity parameters to correct the reversed nature of these parameters (e.g. slower germination results in a larger t_{10} and t_{50}). Table 3.1 provides an overview of all corrections that have been applied.

An analytical protocol was designed, using the popular R/qtl package of R to analyze trait data of recombinant inbred populations with the multiple QTL model approach (Arends *et al.* 2010). When performing a detailed QTL analysis it is important that several steps are performed or checked. Missing genotypic data is imputed and a recombination frequency plot is generated (Figure 3.1A). In the next step, quality of the trait data is investigated. Outliers are detected and removed using a z-score transformation with a user defined threshold. To estimate the effect of data normalization on MQM mapping we have used the distribution analyzer version 1.2 (www.variation.com). A LOD score correlation plot comparing raw and normalized data (Supporting Information, S3.9) clearly shows that this does not affect the output. Therefore, we decided to use non-transformed data instead of fitting a polynomial distribution without proper biological rationale. As an extra control the results of the MQM mapping were always compared to standard interval mapping, using the parametric model with Haley Knott regression (Haley and Knott 1992) (Figure 3.1B). The whole genome additive effect was estimated based on the non-transformed data as half the difference between the phenotypic averages for the two homozygotes (Figure 3.1C).

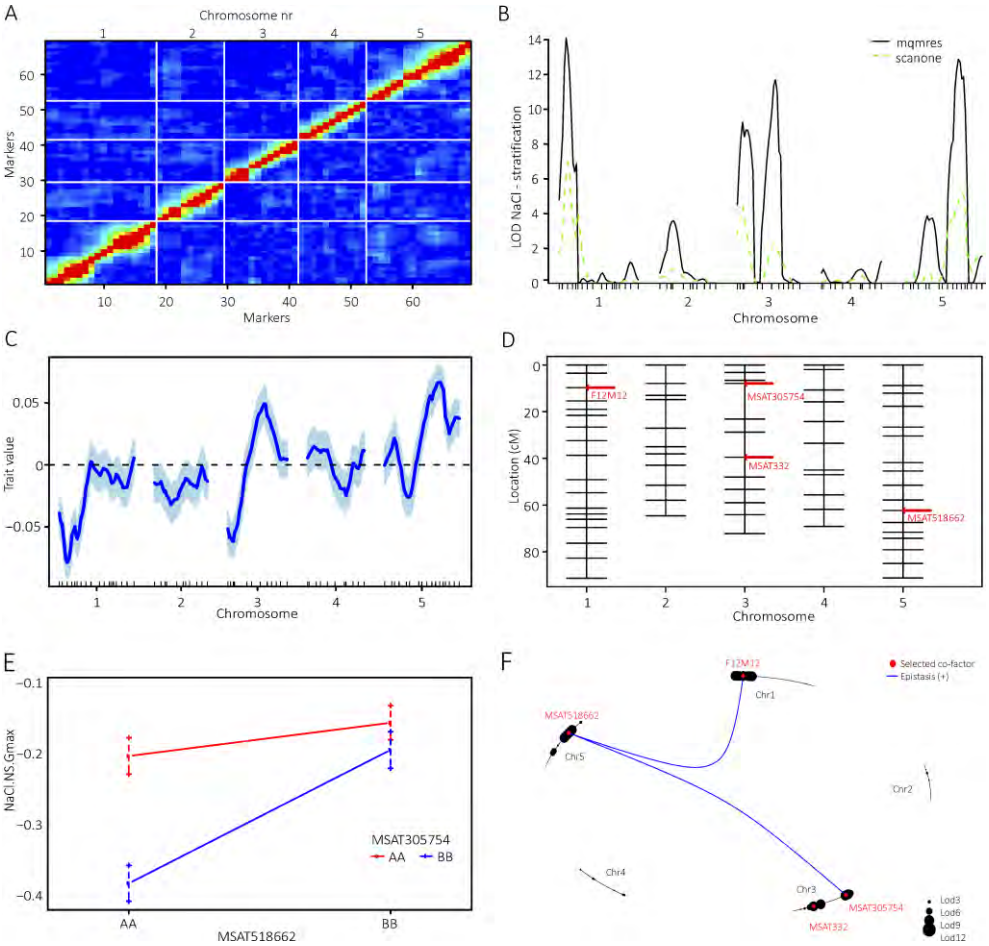


Figure 3.1: R/qtl output for the effect of 100mM NaCl on the maximum germination without stratification. A) Pairwise recombination fractions; B) LOD profile comparison between MQM and Haley-Knott (scan-one) interval mapping; C) Genome wide additive effect based on raw phenotype data D) Genetic map showing the significant QTL markers; E) Interaction plot showing the effect size comparison between markers MSAT305754 and MSAT518662 at the Sha (AA) and Bay-0 (BB) alleles; F) Circle plot showing epistatic interactions between all significant QTL.

R/qtl MQM uses a backward elimination of cofactors. As a rule of thumb one can select a maximum of n-16 initial cofactors with this procedure (Jansen 2008), with n being the number of lines in the RIL population. In our script, a cofactor file can be provided with the selection of the initial cofactors. When no cofactor file is provided, the analysis will be performed without cofactors resulting in an analysis comparable with the composite interval mapping (CIM) method. For the analysis of the Bay-0xSha population we selected 39 out of 69 markers as possible cofactors. Cofactors were selected based on their quality (least amount of missing data or heterozygous status) and physical cM position, attempting

to obtain intervals of about 10 cM. Although the procedure allows the selection of all 69 markers as cofactors, this does not improve mapping and only lowers statistical power due to the multiple testing correction in the permutation analysis. The provided cofactor file is used to perform automated backward elimination of cofactors. Backward elimination is performed to remove cofactors that do not significantly contribute to the fit of the initial model. This is achieved by comparing Akaike's information criterions (AIC) of the different models (Jansen 1993). Using the final selected QTL model, the mapping LOD scores are calculated for all genetic markers. Plots showing all significant markers are produced automatically (Figure 3.1D). We have used the procedure described to map all 327 individual measurements (Supporting Information, S3.2) but to enhance readability of this paper we only show average values for each trait (94 traits) (Supporting Information, S3.3).

Detection of Epistatic Interactions

Detection of epistatic interactions with the relatively limited sample sizes that are common in RIL populations is often cumbersome (Li *et al.* 2010). However, a useful hint of epistasis can be obtained when clear effects are visible between loci and the same interaction can be detected when measuring multiple traits. We calculated epistatic interactions between the QTL loci that are detected with the multiple QTL models from MQM. All possible combinations of the detected QTL loci were used to calculate the estimated interaction effect (Figure 3.1E). Interactions will be reported when their biological effect size is above a user-defined threshold, which is based on the number of standard deviation differences between two interacting loci. In this study we have set this threshold to 8. For each single trait a graphical visualization shows all interactions, using the circle plot routine from R/qtl (Figure 3.1F).

Multi Trait Visualizations

The interpretation of large numbers of phenotypes requires comprehensive visualization methods. Therefore, we produced several outputs that can help to dissect and interpret the data. A correlation plot based on trait values enabled a quick overview of similarities between all input traits (Figure 3.2, bottom left panel). After QTL mapping of all traits, this was also done based on the LOD profiles (Figure 3.2, top right panel). This plot shows a strong correlation between the effect of after-ripening and stratification on fresh seeds, indicating that in the Bay-0 and Sha ecotypes both dormancy breaking treatments resulted in total recovery of germination. All other stress treatments had a negative effect on germination and resulted in negative trait correlations, as compared to after-ripening and stratification. Interestingly, the same structure was observed when studying the LOD profile correlations. This indicates that most of the variation is well captured in the genetic analysis. Neither dry seed size nor flowering time correlated significantly with any of the germination parameters. Imbibed seed size appeared to have a negative correlation with

germination in the presence of ABA and a positive correlation with the rate of germination. This correlation was strongest at LOD-profile levels.

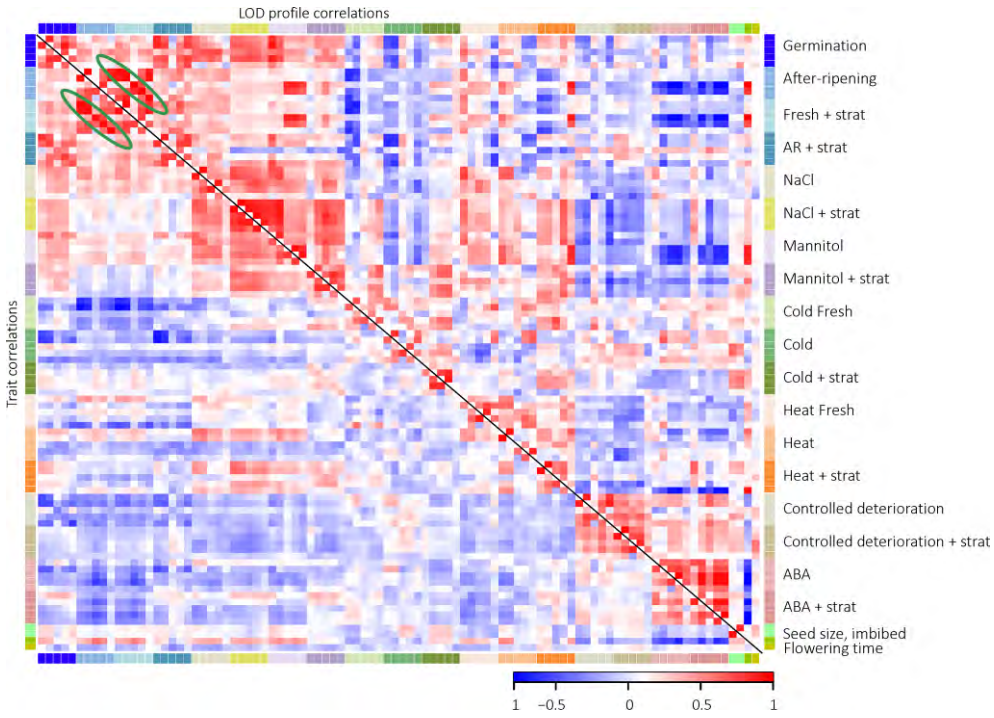


Figure 3.2: Correlation between all trait values (bottom left panel) and between all LOD profiles (top right panel). Five parameters (G_{max}, AUC, t_{50}, t_{10} and U_{8416} resp.) per experiment are shown. A precise description of the trait-values can be found in Table 3.1. Green ellipses indicate an example of the close correlation between 'after-ripened' and 'fresh with stratification' trait values.

Further, the analytical script creates heat maps of all LOD scores. These heat maps can be used to interpret the genetic landscape of the studied traits. A heatmap of all LOD scores clustered by traits (one-way) with Hclust (Murtagh 1985) is created to visualize similarities among different traits (Figure 3.3). According to a procedure described by Breitling *et al.* (2008), the heatmap shows several significant hotspots in the genome. These appear to control different traits and confirm the high level of interconnection among the responses to different environmental stresses and is in agreement with earlier findings of pervasive genetic buffering (Fu *et al.* 2009). They found only a few influential 'hot spot' regions cause major phenotypic variation across a range of environmental conditions whereas the largest fraction of molecular variants is silent at the phenotypic level. Further, a clear separation in the clustering can be observed between dormancy / mannitol / salt / heat / cold QTL compared to ABA and germination after controlled deterioration QTL. This is mainly caused by the large QTL with reversed effect on the bottom of chromosome V.

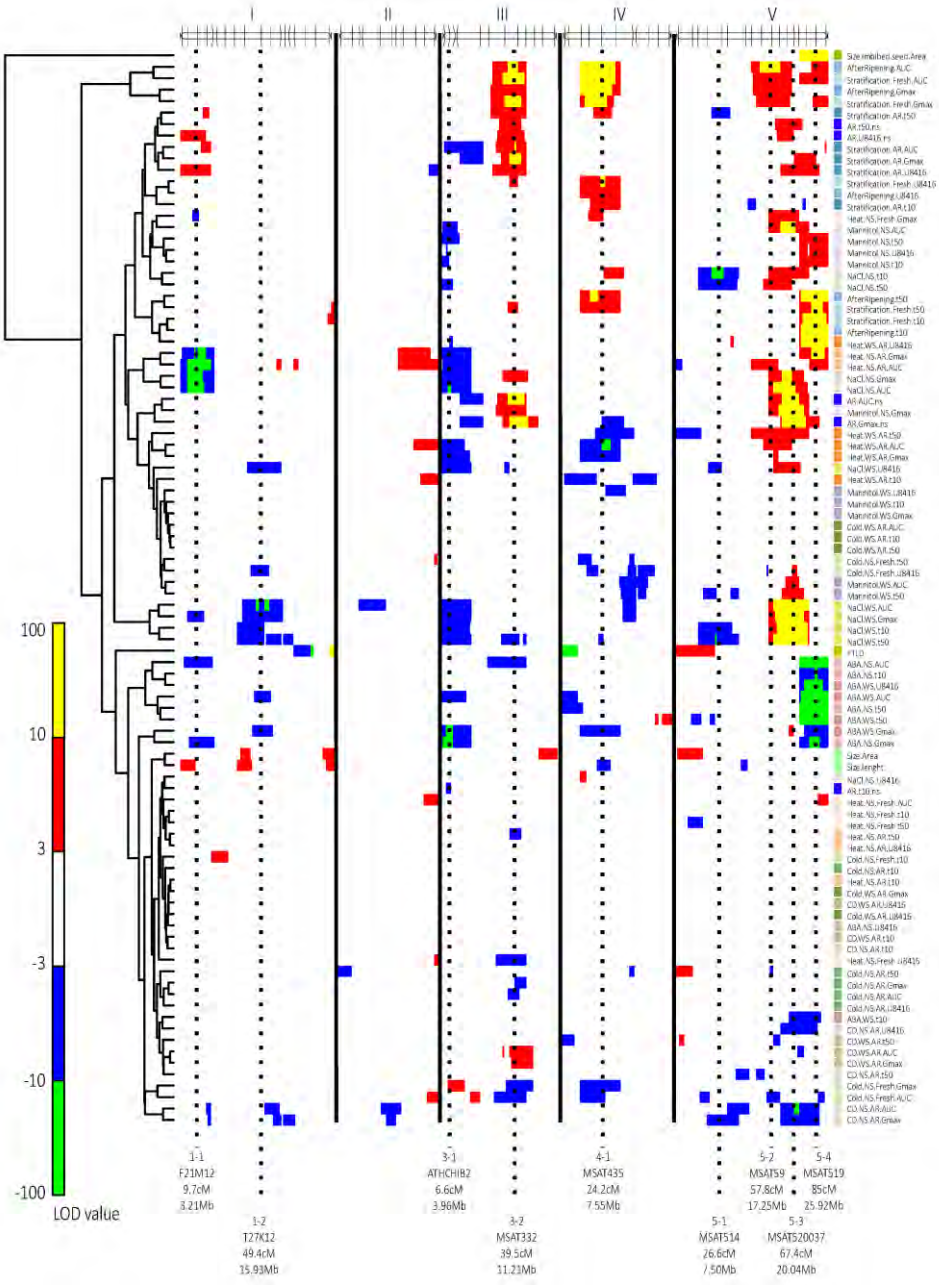


Figure 3.3: A clustered heat map showing the LOD profiles of the measured traits. Columns indicate chromosome position along the 5 chromosomes; rows indicate individual trait LOD profiles. A false color scale is used to indicate the QTL significance. Positive values (yellow and red) represent a larger effect of the treatment in Sha, negative values (blue and green) in Bay-0. Dashed lines indicate major QTL positions which are further discussed in Table 3.4. Clustering on the left shows the correlation between QTL profiles.

The protocol stores all LOD profiles in a single tab-delimited text file, allowing further analysis in most available statistical and/or microarray analysis software suites. A second output file summarizes all QTL results, providing an overview of all detected QTL, their LOD score, position, confidence interval and direction (Supporting Information, S3.4 showing all 327 traits, Supporting Information, S3.5 showing the output for the average traits described here). An adjusted result file in the sif format (simple interaction format) allows direct import in Cytoscape. Using Cytoscape we created a Marker-Trait network of QTL positions, with nodes indicating markers or traits and edges representing LOD scores and directions, allowing an alternative method to visualize many trait QTL in one figure (Figure 3.4A, Supporting Information S3.10). One advantage of loading a QTL network in Cytoscape is the dynamic nature of the program, which allows ordering, filtering and selection in all directions. Figure 3.4B shows an example where a specific marker (MSAT519) was selected to show all traits that have a significant QTL at this locus, visualized by the adjacent edges and connected nodes. In Figure 3.4C we selected a single trait (germination on NaCl with stratification), to show all significant QTL positions for this trait.

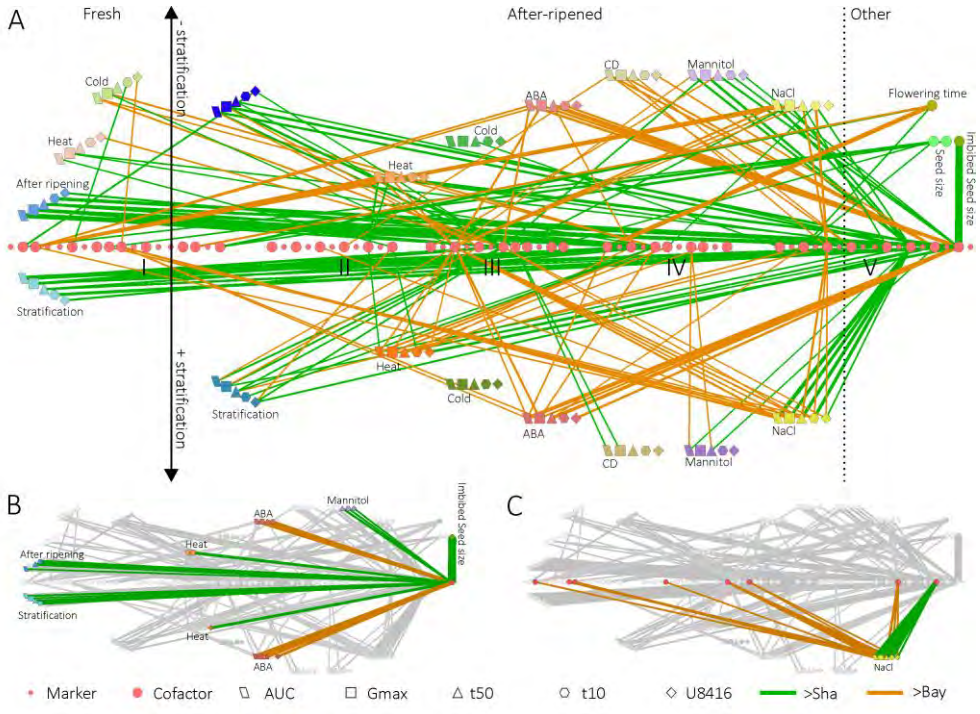


Figure 3.4: Cytoscape Marker-Trait network. A) Significant QTL positions are indicated by a connection between traits and markers, edge colors indicate the direction of the QTL effect, line width indicates the LOD score; B) Sub-network showing all traits with a significant QTL at marker MSAT519; C) Sub-network showing all markers with significant QTL for germination on NaCl with stratification

QTLxQTL Interactions

Epistatic interactions between QTL can help to elucidate meaningful co-localizations and will enable an efficient design of follow up experiments. Besides the visualization of the epistatic interactions per trait (Figure 3.1F) our script creates an output that can help to visualize all detected epistatic interactions in a single plot. This output file in sif format summarizes all detected epistatic interactions (Figure 3.5, Supporting Information S3.11). Among others, clear hotspots of epistatic interactions between QTL loci on chromosome 3, 4 and 5 (resp. ATHCHIB2 + MSAT332, MSAT435 and MSAT520037 + MSAT519) were observed for germination on salt (yellow lines) and dormancy (blue lines). Next to the importance of detecting possible interacting loci this QTLxQTL analysis provides additional arguments for co-locating QTL to be of similar genetic origin. Overall, the creation of this type of summarizing figures is greatly facilitating the interpretation of large datasets.

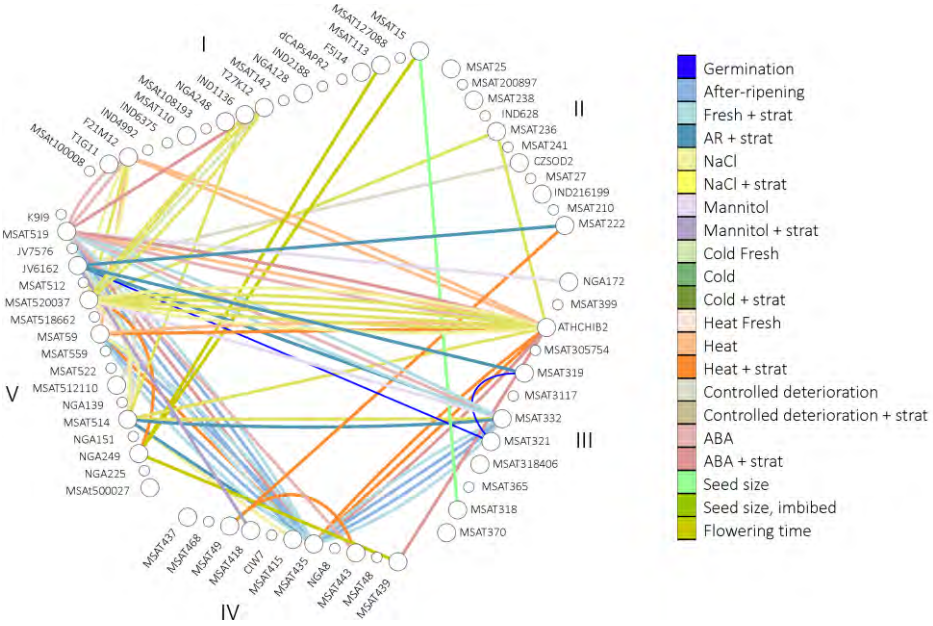


Figure 3.5: Epistatic interaction network. Nodes indicate markers (small circles) and selected cofactors (large circles). Edges represent the detected significant epistatic interactions (edge colors represent traits). A precise description of the trait-values can be found in Table 3.1.

QTLxEnvironment interaction

To obtain a parameter for the response, we had to correct all values with their proper control condition values. This sometimes led to complex interpretation, which can be circumvented by using the non-corrected germination parameters and model them over the various environmental conditions that were tested. Because several environments are taken into account simultaneously the statistical power to detect loci that are affected by several environments increases and interpretation becomes more intuitive as the need for correcting the stress response by the control response is eliminated. By using this approach the sensitivity of a specific QTL for environmental conditions can be determined for each separate germination parameter. Details about the procedure are described in Material and Methods. Results are summarized in Figure 3.6. The final model P-value profiles (top panel, Figure 3.6) clearly show the great consistency between the 5 germination parameters that we measured. However, a closer look also reveals loci that are affecting different germination curve parameters. For example, the QTL on top chromosome 5 is not detected by measuring maximum germination but is well defined when using t_{50} or t_{10} as parameter. As expected, the parameter AUC (Area Under the Curve) is outperforming the others as it represents a combined value for maximum germination percentage, rate and uniformity. For comparison of the environment-specific QTL effects for the 5 different germination parameters (5 lower panels, Figure 3.6) the effects could be compared with germination under control conditions. For example, after-ripened seeds without stratification (AR.NS) can guide as reference for the stress treatments (AR.NS.ABA, AR.NS.CD, AR.NS.Cold, AR.NS.Heat, AR.NS.Mannitol, AR.NS.NaCl). The same analogy holds true for after-ripened seeds without stratification (AR.NS) and freshly harvested seeds without stratification (Fresh.NS). In this way stress specific QTLs on chromosome II and top chromosome III can easily be identified. Interestingly, some QTLs, including germination at low temperature (top chromosome I) and germination in the presence of exogenous ABA (bottom chromosome V) displayed opposite effects on germination when compared to the other treatments. In Table 3.3 the environmental specific effect sizes are summarized for the major loci. A complete overview of effect sizes and explained variances for all detected loci can be found in Supporting Information S3.6.

Figure 3.6: Genome scan for QTLxEnvironment effects for seed germination. The P-values for the main effects of the different germination parameters are shown in the top panel. The red horizontal line is the genome wide significance threshold. The 5 bottom panels show the environment specific QTL effects. The horizontal green bar at the top of each panel indicates significant environment specific effects. For both G_{max} and AUC a bigger effect of the Sha allele is indicated in yellow-red and bigger effect of the Bay-0 allele in cyan-blue. The colors scale is opposite for the t_{50}, t_{10} and U_{8416} parameters due to the inversed nature of these parameters.

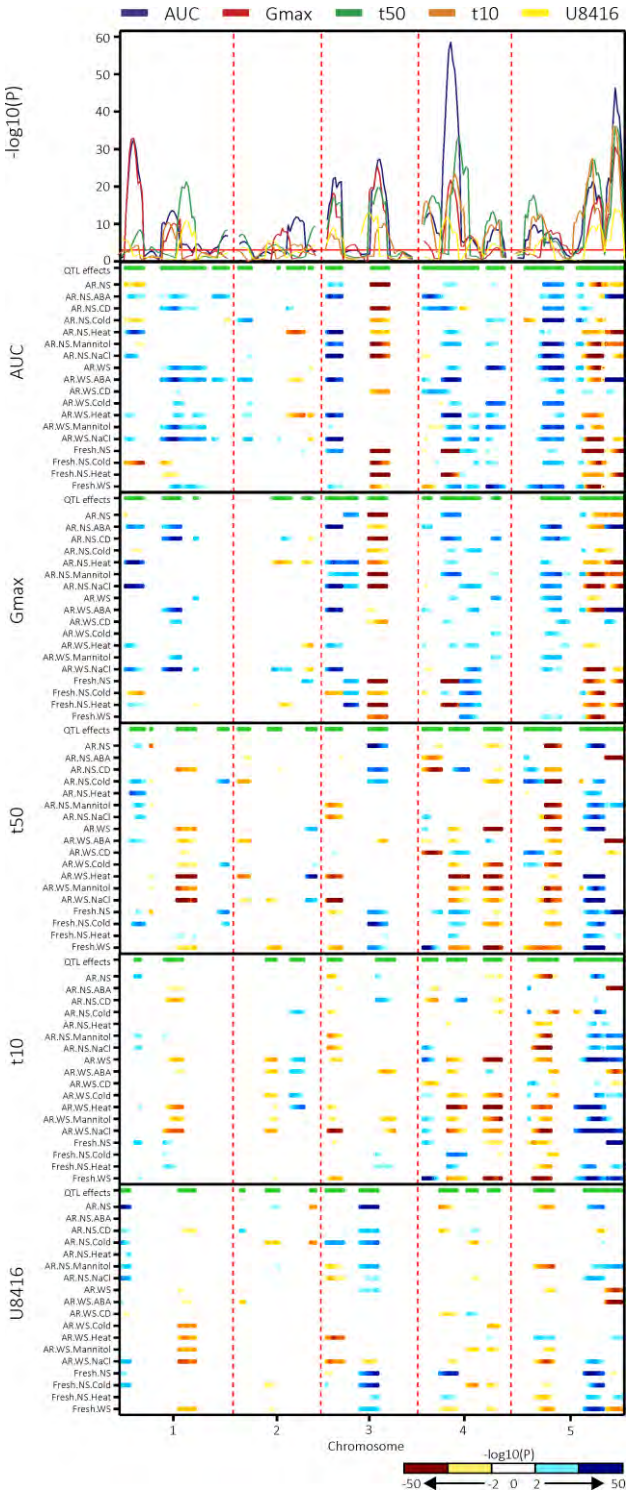

AUC
Gmax
t50
t10
U8416
-log10(P)
60
50
40
30
20
10
0
QTL effects
AR.NS
AR.NS.ABA
AR.NS.CD
AR.NS.Cold
AR.NS.Heat
AR.NS.Mannitol
AR.NS.NaCl
AR.WS
AR.WS.ABA
AR.WS.CD
AR.WS.Cold
AR.WS.Heat
AR.WS.Mannitol
AR.WS.NaCl
Fresh.NS
Fresh.NS.Cold
Fresh.NS.Heat
Fresh.WS
1
2
3
4
5
Chromosome
-log10(P)
-50
-2
0
2
50

Table 3.3: Significant environment-specific QTL effects (P<0.05). Positive values for AUC and G_{max} indicate a larger effect of the Sha allele, negative values for AUC and G_{max} indicate a larger effect of the Bay-0 allele. This is opposite for t_{50}, t_{10} and U_{8416} values due to the inversed nature of these parameters. Environments as mentioned in Table 3.1, loci as indicated in Figure 3.3.

		Environment	AR.NS	AR.NS.ABA	AR.NS.CD	AR.NS.Cold	AR.NS.Heat	AR.NS.Mannitol	AR.NS.NaCl	AR.WS	AR.WS.ABA	AR.WS.CD	AR.WS.Cold	AR.WS.Heat	AR.WS.Mannitol	AR.WS.NaCl	Fresh.NS	Fresh.NS.Cold	Fresh.NS.Heat	Fresh.WS
Locus	cM																			
1-1	9	AUC	1.3	-1.6		0.9	-3.5		-2.1					-1.6		-1.9		2.1		
		Gmax (%)		-5.8			-7.4		-7.8					-2.4		-4.4		3.7	-4.6	
		t50 (h)	-0.8			-1.3	-1.8	-1.9	-1.8		-2.1							-4.2		
		t10 (h)								0.4	-1.9	-1.5				0.9				0.6
		U8416 (h)	-2.0		-1.5	-1.3		-1.9	-2.0		-4.8					-1.7	-2.6	-4.3		
1-2	50	AUC		-1.4	-3.2					-0.8	-4.4				-2.6	-4.6		1.5	2.8	-0.7
		Gmax (%)		-6.1	-6.8						-7.2	-4.5			-2.3	-6.4				
		t50 (h)			2.5					0.6	2.7	1.3	1.1	1.6	1.7	2.7				0.5
		t10 (h)		5.7	3.1					0.5				1.0	1.3	1.8		-2.1	-3.2	
		U8416 (h)			1.4						6.0		1.7	1.2	1.6	2.9				0.6
3-1	5	AUC	-1.3	-2.5			-4.3	-2.6	-4.2		-5.2			-3.5		-4.8	-3.0			-0.8
		Gmax (%)		-9.9			-5.8	-2.9	-6.8		-8.7			-2.9		-5.7		4.5		
		t50 (h)						2.5	2.5					1.1		2.8	1.8			0.6
		t10 (h)	0.8			1.2		1.8	2.0	0.4				0.6	0.8	1.6	1.2			0.5
		U8416 (h)				-1.3		2.2	1.7					1.2		2.4	2.4			
3-2	44	AUC	3.5		5.3	1.3	2.3	3.4	3.9			4.2					7.5	2.8	6.5	0.9
		Gmax (%)	4.6	4.6	9.0	1.4	6.1	11.7	10.4			5.9					15.3	5.1	18.0	0.9
		t50 (h)	-1.8		-3.5	-1.6					2.5						-2.4	-5.3		-0.6
		t10 (h)			-2.1						1.9				0.9	0.7			-3.2	
		U8416 (h)	-2.5		-2.1	-1.4		-2.7	-1.3	-0.6						1.6	-5.9	-7.3	-0.8	-0.6
4-1	24	AUC	-1.2		-2.9		-2.2	-2.0		-0.7	-5.3		-1.2	-4.6	-2.1	-3.1	8.5		5.7	-0.9
		Gmax (%)	-3.1		-7.8	-1.3	-4.2	-6.1		-0.4		-4.8		-2.7			15.8		16.1	
		t50 (h)			-3.5					0.5	2.4	-1.9	1.3	1.6	1.6	1.8	-2.8	-3.5		0.9
		t10 (h)			-2.2					0.5	1.5		1.2	1.4	0.9	1.3		-4.2	-3.6	0.7
		U8416 (h)	1.4		1.9							4.2					-6.7		-1.0	0.6
5-1	14	AUC				1.0														-1.2
		Gmax (%)																		
		t50 (h)			-2.8	-2.2						-2.7	-1.4							0.9
		t10 (h)																		
		U8416 (h)																		
5-2	63	AUC	1.8				2.8	3.1	4.5	1.2	3.8			3.3	2.0	6.5	5.9	2.2	4.2	2.2
		Gmax (%)	1.8	5.2			5.0	6.7	8.6	0.4	8.3			2.6		9.1	7.5	6.5	10.0	1.0
		t50 (h)	-2.0			-1.3	-1.8	-3.3	-3.6	-1.3				-2.2	-2.2	-3.8	-3.0	-4.6	-6.6	-1.8
		t10 (h)								-0.4				-1.1		-1.5			-4.0	-0.6
		U8416 (h)	-1.4					-2.0	-2.1							-2.3	-7.2	-5.5	-1.2	-1.0
5-3	83	AUC	1.6	-4.6			5.2	3.0	1.9		-12.2					2.5	3.7		5.2	
		Gmax (%)	2.8	-12.1			8.6	5.8	5.3		-14.0					3.4	6.2		9.0	
		t50 (h)		23.2				-2.3			10.8	1.8					-4.9			
		t10 (h)		15.7				-1.6	-1.7	-0.7	2.4	1.4				-1.4	-5.2		-4.4	-0.9
		U8416 (h)		14.2				-3.4		1.2	17.8			-0.8			3.1		-1.0	0.8

QTL confirmation

Taking advantage of the residual heterozygosity present in the F6 generation of the Bay-0xSha population, combined with the large population size, we were able to confirm several QTL following the heterogeneous inbred family (HIF) approach. In short, RIL lines which are heterozygous at the locus of interest were selected in the next generation for lines homozygous for both parental alleles. These 'families' are near isogenic lines (NIL) which can be used to confirm the observed allelic effects (Figure 3.7A). We applied this strategy for 7 of the major QTL that we detected in this study and tested the 5 germination

parameters for 11 different conditions. For a single parameter (G_{max}) and a single HIF (line HIF103) the analytical procedure is summarized in Figure 3.7B. Traits that could be confirmed by one or several HIF lines are indicated in Table 3.4. An overview of all HIF results can be found in Supporting Information S3.7.

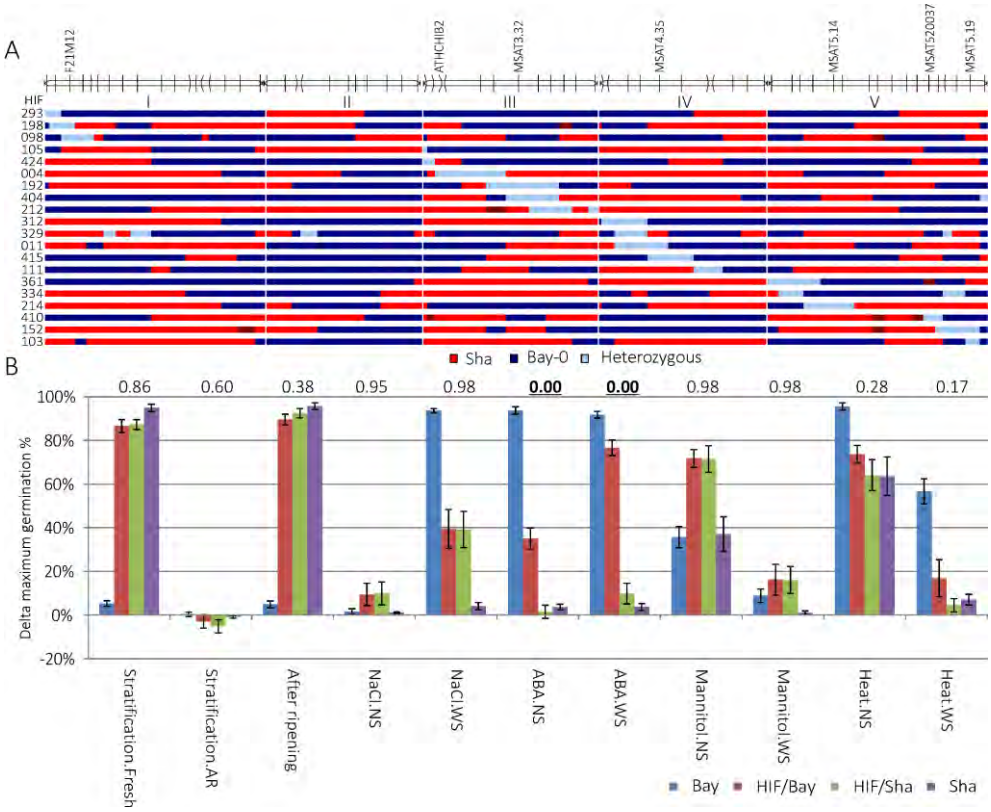


Figure 3.7: Confirmation of QTL with a HIF approach. A) The blue/red bars indicate the allelic distribution of the HIF lines used (blue=Bay-0, red=Sha, light blue=segregating). The 5 chromosomes are indicated at the top with the nearest genetic marker for the 7 major loci. B) Example analysis for HIF103 (segregating at MSAT5.19, bottom chromosome V). Indicated is the response for maximum germination for 11 conditions compared to control. Error bars represent standard error of at least 6 replicates. Responses are calculated by subtracting the test sample from the control sample as indicated in Table 3.1. Numbers above the graph are the t-test significance for the responses as measured between HIF/Bay vs HIF/Sha (significant values (P<0.05) are in bold).

We detected a vast QTL for imbibed seed size at the bottom of chromosome 5, which could be confirmed by the use of HIF103. Upon imbibition seeds swell due to rapid water uptake and possibly because of the expansion of the inner mucilage layer. In Sha, which is a natural mutant for the *mum2* gene (Macquet *et al.* 2007), this swelling did not occur. Also the HIF lines at the *mum2* position showed a clear difference in swelling phenotype which was still significant 24 hours after imbibition (Figure 3.8).

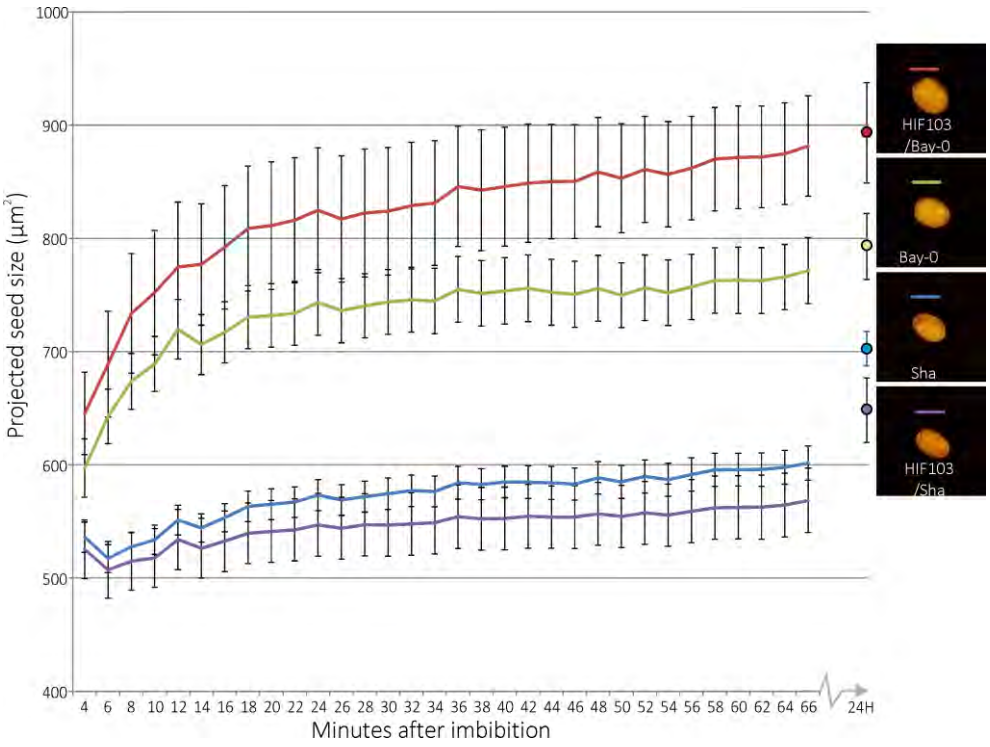


Figure 3.8: Different increase in seed size during start of imbibition for Bay-0 (green), Sha (blue), HIF103/Bay-0 (red) and HIF103/Sha (purple) seeds. Shown is the average projected seed size area of 10 seeds. Error bars represent standard error values. Photographs show 24H imbibed seeds.

Discussion

When analyzing large (RIL) populations, it is hardly feasible to manually count all germination experiments several times a day to obtain germination curves. Therefore, previous studies mostly restricted to counting end-point germination (Quesada *et al.* 2002; Alonso-Blanco *et al.* 2003; Clerkx *et al.* 2004; Laserna *et al.* 2008; Meng *et al.* 2008; Bentsink *et al.* 2010; Galpaz and Reymond 2010; Vallejo *et al.* 2010). A germination curve allows QTL mapping under conditions where rate and uniformity are delayed, but maximum germination is not affected. Therefore, we used the Germinator package (Joosen *et al.* 2010) that enabled measurement of cumulative germination data and extracting 5 germination parameters that describe the resulting germination curve. In the present study we describe several germination QTLs that were not detected before in the Bay-0xSha population. We observed interesting co-localizations for several germination traits and identified the loci that show large effect epistatic interactions. Among these were new loci and loci similar to the ones already found in other RIL populations as summarized in Table 3.4 for the major identified QTL loci.

Table 3.4: Co-locations at the major QTL loci

QTL[1]	Trait	HIF confirmed[2]	Marker	LOD	cM/ Mb	Interval (cM)	Effect	Co-localization with other seed studies
1-1	NaCl.NS.Gmax NaCl.NS.AUC NaCl.WS.Gmax ABA.NS.Gmax ABA.NS.AUC Heat.NS.AR.Gmax Heat.NS.AR.AUC	 098/198 098 098 098 098	F21M12	13.1 14.1 4.6 4.7 6.6 10.1 14.4	9.7/ 3.2	5 - 13 6 - 13 4 - 14 5 - 22 4 - 14 6 - 18 6 - 18	>Bay-0 >Bay-0 >Bay-0 >Bay-0 >Bay-0 >Bay-0 >Bay-0	Cloned QTL for salt and ABA sensitivity during germination and seedling growth in LerxSha, RAS1 (Ren *et al.* 2010) QTL for germination after heat shock in LerxSha (Clerkx et al. 2004)
1-2	NaCl.WS.Gmax NaCl.WS.U8416 NaCl.WS.AUC ABA.WS.Gmax		T27K12	7.9 3.6 9.2 4.9	49.1/ 15.9	42 - 54 38 - 62 41 - 57 44 - 54	>Bay-0 >Bay-0 >Bay-0 >Bay-0	QTL for germination on salt in LerxSha (Galpaz and Reymond 2010)
3-1	Stratification.AR.AUC NaCl.NS.Gmax NaCl.NS.t50 NaCl.NS.AUC NaCl.WS.Gmax NaCl.WS.t50 NaCl.WS.U8416 NaCl.WS.t10 NaCl.WS.AUC Mannitol.NS.AUC ABA.NS.Gmax ABA.WS.Gmax ABA.WS.AUC Cold.NS.Fresh.Gmax Heat.NS.AR.Gmax Heat.NS.AR.AUC Heat.WS.AR.Gmax Heat.WS.AR.U8416 Heat.WS.AR.AUC	 424 004 004 424 424 424/004 105/424/004 424/105 424 424	ATHCHIB2	3.4 6.0 4.4 9.7 5.9 7.9 4.0 4.9 9.6 3.6 10.7 10.8 6.5 3.4 7.6 6.9 4.7 3.4 5.0	6.6/ 3.9	0 - 19 2 - 16 0 - 8 1 - 7.9 1 - 16 0 - 14 0 - 19 0 - 17 1 - 14 0 - 16 1 - 7.9 2 - 7.9 0 - 7.9 0 - 19 3.2 - 8 3.2 - 19 0 - 17 4 - 17 1 – 13	>Bay-0 >Bay-0 >Bay-0 >Bay-0 >Bay-0 >Bay-0 >Bay-0 >Bay-0 >Bay-0 >Bay-0 >Bay-0 >Bay-0 >Bay-0 >Sha >Bay-0 >Bay-0 >Bay-0 >Bay-0 >Bay-0	QTL for germination on ABA and salt in LerxSha (Clerkx et al. 2004) QTL for germination on PEG in Bay-0xSha, OSR2 (Vallejo et al. 2010)
3-2	Stratification.Fresh.Gmax Stratification.Fresh.AUC Stratification.AR.Gmax Stratification.AR.t50 Stratification.AR.U8416 Stratification.AR.AUC After-ripening.Gmax After-ripening.AUC NaCl.WS.t50 Mannitol.NS.Gmax ABA.NS.AUC Cold.NS.Fresh.AUC Heat.NS.Fresh.U8416	192/404 004/192/404 004/192 404 404 004/192/404 404 004 192/404	MSAT332	11.3 12.9 8.0 6.7 8.1 7.8 9.0 10.0 3.8 6.7 8.4 8.0 4.8	39.5/ 11.2	36 - 50 37 - 50 35 - 52 33 - 48 35 - 48 36 - 48 35 - 48 36 - 48 33 - 52 36 - 52 32 - 46 36 - 52 33 - 48	>Sha >Sha >Sha >Sha >Sha >Sha >Sha >Sha >Bay-0 >Sha >Bay-0 >Bay-0 >Bay-0	Delay of Germination, DOG6 (Bentsink et al. 2010) QTL for dormancy and germination on salt in LerxSha (Clerkx et al. 2004) QTL for cold-dark germination in Bay-0 x Sha, CDG-3 (Meng et al. 2008)
4-1	Stratification.Fresh.Gmax Stratification.Fresh.t50 Stratification.Fresh.U8416 Stratification.Fresh.AUC Stratification.AR.t50 Stratification.AR.t10 After-ripening.Gmax After-ripening.t50 After-ripening.U8416 After-ripening.AUC ABA.WS.Gmax ABA.WS.AUC Cold.NS.Fresh.Gmax Cold.NS.Fresh.AUC Heat.NS.Fresh.Gmax Heat.WS.AR.Gmax Heat.WS.AR.AUC	329/415 111 329/415 011 329/011/415/111 011/415/111 011 011	MSAT435	14.6 8.4 9.9 28.8 4.0 5.7 20.4 9.9 8.8 35.8 5.4 4.0 7.4 4.7 3.6 8.1 9.2	24.2/ 7.5	16 - 26 17 - 32 17 - 32 17 - 26 19 - 33 20 - 37 16 - 26 16 - 28 17 - 31 16 - 25 12 - 34 16 - 37 12 - 30 12 - 27 12 - 31 17 - 37 17 - 34	>Sha >Sha >Sha >Sha >Sha >Sha >Sha >Sha >Sha >Sha >Bay-0 >Bay-0 >Bay-0 >Bay-0 >Sha >Bay-0 >Bay-0	Delay of Germination, DOG18 (Bentsink et al. 2010)
5-1	Stratification.AR.t50 NaCl.NS.t50 NaCl.NS.t10 NaCl.WS.t50 NaCl.WS.U8416	 214 361	MSAT514	4.0 8.8 11.3 9.9 3.5	26.6/ 7.5	21 - 37 19 - 29 21 - 30 20 - 30 18 - 30	>Bay-0 >Bay-0 >Bay-0 >Bay-0 >Bay-0	

QTL[1]	Trait	HIF confirmed[2]	Marker	LOD	cM/ Mb	Interval (cM)	Effect	Co-localization with other seed studies
	NaCl.WS.t10			5.4		18 - 30	>Bay-0	
	CD.NS.AR.Gmax			4.7		20 - 39	>Bay-0	
5-2	Stratification.Fresh.Gmax	152	MSAT59	4.6	57.8/	48 - 68	>Sha	Delay of
	Stratification.Fresh.AUC	152		6.0	17.2	48 - 67	>Sha	Germination, DOG1
	After-ripening.AR.Gmax	152		7.7		51 - 67	>Sha	(Bentsink et al.
	After-ripening.AR.AUC	152		11.7		50 - 62	>Sha	2010)
	NaCl.NS.t50			4.3		52 - 70	>Sha	
	Cold.NS.Fresh.U8416			1.3		47 - 61	>Bay-0	QTL for cold-dark
	Heat.NS.AR.AUC	152		5.1		46 - 62	>Sha	germination in Bay-
	Heat.WS.AR.AUC			4.8		53 - 70	>Sha	0 x Sha, CDG-6
								(Meng et al. 2008)
5-3	NaCl.NS.Gmax		MSAT	10.1	67.4/	61 - 72	>Sha	QTL for germination
	NaCl.NS.t10		520037	6.4	20.0	62 - 79	>Sha	on salt in Bay-0 x
	NaCl.NS.AUC			11.6		58 - 72	>Sha	Sha, SSR2 (Vallejo
	NaCl.WS.Gmax	410/152		18.0		63 - 72	>Sha	et al. 2010)
	NaCl.WS.t50	410/152/103		19.6		64 - 72	>Sha	
	NaCl.WS.U8416			5.4		59 - 74	>Sha	QTL for germination
	NaCl.WS.t10			18.4		65 - 74	>Sha	on salt in ShaxCol
	NaCl.WS.AUC			23.2		64 - 73	>Sha	(Galpaz and
	Mannitol.NS.Gmax	152		13.2		63 - 74	>Sha	Reymond 2010)
	Mannitol.NS.AUC	152		10.4		60 - 74	>Sha	
	Mannitol.WS.t50	152		4.0		63 - 77	>Sha	
	Mannitol.WS.AUC	152		3.5		62 - 79	>Sha	
	Heat.NS.Fresh.Gmax			4.8		55 - 73	>Sha	
	Heat.WS.AR.t50	152		8.0		63 - 74	>Sha	
5-4	Stratification.Fresh.Gmax	152	MSAT519	3.9	85/	80 - 91	>Sha	Cloned Mucilage
	Stratification.Fresh.t50			10.4	25.9	79 - 91	>Sha	affecting locus,
	Stratification.Fresh.t10			16.3		79 - 88	>Sha	MUM2 (Macquet et
	Stratification.Fresh.AUC	152		7.3		80 - 91	>Sha	al. 2007)
	After-ripening.t50			14.8		79 - 90	>Sha	
	After-ripening.t10			19.3		79 - 88	>Sha	
	After-ripening.AUC	152		4.7		80 - 91	>Sha	
	Mannitol.NS.t50			9.6		80 - 91	>Sha	
	Mannitol.NS.U8416			5.1		80 - 91	>Sha	
	Mannitol.NS.t10			6.2		79 - 91	>Sha	
	ABA.NS.Gmax	103		9.9		79 - 89	>Bay-0	
	ABA.NS.t50	103		22.8		80 - 89	>Bay-0	
	ABA.NS.t10			9.7		80 - 90	>Bay-0	
	ABA.NS.AUC	103		12.0		73 - 91	>Bay-0	
	ABA.WS.Gmax	410/152/103		8.2		81 - 90	>Bay-0	
	ABA.WS.t50	103		19.5		79 - 88	>Bay-0	
	ABA.WS.U8416	103		11.9		76 - 88	>Bay-0	
	ABA.WS.AUC	410/152/103		22.3		80 - 88	>Bay-0	
	Heat.NS.AR.Gmax	152		11.0		80 - 90	>Sha	
	Heat.NS.AR.AUC	152		9.1		82 - 91	>Sha	
	Heat.WS.AR.U8416	103		13.3		80 - 91	>Sha	
	Size.imbibed.seed.Area	103		72.3		79 - 89	>Sha	

[1] as marked in Figure 3.3, [2] lines indicated in Figure 3.7A, t-test P-value<0.1

Dormancy

Primary dormancy has been studied extensively in various RIL populations (Bentsink *et al.* 2010). These authors quantified primary dormancy with the DSDS50 parameter (days of dry storage to reach 50% germination), which is a good measure for after-ripening related dormancy breaking. Although we only compared the germination characteristics of freshly harvested seeds with those of after-ripened seeds and fresh seeds with and without stratification, we detected large genetic variation. Both dormancy breaking treatments showed strong QTL at positions 3-2, 4-1 and 5-2, co-locating with *DOG6*, *DOG18* and *DOG1*, respectively (Table 3.4). *DOG18* was not detected in a LerxSha population and showed a stronger dormancy in Ler as compared with An-1, Fei-0 and Kas-2 (Bentsink *et al.* 2010). We detected stronger dormancy in Sha as compared to Bay-0 at the

DOG18 locus. This suggests that both Ler and Sha contain an allele of similar strength which is stronger when compared to An-1, Fei-0, Kas-2 and Bay-0.

Remarkably, for both the *DOG6* and *DOG18* location the sensitivity to ABA was higher in Bay-0, whereas dormancy was deeper in Sha, which resulted in a directional change of the QTL effect. The more dormant Sha parent contains higher initial ABA levels (Supporting Information, S3.8) and apparently, after-ripening and stratification reduce the ABA sensitivity to a greater extent as compared to the Bay-0 parent. This effect was not observed for the *DOG1* locus. Further, we identified a strong effect of the dormancy-breaking treatments on the initiation (t_{10}) and rate (t_{50}) of germination at the bottom of chromosome 5 (marker MSAT519, 85 cM). The same was observed for germination on mannitol and germination at higher temperature. A QTL with opposite effect at this position was found for germination on ABA. Interestingly, these co-located with a QTL found for imbibed seed size.

Water uptake

Initiation and rate of germination are highly influenced by the overall water potential of the seed. The mucilage layer surrounding the seed appears to play an important role in the process of water uptake (Penfield *et al.* 2001). Sha is a natural mucilage mutant due to a mutation in the *MUM2* gene, which changes the hydrophilic potential of rhamnogalacturonan I (Macquet *et al.* 2007). Although mucilage has been reported to be dispensable for germination and development under lab conditions (Arsovski *et al.* 2010), a link with germination under reduced water potential conditions was shown by Penfield *et al.* (2001). They showed reduced maximum germination of a mucilage-impaired mutant only on osmotic PEG solutions. In our study, other traits that co-located on the *MUM2* locus were delayed initiation and rate of germination on osmotic mannitol solution but also on water, which clearly shows the advantage of determining a detailed germination curve. We also observed a very strong QTL for swelling of the seed in the first hours of imbibition (imbibed seed size) at the *MUM2* location. Interestingly, exogenous ABA can be used to stimulate mucilage production and *aba1* mutants are affected in mucilage production (Karssen *et al.* 1983). This indicates a regulatory role of ABA in mucilage production and fits with our observation of the co-localization of a QTL for initiation and rate of germination with a QTL with opposite effect for ABA sensitivity. Therefore, we hypothesize that Sha has a slower initiation and rate of germination, combined with reduced ABA sensitivity due to its mutation in the *MUM2* gene. This observation may open new research strategies to define the regulatory role of ABA in mucilage production and its multiple effects on germination parameters.

Germination responses to salt, heat and ABA

At the top of chromosome I, underlying marker F12M12, we detected a strong QTL for maximum germination in the presence of 100 mM NaCl or 0.5µm ABA. A similar

locus has been identified and fine-mapped in a LerxSha population (Ren *et al.* 2010). They identified a premature stop codon in the *Response to ABA and Salt 1* gene (*RAS1*; At1g09950) in Sha that led to a truncated protein and showed its role as a negative regulator of salt tolerance during seed germination and early seedling growth by enhancing ABA sensitivity. Here we show that a similar locus is also inferring tolerance to germination at 30°C. This suggests an additional role for the *RAS1* gene. Increased heat tolerance due to modulation of ABA sensitivity has been shown before for other loci, (Argyris *et al.* 2008; Lee *et al.* 2010). Interestingly, our present study showed a strong effect of stratification which resulted in a strong reduction of significant linkage for NaCl, heat and ABA sensitivity at the F12M12 locus. A specific QTL for germination on NaCl preceded by a cold stratification period was found at the middle of chromosome I (marker T27K12). Also at this locus we found colocation with sensitivity for germination on ABA after stratification. Further fine-mapping at this locus might help to elucidate the effect of stratification on ABA mediated abiotic stress tolerance, as well as the apparent overlap of dormancy and stress responses.

Especially interesting is QTL 5-1 (Table 3.4, Figure 3.3) which mainly influences rate and initiation of germination. We detected this QTL for t_{50} in after-ripened seeds with stratification treatment, but also for t_{10} and t_{50} for germination on salt, regardless of a preceding cold stratification and for maximum germination after an accelerated aging treatment. One of the genes underlying this QTL interval is a nicotinamidase gene (*NIC2*, At5g23230), the mutant of which has retarded germination and impaired germination potential (Hunt *et al.* 2007). These authors suggested that NIC2 is normally metabolizing nicotinamide during moist chilling or after-ripening, which relieves inhibition of poly(ADP-ribose) polymerase (PARP enzyme) activity and allows DNA repair to occur prior to germination. Both accelerated aging and germination under salt stress conditions might require optimal functioning of this DNA repair mechanism. Further research is needed to determine whether *NIC2* is causal for this QTL.

Detection of epistatic interactions in genetic studies can enhance the understanding of underlying molecular mechanisms. Recently, Galpaz and Reymond (2010) showed strong epistasis in the genetic network controlling germination under salt stress in Arabidopsis. Due to careful dissection of the epistatic relationships they were able to show that three detected QTL rely on the presence of a Columbia allele at a QTL on top of chromosome I. This observation led to the hypothesis that RAS1 (Ren *et al.* 2010) functions as a switch of the genetic network by regulating the expression of the other QTL. In another study it was found that epistasis significantly influences both fitness and germination in *Arabidopsis* (Huang *et al.* 2010) and novel allele combinations were identified that resulted in higher fitness. In our study we detected clear hotspots of epistatic interactions between QTL loci on chromosome 3, 4 and 5 (ATHCHIB2, MSAT332, MSAT435, MSAT520037 + MSAT519, respectively). This observation strengthens the hypothesis that some of the traits with strong QTL co-localizations indeed rely on the same underlying genetic networks.

Conclusion

In conclusion, we analyzed natural variation for many seed germination characteristics and showed their correlation, (shared) QTL positions and epistatic interactions, using a high-throughput phenotyping approach and subsequent high-throughput QTL mapping. Using the HIF approach, confirmation of some major QTL hotspots was demonstrated, which allows a fast but solid confirmation of a QTL position. Together with results from several other studies focusing on genetic variation in seed traits, this study has generated an extensive QTL database for *Arabidopis* and proposed a method of analysis to visualize the genetic landscape of seed performance. This database is a solid resource for further study. For most of the found loci in this and other studies further characterization, and in most cases fine mapping, must be undertaken to elucidate the causal molecular mechanisms. Further, we have designed a free available analysis protocol to perform detailed high-throughput QTL analysis based on the R/qtl MQM routine. In this era of large-scale phenotyping we regard a detailed analysis of QTL, QTLxQTL and QTLxEnvironment interactions as indispensable steps to allow visualization and interpretation of multiple traits. Finally, there is great potential in combining extensive phenotyping of RIL populations with available -omics approaches to increase the speed of causal allele detection (Joosen *et al.* 2009).

Materials and Methods

Plant Growth

Seeds from the core population (165 lines) of an Arabidopsis Bay-0 x Sha recombinant inbred population (Loudet, et al., 2002) were obtained from the Versailles Biological Resource Centre for Arabidopsis (http://dbsgap.versailles.inra.fr/vnat/). The population is mapped with 69 markers with an average distance between the markers of 6.1 cM (Loudet *et al.* 2002). Maternal plants were grown twice in a fully randomized setup. In the first round we separated the harvest in 3 groups (A,B and C), each containing 3-5 plants/RIL. In the second round we pooled the harvest of 4-7 plants/RIL (D). Plants were grown on 4x4 cm rockwool plugs (MM40/40, Grodan B.V.) and watered with 1 g/l Hyponex fertilizer (NPK=7:6:19, http://www.hyponex.co.jp) in a climate chamber (20°C day, 18°C night) with 16 hours of light (35W/m2) at a relative humidity of 70%. Seeds were bulk harvested and after-ripened at room temperature and ambient relative humidity until they reached their maximum germination potential after 5 d of imbibition.

Seed Germination Assays

Germination experiments were performed as described previously (Joosen *et al.* 2010). Germination was scored using the Germinator package. When mentioned, a cold stratification period of 4 days at 4°C in the dark was applied before transferring the trays to

the germination incubator (20°C, continuous light). A temperature of 10°C was used for testing 'cold'-germination whereas 30°C was use for 'heat'-germination. Salt stress was applied by replacing the water by an NaCl (Sigma Aldrich, #S-3014) solution (100 mM for non-stratified, 125 mM for stratified seeds, as stratification reduces the sensitivity to NaCl). A solution of -0.6 MPa mannitol (Sigma Aldrich, #15719) was used to test for osmotic stress. ABA (Duchefa Biochemie, A0941) was initially dissolved in a few drops of 1N NaOH from which stock solutions were prepared in 10 mM MES buffer, pH 5.9. ABA was used at a final concentration of 0.5 µM. Accelerated aging of the seeds was performed using a closed container with a saturated NaCl solution to obtain 75% relative humidity (RH). Seeds were equilibrated for 7 days in this humidity at 20°C in the dark, followed by 30 days at 75% RH, 32.5°C in the dark (Hundertmark *et al.* 2011). For each measurement we used at least 2 replicates for every harvest that was tested (Table 3.1) to determine the germination characteristics. All germination tests were performed in a fully randomized setup. Averages were calculated and corrected for their proper control (Table 3.1). For G_{max} and AUC the stress condition was subtracted from the control condition. Because t_{10}, t_{50} and U_{8416} are reversed parameters, we subtracted control conditions from stress conditions. Dry seed size was determined by taking close-up photographs from ~100-200 seeds using a Nikon D80 camera with a 50mm Macro objective. Imbibed seedsize was extracted from the first images acquired within the Germinator setup (100-200 seeds). For Figure 3.8, 10 seeds of each line were photographed with maximum magnification (using a 50mm Macro objective). The photographs were analyzed using the open source image analysis suite ImageJ (http://rsbweb.nih.gov/ij/) by using color-thresholds combined with particle analysis.

Single trait QTL mapping using R/qtl

This protocol is provided as an R script and performs the following analysis and transformations (steps specified with an (o) are optional, (c) means that the steps can be configured by the user):

Preparing and starting:

To run an analysis two kind of files need to be provided. (1) the RAW data file formatted in R/qtl cross format (See the manual of R/qtl for more information) and (2) a configuration file. The description of the allowed (and necessary) parameters is available in the manual (Supporting Information, S3.12). After having prepared these files the user can start R and change to the directory the script is located by using the setwd command:

```
> setwd('d:/script')
```

Now the script can be loaded by using the source command:

```
> source('qtl_analysis_script.R')
```

Then the analysis is started by the doAnalysis command, this command takes two parameters a filename and a directory (only needed when different from the script directory):

> doAnalysis('myconfig.txt','d:/configfiles')

The script will now start the analysis.

Analysis protocol – loading and preprocessing:

The script starts by reading the configuration and data file and then performs the following analysis:

(o,c) Use a Z-transformation to check data distribution and remove any outliers;

(o,c) Automated phenotype normalization / User supplied normalization;

(c) Plotting of basic genetic statistics like genetic map, recombination frequencies, trait correlation plots.

Main analysis loop (performed for all traits)

In the main loop we analyze traits independently. For each trait a directory in the output folder is created to store the per trait results. We used the following steps:

Plotting of basic per phenotype statistics: Distributions of raw and normalized data;

(c) QTL modeling using backward elimination on genetic cofactors, followed by interval mapping using multiple cofactors (multiple QTL mapping;);

(o) QTL by single marker mapping using Hailey-Knott regression (scan.one);

(o) Whole genome interaction scan heatmap made by using the scan.two QTL interaction mapping routine from R/qtl;

(o) Single trait permutation using the mqmpermutation routine;

Generation of various plots related to the single trait QTL results:

Raw phenotype effect plots;

QTL model by backward elimination;

QTL profiles showing scan.one, scan.two and the MQM interval mapping results;

Interaction effect plots.

Furthermore at the end of the analysis for each phenotype additional information is saved. Also an Rdata file containing the output object from the MQMscan is saved to enable the user to cancel the analysis and resume at a later time.

Multi trait analysis and plots

After all phenotypes have been analyzed the script provides additional plots based on the aggregated data:

(o,c) Circleplots showing selected cofactors and possible interactions;

(o,c) Combined heatmaps and HClust clustering of all QTL profiles;

(o,c) Extraction of the clusters and plotting of the grouped QTL results;

(o) The user can output sif (simple interaction format) formatted files which can be used to create network overviews in Cytoscape (or other visualization tools). We provide two networks:

- The QTL network: Genetic marker network based on QTL data
- The epistatic interaction network produced by summarizing the interactions found between selected cofactors from the MQM algorithm.

To visualize the created .sif files download and install Cytoscape, launch Cytoscape and load the network using File | Import | Network (Multiple file types).

QTL x Environment analysis

By using Genstat version 14, the variance covariance (VCOV) model is calculated for the G+GxE variation in the phenotypic data based on an unstructured model. Given this VCOV model a simple interval mapping (SIM) procedure was started followed by two rounds of composite interval mapping (CIM) using detected QTLs as cofactors, but omitting these covariables in windows around the SIM QTLs. A final multi-QTL model is created using a backward elimination for significant cofactors. For imputing virtual markers along the chromosomes a step size of 2 cM was used. Minimum cofactor proximity as well as minimum separation for selected QTL were set to 16 cM.

Acknowledgements

We would like to thank Suzanne Abrams and Irina Zaharia from the Plant Hormone Profiling Lab at NRC PBI for carrying out hormone measurements (http://www.nrc-cnrc.gc.ca/eng/facilities/pbi/plant-hormone.html). We would like to thank Martin Boer and Linus van der Plas for their useful comments and critical reading of the document.

Supporting Information

Supporting information can be downloaded from either the online version of this article (Joosen *et al.* 2012) or from:
www.wageningenseedlab.nl/thesis/rvljoosen/SI/chapter3

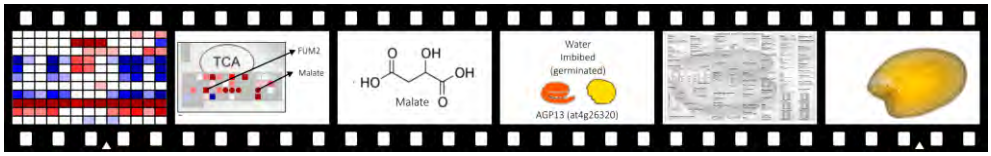
Table S3.1: Pearson correlation matrix for Area Under the Curve (AUC) trait values measured in 4 different harvests, Microsoft Excel (xls) format
Table S3.2: Crossobject containing all data for the 327 measured traits (in columns) in the 165 RIL lines (in rows) which can be used as input file for the script, comma separated value (csv) format
Table S3.3: Crossobject containing data for the average values of the 94 traits (in columns) in the 165 RIL lines (in rows) that are described in this paper. The crossobject can be used as input file for the script, comma separated value (csv) format
Table S3.4: output file that summarizes all QTL results, providing an overview of all detected QTL, their peak LOD score, position, confidence interval and direction for all 327 traits measured, Microsoft Excel (xls) format
Table S3.5: output file that summarizes all QTL results, providing an overview of all detected QTL, their peak LOD score, position, confidence interval and direction for all 94 traits described in this paper, Microsoft Excel (xls) format
Table S3.6: Genstat version 14 output summary for effect size and explained variance for QTLxEnvironment effects, Microsoft Excel (xls) format
Table S3.7: Overview of t-test P-values for all contrasts from the HIF results
Table S3.8: Results and methodology from ABA measurements in dry Bay-0 and Shahdara seeds
Figure S3.9: LOD score correlation plot comparing raw and transformed data (fitted to a normal distribution)
File S3.10: QTLnetwork.cys. Cytoscape file containing the interactive QTL network shown in Figure 3.4
File S3.11: Interactionnetwork.cys. Cytoscape file containing the epistatic interaction network shown in Figure 3.5
File S3.12: Package.zip. Zip file containing all necessary files for re-analysis of the presented data, including a manual with installation and analysis instructions

“With a little seed of imagination you can grow a field of hope.”

4 Visualization of molecular processes associated with seed dormancy and germination using MapMan

Joosen RVL, Ligterink W, Dekkers BJW, Hilhorst HWM (2011)

Seed Science Research. Vol. 21: 143-152.



Abstract

Seed dormancy and germination involve the concerted operation of molecular and biochemical programs. It has become feasible to study these processes in great detail, using the current methods for transcriptome, proteome and metabolome analysis. Yet, the large amounts of data generated by these methods are often dazzling and demand efficient tools for data visualization. We have used the freely available PageMan/MapMan package (http://MapMan.gabipd.org) to visualize transcriptome and metabolome changes in *Arabidopsis thaliana* seeds during dormancy and germination. Using this package we developed two seed-specific MapMan pathways, which efficiently capture the most important molecular processes in seeds. The results demonstrated the usefulness of the PageMan/MapMan package for seed research.

Introduction

The Arabidopsis community has developed comprehensive databases for gene description, annotation and expression analysis (Brazma *et al.* 2003; Toufighi *et al.* 2005; Zimmermann *et al.* 2005). The available information is not limited to transcriptome but is expanded to proteome and metabolome data as well (De Vos *et al.* 2007; Meyer *et al.* 2007; Baerenfaller *et al.* 2008). Arabidopsis is an important model also for seed science and provides valuable insight into the processes underlying germination, dormancy and stress resistance ((Finkelstein *et al.* 2008; Holdsworth *et al.* 2008a; Holdsworth *et al.* 2008b).

The seed represents a critical stage in the plant life cycle. After fertilization the embryo is formed, which is surrounded by the endosperm and seed coat (Liu *et al.* 2005). Seeds acquire desiccation tolerance and dormancy during maturation to survive under harsh conditions after seed dispersal (Bewley 1997). Germination at an appropriate timing is a critical step for the initiation of the plant life cycle (Huang *et al.* 2010; Moyers and Kane 2010). Seeds are equipped with accurate sensors for water, light and temperature to monitor optimal seasonal timings for germination, successful seedling establishment and further plant development (Franklin and Quail 2010). Upon imbibition protein synthesis and DNA transcription are resumed and cell wall expansion and degradation facilitate the penetration of the radicle through the endosperm and seed coat (Nonogaki *et al.* 2007). Finally, energy sources are remobilized to enable a fast growth of the emerging seedling (Nonogaki 2006). The concerted operation of these molecular processes is organized by plant hormones, hormone- or photo-receptors and transcription factors (Holdsworth *et al.* 2008). Modern 'omics' tools can provide valuable insight into the function and regulation mechanisms of these molecular processes (Joosen *et al.* 2009).

Many tools to analyze transcriptome, proteome or metabolome data rely on approaches to detect co-expression or co-existence. Such clustering methods and principal component analysis are efficient tools to summarize data and detect groups of genes, proteins and metabolites with similar behavior (Rensink and Hazen 2006). However, more insights into multiple biological processes can be captured by organizing annotations in such way that profiling datasets are integrated with pre-existing biological knowledge (Zhou and Su 2007). A good example of this type of approach is provided in Taggit in which the creation of seed-specific annotations can be combined with filtered gene-expression datasets (Carrera *et al.* 2007). Taggit provides pie diagrams visualizing relative proportions of functional categories affected by the treatments or developmental stages of interest. A more comprehensive tool that uses a similar approach is called MapMan (Thimm *et al.* 2004). This tool allows users to display genomic datasets onto pictorial diagrams. The diagrams can be fully customized to depict the biological processes of interest. One of the most critical points in using pre-existing knowledge is the quality of the annotation of genes, proteins and metabolites in terms of functional classes. The MapMan tool uses information from the TIGR database (http://compbio.dfci.harvard.edu/tgi/) and input from a number of experts to curate specific biological processes. It has been employed in

different studies and different plant species, such as barley grain maturation and germination (Sreenivasulu *et al.* 2008) and diurnal changes in Arabidopsis (Blasing *et al.* 2005).

In this article we describe the development of two new diagrams that can be used in MapMan and that are focused on biological processes important for seed dormancy and germination. By using Pageman (Usadel *et al.* 2006), a tool combined with the MapMan package, we defined the most informative functional categories. We combined these categories in the first diagram which summarizes transcript and/or metabolite level changes in the pathways important for seed germination. The second diagram provides a focused view of cell wall modification and degradation that are key processes for the completion of seed germination. This comprehensive approach, using the MapMan tools offers the seed science community an easy way to analyze and visualize transcriptome and metabolome data for Arabidopsis.

Methods

We used publicly available data sources that describe seed dormancy and germination. To study the dormancy transcriptome we used data from Finch-Savage *et al.* (2007) and Cadman *et al.* (2006). They compared gene expression in dormant seeds with that in non-dormant seeds under a variety of conditions. Transcriptome changes during seed germination are accurately profiled by data sets from Nakabayashi *et al.* (2005) and polar metabolite changes by data from Fait *et al.* (2006). Penfield *et al.* (2006) dissected Arabidopsis seeds into the embryo and endosperm shortly after radicle protrusion to analyze gene expression. Their transcriptome data sets were also used. In total, we gathered data of 20 seed-specific transcriptome analyses (Table 4.1).

All microarray data was normalized using MAS 5.0 and raw expression values were filtered to display expression above a background value of 50 in four or more experiments. The initial screening filter yielded 11,443 seed-expressed genes (Supporting Information, S4.1). The Pageman tool v0.12 (http://MapMan.gabipd.org; Usadel et al., 2006) was used to identify functional categories with significant enrichment or depletion of up-regulated genes. Within the PageMan package we made use of a Wilcoxon test combined with Benjamin-Hochberg filtering to calculate P-values for enriched categories. The obtained P-values were transformed to z-scores and plotted as heat map (Figure 4.1). Only significant functional categories are shown in the figure.

Table 4.1: Transcriptome and metabolome data sets used in this study

Abbreviation	Description	Ecotype	Replicates	Reference
PD24H	Primary dormant seeds imbibed for 24 h in the dark	Cvi	3	(Finch-Savage et al. 2007)
PD48H	Primary dormant seeds imbibed for 48 h in the dark	Cvi	3	(Cadman et al. 2006)
PD30D	Primary dormant seeds imbibed for 30 days in the dark	Cvi	3	(Cadman et al. 2006)
SD1	Secondary dormant DL seeds imbibed in the dark for 24 days	Cvi	3	(Cadman et al. 2006)
SD2	Secondary dormant SD1 seeds imbibed at 3°C in the dark for 20 days	Cvi	3	(Cadman et al. 2006)
LIG	Dry after-ripened seeds imbibed for 20 h in the dark and then 4 h in red light	Cvi	3	(Finch-Savage et al. 2007)
PDD	Primary dormant dry seeds	Cvi	3	(Finch-Savage et al. 2007)
NDD	Non dormant dry seeds	Cvi	3	(Finch-Savage et al. 2007)
PDL	Primary dormant seeds imbibed for 24 h in the light	Cvi	3	(Finch-Savage et al. 2007)
PDN	Primary dormant imbibed for 24 h on a 10 mM KNO3 solution	Cvi	3	(Finch-Savage et al. 2007)
PDLN	Primary dormant seeds imbibed in white light for 24 h on a 10 mM KNO3 solution	Cvi	3	(Finch-Savage et al. 2007)
PDC	Primary dormant seeds imbibed for 4 days at 3°C	Cvi	3	(Finch-Savage et al. 2007)
Dry seed	Dry after-ripened seeds	Col-0	2	(Nakabayashi et al. 2005)
1H IMB	After-ripened seeds imbibed for 1H under continues white light	Col-0	2	(Nakabayashi et al. 2005)
3H IMB	After-ripened seeds imbibed for 3H under continues white light	Col-0	2	(Nakabayashi et al. 2005)
6H IMB	After-ripened seeds imbibed for 6H under continues white light	Col-0	2	(Nakabayashi et al. 2005)
12H IMB	After-ripened seeds imbibed for 12H under continues white light	Col-0	2	(Nakabayashi et al. 2005)
24H IMB	After-ripened seeds imbibed for 24H under continues white light	Col-0	2	(Nakabayashi et al. 2005)
Endosperm	Isolated endosperms from stratified and germinated seeds	Ler-0	3	(Penfield et al. 2006)
Embryo	Isolated embryos from stratified and germinated seeds	Ler-0	3	(Penfield et al. 2006)
D/G*	Ratios from dry versus stratified, 24H imbibed seeds	Ws	3	(Fait et al. 2006)

* Polar metabolite profiling with GC-MS. All other samples consist of expression profiling using the Affymetrix AtH1 microarray

Because we intended to select the categories with a general role only in dormancy and germination, we excluded mutant transcriptome datasets from the Pageman analysis. To create a detailed view of the enriched functional categories that we identified with Pageman, we made two custom pathway images (using CorelDRAW graphics suite X4, www.corel.com) which can be used in the MapMan tool v3.5.0 (MapMan.gabipd.org). First, we created an 'Arabidopsis seed - Molecular Networks' diagram including all enriched functional categories (Figure 4.2-Figure 4.4). The diagram of hormonal regulation was adopted from Finkelstein et al. (2002) and was simplified to depict hormone signaling. More detailed information about hormone signaling can be found in Kucera et al. (2005) and Holdsworth et al. (2008). Two functional categories describing genes that were linked to dormancy or germination were added to the mapping file (data derived from Taggit ontology, Carrera et al. 2007). Secondly, we created an 'Arabidopsis seed - Cell wall Networks' diagram that allows a focused view of cell wall changes (synthesis, modification, degradation and proteins). For this second diagram some subdivisions were made within the original 'Cell wall' bin. The 'Cell wall' bin 10.5.1, 'Cell wall Proteins AGP (arabinogalactan proteins)' was further divided to 'AGPs', 'FLA (fasciclinlike arabinogalactan proteins)' and 'AGP Other' and 'Cell wall' bin 10.7, 'Cell wall Modification' was subdivided to 'Expansin A', 'Expansin B' and 'Xyloglucan' (Figure 4.5). All these files are freely available at http://mapman.gabipd.org/. Both transcript and metabolite levels can be visualized with this user friendly package. All individual genes within a functional category are represented as a square box and their expression levels are shown in a color (blue-red) scale. Metabolites are represented as colored circles (Figure 4.2B). Users can load raw expression levels as well as expression ratios. We calculated expression ratios by dividing Log2 expression values and subtracting -1 for scaling around 0 ('Log2-1'). A two-tailed paired t-test was used to calculate P-values for all expression ratios (Supporting Information, S4.2). AGI codes or metabolite names are used to match the data with a mapping file that contains the functional categorization of genes and metabolites.

Here we show the power of efficient data visualization of changes in transcriptome and/or metabolome using four examples:

- dry seeds vs imbibed seeds resulting in germination (dry vs. 24h)
- dormant imbibed versus non-dormant germinating seeds (PD24 vs. LIG)
- 24h imbibed, stored Ler vs 24h imbibed, stored cts-1 seeds
- embryo versus endosperm tissue

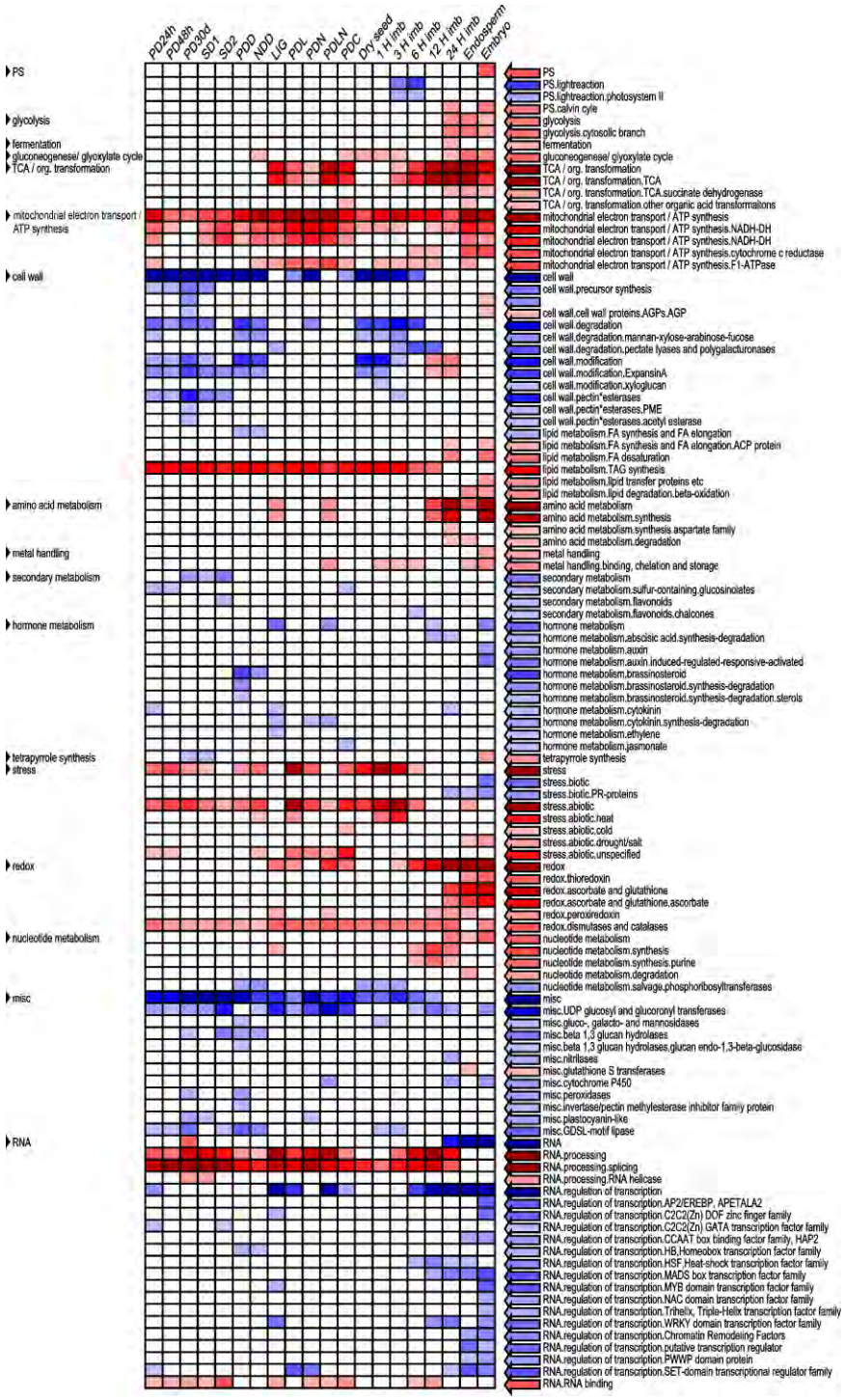


Figure continued on next page

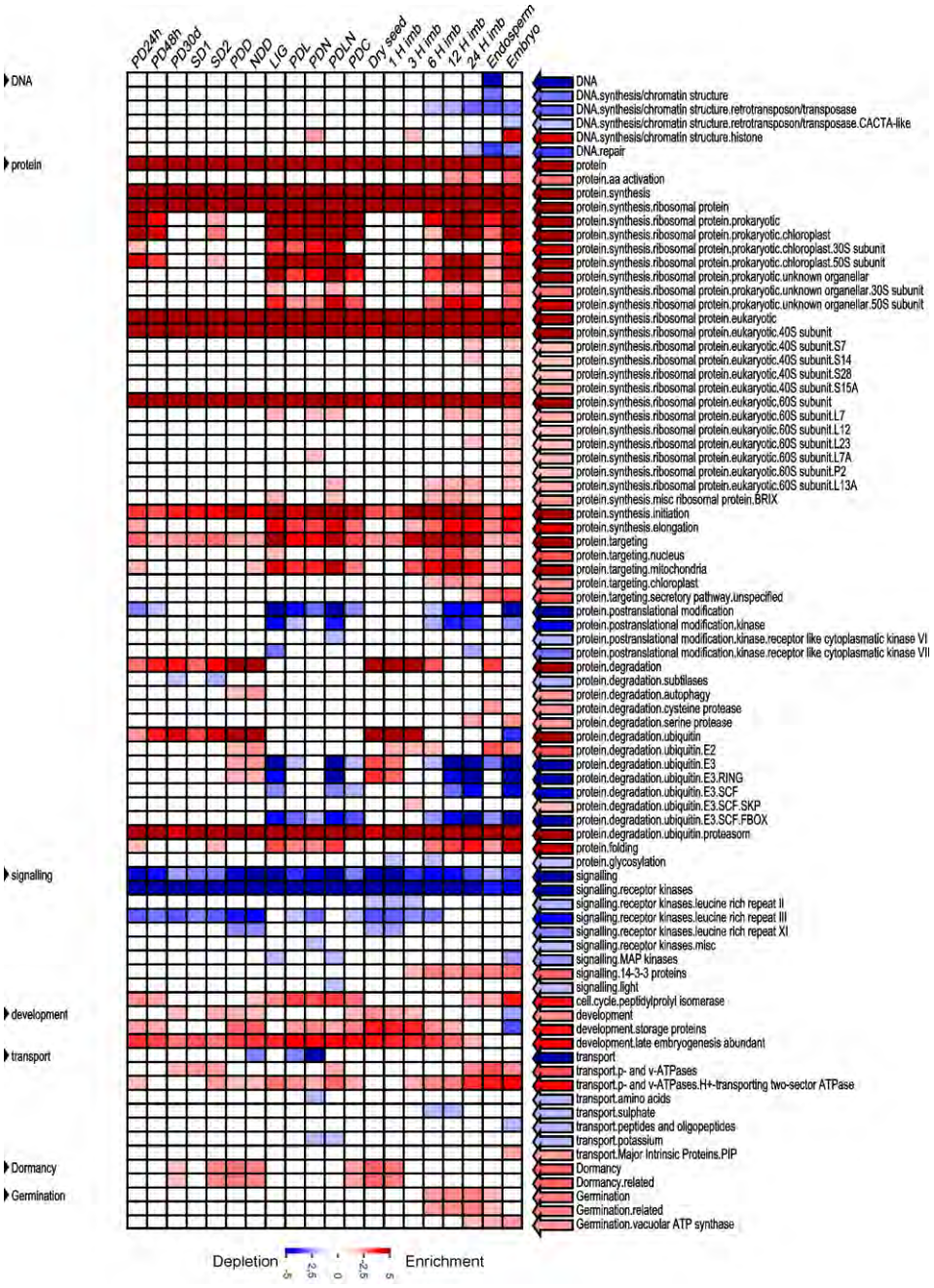


Figure 4.1: Pageman display of coordinated changes of gene function categories during Arabidopsis seed dormancy and germination. The Affymetrix ATH1 normalized gene expression data were subjected to an analysis to identify overrepresented functional categories using Pageman. Red color indicates significant enrichment of up-regulated genes, blue indicates significant depletion of up-regulated genes. Only significant gene function categories are shown. See Table 4.1 for detailed description of sample abbreviations.

Results and Discussion

To examine the efficiency of MapMan data visualization, expression ratios were calculated for dry- versus 24-h-imbibed seeds (Figure 4.2A). In the diagram global transcriptome changes are obvious at a first glance. For example, strong up-regulation of genes related to amino acid biosynthesis ('Amino acid'), energy metabolism ('Energy') and cell wall modification ('Cell wall') in 24-h–imbibed seeds were visualized (red squares). In contrast, transcripts related to late embryogenesis abundant (LEA) and seed storage proteins (Seed storage proteins) rapidly decline (blue squares). Also a decline in stress-related transcripts was observed. Most likely these transcripts accumulated at the end of seed maturation and dehydration and were rapidly lost upon imbibition. In this figure we combined both transcript and metabolite level ratios for dry versus germinating seeds, which allowed us to analyze changes at the metabolite levels in relation to transcriptional changes. For example, several enzymes in the TCA cycle in the 'Energy' category were up-regulated in 24-h-imbibed seeds (Figure 4.2A), which is consistent with higher levels of TCA intermediates, such as citrate, iso-citrate, 2-oxoglutarate and malate, known to occur in imbibed Arabidopsis seeds (Fait *et al.* 2006). In Figure 4.2B, we depicted an example; the concomitant accumulation of malate and a transcript (FUM2) encoding a fumarase that catalyzes the conversion of fumarate to malate. This particular example should be interpreted with some caution since transcript levels of Arabidopsis Columbia (Col) seeds were compared with metabolite levels of stratified Arabidopsis Wassilewskija (Ws) seeds in this case. However, this type of analysis opened the possibilities of combining transcript and metabolite data using MapMan.

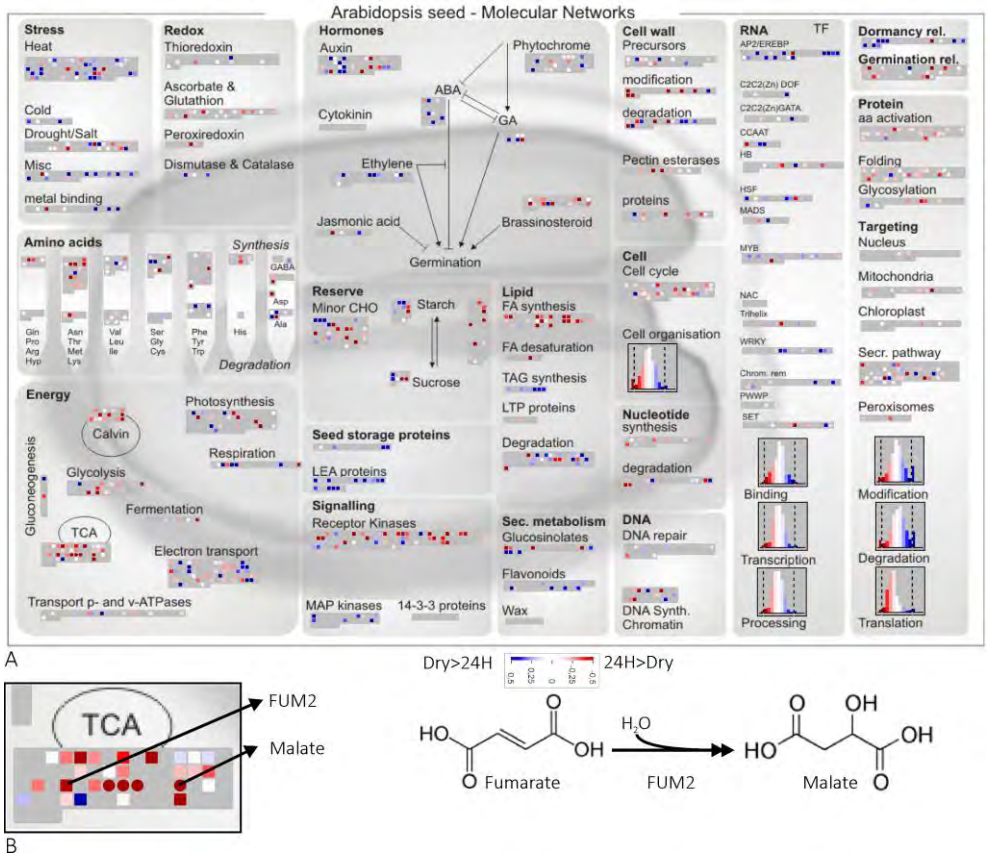


Figure 4.2: (A) MapMan Arabidopsis seed - Molecular Networks map. The Molecular Networks map shows differences in both transcript (colored squares) and metabolite (colored circles) levels. Squares and circles can be clicked to retrieve gene or metabolite data. (B) Example of close-up view of one category. The TCA cycle in the 'Energy' category is shown to depict the accumulation of malate (red circle) and the up-regulation of a gene (FUM2) encoding a fumarase, which catalyzes one of the steps of the TCA cycle and converts fumarate to malate, in 24-h-imbibed seeds. Log2-1 ratios are used to express relative levels of transcripts and metabolites in dry versus 24-h-imbibed seeds using a color scale. Red, higher levels in 24-h-imbibed seeds; blue, higher levels in dry seeds. Only ratios with a P-value <0.05 are presented.

Arabidopsis seeds show certain levels of primary dormancy immediately after seed harvest. In our second example, we visualized changes in molecular processes that are affected by dormancy (Figure 4.3). Therefore, we plotted the expression ratios for primary dormant seeds that were imbibed for 24 h (PD24H) and seeds that were after-ripened for 120 days and imbibed for 24 h in the dark with a 4-h pulse of red light (LIG). The PD24H seeds will not complete germination in contrast to the LIG-treated seeds which do complete germination. When dormant and non-dormant seeds were compared, obvious transcriptional differences were observed in the gene clusters, 'Cell wall', 'Stress', 'Secondary metabolism' and 'Hormones'. Surprisingly, relatively small changes were observed in Taggit gene clusters dormancy-related and germination-related.

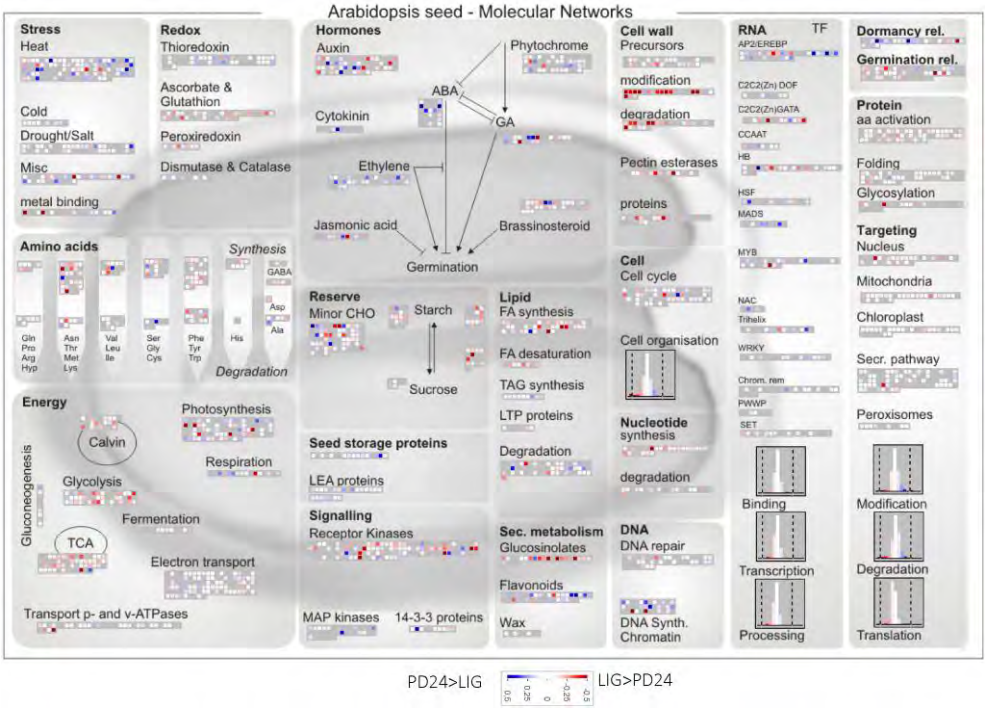


Figure 4.3: MapMan Seed Molecular Networks map showing differences in transcript (colored squares) levels. Log2-1 ratios are used to express relative levels of transcript in the primary dormant seeds (PD24H) versus after-ripened seeds that were imbibed for 24 h (LIG). Red, higher levels in LIG seeds; blue, higher levels in PD24H. Only ratios with a P-value <0.05 are presented.

In Figure 4.2 and Figure 4.3, the same sets of data were used for the initial selection of gene function categories in Pageman. Because this could potentially lead to a self-fulfilling, re-detection of differentially regulated genes, we also analyzed a transcriptome dataset which was not used for our pathway selection. We compared transcript profiles of 24-h-imbibed, stored wild-type Landsberg erecta (Ler) seeds (Ler AR) and 24-h-imbibed, stored comatose (cts)-1 mutant seeds (cts-1 S) using our Mapman diagram (Figure 4.4). As expected, CTS-1 levels were strongly reduced in the mutant. Consistent with the results described by Carrera *et al.* (2007), effects on a production of anthocyanin pigment 2 protein (PAP2=MYB90), GA-responsive GAST1 protein homologs (GASA1, GASA4) and the flavonoid pathway were clearly visible (Figure 4.4, 'RNA' and 'Hormones'). Theodoulou *et al.* (2005) described the jasmonic acid (JA)-deficient phenotype of the cts-1 mutant. Our results suggest that up-regulation of a seed-specific JA biosynthesis gene (putative 12-oxophytodienoic acid reductase,'OPR') could be part of this mechanism (Figure 4.4, 'Hormones'). The up-regulation of several photosynthesis pathway genes in the 'Energy' category in the mutant is noteworthy and might be an intriguing starting point for new research.

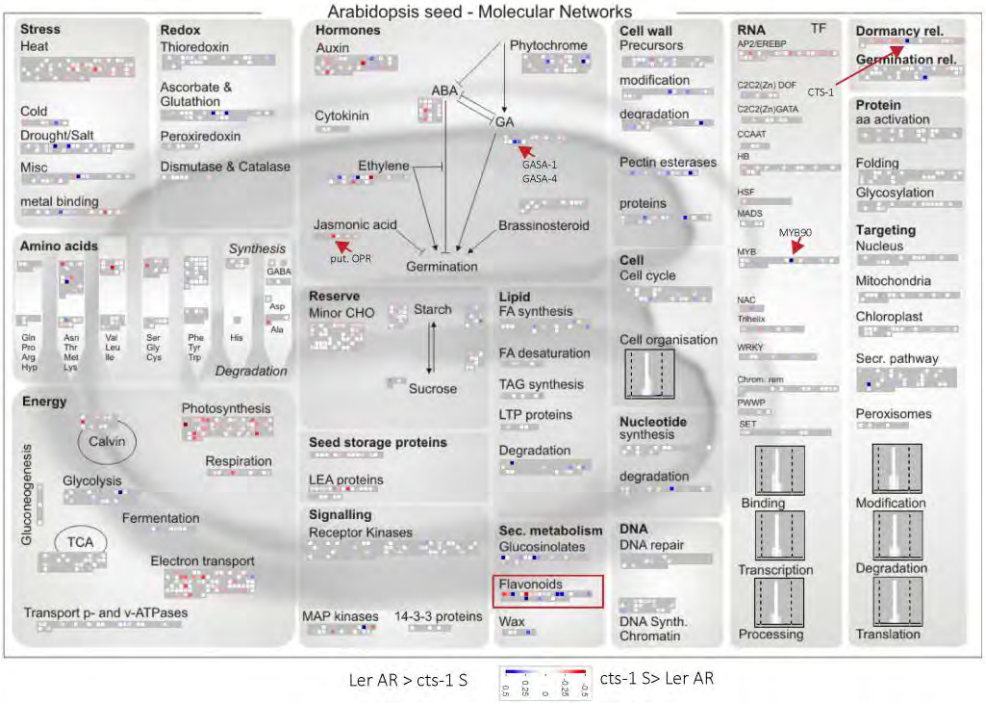


Figure 4.4: MapMan Seed Molecular Networks map showing differences in transcript (colored squares) levels. Log2-1 ratios are used to express relative transcript levels in of 24-h-imbibed, stored wild-type Landsberg erecta (Ler) seeds versus 24-h-imbibed stored cts-1 seeds. Red, higher levels in cts-1 seeds; blue, higher levels in Ler seeds. Only ratios with a P-value <0.05 are presented.

Several studies describe the important role of the endosperm layer in the regulation of seed germination (e.g. Müller *et al.* 2006; Penfield *et al.* 2006). We calculated ratios of gene expression levels between endosperm and embryo to visualize affected genes and molecular processes (Supporting Information, S4.3). As expected, genes related to photosynthesis and DNA synthesis are mainly expressed in the embryo. Since cell wall changes are crucial for the completion of germination, we created a diagram to specifically map cell wall changes and used it to highlight differences in cell wall modification between embryo and endosperm (Figure 4.5). AGPs were more linked to embryo cell walls. AGPs are a diverse class of cell wall proteins which are implicated in growth, development and plant-microbe interactions (Seifert and Roberts 2007), and therefore it is not surprising to find these transcripts more abundantly in embryo tissues.

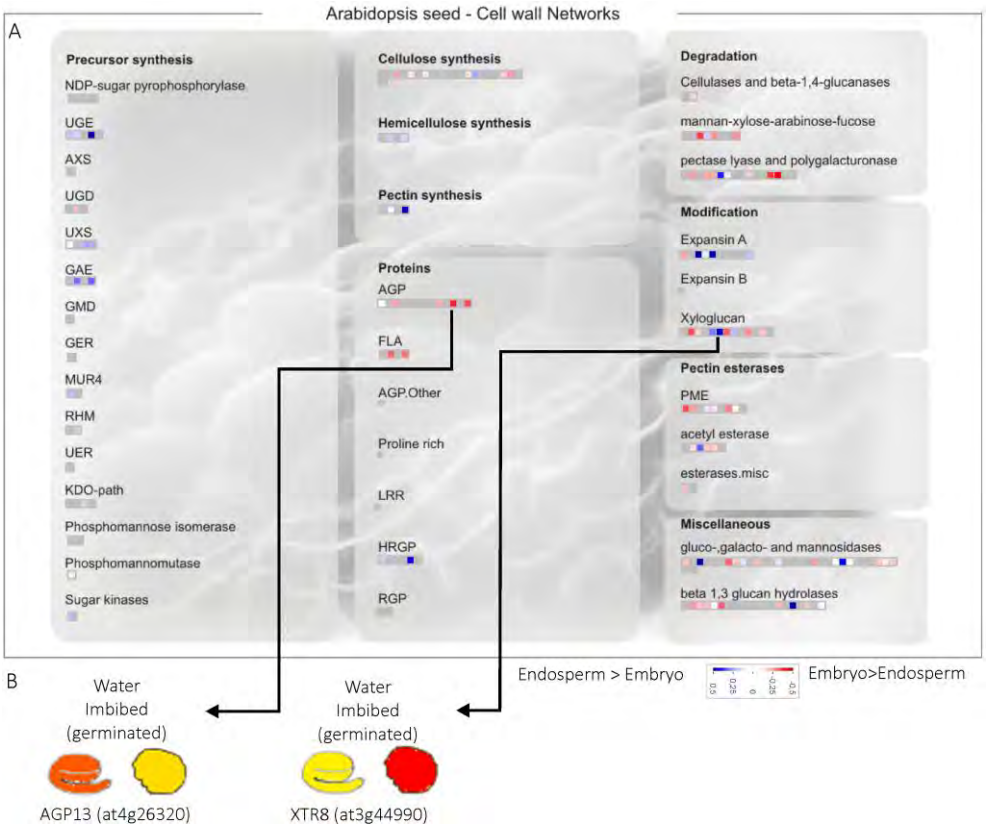


Figure 4.5: (A) MapMan Cell wall map showing differences in transcript levels between the embryo and endosperm. Log2-1 ratios are used to express relative transcript levels of the endosperm versus the embryo shortly after radicle protrusion. Red, higher levels in the embryo; blue, higher levels in the endosperm (A). Only ratios with a P-value <0.05 are presented. (B) Images of the eFP browser (BAR website: http://bar.utoronto.ca; (Winter et al. 2007; Bassel et al. 2008)) are included as examples of tissue-specific expression. Left and right, images of the enhanced expression of AGP13 (at4g26320) in the embryo and the expression of XTR8 (at3g44990) in the endosperm, respectively.

It is a major challenge to interpret the overwhelming amount of information currently available for genes, proteins and metabolites and understand their function in various biological processes. Combining the information about gene expression levels with known biological function of genes or gene classification can be very helpful in creating a categorization or applying priority within the data. The freeware tool MapMan has proved to be an easy-to-use and helpful tool to visualize multilevel data (transcriptomics, metabolomics and proteomics).

The data used in this study (summarized in Table 4.1) and the way of data visualization (Figure 4.1) that we examined turned out to be very informative as it clearly summarizes the molecular processes that are affected. For example, the importance of triacylglycerol (TAG) and protein synthesis during germination can readily be inferred from

the diagram. This is in agreement with the role of TAG metabolism in germination control and seedling establishment (Penfield *et al.* 2007) and the essential role described for translation in the completion of germination (Rajjou *et al.* 2004).

When transcript levels of dry seeds are compared to seeds that have been imbibed for 24 h many important molecular processes can be recognized (Figure 4.2A). Genes in the Calvin cycle, glycolysis and TCA cycle seem to be up-regulated, as well as in redox modification, cell cycle, cell wall -modification, -degradation and protein activation and folding. Contrarily, transcripts for seed storage proteins, LEA proteins and TAG synthesis are severely decreased in imbibed seeds. Noteworthy, genes involved in electron transport and respiration seem to be lower expressed in 24 h imbibed seeds compared to dry seeds. Our analysis indicated that the major differences between dormant and non-dormant seeds were in the Hormone and Cell wall clusters (Figure 4.3). Interestingly, only minor differences were observed in other processes such as seed storage proteins and LEA proteins between dormant and non-dormant seeds. This is due to the decrease in the transcript levels of these proteins to a similar extent in both dormant and non-dormant seeds upon imbibition. It is possible that seed storage proteins and LEA transcripts are remnants from the seed developmental stages, which are probably not necessary during imbibition anymore. These MapMan diagrams allow users not only to observe global changes but also to retrieve detailed information about gene annotation and expression, because the users can click on an individual process or gene in the interactive MapMan tools.

By plotting the transcriptome differences between the comatose-1 mutant and wild-type we showed that our selection of molecular processes has the potential to clearly visualize the affected genes and pathways as they were previously described (Figure 4.4) (Theodoulou *et al.* 2005; Carrera *et al.* 2007). For a more detailed view on a certain process or metabolic pathways one can make a customized diagram (as we showed for cell wall changes) or use already available diagrams in the MapMan package.

We have explored the possibility to use MapMan for multi-level data by combining transcriptome data from Nakabayashi *et al.* (2005) (dry vs. 24h imbibed seeds) with metabolome data from Fait *et al.* (2006) (dry vs. germinating seeds). In this way, relationships between gene expression and metabolome changes can easily be visualized as we depicted for the TCA cycle genes and metabolites (Figure 4.2B).

While our analysis demonstrated the usefulness of the Seed - Molecular Networks diagram, it does not cover all functional categories in every possible seed experiment. Besides, one should bear in mind that non-annotated genes are rarely selected for visualization, which hampers the discovery of new genes with unknown function. Also, it can be misleading when the original functional annotation is incorrect. Despite the aforementioned issues, the annotation used in MapMan has attained a high quality level and its usability will only improve, because knowledge about many genes and biological processes is rapidly increasing.

In conclusion, the MapMan tool allows a quick identification of the molecular processes that are regulated during a developmental program of interest, for which candidate genes with known annotation can easily be identified. This way of data visualization and the two pathway files that we have created provide a solid base for a next level of statistical data analysis and are useful tools for the seed science community.

Acknowledgements

This work was supported by the Technology Foundation STW (RJ, WL) and the ERA-NET Plant Genomics grant vSEED (BD). Raw microarray data were kindly provided by N. Provart and H. Nahal, University of Toronto. The background photograph in the cell-wall diagram was kindly provided by N. Everitt, N. Weston and S. Pearce, University of Nottingham.

Supporting Information

Supporting information can be downloaded from either the online version of this article (Joosen *et al.* 2011) or from:
www.wageningenseedlab.nl/thesis/rvljoosen/SI/chapter4

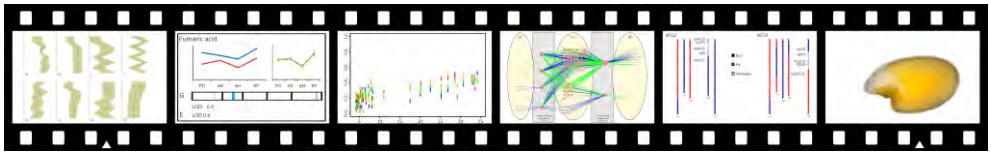
File S4.1: Transcriptome data of 20 microarray experiments used for Pageman analysis
File S4.2: Log2 ratios with P-values used for MapMan analysis
Figure S4.3: Seed-Molecular Networks map showing differences in transcript levels between embryo and endosperm. Log(2-1) ratios of endosperm vs embryo samples of seeds shortly after radicle protrusion are used to express level differences with help of a false color scale. Red indicates higher levels in embryo, blue indicates higher levels in endosperm. Only ratios with a P-value <0.05 are represented.

"All the flowers of tomorrow are in the seeds of yesterday."

5 Identifying genotype-by-environment interactions in the metabolism of germinating seeds using Generalized Genetical Genomics

Joosen RVL, Arends D*, Li Y*, Willems LAJ, Keurentjes JJB, Ligterink W, Jansen RC, Hilhorst HWM*

Accepted for publication in Plant Physiology
**Equal contribution*



Abstract

A complex phenotype such as seed germination is the resultant of several genetic and environmental cues and requires the concerted action of many genes. The use of well-structured recombinant inbred lines in combination with omics analysis can help to disentangle the genetic basis of such quantitative traits. This so called genetical genomics approach can effectively capture both genetic (G) and epistatic interactions (G:G). However, to understand how the environment interacts with genomic encoded information (G:E) a better understanding of the perception and processing of environmental signals is needed. In a classical genetical genomics setup this requires replication of the whole experiment in different environmental conditions. A novel generalized setup overcomes this limitation and includes environmental perturbation within a single experimental design. We developed a dedicated QTL mapping procedure to implement this approach and used existing phenotypical data to demonstrate its power. Additionally, we studied the genetic regulation of primary metabolism in dry and imbibed Arabidopsis seeds. Many changes were observed in the metabolome which are both under environmental and genetic control and their interactions. This concept offers unique reduction of experimental load with minimal compromise of statistical power and is of great potential in the field of systems genetics which requires a broad understanding of both plasticity and dynamic regulation.

Introduction

The use of natural variation to disentangle the genetic mechanisms underlying phenotypic differences has been very successful both in crop plants and in the model plant *Arabidopsis thaliana* (Alonso-Blanco *et al.* 2009). Most of the variation within wild or domesticated plant species is of quantitative nature determined by genetic polymorphisms at multiple loci. Such quantitative trait loci (QTL) can be analyzed efficiently using experimental mapping populations like recombinant inbred lines (RILs) derived from directed crosses. Nowadays, many well-structured RIL populations are available, often accompanied with detailed studies of phenotypic variation (Mitchell-Olds and Schmitt 2006). The complexity of quantitative traits is further determined by genetic interactions between genomic loci, i.e. epistasis (G:G), and between the genotype and the environment (G:E). While epistasis can be effectively identified in QTL analyses, albeit with lower power than main effects, the detection of G:E interactions requires experimentation in multiple conditions of interest. Because of the large population sizes often needed to obtain sufficient statistical power for QTL detection, G:E interactions are usually ignored in experimental setups. However, a better understanding of the perception and processing of environmental signals is needed, since interactions provide important insights in adaptation mechanisms and evolutionary constraints such as balancing and disruptive selection.

To obtain a more detailed view of the molecular mechanisms underlying phenotypic variation, genetical genomics studies, in which molecular traits are genetically analyzed, have been successfully applied to enhance a directed strategy to identify causal relationships (Kliebenstein *et al.* 2006; Keurentjes *et al.* 2007; Van Leeuwen *et al.* 2007; Wentzell *et al.* 2007; West *et al.* 2007; Rowe *et al.* 2008). The observed phenotype is often the resultant of a functional cascade of gene transcription followed by protein translation and modification which finally results in a highly dynamic metabolome underlying emergent properties (Kooke and Keurentjes 2011). With the technological advances made in genomic analytical platforms, such as transcriptomics, proteomics, and metabolomics, the large-scale, high-throughput analyses needed for quantitative genetic approaches have become feasible (Jansen and Nap 2001; Keurentjes *et al.* 2008). To incorporate developmental and environmental perturbation in the often expensive and laborious omic analyses, an alternative experimental setup, coined generalized genetical genomics (GGG), using balanced fractions of a RIL population has been proposed (Li *et al.* 2008). It provides an inexpensive experimental setup for hypothesis generating research in multiple environments. Such an approach aims at the creation of sub-populations of RILs, one for each environment to be tested, with an optimal distribution of parental alleles over all available markers (Li *et al.* 2009). When these sub-populations are subjected to environmental perturbation the emerging phenotypes can be explained by several sources of variation: 1) genetic variation 2) environmental variation and 3) genetic x environmental variation. Whenever the resulting phenotype is not or only mildly affected by environmental interactions (G:E), the analysis of the different sub-populations can be

combined gaining the full power of a complete population. However, when a trait shows strong G:E interaction, e.g. those that only express genetic variation in specific environments, the power to detect QTL is dependent on those sub-populations expressing the genetic variation. Although G:E interactions have been detected previously in genetical genomics studies for expression (Li *et al.* 2006; Smith and Kruglyak 2008; Gerrits *et al.* 2009; Yeung *et al.* 2011) and metabolite content (Zhu *et al.* 2012) by analyzing all lines in a population under different environments, the GGG concept offers unique reduction in experimental load with minimal compromise to statistical power and is of great potential in the field of systems genetics in which a broad understanding of both plasticity and dynamics is required (Li *et al.* 2008). As a proof of principle we present experimental data on the genetic regulation of primary metabolism in dry and imbibed *Arabidopsis* seeds using a GGG design and discuss the application and implications of such a strategy.

Plants are extremely rich in biochemical compounds and major roles in plant development, adaptation and defense have been identified for biosynthesis pathways and their products (Binder 2010). In *Arabidopsis*, genetic variation for many metabolic compounds has been observed, but G:E interactions were ignored in these studies (Kliebenstein *et al.* 2001; Keurentjes *et al.* 2006; Rowe *et al.* 2008) and are only addressed by Chan *et al.* (2011). Here we report on the interaction of four different physiological environments, i.e. developmental stages, in dry and imbibed seeds with two founder genotypes in a RIL population. To detect the majority of the most prominent primary metabolites we used gas chromatography-mass spectrometry (GC-MS) of polar extracts (Roessner *et al.* 2000; Lisec *et al.* 2008). These include essential metabolites such as sugars, amino acids, and organic acids, which are key compounds in reserve storage and catabolism, growth and energy metabolism. The biosynthetic pathways of primary metabolites are well-studied and often well-conserved between different taxa (Peregrin-Alvarez *et al.* 2009). Nonetheless, quantitative variation for many of these compounds can be observed between natural variants which might be reflected in their different growth characteristics. The analysis of single gene mutants, for example, has unraveled many key components in biochemical pathways and has demonstrated their role in phenotypic traits (Fiehn *et al.* 2000). Metabolic profiling at different growth stages has further revealed important fluxes that regulate plant development and adaptation (de Oliveira Dal'Molin *et al.* 2010). Using the accumulated historical mutations that occur in natural variants in combination with metabolic profiling in a generalized design offers the unique possibility of identifying genetic effects over a series of developmental stages.

The switch from a dry seed, which is equipped for optimal survival and storage of reserves, towards an imbibed seed, in which energy needed for germination is released and which prepares for autotrophic production is remarkable. Reserves that have been stored during seed maturation are degraded and remobilized during germination (Bewley 1997; Shu *et al.* 2008), a process that is heavily influenced by the capacity of C/N partitioning of a maturing seed (Dowdle *et al.* 2007). Arabidopsis mutants affected in their oil reserve content or its mobilization show delayed, but not full inhibition of germination (Kinnersley

and Turano 2000; Bouche and Fromm 2004; Shu *et al.* 2008; Kelly *et al.* 2011). This suggests an additional metabolic switch that occurs during seed desiccation after seed maturation, involving a change from accumulation of oil and storage proteins to the synthesis of free amino acids, sugars, fatty acids and their degradation products functioning to prepare for rapid metabolic recovery during imbibition (Fait *et al.* 2006; Angelovici *et al.* 2010). Imbibition of mature seeds specifically shows reduction of the metabolites that accumulate during the desiccation period. Upon germination, an increase of many metabolites, including amino acids, sugars and organic acids, can be observed again, which reflects the increase of autotrophic activity (Fait *et al.* 2006). Profiling the primary metabolome over different developmental stages in a mapping population is therefore expected to reveal the dynamic genetic regulation of many of these important processes. We will demonstrate here that much of the observed variation in biochemical profiles can be attributed to genotype-by-environment interactions which can be effectively identified in a generalized genetical genomics approach.

Results and Discussion

Experimental design

Previous studies which focused on the comparative analysis of developmental and metabolic variation suggest a link between central metabolism and plant physiology, but genetic co-regulation is not frequently observed (Keurentjes *et al.* 2006; Meyer *et al.* 2007). That said, in several studies in Arabidopsis a major metabolite QTL cluster is associated with the ERECTA locus, representing a strong regulator of development which is known for its pleiotropic effects (Fu *et al.* 2009). To circumvent this strong bias we used two natural variants, Bayreuth-0 (Bay-0) and Shahdara (Sha), which are not polymorphic for the ERECTA locus. The Bay-0 x Sha RIL population (Loudet *et al.* 2002) has previously been shown to contain genetic variation for seed germination (Joosen *et al.* 2012) and other physiological traits (Loudet *et al.* 2003; Barriere *et al.* 2005; Loudet *et al.* 2005; Diaz *et al.* 2006; Reymond *et al.* 2006; Loudet *et al.* 2008; Meng *et al.* 2008), anion strength (Loudet *et al.* 2003), carbohydrate content (Calenge *et al.* 2006), gene expression (West *et al.* 2007) and primary (Rowe *et al.* 2008) and secondary metabolite levels (Wentzell *et al.* 2007).

Powerful mapping of genetic variation in a RIL population is dependent on the size of the population, the level of recombination and on an evenly genome-wide distribution of the parental alleles. A core set of the Bay-0 x Sha RIL population (Loudet *et al.* 2002) consisting of 165 lines and optimized for the aforementioned factors was used in this study. This core population was divided in four sub-populations optimized for the distribution of parental alleles using the R-package DesignGG, aiming at the most accurate estimate of genetic and G:E effects (Li *et al.* 2009) (Supporting Information, S5.6).

Comparison of different designs using classic phenotypes

Standard QTL mapping procedures can efficiently capture genetic variation and epistasis, but do not take environmental perturbation into consideration. Appropriate modeling of the genetic variance-covariance (VCOV) in the data is of great importance when combining information from different environments in QTL analysis (Churchill 2002). Linear models are particularly well suited for this. Here environmental differences are incorporated as an additional variable in a generalized design (GGG design). To enable mapping of the observed trait variation and taking the four developmental stages into consideration an R-script was developed which use functions and data structures from the R/qtl package (Broman *et al.* 2003; Arends *et al.* 2010) (Supporting Information, S5.3). The R-script uses a linear model to calculate the likelihood of genotype to phenotype linkage for each marker with the following formula:

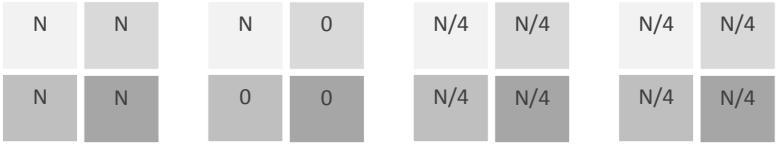
$$y_i = \beta_0 + \beta_1 e_i + \beta_2 g_i + \beta_3 e_i{:}g_i + \varepsilon_i$$

where y_i is the i^{th} observation of the studied phenotype, variable g_i is the genotype, e_i is a vector with seed conditions, and $e_i{:}g_i$ the interaction term. The values β_j represent parameters to be estimated, and ε_i is the error term. The simplified description Y = E + G + G:E + ε of this linear model will be used henceforward. Separate likelihood estimates (–log10 Probability, henceforth LOD scores) are generated for the environmental (E), genetic (G) and genetic x environmental (G:E) effects.

To validate the use of a GGG design, we studied the genetic (G) and the interacting effects between G and E (G:E) on phenotypes in four different environmental conditions (E). These phenotypes were obtained by studying different germination parameters under different environmental conditions (Joosen *et al.* 2012). In total we compared the power of different designs by performing QTL analysis for 96 classic phenotypes under 4 different environments (Joosen *et al.* 2012). Furthermore, we also investigated the interacting effect between genotype and environment. The full model mapping (Y = E + G + G:E + ε) was applied to a full block design, random design and GGG design. Single maker mapping (Y = G + ε) was applied to a single block design. The number of detected QTL and interacting QTL (FDR = 0.05, based on >10000 runs permutation) within the different designs are shown in Table 5.1. In the full block design all samples were allocated to the four conditions. Obviously, this is the most expensive way of performing the experiment as the required resources and effort are quadrupled (4 x N). As a consequence of the size of the experiment, the power of detecting genetic effects is the best for this design. Unfortunately, we cannot afford such expensive experiments in many situations due to limited resources and time. The single block design only focuses on one of the four conditions, as in most published genetical genomics studies to date. In this way the samples size for the selected condition is N and we will have equal power as in the full block design for detecting the genetic effects for this particular condition. Clearly, this design will miss

the information from the other three conditions and interacting effects between genetic and environmental factors cannot be investigated. In order to study both genetic and interacting effects with a limited budget, the random and the GGG design allocate the N different samples to the four environments evenly, measuring N/4 samples in each condition. Although the possibility to detect genetic effects is only slightly better for the GGG design, the detection of interacting QTL is clearly improved in the GGG design as compared to the random design. These results show that the optimal allocation of samples in the GGG design clearly improves the ability to detect both genetic and interacting effects and that the GGG design results in the maximization of detected variation in relation to the necessary resources with only a minimal compromise of statistical power as compared to the full block design.

Table 5.1: Comparing different experimental designs. Comparison of different experimental designs to study G and G: E effects on phenotypes in four different conditions. Each environmental condition is indicated with different gradients of grey in the blocks. In total there are N (=164) genetically different RILs and the data was analysed in 4 different ways.



Design	Full block design	Single block design	Random design	GGG design
	Best power for G	Same power for G in the selected condition	Limited power for G	Optimal power for G
	Most expensive	Less expensive	Less expensive	Less expensive
	Best power for GxE	Missing GxE	Limited power for GxE	Optimal power for GxE
QTL	96	93	78	81
Interacting QTL	30	0	17	27

Metabolic analyses

To study the metabolic status of Arabidopsis seeds during germination, four biologically important developmental stages of seed germination with expected variation in metabolite levels to different extent were selected. The first two stages, being freshly harvested primary dormant (PD) and after-ripened (AR) non-dormant dry seeds, respectively, are expected to comprise a very similar metabolome as most, if not all, metabolic fluxes are arrested in the dry seed. The oil rich (~40%) Arabidopsis seeds (Hobbs *et al.* 2004) typically desiccate to moisture contents below 5% which results in an arrest of all enzymatic reactions due to the lack of free water. The other two stages represented early imbibition of seeds, imbibed for 6 hours (6H), and seeds at radical protrusion (RP),

respectively. Full rehydration of dry seeds typically completes in less than 2 hours and although developmental differences are not yet expected, many metabolic processes will have started after 6 hours of imbibition (Nakabayashi *et al.* 2005; Howell *et al.* 2009). Radicle protrusion marks the end-point of germination sensu stricto and is known to be accompanied by a major switch of both the transcriptome and metabolome (Nakabayashi *et al.* 2005; Fait *et al.* 2006). These four developmental stages are anticipated to vary to different degrees in their metabolic profiles, hardly any difference between dry seed samples, some differences between dry and imbibed seeds and very pronounced differences between dry seeds and seeds at radicle protrusion.

To determine the metabolic status of genetic variants in these different developmental stages, all individuals in the four sub-populations and their parental accessions were subjected to GC-TOF-MS. Each sample consists of the polar fraction of a methanol extract of a bulk of approximately 700-1000 seeds (20 mg). Samples were analyzed in random order and interspersed with pooled sample controls to control for experimental errors. The metabolic profiling of the segregating RILs was performed and the use of segregation population provides an intrinsic replication for each genotypic marker (Jansen and Nap 2001). In total 7537 mass peaks were detected, representing 161 metabolites according to centrotyping based on retention time and correlation structure (Tikunov *et al.* 2011). In total 63 metabolites could be annotated using an in-house constructed library and a publicly available mass spectra library (Schauer *et al.* 2005) (Supporting Information, S5.1).

The parental accessions Bay-0 and Sha were measured in duplicate for all four developmental stages allowing us to model the influence of condition and accession using a multi-factor univariate analysis of variance (ANOVA).

$$y_i = \beta_0 + \beta_1 \text{condition}_i + \beta_2 \text{accession}_i + \varepsilon_i$$

Analysis of variance for the parental samples identified 108 metabolites showing significant variation (FDR < 0.05) between developmental stages (E) and 85 showing variation between the parents (G) with an overlap of 54 metabolites showing variation between both variables in an interactive way (G:E) (Supporting Information, S5.2). For 37 metabolites no significant variation was detected between the parental accessions or in any of the developmental stages. A self-organizing map (SOM), created from the metabolites showing significant variation between the parents, groups different metabolites according to their accumulation pattern over different genotypes and developmental stages (Figure 5.1). Clearly different patterns of variation can be observed, namely genetic in panel A and H; environmental in panel C and D; genetic + environmental in panel B and G and genetic x environmental in panel E and F, illustrating the complex regulation of metabolic processes and the need for sophisticated analysis methods.

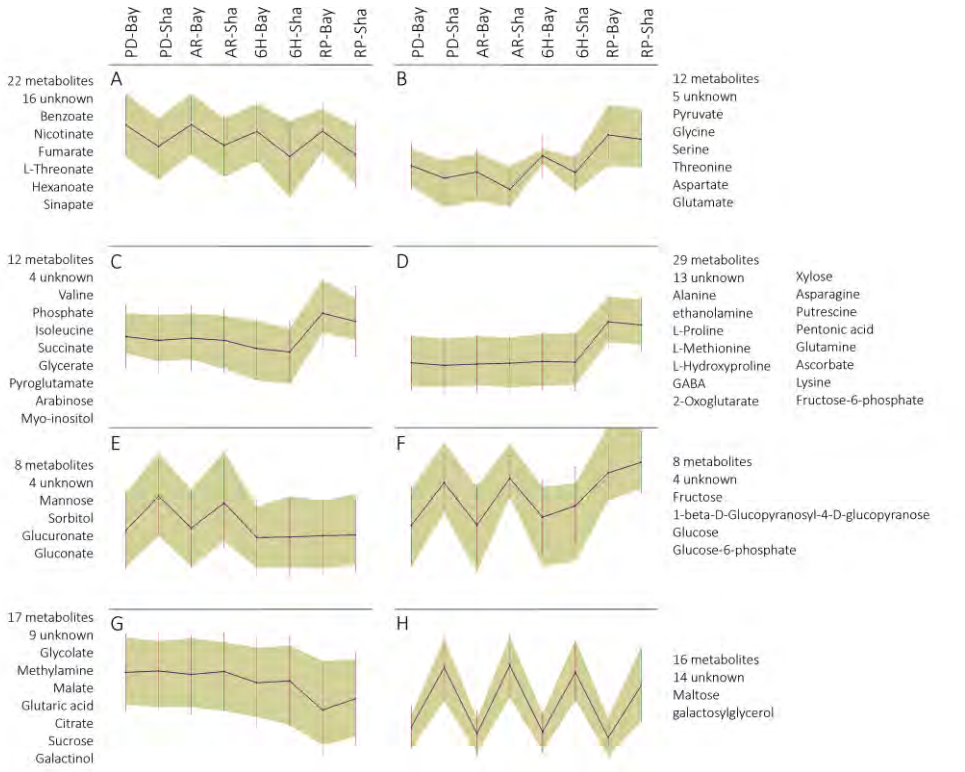


Figure 5.1: Self organizing map, grouping different metabolites according to their accumulation pattern over different genotypes and developmental stages of significantly variable metabolites (Anova F pr. < 0.05) measured in the parental lines Bay-0 and Sha in four developmental stages. PD=Primary dormant, AR=After-ripened, 6H=6 hour imbibed, RP=seeds at radicle protrusion. Two independent biological replicates were measured for each combination of parent and developmental stage.

Because metabolite levels are varying between both parents and between the chosen seed germination stages a segregation of metabolic accumulation can be expected in the RIL population of 164 lines. A principle component analysis of the metabolic profiles, revealing the internal structure in the data, shows that the first component clearly separates 6-hour imbibed seeds and seeds at radicle protrusion from both primary dormant and after-ripened seeds, explaining 37% of the total variation (Supporting Information, S5.7). This confirms the large metabolic changes accompanying the transition from dry arrested seeds to the imbibed and germinating developmental stages. As expected, no obvious differences could be detected between the metabolomes of primary dormant and after-ripened dry seeds. The second component, explaining 11% of the total variation, sharply separates the parental accessions, indicating that this component explains most of the genetic variation in metabolic profiles. These results demonstrate that Bay-0 and Sha possess substantial genetic variation for the accumulation of primary

metabolites which segregates in their recombinant offspring and which is strongly influenced by the developmental stage used for profiling. Transgressive segregation was visualized by comparing parental and RIL metabolite level distributions (Supporting Information, S5.8). Some positive and negative transgression is observed for most of the metabolites in which the metabolite accumulation in a RIL is respectively higher or lower compared to the respectively highest or lowest parent. In addition, 15 metabolites were detected in RILs which were not present in either parent. This suggests that new allele combinations in the RIL population resulted in enhanced accumulation or even novel formation of metabolites.

Genetic mapping in a generalized genomics design

In the experimental setup of this study, the environmental variation is defined as variation observed between the four developmental stages (PD, AR, 6H and RP). Significance thresholds, determined by permutation analysis (n=1000, $p<0.01$) for each metabolite, ranged from LOD 3.43 to LOD 3.50 and was stringently set to LOD 4 for all analyses. Mapping resulted in 120 significant QTLs in the genetic (G) component for 83 metabolites and 31 genetic x environmental (G:E) QTLs for 27 metabolites, ranging from one to four QTLs per metabolite. Thirteen of the G:E QTLs are significant in the G component as well. For 66 metabolites no significant QTL was detected. Clustered heatmaps for both the G and the G:E QTL profiles were created (Supporting Information, S5.9-S5.10).

To test the performance of the generalized mapping procedure, QTLs detected in individual environments (using the linear model $y_i=\beta_0+\beta_1 g_i+\varepsilon_i$, henceforth $Y=G+\varepsilon$) were compared to QTLs detected in the combined mapping approach (using the linear model $Y=E+G+G:E+\varepsilon$). QTLs were binned in upper or lower chromosome arms to reduce the effects of small positional shifts. Results were plotted in a network with nodes representing QTLs connected with edges to nodes representing the mapping populations in which they were detected (Figure 5.2). QTLs are grouped in three panels according to their detection in the different mapping procedures. The middle panel shows 73 QTLs that were detected in both the $Y=E+G+G:E+\varepsilon$ model and in one or more single environment mappings using the $Y=G+\varepsilon$ model. This shows that most of the genetic variation present in the single environments can effectively be captured by using the generalized model. The presence of 60 QTLs that were only significantly detected in the $Y=E+G+G:E+\varepsilon$ model (right panel) shows the combined power of the generalized approach and the usage of more genotypes. These QTLs are not detected in the single environment mapping in which only 41 individuals were used. Combining all data across all environments in the linear model increases power to detect QTLs, but it should be noted that there are also 20 minor QTLs (left panel) which are only significant in the single environment mapping with the $Y=G+\varepsilon$ model. These QTLs are not detected in the $Y=E+G+G:E+\varepsilon$ model. This can be explained by two factors: 1) environments in which the genetic variation is not expressed introduce

noise in the experimental data and thereby decrease mapping power, and 2) deviations from a balanced allele distribution in the different subpopulations can introduce some stochasticity around the threshold level.

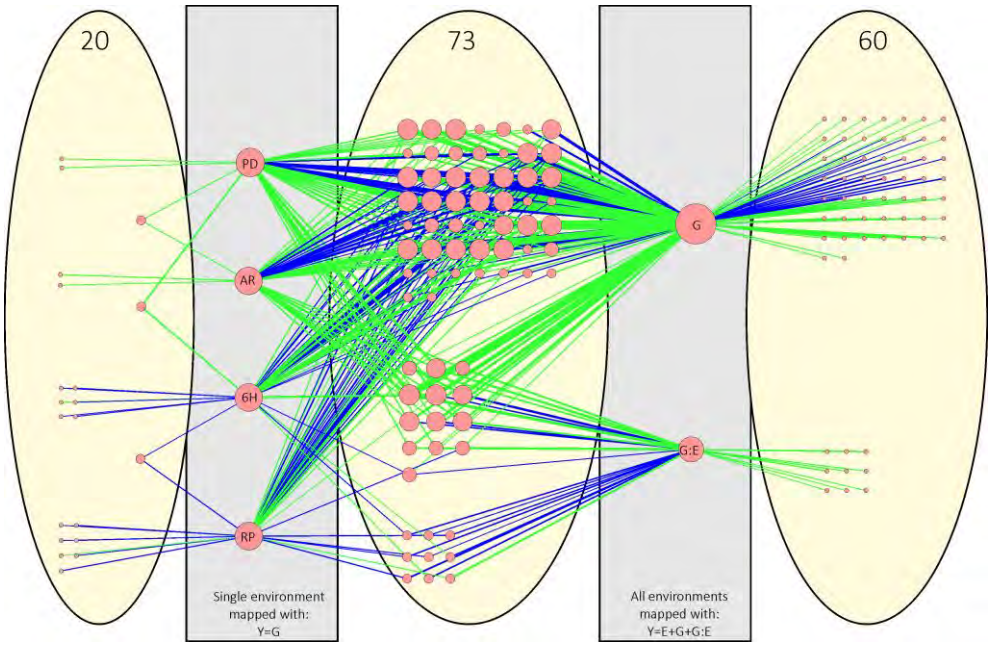


Figure 5.2: Comparison of QTLs detected within single environments (PD, AR, 6H and RP) by using the Y=G+ε model with QTLs detected when combining environments via the Y=E+G+G:E+ε model. QTLs were binned to two regions per chromosome. Nodes indicate metabolite QTLs and node size shows the degree of connectivity. Nodes are connected by edges which show the link between a QTL and a mapping population (single environments versus multiple environments). Separate nodes are created for the genetic (G) component and the genetic x environmental (G:E) component. Edge line color represents direction of the QTLs, green for higher levels in Sha; blue for higher levels in Bay-0. Line width indicates increasing LOD scores.

Importantly, all major to moderate effect size QTLs could be detected using the generalized model even when these QTLs were not detected in the separate environment models. Although it is difficult to compare power with the latter models, because population sizes differ, the generalized design efficiently identifies all relevant QTLs which were detected by the four separate models and in addition it detects G:E interactions. In a general exploratory study, the reduction in experimental burden therefore amply outweighs the incidental failure to detect the limited number of small-effect QTLs. The application of a GGG design can thus be an important advancement in evolutionary and ecological studies assessing the contribution of genetic and environmental effects to natural variation in life history traits.

For breeding purposes the allelic effect size is an important measure and differentiation of the environment in which the allelic effect is expressed can be very useful. In the generalized setup the allelic effect size of those metabolites with significant

QTLs is separated per environment (Supporting Information, S5.4-S5.5). For every QTL a LOD score for genetic effect is obtained from full model mapping. For these QTLs, normalized allelic effect sizes are calculated by Z-score transformations for each environment (Figure 5.3). QTLs detected in the G component of the linear model (Figure 5.3A) show an expected linear relationship between LOD score and effect size in all measured environments. This correlation is much weaker for QTLs detected in the G:E component of the linear model (Figure 5.3B), because the genetic variation is not expressed in all environments. QTLs of metabolites with strong G:E interaction, therefore, display larger effect sizes in fewer environments compared to G component QTLs of similar significance levels.

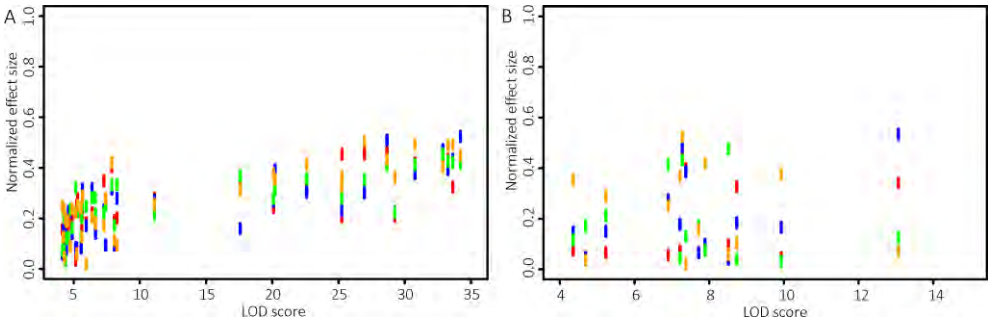


Figure 5.3: Effect sizes for each individual developmental stages are plotted against the derived LOD score. A: normalized allelic effect size per environment against LOD scores from the genetic (G) component and B: normalized allelic effect size per environment against LOD scores from the genetic x environmental interaction (G:E) component. Colors indicate the developmental stages (red = primary dormant (PD); blue = after-ripened (AR); green = 6 hours imbibed (6H); orange = seeds at radicle protrusion (RP).

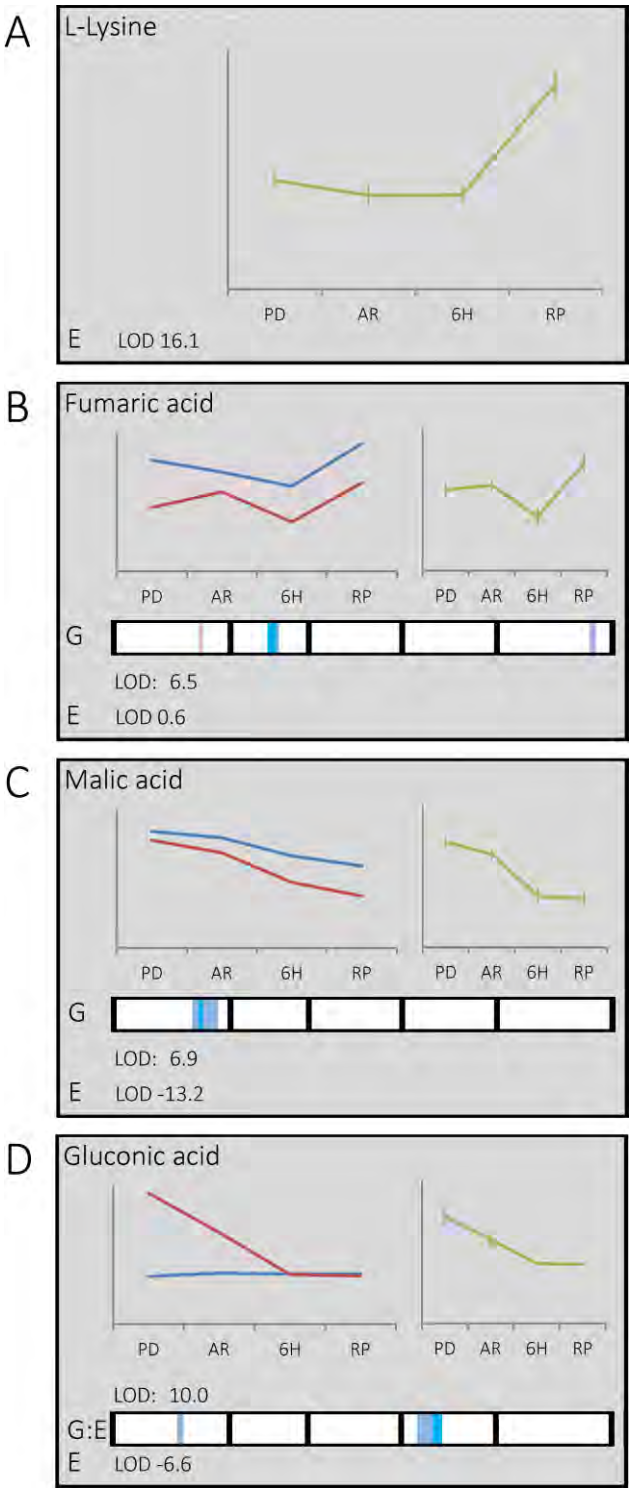
Clearly, the choice of environments used in such study is crucial. Limited power can be expected when environments vary too much and no overlapping genetic variation is present and contrarily there is hardly additive value of the design when using very similar environments. In this study we varied the environment by using a range of developmental stages starting from primary dormant dry seeds to seeds at the point of radicle protrusion. Different levels of environmental variation were obtained and could be mapped by the genetic (G) and/or genetic x environment (G:E) component of the linear model.

Genetic regulation of metabolic traits

One of the most rewarding benefits of the generalized approach is the possibility to analyze metabolic fluxes over different environments or developmental stages in addition to the effect of genetic variation. The acquired information of both sources of variation can be effectively displayed in so-called flash cards in which line graphs illustrate the genetic and environmental effect and detected QTLs are plotted in heat bars (Figure 5.4; Supporting Information S5.11). The individual components of the linear model

Y=E+G+G:E+ε provide the valuable measures for the various sources of variation. For example lysine content strongly increases in germinating seeds, indicated by a significant LOD score of 16.1 for the environmental effect, but no genetic variation for lysine could be detected (Figure 5.4A). For this metabolite genetic variants vary indistinguishable from each other over different environments. In contrast, fumaric acid shows little variation between the developmental stages (LOD 0.6), but displays strong genetic variation explained by a highly significant QTL (LOD 6.5) for the genetic effect at the center of chromosome II. Higher levels for fumaric acid are detected in all developmental stages for those lines harbouring the Bay-0 allele (Figure 5.4B). An example of the additive effect of environmental and genetic factors is the decrease in levels of malic acid in imbibed seeds. Here a strong environmental effect (LOD 13.2) is accompanied with an additional genetic effect explained by a G QTL (LOD 6.9) at the bottom of chromosome I. Note that the genetic effect here is similar in all environments (Figure 5.4C). This is not the case for gluconic acid which levels are strongly affected by the interaction between the genotype and the environment. A strong G:E QTL (LOD 10) is detected at the top of chromosome IV. The Sha allele at this position causes higher levels of gluconic acid in dry seeds, but not in imbibed seeds (Figure 5.4D). This strong negative environmental effect (LOD 6.6) is also responsible for the apparent directional shift of the G:E QTL effect.

Figure 5.4: Normalized metabolite changes during 4 developmental stages (PD, AR, 6H and RP). Each panel represents a single metabolite and contains information about environmental variation (green line plot, average over all lines within a single developmental stage) and genetic variation (blue lines represent the metabolite levels for lines carrying the Bay-0 allele for the most significant QTL and red lines those for the Sha allele carrying lines). QTL profiles for metabolites with either genetic (G) or genetic x environmental (G:E) variation are indicated at the bottom of each panel by a heat bar representing the 5 chromosomes. Environmental (E) variation is expressed as LOD score in the lower left corner. Depending on the most significant variation either genetic (G) or interaction (G:E) effects are also indicated with LOD scores in the lower left corner. A: L-Lysine showing only Environmental (E) variation; B: Fumaric acid: showing Genetic (G) variation; C: Malic acid showing both Environmental and Genetic variation (G+E); D: Gluconic acid showing interaction between Environment and Genetic variation (G:E).


A
L-Lysine
PD
AR
6H
RP
E
LOD 16.1
B
Fumaric acid
PD
AR
6H
RP
PD
AR
6H
RP
G
LOD: 6.5
E
LOD 0.6
C
Malic acid
PD
AR
6H
RP
PD
AR
6H
RP
G
LOD: 6.9
E
LOD -13.2
D
Gluconic acid
PD
AR
6H
RP
PD
AR
6H
RP
LOD: 10.0
G:E
E
LOD -6.6

Similar to the self-organizing maps in Figure 5.1 flashcards can be instrumental in the identification of metabolic relationships with the added value of genetic regulatory information. This is illustrated by integrating flashcards of all metabolites that were identified in this study with a general Arabidopsis metabolic pathway diagram (http://www.KEGG.jp, Supporting Information S5.12). For instance, several pathways in carbohydrate metabolism, such as the biosynthesis routes for galactose, pentose phosphate, starch/sucrose and amino and nucleotide sugars, are highly interconnected and are therefore subject to co-regulation mechanisms. A number of compounds involved in different subparts of the carbohydrate network module (e.g. glucose-6-phoshate, maltose, mannose, glucuronic and gluconic acid) indeed share a strong QTL at the top of chromosome IV. This suggests that the observed variation for these compounds has a single genetic basis, possibly affecting competition for a general precursor or directing feedback loops. In addition many of these compounds show strong positive or negative correlation due to environmental control. Genetic co-regulation was also observed for amino acid metabolism. Amino acids are substrate for the synthesis of aminoacyl-tRNAs which in turn are essential substrates for translation (Sheppard *et al.* 2008). A single G:E QTL at the bottom of chromosome I was detected for eight amino acids explaining a large part of the observed genetic variation. The joined analysis of environmentally and genetically induced variation in metabolic profiles can thus identify causal relationships between different modular parts of metabolic networks and associate these connections with relevant biological processes.

Regulatory hotspots and physiological co-regulation

As noted, the accumulation of several metabolites maps to identical positions suggesting that these might be regulated by a common genetic factor. Although co-locating QTLs can be the result of independent closely linked genetic factors, such coinciding QTLs are expected to occur more or less randomly by chance. Any deviation from expected frequency distributions along the genome thus hints at genetic co-regulation (Breitling *et al.* 2008). When plotted against their genomic position eight of such suggestive QTL hotspots can be seen (Figure 5.5) of which the two major ones (Chromosome IV-MSAT4.8 and Chromosome V-NGA139) co-locate with previously identified hotspots for metabolic regulation (Kliebenstein *et al.* 2001; Keurentjes *et al.* 2006; Wentzell *et al.* 2007; Rowe *et al.* 2008). Interestingly, both these loci have been shown to play a role in glucosinolate biosynthesis. The AOP locus at chromosome IV regulates side chain modification while the MAM locus at chromosome V determines chain elongation, but these compounds are not targeted for in GC-MS analysis which predominantly detects primary metabolites. As for many glucosinolates, for some metabolites, including GABA and maltose, QTLs were detected at both positions. In other cases a single QTL was detected at chromosome IV or V, e.g. glucose-6-phosphate and tyrosine, respectively. Although the identified primary metabolites are not directly connected with the glucosinolate biosynthesis pathway such

associations have been reported before (Rowe *et al.* 2008). These results might suggest alternative functions for AOP and MAM or a role in resource competition and allocation in central metabolism. This suggestion is further supported by the fact that these loci link to flowering time and the circadian clock regulation in the Bay-0 x Sha population (Chan *et al.* 2011). It also cannot be ruled out that other genes overlapping the AOP or MAM regions are causal for the observed variation.

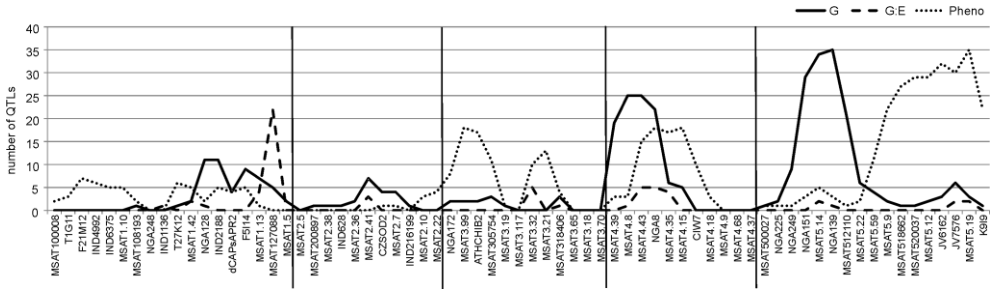


Figure 5.5: Number of significant QTLs plotted against the genetic location. Metabolic QTLs are represented by the solid (genetic component; G) and dashed (genetic x environmental component; G:E) lines. Germination related QTLs (Joosen *et al.* 2012) are shown by the dotted line.

Since many metabolites appear to be co-regulated, the strong impact of some loci on central metabolism might also exert its effect on physiological traits. Recently, the genetic landscape of seed germination in the same population has been described for which seed germination parameters were acquired under a wide range of environmental conditions (Chapter 3, Joosen *et al.* 2012). A comparison between variation in germination characteristics and metabolite levels might reveal compounds involved in the process of germination. Although no clear co-location of hotspots for germination and metabolite QTLs could be observed, incidental coincidence between isolated QTLs of both types of traits did occur. For instance, genetic variation for seed size co-locates with a large metabolic QTL cluster on the lower arm of chromosome I (~75 cM). This cluster contains many QTLs for amino acids, but also for components of the TCA cycle (e.g. fumarate and malate). In plants, leucine, isoleucine and valine, can be broken down and the end products of their catabolic pathways enter the TCA cycle to generate energy. It has been shown that these amino acids promote their own degradation, but only during seed germination, senescence, or under sugar starvation (Binder 2010). This suggests that the degradation pathways provide alternative carbon sources for the plant in extreme conditions. In addition, branched-chain amino acids and their derived alpha-keto acids are cytotoxic and preventing accumulation through degradation may be an important detoxification mechanism (Fujiki *et al.* 2000). Higher levels of both fumarate and malate, as a result of the degradation of a surplus of amino acids, might thus be indicative for larger seed sizes. A second QTL for seed size on chromosome V co-locates with a QTL of opposite effect for GABA accumulation. Interestingly, Bay-0 alleles at both QTLs confer larger seed size, suggesting directed evolution, as was also observed in a different population (Alonso-

Blanco *et al.* 1999). However, where levels of fumarate and malate are increased in larger seeds, the accumulation of GABA is decreased. GABA is known to be involved in a range of cellular processes (Palanivelu *et al.* 2003) and is rapidly accumulated in response to biotic and abiotic stresses (Kinnersley and Turano 2000). It has been postulated that it has roles in herbivore deterrence, pH and redox regulation, energy production and maintenance of carbon/nitrogen (C/N) balance (Bouche and Fromm 2004). In a recent study, GABA levels in seeds were shown to increase by expressing glutamate decarboxylase (GAD) under a seed maturation-specific phaseolin promoter (Fait *et al.* 2011). In accordance with our findings this resulted in smaller seed size and reduced seed vigor in T3 plants. No opposite seed size effect could be detected at a GABA QTL with increased levels due to the Bay-0 allele at the top of chromosome four, but co-locating genetic variation for germination on ABA, heat sensitivity and dormancy was observed at this position. These cases illustrate the power of joined genetic analyses of metabolic and physiological traits for generation of hypotheses that can help in the functional annotation of plant metabolites and their possible role in the regulation of important physiological processes.

Confirmation of mQTLs

To independently confirm the effect of a single locus it must be isolated and tested in an isogenic background. Several methods can be followed to perform such an independent confirmation of QTLs. A powerful approach is the use of residual heterozygosity in early generations of RILs. The Bay-0xSha RIL population (420 lines in total) was genotyped at F6 in which approximately 97% homozygosity is reached in each line. This resulted in the presence of residual heterozygosity in at least a single RIL at almost all genome positions. Those heterozygous regions are segregating in a Mendelian fashion in the next generation and can be used to confirm QTL positions, as it provides a possibility to study both parental alleles at the locus of interest in an otherwise homozygous background (Tuinstra *et al.* 1997). In a heterogeneous inbred family (HIF) those heterozygous regions are fixed and two separate lines containing the alleles of both parents respectively are maintained.

HIF312 and HIF214 are segregating for regions at the top of chromosome IV and V (Figure 5.6A), respectively, and cover the region in which the two major metabolite hotspots were detected. Because many of the QTLs detected in this region showed a large-effect size at the dry seed stages, after-ripened dry seeds were used to profile the HIFs for metabolic content. Significant differences between parental alleles using 4 replicates were defined by a two-tailed t-test ($p<0.05$). In total 34 out of 64 QTLs could be confirmed using this approach (Supporting Information, S5.13). For maltose for instance, two QTLs with opposite direction were found (Figure 5.6B) which both could be confirmed using the two distinct HIFs (Figure 5.6C).

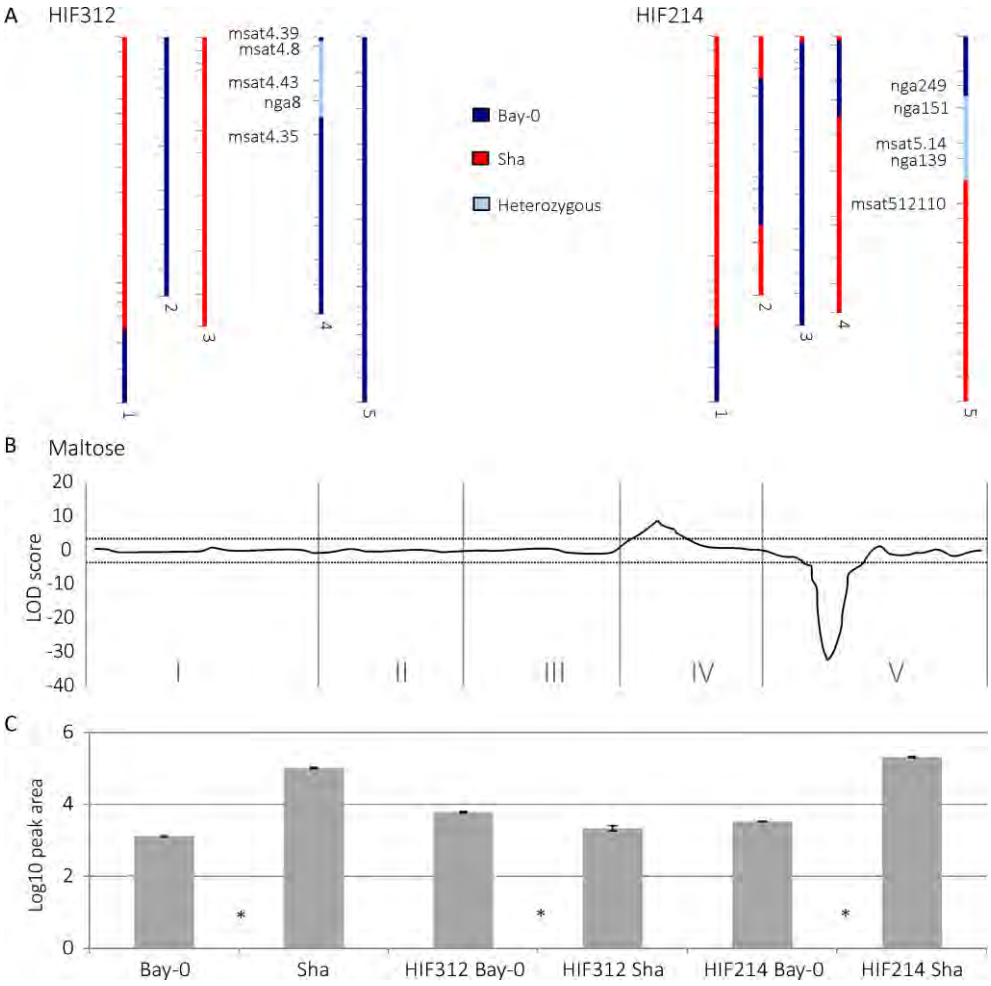


Figure 5.6: QTL confirmation for maltose using the heterogeneous inbred family (HIF) approach. Two QTL regions (top chromosome IV and top chromosome V) were analyzed using after-ripened (AR) seeds of lines HIF312 and HIF214 (A). The QTL profile for maltose (B) shows two significant QTLs (dashed line indicates the LOD 4 significance threshold). The lower panel (C) shows the parental levels for maltose and the confirmation for both QTLs by the segregating HIF lines (either fixed for Bay-0 or Sha alleles at the heterozygous interval). Significant differences (t-test $p<0.05$) are indicated with * in-between the two contrasting samples.

In a number of cases a HIF effect was observed that was not detected significantly in the RIL population (e.g. Digalactosylglycerol). This might be the result from the higher power in near isogenic lines due to the absence of epistatic interactions (Keurentjes *et al.* 2007). Nonetheless, a substantial number of QTLs could not be confirmed by the HIF lines. The enrichment for small-effect QTLs in the unconfirmed class suggests that four replicates generate insufficient power to identify significant differences for these metabolites in the HIF experiments, although we cannot rule out that they are false positives from the QTL analysis. Furthermore, QTLs depending on epistatic interactions cannot be detected in

some near isogenic lines. In addition, a number of QTL support intervals are broader than the region covered by the HIF and thus the causal genetic polymorphism within the QTL interval, but outside the region covered by the HIF, would have been missed.

The analyses of the HIF lines indicate that most of the large-effect QTLs can be accurately detected using a generalized genomics approach. Although an underestimation of small-effect QTLs can be expected this is largely compensated by the higher power of detecting genetic and environmental interactions.

Conclusions

The use of natural variation is a valuable tool to dissect the genetics of complex traits and the addition of powerful 'omics' analysis provides a great resource to disentangle molecular mechanisms. However, the expensive nature of many 'omics' experiments limits researchers to deploy perturbation of either environment or development. New strategies are needed to enable the switch from genetical genomics to system genetics. Here we have reported on a strategy to divide a RIL population in well-defined sub-populations and to use those to perturb the environment or developmental stage. To this end a novel R-script has been created to enable QTL mapping using a linear model that includes the possibility to account for genetic and environmental variation. This R-script is fast enough to analyze hundreds to thousands of traits and creates possibilities to extend the generalized genetical genomics strategy to whole genome gene expression analysis by either microarray or next generation sequence approaches (Joosen *et al.* 2009; Ligterink *et al.* 2012).

Efficient QTL mapping is strongly dependent on the population size and recombination frequency. Keurentjes *et al.* (2007) studied the effect of the population size and showed a linear relationship between the number of individuals used for mapping and the smallest detectable genetic effect. In this light it might seem undesirable to split a RIL population in smaller sub-populations. This is true when genetic variation is only detectable in a single unique environment or developmental stage leading to a strong genetic x environment interaction. More often, variation is subject to the environment without a complete abolishment of the genetic variation. In those cases the environmental effects can be normalized and the power of detecting a QTL is increased to the total number of lines used in the different sub-populations.

The availability of a genome wide set of heterogeneous inbred family (HIF) lines of the Bay-0 x Sha RIL population provides a solid and fast way to confirm QTLs. By using this approach we tested two of the observed QTL hotspots and were able to confirm many of the detected QTLs. When resources are limited this can be regarded as a good alternative for replicating the whole experiment for e.g. different growth seasons.

Many studies have shown the highly dynamic nature of molecular mechanisms leading towards seed germination (e.g. reviewed in Catusse *et al.* 2008; Daszkowska-Golec 2011; Weitbrecht *et al.* 2011). Performing expensive genetical genomic experiments without any perturbation of the environment will therefore always raise questions about

the possible extrapolation of the results when slightly different conditions are used. Information about the flux of a metabolite within a range of developmental stages or within a range of environments allows a much more precise interpretation of the molecular effects. By using the generalized strategy we showed that it is possible to deduct the metabolic fluxes (Figure 5.4). This extra level of information is a very valuable addition and helps to interpret the effect of genetic variation in the context of a dynamic and constantly changing metabolome.

Metabolite hotspots can reveal important loci involved in major metabolic pathway differences between two natural variants. In several studies the detected 'omics' hotspots did not co-locate more than expected by chance with phenotypic hotspots (Keurentjes *et al.* 2006; Meyer *et al.* 2007). However, in this study we detected some co-locating QTLs which might be explained by the narrow developmental window in which both metabolite and phenotypic QTLs (Chapter 3, Joosen *et al.* 2012) were gathered. We detected overlapping QTLs for amino-acid synthesis, TCA cycle compounds and seed size at the bottom of chromosome I and also co-location between QTLs for GABA, seed size and germination under stress conditions at the top of chromosome 5 (Chapter 3, Joosen *et al.* 2012). These co-locating QTLs are interesting leads for further research which is necessary to elucidate the true causal molecular mechanisms.

In conclusion, in the era of large systems genetics initiatives, we propose to consider the use of a generalized design for genetical genomics studies. The simultaneous acquisition of both genetic variation and developmental fluxes is a cost effective approach enabling a much better understanding of the processes involved. We see great potential in further exploration of the generalized design for transcriptome or other 'omics' related studies.

Material and methods

Plant material

Seeds from the core population (165 lines) of the *Arabidopsis* Bay-0 x Sha recombinant inbred line population (Loudet *et al.* 2002) and heterogeneous inbred family (HIF) lines were obtained from the Versailles Biological Resource Centre for *Arabidopsis* (http://dbsgap.versailles.inra.fr/vnat/). The population is mapped with 69 markers with an average distance between the markers of 6.1 cM (Loudet *et al.* 2002). Maternal plants were grown in a fully randomized setup and seeds from 4-7 plants/RIL were bulk harvested. Plants were grown on 4x4 cm rockwool plugs (MM40/40, Grodan B.V.) and watered with 1 g/l Hyponex fertilizer (NPK=7:6:19, http://www.hyponex.co.jp) in a climate chamber (20°C day, 18°C night) with 16 hours of light (35 W/m2) at a relative humidity of 70%. Seeds were either stored at -80°C 1 week after harvest (primary dormant; PD) or after-ripened at room temperature and ambient relative humidity until maximum germination potential after 5 d of imbibition was reached (after-ripened; AR). After-ripened seeds were imbibed on water

saturated filter paper at 20°C for 6H and quickly transferred to a dry filter paper for 1 minute to remove excess of water (6 hours imbibed seeds; 6H) Manual selection with help of a binocular was carried out to harvest seeds with the radicle at the point of protrusion (radicle protrusion; RP). Three radicle protrusion lines failed the metabolite analysis and were replaced by dry primary dormant samples.

Metabolite analysis

The metabolite extraction was performed based on a previously described method (Roessner *et al.* 2000) with some modifications. Seeds (20 mg) were homogenized using a micro dismembrator (Sartorius) in 2 ml tubes with 2 iron balls (2,5 mm), precooled in liquid nitrogen. 700 µl methanol/chloroform (4:3) was added together with the standard (0.2 mg/ml ribitol) and mixed thoroughly. After 10 minutes of sonication 200 µl MQ was added to the mixture followed by vortexing and centrifugation (5 min., 13500 rpm). The methanol phase was collected in a glass vial. 500 µl methanol/chloroform was added to the remaining organic phase and kept on ice for 10 min. 200 µl MQ was added followed by vortexing and centrifugation (5 min., 13500 rpm). Again the methanol phase was collected and mixed with the other collected phase. 100 µl was dried overnight using a speedvac (35°C Savant SPD121).

A GC-TOF-MS method (Carreno-Quintero *et al.* 2012) was used with some minor modifications. Detector voltage was set at 1600V. Raw data was processed using the chromaTOF software 2.0 (Leco instruments) and further processed using the Metalign software (Lommen 2009), to extract and align the mass signals. A signal to noise ratio of 2 was used. The output was further processed by the Metalign Output Transformer (METOT; Plant Research International, Wageningen) and mass signals that where present in less than 3 RIL's where discarded. Centrotypes were created using the MSclust program (Tikunov *et al.* 2011). The mass spectra of these centrotypes were used for the identification by matching to an in-house constructed library and the NIST05 (National Institute of Standards and Technology, Gaithersburg, MD, USA; http://www.nist.gov/srd/mslist.htm) libraries. This identification is based on spectra similarity and comparison of retention indices calculated by using a 3^{th} order polynomial function (Strehmel *et al.* 2008).

QTL mapping

Data was preprocessed using a log10 transformation and per phenotype outliers were removed after Z-transformation (Z-scores > 3). With the open source statistical package R (version 2.14.1) we fitted a basic linear model ($y_i=\beta_0+\beta_1 g_i+\varepsilon_i$) on the 4 conditions separately. This was followed by a combined mapping allowing for a developmental covariate and interaction term between the genetic marker and developmental stage ($y_i=\beta_0+\beta_1 e_i+\beta_2 g_i+\beta_3 e_i{:}g_i+\varepsilon_i$). P-values from all mappings are transformed into LOD scores by taking the –log10. Additionally, raw and normalized effects were calculated for each individual environment. Normalized effects were calculated by dividing the difference

between the maximum and the minimum value for that trait by the mean effect at the marker. LOD significance was determined using permutations for the combined mapping of the 4 environments: a LOD score of 4 was found to be significant (Breitling *et al.* 2008). Supporting Information S5.3 contains the R script used for the data analysis.

Acknowledgements

This work was supported by the Technology Foundation STW, the Applied Science Division of NWO (RVLJ, LAJW, WL). The Centre for BioSystems Genomics (CBSG) and the Netherlands Consortium of Systems Biology (NCSB), both of which are part of the Netherlands Genomics Initiative / Netherlands Organisation for Scientific Research (DA); the EU 7th Framework program under the Research Project PANACEA [222936] (RJ).

Supporting Information

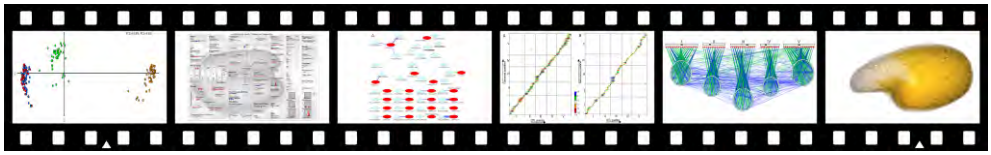
Supporting information can be downloaded from an online storage located at: www.wageningenseedlab.nl/thesis/rvljoosen/SI/chapter5

File S5.1: Metabolite centrotype data. Including peak numbers, retention time, hit quality, probability and mass used for quantification.
File S5.2: ANOVA results from metabolic profiling of the parental lines Bay-0 and Sha.
File S5.3: R-script with original data files allowing re-analysis of all data provided in this paper.
File S5.4: Summary of all detected metabolic G QTLs.
File S5.5: Summary of all detected metabolic G:E QTLs.
Figure S5.6: Allele distribution within the Bay-0 x Sha RIL population and the 4 selected sub-populations.
Figure S5.7: Principal component analysis plot showing the first two principal components of the metabolite analysis in the Bay-0 x Sha RIL population. Colors indicate the developmental stage (red = primary dormant (PD); blue = after-ripened (AR); green = 6 hours imbibed (6H); orange = seeds at radicle protrusion (RP), parental lines are indicated by triangles (Sha) or squares (Bay-0).
Figure S5.8: Transgression plot. Graph with scaled metabolite levels per RIL and parental levels.
Figure S5.9: Clustered heatmap from the genetic (G) component showing all metabolites.
Figure S5.10: Clustered heatmap from the genetic x environmental (G:E) component showing all metabolites.
Figure S5.11: Flashcards of all identified metabolites.
Figure S5.12: KEGG metabolic pathway with flashcards overlay of the metabolites identified in this study.
Figure S5.13: Overview from HIF analysis with all metabolites with significant QTL confirmation.

"No matter how much you eat, save some seeds for sowing."

6 Next generation eQTL mapping; a sneak preview

Joosen RVL, Willems LAJ, Provart NJ, Ligterink W, Hilhorst HWM



Abstract

Gene expression can vary between accessions due to an evolutionary accumulation of polymorphisms. Such quantitative variation can be used to identify the responsible loci; a procedure called eQTL mapping. In the era of microarray technology it became feasible to perform large scale genome wide expression analysis and several studies have presented the transcriptome architecture of single developmental stages. In this 'sneak preview' we present the first results of a newly generated dataset using the Arabidopsis Bay-0 x Sha recombinant inbred lines (RIL) population. Gene expression was profiled in seeds using a tiling microarray which allows full genome expression profiling and enables identification of genetic variation for alternatively spliced exons and anti-sense transcripts in future analysis. The used experimental design allows for perturbation of the environment within a single hybridization of the RIL population and has been applied to study four developmental stages during seed germination. We show that such design is effective to capture both genetic en environmental variation of gene expression.

Introduction

Natural variation provides a great resource to study the genetics of complex physiological traits which are determined by a concerted action of multiple genes. Quantitative genetics has been applied to study a wide range of agriculturally important traits (Alonso-Blanco *et al.* 2009). However, the cloning of the causal genetic polymorpishm requires finemapping which is labor intense and time consuming (Weigel and Nordborg 2005). Like many physiological traits, variation in gene expression often shows a quantitative distribution. This opens the possibility to subject expression variation to linkage analysis, a concept called 'genetical genomics' (Jansen and Nap 2001). Experiments following this concept combine a genotyped segregating population and genome wide expression profiling in order to formulate hypothetic regulatory pathways and disentangle complex traits in a more high-throughput manner. Several studies in a broad range of taxonomic kingdoms have now been conducted and demonstrate the power to refine molecular pathways and to identify key driver genes thereof (reviewed in Wittkopp *et al.* 2004; Mitchell-Olds and Schmitt 2006).

The proportion of genetically regulated genes that can be observed is depending on the genotypic diversity present in the mapping population and on biological factors that influence gene expression. Due to the expensive nature of whole genome expression profiling, most genetical genomic studies are confined to a single tissue, developmental stage or environmental condition. Increased understanding of the interplay between environment and genetic factors will allow a more precise prediction of the physiological effects of gene expression variation. An alternative design, called generalized genetical genomics, allows a cost efficient perturbation of the environment, tissue or developmental stage (Chapter 5, Li *et al.* 2008). The biological context in which regulatory networks function often determines the information about spatial or temporal variation.

Seed germination is a complex trait and is the result of an interaction between the genome and the environment encountered during seed development and maturation. This interaction is part of the adaptation of plants to a varying environment and is aimed at maximizing the probability of successful offspring (Penfield and King 2009; Huang *et al.* 2010). In *Arabidopsis thaliana* different QTLs were found for dormancy and for a range of germination characteristics but a detailed understanding of the molecular mechanisms that are affected during the transition from a dry dormant seed towards completion of germination is largely lacking (Bentsink *et al.* 2000; Clerkx *et al.* 2004; Bentsink *et al.* 2010; Galpaz and Reymond 2010; Joosen *et al.* 2012).

In this paper we describe a generalized genetical genomics experiment using the Bay-0 x Sha recombinant inbred lines (RIL) population (Loudet *et al.* 2002). This population was used in previous studies to characterize a broad range of seed germination related traits, metabolite and flavonoid levels (Rowe *et al.* 2008; Joosen *et al.* 2012; Routaboul *et al.* 2012). Further, it has been used for expression profiling of plants at the rosette stage and 10 day old siliques (West *et al.* 2007; Cubillos *et al.* 2012). Here, we expand the

expression profiling data with four different developmental stages during seed germination. The use of an Affymetrix tiling microarray enables a very detailed analysis of all annotated genes in the *Arabidopsis* genome (Zhang and Borevitz 2009).

Results and discussion

Experimental setup

In this study the core set of the Bay-0 x Sha RIL population (Loudet *et al.* 2002) consisting of 165 lines was used. To reduce environmental variation derived from the maternal plants as much as possible the population was grown in a randomized setup under fully controlled conditions and seeds from 4-7 plants/RIL were pooled. These seeds were extensively tested for their germination behavior and showed very good heritability scores when different harvests were compared (Joosen *et al.* 2012).

The Bay-0 x Sha RIL population was divided in four sub-populations optimized for the distribution of parental alleles using the R-procedure DesignGG (Li *et al.* 2009). Four biologically important developmental stages of seed germination were selected (Supporting Information, S6.1). The first two stages, freshly harvested primary dormant (PD) and after-ripened (AR) non-dormant dry seeds are expected to comprise a very similar transcriptome. The other two stages represented seed imbibed for 6 hours (6H) and seeds at radical protrusion (RP). Full rehydration of dry seeds is completed typically in less than 2 hours and although developmental differences are not yet expected, many metabolic processes will have started after 6 hours of imbibition (Nakabayashi *et al.* 2005; Howell *et al.* 2009). Radicle protrusion marks the end-point of germination *sensu stricto* and is known to be accompanied by a major shift in the transcriptome (Nakabayashi *et al.* 2005; Fait *et al.* 2006).

Full genome expression profiling was performed using the Affymetrix AtSNPtile microarray (Zhang and Borevitz 2009). This microarray contains 1.7 million unique 25mer tiling probes in sense and antisense direction covering the non-repetitive part of the genome at 35 bp resolution. Genomic DNA from the parental Bay-0 and Sha lines was hybridized in a triplicated experiment to enable filtering of probes with bad hybridisation characteristics. This prevents false positive eQTL detection which are solely caused by polymorphisms that lead to hybridisation differences. Tair9.0 (www.arabidopsis.org) gene annotation was used to extract antisense exon probes for each gene. The hybridisation levels of all selected probes per gene were averaged to obtain a solid measure for gene expression. In total we extracted expression levels for 29.304 genes from 180 microarrays (Supporting Information, S6.2). A principal component analysis of the expression profiles, revealing the internal structure in the data, shows clear separation patterns (Figure 6.1). The first component, explaining 54.6% of the total variation, separates 6 hours imbibed seeds and seeds at radicle protrusion from both primary dormant and after-ripened seeds. This confirms the large transcriptome changes accompanying the transition from dry

arrested seeds to the active imbibed and germination developmental stages. The second component, explaining 9.6% of the total variation, might be expected to separate the observed genetic variation. However, it should be noted that both parental accessions are not clearly separated in the second component. This might indicate that the non-shuffled parental genomes exhibit a relatively robust transcriptome, which is destabilized in the RILs.

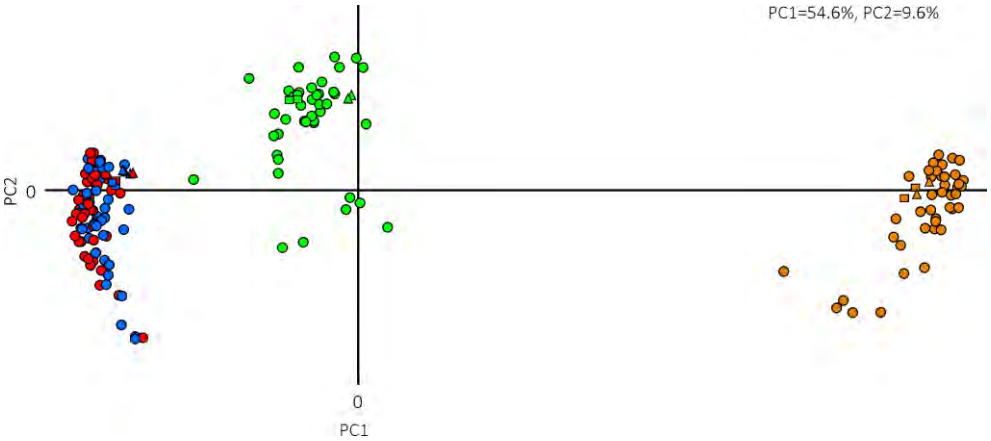


Figure 6.1: Principal component plot showing the first and second component of the explained variation derived from the whole transcriptome measurements per RIL line and both parental lines. Developmental stages are indicated by colors (red= Primary dormant (PD), blue = After-ripened (AR), green = 6H imbibed (6H), orange = seeds at radicle protrusion) and shape indicates the genotype background (circle = RIL lines, triangle = Bay-0, square = Sha).

Developmental variation

Although the primary purpose of this experiment was focused on detection of the genetic variation in gene expression, it offered great opportunities to acquire insight in developmental variation of transcription as well. For each developmental stage a total of 45 microarrays were analyzed. When comparing the transcriptomes of dry primary dormant (PD) with dry after-ripened (AR) seeds, only small expression differences were detected but a striking over-representation of down regulated cell-wall related genes can be observed (Table 6.1). Most probably, those genes are mainly located in the seed-coat in which the RNA might be less protected to oxidative degradation compared to the embryonic tissues. Whether the degradation of these genes is causally linked with loss of dormancy is an interesting subject for further research and can be tested by evaluating gene expression levels during seed after-ripening in lines with various dormancy levels.

As many as 6559 genes were found to be differently expressed when comparing the transcriptomes of dry after-ripened seeds against 6H imbibed seeds (Bonferroni corrected t-test P-value < $1.7\ 10^{e-7}$, Supporting Information S6.3).

Table 6.1: Overview of the 25 most significantly changed genes during seed dry storage. Average and t-test P-values of 45 microarrays performed on primary dormant (PD) seeds and after-ripened (AR) seeds each.

			P (t-test)	Average Expression		Standard error	
AGI-ID	Symbol	Description	PD-AR	PD	AR	PD	AR
AT3G13400	SKU5 SIMILAR 13 (sks13)	SKU5 similar 13 (sks13); FUNCTIONS IN: oxidoreductase activity, copper ion binding	2.3E-07	5.77	5.55	0.03	0.02
AT3G62730		Unknown protein;	6.2E-07	5.27	5.14	0.02	0.02
AT2G47050		Plant invertase/pectin methylesterase inhibitor superfamily protein; FUNCTIONS IN: enzyme inhibitor activity, pectinesterase inhibitor activity, pectinesterase activity;	1.2E-06	5.68	5.45	0.04	0.02
AT5G19580		Glyoxal oxidase-related protein;	2.2E-06	5.60	5.48	0.02	0.01
AT1G02790	POLYGALACT-URONASE 4 (PGA4)	Encodes a exopolygalacturonase.	5.4E-06	5.83	5.59	0.04	0.03
AT3G28750		Unknown protein; LOCATED IN: endomembrane system	5.7E-06	5.41	5.23	0.03	0.02
AT3G62170	VANGUARD 1 HOMOLOG 2 (VGDH2)	VANGUARD 1 homolog 2 (VGDH2); FUNCTIONS IN: enzyme inhibitor activity, pectinesterase activity	1.9E-05	5.30	5.20	0.02	0.01
AT3G01270		Pectate lyase family protein; FUNCTIONS IN: lyase activity, pectate lyase activity	2.9E-05	5.34	5.21	0.03	0.02
AT2G47040	VANGUARD1 (VGD1)	Share high homologies with a group of pectin methylesterases (PME)	3.3E-05	5.58	5.42	0.03	0.02
AT1G61563	RALF-LIKE 8 (RALFL8)	Member of a diversely expressed predicted peptide family showing sequence similarity to tobacco Rapid Alkalinization Factor (RALF)	5.1E-05	5.84	5.63	0.03	0.03
AT5G20390		Glycosyl hydrolase superfamily protein; FUNCTIONS IN: cation binding, hydrolase activity, hydrolyzing O-glycosyl compounds, catalytic activity	7.2E-05	5.17	5.08	0.02	0.02
AT5G14380	ARABINOGALACTAN PROTEIN 6 (AGP6)	Encodes an arabinogalactan protein that is expressed in pollen, pollen sac and pollen tube.	1.0E-04	5.84	5.69	0.03	0.02
AT5G07430		Pectin lyase-like superfamily protein; FUNCTIONS IN: pectinesterase activity	1.2E-04	5.04	4.92	0.03	0.02
AT5G26700		RmlC-like cupins superfamily protein; FUNCTIONS IN: manganese ion binding, nutrient reservoir activity	1.6E-04	5.62	5.48	0.03	0.02
AT3G17060		Pectin lyase-like superfamily protein; FUNCTIONS IN: pectinesterase activity	2.4E-04	4.95	4.86	0.02	0.01
AT1G75335		Unknown protein	2.6E-04	4.97	5.08	0.02	0.02
AT5G07410		Pectin lyase-like superfamily protein; FUNCTIONS IN: pectinesterase activity	3.7E-04	5.37	5.22	0.03	0.03
AT2G47030	(VGDH1)	VGDH1; FUNCTIONS IN: enzyme inhibitor activity, pectinesterase activity	4.2E-04	5.04	4.97	0.02	0.02
AT3G28830		Protein of unknown function	4.4E-04	5.05	4.96	0.02	0.01
AT1G61566	RALF-LIKE 9 (RALFL9)	Member of a diversely expressed predicted peptide family showing sequence similarity to tobacco Rapid Alkalinization Factor (RALF)	5.3E-04	5.73	5.48	0.05	0.04
AT3G62710		Glycosyl hydrolase family protein; FUNCTIONS IN: xylan 1,4-beta-xylosidase activity, hydrolase activity, hydrolyzing O-glycosyl compounds	6.7E-04	5.23	5.14	0.02	0.01
AT5G48140		Pectin lyase-like superfamily protein; FUNCTIONS IN: polygalacturonase activity	7.2E-04	4.88	4.80	0.02	0.01
AT5G50030		Plant invertase/pectin methylesterase inhibitor superfamily protein	7.5E-04	5.40	5.27	0.03	0.02
AT4G18596		Pollen Ole e 1 allergen and extensin family protein	8.5E-04	5.69	5.57	0.02	0.02
AT1G08310		Alpha/beta-Hydrolases superfamily protein	9.0E-04	4.89	4.93	0.01	0.01

A subset of the most differentially expressed genes was used to visualize the affected molecular processes using the MapMan Arabidopsis-seed map (Figure 6.2) (Chapter 4, Joosen *et al.* 2011). This shows clearly that 6 hours of imbibition is enough to

induce genes involved in glycolysis, TCA, cell cycle, DNA synthesis, RNA transcription or translation and GA anabolism. Genes coding for LEA proteins, starch and ABA anabolism were downregulated.

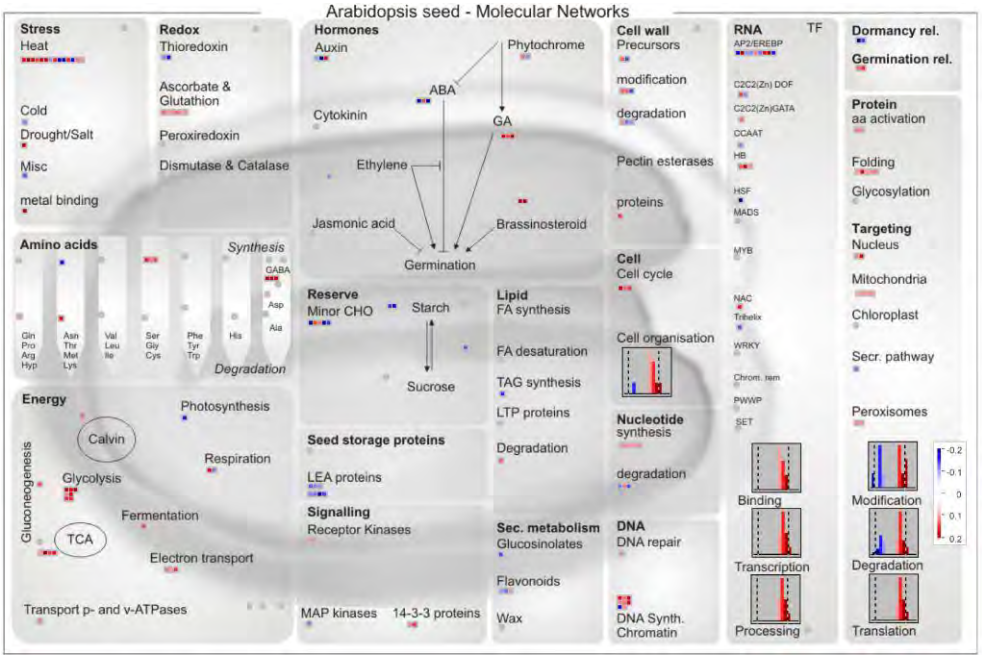


Figure 6.2: MapMan Seed Molecular Networks map showing differences in transcript levels (colored squares) between dry after-ripened seeds and 6 hours imbibed seeds. Ratios are used to express differences (red = upregulated in 6H imbibed seeds, blue = downregulated in 6H imbibed seeds). Only ratios <-0.1 or >0.1 with P-values < 1.7 10^{e-7} are presented.

Comparing the transcriptomes of dry seeds and seeds at the point of radicle protrusion resulted in the detection of 13058 genes with differential expression (Supporting Information, S6.4). Again, a subset of the most differentially expressed genes was used to visualize the affected molecular processes using the MapMan Arabidopsis-seed map (Figure 6.3). Next to the molecular processes that had already started 6 hours after imbibition a clear up-regulation of genes involved in cell wall modification and degradation, photosynthesis, Calvin cycle and fatty acid (FA) synthesis was observed. Genes encoding for seed storage proteins and TAG synthesis were downregulated. The affected molecular processes are in agreement with the expected mechanisms that initiate the transition from a dry dormant seed, which is equipped for optimal survival and storage of reserves, towards an imbibed seed, in which energy needed for germination is released and which prepares for autotrophic growth.

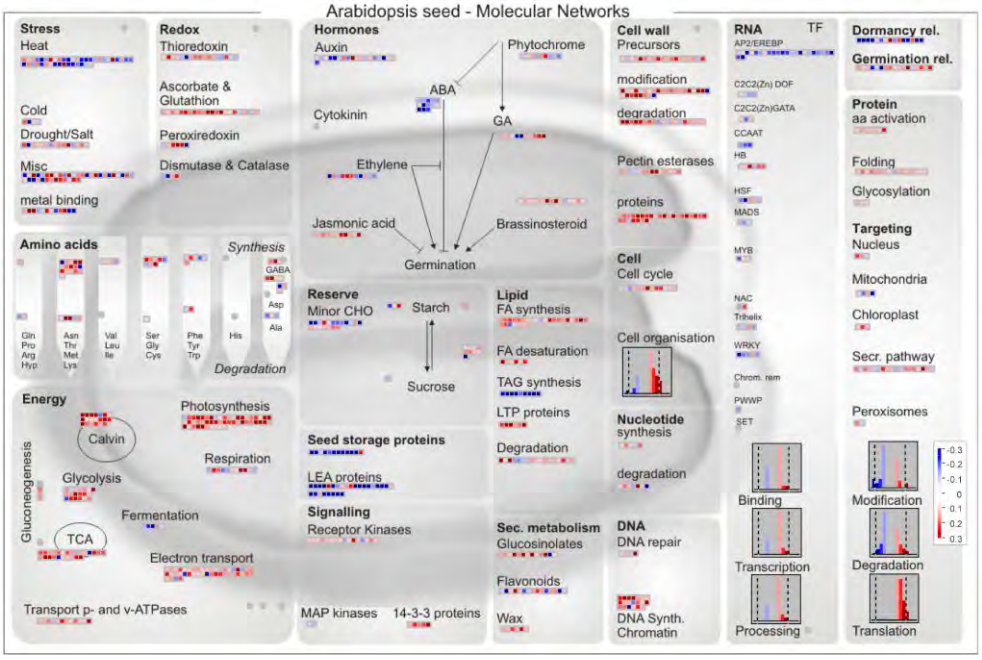


Figure 6.3: MapMan Seed Molecular Networks map showing differences in transcript levels (colored squares) between dry seed and seeds at the point of radicle protrusion (RP). Ratios are used to express differences (red = upregulated in RP, blue = downregulated in RP). Only ratios <-0.1 or >0.1 with P-values < $1.7\ 10^{e-7}$ are presented.

Genetic variation

Capturing genetic variation in a genetical genomics experiment combined with different developmental stages requires an alternative QTL mapping method. Linear models are well suited to build a variance-covariance model (VCOV) using the following formula:

$$y_i = \beta_0 + \beta_1 e_i + \beta_2 g_i + \beta_3 e_i{:}g_i + \varepsilon_i$$

where y_i is the i^{th} observation of gene expression, variable g_i is the genotype, e_i is a vector with developmental stages, and $e_i{:}g_i$ the interaction term. The values β_j represent parameters to be estimated, and ε_i is the error term. The simplified description Y = E + G + G:E + ε of this linear model will be used henceforward. Analysis is performed with the open source R-statistics program (http://www.r-project.org) using the R-scripts described in chapter 5 of this thesis. Separate likelihood estimates (–log10 probability, henceforth LOD scores) are generated for the environmental (E), genetic (G) and genetic x environmental (G:E) effects (Supporting Information, S6.5). In this study's experimental setup the environmental variation is defined as variation observed between the four developmental stages (PD, AR, 6H and RP). The significance threshold is determined by a rather

conservative Bonferroni correction at P=0.05 for 30.000 tests and equals LOD 5.8. In total 2006 eQTLs were detected in the genetic (G) component, representing 1990 genes (Figure 6.4A). At most, 2 QTLs were found for an individual gene. For 517 genes a strong interaction between the genetic x environmental (G:E) variation was observed, which resulted in the detection of 529 eQTLs (Figure 6.4B). Genes with a strong genetic effect in a specific environment will often result in QTLs in both the G and G:E component. However, 250 G:E specific eQTLs were detected. From the total of 29304 profiled genes, 15310 showed significant variation between the tested developmental stages (E).

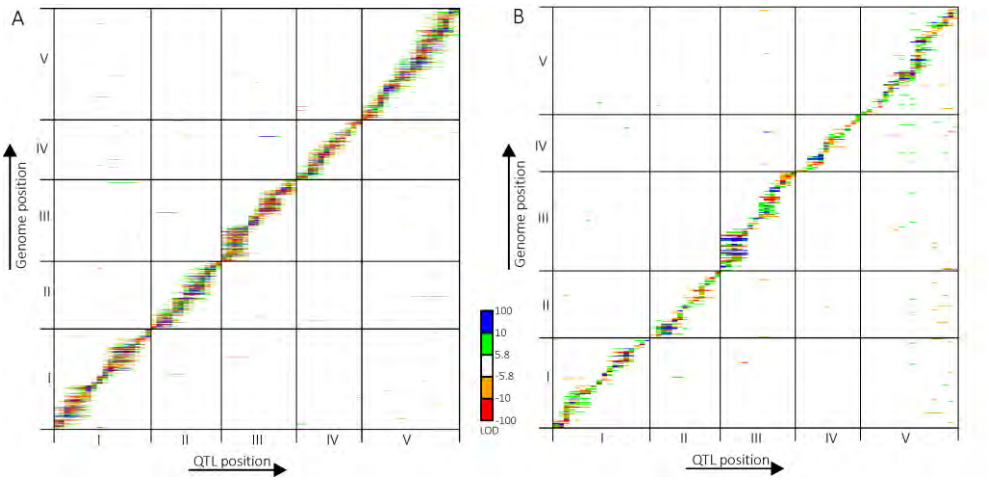


Figure 6.4: Distribution of mapped genes versus the position of their accompanying eQTL. Positions of detected eQTLs are plotted against the position of the gene for which that eQTL was found. QTLs were multiplied by the sign of effect resulting in a color distribution ranging from red (larger effect in Bay-0) to blue (larger effect in Sha). Chromosomal borders are depicted as horizontal and vertical lines. Panel A represents genes with genetic variation (G); panel B represents genes with genetic x environmental variation (G:E).

Expression QTLs in the fully sequenced Arabidopsis can be classified according to the position of the causal polymorphisms. Local eQTLs can be the result of closely linked *trans*-acting factors but in the majority of cases result from *cis*-regulatory variation in the genes under study (Chapter 1, Joosen *et al.* 2009). By definition eQTLs acting *in cis* affect transcription initiation, rate and/or transcript stability in an allele-specific manner. In addition, *cis*-regulated genes might encode regulators affecting the expression of downstream target genes *in trans* (Rockman and Kruglyak 2006). Because of the multiplicity of regulators and the often-observed epistasis between them, each *trans* eQTL can have a relatively small effect. As a result the detected number of *trans* eQTLs relative to the number of *cis* eQTLs drops when the stringency for detection is increased (Doss *et al.* 2005). The term *cis* eQTLs should be used with care because eQTL support intervals may contain multiple genes and as a result it can overlap with local-*trans* eQTLs. We therefore prefer using the terms local and distant eQTLs. Local eQTLs were defined by comparing eQTL locations with genome positions allowing an interval of 2.4 mega bases (approx. 10

cM). Genes with eQTLs outside this interval are classified as distant eQTLs. Of the 2006 eQTLs that were detected to have genetic (G) variation, 1809 were classified as local and 197 as distant. Accordingly, 447 local and 82 distant eQTLs were found for genes with genetic x environmental (G:E) variation.

Visualizing the eQTL network

Applying the generalized genetical genomics approach on 4 developmental stages combined with full genome transcription profiling results in a large and highly information dense dataset. Efficient visualization can be of great help to detect patterns and to query the data. Here, we used the program Cytoscape, which is an open source platform for complex network analysis and visualization, to create a marker-trait network. Within the R analysis procedure a peak-detection is performed to find the significant eQTL positions; this information was subsequently used to connect genes and markers. Network nodes indicate either genetic markers or genes which are connected by edges which represent the LOD score and the direction of the detected eQTL. Further, a Spearman correlation (cutoff 0.9) has been applied to detect genes which are co-expressed over both the 4 developmental stages and the various genotypes. In total, co-expression was determined over 180 microarrays (164 RIL + 16 parent hybridizations) and 430 gene pairs showed correlation above the threshold of 0.9 (Figure 6.5). Additional attributes, such as gene description, number of probes, GO-annotation, ratio between dry and imbibed seed expression, and transcription family category can be added to further describe the data. A large advantage of loading a QTL network in Cytoscape is the dynamic nature of the program (Supporting Information, S6.6). It allows ordering, filtering and selection in all directions which is of great help to query the data.

Seed development, dormancy and germination are regulated by many genes and several of them have been identified and characterized by reverse genetics approaches. Using the AmiGo browser (http://amigo.geneontology.org) a list of 211 genes with a known function in seeds was extracted. This list of a priori candidate genes was compared against our dataset and 24 genes showed significant expression variation between Bay-0 and Sha (Figure 6.6A). Expanding the selection by adding genes that are co-expressed with the 24 selected candidate genes (Figure 6.6B) resulted in 34 genes. Interestingly, we identified *DOG1*, the major regulator of dormancy, showing a local eQTL with strong genetic x environment interaction on chromosome V at marker MSAT518662 overlapping the *DOG1* QTL for seed dormancy. Six genes are tightly co-expressed with *DOG1* in our dataset and thus might indicate a shared functional pathway.

Although a detailed discussion for all identified candidate genes would transcend the scope of this paper we would like to elaborate on one very interesting candidate: *MIPS2* (AT2G22240).

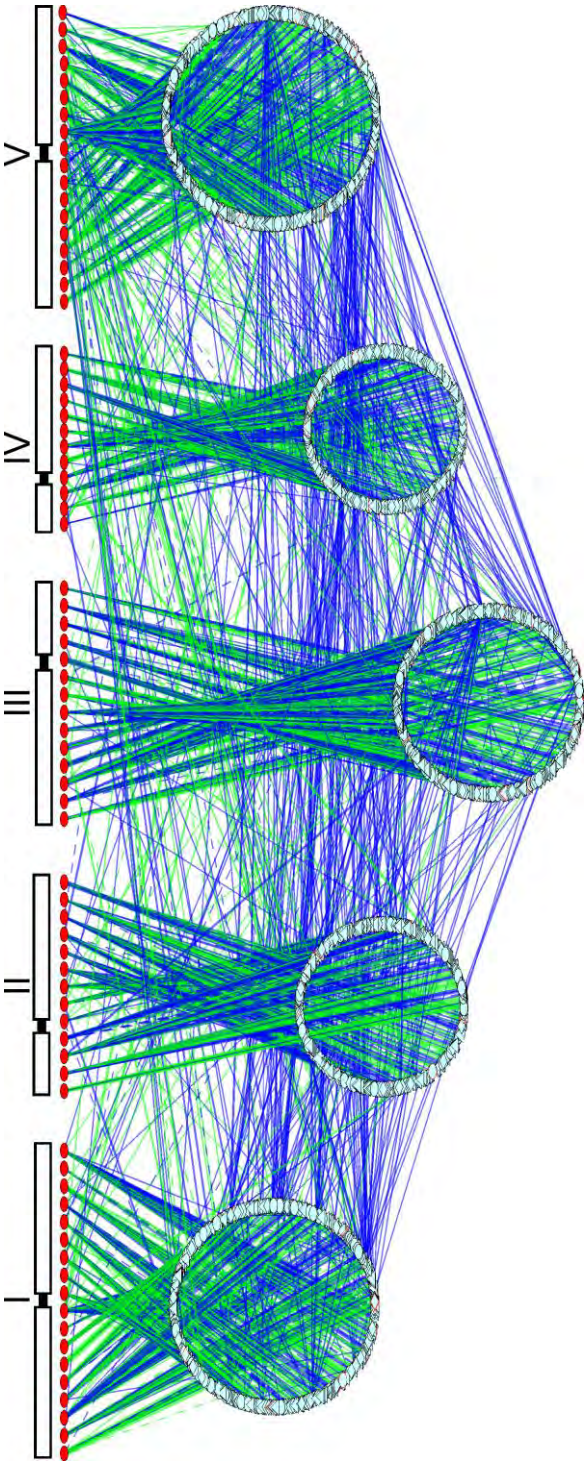
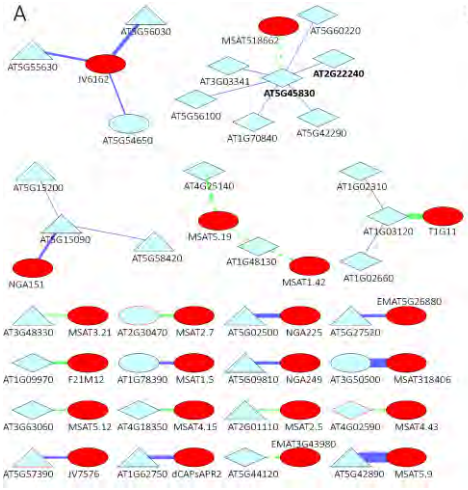


Figure 6.5: Cytoscape marker-trait network. Significant QTLs (>LOD5.8) are indicated by a connection between genes (as light blue nodes in 5 large circles representing the chromosomes) and markers (red nodes). Markers are sorted along the 5 chromosomes. Co-expressed genes are indicated by connections between genes. Edge colors indicate either the direction of the QTL (green = larger effect in Bay-0, blue = larger effect in Sha) or the spearman correlation coefficient (negative values = green, positive values = blue). The LOD-score is represented by the line width. Solid lines represent genetic (G) QTLs while genetic x environment interactions (G:E) are indicated by dashed lines.

This myo-inositol-1-phosphate synthase, catalyzes the rate limiting step in the de-novo synthesis of myo-inositol (Donahue *et al.* 2010). An eQTL for *MIPS2* was found on chromosome II, marker 2.36 and overlaps an mQTL found for myo-inositol concentration in the same population (this thesis, Chapter 5). MIPS proteins localize in the cytosol of Arabidopsis seed endosperm during seed maturation (Mitsuhashi *et al.* 2008). Myo-inositol-6-phosphate is the predominant form of phosphorus found in seeds and is hydrolyzed into myo-inositol and inorganic phosphates (Pi) by phytase during imbibition and subsequent seedling growth (Loewus and Murthy 2000). Further, it has been shown to act as a co-factor for auxin binding to the TIR1 auxin receptor (Tan *et al.* 2007). MIPS proteins are suggested to fulfill a role for auxin-regulated embryogenesis and show embryo-lethality in double and triple mutants (Luo *et al.* 2011). Together this suggests a possible role for the detected *MIPS2* gene in the process of seed germination but so far, a detailed analysis of germination characteristics including determination of dormancy is lacking for MIPS mutants. Our observation may open new research strategies to study the role of *MIPS* genes during seed germination or dormancy including a possible interaction with the *DOG1* gene.



B

ID	Symbol	GO-category
AT4G02590	UNE12	GO:0009567 : double fertilization forming a zygote and endosperm
AT2G01110	APG2	GO:0009567 : double fertilization forming a zygote and endosperm
AT1G78390	NCED9	GO:0009688 : abscisic acid biosynthetic process
AT4G18350	NCED2	GO:0009688 : abscisic acid biosynthetic process
AT5G09810	ACT7	GO:0009845 : seed germination
AT5G42890	SCP2	GO:0009845 : seed germination
AT1G62750	SCO1	GO:0009845 : seed germination
AT5G57390	AIL5	GO:0009845 : seed germination
AT1G03120	RAB28	GO:0009845 : seed germination
AT1G09970	LRR XI-23	GO:0009845 : seed germination
AT3G48330	PIMT1	GO:0009845 : seed germination
AT5G54650	Fh5	GO:0009960 : endosperm development
AT3G50500	SNRK2.2	GO:0010029 : regulation of seed germination
AT5G15090	VDAC3	GO:0010029 : regulation of seed germination
AT3G63060	EDL3	GO:0010029 : regulation of seed germination
AT5G55630	ATKCO1	GO:0010029 : regulation of seed germination
AT5G02500	HSC70-1	GO:0010187 : negative regulation of seed germination
AT5G56030	HSP81-2	GO:0010187 : negative regulation of seed germination
AT1G48130	PER1	GO:0010231 : maintenance of seed dormancy
AT4G25140	OLEO1	GO:0010344 : seed oilbody biogenesis
AT5G44120	CRA1	GO:0010431 : seed maturation
AT5G27520	PNC2	GO:0090351 : seedling development
AT5G45830	DOG1	GO:2000033 : regulation of seed dormancy
AT2G30470	HSI2	GO:2000034 : regulation of seed maturation

Figure 6.6: Cytoscape marker-trait network showing a selection of *a priori* seed germination related candidate genes with significant variation in expression between Bay-0 and Sha and their associated QTL positions. A second order selection of the candidate genes including co-expressed genes (Spearman correlation > 0.9) is shown in panel (A). Gene-node shape indicate whether the gene is up or down-regulated upon seed imbibition (up=triangle, down=diamond, no change = ellipse). Co-expressed genes are indicated by connections between genes. Edge colors indicate either the direction of the QTL (green = larger effect in Bay-0, blue = larger effect in Sha) or the Spearman correlation coefficient (negative values = green, positive values = blue). The LOD-score is represented by the line width. Solid lines represent genetic (G) QTLs while genetic x environment interactions (G:E) are indicated by dashed lines, markers are indicated as red ellipses. Panel (B) shows a list of the selected genes with their GO category.

A polymorphism in a major regulator might cause expression variation in a whole cascade of downstream genes. The presented marker-trait network can be used to extract

all eQTLs at a certain genome interval followed by gene ontology over-representation analysis using the Cytoscape BINGO plugin. For example, after selecting all eQTLs at marker MSAT5.59 a clear over-representation was observed for genes involved in light response and photosynthesis (Table 6.2). The Cytoscape network can be used to divide genes according to their eQTL position (Figure 6.7). Four genes with a local eQTL and 7 genes with a distant eQTL were detected. All seven genes with a distant eQTL were up regulated upon seed imbibition and showed a G:E eQTL affected at the Sha allele. Two genes with a local eQTL showed the same characteristics and might therefore be considered as candidate cis-regulators of the observed molecular response. In this example, a large local eQTL with similar characteristics compared to the identified distant eQTLs is found for AT5G38430; which encodes a member of the rubisco small subunit (*RBCS1B*) multigene family (Table 6.3). Rubisco plays a crucial role in carbon fixation and is often the rate limiting step for photosynthesis. Photosynthesis is restarted prior to seed germination and differences in the rate of initiation might have large influence on seedling establishment. Expression variation in *RBCS1B* can therefore explain the observed effect on genes involved in light response and photosynthesis. It should be noted that a *cis*-regulator not necessarily needs to have an expression difference or a similar expression profile. Both examples illustrate the power of the generalized genetical genomics approach combined with powerful filtering and sorting capabilities of a marker-trait network.

Table 6.2: Over-represented GO categories between genes with an eQTL at marker 5.59. The P-value is based on a hypergeometric test on the Arabidopsis GO biological process annotation. Cluster frequency indicates the number of genes within each GO sub category from the selected set of genes. Total frequency indicates the number of genes within each GO sub category from the total genome.

GO-ID	Description	P-value	cluster freq	total freq	genes
9637	response to blue light	2.59E-06	4/43 9.3%	50/22304 0.2%	AT4G10340, AT2G30520 AT5G38150, AT5G38430
10114	response to red light	3.53E-06	4/43 9.3%	54/22304 0.2%	AT4G10340, AT4G36880 AT4G14690, AT5G38430
9628	response to abiotic stimulus	5.87E-05	10/43 23.2%	1168/22304 5.2%	AT4G10340, AT5G20250 AT5G65020, AT2G30520 AT4G36880, AT4G14690 AT5G38470, AT5G38150 AT5G38430, AT5G43060
10218	response to far red light	6.76E-05	3/43 6.9%	41/22304 0.1%	AT4G10340, AT4G14690 AT5G38430
9416	response to light stimulus	2.24E-04	6/43 13.9%	455/22304 2.0%	AT4G10340, AT2G30520 AT4G36880, AT4G14690 AT5G38150, AT5G38430
9639	response to red or far red light	2.47E-04	4/43 9.3%	159/22304 0.7%	AT4G10340, AT4G36880 AT4G14690, AT5G38430
9314	response to radiation	2.70E-04	6/43 13.9%	471/22304 2.1%	AT4G10340, AT2G30520 AT4G36880, AT4G14690 AT5G38150, AT5G38430
15979	photosynthesis	1.35E-03	3/43 6.9%	113/22304 0.5%	AT4G10340, AT3G47470 AT5G54270

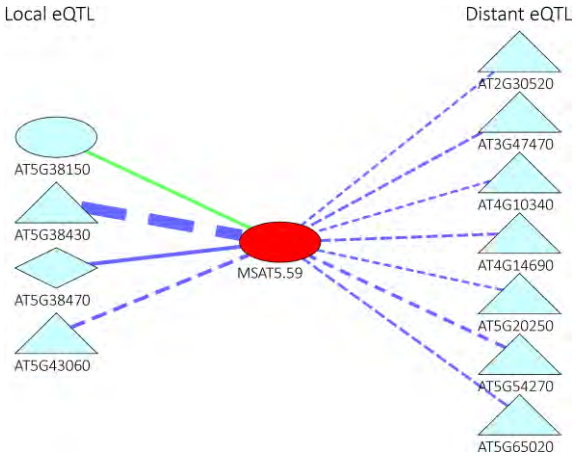


Figure 6.7: Marker-trait network showing over-represented genes for light response and photosynthesis at marker MSAT5.59. Genes are separated based on their QTL position (local = left, distant = right). Gene descriptions are presented in Table 6.3. Edge colors indicate the direction of the QTL (green = larger effect in Bay-0, blue = larger effect in Sha). The LOD-score is represented by the line width. Solid lines represent genetic (G) QTLs while genetic x environment interactions (G:E) are indicated by dashed lines.

Table 6.3: Gene descriptions for the genes shown in Figure 6.7.

ID	Symbol	Description
AT2G30520	RPT2	Light inducible root phototropism 2 encoding a signal transducer of the phototropic response in Arabidopsis
AT3G47470	CAB4	Encodes a chlorophyll a/b-binding protein that is more similar to the PSI Cab proteins than the PSII cab proteins
AT4G10340	LHCB5	Photosystem II encoding the light-harvesting chlorophyll a/b binding protein CP26 of the antenna system of the photosynthetic apparatus
AT4G14690	ELIP2	Encodes an early light-induced protein. ELIPs are thought not to be directly involved in the synthesis and assembly of specific photosynthetic complexes, but rather affect the biogenesis of all chlorophyll-binding complexes
AT5G20250	DIN10	Encodes a member of glycosyl hydrolase family 36. Expression is induced within 3 hours of dark treatment, in senescing leaves and treatment with exogenous photosynthesis inhibitor
AT5G38150	PMI15	Involved in chloroplast avoidance movement under high-light intensities
AT5G38430	RBCS1B	Ribulose bisphosphate carboxylase (small chain) family protein INVOLVED IN: carbon fixation, response to blue light, response to red light, response to far red light
AT5G38470	RAD23D	Encodes a member of the RADIATION SENSITIVE23 (RAD23) family. RAD23 proteins play an essential role in the cell cycle, morphology, and fertility of plants through their delivery of UPS substrates to the 26S proteasome
AT5G43060		Granulin repeat cysteine protease family protein
AT5G54270	LHCB3*1	Lhcb3 protein is a component of the main light harvesting chlorophyll a/b-protein complex of Photosystem II (LHC II)
AT5G65020	ANNAT2	Annexins are calcium binding proteins that are localized in the cytoplasm. They may be involved in the Golgi-mediated secretion of polysaccharides

In conclusion, we analyzed natural variation of gene expression in the Bay-0xSha RIL population in four developmental stages during seed germination. This approach is effective in determining differences in gene expression due to genetic and environmental variation. Summarizing the data with the use of a marker-trait network constructed with

Cytoscape offers great opportunities to mine and combine data with public available resources such as gene annotation.

Future prospects

In this study we assessed gene expression with the help of a tiling microarray. The described data represents overall gene expression levels acquired by averaging all anti-sense exon probes per gene. However, a more detailed analysis is required to analyze this data to its full potential. The used microarray, AtSNPtile, contains 1.7 million unique 25mer tiling probes in sense and antisense direction covering the non-repetitive part of the genome at 35 bp resolution. Currently, statistical procedures are in development which uses individual probe expression levels to investigate expression variation. Next to a more accurate estimation of transcript expression this will provide information about alternative splicing events and detect natural antisense transcription. Both alternative splicing and anti-sense transcription are important mechanisms for gene expression under both normal and stress conditions (Jin *et al.* 2008). Genetic variation for these events between Bay-0 and Sha might therefore provide important insight in such regulatory mechanisms.

The Bay-0xSha RIL population has been used in numerous studies to map QTL positions. Often, comparisons between studies are hampered because different developmental stages or plant growing conditions were used. With the data described in this paper we finalized a comprehensive study on seed traits at the morphological, germination potential (Joosen *et al.* 2012), metabolic (this thesis, Chapter 5) and gene expression level. Combining these datasets might provide insight in the molecular processes that underlie differences observed for seed germination potential between Bay-0 and Sha.

Materials and methods

Plant material

Seeds from the core population (164 lines) of the *Arabidopsis* Bay-0 x Sha recombinant inbred line population (Loudet *et al.* 2002) and heterogeneous inbred family (HIF) lines were obtained from the Versailles Biological Resource Centre for *Arabidopsis* (http://dbsgap.versailles.inra.fr/vnat). The population is mapped with 69 markers with an average distance between the markers of 6.1 cM (Loudet *et al.* 2002). Maternal plants were grown in a fully randomized setup and seeds from 4-7 plants/RIL were bulk harvested. Plants were grown on 4x4 cm rockwool plugs (MM40/40, Grodan B.V.) and watered with 1 g/l Hyponex fertilizer (NPK=7:6:19, http://www.hyponex.co.jp) in a climate chamber (20°C day, 18°C night) with 16 hours of light (35 W/m2) at a relative humidity of 70%.

Sample preparation

The Bay-0 x Sha RIL population was divided in four sub-populations optimized for the distribution of parental alleles using the R-procedure DesignGG (Li *et al.* 2009). The four sub-populations are used to represent four different developmental seed stages. Seeds were either stored at -80°C 1 week after harvest (primary dormant; PD) or after-ripened at room temperature and ambient relative humidity until maximum germination potential after 5 d of imbibition was reached (after-ripened; AR). After-ripened seeds were imbibed on water saturated filter paper at 20°C for 6H and quickly transferred to a dry filter paper for 1 minute to remove excess of water (6 hours imbibed seeds; 6H). Manual selection with help of a binocular was carried out to harvest seeds with the radicle at the point of protrusion (radicle protrusion; RP).

RNA isolation

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

Microarray analysis

RNA was processed for use on Affymetrix Arabidopsis SNPtile array (atSNPtilx520433) as described by the manufacturer. Briefly, 1 mg of total RNA was reverse transcribed using a T7-Oligo(dT) Promoter Primer in the first-strand cDNA synthesis reaction. Following RNase H-mediated second-strand cDNA synthesis, the double-stranded cDNA was purified and served as template in the subsequent in vitro transcription (IVT) reaction. The IVT reaction was carried out in the presence of T7 RNAPolymerase and a biotinylated nucleotide analog/ribonucleotide mix for complementary RNA (cRNA) amplification and biotin labeling. The biotinylated cRNA targets were then cleaned up, fragmented, and hybridized to the SNPtile array.

Data analysis

The hybridization data was extracted using an R-script with the help of an annotation-file based on TAIR9 annotation (http://aquilegia.uchicago.edu). Expression levels of anti-sense exon probes were averaged for each annotated gene. Data was normalized in R using quantile normalization. QTL analysis has been performed using the R-procedure described in this thesis, Chapter 5. After an initial QTL analysis using the original 69 markers a selection of strong local eQTLs was used to improve the genetic map. In total 6 new markers were added and 67 missing alleles were imputed. This improved genetic map (Supporting Information, S6.7) was used for the presented eQTL analysis.

Supporting Information

Supporting information can be downloaded from an online storage located at www.wageningenseedlab.nl/thesis/rvljoosen/SI.

Table S6.1: Microarray hybridization sample list including the RIL division in four developmental stages

Table S6.2: Gene expression levels for all 180 microarray hybridizations

Table S6.3: Differential genes between dry seed and 6H imbibed seeds

Table S6.4: Differential genes between dry seed and seeds at radicle protrusion

Table S6.5: LOD scores for environmental, genetic and genetic x environmental variation

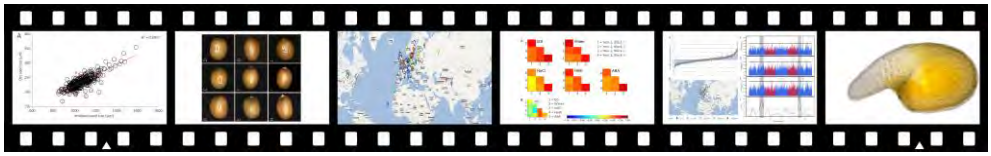
File S6.6: Cytoscape file for the model presented in Figure 6.5.

Table S6.7: Improved genetic map the Bay-0 x Sha RIL population

“They sowed the seed of an "if', but it didn't germinate.”

7 Comparing Genome Wide Association and Linkage analysis for seed traits

Joosen RVL, Willems LAJ, Lajo Morgan G, Kruijer W, Keurentjes JJB, Ligterink W, Hilhorst HWM
Submitted to Plos One



Abstract

Association mapping is rapidly becoming an important method to explore the genetic architecture of complex traits in plants. Over the past decades, a large amount of accessions has been collected for *Arabidopsis thaliana*. Ultra-high density genotyping followed by careful family-structure analysis has resulted in the assembly of a core-population consisting of 360 accessions. In this study we used this population to quantify seed size, germination on water, as well as the germination response to salt, heat and ABA. Experiments were replicated with seeds harvested in two successive years and a high level of heritable variation was observed. Five new natural seed coat mucilage mutants were discovered by analyzing the correlation between dry and imbibed seed size. Interesting correlations between the measured phenotypes and latitude or longitude positions were found. Such local adaptation is the most compelling source of evidence for natural selection during evolution. The combination of highly heritable phenotype measures and an optimally designed population increases the probability to find statistically significant SNP associations. However, no significant SNPs were detected when applying a Bonferroni multiple testing correction, which complicates the discovery of true associations. Due to the importance of seed germination in a plant's life cycle, a robust biological system is needed in which many loci with small additive effects may determine the final output. Thus, genome wide association can easily be underpowered to efficiently detect such relatively small effect loci. A comparison with traditional linkage mapping in a Bay-0 x Sha RIL population was made in an attempt to enforce the discovery of true associations. *De novo* candidate genes are listed and prioritized using available expression and annotation data.

Introduction

Accumulation of mutations during evolution is the powering force of adaptation to a large variety of environmental cues. In the early 60s plant biologists realized that a functional dissection of these mutations can be realized by creating structured populations (Thoday 1961). A particularly powerful approach is the creation of so-called recombinant inbred lines (RILs). These are made by crossing two homozygous founder lines which are adapted to distinct environments. The F1 will undergo recombination during meiosis which is fixed by several rounds of self-pollination. This results in an immortal population that can be tested for phenotypic variation in different developmental stages and under a large array of environmental perturbations. Subsequent linkage mapping allows identification of chromosomal regions, i.e. quantitative trait loci (QTL), which contain genetic variation important for the trait under investigation (reviewed by Doerge 2002). For complex and polygenic traits this often results in the detection of several QTLs with varying effect on the final phenotype. This approach has proven its power in numerous studies for a wide range of phenotypic traits and dozens of causal genes have subsequently been identified by a procedure called fine mapping (Alonso-Blanco *et al.* 2009). Fine mapping relies on narrowing the genomic region of a QTL by searching new recombinants within the linkage interval. This is a time consuming and laborious task which confines this classical approach for QTL mapping to a low throughput technique. Another disadvantage is the limited amount of genetic variation that can be analyzed when only two founder parents are used. Advanced crossings using multiple parental lines can partly overcome this limitation (Kover *et al.* 2009; Huang *et al.* 2011).

Linkage mapping requires controlled crosses which are not desired to apply on humans. Therefore an alternative procedure called genome wide association (GWA) was developed for human genetics. It requires a large set of genetically variable individuals and a detailed inventory of polymorphisms that are inherited and shuffled by recombination (Hirschhorn and Daly 2005). In this type of studies the genomic regions that are associated with a phenotype are often very small because the accumulated genetic variation results in small genomic regions that are independent of each other. This phenomenon is also referred to as 'fast decay of linkage disequilibrium'. The level of linkage disequilibrium is not only influenced by genetic linkage but also by the rate of recombination, rate of mutation, genetic drift, non-random mating and population structure. Individual single nucleotide polymorphisms (SNP) or multimarker combinations (haplotypes) can be used to identify the genetic variation within a genomic region. GWA studies require a high density of SNP variants to survey the size of genome regions that are in linkage disequilibrium. Another important factor to consider is the population structure (also referred to as population stratification). Population structure can lead to an over-representation of subgroups within the population that differ in trait prevalence. This phenomenon usually causes confounding factors which results in the association of markers with phenotypic

effects without being truly related with the causes of the phenotype. Methods to correct for population structure are reviewed in Price *et al.* (2010).

Already in 1937, Friedrich Laibach began collecting local ecotypes of the wild crucifer *Arabidopsis thaliana,* and his initiative was continued by many other plant scientists. This resulted in a large set of ecotypes which were globally collected at natural sites. A high density genotyping microarray (AtSNPtile) containing 250.000 SNP probes was developed (Zhang and Borevitz 2009) and used to genotype 1307 lines (https://cynin.gmi.oeaw.ac.at/home/resources/atpolydb). Since the genome size of *Arabidopsis* is around 120 megabases this resulted in approximately one SNP in every 500 bp. The first study in *Arabidopsis* showing the feasibility of GWA mapping was performed on 191 ecotypes and describes 107 phenotypes (Atwell *et al.* 2010). One of the important observations in this study was related to the required sample size. In contrast to human studies, the environmental conditions during growth and phenotyping assays can be well controlled and replicated in plants. This leads to high trait heritability and thus allows GWA studies on a much smaller number of individuals compared to GWA studies in humans. However, a strong and complex population structure was detected which can be expected to generate a high rate of false positive associations (Atwell *et al.* 2010). This confounding factor was corrected by using a mixed-model approach including the kinship between the natural accessions. The success of such strategy is hard to evaluate, but it was shown that some highly expected associations (e.g. the FLC gene for flowering time) could still be detected after correction for the population structure. It was suggested that both a larger sample size and a reduced complexity of the population structure would increase the power to detect true associations. Global population structure was therefore estimated within 5707 ecotypes with a set of 137 SNPs (Platt *et al.* 2010) and this data was used to define an *Arabidopsis* population consisting of 360 ecotypes with maximized diversity and elimination of unusually close relationships (Li *et al.* 2010). This *Arabidopsis* core 360 HapMap population was used in the study presented in this paper.

In general, genome wide association has become a promising tool to dissect natural genetic variation but it remains hard to distinguish true from false positives and detect rare alleles. However, in plants we can benefit from a combination of genome wide association and traditional linkage mapping (Brachi *et al.* 2010). GWA benefits from a high level of genetic variation resulting in high resolution but it has limiting power to detect rare alleles whereas linkage mapping in a RIL population has limited genetic variation, low resolution but high power to detect QTL. Combining the two approaches can be powerful to validate associations, increase power to detect rare alleles and reduce the number of candidate genes. In our study we compared GWA in the core 360 HapMap population with traditional linkage mapping in the Bay-0 x Sha RIL population.

The change from a seed to seedling marks an important phase transition with evolutionary importance (Barua *et al.* 2011). In order to optimize the chance of successful reproduction, seed germination must be timed accurately. This timing is determined by seed dormancy which is relying on signals received from the environment. Also genetic

differences are detected for dormancy between several *Arabidopsis* ecotypes with the help of QTL analysis (Bentsink *et al.* 2010). Not only the timing of germination, but also the adaptation to particular environments is an important selection criterion. Therefore, the control of seed germination is a consequence of complex genome by environment interactions. Environmental factors such as light, temperature, nutrient availability and the duration of seed after-ripening generate integrated signals that interact with endogenous factors (Bewley 1997). These include the control of genes responsible for signaling pathways and metabolism of the hormones ABA and GA as well as cell wall weakening enzymes, the circadian clock and phytochrome-interacting factors, energy metabolism, reactive oxygen scavenging and many others (Penfield and King 2009). The role of many of these genes has not been elucidated thus far and additional research is required to improve our understanding of seed germination. Both traditional linkage mapping and genome wide association analysis are methods well suited to investigate the evolutionary important and complex adaptations important for seed germination and we believe that there is a clear advantage of using the combination of these methods for dissecting the molecular pathways influencing this trait. In this paper we describe the analysis of seed size and seed germination under optimal conditions, under salt (125 mM NaCl) and heat (30°C) stress. ABA is an important regulator of stress responses and germination. Therefore, we also tested the natural variation for sensitivity of seed germination for externally applied ABA.

Results and Discussion

Phenotyping the HapMap population for seed characteristics

In *Arabidopsis* seed size differences can be the result of changes in e.g. seed coat thickness or embryo/cotyledon size. It might therefore be expected that a clear correlation exists between seed germination characteristics and seed size. However, several studies showed that this correlation is only observed for seedling growth and not present at the level of seed germination. The results of those studies can be biased because they were performed in recombinant inbred populations derived from only two genotypically different parents (Joosen *et al.* 2012; Kazmi *et al.* 2012; Khan *et al.* 2012). GWA offers the unique possibility to assess this correlation in a much broader genotype perspective. Seed size was determined with the help of the free image analysis tool imageJ. Pictures that were taken during germination could directly be used to determine seed size of imbibed seeds. To test whether imbibed seed size reflects initial dry seed size we compared the size of dry mature seeds to the size from 12 hours imbibed seeds (Figure 7.1A). Apart from several outliers this showed a good correlation (r^2 = 0.6917) for normal mucilage forming seeds (Figure 7.1B-D). Interestingly, a closer inspection of the outliers revealed that those seeds did not form a proper mucilage layer after imbibition (Figure 7.1E-J). A crippled mucilage formation may result in either slower imbibition leading to reduced seed swelling, or an imaging artifact caused by diffraction due the shiny mucilage layer surrounding the

imbibed seed. By using the outliers from the correlation between dry and imbibed seed size we were able to identify 6 natural mucilage mutants. Only one of them, Shahdara, has been described before as a natural mutant for the MUM2 protein (Macquet *et al.* 2007). The imbibed seed size measures were used for further genome wide association.

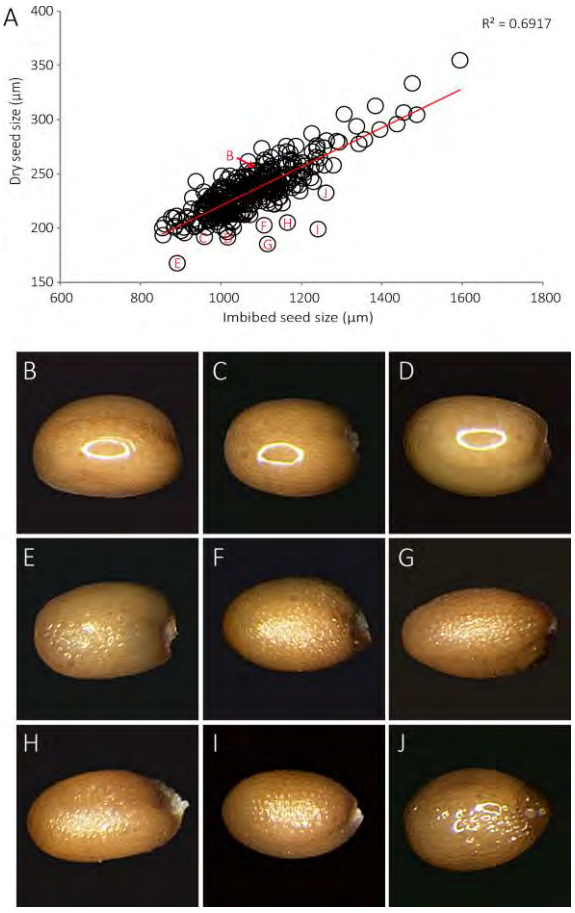


Figure 7.1: Natural mucilage mutant detection with outliers of dry vs. imbibed seed size correlation (A). Normal mucilage formation in B: Col-0 (CS76113), C: ALL1-3 (CS76090), D: LDV-34 (CS76162). Crippled mucilage formation in E: Lc-0 (CS76159), F: Eden-2 (CS76125), G: Sha (CS76227), H: Tad01 (CS76243), I: Var2-1 (CS76298), J: Lov-5 (CS76175).

Seed germination is the final output from a complex signaling cascade, which is heavily influenced by the environment. Freshly harvested Arabidopsis seeds are often not germinating when exposed to conditions optimal for germination. This phenomenon, called 'primary dormancy', will substantially interfere when measuring germination performance. Primary dormancy is slowly released in time by a process commonly referred to as 'after-ripening'. Another dormancy breaking treatment is the application of a period of 4 days at 4°C in the dark to imbibed seeds ('cold stratification'). In this study we wanted to test seed performance and reduce the effect of primary dormancy as much as possible. Therefore we combined both dormancy breaking treatments. First we tested germination on water at

20°C with continuous light, which is considered to be the optimal condition for *Arabidopsis* seed germination (Toorop *et al.* 2005). Seed germination was monitored for 5 days and cumulative germination data was obtained with the use of the Germinator setup (Joosen *et al.* 2010). The integration of the area under such cumulative germination curves (AUC) can efficiently summarize germination performance because it is affected by the rate, uniformity and maximum germination. To evaluate the ability of seeds to germinate under stress conditions we tested germination in the presence of 125 mM NaCl and germination at 30°C. Both salinity and heat tolerance can be regarded as important mechanisms for evolutionary adaptation and speciation. One of the plant hormones with major impact on seed germination and stress tolerance is abscisic acid (ABA). Applied ABA inhibits cell wall degrading enzymes often resulting in an incomplete protrusion of the radicle through the surrounding layers (Müller *et al.* 2006). Further, the proper establishment of seedlings is often inhibited. In agreement with the other germination measures only complete radicle protrusion through the endosperm layer was scored as germination. The response to any of the applied stresses is expressed as the difference in the AUC between germination on water at 20°C and the AUC of the stressed seeds. A z-score transformation was applied to standardize the comparison between the harvests from consecutive years and blocks. Correlation analysis (Figure 7.2A) shows the quality of the replications between the two blocks and two successive years. In the response to NaCl (block I from year 1) one experiment resulted in a low correlation with the other replicates. When comparing the correlation between the different experiments (Figure 7.2B) a clear resemblance can be observed between germination under both stress conditions. This might indicate that the signaling cascade under both NaCl and heat stress comprises similar components. No clear correlation could be observed between seed size and germination performance which also confirms the previous reported lack of correlation in a much broader genetic perspective.

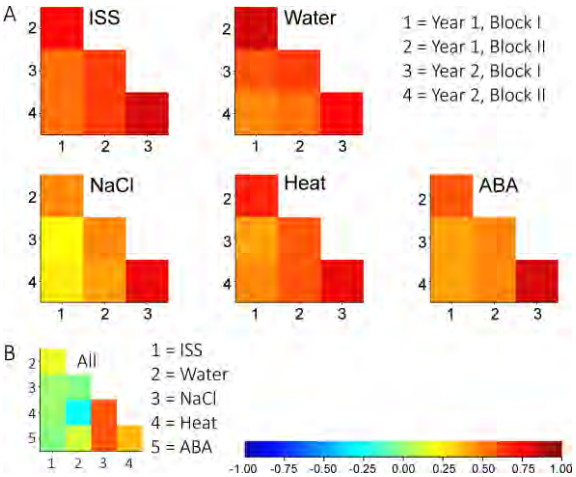


Figure 7.2: Correlation analysis of replicated phenotype measurements in the HapMap population. A) correlation between replicated experiments (block I+II year 1+2). B) Correlation between the average values for each experiment. The false color scale represents the Pearson correlation coefficient. ISS = imbibed seed size, Water = AUC of germination at 20°C on water, NaCl = AUC of germination on 125 mM NaCl, Heat = AUC of germination at 30°C, ABA = AUC of germination on 0.5 µM ABA.

Phenotype-location correlation

Local adaptation is the most convincing source of evidence for natural selection during evolution. The HapMap population used in this study provides very suitable material to study patterns in global distribution. Detailed information about the location of origin is available for all accessions and can be used for correlation analysis with the observed phenotypes. Here, we restricted the analysis to the correlation between phenotype and latitude or longitude.

Table 7.1: Linear regression results comparing phenotype against latitude and longitude. Significant P values ($p<0.001$) are indicated in bold. ISS = imbibed seed size, Water = AUC of germination at 20°C on water, deltaNaCl = AUC of germination on water – AUC of germination on 125 mM NaCl, deltaHeat = AUC of germination on water – AUC of germination at 30°C, deltaABA = AUC of germination on water - AUC of germination on 0.5 µM ABA.

	Latitude correlations			Longitude correlations		
Trait	**Reg.coeff**	**SE**	**P value**	**Reg.coeff**	**SE**	**P value**
ISS	-1.5072	0.3309	7.29e-06	-1.0563	1.9017	0.579
Water	-0.4666	0.3540	0.188	-4.2548	1.9682	0.0313
deltaNaCl	-1.5980	0.3771	2.9e-05	-10.2908	2.0875	1.28e-06
deltaHeat	-1.6274	0.3500	4.75e-06	-7.4671	1.9744	1.83e-4
deltaABA	-0.3186	0.3569	0.373	-4.7036	1.9789	0.018

Imbibed seed size, germination response to salt and germination response at high temperature show negative correlations with the latitude of the origin of the accessions. This indicates that accessions from the northern regions tend to have smaller seeds that withstand salt and heat stress better, as compared to the bigger seeds from the more southern regions. Germination on salt and high temperatures also show negative correlation with the longitude of the accessions origin, indicating that accessions from the eastern Eurasian regions withstand salt and heat stress better, as compared with the accessions from western Europe and US regions. For each trait a world map with positions of the accessions and a color coded phenotype ranking is presented in Figure 7.3B-Figure 7.7B and discussed in more detail hereafter. More elaborate correlation analysis involving e.g. climatological data or soil composition might lead to a better ecological interpretation but is a future perspective.

Genetic mapping

A complicating factor in GWA is the presence of population structure. Although the population used in this study was especially selected to contain a low level of confounding structure it should still be taken into account. To allow proper correction an identity-by-state (IBS) matrix was calculated using 214051 SNP markers. Such an IBS matrix facilitates quality control of genomic data, e.g. plants with high IBS values indicate close relatedness. An advantage for plant geneticists is the possibility to control environmental variation by replicating experiments under strictly regulated conditions. This allows an

optimal evaluation of the present genetic variation for the traits under study. Calculating the trait heritability provides a good measure for the fraction of phenotype variation that can be attributed to genetic variation. Here, the broad sense heritability was calculated by taking the IBS matrix into account using the calculations provided by Kang *et al.* (2010) and MacKenzie and Hackett (2012). Overall trait heritability (Table 7.2) was high, indicating that a substantial part of the observed variation could be attributed to genetic variation.

Table 7.2: Trait heritability scores for seed size and germination performance expressed by the deltaAUC, calculated from replicated experiments (two years with two blocks).

Trait	Heritability(H^2)
Imbibed seed size	0.673
Germination on Water	0.628
Germination response to 125 mM NaCl	0.431
Germination response to high temperature (30°C)	0.553
Germination response to 0.5 µM ABA	0.580
Seedling establishment on 0.5 µM ABA	0.651

These observed levels of heritability provide a good starting point for GWA. A modified version of the EMMAX procedure (Kang *et al.* 2010) was created using the open source statistics package R in combination with C++. Basically we used a mixed model that takes genetic similarity (using the IBS matrix) into account and which incorporates cofactors such as block or replicate effects in the analysis. Similar to EMMAX, p-values derived from our procedure might be affected by SNPs with rare alleles. Therefore a stringent cutoff of 10% minor allele frequency (MAF) has been used throughout the analysis as has been used in simulations by (Kang *et al.* 2008).

For each trait the genome wide association was calculated for both the year 1 and year 2 experiments separately as well as for the average of the two years. The observed variation between block and year replicates was taken into consideration. No significant SNPs were detected when applying a Bonferroni multiple testing correction, resulting in a threshold of –Log10 = 6.5 ($p>0.05$ / number of SNPs tested = 171.935). Therefore, only genomic positions with 2 or more SNPs with p-values above 4 (–log10) within an interval of 20 kb in the average between year 1 and 2 were considered for further analysis. Given the fact that average linkage disequilibrium decays at 10 kb in *Arabidopsis* (Clark *et al.* 2007) we took a conservative interval of 20 kb to search for associated candidate genes. The high density genotyping with 250.000 SNPs results in an average 1 SNP for every 500 bp and can thus be expected to reflect haplotype structures. Therefore we considered single associated SNPs within a 20 kb interval more likely to be false-positive compared to those instances were multiple SNPs within the interval were found to be associated. Nevertheless, it remains cumbersome to distinguish true from false associations in GWA mapping.

Brachi *et al.* (2010) have shown that a combination of traditional linkage mapping and association mapping can outperform each individual method. Recently, the Bay-0xSha

RIL population was extensively phenotyped for seed performance under a wide range of environmental conditions (Joosen *et al.* 2012). These results were compared with our GWA analysis by overlapping the 2LOD support intervals with the GWA results. All SNPs located in those intervals with a –log10 p-value > 4 in the average between year 1 and 2 were considered possible candidates.

Seed size

As shown in Figure 7.1A dry seed size correlated very well with imbibed seed size (r^2=0.6917) except for six outliers for which we observed a deformed mucilage phenotype. Seed size varied considerably (almost doubled from smallest to largest seed) and despite a large maternal effect on seed size the trait heritability reached a level of 0.673. Linear regression revealed a trend for imbibed seed size and latitude, indicating that seeds from the northern regions are smaller compared to seeds from the southern regions (Table 7.1). Figure 7.3B shows the global distribution of all accessions used in this study colored according to their rank in seed size. From this figure it is obvious that European accessions are overrepresented in the HapMap population which implicates that the North-South correlation with seed size is mainly based on European distribution. GWA (Figure 7.3C-E) was compared to QTL analysis in the Bay-0 x Sha RIL population (Figure 7.3F). However, in this population imbibed seed size yielded a single large QTL at the bottom of chromosome V. This is most probably caused by a mutation in the Sha allele of the *mum2* gene which is involved in the modification of mucilage preventing its expansion after hydration (Dean *et al.* 2007). Considering the good correlation between dry and imbibed seed sizes in the GWA panel we used QTL analysis of dry seed size in Bay-0 x Sha to compare with the imbibed seed size in the GWA analysis. Table 7.3 shows selected candidate genes according to the aforementioned criteria (2 or more SNPs with –log10 p-values above 4 within an interval of 20 kb in the average between year 1 and 2, or single SNPs with –log10 p-values above 4 within an interval of 20 kb in the average between year 1 and 2 when overlapping with a Bay-0 x Sha QTL interval).

Genes that influence seed size are expected to be expressed during seed development. A careful microdissection of developing seeds followed by transcriptome analysis was performed by Le *et al.* (2010) and can be queried via the Bio-Array Resource e-northern facility (Toufighi *et al.* 2005). Expression levels of this transcriptome analysis that reached a level > 200 in any of the tissues during seed development are marked with (+) in Table 7.3. This might allow a more precise selection of the candidate gene within each 20 kb interval. However, it should be noted that it only provides an overview of expression in the Columbia ecotype and might therefore lead to false negative interpretations for genes which lost their expression in Columbia. Three genes within the 20 kb interval on chromosome I exhibit expression during seed development (AT1G68580, AT1G68590 and AT1G68640). *PERIANTHIA* (*PAN*; AT1G68640) mutants are affected in their floral organ patterning and have a pentamerous pattern of 5 sepals, 5 petals 5 stamens, and 2 carpels

(Running and Meyerowitz 1996). This phenotype is not observed in the HapMap association panel and might thus indicate that a knock-down mutation for *PERANTHIA* is more common than complete knock-out mutations. *PERIANTHIA* has been shown to have overlapping functions with *ULTRAPETALA* and the *CLAVATA* signal transduction pathway in controlling shoot and floral meristem size and meristem determinacy (Fletcher 2001) and evolutionary diversification in this function cannot be ruled out.

One of the most plausible candidate genes for the region below the association peak on chromosome V is *ANGUSTIFOLIA* (*AN3*; AT5G28640). *ANGUSTIFOLIA* is a transcription coactivator which interacts with the *GROWTH-REGULATING FACTOR 1* (GRF1) transcription factor. Its function in the control of cell proliferation via cell cycle regulation has been well studied (Lee *et al.* 2009).

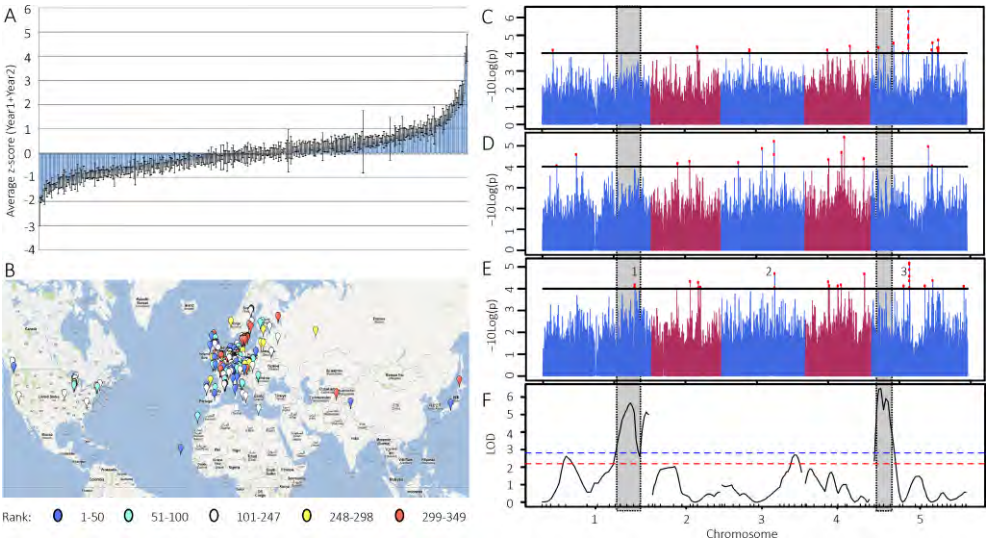


Figure 7.3: Analysis of imbibed seed size. Panel A shows the z-score adjusted distribution of the average imbibed seed size measured in year 1 and 2 sorted from small to large over all 349 HapMap accessions. Error bars indicate the standard error. Global distribution of the accessions (ranked from large to small z-score values with indicated color code; blue = big seeds, red = small seeds) is shown in panel B. Genome wide association corrected for the IBS population structure (Year 1: C, Year 2: D, Average Year 1 and Year 2: E). SNPs with –log10 p-values > 4 are marked with red dots. Selected peaks for further analysis are indicated with numbers in panel E. Linkage mapping results for dry seed size in the Bay-0 x Sha RIL population are shown in panel F. Significance thresholds, determined by permutation (n=1000) are indicated with dashed lines (Red=85%, Blue=95%). Positions shaded in grey visualize the overlap of genome positions determined by the 2LOD drop of significant QTLs in the Bay-0 x Sha population.

The *AN3* mutant has reduced cell numbers of lateral organs, such as leaves and flowers as well as cotyledons and overexpression results in 20-30% increase in leaf size (Horiguchi *et al.* 2005). *AN3* is well expressed during seed development but no effects on seed size were reported, to our knowledge. Confirmation of its influence on seed size can be studied using the available mutants and overexpression lines. With the ongoing sequence initiatives (www.1001genomes.org/) it will become feasible in the near future to evaluate neutral,

nonsense and missense mutations present in the candidate genes in all the ecotypes used for association. Further research is required to confirm the involvement in seed size regulation of the candidate genes described in Table 7.3.

Table 7.3: Candidate genes for variation in imbibed seed size selected from the GWA analysis. Peak numbers are shown in Figure 7.3E. All annotated (TAIR10) genes overlapping the selected 20 kb regions are presented. The column (Expr.) indicates whether expression reaches levels > 200 during seed development queried using the Bio-Array Resource e-northern facility (Toufighi *et al.* 2005).

Peak	ID	Symbol	Description	Expr.
1	AT1G68580		Agenet domain-containing protein / bromo-adjacent homology (BAH) domain-containing protein	+
1	AT1G68585		Unknown protein	-
1	AT1G68590		Ribosomal protein PSRP-3/Ycf65	+
1	AT1G68600		Aluminium activated malate transporter family protein	-
1	AT1G68610	(PCR11)	Target promoter of the male germline-specific transcription factor DUO1	-
1	AT1G68620		Alpha/beta-Hydrolases superfamily protein	-
1	AT1G68630		PLAC8 family protein	-
1	AT1G68640	(PAN)	Encodes bZIP-transcription factor. Mutant plants have extra floral organs. PAN is essential for AG activation in early flowers of short-day-grown plants	+
2	AT3G42800		unknown protein	-
3	AT5G28640	(AN3)	Encodes a protein with similarity to mammalian transcriptional coactivator, the gene is also shown to be involved in cell proliferation during leaf and flower development. Loss of function mutations have narrow, pointed leaves and narrow floral organs. AN3 interacts with members of the growth regulating factor (GRF) family of transcription factors	+
3	AT5G28646	(WVD2)	Encodes a novel protein. The wvd2 gain-of-function mutant has impaired cell expansion and root waving, and changed root skewing	n.d.
3	AT5G28650	(WRKY74)	Member of WRKY Transcription Factor- group II-d	-

n.d. = not determined because probeset is not available for this gene on the AtH1 microarray

Seed germination

Germination was evaluated to test the optimal performance of a seed batch. Seed germination may be affected substantially by the primary dormancy level of a seed batch. To eliminate this effect as much as possible, seeds were after-ripened and perceived a cold stratification period. Seeds were germinated on excess water at 20°C with continuous light which can generally be considered optimal conditions for *Arabidopsis* germination (Toorop *et al.* 2005). It should be noted that these optimal conditions can vary considerably for the large range of genotypes tested. Only 23 accessions showed significantly less than 90% maximum germination after 5 days in the four replicates. However, large differences in both rate and uniformity of germination were observed. To take these factors into account the area under the germination curve (AUC) (Joosen *et al.* 2010) was used as a measure for seed performance (Figure 7.4). Heritability reached a level of 0.628 (Table 7.2). No clear trend could be observed in the global distribution of the accessions (Table 7.1, Figure 7.4B). QTL analysis of the AUC of seed germination from cold stratified after-ripened seeds from

the Bay-0 x Sha RIL population was used to compare with the GWA results (Figure 7.4F). Table 7.4 shows the selected candidate genes according to the aforementioned criteria.

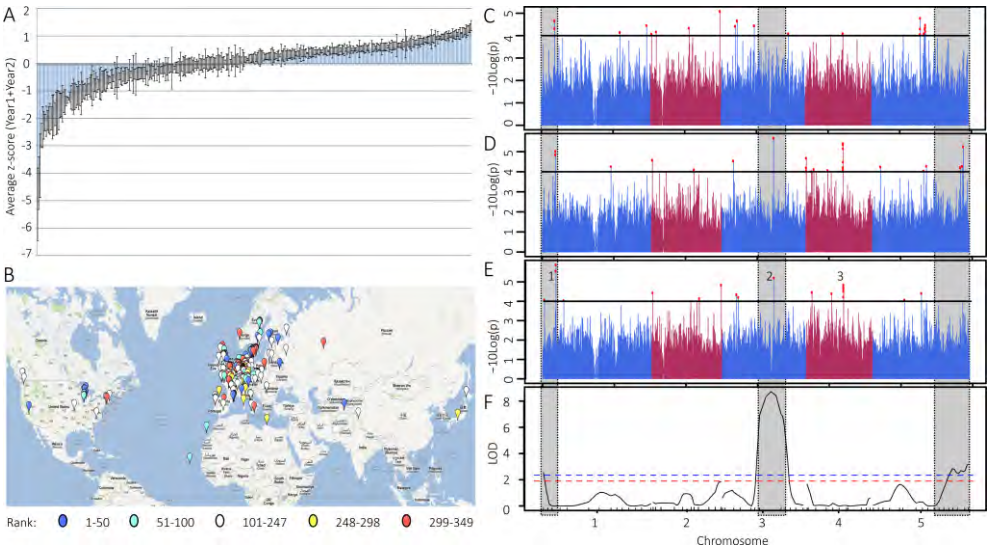


Figure 7.4: Analysis of seed germination on water at 20°C after 4 days of stratification. Panel A shows the z-score adjusted distribution of the average area under the germination curve (AUC) in year 1 and year 2 sorted from small to large over all 349 HapMap accessions. Error bars indicate the standard error. Global distribution of the accessions (ranked from large to small z-score values with indicated color code; blue = good germination, red = bad germination) is shown in panel B. Genome wide association corrected for the IBS population structure (Year 1: C, Year 2: D, Average Year 1 and Year 2: E). SNPs with –log10 p-values > 4 are marked with red dots. Selected peaks for further analysis are indicated with numbers in panel E. Linkage mapping results for seed germination on water at 20°C after 4 days of stratification in the Bay-0 x Sha RIL population are shown in panel F. Significance thresholds, determined by permutation (n=1000) are indicated with dashed lines (Red=85%, Blue=95%). Positions shaded in grey visualize the overlap of genome positions determined by the 2LOD drop of significant QTLs in the Bay-0 x Sha population.

STRUBBELIG (*SUB*; AT1G11130, Table 7.4) also known as *SCRAMBLED* encodes a leucine-rich repeat transmembrane receptor-like kinase that is required for floral organ shape, the development of the outer integument of ovules, stem development and specification of epidermal root hairs (Yadav *et al.* 2008). The ovule phenotype of *SUB* mutants has been described in detail by Chevalier *et al.* (2005) who showed that it is variable, ranging from severely affected to normal and fertile ovules, and sensitive to ecotype background. A germination related phenotype has not been described yet but can be expected according to the ovule development phenotype. Expression analysis via the Bio-Array Resource e-northern facility (Toufighi *et al.* 2005) shows that *NLM1* (AT4G19030) is highly expressed upon seed imbibition and shows light responsive expression. Further, it has been detected in a screen for *PIF3* targets (Feng *et al.* 2008). Together this implies a role for *NLM1* in the light responsive *PIF3/DELLA* coordinated GA-signaling with obvious

consequences for seed germination. More research is required to confirm the involvement in seed germination of the candidate genes described in Table 7.4.

Table 7.4: Candidate genes for variation in seed germination selected from the GWA analysis. Peak numbers are shown in Figure 7.4E. All annotated (TAIR10) genes overlapping the selected 20 kb region are presented. The column (Expr.) indicates whether expression reaches levels > 200 during seed germination queried using the Bio-Array Resource e-northern facility (Toufighi *et al.* 2005).

Peak	ID	Symbol	Description	Expr.
1	AT1G11120		Unknown protein	n.d.
1	AT1G11125		Unknown protein	-
1	AT1G11130	(SUB)	Encodes a receptor-like kinase. Regulates expression of GLABRA2, CAPRICE, WEREWOLF, and ENHANCER OF GLABRA	-
1	AT1G11145		Protein of unknown function (DUF674)	n.d.
2	AT3G42310		Unknown protein	-
3	AT4G19000	(IWS2)	The C-terminal portion of this protein has homology to the C-termini of the IWS1 (Interacts With Spt6) proteins found in yeast and humans	-
3	AT4G19003	(VPS25)	VPS25	n.d.
3	AT4G19006		Proteasome component (PCI) domain protein	n.d.
3	AT4G19010		AMP-dependent synthetase and ligase family protein	-
3	AT4G19020	(CMT2)	Chromomethylase 2 (CMT2)	n.d.
3	AT4G19030	(NLM1)	An aquaporin whose expression level is reduced by ABA, NaCl, dark, and dessication. Is expressed at relatively low levels under normal conditions. Also functions in arsenite transport and tolerance	+

n.d. = not determined because probe set is not available for this gene on the AtH1 microarray

Germination on Salt

GWA is expected to be particularly powerful for traits that require evolutionary adaptation to environmental conditions. Plant salinity tolerance can be regarded as such and has been intensively studied (Munns and Tester 2008). A clear association peak was detected in a recent study focusing on leaf Na^+ accumulation in *Arabidopsis* (Baxter *et al.* 2010). The peak overlapped the *AtHKT1*:1 gene (AT4G10310), a Na^+ transporter which was shown to be responsible for salt tolerance. They also observed a clear overrepresentation of salt tolerant plants growing in coastal regions and on saline soils which suggests adaptation to the elevated salinity of their local environment. We studied salt tolerance by scoring seed germinating in the presence of 125 mM NaCl. Maximum germination was severely affected when compared to germination on water resulting in a normally distributed germination range between 0 and 93%. Heritability reached a level of 0.431 (Table 7.2). A negative correlation with both latitude and longitude was observed (Table 7.1) which indicates that accessions from the north-eastern regions outperform accessions from the south-western regions (Figure 7.5B). A more detailed correlation study accounting for e.g. coastal regions and saline soils is needed to gain better understanding of the ecological implications of this observation. Several SNPs showed clear association but the expected candidate *AtHKT1:1* was not identified. However, expression analysis via the Bio-

Array Resource e-northern facility (Toufighi *et al.* 2005) shows that *AtHKT1:1* is mainly expressed in roots and flowers and no expression is detected in seeds. This tissue specificity may indicate that other genes are responsible for salt tolerance adaptation during seed germination.

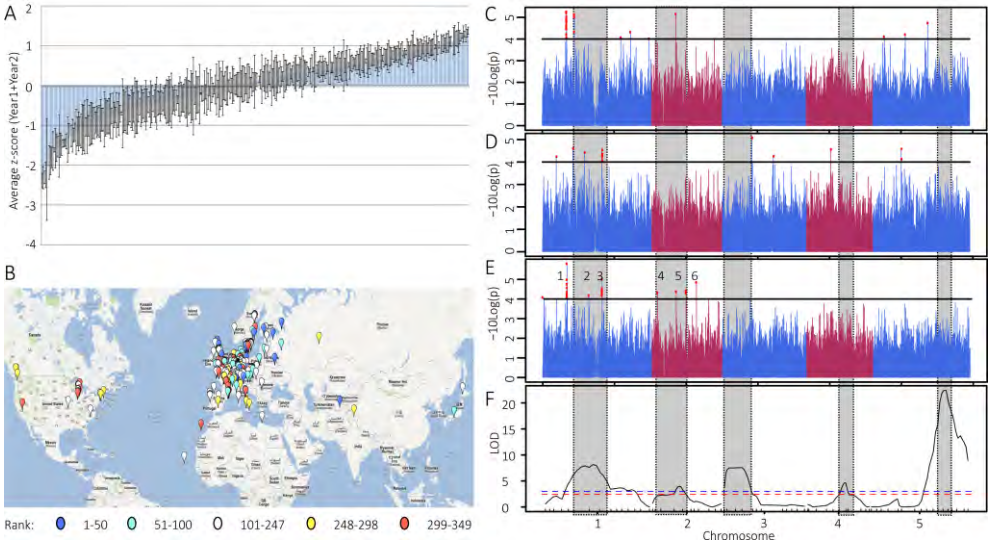


Figure 7.5: Analysis of seed germination on 125 mM NaCl at 20°C after 4 days of stratification. Panel A shows the z-score adjusted distribution of the average delta (water-NaCl) area under the germination curve (AUC) in year 1 and year 2 sorted from small to large over all 349 HapMap accessions. Error bars indicate the standard error. Global distribution of the accessions (ranked from small to large z-score values with indicated color code; blue = good germination on NaCl, red = bad germination on NaCl) is shown in panel B. Genome wide association corrected for the IBS population structure (Year 1: C, Year 2: D, Average Year 1 and Year 2: E). SNPs with –log10 p-values > 4 are marked with red dots. Selected peaks for further analysis are indicated with numbers in panel E. Linkage mapping results for seed germination on 125 mM NaCl at 20°C after 4 days of stratification in the Bay-0 x Sha RIL population are shown in panel F. Significance thresholds, determined by permutation (n=1000) are indicated with dashed lines (Red=85%, Blue=95%). Positions shaded in grey visualize the overlap of genome positions determined by the 2LOD drop of significant QTLs in the Bay-0 x Sha population.

The large number of significant SNPs around peak 1 results in an interval that contains 27 genes, peak number two only matched transposable elements. Several of the identified genes are highly expressed in either developing and/or germinating seeds and based on their annotation some of those can easily be regarded as potential causal candidates for the salt-tolerance phenotype. The unknown protein At1G19530 is a plasma-membrane associated protein that is upregulated by a repressor of GA (unpublished annotation detail; TAIR). GA is a major regulator of seed germination. Excess of NaCl can induce oxidative stress and a role for genes involved in the detoxification of reactive oxygen species can therefore be hypothesized (AT1G19550, AT1G19570, AT2G22420, AT2G15620). *AtHB6* (AT2G22430) encodes a homeodomain leucine zipper class I (HD-Zip I) protein that is a target of the protein phosphatase ABI1 and is a negative regulator of the ABA signaling pathway (Himmelbach *et al.* 2002). An extensive survey of gene expression during abiotic

stress was performed by Kilian *et al.* (2007). This data can be accessed via the Bio-Array Resource e-northern facility (Toufighi *et al.* 2005). Three of our candidate genes showed clear upregulation during salt stress in the Kilian dataset (AT1G19550, AT1G19570 and AT2G22430). Further research is required to confirm the involvement in salt tolerance of the candidate genes described in Table 7.5.

Table 7.5: Candidate genes for variation in seed germination on 125 mM NaCl selected from the GWA analysis. Peak numbers are shown in Figure 7.5. All annotated (TAIR10) genes overlapping the selected 20 kb region are presented. The column (Expr.) indicates whether expression increased during salt stress queried using the Bio-Array Resource e-northern facility (Toufighi *et al.* 2005).

Peak	ID	Symbol	Description	Expr.
1	AT1G19464	MIR864A	Encodes a microRNA of unknown function	n.d.
1	AT1G19470		Galactose oxidase/kelch repeat superfamily protein	-
1	AT1G19480		DNA glycosylase superfamily protein	-
1	AT1G19485		Transducin/WD40 repeat-like superfamily protein	-
1	AT1G19490		Basic-leucine zipper (bZIP) transcription factor family protein	n.d.
1	AT1G19500		Unknown protein	-
1	AT1G19510	(RL5)	RAD-like 5 (RL5)	-
1	AT1G19520	(NFD5)	NUCLEAR FUSION DEFECTIVE 5 (NFD5)	-
1	AT1G19530		Unknown protein	-
1	AT1G19540		NmrA-like negative transcriptional regulator	-
1	AT1G19550		Glutathione S-transferase family protein	+
1	AT1G19560		Pseudogene, putative CHP-rich zinc finger	n.d.
1	AT1G19570	(DHAR1)	Encodes a member of the dehydroascorbate reductase gene family	+
1	AT1G19580	(GAMMA CA1)	Encodes mitochondrial gamma carbonic anhydrase	-
1	AT1G19600		PfkB-like carbohydrate kinase family protein	-
1	AT1G19610	(PDF1.4)	Predicted to encode a PR (pathogenesis-related protein)	-
1	AT1G19620		Unknown protein	-
1	AT1G19630	(CYP722A1)	Member of CYP722A	-
1	AT1G19640	(JMT)	Encodes a S-adenosyl-L-methionine: jasmonic acid carboxyl methyltransferase	-
1	AT1G19650		Sec14p-like phosphatidylinositol transfer family protein	-
1	AT1G19660		Wound-responsive family protein	-
1	AT1G19680		RING/U-box superfamily protein	-
1	AT1G19690		NAD(P)-binding Rossmann-fold superfamily	-
1	AT1G19700	(BEL10)	Encodes a member of the BEL family of homeodomain proteins	-
1	AT1G19710		UDP-Glycosyltransferase superfamily protein	-
1	AT1G19715		Mannose-binding lectin superfamily protein	-
1	AT1G19720		Pentatricopeptide repeat (PPR-like) superfamily	n.d.
3	AT1G43766		Pseudogene, putative phosphofructokinase beta subunit	n.d.
3	AT1G43770		RING/FYVE/PHD zinc finger superfamily protein	-

Peak	ID	Symbol	Description	Expr.
4	AT2G04630	(NRPB6B)	Subunit of nuclear DNA-dependent RNA polymerases II and V	-
4	AT2G04650		ADP-glucose pyrophosphorylase family protein	-
4	AT2G04660	(APC2)	Ubiquitin-protein ligase involved in cell cycle regulation	-
5	AT2G15620	(NIR1)	Involved in the second step of nitrate assimilation	-
5	AT2G15630		Pentatricopeptide repeat (PPR) superfamily	-
5	AT2G15640		F-box family protein	-
5	AT2G15670		Best Arabidopsis thaliana protein match is: SEC14 cytosolic factor family protein	-
6	AT2G22420		Peroxidase superfamily protein	-
6	AT2G22425		Microsomal signal peptidase 12 kDa subunit (SPC12)	-
6	AT2G22426		Unknown protein	n.d.
6	AT2G22430	(HB6)	Encodes a homeodomain leucine zipper class I (HD-Zip I)	+
6	AT2G22440		Matches Ribonuclease H-like superfamily	-

n.d. = not determined because probe set is not available for this gene on the ATH1 microarray

Germination at high temperature

An increase in global temperatures has made heat stress an increasing problem in agriculture. It affects plant growth and development and may lead to a drastic reduction in economic yield. Heat stress affects plant growth throughout its ontogeny. Depending on the duration and intensity it might slow down or totally inhibit germination. Plants have developed an array of mechanisms to cope with heat stress, including maintenance of membrane stability, scavenging of reactive oxygen species (ROS), production of antioxidants, accumulation and adjustment of compatible solutes, induction of mitogen-activated protein kinase (MAPK) and calcium-dependent protein kinase (CDPK) cascades, and chaperone signaling and transcriptional activation (Wahid *et al.* 2007). We studied heat tolerance by germinating seeds at a constant temperature of 30°C (Figure 7.6). Maximum germination was severely affected when compared to germination at 20°C and resulted in a normally distributed germination range between 0 and 99%. Heritability reached a level of 0.553. A negative correlation with both latitude and longitude was observed (Table 7.1) which indicates that accessions from the north-eastern regions outperform accessions from the south-western regions (Figure 7.6B). Surprisingly, most heat tolerant ecotypes were found in the colder northern regions whereas the heat sensitive ecotypes were found in the warmer southern regions. However, from an ecological point of view this might be explained by considering the expected length and intensity of the heat stress. In the northern regions, 30°C can be a positive trigger because it can be regarded as an extreme temperature which indicates the presence of a mild summer with regular rainfall and good opportunities for plant survival and reproduction. Contrarily, in the southern regions 30°C can serve as a warning message because it indicates a long hot and dry summer with low probability for plant survival and reproduction.

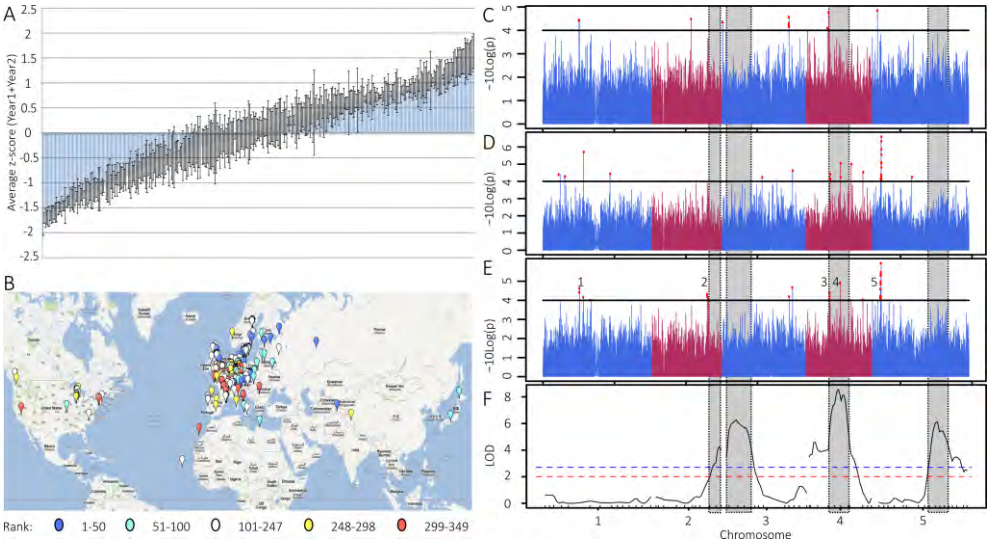


Figure 7.6: Analysis of seed germination at 30°C after 4 days of stratification. Panel A shows the z-score adjusted distribution of the average delta (water-30°C heat) area under the germination curve (AUC) in year 1 and year 2 sorted from small to large over all 349 HapMap accessions. Error bars indicate the standard error. Global distribution of the accessions (ranked from small to large z-score values with indicated color code; blue = good germination at 30°C, red = bad germination at 30°C) is shown in panel B. Genome wide association corrected for the IBS population structure (Year 1: C, Year 2: D, Average Year 1 and Year 2: E). SNPs with –log10 p-values > 4 are marked with red dots. Selected peaks for further analysis are indicated with numbers in panel E. Linkage mapping results for seed germination at 30°C after 4 days of stratification in the Bay-0 x Sha RIL population are shown in panel F. Significance thresholds, determined by permutation (n=1000) are indicated with dashed lines (Red=85%, Blue=95%). Positions shaded in grey visualize the overlap of genome positions determined by the 2LOD drop of significant QTLs in the Bay-0 x Sha population.

Plants cope with heat stress via a broad range of molecular mechanisms. This makes it difficult to prioritize the genes listed in Table 7.6. However, several genes specifically involved in response to stress conditions can be recognized. *BAM3* (AT4G17090) encodes a beta-amylase targeted to the chloroplast. Starch hydrolysis was correlated with maltose accumulation during cold shock and increased expression of *BAM3* (Sicher 2011). Genes involved in scavenging of ROS and production of antioxidants (*PMSR2*; AT5G07460, *PMSR3*; AT5G07470 and *KUOX1*; AT5G07480) can be regarded as important candidates regarding heat stress tolerance. *SMC6A* (AT5G07660) encodes for a component of the SMC5/6 complex. SMC5/6 complex promotes sister chromatid alignment and homologous recombination after DNA damage. *SMC5* has been characterized to be essential for proper seed development (Watanabe *et al.* 2009). None of the genes showed increased expression during the heat stress experiments performed by Kilian *et al.* (2007) but several of the genes are highly expressed during seed germination as indicated in Table 7.6.

Table 7.6: Candidate genes for variation in seed germination at 30°C selected from the GWA analysis. Peak numbers are shown in Figure 7.6. All annotated (TAIR10) genes overlapping the selected 20 kb region are presented. The column (Expr.) indicates whether expression reaches levels > 200 during seed germination queried using the Bio-Array Resource e-northern facility (Toufighi *et al.* 2005).

Peak	ID	Symbol	Description	Expr.
1	AT1G29300	(UNE1)	Unfertilized embryo sac 1 (UNE1)	n.d.
1	AT1G29310		SecY protein transport family protein	+
1	AT1G29320		Transducin/WD40 repeat-like superfamily protein	n.d.
1	AT1G29330	(AERD2)	Similar to endoplasmic reticulum retention signal receptor	+
1	AT1G29340	(PUB17)	Has E3 ubiquitin ligase activity	+
2	AT2G36960	(TKI1)	Arabidopsis thaliana myb/SANT domain protein	+
2	AT2G36970		UDP-Glycosyltransferase superfamily protein	+
2	AT2G36980		Tetratricopeptide repeat (TPR)-like superfamily protein	-
2	AT2G36985	(ROT4)	Encodes ROTUNDIFOLIA4	n.d.
2	AT2G36990	(SIGF)	Encodes a general sigma factor in chloroplasts	-
2	AT2G37010	(NAP12)	Member of NAP subfamily	-
2	AT2G37020		Translin family protein	+
2	AT2G37025	(TRFL8)	TRF-like 8 (TRFL8)	n.d.
2	AT2G37030		SAUR-like auxin-responsive protein family	-
3	AT4G10780		LRR and NB-ARC domains-containing disease resistance protein	-
3	AT4G10790		UBX domain-containing protein	+
3	AT4G10800		Best match is: BTB/POZ domain-containing protein	-
3	AT4G10810		Unknown protein	+
3	AT4G10820		F-box family protein	-
3	AT4G10840		Tetratricopeptide repeat (TPR)-like superfamily protein	+
4	AT4G17070		peptidyl-prolyl cis-trans isomerases	-
4	AT4G17080		Histone H3 K4-specific methyltransferase SET7/9 family protein	-
4	AT4G17090	(BAM3)	Encodes a beta-amylase targeted to the chloroplast	-
4	AT4G17098		Natural antisense gene, locus overlaps with AT4G17100	n.d.
4	AT4G17100		Unknown protein	n.d.
5	AT5G07450	(CYCP4;3)	Cyclin p4;3 (CYCP4;3)	-
5	AT5G07460	(PMSR2)	Ubiquitous enzyme that repairs oxidatively damaged proteins	+
5	AT5G07470	(PMSR3)	Ubiquitous enzyme that repairs oxidatively damaged proteins	+
5	AT5G07480	(KUOX1)	KAR-UP oxidoreductase 1 (KUOX1)	-
5	AT5G07490		Unknown protein	n.d.
5	AT5G07510	(GRP14)	Encodes a glycine-rich protein	-
5	AT5G07630		Lipid transporters	-
5	AT5G07640		RING/U-box superfamily protein	-
5	AT5G07650		Actin-binding FH2 protein	n.d.
5	AT5G07660	(SMC6A)	Encodes SMC6A (STRUCTURAL MAINTENANCE OF CHROMOSOMES)	-
5	AT5G07670		RNI-like superfamily protein	n.d.
5	AT5G07680	(NAC080)	NAC domain containing protein 80 (NAC080)	-

n.d. = not determined because probe set is not available for this gene on the ATH1 microarray

Germination response to exogenous ABA

Abscisic acid (ABA) is a phytohormone which is a positive regulator of seed dormancy; it inhibits seed germination and has a role during after-ripening (Kucera *et al.* 2005). Seed germination requires that the growth potential of the radicle overcomes the tissue resistance of the seed covering layers. In *Arabidopsis* and many other species testa rupture and endosperm rupture are two sequential steps during germination (Müller *et al.* 2006). ABA sensitivity depends on genotype and seed age and a range of phenotypes might be observed. Externally applied ABA can either completely inhibit germination, prevent radicle protrusion through the endosperm or inhibit seedling greening and growth. We studied response to ABA by germinating seeds in the presence of 0.5 µM ABA. Maximum germination was severely affected when compared to germination at 20°C resulting in a normally distributed germination range between 0 and 99%. Heritability reached a level of 0.58. No clear trend could be observed in the global distribution of the accessions (Table 7.1, Figure 7.7B).

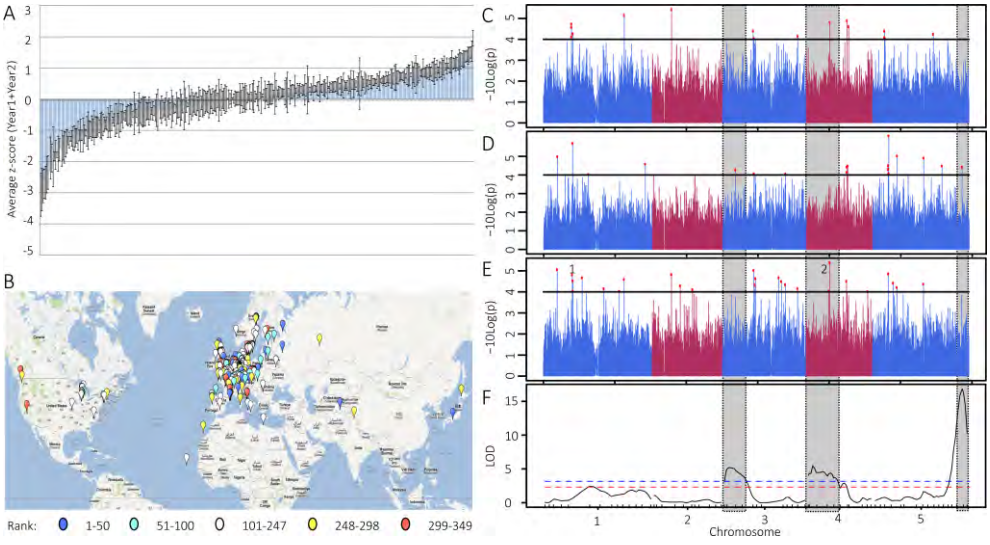


Figure 7.7: Analysis of seed germination on 0.5 µM ABA after 4 days of stratification. Panel A shows the z-score adjusted distribution of the average delta (water-ABA) area under the germination curve (AUC) in year 1 and year 2 sorted from small to large over all 349 HapMap accessions. Error bars indicate the standard error. Global distribution of the accessions (ranked from small to large z-score values with indicated color code; blue = good germination on ABA, red = bad germination on ABA) is shown in panel B. Genome wide association corrected for the IBS population structure (Year 1: C, Year 2: D, Average Year 1 and Year 2: E). SNPs with –log10 p-values > 4 are marked with red dots. Selected peaks for further analysis are indicated with numbers in panel E. Linkage mapping results for seed germination on 0.5 µM ABA after 4 days of stratification in the Bay-0 x Sha RIL population are shown in panel F. Significance thresholds, determined by permutation (n=1000) are indicated with dashed lines (Red=85%, Blue=95%). Positions shaded in grey visualize the overlap of genome positions determined by the 2LOD drop of significant QTLs in the Bay-0 x Sha population.

Despite the high heritability and the association of several SNPs only two peaks met our selection criteria (2 or more SNPs with –log10 p-values > 4 within an interval of 20 kb in the average between year 1 and 2, or single SNPs with –log10 p-values above 4 within an interval of 20 kb in the average between year 1 and 2 when overlapping with a Bay-0 x Sha QTL interval). *GH9B6* (AT1G23210, Table 7.7) encodes a glycosyl hydrolase and might be involved in cell wall hydrolysis upon germination but it should be noted that expression analysis using the Bio-Array Resource e-northern facility (Toufighi *et al.* 2005) shows a very specific pollen expression of GH9B6 without any expression during seed germination. The dynein light chain type 1 family protein is part of a microtubule associated complex and is annotated to be involved in the catalysis of movement along a microtubule. *AtOPT7* (AT4G10770) is expressed in the embryonic cotyledons prior to root radicle emergence and is suggested to have distinct cellular roles including nitrogen mobilization during germination and senescence, ovule development, seed formation and metal transport (Stacey *et al.* 2006). None of the possible candidate genes mentioned in table 7 was previously shown to be affected by ABA.

Table 7.7: Candidate genes for variation in seed germination on 0.5 µM ABA selected from the GWA analysis. Peak numbers are shown in Figure 7.7. All annotated (TAIR10) genes overlapping the selected 20 kb region are presented. The column (Expr.) indicates whether expression reaches levels > 200 during seed germination queried using the Bio-Array Resource e-northern facility (Toufighi *et al.* 2005).

Peak	ID	Symbol	Description	Expr.
1	AT1G23205		Plant invertase/pectin methylesterase inhibitor superfamily protein	-
1	AT1G23210	(GH9B6)	Glycosyl hydrolase 9B6 (GH9B6)	-
1	AT1G23220		Dynein light chain type 1 family protein	+
1	AT1G23230		CONTAINS InterPro DOMAIN/s	-
2	AT4G10767	(SCRL21)	Member of a family of small, secreted, cysteine rich proteins	n.d.
2	AT4G10770	(OPT7)	Oligopeptide transporter	+
2	AT4G10780		LRR and NB-ARC domains-containing disease resistance protein	-
2	AT4G10790		UBX domain-containing protein	+
2	AT4G10800		Best match is: BTB/POZ domain-containing protein	-

n.d. = not determined because probeset is not available for this gene on the ATH1 microarray

Conclusions

We tested seed germination related traits and compared GWA results with conventional mapping in the Bay-0 x Sha RIL population. Phenotyping of such a diverse panel of accessions resulted in several interesting observations. For example, we discovered 5 new mucilage mutants as being outliers in the correlation between imbibed and dry seed size and we noticed that accessions from the northern regions were less affected by high temperatures compared to accessions from the southern regions. Phenotypes were accurately assessed by using 4 replicates (2 years, 2 blocks) and high heritability scores (ranging from 0.431 to 0.673) were obtained. In a large and well-defined

population such as the core360 HapMap in which the population structure has been minimized one would expect enough power to efficiently map genetic variation for traits expressing such levels of genetic variation. Despite the high heritability scores we must conclude that none of the traits resulted in clear association with –log10 p-values above the Bonferroni multiple test corrected threshold of 6.5. Seed germination is a crucial step in the life cycle and the involved loci might be under strong evolutionary selection as they are of great relevance for adaptation to new locations (Huang *et al.* 2010). Contrarily, due to the importance of seed germination, a robust system is needed in which many loci with small additive effects determine the final output. Genome wide association can easily be underpowered to efficiently detect such relatively small effect loci. Effects are bigger in structured biparental populations but such populations will also lack genetic variation for many loci. Combining GWA analysis and traditional linkage mapping can be an efficient approach to validate associations, increase power to detect rare alleles and reduce the number of candidate genes (Brachi *et al.* 2010). However, no clear peaks of multiple associated SNPs in the GWA co-located with any of the QTL intervals detected in the Bay-0 x Sha RIL population. New RIL populations created from extreme ecotypes discovered in our HapMap phenotyping could be an interesting lead for further research. In conclusion, genome wide association is a promising tool to dissect natural genetic variation but it needs further development of the procedures of analysis and population definition to overcome the lack of power encountered and described in this paper.

Materials and methods

Mapping population

The *Arabidopsis* population used in this study is composed of a selection of 360 accessions (Li *et al.* 2010) and were obtained via the Arabidopsis Biological Resource Centre. Large differences in flowering time exist within this population which might affect the conditions during seed maturation. To reduce the effect of flowering time all plants were vernalized (5°C, 16 hour day) for 8 weeks before transferring them to the greenhouse. This resulted in a uniform start of flowering restricted to a period of 2 weeks. All available accessions (349) were grown in duplicate in two blocks for two consecutive years. Plants were randomly distributed in two blocks of two plants per accession. Seeds were harvested after complete maturation on the plant and bulked for each block. Two growth seasons (2010 and 2011) were used with the same setup as described above but with a different growing substrate. In season 2010 fertilized soil was used while in season 2011 rockwool plugs fertilized with 1 g/l Hyponex was used.

Germination experiments

Germination experiments were performed by using the Germinator setup. Seeds were after-ripened at ambient laboratory conditions (~30 RH%, 20°C); seeds from harvest year 1 for 7 months and seeds from harvest year 2 for 6 months. A cold stratification period of 4 days at 4°C in the dark was applied before transferring the trays to the germination incubator (20°C, continuous light). A temperature of 30°C was used for 'heat'-germination. Salt stress was applied by replacing the water by a 125 mM NaCl (Sigma Aldrich, #S-3014) solution. ABA (Duchefa Biochemie, A0941) was initially dissolved in a few drops of 1N NaOH from which stock solutions were prepared in 10 mM MES buffer, pH 5.9. ABA was used at a final concentration of 0.5 µM. All germination tests were performed in a fully randomized setup. Averages were calculated and corrected for their proper control.

Seed size measurements

Dry seed size was determined by taking close-up photographs from ~100-200 seeds using a Nikon D80 camera with a 50mm Macro objective. Imbibed seed size was extracted from the first images acquired within the Germinator setup (100-200 seeds). The photographs were analyzed using the open source image analysis suite ImageJ (http://rsbweb.nih.gov/ij/) by using color-thresholds combined with particle analysis (Joosen *et al.* 2010).

GWA mapping

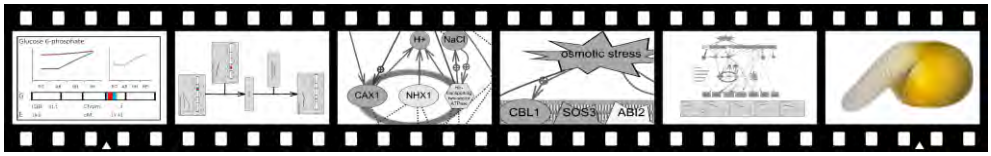
Genome wide association mapping was performed using a custom R-script and C+ program (ScanGLS) which was tailored to perform analysis on replicated phenotype measurements. The used procedure including the construction of the IBS matrix exactly follows the procedure described by Kang *et al.* (2010). All scripts will be implemented at the x-QTL workbench (www.xqtl.org).

The mouth who ate the seeds asks; "Now what shall I plant?"

8 General discussion

(R)evolution of seed quality research

Joosen RVL



Seed Quality

Seed quality is a collective term for the condition of a seed batch, comprising attributes such as genetic and physical purity, viability, germination, dormancy, vigor, storability, uniformity in size, freedom from seedborne diseases and absence of mechanical damage, heat damage and preharvest sprouting. The practical definition of seed quality can differ, depending on the end user. A farmer or plant grower may desire high-quality seed that produces rapidly growing plants of uniform size, with high yielding capacity under a wide range of field conditions, whereas a producer of oil seed crop may desire seed with a certain stable fatty acid composition. Since seed quality is such a complex 'trait', testing is in many cases at best an 'educated guess' to predict behavior in the field. In addition, seed producers have redefined the term 'seed quality' to include important attributes, such as 'usable plants' and 'seedling and crop establishment', whereas the traditional notion of seed quality is predominantly seed-centered, e.g. related to germination or storability. This poses important questions, such as: how is seedling- or plantlet performance established during seed development or dependent on germination/growth conditions? Furthermore, any study of seed quality must be aware of the fact that changes in seed performance may occur during dry storage, even under the most optimal conditions.

Seed quality is largely acquired during seed development and mainly during the maturation phase. The resulting performance of the seed is a function of the complex interaction between the genome and environment during development. Thus, quality of seed lots, as they are received by seed companies, often from locations all over the world, vary among regions and production seasons and years. In the present seed production practice the emphasis is on harvest timing and methods and post-harvest treatments since it is difficult to influence the production environment, even under greenhouse conditions where the outside climate also has it influences on the environment. Moreover, the effect of the environment on seed quality is largely unknown. So far, genetic components of seed quality have hardly been used in breeding programs. Some quantitative trait loci (QTLs) related to germination, storability and stress tolerance have been found in Arabidopsis and

tomato (Foolad *et al.* 2003; Clerkx *et al.* 2004; Kazmi *et al.* 2012) but a systematic study of the genetics of seed quality is lacking.

With current technological advances it has become possible to combine quantitative genetics with genomics. In our study we aimed at an integration of genetics with detailed phenotyping at the physiological, metabolomic and genomic level. If genes, or rather gene sets, associated with seed quality parameters become available, they may be used as diagnostic tools to assess seed quality, in marker-assisted breeding, or for genetic modification to enhance seed quality.

Development of a new toolbox was needed to enable high-throughput characterization of the genetics of seed performance. First a method was developed to allow automatic scoring of seed germination (Chapter 2). With the use of this method a broad range of environmental conditions was screened to identify QTLs for germination capacity (Chapter 3). New diagrams were developed for the MapMan tool to allow combined analysis of gene expression and metabolite fluctuations with a focused view on molecular processes related to seed dormancy and germination (Chapter 4). This was followed by large-scale QTL analysis and efficient environmental perturbation in genetical genomics experiments, new analysis scripts to allow data-analysis are developed and described (Chapter 5 and 6). The metabolome and gene-expression study was restricted to dry mature seeds (dormant and non-dormant), 6-hour imbibed seeds and seeds at the moment of radicle protrusion. Recently, the expression QTL landscape of developing seeds has been described in the same RIL population as in the present study (Cubillos et al. 2012). An integration of both datasets is now possible and is highly recommended for future research. Finally we assessed the possibilities to use genome wide association to dissect the genetics of seed performance and compared the obtained results to QTLs that were detected in the Bay-0 x Sha RIL population (Chapter 7).

Measuring seed germination

An important instrument to assess the performance of a seed lot is the accurate quantification of germination by collecting cumulative germination data. Completion of germination is defined as the protrusion of the radicle through the endosperm and seed coat (Bewley 1997). We have created a high resolution time-lapse movie of this process which was published at the Arabidopsis Information Resource (www.arabidopsis.org). For accurate scoring of seed germination a careful discrimination should be made between the moment of testa and subsequent endosperm rupture because the intermittent lag phase may vary among germination conditions and treatments (Liu et al. 2005; Finch-Savage and Leubner-Metzger 2006; Müller et al. 2006). Although commonly used, the total percent germination after a nominated period of time is not very explanatory. It lacks information about start, rate and uniformity of germination, which are essential parameters of a normally distributed seed population for many traits such as dormancy, stress tolerance and seed aging. Information about germination at various time intervals is required to

calculate a cumulative germination curve, but the number of samples that can be handled with manual counting is usually a limiting factor. Therefore, we have developed the Germinator system, an automated procedure that enables high-throughput germination screening (Chapter 2, Joosen *et al.* 2010). Many experiments described in this thesis (Chapter 3 and 7) have relied heavily on high-throughput characterization of seed germination and would not have been possible without the use of the Germinator. The system was optimized for use with Arabidopsis seeds but it can be used for other species as well: optimization for tomato, Brassica and lettuce is currently under development. The system was complemented with a high-throughput module for analyzing cumulative germination curves. The use of this curve-fitting module is not restricted to any species and has been supplied to over 300 researchers from various universities and commercial seed companies all over the world. Further extensions using a similar approach, but with a more advanced image analysis algorithm called 'expectation maximization' are now in development. This new procedure has the potential to detect both germination sensu stricto and the development of green cotelydons. Developmental arrest after germination is an important seed quality characteristic and simultaneous detection of both parameters will allow a more precise description of the performance potential of a seed batch.

Using natural variation

Two approaches are used to study natural variation present for seed performance (Chapter 3 and Chapter 7). We first used a recombinant inbred line (RIL) population derived from two distinct Arabidopsis thaliana ecotypes: Bayreuth (Bay-0) which originates from a fallow land habitat in Germany and Shahdara (Sha) which grows at high altitude in the Pamiro-Alay mountains in Tadjikistan (Loudet et al. 2002). The Sha parent has successfully adapted to the harsh environment which it encountered in Tadjikistan and is often found to be remarkably stress tolerant. With the use of the Germinator system, germination sensu stricto was determined under a wide range of environmental conditions. In total this analysis resulted in 327 trait scores over several different harvests. Evaluation of these high numbers of phenotypes demanded methods of QTL analysis that extended beyond individual trait mapping and that allows comprehensive and comprehensible visualization. Analysis of loci with genetic variation can efficiently be done using the statistical language R which includes the R/qtl package (Broman et al. 2003). This package contains an array of different QTL mapping methods, including Single Marker Mapping, Interval Mapping and Multiple QTL Mapping (MQM) (Arends et al. 2010). Although all possibilities to perform a detailed QTL analysis including data preprocessing and output formatting are present in R, it requires extensive knowledge of the R-syntax to combine all necessary steps in a single analysis protocol that can loop through hundreds or thousands of traits. We created a script that combined all these necessary steps and that can perform automated QTL mapping including the necessary data preprocessing and output formatting (Chapter 3, Joosen *et al.* 2012). This type of automated analysis combined with efficient data

visualization is a necessary step to keep up with the ever increasing rate of biological data production.

The detailed phenotyping of such a wide range of seed germination characteristics has yielded a comprehensive inventory of loci with genetic variation. This 'genetic landscape of seed performance' not only provides information about the various QTL positions (G) but also about the interaction between loci (GxG) and between loci and the environment (GxE). Many QTLs showed overlap with previously reported loci in either the same or different RIL populations. Confirmation of some major QTL hotspots was demonstrated using the heterogenous inbred family (HIF) approach (Tuinstra et al. 1997). However, examples of clear causality are sparse and QTL intervals often span large genomic regions that can easily contain up to 1000 annotated genes. Refining the QTL intervals by classical fine mapping procedures is labor intensive and should be regarded as a low throughput technique. Therefore, alternative procedures to obtain insight in the molecular processes underlying the detected QTLs were studied.

The ultimate mapping population to explore the comprehensive reservoir of natural variation consists of a worldwide collection of accessions. In our study in Chapter 7 we used the Arabidopsis HapMap population, which is an assembly of the most diverse 360 accessions found worldwide (Li et al. 2010). The low level of linkage disequilibrium (LD) in this population allows a high resolution mapping which confines the causal gene detection to only a few genes. It is thought that the relatively low LD reflects a history of frequent outcrossing together with rapid dispersal enabled by the selfing mode of reproduction (Weigel 2012). We assessed seed size and germination capacity in the HapMap population by using 4 replicates and observed large genetic variation. However, we must conclude that, despite the high heritability scores, none of the traits resulted in significant association after a Bonferroni correction (Chapter 7). This lack of statistical power can be the consequence of several technical reasons such as the quality of the SNP genotype data or the statistics used to determine the linkage. Soon the SNP genotype data for this population will be replaced by full genome sequence data which might allow a more powerful analysis (Nordborg and Weigel 2008). However, for crucial steps in the plant's life cycle or survival a robust biological process is needed. In those cases, the evolutionary selection might be directed towards a system with many loci with small additive effects. Genome wide association can easily be underpowered to efficiently detect such small effect loci. Further, functional variants at low frequency have little influence on the population as a whole, and their signal is therefore difficult to detect (Myles et al. 2009). Methods to detect genes with epistatic interaction or genes with multiple alleles with similar effects on the phenotype are still in development (Weigel 2012). Effects are bigger in structured biparental populations but such populations will also lack genetic variation for many loci. Combining GWA analysis and traditional linkage mapping can be an efficient approach to validate associations, increase power to detect rare alleles and reduce the number of candidate genes (Brachi et al. 2010). Overlap between multiple associated SNPs in the GWA with QTL intervals that were detected in the Bay-0 x Sha RIL population are

discussed in Chapter 7. New RIL populations created from ecotypes with extreme phenotypes discovered in the HapMap phenotyping could be an interesting lead for further research. In conclusion, genome wide association is a promising tool to dissect natural genetic variation but it needs further development of the procedures of analysis and population definition to overcome the lack of power encountered and described in this thesis.

Genetical 'omics'

Another approach to close in on the molecular mechanisms underlying the genetic variation that was found for seed germination in the Bay-0 x Sha RIL population is the use of 'omics' technology. Genetical genomics studies, in which molecular traits are genetically analyzed, have been successfully applied to enhance a directed strategy to identify causal relationships (Kliebenstein *et al.* 2006; Keurentjes *et al.* 2007; Van Leeuwen *et al.* 2007; Wentzell *et al.* 2007; West *et al.* 2007; Rowe *et al.* 2008; Cubillos *et al.* 2012). Many studies have shown an effect of variable environments on these molecular traits and feedback mechanisms between different levels of organization are well known. To incorporate developmental or environmental perturbation in the often expensive and laborious omics analyses, an alternative experimental setup, coined 'generalized genetical genomics' (GGG) has been proposed. This approach aims at the creation of sub-populations of RILs, one for each environment to be tested, with an optimal distribution of parental alleles over all available markers (Li *et al.* 2009). This concept offers unique reduction in experimental load with minimal compromise to statistical power and is of great potential in the field of systems genetics in which a broad understanding of both plasticity and dynamics is required (Li *et al.* 2008). We divided the Bay-0 x Sha RIL population in four well balanced subpopulations consisting out of 41 lines each. This created the possibility to profile four different environmental conditions; 1) primary dormant dry seeds, 2) after-ripened dry seeds, 3) 6-hour imbibed seeds and 4) seeds at the time of radicle protrusion. An R-procedure using linear models was developed which enables a fast QTL mapping for this type of design which takes the environmental perturbations into consideration (Chapter 5). Different levels of variation were obtained and could be mapped by the environmental (E), genetic (G) and/or genetic x environment (G:E) component of the linear model. Often, the observed variation is subject to the environment without a complete abolishment of the genetic variation. Environmental effects can be normalized in those cases and the power of detecting a QTL is restored to the total number of lines used in the different sub-populations. However, if genetic variation is only detectable in a single unique environment or developmental stage, QTL detection is limited by the number of lines used for that specific environment. A careful selection of the environments used for a GGG experiment is therefore crucial. Limited power can be expected when environments vary too much and no overlapping genetic variation is present. Contrarily, there is hardly additive value of the design when using very similar environments.

One of the benefits of the generalized approach is the possibility to simultaneously analyse fluxes of the trait under study. Many studies have shown the highly dynamic nature of molecular mechanisms leading towards seed germination ((e.g. reviewed in Catusse *et al.* 2008; Daszkowska-Golec 2011; Weitbrecht *et al.* 2011). Performing expensive genetical genomic experiments without any perturbation of the developmental stage will therefore always raise questions about the possible extrapolation of the results when slightly different conditions are used. To enhance the visualization of the molecular processes that underlie seed germination we describe in Chapter 4 (Joosen *et al.* 2011) the creation of two new diagrams which can be used in the program MapMan (Usadel *et al.* 2006). Pre-existing biological knowledge was used to group genes and metabolites in functional categories. These categories were combined in a single diagram which summarizes transcript and/or metabolite level changes in the pathways important for seed germination. A second diagram provides a focused view of cell wall modification and degradation that are key processes for the completion of seed germination (Lee *et al.* 2012). This approach, using the MapMan tools, offers the seed science community an easy way to analyze and visualize transcriptome and metabolome data for *Arabidopsis*. The produced pathways are also useful for the analysis of genetical genomics data described in chapters 5 and 6. An overlay of the annotated metabolites and genes that show genetic variation at a specific locus can be visualized and allows a broad perspective of the molecular processes that might affect a co-located phenotypic QTL.

The feasibility of the generalized genetical genomics concept was first tested by using metabolite analysis. Polar fractions of a methanol extract from all individuals of the Bay-0 x Sha RIL population at the aforementioned four different developmental stages were subjected to GC-TOF-MS (Chapter 5). In total, 161 metabolites were detected of which 63 could be annotated. Further improvement in the development of mass identification libraries is important as it would increase the number of identified compounds. The unraveling of metabolic pathways requires proper identification of the detected compounds and would benefit much from such improvement. We were able to detect 83 metabolites with genetic variation and 27 metabolites with a clear interaction between the genetic and environmental variation. Several QTLs were confirmed by using the heterogeneous inbred family (HIF) approach (Tuinstra *et al.* 1997). Overlapping QTL positions for several metabolites were observed and could be explained by the fact that they play a role in highly interconnected pathways from which the individual compounds are most likely subject to co-regulation. Since some metabolites appear to be co-regulated, the strong impact of the involved loci on central metabolism might also exert its effect on physiological traits. If true, a comparison between the variation in germination characteristics described in Chapter 3 (Joosen *et al.* 2012) and metabolite levels (Chapter 5) would reveal compounds involved in the process of germination. However, no clear co-location of hotspots for germination and metabolite QTLs could be observed but only incidental overlap between QTLs of both types of traits was detected. Several hypotheses regarding the physiological effects of variation in amino-acids, fumarate, malate and GABA

levels could be formulated based on the overlapping QTL positions and these are interesting leads for further research.

After the experience gathered with the metabolomics experiments the same generalized genetical genomics approach was used to perform gene expression profiling (Chapter 6). The exact same material and developmental stages were used to allow optimal comparison. Expression profiling was performed using the Affymetrix AtSNPtile microarray. This array contains 1.7 million unique 25-mer tiling probes in sense and antisense direction covering the whole genome at a 35 bp resolution. Expression QTLs for Arabidopsis can be classified according to the position of the causal polymorphisms. In our study we detected 2006 eQTLs that had genetic (G) variation from which 1809 were classified as local and 197 as distant. Accordingly, 447 local and 82 distant eQTLs were found for genes with genetic x environmental (G:E) variation. Efficient modeling of molecular processes influenced by a strong local eQTL requires overlapping detection of the small effect distant eQTLs (Jimenez-Gomez *et al.* 2010). Unfortunately, only few examples of such co-expressed molecular processes were detected. Possibly more replications of the expression profiling are needed to increase the statistical power to enable detection of many more small effect distant eQTLs. Improvement of the statistical methods used for the eQTL detection in the generalized design (e.g. by allowing cofactors as used in MQM mapping) will also enhance the chance to detect small size effects (Jansen *et al.* 1994). Screening an a-priory list of 211 candidate genes with known functions in seed germination resulted in 24 genes that showed significant expression variation between Bay-0 and Sha. For example, we identified *DOG1*, the major regulator of seed dormancy, showing a local eQTL with strong genetic x environment interaction on chromosome V overlapping the DOG1 QTL for seed dormancy.

Despite our efforts to analyze this huge dataset in great detail we were not able to exploit its full potential. For example, the tiling array allows to analyze probe specific QTLs. Currently, statistical procedures are in development which use these individual probe expression levels to investigate expression variation. Next to a more accurate estimation of transcript expression this will provide information about alternative splicing events and detect natural antisense transcription. Both alternative splicing and anti-sense transcription are important mechanisms for gene expression under both normal and stress conditions (Jin *et al.* 2008). Genetic variation for these events between Bay-0 and Sha might therefore provide important insight in such regulatory mechanisms.

Data integration

The Bay-0 x Sha RIL population has been used in numerous studies to map QTL positions which provide great opportunities for future research to integrate data ranging from physiological, metabolic and gene expression levels. Often comparisons between studies are hampered because different developmental stages or plant growing conditions were used. With the data described in this thesis we finalized a comprehensive study on seed traits at the physiological (Chapter 3, Joosen *et al.* 2012), metabolic (Chapter 5) and

gene expression level (Chapter 6). These three studies were carried out on the exact same biological material, providing a solid base to elucidate the phenotype-to-genotype relationship. Overlapping QTL positions may be indicative for common regulatory processes but must be interpreted with great caution because assuming causal relations only based on overlapping QTL intervals should be regarded with skepticism (Li et al. 2010). However, performing an in-depth analysis using prior knowledge of interrelated biological data can improve the interpretation of possible phenotype-to-genotype relationships. A great information resource for metabolic pathways including predicted enzymes and coding genes for Arabidopsis is available at AraCyc (http://pmn.plantcyc.org). This information can help to recognize possible relations between metabolite- and expression QTLs found in our studies.

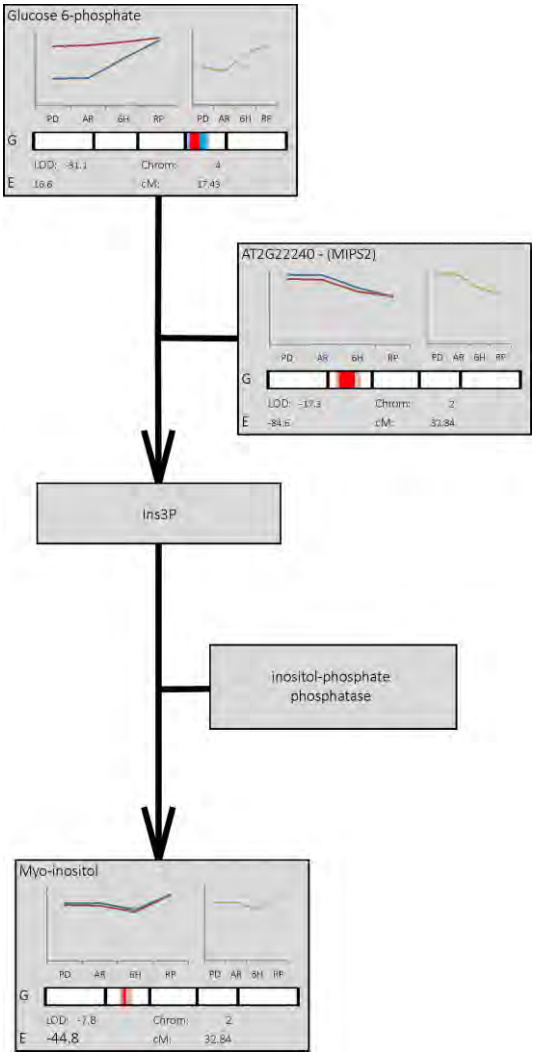


Figure 8.1: Myo-inositol biosynthesis pathway. Metabolite QTLs are detected for glucose-6-phosphate and myo-inositol. An expression QTL is detected for *MIPS2*. Each panel contains information about environmental variation (green line plot, average over all lines within a single developmental stage) and genetic variation (blue lines represent the metabolite levels for lines carrying the Bay-0 allele for the most significant QTL and red lines those for the Sha allele carrying lines). QTL LOD profiles are indicated at the bottom of each panel by a heat bar representing the 5 chromosomes. Genetic (LOD) and Environmental (E) variation is expressed as LOD score in the lower left corner.

For example, several components were detected in our studies for the myo-inositol biosynthesis pathway: the metabolites glucose-6-phosphate, myo-inositol and three genes encoding myo-inositol-1-phosphate synthase (MIPS) (details in Chapter 6). The global expression patterns of these genes can be evaluated using the eFP-browser (http://bar.utoronto.ca/efp/cgi-bin/efpWeb.cgi) and show that MIPS1 is mainly expressed during seed development, MIPS2 at the end of seed maturation and in dry seeds and MIPS3 in young seedlings and throughout the rest of vegetative plant life. Comparing the QTL LOD profiles show that MIPS2 (At2G22240) overlaps with the metabolite QTL found for myo-inositol at chromosome II, marker 2.36 (Figure 8.1). Because 4 different developmental stages were used in our study the allelic effect can be separated accordingly. This reveals a comparable expression pattern for MIPS2 and myo-inositol at the dry primary dormant (PD), dry after-ripened (AR) and 6-hour imbibed (6H) stages with higher expression in lines with the Bay-0 allele. However, the expression for MIPS2 is decreasing at the point of radicle protrusion (RP) while the level of myo-inositol is increased at this stage. Here MIPS3 is highly expressed and may take over the function of MIPS2. MIPS3 does not show expression variation between Bayreuth and Shahdara which could explain the converging allelic effect lines at the point of radicle protrusion. This example shows both the high level of information enclosed in the data described in this thesis and the potential of including prior knowledge to hypothesize causal relationships.

A particular helpful software tool to include prior biological knowledge is Pathway Studio (www.ariadnegenomics.com) which can be used to add biological perspective based upon knowledge extracted from scientific literature. It can find common regulators and associated pathway components for biological processes validated by literature citations. As an example we used Pathway Studio to visualize molecular processes involved in osmotic stress regulation (Figure 8.2). Elements of this process are expected to play crucial roles during seed desiccation and germination under non-favorable conditions. Identifying the components of this pathway that showed natural variation between Bayreuth and Shahdara accessions may help to explain their differential response to salt exposure during seed germination. Five genes in this pathway showed significant expression differences between the two accessions, resulting in either a local or distant eQTL under the non-stressed conditions used for the expression QTL analysis. A strong local eQTL was detected for ABI2 which encodes a protein phosphatase 2C and is involved in ABA signal transduction (Finkelstein and Somerville 1990). ABI2 mutants are abscisic acid tolerant, contain increased levels of endogenous ABA during seed development but are reduced in seed dormancy and have reduced sensitivity to salt and osmotic stress during germination. The detected eQTL for ABI2 overlaps with a QTL detected for seed germination in the presence of NaCl and mannitol. In agreement with the ABI2 mutant phenotype, a lower ABI2 expression level was detected in Shahdara compared to Bayreuth. Surprisingly, RAS1, a gene recently discovered to play a role in ABA signaling during salt stress (Ren et al. 2010) showed a distant expression QTL overlapping with the ABI2 eQTL. This suggests an interaction between these two genes which is supported by an epistatic interaction

between two QTLs that were detected in chapter 3 for seed germination under salt exposure which overlaps with both the RAS1 and ABI2 positions (Joosen *et al.* 2012). However, the complexity of this type of interpretation is shown by the fact that the ABI2 eQTL also overlaps a phenotypic QTL for seed germination in the presence of ABA at which the Bay-0 accession shows higher tolerance to ABA. Although, this type of opposite effects are expected for complex traits which are regulated by many genes it complicates associative studies which are based on co-location of QTLs. A strong phenotypic QTL for the rate of seed germination on salt was also detected on the top of chromosome V with higher tolerance for the Bay-0 allele. Two genes in the osmotic stress pathway (NHX1; AT5G27150 and H^+ ATPase; AT5G08690) are flanking this QTL but it must be noted that both only partly overlap with the phenotypic QTL interval. NHX1 encodes a vacuolar sodium/proton antiporter involved in salt tolerance (Barragan et al. 2012). The detected NHX1 local eQTL shows higher expression levels in Bayreuth which is in agreement with the detected phenotypic QTL and the reported improved salt and drought tolerance in NHX1 overexpression lines in various species (Leidi *et al.* 2010; Teakle *et al.* 2010; Zhou *et al.* 2010; Asif *et al.* 2011). The H^+ ATPase encodes a mitochondrial ATP synthase beta-subunit and functions in mitochondrial oxidative phosphorylation. The increased activity of H^+ ATPases creates the driving force for Na^+ transport by membrane salt overly sensitive proteins 1 (SOS1) (Zhu 2003) and a proteomics study showed clear induction of this protein after NaCl treatment in Arabidopsis cell suspension cultures (Ndimba *et al.* 2005). Finally, we identified one member of the PLC gene family with expression variation between Bayreuth and Shahdara. The Arabidopsis genome contains nine AtPLC genes (Tasma et al. 2008). Members of this gene family are differentially expressed in Arabidopsis organs and it has been shown that a majority of the AtPLC genes are induced in response to various environmental stimuli, including cold, salt, dehydration, and ABA. Transcriptional activation of the AtPLC gene family is considered to be important for the adaptation of plants to stressful environments. A local eQTL showing expression variation between Bayreuth and Shahdara was detected for PLC8. Expression patterns and phylogenetic relationships in the ecotype Columbia-0 indicate that AtPLC gene members AtPLC8 and AtPLC9 may represent a recent duplication; this observation could not be confirmed by our eQTL data. This example clearly shows the complexity enclosed in expression QTL data. Together, the integrated analysis combining the expression QTL data, phenotypic QTL data and available biological knowledge shows its potential of identifying possible candidate genes and interactions. While the expression QTL study was performed under non-stressed conditions it appears possible to examine molecular processes involved in osmotic stress signaling.

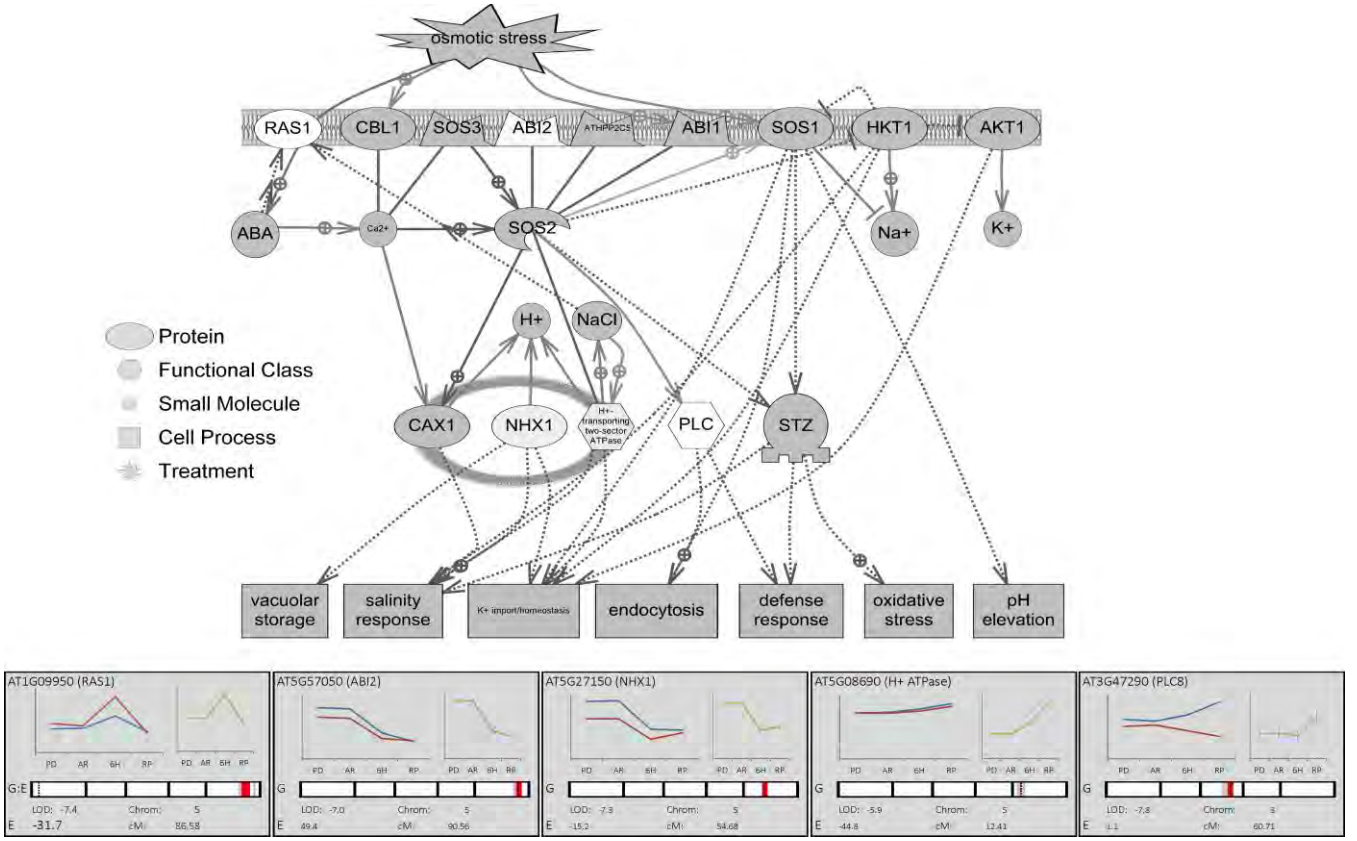


Figure 8.2: Curated Pathway Studio representation of osmotic stress signaling combined with the expression variation during seed germination measured in the Bay-0 x Sha RIL population. Each of the 5 lower panels contains information about environmental variation (green line plot, average over all lines within a single developmental stage) and genetic variation (blue lines represent the expression levels for lines carrying the Bay-0 allele for the most significant QTL and red lines those for the Sha allele carrying lines). QTL LOD profiles are indicated at the bottom of each panel by a heat bar representing the 5 chromosomes. Genetic (LOD) and Environmental (E) variation is expressed as LOD score in the lower left corner.

Future considerations

In conclusion, the study described in this thesis shows that the newly developed tools allow a dissection of seed germination genetics. However, it was expected that overlapping QTL positions between phenotype, metabolite and gene expression variation would uncover many more relationships that could provide leads to disentangle the molecular mechanisms behind the observed phenotypic variation. It can be argued that this is a consequence of several technical reasons such as the used population, repeatability or statistical power consequences due to the used environmental perturbation. Moreover, it might be a warning to retain a certain level of reticence when trying to understand the consequences of evolution within a limited set of experiments. The overwhelming complexity of the genetics which is dissected via either a recombinant inbred population or, even worse, a genome wide association panel of natural variants, should not be underestimated and might therefore always suffer from lack of statistical power to detect causal relations and interactions. Recently, a systematic comparison between the transcriptome architecture in leaf tissue of two RIL populations obtained from a connected-cross design involving 3 commonly used *Arabidopsis* accessions demonstrated the extensive diversity and moderately conserved eQTL landscape between crosses (Cubillos *et al.* 2012). This stresses the need for a wider spectrum of diversity to fully understand expression trait variation within a species. Therefore the trend in genetic analysis is directed towards the use of complex cross populations and genome wide association which allows studying a broad range of natural variants. However, it must be noted that the increase in genetic variation in these populations also increases the complexity to understand possible relations and variation due to developmental or environmental changes. A reduction in genetic complexity might therefore be beneficial and can be achieved by the construction of chromosome substitution lines (Koumproglou *et al.* 2002). The recent developments in reverse breeding, which includes an optimized procedure to construct chromosome substitution lines, holds a great promise for crop improvement and will be a valuable complement to the existing approaches to study polygenic traits in the future (Wijnker *et al.* 2012).

With the recent developments in sequence technology there is evident need for methods that increase our understanding of genotype to phenotype relationships in a high-throughput manner. A wealth of methods to produce high density genotype information is currently available and is not the limiting factor for this type of research. A range of methods for large scale phenotyping using automated screens became available recently and allows in depth analysis of phenotype to environment interactions. For example, the development of an automated method to phenotype seed germination (Chapter 2) turned out to be indispensable for the research described in this thesis. In chapter 6, full genome gene expression phenotyping has been conducted by using microarray analysis. RNA-sequencing will soon become the method of choice for expression profiling. This will generate high quality expression data which can be used to detect sequence

polymorphisms simultaneously. With this wealth of data the need for bio-statistics increases even further. Efficient statistical procedures such as the R/qtl loop (Chapter 3) and the linear model to evaluate genetic x environmental interactions in generalized genetical omics studies (Chapter 5) are needed and should always be combined with effective data visualization (e.g. Chapter 3 and 4) to allow interpretation of the often complex results. We detected new QTLs for seed performance (Chapter 3) which can be co-located with QTLs detected for metabolites (Chapter 5) and gene expression (Chapter 6). The potential of such an integrated approach is shown by two examples in this Chapter. Several new mucilage mutants are discovered while screening the HapMap population (Chapter 7) and interesting global distribution patterns of adaptation were observed.

In conclusion, the tools and concepts described in this thesis should be fueling future research and provide opportunities for efficient improvement of seed performance in crop species.

References

Alonso-Blanco, C., M. G. Aarts, L. Bentsink, J. J. Keurentjes, M. Reymond, et al. (2009). What has natural variation taught us about plant development, physiology, and adaptation? Plant Cell **21**(7): 1877-1896.

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

Alonso-Blanco, C., H. Blankestijn-de Vries, C. J. Hanhart and M. Koornneef (1999). Natural allelic variation at seed size loci in relation to other life history traits of Arabidopsis thaliana. Proceedings of the National Academy of Sciences of the United States of America **96**(8): 4710-4717.

Alonso-Blanco, C. and M. Koornneef (2000). Naturally occurring variation in Arabidopsis: an underexploited resource for plant genetics. Trends Plant Science **5**(1): 22-29.

An, C., S. Saha, J. N. Jenkins, B. E. Scheffler, T. A. Wilkins, et al. (2007). Transcriptome profiling, sequence characterization, and SNP-based chromosomal assignment of the EXPANSIN genes in cotton. Molecular Genetics Genomics **278**(5): 539-553.

Angelovici, R., G. Galili, A. R. Fernie and A. Fait (2010). Seed desiccation: a bridge between maturation and germination. Trends Plant Science **15**(4): 211-218.

Arends, D., P. Prins, R. C. Jansen and K. W. Broman (2010). R/qtl: high-throughput multiple QTL mapping. Bioinformatics **26**(23): 2990-2992.

Argyris, J., P. Dahal, E. Hayashi, D. W. Still and K. J. Bradford (2008). Genetic variation for lettuce seed thermoinhibition is associated with temperature-sensitive expression of abscisic Acid, gibberellin, and ethylene biosynthesis, metabolism, and response genes. Plant Physiology **148**(2): 926-947.

Armengaud, P., K. Zambaux, A. Hills, R. Sulpice, R. J. Pattison, et al. (2009). EZ-Rhizo: Integrated software for the fast and accurate measurement of root system architecture. Plant Journal **57**(5): 945-956.

Arsovski, A. A., G. W. Haughn and T. L. Western (2010). Seed coat mucilage cells of Arabidopsis thaliana as a model for plant cell wall research. Plant Signaling and Behavior **5**(7): 796-801.

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

Asif, M. A., Y. Zafar, J. Iqbal, M. M. Iqbal, U. Rashid, et al. (2011). Enhanced expression of AtNHX1, in transgenic groundnut (Arachis hypogaea L.) improves salt and drought tolerence. Mol Biotechnol **49**(3): 250-256.

Atwell, S., Y. S. Huang, B. J. Vilhjalmsson, G. Willems, M. Horton, et al. (2010). Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature **465**(7298): 627-631.

Ayroles, J. F., M. A. Carbone, E. A. Stone, K. W. Jordan, R. F. Lyman, et al. (2009). Systems genetics of complex traits in Drosophila melanogaster. Nature Genetics **41**(3): 299-307.

Baecker, P. and P. Travo (2006). Cell Image Analyzer: A visual scripting interface for ImageJ and its usage at the microscopy facility Montpellier RIO. Proceedings of the ImageJ User and Developer Conference **1**: 105-110.

Baerenfaller, K., J. Grossmann, M. A. Grobei, R. Hull, M. Hirsch-Hoffmann, et al. (2008). Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. Science **320**(5878): 938-941.

Barragan, V., E. O. Leidi, Z. Andres, L. Rubio, A. De Luca, et al. (2012). Ion exchangers NHX1 and NHX2 mediate active potassium uptake into vacuoles to regulate cell turgor and stomatal function in Arabidopsis. Plant Cell **24**(3): 1127-1142.

Barriere, Y., A. Laperche, L. Barrot, G. Aurel, M. Briand, et al. (2005). QTL analysis of lignification and cell wall digestibility in the Bay-0 x Shahdara RIL progeny of Arabidopsis thaliana as a model system for forage plant. Plant Science **168**(5): 1235-1245.

Barua, D., C. Butler, T. E. Tisdale and K. Donohue (2011). Natural variation in germination responses of Arabidopsis to seasonal cues and their associated physiological mechanisms. Annals of Botany.

Bassel, G. W., P. Fung, T. F. Chow, J. A. Foong, N. J. Provart, et al. (2008). Elucidating the germination transcriptional program using small molecules. Plant Physiology **147**(1): 143-155.

Baxter, I., J. N. Brazelton, D. Yu, Y. S. Huang, B. Lahner, et al. (2010). A coastal cline in sodium accumulation in Arabidopsis thaliana is driven by natural variation of the sodium transporter AtHKT1;1. PLoS Genetics **6**(11): e1001193.

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

Bentsink, L., J. Hanson, C. J. Hanhart, H. Blankestijn-de Vries, C. Coltrane, et al. (2010). Natural variation for seed dormancy in Arabidopsis is regulated by additive genetic and molecular pathways. Proceedings of the National Academy of Sciences of the United States of America **107**(9): 4264-4269.

Bewley, J. D. (1997). Seed Germination and Dormancy. Plant Cell **9**(7): 1055-1066.

Binder, S. (2010). Branched-Chain Amino Acid Metabolism in Arabidopsis thaliana. The Arabidopsis Book: e0137.

Bing, N. and I. Hoeschele (2005). Genetical genomics analysis of a yeast segregant population for transcription network inference. Genetics **170**(2): 533-542.

Blasing, O. E., Y. Gibon, M. Gunther, M. Hohne, R. Morcuende, et al. (2005). Sugars and Circadian Regulation Make Major Contributions to the Global Regulation of Diurnal Gene Expression in Arabidopsis. Plant Cell **17**(12): 3257-3281.

Boer, M. P., D. Wright, L. Feng, D. W. Podlich, L. Luo, et al. (2007). A mixed-model quantitative trait loci (QTL) analysis for multiple-environment trial data using environmental covariables for QTL-by-environment interactions, with an example in maize. Genetics **177**(3): 1801-1813.

Borevitz, J. O. and J. Chory (2004). Genomics tools for QTL analysis and gene discovery. Current Opinion in Plant Biology **7**(2): 132-136.

Bouche, N. and H. Fromm (2004). GABA in plants: just a metabolite? Trends Plant Science **9**(3): 110-115.

Brachi, B., N. Faure, M. Horton, E. Flahauw, A. Vazquez, et al. (2010). Linkage and association mapping of Arabidopsis thaliana flowering time in nature. PLoS Genetics **6**: e1000940.

Brazma, A., H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, et al. (2003). ArrayExpress - A public repository for microarray gene expression data at the EBI. Nucleic Acids Research **31**(1): 68-71.

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

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

Broman, K. W., H. Wu, S. Sen and G. A. Churchill (2003). R/qtl: QTL mapping in experimental crosses. Bioinformatics **19**(7): 889-890.

Brown, R. F. and D. G. Mayer (1988). Representing cumulative germination. 2. The use of the Weibull function and other empirically derived curves. Annals of Botany **61**(2): 127-138.

Buckler, E. and M. Gore (2007). An Arabidopsis haplotype map takes root. Nature Genetics **39**(9): 1056-1057.

Burke, J. M., J. C. Burger and M. A. Chapman (2007). Crop evolution: from genetics to genomics. Current Opinion in Genetics and Development **17**(6): 525-532.

Bystrykh, L., E. Weersing, B. Dontje, S. Sutton, M. T. Pletcher, et al. (2005). Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. Nature Genetics **37**(3): 225-232.

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

Calenge, F., V. Saliba-Colombani, S. Mahieu, O. Loudet, F. Daniel-Vedele, et al. (2006). Natural variation for carbohydrate content in Arabidopsis. Interaction with complex traits dissected by quantitative genetics. Plant Physiology **141**(4): 1630-1643.

Carreno-Quintero, N., A. Acharjee, C. Maliepaard, C. Bachem, R. Mumm, et al. (2012). Untargeted metabolic quantitative trait loci (mQTL) analyses reveal a relationship between primary metabolism and potato tuber quality. Plant Physiology.

Carrera, E., T. Holman, A. Medhurst, W. Peer, H. Schmuths, et al. (2007). Gene expression profiling reveals defined functions of the ATP-binding cassette transporter COMATOSE late in phase II of germination. Plant Physiology **143**(4): 1669-1679.

Catusse, J., C. Job and D. Job (2008). Transcriptome- and proteome-wide analyses of seed germination. Comptes Rendus - Biologies **331**(10): 815-822.

Chan, E. K., H. C. Rowe, J. A. Corwin, B. Joseph and D. J. Kliebenstein (2011). Combining genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in Arabidopsis thaliana. PLoS Biol **9**(8): e1001125.

Chesler, E. J., L. Lu, S. Shou, Y. Qu, J. Gu, et al. (2005). Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. Nature Genetics **37**(3): 233-242.

Chevalier, D., M. Batoux, L. Fulton, K. Pfister, R. K. Yadav, et al. (2005). STRUBBELIG defines a receptor kinase-mediated signaling pathway regulating organ development in

Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America **102**(25): 9074-9079.

Chevalier, F., O. Martin, V. Rofidal, A. D. Devauchelle, S. Barteau, et al. (2004). Proteomic investigation of natural variation between Arabidopsis ecotypes. Proteomics **4**(5): 1372-1381.

Chiang, G. C., D. Barua, E. M. Kramer, R. M. Amasino and K. Donohue (2009). Major flowering time gene, flowering locus C, regulates seed germination in Arabidopsis thaliana. Proceedings of the National Academy of Sciences of the United States of America **106**(28): 11661-11666.

Chinnusamy, V., K. Schumaker and J. K. Zhu (2004). Molecular genetic perspectives on cross-talk and specificity in abiotic stress signalling in plants. Journal of Experimental Botany **55**(395): 225-236.

Churchill, G. A. (2002). Fundamentals of experimental design for cDNA microarrays. Nat Genet **32 Suppl**: 490-495.

Clark, R. M., G. Schweikert, C. Toomajian, S. Ossowski, G. Zeller, et al. (2007). Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. Science **317**(5836): 338-342.

Clerkx, E. J. M., M. E. El-Lithy, E. Vierling, G. J. Ruys, H. Blankestijn-De Vries, et al. (2004). Analysis of natural allelic variation of Arabidopsis seed germination and seed longevity traits between the accessions Landsberg erecta and Shakdara, using a new recombinant inbred line population. Plant Physiology **135**(1): 432-443.

Consortium, T. G. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. Nature **485**(7400): 635-641.

Cowles, C. R., J. N. Hirschhorn, D. Altshuler and E. S. Lander (2002). Detection of regulatory variation in mouse genes. Nature Genetics **32**(3): 432-437.

Cubillos, F., J. Yansouni, H. Khalili, S. Balzergue, S. Elftieh, et al. (2012). Expression variation in connected recombinant populations of Arabidopsis thaliana highlights distinct transcriptome architectures. BMC Genomics **13**(1): 117.

Daszkowska-Golec, A. (2011). Arabidopsis Seed Germination Under Abiotic Stress as a Concert of Action of Phytohormones. OMICS.

de Oliveira Dal'Molin, C. G., L. E. Quek, R. W. Palfreyman, S. M. Brumbley and L. K. Nielsen (2010). AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis. Plant Physiology **152**(2): 579-589.

De Vos, R. C. H., S. Moco, A. Lommen, J. J. B. Keurentjes, R. J. Bino, et al. (2007). Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. Nature Protocols **2**(4): 778-791.

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

Dechaine, J. M., G. Gardner and C. Weinig (2009). Phytochromes differentially regulate seed germination responses to light quality and temperature cues during seed maturation. Plant Cell Environment **32**(10): 1297-1309.

DeCook, R., S. Lall, D. Nettleton and S. H. Howell (2006). Genetic regulation of gene expression during shoot development in Arabidopsis. Genetics **172**(2): 1155-1164.

Dell'Aquila, A. (2004). Cabbage, lentil, pepper and tomato seed germination monitored by an image analysis system. Seed Science and Technology **32**(1): 225-229.

Dell'Aquila, A. (2005). The use of image analysis to monitor the germination of seeds of broccoli (Brassica oleracea) and radish (Raphanus sativus). Annals of Applied Biology **146**(4): 545-550.

Dell'Aquila, A. (2007). Pepper seed germination assessed by combined X-radiography and computer-aided imaging analysis. Biologia Plantarum **51**(4): 777-781.

Dell'Aquila, A. (2009). Digital imaging information technology applied to seed germination testing. A review. Agronomy for Sustainable Development **29**(1): 213-221.

Dell'Aquila, A., J. W. Van Eck and G. W. A. M. Van Der Heijden (2000). The application of image analysis in monitoring the imbibition process of white cabbage (Brassica oleracea L.) seeds. Seed Science Research **10**(2): 163-169.

Dias, P. M., S. Brunel-Muguet, C. Durr, T. Huguet, D. Demilly, et al. (2011). QTL analysis of seed germination and pre-emergence growth at extreme temperatures in Medicago truncatula. Theoretical and Applied Genetics **122**(2): 429-444.

Diaz, C., V. Saliba-Colombani, O. Loudet, P. Belluomo, L. Moreau, et al. (2006). Leaf yellowing and anthocyanin accumulation are two genetically independent strategies in response to nitrogen limitation in Arabidopsis thaliana. Plant and Cell Physiology **47**(1): 74-83.

Dickson, M. H. (1980). Genetic aspects of seed quality. Horticultural Science **15**: 771-774.

Dixon, A. L., L. Liang, M. F. Moffatt, W. Chen, S. Heath, et al. (2007). A genome-wide association study of global gene expression. Nature Genetics **39**(10): 1202-1207.

Doerge, R. W. (2002). Mapping and analysis of quantitative trait loci in experimental populations. Nature Reviews Genetics **3**(1): 43-52.

Doganlar, S., A. Frary and S. D. Tanksley (2000). The genetic basis of seed-weight variation: tomato as a model system. Theoretical and Applied Genetics **100**(8): 1267-1273.

Donahue, J. L., S. R. Alford, J. Torabinejad, R. E. Kerwin, A. Nourbakhsh, et al. (2010). The Arabidopsis thaliana Myo-inositol 1-phosphate synthase1 gene is required for Myo-inositol synthesis and suppression of cell death. Plant Cell **22**(3): 888-903.

Doss, S., E. E. Schadt, T. A. Drake and A. J. Lusis (2005). Cis-acting expression quantitative trait loci in mice. Genome Research **15**(5): 681-691.

Dowdle, J., T. Ishikawa, S. Gatzek, S. Rolinski and N. Smirnoff (2007). Two genes in Arabidopsis thaliana encoding GDP-L-galactose phosphorylase are required for ascorbate biosynthesis and seedling viability. Plant Journal **52**(4): 673-689.

Ducournau, S., A. Feutry, P. Plainchault, P. Revollon, B. Vigouroux, et al. (2005). Using computer vision to monitor germination time course of sunflower (Helianthus annuus L.) seeds. Seed Science and Technology **33**(2): 329-340.

Eeuwijk, F. A. V., M. Malosetti and M. P. Boer (2006). Modelling the genetic basis of response curves underlying genotype x environment interaction. Wageningen UR Frontis Series, Scale and Complexity in Plant Systems **21**(Research: Gene-Plant-Crop Relations. 2006).

El-Kassaby, Y. A., I. Moss, D. Kolotelo and M. Stoehr (2008). Seed germination: Mathematical representation and parameters extraction. Forest Science **54**(2): 220-227.

Elwell, A. L., D. S. Gronwall, N. D. Miller, E. P. Spalding and T. L. Brooks (2011). Separating parental environment from seed size effects on next generation growth and development in Arabidopsis. Plant Cell Environment **34**(2): 291-301.

Emilsson, V., G. Thorleifsson, B. Zhang, A. S. Leonardson, F. Zink, et al. (2008). Genetics of gene expression and its effect on disease. Nature **452**(7186): 423-428.

Espinosa-Ruiz, A., J. M. Bellés, R. Serrano and F. A. Culiáñez-Macià (1999). Arabidopsis thaliana AtHAL3: A flavoprotein related to salt and osmotic tolerance and plant growth. Plant Journal **20**(5): 529-539.

Fait, A., R. Angelovici, H. Less, I. Ohad, E. Urbanczyk-Wochniak, et al. (2006). Arabidopsis Seed Development and Germination Is Associated with Temporally Distinct Metabolic Switches. Plant Physiology **142**(3): 839-854.

Fait, A., A. N. Nesi, R. Angelovici, M. Lehmann, P. A. Pham, et al. (2011). Targeted Enhancement of Glutamate-to-gamma-Aminobutyrate Conversion in Arabidopsis Seeds Affects Carbon-Nitrogen Balance and Storage Reserves in a Development-Dependent Manner. Plant Physiology **157**(3): 1026-1042.

Feng, S., C. Martinez, G. Gusmaroli, Y. Wang, J. Zhou, et al. (2008). Coordinated regulation of Arabidopsisthaliana development by light and gibberellins. Nature **451**(7177): 475-479.

Fiehn, O., J. Kopka, P. Dormann, T. Altmann, R. N. Trethewey, et al. (2000). Metabolite profiling for plant functional genomics. Nature Biotechnol **18**(11): 1157-1161.

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

Finch-Savage, W. E. and G. Leubner-Metzger (2006). Seed dormancy and the control of germination. New Phytology **171**(3): 501-523.

Finkelstein, R., W. Reeves, T. Ariizumi and C. Steber (2008). Molecular aspects of seed dormancy. Annual Review of Plant Biology. **59:** 387-415.

Finkelstein, R. R., S. S. Gampala and C. D. Rock (2002). Abscisic acid signaling in seeds and seedlings. Plant Cell **14 Suppl**: S15-45.

Finkelstein, R. R. and C. R. Somerville (1990). Three Classes of Abscisic Acid (ABA)-Insensitive Mutations of Arabidopsis Define Genes that Control Overlapping Subsets of ABA Responses. Plant Physiology **94**(3): 1172-1179.

Fletcher, J. C. (2001). The ULTRAPETALA gene controls shoot and floral meristem size in Arabidopsis. Development **128**(8): 1323-1333.

Foolad, M. R., P. Subbiah and L. Zhang (2007). Common QTL affect the rate of tomato seed germination under different stress and nonstress conditions. International Journal of Plant Genomics **2007**: 97386.

Foolad, M. R., L. P. Zhang and P. Subbiah (2003). Genetics of drought tolerance during seed germination in tomato: inheritance and QTL mapping. Genome **46**(4): 536-545.

Franklin, K. A. and P. H. Quail (2010). Phytochrome functions in Arabidopsis development. Journal of Experimental Botany **61**(1): 11-24.

French, A., S. Ubeda-Tomas, T. J. Holman, M. J. Bennett and T. Pridmore (2009). High-throughput quantification of root growth using a novel image-analysis tool. Plant Physiology **150**(4): 1784-1795.

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

Fujiki, Y., T. Sato, M. Ito and A. Watanabe (2000). Isolation and characterization of cDNA clones for the e1beta and E2 subunits of the branched-chain alpha-ketoacid dehydrogenase complex in Arabidopsis. Journal Biological Chemistry **275**(8): 6007-6013.

Gallardo, K., C. Job, S. P. C. Groot, M. Puype, H. Demol, et al. (2001). Proteomic analysis of Arabidopsis seed germination and priming. Plant Physiology **126**(2): 835-848.

Galpaz, N. and M. Reymond (2010). Natural variation in Arabidopsis thaliana revealed a genetic network controlling germination under salt stress. PLoS ONE **5**(12): e15198.

Gerrits, A., Y. Li, B. M. Tesson, L. V. Bystrykh, E. Weersing, et al. (2009). Expression quantitative trait loci are highly sensitive to cellular differentiation state. PLoS Genet **5**(10): e1000692.

Gibson, G. and B. Weir (2005). The quantitative genetics of transcription. Trends in Genetics **21**(11): 616-623.

Gilad, Y. and J. Borevitz (2006). Using DNA microarrays to study natural variation. Current Opinion in Genetics and Development **16**(6): 553-558.

Goda, H., E. Sasaki, K. Akiyama, A. Maruyama-Nakashita, K. Nakabayashi, et al. (2008). The AtGenExpress hormone and chemical treatment data set: Experimental design, data evaluation, model data analysis and data access. Plant Journal **55**(3): 526-542.

Goff, S. A., D. Ricke, T. H. Lan, G. Presting, R. Wang, et al. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science **296**(5565): 92-100.

Göring, H. H. H., J. E. Curran, M. P. Johnson, T. D. Dyer, J. Charlesworth, et al. (2007). Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nature Genetics **39**(10): 1208-1216.

Gupta, V., S. Mathur, A. U. Solanke, M. K. Sharma, R. Kumar, et al. (2009). Genome analysis and genetic enhancement of tomato. Critical reviews in biotechnology **29**(2): 152-181.

Gutterman, Y. (2000). Maternal effects on seeds during development. The ecology of regeneration in plant communities. M.Fenner. Wallingford, CABI: 59-84.

Haley, C. S. and S. A. Knott (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69**(4): 315-324.

Hansen, B. G., B. A. Halkier and D. J. Kliebenstein (2008). Identifying the molecular basis of QTLs: eQTLs add a new dimension. Trends in Plant Science **13**(2): 72-77.

Harada, J. J. (1997). Seed maturation and control of germination. Cellular and molecular biology of plant seed development **Larkins, B and Vasil, I (Ed.)**(Dordrecht, Kluwer Academic publishers): 545-592.

Hilhorst, H. W. M. and P. E. Toorop (1997). Review on Dormancy, Germinability, and Germination in Crop and Weed Seeds. Advances in Agronomy. **61:** 111-165.

Himmelbach, A., T. Hoffmann, M. Leube, B. Hohener and E. Grill (2002). Homeodomain protein ATHB6 is a target of the protein phosphatase ABI1 and regulates hormone responses in Arabidopsis. EMBO Journal **21**(12): 3029-3038.

Hirschhorn, J. N. and M. J. Daly (2005). Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics **6**(2): 95-108.

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

Holdsworth, M. J., L. Bentsink and W. J. J. Soppe (2008a). Molecular networks regulating Arabidopsis seed maturation, after-ripening, dormancy and germination. New Phytologist **179**(1): 33-54.

Holdsworth, M. J., W. E. Finch-Savage, P. Grappin and D. Job (2008b). Post-genomics dissection of seed dormancy and germination. Trends in Plant Science **13**(1): 7-13.

Hong, S. W. and E. Vierling (2000). Mutants of Arabidopsis thaliana defective in the acquisition of tolerance to high temperature stress. Proceedings of the National Academy of Sciences of the United States of America **97**(8): 4392-4397.

Horiguchi, G., G. T. Kim and H. Tsukaya (2005). The transcription factor AtGRF5 and the transcription coactivator AN3 regulate cell proliferation in leaf primordia of Arabidopsis thaliana. Plant Journal **43**(1): 68-78.

Howell, K. A., R. Narsai, A. Carroll, A. Ivanova, M. Lohse, et al. (2009). Mapping metabolic and transcript temporal switches during germination in rice highlights specific transcription factors and the role of RNA instability in the germination process. Plant Physiology **149**(2): 961-980.

Hsieh, W. P., G. Passador-Gurgel, E. A. Stone and G. Gibson (2007). Mixture modeling of transcript abundance classes in natural populations. Genome Biology **8**(6).

Huang, X., M. J. Paulo, M. Boer, S. Effgen, P. Keizer, et al. (2011). Analysis of natural allelic variation in Arabidopsis using a multiparent recombinant inbred line population. Proceedings of the National Academy of Sciences of the United States of America **108**(11): 4488-4493.

Huang, X., J. Schmitt, L. Dorn, C. Griffith, S. Effgen, et al. (2010). The earliest stages of adaptation in an experimental plant population: Strong selection on QTLS for seed dormancy. Molecular Ecology **19**(7): 1335-1351.

Hubner, N., C. A. Wallace, H. Zimdahl, E. Petretto, H. Schulz, et al. (2005). Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nature Genetics **37**(3): 243-253.

Hundertmark, M., J. Buitink, O. Leprince and D. K. Hincha (2011). The reduction of seed-specific dehydrins reduces seed longevity in Arabidopsis thaliana. Seed Science Research **21**(03): 165-173.

Hunt, L., M. J. Holdsworth and J. E. Gray (2007). Nicotinamidase activity is important for germination. Plant Journal **51**(3): 341-351.

Jaillon, O., J. M. Aury, B. Noel, A. Policriti, C. Clepet, et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature **449**(7161): 463-467.

Jansen, R. C. (1993). Interval mapping of multiple quantitative trait loci. Genetics **135**(1): 205-211.

Jansen, R. C. (2008). Quantitative Trait Loci in Inbred Lines. Handbook of Statistical Genetics, John Wiley & Sons, Ltd: 587-622.

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

Jansen, R. C., J. W. Van Ooijen, P. Stam, C. Lister and C. Dean (1994). Genotype-by-environment interaction in genetic mapping of multiple quantitative trait loci. Theoretical and Applied Genetics **91**(1): 33-37.

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

Jin, H., V. Vacic, T. Girke, S. Lonardi and J. K. Zhu (2008). Small RNAs and the regulation of cis-natural antisense transcripts in Arabidopsis. BMC Molecular Biology **9**.

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

Joosen, R. V. L., J. Kodde, L. A. Willems, W. Ligterink, L. H. van der Plas, et al. (2010). Germinator: a software package for high-throughput scoring and curve fitting of Arabidopsis seed germination. Plant Journal **62**(1): 148-159.

Joosen, R. V. L., W. Ligterink, B. J. W. Dekkers and H. W. M. Hilhorst (2011). Visualization of molecular processes associated with seed dormancy and germination using MapMan. Seed Science Research **21**(02): 143-152.

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

Jordan, M. C., D. J. Somers and T. W. Banks (2007). Identifying regions of the wheat genome controlling seed development by mapping expression quantitative trait loci. Plant Biotechnol Journal **5**(3): 442-453.

Juenger, T. E., J. K. McKay, N. Hausmann, J. J. B. Keurentjes, S. Sen, et al. (2005). Identification and characterization of QTL underlying wholeplant physiology in Arabidopsis thaliana: ?13C, stomatal conductance and transpiration efficiency. Plant, Cell and Environment **28**(6): 697-708.

Juenger, T. E., T. Wayne, S. Boles, V. V. Symonds, J. McKay, et al. (2006). Natural genetic variation in whole-genome expression in Arabidopsis thaliana: The impact of physiological QTL introgression. Molecular Ecology **15**(5): 1351-1365.

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S. Y. Kong, et al. (2010). Variance component model to account for sample structure in genome-wide association studies. Nature Genetics **42**(4): 348-354.

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, et al. (2008). Efficient control of population structure in model organism association mapping. Genetics **178**(3): 1709-1723.

Karssen, C. M., D. L. C. Brinkhorst-van der Swan, A. E. Breekland and M. Koornneef (1983). Induction of dormancy during seed development by endogenous abscisic acid: studies on abscisic acid deficient genotypes of <i>Arabidopsis thaliana</i> (L.) Heynh. Planta **157**(2): 158-165.

Kazmi, R. H., N. Khan, L. A. Willems, V. A. N. H. AW, W. Ligterink, et al. (2012). Complex genetics controls natural variation among seed quality phenotypes in a recombinant inbred population of an interspecific cross between Solanum lycopersicum x Solanum pimpinellifolium. Plant Cell Environment **35**(5): 929-951.

Kelly, A. A., A. L. Quettier, E. Shaw and P. J. Eastmond (2011). Seed storage oil mobilization is important but not essential for germination or seedling establishment in Arabidopsis. Plant Physiology **157**(2): 866-875.

Keurentjes, J. J., J. Fu, I. R. Terpstra, J. M. Garcia, G. van den Ackerveken, et al. (2007). Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. Proceedings of the National Academy of Sciences of the United States of America **104**(5): 1708-1713.

Keurentjes, J. J., M. Koornneef and D. Vreugdenhil (2008). Quantitative genetics in the age of omics. Current Opinion in Plant Biology **11**(2): 123-128.

Keurentjes, J. J. and R. Sulpice (2009). The role of natural variation in dissecting genetic regulation of primary metabolism. Plant Signaling and Behavior **4**(3): 244-246.

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

Keurentjes, J. J. B., J. Fu, C. H. R. De Vos, A. Lommen, R. D. Hall, et al. (2006). The genetics of plant metabolism. Nature Genetics **38**(7): 842-849.

Keurentjes, J. J. B., R. Sulpice, Y. Gibon, M. C. Steinhauser, J. Fu, et al. (2008). Integrative analyses of genetic variation in enzyme activities of primary carbohydrate metabolism reveal distinct modes of regulation in Arabidopsis thaliana. Genome Biology **9**(8).

Khajeh-Hosseini, M., A. Lomholt and S. Matthews (2009). Mean germination time in the laboratory estimates the relative vigour and field performance of commercial seed lots of maize (Zea mays L.). Seed Science and Technology **37**: 446-456.

Khan, N., R. H. Kazmi, L. A. Willems, A. W. van Heusden, W. Ligterink, et al. (2012). Exploring the natural variation for seedling traits and their link with seed dimensions in tomato. PLoS ONE **7**(8): e43991.

Kilian, J., D. Whitehead, J. Horak, D. Wanke, S. Weinl, et al. (2007). The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. Plant Journal **50**(2): 347-363.

Kim, Y. O., J. S. Kim and H. Kang (2005). Cold-inducible zinc finger-containing glycine-rich RNA-binding protein contributes to the enhancement of freezing tolerance in Arabidopsis thaliana. Plant Journal **42**(6): 890-900.

Kinnersley, A. M. and F. J. Turano (2000). Gamma Aminobutyric Acid (GABA) and Plant Responses to Stress. Critical Reviews in Plant Sciences **19**(6): 479-509.

Kirst, M., C. J. Basten, A. A. Myburg, Z. B. Zeng and R. R. Sederoff (2005). Genetic architecture of transcript-level variation in differentiating xylem of a eucalyptus hybrid. Genetics **169**(4): 2295-2303.

Kirst, M., A. A. Myburg, J. P. De Leon, M. E. Kirst, J. Scott, et al. (2004). Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of eucalyptus. Plant Physiology **135**(4): 2368-2378.

Kitano, H. (2004). Biological robustness. Nature Reviews Genetics **5**(11): 826-837.

Kliebenstein, D. J., J. Gershenzon and T. Mitchell-Olds (2001). Comparative quantitative trait loci mapping of aliphatic, indolic and benzylic glucosinolate production in Arabidopsis thaliana leaves and seeds. Genetics **159**(1): 359-370.

Kliebenstein, D. J., V. M. Lambrix, M. Reichelt, J. Gershenzon and T. Mitchell-Olds (2001). Gene duplication in the diversification of secondary metabolism: tandem 2-oxoglutarate-dependent dioxygenases control glucosinolate biosynthesis in Arabidopsis. Plant Cell **13**(3): 681-693.

Kliebenstein, D. J., M. A. L. West, H. van Leeuwen, O. Loudet, R. W. Doerge, et al. (2006). Identification of QTLs controlling gene expression networks defined a priori. BMC Bioinformatics **7**.

Kooke, R. and J. J. Keurentjes (2011). Multi-dimensional regulation of metabolic networks shaping plant development and performance. Journal of Experimental Botany.

Koumproglou, R., T. M. Wilkes, P. Townson, X. Y. Wang, J. Beynon, et al. (2002). STAIRS: a new genetic resource for functional genomic studies of Arabidopsis. Plant Journal **31**(3): 355-364.

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

Kucera, B., M. A. Cohn and G. Leubner-Metzger (2005). Plant hormone interactions during seed dormancy release and germination. Seed Science Research **15**(4): 281-307.

Lan, H., M. Chen, J. B. Flowers, B. S. Yandell, D. S. Stapleton, et al. (2006). Combined expression trait correlations and expression quantitative trait locus mapping. PLoS Genetics **2**(1): 51-61.

Laserna, M. P., R. A. Sanchez and J. F. Botto (2008). Light-related Loci Controlling Seed Germination in Ler x Cvi and Bay-0 x Sha Recombinant Inbred-line Populations of Arabidopsis thaliana. Annals of Botany.

Laubinger, S., G. Zeller, S. R. Henz, T. Sachsenberg, C. K. Widmer, et al. (2008). At-TAX: A whole genome tiling array resource for developmental expression analysis and transcript identification in Arabidopsis thaliana. Genome Biology **9**(7).

Le, B. H., C. Cheng, A. Q. Bui, J. A. Wagmaister, K. F. Henry, 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 of the United States of America **107**(18): 8063-8070.

Leach, L. J., Z. Zhang, C. Lu, M. J. Kearsey and Z. Luo (2007). The role of cis-regulatory motifs and genetical control of expression in the divergence of yeast duplicate genes. Molecular Biology and Evolution **24**(11): 2556-2565.

Lee, B. H., J. H. Ko, S. Lee, Y. Lee, J. H. Pak, et al. (2009). The Arabidopsis GRF-INTERACTING FACTOR gene family performs an overlapping function in determining organ size as well as multiple developmental properties. Plant Physiology **151**(2): 655-668.

Lee, K. J., B. J. Dekkers, T. Steinbrecher, C. T. Walsh, A. Bacic, et al. (2012). Distinct cell wall architectures in seed endosperms in representatives of the brassicaceae and solanaceae. Plant Physiology **160**(3): 1551-1566.

Lee, S. I., A. M. Dudley, D. Drubin, P. A. Silver, N. J. Krogan, et al. (2009). Learning a prior on regulatory potential from eQTL data. PLoS Genetics **5**(1).

Lee, S. J., J. Y. Kang, H. J. Park, M. D. Kim, M. S. Bae, et al. (2010). DREB2C interacts with ABF2, a bZIP protein regulating abscisic acid-responsive gene expression, and its overexpression affects abscisic acid sensitivity. Plant Physiology **153**(2): 716-727.

Leidi, E. O., V. Barragan, L. Rubio, A. El-Hamdaoui, M. T. Ruiz, et al. (2010). The AtNHX1 exchanger mediates potassium compartmentation in vacuoles of transgenic tomato. Plant Journal **61**(3): 495-506.

Leon-Kloosterziel, K. M., G. A. van de Bunt, J. A. Zeevaart and M. Koornneef (1996). Arabidopsis mutants with a reduced seed dormancy. Plant Physiology **110**(1): 233-240.

Li, Y., O. A. Alvarez, E. W. Gutteling, M. Tijsterman, J. Fu, et al. (2006). Mapping determinants of gene expression plasticity by genetical genomics in C. elegans. PLoS Genetics **2**(12).

Li, Y., R. Breitling and R. C. Jansen (2008). Generalizing genetical genomics: getting added value from environmental perturbation. Trends in Genetics **24**(10): 518-524.

Li, Y., Y. Huang, J. Bergelson, M. Nordborg and J. O. Borevitz (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**(49): 21199-21204.

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

Li, Y., B. M. Tesson, G. A. Churchill and R. C. Jansen (2010). Critical reasoning on causal inference in genome-wide linkage and association studies. Trends in Genetics **26**(12): 493-498.

Ligterink, W., R. V. L. Joosen and H. W. M. Hilhorst (2012). Unravelling the complex trait of seed quality: using natural variation through a combination of physiology, genetics and -omics technologies. Seed Science Research **22**(SupplementS1): S45-S52.

Lin, Y., L. Sun, L. V. Nguyen, R. A. Rachubinski and H. M. Goodman (1999). The Pex16p Homolog SSE1 and Storage Organelle Formation in Arabidopsis Seeds. Science **284**(5412): 328-330.

Linkies, A., K. Muller, K. Morris, V. Tureckova, M. Wenk, et al. (2009). Ethylene interacts with abscisic acid to regulate endosperm rupture during germination: a comparative approach using Lepidium sativum and Arabidopsis thaliana. Plant Cell **21**(12): 3803-3822.

Lisec, J., R. C. Meyer, M. Steinfath, H. Redestig, M. Becher, et al. (2008). Identification of metabolic and biomass QTL in Arabidopsis thaliana in a parallel analysis of RIL and IL populations. Plant Journal **53**(6): 960-972.

Liu, P. P., N. Koizuka, T. M. Homrichhausen, J. R. Hewitt, R. C. Martin, et al. (2005). Large-scale screening of Arabidopsis enhancer-trap lines for seed germination-associated genes. Plant Journal **41**(6): 936-944.

Loewus, F. A. and P. P. N. Murthy (2000). myo-Inositol metabolism in plants. Plant Science **150**(1): 1-19.

Lommen, A. (2009). MetAlign: Interface-Driven, Versatile Metabolomics Tool for Hyphenated Full-Scan Mass Spectrometry Data Preprocessing. Analytical Chemistry **81**(8): 3079-3086.

Loudet, O., S. Chaillou, C. Camilleri, D. Bouchez and F. Daniel-Vedele (2002). Bay-0 x Shahdara recombinant inbred line population: A powerful tool for the genetic dissection of complex traits in Arabidopsis. Theoretical and Applied Genetics **104**(6-7): 1173-1184.

Loudet, O., S. Chaillou, A. Krapp and F. Daniel-Vedele (2003). Quantitative trait loci analysis of water and anion contents in interaction with nitrogen availability in Arabidopsis thaliana. Genetics **163**(2): 711-722.

Loudet, O., S. Chaillou, P. Merigout, J. Talbotec and F. Daniel-Vedele (2003). Quantitative trait loci analysis of nitrogen use efficiency in Arabidopsis. Plant Physiology **131**(1): 345-358.

Loudet, O., V. Gaudon, A. Trubuil and F. Daniel-Vedele (2005). Quantitative trait loci controlling root growth and architecture in Arabidopsis thaliana confirmed by heterogeneous inbred family. Theoretical and Applied Genetics **110**(4): 742-753.

Loudet, O., T. P. Michael, B. T. Burger, C. Le Mette, T. C. Mockler, et al. (2008). A zinc knuckle protein that negatively controls morning-specific growth in Arabidopsis thaliana. Proceedings of the National Academy of Sciences of the United States of America **105**(44): 17193-17198.

Loudet, O., V. Saliba-Colombani, C. Camilleri, F. Calenge, V. Gaudon, et al. (2007). Natural variation for sulfate content in Arabidopsis thaliana is highly controlled by APR2. Nature Genetics **39**(7): 896-900.

Luo, Y., G. Qin, J. Zhang, Y. Liang, Y. Song, et al. (2011). d-myo-Inositol-3-Phosphate Affects Phosphatidylinositol-Mediated Endomembrane Function in Arabidopsis and Is Essential for Auxin-Regulated Embryogenesis. The Plant Cell Online **23**(4): 1352-1372.

MacKenzie, K. and C. Hackett (2012). Association mapping in a simulated barley population. Euphytica **183**(3): 337-347.

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

Malosetti, M., J. Voltas, I. Romagosa, S. E. Ullrich and F. A. van Eeuwijk (2004). Mixed models including environmental covariables for studying QTL by environment interaction. Euphytica **137**(1): 139-145.

Matsui, A., J. Ishida, T. Morosawa, Y. Mochizuki, E. Kaminuma, et al. (2008). Arabidopsis transcriptome analysis under drought, cold, high-salinity and ABA treatment conditions using a tiling array. Plant and Cell Physiology **49**(8): 1135-1149.

Meng, P.-H., A. Macquet, O. Loudet, A. Marion-Poll and H. M. North (2008). Analysis of Natural Allelic Variation Controlling Arabidopsis thaliana Seed Germinability in Response to Cold and Dark: Identification of Three Major Quantitative Trait Loci. Molecular Plant **1**(1): 145-154.

Meyer, R. C., M. Steinfath, J. Lisec, M. Becher, H. Witucka-Wall, et al. (2007). The metabolic signature related to high plant growth rate in Arabidopsis thaliana. Proceedings of the National Academy of Sciences of the United States of America **104**(11): 4759-4764.

Ming, R., S. Hou, Y. Feng, Q. Yu, A. Dionne-Laporte, et al. (2008). The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). Nature **452**(7190): 991-996.

Mitchell-Olds, T. and J. Schmitt (2006). Genetic mechanisms and evolutionary significance of natural variation in Arabidopsis. Nature **441**(7096): 947-952.

Mitsuhashi, N., M. Kondo, S. Nakaune, M. Ohnishi, M. Hayashi, et al. (2008). Localization of myo-inositol-1-phosphate synthase to the endosperm in developing seeds of Arabidopsis. Journal of Experimental Botany **59**(11): 3069-3076.

Monks, S. A., A. Leonardson, H. Zhu, P. Cundiff, P. Pietrusiak, et al. (2004). Genetic inheritance of gene expression in human cell lines. American Journal of Human Genetics **75**(6): 1094-1105.

Moreau, L., A. Charcosset and A. Gallais (2004). Use of trial clustering to study QTL x environment effects for grain yield and related traits in maize. Theoretical and Applied Genetics **110**(1): 92-105.

Moyers, B. T. and N. C. Kane (2010). The genetics of adaptation to novel environments: Selection on germination timing in Arabidopsis thaliana. Molecular Ecology **19**(7): 1270-1272.

Müller, K., S. Tintelnot and G. Leubner-Metzger (2006). Endosperm-limited Brassicaceae seed germination: abscisic acid inhibits embryo-induced endosperm weakening of Lepidium sativum (cress) and endosperm rupture of cress and Arabidopsis thaliana. Plant Cell Physiology **47**(7): 864-877.

Munnik, T., W. Ligterink, I. Meskiene, O. Calderini, J. Beyerly, et al. (1999). Distinct osmo-sensing protein kinase pathways are involved in signalling moderate and severe hyper-osmotic stress. Plant Journal **20**(4): 381-388.

Munns, R. and M. Tester (2008). Mechanisms of salinity tolerance. Annual Review of Plant Biology **59**: 651-681.

Murtagh, F. (1985). A Survey of Algorithms for Contiguity-constrained Clustering and Related Problems. The Computer Journal **28**(1): 82-88.

Myers, A. J., J. R. Gibbs, J. A. Webster, K. Rohrer, A. Zhao, et al. (2007). A survey of genetic human cortical gene expression. Nature Genetics **39**(12): 1494-1499.

Myles, S., J. Peiffer, P. J. Brown, E. S. Ersoz, Z. Zhang, et al. (2009). Association mapping: critical considerations shift from genotyping to experimental design. Plant Cell **21**(8): 2194-2202.

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

Ndimba, B. K., S. Chivasa, W. J. Simon and A. R. Slabas (2005). Identification of Arabidopsis salt and osmotic stress responsive proteins using two-dimensional difference gel electrophoresis and mass spectrometry. Proteomics **5**(16): 4185-4196.

Nesi, N., R. Delourme, M. Bregeon, C. Falentin and M. Renard (2008). Genetic and molecular approaches to improve nutritional value of Brassica napus L. seed. Comptes rendus biologies **331**(10): 763-771.

Nonogaki, H. (2006). Seed germination - The biochemical and molecular mechanisms. Breeding Science **56**(2): 93-105.

Nonogaki, H., F. Chen and K. J. Bradford (2007). Mechanisms and Genes Involved in Germination *Sensu Stricto*. Annual Plant reviews, Seed development, dormancy and germination **27**(chapter 11): 40.

Nordborg, M. and D. Weigel (2008). Next-generation genetics in plants. Nature **456**(7223): 720-723.

Orsi, C. H. and S. D. Tanksley (2009). Natural variation in an ABC transporter gene associated with seed size evolution in tomato species. PLoS Genetics **5**(1): e1000347.

Ossowski, S., K. Schneeberger, R. M. Clark, C. Lanz, N. Warthmann, et al. (2008). Sequencing of natural strains of Arabidopsis thaliana with short reads. Genome Research **18**(12): 2024-2033.

Palanivelu, R., L. Brass, A. F. Edlund and D. Preuss (2003). Pollen tube growth and guidance is regulated by POP2, an Arabidopsis gene that controls GABA levels. Cell **114**(1): 47-59.

Paran, I. and D. Zamir (2003). Quantitative traits in plants: Beyond the QTL. Trends in Genetics **19**(6): 303-306.

Paterson, A. H., J. E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood, et al. (2009). The Sorghum bicolor genome and the diversification of grasses. Nature **457**(7229): 551-556.

Payne, R. W., S. A. Harding, D. A. Murray, D. M. Soutar, D. B., et al. (2011). The Guide to GenStat Release 14. Part 2: Statistics. VSN International, Hemel Hempstead, UK.

Penfield, S. and J. King (2009). Towards a systems biology approach to understanding seed dormancy and germination. Proceedings of the Royal Society: Biological Sciences.

Penfield, S., Y. Li, A. D. Gilday, S. Graham and I. A. Graham (2006). Arabidopsis ABA INSENSITIVE4 regulates lipid mobilization in the embryo and reveals repression of seed germination by the endosperm. Plant Cell **18**(8): 1887-1899.

Penfield, S., R. C. Meissner, D. A. Shoue, N. C. Carpita and M. W. Bevan (2001). MYB61 is required for mucilage deposition and extrusion in the arabidopsis seed coat. Plant Cell **13**(12): 2777-2791.

Penfield, S., H. Pinfield-Wells and I. A. Graham (2007). Lipid Metabolism in Seed Dormancy. Annual Plant reviews, Seed development, dormancy and germination **27**(chapter 6): 19.

Peregrin-Alvarez, J. M., C. Sanford and J. Parkinson (2009). The conservation and evolutionary modularity of metabolism. Genome Biology **10**(6): R63.

Ping Lou, Jianjun Zhao, Hongju He, Corrie Hanhart, Dunia Pino Del Carpio, et al. (2008). Quantitative trait loci for glucosinolate accumulation in Brassica rapa leaves. New Phytologist **179**(4): 1017-1032.

Platt, A., M. Horton, Y. S. Huang, Y. Li, A. E. Anastasio, et al. (2010). The scale of population structure in Arabidopsis thaliana. PLoS Genetics **6**(2): e1000843.

Poormohammad Kiani, S., P. Grieu, P. Maury, T. Hewezi, L. Gentzbittel, et al. (2007). Genetic variability for physiological traits under drought conditions and differential expression of water stress-associated genes in sunflower (Helianthus annuus L.). Theoretical and Applied Genetics **114**(2): 193-207.

Potokina, E., A. Druka, Z. Luo, R. Wise, R. Waugh, et al. (2008). Gene expression quantitative trait locus analysis of 16 000 barley genes reveals a complex pattern of genome-wide transcriptional regulation. Plant Journal **53**(1): 90-101.

Powell, A. A. (2006). Seed vigor and its assessment. Handbook of seed science and technology Basra, A.S. (Ed.)(Binghamton, USA, Food Products Press.): 603-648.

Price, A. L., N. A. Zaitlen, D. Reich and N. Patterson (2010). New approaches to population stratification in genome-wide association studies. Nature Reviews Genetics 11(7): 459-463.

Quesada, V., S. García-Martínez, P. Piqueras, M. R. Ponce and J. L. Micol (2002). Genetic architecture of NaCl tolerance in Arabidopsis. Plant Physiology 130(2): 951-963.

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

Ren, Z., Z. Zheng, V. Chinnusamy, J. Zhu, X. Cui, et al. (2010). RAS1, a quantitative trait locus for salt tolerance and ABA sensitivity in Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America 107(12): 5669-5674.

Rensink, W. A. and S. P. Hazen (2006). Statistical issues in microarray data analysis. Methods in molecular biology 323: 359-366.

Reymond, M., S. Svistoonoff, O. Loudet, L. Nussaume and T. Desnos (2006). Identification of QTL controlling root growth response to phosphate starvation in Arabidopsis thaliana. Plant, Cell and Environment 29(1): 115-125.

Rivero-Lepinckas, L., D. Crist and R. Scholl (2006). Growth of plants and preservation of seeds. Methods in molecular biology 323: 3-12.

Rockman, M. V. (2008). Reverse engineering the genotype-phenotype map with natural genetic variation. Nature 456(7223): 738-744.

Rockman, M. V. and L. Kruglyak (2006). Genetics of global gene expression. Nature Reviews Genetics 7(11): 862-872.

Roessner, U., C. Wagner, J. Kopka, R. N. Trethewey and L. Willmitzer (2000). Simultaneous analysis of metabolites in potato tuber by gas chromatography–mass spectrometry. Plant Journal 23(1): 131-142.

Routaboul, J. M., C. Dubos, G. Beck, C. Marquis, P. Bidzinski, et al. (2012). Metabolite profiling and quantitative genetics of natural variation for flavonoids in Arabidopsis. Journal of Experimental Botany 63(10): 3749-3764.

Routaboul, J. M., L. Kerhoas, I. Debeaujon, L. Pourcel, M. Caboche, et al. (2006). Flavonoid diversity and biosynthesis in seed of Arabidopsis thaliana. Planta 224(1): 96-107.

Rowe, H. C., B. G. Hansen, B. A. Halkier and D. J. Kliebenstein (2008). Biochemical networks and epistasis shape the Arabidopsis thaliana metabolome. Plant Cell 20(5): 1199-1216.

Running, M. P. and E. M. Meyerowitz (1996). Mutations in the PERIANTHIA gene of Arabidopsis specifically alter floral organ number and initiation pattern. Development 122(4): 1261-1269.

Salvi, S. and R. Tuberosa (2005). To clone or not to clone plant QTLs: Present and future challenges. Trends in Plant Science 10(6): 297-304.

Schadt, E. E., S. A. Monks, T. A. Drake, A. J. Lusis, N. Che, et al. (2003). Genetics of gene expression surveyed in maize, mouse and man. Nature 422(6929): 297-302.

Schauer, N., Y. Semel, U. Roessner, A. Gur, I. Balbo, et al. (2006). Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. Nature Biotechnol **24**(4): 447-454.

Schauer, N., D. Steinhauser, S. Strelkov, D. Schomburg, G. Allison, et al. (2005). GC–MS libraries for the rapid identification of metabolites in complex biological samples. FEBS letters **579**(6): 1332-1337.

Schnable, P. S., D. Ware, R. S. Fulton, J. C. Stein, F. Wei, et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science **326**(5956): 1112-1115.

Seifert, G. J. and K. Roberts (2007). The biology of arabinogalactan proteins. Annual Review of Plant Biology **58**: 137-161.

Sheppard, K., J. Yuan, M. J. Hohn, B. Jester, K. M. Devine, et al. (2008). From one amino acid to another: tRNA-dependent amino acid biosynthesis. Nucleic Acids Research **36**(6): 1813-1825.

Shi, C., A. Uzarowska, M. Ouzunova, M. Landbeck, G. Wenzel, et al. (2007). Identification of candidate genes associated with cell wall digestibility and eQTL (expression quantitative trait loci) analysis in a Flint x Flint maize recombinant inbred line population. BMC Genomics **8**.

Shirley, B. W., W. L. Kubasek, G. Storz, E. Bruggemann, M. Koornneef, et al. (1995). Analysis of Arabidopsis mutants deficient in flavonoid biosynthesis. Plant Journal **8**(5): 659-671.

Shu, X. L., T. Frank, Q. Y. Shu and K. H. Engel (2008). Metabolite profiling of germinating rice seeds. Journal of Agricultural Food Chemistry **56**(24): 11612-11620.

Sicher, R. (2011). Carbon partitioning and the impact of starch deficiency on the initial response of Arabidopsis to chilling temperatures. Plant Science **181**(2): 167-176.

Siva, N. (2008). 1000 Genomes project. Nature biotechnology **26**(3): 256.

Smith, E. N. and L. Kruglyak (2008). Gene-environment interaction in yeast gene expression. PLoS Biol **6**(4): e83.

Sorkheh, K., L. V. Malysheva-Otto, M. G. Wirthensohn, S. Tarkesh-Esfahani and P. Martínez-Gómez (2008). Linkage disequilibrium, genetic association mapping and gene localization in crop plants. Genetics and Molecular Biology **31**(4): 805-814.

Sreenivasulu, N., B. Usadel, A. Winter, V. Radchuk, U. Scholz, et al. (2008). Barley grain maturation and germination: metabolic pathway and regulatory network commonalities and differences highlighted by new MapMan/PageMan profiling tools. Plant Physiology **146**(4): 1738-1758.

Stacey, M. G., H. Osawa, A. Patel, W. Gassmann and G. Stacey (2006). Expression analyses of Arabidopsis oligopeptide transporters during seed germination, vegetative growth and reproduction. Planta **223**(2): 291-305.

Stranger, B. E., A. C. Nica, M. S. Forrest, A. Dimas, C. P. Bird, et al. (2007). Population genomics of human gene expression. Nature Genetics **39**(10): 1217-1224.

Street, N. R., O. Skogstrom, A. Sjodin, J. Tucker, M. Rodriguez-Acosta, et al. (2006). The genetics and genomics of the drought response in Populus. Plant Journal **48**(3): 321-341.

Strehmel, N., J. Hummel, A. Erban, K. Strassburg and J. Kopka (2008). Retention index thresholds for compound matching in GC–MS metabolite profiling. Journal of Chromatography B **871**(2): 182-190.

Stylianou, I. M., J. P. Affourtit, K. R. Shockley, R. Y. Wilpan, F. A. Abdi, et al. (2008). Applying gene expression, proteomics and single-nucleotide polymorphism analysis for complex trait gene identification. Genetics **178**(3): 1795-1805.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America **102**(43): 15545-15550.

Tamura, N., T. Yoshida, A. Tanaka, R. Sasaki, A. Bando, et al. (2006). Isolation and characterization of high temperature-resistant germination mutants of Arabidopsis thaliana. Plant Cell Physiology **47**(8): 1081-1094.

Tan, X., L. I. Calderon-Villalobos, M. Sharon, C. Zheng, C. V. Robinson, et al. (2007). Mechanism of auxin perception by the TIR1 ubiquitin ligase. Nature **446**(7136): 640-645.

Tasma, I. M., V. Brendel, S. A. Whitham and M. K. Bhattacharyya (2008). Expression and evolution of the phosphoinositide-specific phospholipase C gene family in Arabidopsis thaliana. Plant Physiology Biochemistry **46**(7): 627-637.

Teakle, N. L., A. Amtmann, D. Real and T. D. Colmer (2010). Lotus tenuis tolerates combined salinity and waterlogging: maintaining O2 transport to roots and expression of an NHX1-like gene contribute to regulation of Na+ transport. Physiologia Plantarum **139**(4): 358-374.

Teixeira, P. C. N., J. A. Coelho Neto, H. Rocha and J. M. De Oliveira (2007). An instrumental set up for seed germination studies with temperature control and automatic image recording. Brazilian Journal of Plant Physiology **19**(2): 99-108.

Theodoulou, F. L., K. Job, S. P. Slocombe, S. Footitt, M. Holdsworth, et al. (2005). Jasmonic acid levels are reduced in COMATOSE ATP-binding cassette transporter mutants. Implications for transport of jasmonate precursors into peroxisomes. Plant Physiology **137**(3): 835-840.

Thimm, O., O. Bläsing, Y. Gibon, A. Nagel, S. Meyer, et al. (2004). mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant Journal **37**(6): 914-939.

Thoday, J. M. (1961). Location of Polygenes. Nature **191**(4786): 368-370.

Tikunov, Y. M., S. Laptenok, R. D. Hall, A. G. Bovy and C. H. d. Vos (2011). MSClust: a tool for unsupervised mass spectra extraction of chromatography-mass spectrometry ion-wise aligned data (online first). Metabolomics.

Toorop, P. E., R. M. Barroco, G. Engler, S. P. C. Groot and H. W. M. Hilhorst (2005). Differentially expressed genes associated with dormancy or germination of <i>Arabidopsis thaliana seeds. Planta **221**(5): 637-647.

Toufighi, K., S. M. Brady, R. Austin, E. Ly and N. J. Provart (2005). The botany array resource: e-Northerns, expression angling, and promoter analyses. Plant Journal **43**(1): 153-163.

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

Tuskan, G. A., S. Difazio, S. Jansson, J. Bohlmann, I. Grigoriev, et al. (2006). The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science **313**(5793): 1596-1604.

Usadel, B., A. Nagel, D. Steinhauser, Y. Gibon, O. E. Blasing, et al. (2006). PageMan: an interactive ontology tool to generate, display, and annotate overview graphs for profiling experiments. BMC Bioinformatics **7**: 535.

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

van Der Schaar, W., C. Alonso-Blanco, K. M. Leon-Kloosterziel, R. C. Jansen, J. W. van Ooijen, et al. (1997). QTL analysis of seed dormancy in Arabidopsis using recombinant inbred lines and MQM mapping. Heredity **79 (Pt 2)**: 190-200.

Van Leeuwen, H., D. J. Kliebenstein, M. A. L. West, K. Kim, R. Van Poecke, et al. (2007). Natural variation among Arabidopsis thaliana accessions for transcriptome response to exogenous salicylic acid. Plant Cell **19**(7): 2099-2110.

Vandecasteele, C., B. Teulat-Merah, M. C. Morere-Le Paven, O. Leprince, B. Ly Vu, et al. (2011). Quantitative trait loci analysis reveals a correlation between the ratio of sucrose/raffinose family oligosaccharides and seed vigour in Medicago truncatula. Plant Cell Environment **34**(9): 1473-1487.

Venu, R. C., Y. Jia, M. Gowda, M. H. Jia, C. Jantasuriyarat, et al. (2007). RL-SAGE and microarray analysis of the rice transcriptome after Rhizoctonia solani infection. Molecular Genetics Genomics **278**(4): 421-431.

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

Vuylsteke, M., H. Van Den Daele, A. Vercauteren, M. Zabeau and M. Kuiper (2006). Genetic dissection of transcriptional regulation by cDNA-AFLP. Plant Journal **45**(3): 439-446.

Wahid, A., S. Gelani, M. Ashraf and M. R. Foolad (2007). Heat tolerance in plants: An overview. Environmental and Experimental Botany **61**(3): 199-223.

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

Watanabe, K., M. Pacher, S. Dukowic, V. Schubert, H. Puchta, et al. (2009). The STRUCTURAL MAINTENANCE OF CHROMOSOMES 5/6 Complex Promotes Sister Chromatid Alignment and Homologous Recombination after DNA Damage in Arabidopsis thaliana. The Plant Cell Online **21**(9): 2688-2699.

Weckwerth, W., M. E. Loureiro, K. Wenzel and O. Fiehn (2004). Differential metabolic networks unravel the effects of silent plant phenotypes. Proceedings of the National Academy of Sciences of the United States of America **101**(20): 7809-7814.

Wehmeyer, N. and E. Vierling (2000). The expression of small heat shock proteins in seeds responds to discrete developmental signals and suggests a general protective role in desiccation tolerance. Plant Physiology **122**(4): 1099-1108.

Weigel, D. (2012). Natural variation in Arabidopsis: from molecular genetics to ecological genomics. Plant Physiology **158**(1): 2-22.

Weigel, D. and M. Nordborg (2005). Natural variation in Arabidopsis. How do we find the causal genes? Plant Physiology **138**(2): 567-568.

Weitbrecht, K., K. Muller and G. Leubner-Metzger (2011). First off the mark: early seed germination. Journal of Experimental Botany.

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

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

West, M. A. L., H. Van Leeuwen, A. Kozik, D. J. Kliebenstein, R. W. Doerge, et al. (2006). High-density haplotyping with microarray-based expression and single feature polymorphism markers in Arabidopsis. Genome Research **16**(6): 787-795.

Wijnker, E., K. van Dun, C. B. de Snoo, C. L. Lelivelt, J. J. Keurentjes, et al. (2012). Reverse breeding in Arabidopsis thaliana generates homozygous parental lines from a heterozygous plant. Nature Genetics **44**(4): 467-470.

Winter, D., B. Vinegar, H. Nahal, R. Ammar, G. V. Wilson, et al. (2007). An "Electronic Fluorescent Pictograph" browser for exploring and analyzing large-scale biological data sets. PLoS ONE **2**(1): e718.

Wittkopp, P. J., B. K. Haerum and A. G. Clark (2004). Evolutionary changes in cis and trans gene regulation. Nature **430**(6995): 85-88.

Xiong, L., K. S. Schumaker and J. K. Zhu (2002). Cell signaling during cold, drought, and salt stress. Plant Cell **14 Suppl**: S165-183.

Yadav, R. K., L. Fulton, M. Batoux and K. Schneitz (2008). The Arabidopsis receptor-like kinase STRUBBELIG mediates inter-cell-layer signaling during floral development. Developmental Biology **323**(2): 261-270.

Yeung, K. Y., K. M. Dombek, K. Lo, J. E. Mittler, J. Zhu, et al. (2011). Construction of regulatory networks using expression time-series data of a genotyped population. Proc Natl Acad Sci U S A **108**(48): 19436-19441.

Yu, J., S. Hu, J. Wang, G. K. S. Wong, S. Li, et al. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science **296**(5565): 79-92.

Yvert, G., R. B. Brem, J. Whittle, J. M. Akey, E. Foss, et al. (2003). Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. Nature Genetics **35**(1): 57-64.

Zeng, D. L., L. B. Guo, Y. B. Xu, K. Yasukumi, L. H. Zhu, et al. (2006). QTL analysis of seed storability in rice. Plant Breeding **125**(1): 57-60.

Zhang, X. and J. O. Borevitz (2009). Global analysis of allele-specific expression in Arabidopsis thaliana. Genetics **182**(4): 943-954.

Zhang, X., E. J. Richards and J. O. Borevitz (2007). Genetic and epigenetic dissection of cis regulatory variation. Current Opinion in Plant Biology **10**(2): 142-148.

Zhou, S., Z. Zhang, Q. Tang, H. Lan, Y. Li, et al. (2010). Enhanced V-ATPase activity contributes to the improved salt tolerance of transgenic tobacco plants overexpressing vacuolar Na+/H+ antiporter AtNHX1. Biotechnology Letters: 1-6.

Zhou, X. and Z. Su (2007). EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species. BMC Genomics 8: 246.

Zhu, J., P. Sova, Q. Xu, K. M. Dombek, E. Y. Xu, et al. (2012). Stitching together Multiple Data Dimensions Reveals Interacting Metabolomic and Transcriptomic Networks That Modulate Cell Regulation. PLoS Biol 10(4): e1001301.

Zhu, J., B. Zhang, E. N. Smith, B. Drees, R. B. Brem, et al. (2008). Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nature Genetics 40(7): 854-861.

Zhu, J. K. (2002). Salt and drought stress signal transduction in plants. Annual Review of Plant Biology. 53: 247-273.

Zhu, J. K. (2003). Regulation of ion homeostasis under salt stress. Current Opinion in Plant Biology 6(5): 441-445.

Zimmermann, P., L. Hennig and W. Gruissem (2005). Gene-expression analysis and network discovery using Genevestigator. Trends in Plant Science 10(9): 407-409.

Summary

The Netherlands has a long history of plant breeding which has resulted in a leading position in the world with respect to the sales of vegetable seeds. Nowadays high-tech methods are used for crop-production which demands high standards for the quality of the starting materials. While breeding has mainly focused on crop yield and disease resistance in the past, it now becomes equally important to create seeds that rapidly and uniformly germinate under a wide range of production environments. A better understanding of the molecular processes that are underlying seed quality is a crucial first step to enable targeted breeding. In this thesis we describe the results of new methods that were used to map the genetics of seed germination.

For this research we have used the leading plant science model species *Arabidopsis thaliana* which has a short generation time and a fully sequenced genome. Further, the large scientific community working on this model species is providing a wealth of resources ranging from large collections of worldwide accessions, genetic mapping populations, mutants and knowledge about gene, protein and metabolite action. A disadvantage of using Arabidopsis is the small size of the seeds, which requires evaluation of the germination of individual seeds with the use of magnifying glasses. This problem has been solved by using image analysis to create an automated procedure to obtain detailed information for parameters such as rate, uniformity and maximum germination. This procedure, called 'the Germinator', is described in Chapter 2 and has been enthusiastically adopted by the seed community.

Plants cannot walk away from the environment at which the seed is dispersed. To survive and to enable reproduction, plants adapt to the prevailing environment which results in considerable genetic variation. This 'natural variation' is a great resource to study the mechanisms of adaptation. In Chapter 3 we have used two distinct Arabidopsis accessions, one originating from Germany (Bayreuth) and the other from high altitude in the Pamiro-Alay Mountains in Tadjikistan (Shahdara). In contrast to the Bayreuth accession, the Shahdara accession is well adapted to survive harsh conditions and is known to be stress tolerant to a range of environments. A genetic mapping (recombinant inbred line; RIL) population, consisting of 165 lines, that was derived from these two accessions is therefore particularly suitable to locate the genomic regions with genetic differences that influence seed germination. Such genomic regions are commonly referred to as quantitative trait loci (QTL). With help of the Germinator system we were able to evaluate germination of this RIL population under many different conditions. This resulted in a description of the 'genetic landscape of seed performance' in which we identified many QTLs for Arabidopsis seed germination.

QTL regions are often large and identification of the causal gene requires intensive follow up research. We therefore aimed for a high throughput analysis using modern 'omics' techniques to analyze differences in metabolite levels and gene expression between the lines. A method to classify and visualize the vast amount of data derived from such an

approach is described in Chapter 4. The so called genetical 'omics' experiments are expensive and therefore often force researchers to limit their study to a single developmental stage or environment only. A novel generalized setup overcomes this limitation and was tested for metabolite level changes in Chapter 5. This setup offers a unique reduction of experimental load with minimal effect on statistical power and is of great potential in the field of system genetics. Four different developmental stages of seed germination were tested in the RIL population. This approach resulted in a large dataset for which efficient analytical procedures were lacking. Thus, Chapter 5 also includes a description of a newly developed statistical procedure to analyze this type of data. The same approach and material were used in Chapter 6 to evaluate the genetics of genome wide gene expression.

Another approach to zoom in on the molecular mechanisms underlying seed performance is described in Chapter 7. Here, the genetic diversity was maximized by using 360 different Arabidopsis accessions which had been subjected to ultra-high density genotyping. In potential, such a genome wide association (GWA) study can provide high resolution mapping of genetic variation resulting in only a few candidate genes per association for the phenotype under study. Although we were able to replicate experiments over two years with a high level of heritability, no significant associations were found. This emphasizes the need to critically review the power of such an approach for traits that are expected to be determined by many small effect loci.

Finally, closing in on the molecular mechanisms underlying the seed traits that we studied might be possible by a full integration of the datasets that were described in the different chapters. Two examples that show the potential and the complexity of such integration are described in the General Discussion (Chapter 8). Research focused on seed quality does not end here but has gained an impulse by the described new methods and hypotheses to continue on both the fundamental and applied level in the coming years.

Samenvatting

Nederland kent een lange geschiedenis in de plantenveredeling. Dit heeft geresulteerd in een leidende positie in de wereld voor de verkoop van groentezaden. De hightech methodes die momenteel worden gebruikt voor gewasproductie vereisen zeer hoge standaarden voor de kwaliteit van het uitgangsmateriaal. Terwijl de veredeling zich in het verleden vooral heeft gericht op eigenschappen zoals opbrengst en ziekteresistentie, is het nu ook van groot belang om zaden te creëren die snel en uniform kiemen in verschillende productieomgevingen. Een verbeterd inzicht in de moleculaire processen die ten grondslag liggen aan zaadkwaliteit is een cruciale eerste stap om gerichte veredeling mogelijk te maken. In dit proefschrift worden de resultaten beschreven van nieuwe methodes waarmee de genetica van kiemingseigenschappen nauwkeurig in kaart kan worden gebracht.

Voor dit onderzoek hebben wij gebruik gemaakt van de populaire modelplant de Zandraket (*Arabidopsis thaliana*). Deze heeft een zeer korte generatietijd en de volledige DNA volgorde van het genoom is bekend. Daarnaast is er een uitgebreide collectie van wereldwijd verkregen accessies, genetische kartering populaties, mutanten en kennis over gen, eiwit en metaboliet functie beschikbaar. Het nadeel van het werken met de Zandraket is dat de zaden zeer klein zijn waardoor het moment van kieming enkel met behulp van vergrotingsapparatuur kan worden bepaald. Om dit probleem op te lossen hebben we een geautomatiseerd systeem van beeldanalyse ontwikkeld waarmee gedetailleerde informatie over de snelheid, uniformiteit en maximale kieming verkregen kan worden. De hierbij gebruikte procedure, genaamd 'de Germinator', is beschreven in Hoofdstuk 2 en wordt momenteel enthousiast gebruikt door vele onderzoekers.

Planten kunnen niet ontsnappen uit de omgeving waarin het zaad terecht is gekomen en hebben zich daarom tijdens de evolutie aangepast aan hun omgeving om te kunnen overleven en zichzelf te kunnen vermeerderen. Dit heeft geresulteerd in grote genetische variatie. Het bestuderen van deze variatie geeft geweldige mogelijkheden om de mechanismen van aanpassing aan de omgeving te onderzoeken. In Hoofdstuk 3 hebben we hiervoor twee accessies van de Zandraket gebruikt, de eerste werd aangetroffen in Duitsland (Bayreuth) terwijl de tweede op grote hoogte groeit in het Pamiro-Alay gebergte in Tadzjikistan (Shahdara). Deze Shahdara accessie heeft zich, in tegenstelling tot Bayreuth, erg goed aangepast aan overleving in moeilijke omstandigheden. Een genetische karteringspopulatie (recombinante inteelt lijnen; RIL) van 165 lijnen afkomstig van deze beide ondersoorten, is daarom uitermate geschikt om genetische verschillen op het genoom te lokaliseren die betrokken zijn bij de regulatie van kieming. Dit soort regio's worden QTLs (Quantitative Trait Loci) genoemd. Met behulp van het eerder ontwikkelde 'Germinator' systeem waren we in staat om de kiemingseigenschappen van deze RIL populatie onder een groot aantal verschillende omstandigheden te onderzoeken. Dit heeft geresulteerd in een beschrijving van het genetische landschap van zaad eigenschappen

waarin we vele QTLs hebben gevonden die betrokken zijn bij de regulatie van de kieming van Zandraket zaden.

De beschreven genoomregio's zijn echter behoorlijk groot, waardoor het aanwijzen van het causale gen veel vervolgonderzoek vereist. Om dit te omzeilen hebben we ons onderzoek toegespitst op het gebruik van moderne 'omics' technologieën. Hiermee is het mogelijk om de verschillen in een groot aantal metabolieten en genen tussen de lijnen nauwkeurig te analyseren. Een methode om al deze informatie te classificeren en te visualiseren is beschreven in Hoofdstuk 4.

De 'genetische omics' experimenten zijn duur, waardoor onderzoekers er vaak voor kiezen om een studie te beperken tot één enkel weefsel of één enkele conditie. Een nieuwe gegeneraliseerde experimentele opzet voorkomt dit probleem en is in Hoofdstuk 5 getest voor verschillen in metaboliet niveaus. Deze nieuwe opzet leidt tot een vermindering van het benodigde aantal metingen met slechts een minimale invloed op de onderliggende statistiek en heeft veel potentie voor toekomstig onderzoek in het veld van de systeemgenetica. Vier verschillende ontwikkelingsstadia tijdens zaadkieming zijn op deze manier getest in de 165 lijnen van de RIL populatie. Deze aanpak resulteerde in een grote hoeveelheid data waarvoor nog geen efficiënte analyse methode beschikbaar was. Hoofdstuk 5 beschrijft daarom ook een statistische procedure om dit type data te kunnen analyseren. Dezelfde aanpak en hetzelfde materiaal zijn in Hoofdstuk 6 gebruikt om de genetica van de gen expressie van het gehele genoom in kaart te brengen.

Een andere aanpak om de moleculaire mechanismen van zaadkieming te onderzoeken is beschreven in Hoofdstuk 7. Hier hebben we de genetische variatie gemaximaliseerd door gebruik te maken van 360 accessies van de Zandraket waarvan een genotypering met hoge dichtheid beschikbaar is (Genoom Brede Associatie). Dit type onderzoek kan de genetische variatie ontrafelen met een zeer hoge resolutie waardoor er per associatie slechts enkele kandidaat genen overblijven die verantwoordelijk kunnen zijn voor het fenotype dat wordt bestudeerd. Hoewel we in staat waren om de experimenten met een hoge nauwkeurigheid te herhalen, werden er geen significantie associaties gevonden. Dit benadrukt de noodzaak om deze methode kritisch te evalueren voor eigenschappen waarvan verwacht mag worden dat ze gereguleerd worden door veel genen met voor elk gen afzonderlijk een klein effect.

Tot slot hebben we de mogelijkheid onderzocht om meer kennis te verkrijgen van de moleculaire mechanismen die ten grondslag liggen aan de eigenschappen die we hebben bestudeerd door alle datasets uit de verschillende hoofstukken te integreren. In de algemene discussie (Hoofdstuk 8) zijn twee voorbeelden beschreven die de potentie, maar tegelijkertijd ook de complexiteit van een dergelijke benadering laten zien. Het onderzoek naar zaadkwaliteit is hiermee niet beëindigd maar heeft een impuls gekregen door de ontwikkelde nieuwe methoden en hypotheses waarmee zowel op fundamenteel als op toegepast niveau veel vervolgonderzoek gedaan kan worden in de komende jaren.

Dankwoord

Het proefschrift dat voor u ligt is het gevolg van een eenvoudige boswandeling. De vraag om wel of niet te gaan promoveren had me nachten lang wakker gehouden, want ik had het enorm naar mijn zin bij Plant Research International waar ik geweldige collega's en uitdagend werk had. Toch heb ik dankzij de bemoedigende woorden van Joost en Miranda tijdens die boswandeling het besluit genomen om de kans te grijpen en een poging te wagen om een promotieonderzoek te gaan doen. Ik heb er zeker geen spijt van gehad!

Henk, jij nam me op in jouw team en ik voelde me meteen welkom. Je kunt als geen ander een team leiden, ziet kansen en weet mensen te verbinden. Ik heb van jou geleerd om niet de techniek maar de onderzoeksvraag centraal te stellen. Daarmee heb je de richting van het onderzoek bepaald. Hoe vaak heb je me niet gevraagd: "waarom?", waarop ik dan steevast het verkeerde antwoord gaf: "..omdat het kan". Wilco, als dagelijks begeleider was jij mijn wetenschappelijk anker. Simpelweg een knal voor mijn kop vanwege een stom idee, maar tegelijkertijd verder kijken en nieuwe oplossingen verzinnen: zo breng je wetenschap tot leven. Je enthousiasme, optimisme en je eigenwijsheid leverde vurige discussies op, waarin we elkaar zonder omhaal op het scherpst van de snede uitdaagden. We misten bijna een vliegtuig omdat we zaten te fantaseren over de eindeloze mogelijkheden van onze geweldige datasets. Kortom, wetenschap zoals het behoort te zijn: spannend, leuk, vernieuwend en dat zonder ooit een chagrijnig gezicht gezien te hebben! Leo, zonder jou was ik sneller klaar geweest met het schrijven, want dan was er veel minder data gegenereerd. Nooit klagen en keihard werken is de beste typering voor jou. Toch had je redenen genoeg, we hebben samen tienduizenden zaden uitgestrooid en gefotografeerd en hoewel dat zaaddodend werk was heb je daar absoluut geen last van gehad! Ook niet als we na een nachtje doorzakken en veel te weinig slaap letterlijk slaapdronken in het lab stonden. "Het leven is een toverbal" klonk het tussen de planten; wie weet hebben de alcoholdampen de proeven toch beïnvloed. Ik ben blij dat je ook tijdens de promotie als paranimf naast me staat. Rashid and Noorullah, both of you were working on the most important part of the project. A real crop! We had great times both in the lab and during the exciting congresses we visited. I am looking back on a great collaboration in which we shared many ideas on how to analyze the extremely big datasets we gathered. My two students, Leticia and Gabriela were both of great help and enabled the analysis of many aspects of the never ending wish list.

Trots ben ik op het feit dat Prof. Linus van der Plas mijn promotor wilde zijn. Tijdens mijn HLO-stage heb je me begeleid, je was leerstoelhouder van de vakgroep plantenfysiologie toen ik daar als analist werkte en je hebt nu vier jaar lang aan mijn zijde gestaan als promotor. Linus, dank voor alle inspirerende gesprekken, alle hulp en voor de zeer nauwgezette controle van de teksten. Veel dank ben ik ook verschuldigd aan Prof. Maarten Koornneef. Wij hebben vaak overleg gehad en telkens wist jij me binnen een half uur meerdere weken werk te verschaffen door te wijzen op belangrijke ontwikkelingen in de vakliteratuur. Hetzelfde geldt voor Joost, zowel als vriend maar zeker ook in de rol van

wetenschappelijk adviseur heb jij een grote bijdrage geleverd aan de totstandkoming van dit proefschrift. Martijn, als oud-collega's mochten wij nooit een kantoor delen. Kinderachtig natuurlijk, maar niet onbegrijpelijk want als duo zijn we niet te hanteren. We hebben onze "raakvlakken" inmiddels verplaatst naar de squashbaan. Ik ben dan ook blij dat jij als paranimf aan mijn zijde staat.

Next to the project team I am left with great memories about all members of the Wageningen seed lab. Our sometimes hilarious work discussions but also the many lunches, trips, dinners and parties will always bring back a smile. Now I understand why so many of our guests were crying during their farewell. Rina, de steun en toeverlaat van ons team wist de weg in de onoverkomelijke administratieve rompslomp en was een grote hulp bij het boeken van allerlei reizen. De lijfstraf met de houten liniaal zal ik niet snel vergeten! Ruim 5000 planten zijn er opgegroeid onder toeziend oog van Taede en Gerard. Jullie waren meteen enthousiast over mijn wens om steenwol te gaan gebruiken. Mede dankzij dit enthousiasme staan de kassen nu vol met zandraketten op steenwol.

Het is een eenvoudige hypothese dat er een significante kans bestaat dat ik hier helemaal niet zou staan zonder extra statistische hulp. Regelmatig reisde ik vol vragen af naar Groningen waar Danny, Yang en Ritsert altijd weer helderheid wisten te verschaffen. In Wageningen waren Willem en Martin van cruciaal belang om de GWAS experimenten te kunnen analyseren.

Science is not bound to local borders and international collaboration is often the key to enhance research. I am looking back on great collaborations with Nick Provart and Thanh Nguyen from the University of Toronto who hybridized the 180 microarrays for our expression QTL study. Richard Pridmore and Tingting Wang from Nottingham University created valuable enhancements of the Germinator system, which are now close to implementation. Further I was inspired a lot by the many discussions I had during congresses and visits to scientists in Poland, France, Italy, Germany, United Kingdom, United States of America and Brazil. Naast de inspirerende congressen heb ik geweldige herinneringen aan vele discussies, lunches, kerstdiners en labuitjes met de collega's van de leerstoelgroep plantenfysiologie.

Tot slot wil ik graag nog een aantal mensen buiten het laboratorium noemen. Allereerst mijn moeder die dit moment helaas niet meer mocht meemaken. Je was er altijd, vol belangstelling en liefde. Mam, ik weet hoe trots je bent. Pap, jij was degene die me nieuwsgierig gemaakt heeft voor techniek. Je hebt me altijd aangemoedigd om toch vooral te blijven studeren. Je geduld is beloond; na 41 jaar sta ik dan eindelijk hier. Daarnaast een woord van dank aan iedereen die dicht bij me staat. Jullie hebben me altijd op de voet gevolgd en steeds geïnteresseerd geluisterd naar mijn eindeloze verhalen over de vorderingen. Rianne, de laatste zinnen heb ik voor jou bewaard. Jouw impulsiviteit maakt het leven spannend. Je hebt me ontzettend veel werk uit handen genomen, me altijd gesteund tijdens het schrijven en je creativiteit losgelaten op het ontwerp van de kaft. Ik kijk ontzettend uit naar wat de toekomst ons brengt. Je bent super!

Curriculum vitae

Ronny Joosen was born on the 10th of July 1972 in Haaksbergen, the Netherlands. In 1991 he obtained his secondary school diploma (HAVO, De Bouwmeester, Haaksbergen) and he finished a bachelor study plant biotechnology at the International Agricultural College Larenstein in Wageningen in 1996. After his graduation he started working as research assistant on a project related to monitoring water and metabolic activity in tomato seeds at Wageningen University (department Plant Physiology). After 3 years the project was finished and he moved to Plant Research International (Wageningen, business unit plant developmental systems). In this period he was involved in a range of projects related to embryogenesis, apomixis and androgenesis and had the opportunity to train his skills in many state of the art molecular techniques. In 2008 he was asked by Dr. Henk Hilhorst to join the project entitled 'Genes for seed quality' as a PhD student. From august 2012 he started working as a researcher quantitative genetics at the breeding company Rijk Zwaan.

Publication list

Related to this thesis

Peer reviewed

Joosen, R.V.L., Ligterink, W., Hilhorst, H.W., Keurentjes, J.J.B. (2009). Advances in genetical genomics of plants. Current Genomics *10*, 540-549.

Joosen, R.V.L., Kodde, J., Willems, L.A.J., Ligterink, W., van der Plas, L.H., Hilhorst, H.W.M. (2010). Germinator: a software package for high-throughput scoring and curve fitting of Arabidopsis seed germination. Plant Journal *62*, 148-159.

Joosen, R.V.L., Ligterink, W., Dekkers, B.J.W., Hilhorst, H.W.M. (2011). Visualization of molecular processes associated with seed dormancy and germination using MapMan. Seed Science Research *21*, 143-152.

Joosen, R.V.L*., Arends, D*., Willems, L.A.J., Ligterink, W., Jansen, R.C., and Hilhorst, H.W.M. (2012). Visualizing the genetic landscape of Arabidopsis seed performance. Plant Physiology *158*, 570-589.

Ligterink, W., **Joosen, R.V.L.**, and Hilhorst, H.W.M. (2012). Unravelling the complex trait of seed quality: using natural variation through a combination of physiology, genetics and -omics technologies. Seed Science Research *22*, 45-52.

Joosen, R.V.L., W. Ligterink, H.W.M. Hilhorst, J.J.B. Keurentjes (2013). Genetical genomics of plants; from Genotype to Phenotype. E-book: Advances in Genome Science (Vol. 1), in press.

Joosen R.V.L.*, Arends D*.,Li Y*., Willems L.A.J., Keurentjes J.J.B., Ligterink W., Jansen R.C., Hilhorst H.W.M. Identifying genotype-by-environment interactions in the metabolism of germinating seeds using generalized genetical genomics. Accepted for publication, Plant Physiology.

Joosen R.V.L, Willems L.A.J., Lajo Morgan G., Kruijer W., Keurentjes J.J.B., Ligterink W., Hilhorst H.W.M. Comparing Genome Wide Association and Linkage analysis for seed traits. Submitted to Plos One.

Non peer reviewed

Joosen R.V.L., Kodde J., Willems L.A.J.,Ligterink W., Hilhorst H.W.M. (2010). The Germinator automated germination scoring system. Seed Testing International, ISTA News Bulletin No. 140, 4-9.

*Equal contribution

Other publications

Spoelstra, P., **Joosen, R.V.L.**, Van der Plas, L.H.W., and Hilhorst, H.W.M. (2002). The distribution of ATP within tomato (Lycopersicon esculentum Mill.) embryos correlates with germination whereas total ATP concentration does not. Seed Science Research 12, 231-238.

Liu, C.M., McElver, J., Tzafrir, I., **Joosen, R.V.L.**, Wittich, P., Patton, D., Van Lammeren, A.A.M., and Meinke, D. (2002). Condensin and cohesin knockouts in Arabidopsis exhibit a titan seed phenotype. Plant Journal 29, 405-415.

Joosen, R.V.L., Lammers, M., Balk, P.A., Brønnum, P., Konings, M.C.J.M., Perks, M., Stattin, E., Van Wordragen, M.F., and Van Der Geest, A.H.M. (2006). Correlating gene expression to physiological parameters and environmental conditions during cold acclimation of Pinus sylvestris, identification of molecular markers using cDNA microarrays. Tree Physiology 26, 1297-1313.

Joosen, R.V.L., Cordewener, J., Supena, E.D.J., Vorst, O., Lammers, M., Maliepaard, C., Zeilmaker, T., Miki, B., America, T., Custers, J., and Boutilier, K. (2007). Combined transcriptome and proteome analysis identifies pathways and markers associated with the establishment of rapeseed microspore-derived embryo development. Plant Physiology 144, 155-172.

Srinivasan, C., Liu, Z., Heidmann, I., Supena, E.D.J., Fukuoka, H., **Joosen, R.V.L.**, Lambalk, J., Angenent, G., Scorza, R., Custers, J.B.M., and Boutilier, K. (2007). Heterologous expression of the BABY BOOM AP2/ERF transcription factor enhances the regeneration capacity of tobacco (Nicotiana tabacum L.). Planta 225, 341-351.

Passarinho, P., Ketelaar, T., Xing, M., van Arkel, J., Maliepaard, C., Hendriks, M.W., **Joosen, R.V.L.**, Lammers, M., Herdies, L., den Boer, B., van der Geest, L., and Boutilier, K. (2008). BABY BOOM target genes provide diverse entry points into cell proliferation and cell growth pathways. Plant Molecular Biology 68(3), 225-237.

Stasolla, C., Belmonte, M.F., Tahir, M., Elhiti, M., Khamiss, K., **Joosen, R.V.L.**, Maliepaard, C., Sharpe, A., Gjetvaj, B., and Boutilier, K. (2008). Buthionine sulfoximine (BSO)-mediated improvement in cultured embryo quality in vitro entails changes in ascorbate metabolism, meristem development and embryo maturation. Planta 228, 255-272.

Education Statement of the Graduate School

Experimental Plant Sciences

Issued to: Ronny Joosen
Date: 24 May 2013
Group: Plant Physiology, Wageningen University

1) Start-up phase	*date*
► **First presentation of your project**	
Genes for Seed Quality; Physiological Genetical Genomics	Jun 24, 2008
► **Writing or rewriting a project proposal**	
Arabidopsis Seed Quality Genes	Jun 20, 2008
► **Writing a review or book chapter**	
Advances in Genetical Genomics of Plants, Current Genomics	Jun 2009
Unravelling the Complex Trait of Seed Quality, Seed Science	Jan 2012
► **MSc courses**	
► **Laboratory use of isotopes**	
Subtotal Start-up Phase	*13,5 credits**

2) Scientific Exposure	*date*
► **EPS PhD student days**	
EPS PhD stundent day (Naturalis), Leiden Univeristy	Feb 26,2009
EPS PhD student day, Wageningen University	May 20, 2011
► **EPS theme symposia**	
EPS theme 1 symposium 'Developmental Biology of Plants'	Jan 22, 2011
EPS Theme 3 symposium 'Metabolism and Adaptation'	Mar 22, 2013
► **NWO Lunteren days and other National Platforms**	
NWO-ALW meeting 'Experimental Plant Sciences', Lunteren	Apr 06-07, 2009
NWO-ALW meeting 'Experimental Plant Sciences', Lunteren	Apr 19-20, 2010
NWO-ALW meeting 'Experimental Plant Sciences', Lunteren	Apr 04-05, 2011
NWO-ALW meeting 'Experimental Plant Sciences', Lunteren	Apr 02-03, 2012
► **Seminars (series), workshops and symposia**	
Seminars (Nicolas Provart, Jian-Kang Zhu)	Jun-Nov 2008
Seminar Pamela J. Hines (Science from an editors view)	Nov 06, 2008
Seminars (Wallace A. Cowling, M. Vuylsteke, H. Nonogaki a.o)	Jun-Oct 2009
Symposium Series NIOO-KNAW Revolution in Evolution?	Sep 18, 2009
EPS Symposium 'Ecology and Experimental Plant Sciences 2'	Sep 22, 2009
Seminars (Justin Borevitz, Glenda Willems, Regina Delourme)	Jan-Oct 2010
Keys Seminars (Keygene) Vincet Colot & Marc Block	Oct 2010
Mini symposium 'How to write a world class paper'	Apr 19, 2011
Seminars (L. Summer, C. Wagstaff, J. Jimenez Gomez a.o)	Aug-Dec 2011
ServiceXS Seminar, Wageningen	Sep 2011
Seminars (Luc Janss, Ian Henderson)	Jan-Feb 2012
Cost meeting, modeling and databases for terpenes	Feb 16, 2012

► **Seminar plus**	
► **International symposia and congresses**	
ISSS Congres, Olsztyn, Poland	Jul 06-11, 2008
QTL mas, Wageningen, The Netherlands	Apr 20-21, 2009
ICAR, Edinburgh, Scotland	Jun 30-Jul 04, 2009
ISSS congress York, UK	Jul 18-22, 2010
ISSS congress Bahia, Brasil	Apr 10-15, 2011
ASPB , Minneapolis, USA	Aug 06-09, 2011
► **Presentations**	
PhD course Natural Variation (Oral)	Aug 26, 2008
INRA Anger, France (Oral)	Sep 09, 2008
INRA Versailles, France (Oral)	Sep 26, 2008
User commitee STW (Oral)	Dec 11, 2008
Breedwise course (Oral)	Apr 01, 2009
NWO-ALW meeting 'Experimental Plant Sciences', Lunteren	Apr 07, 2009
User commitee STW (Oral)	Apr 09, 2009
QTL mas, Wageningen, The Netherlands (Oral)	Apr 20, 2009
ICAR meeting (Poster)	Jun 29, 2009
User committee STW (Oral)	Oct 15, 2009
Masterclass Seed Science (Oral)	Oct 29, 2009
NWO-ALW meeting 'Experimental Plant Sciences', Lunteren	Apr 19, 2010
PRI Bioscience (Oral)	May 11, 2010
Masterclass Seed Science (Oral)	Jun 03, 2010
Breedwise course (Oral)	Sep 29, 2010
EPS theme 1 symposium 'Developmental Biology of Plants'	Jan 22, 2011
NWO-ALW meeting 'Experimental Plant Sciences', Lunteren	Apr 04-05, 2011
Seed Science Masterclass (Oral)	May, 2011
Warwick University, Coventry, UK (Oral)	Nov 15, 2011
IPK Gatersleben, Germany (Oral)	Dec 14, 2011
NWO-ALW meeting 'Experimental Plant Sciences', Lunteren	Apr 02-03, 2012
► **IAB interview**	Feb 18, 2011
► **Excursions**	
KeyGene (organised by EPS PhD student council)	Jan 26, 2012
Subtotal Scientific Exposure	*36,3 credits**

3) In-Depth Studies	*date*
► **EPS courses or other PhD courses**	
Master Class 'Seed technology' (Hilhorst & Groot)	Jun 09-12, 2008
PhD course 'Natural Variation' (Koorneef & Aarts)	Aug 26-29, 2008
Plant Systems Biology Summer School (Warwick University)	Sep 12-16, 2011
Association mapping for Learning from Nature data	Feb 2012
► **Journal club**	
Participation in literature discussion group at WU-PPH	2008-2012
► **Individual research training**	
Subtotal In-Depth Studies	*7,2 credits**

4) Personal development	*date*
► Skill training courses	
Course R-statistics	Oct 23-24, 2008
Biometris QTL mapping	Jun 2011
Expectations day EPS	Nov 18, 2011
Presentation skills, Wageningen Language Services	Oct 14-28, 2011
► Organisation of PhD students day, course or conference	
► Membership of Board, Committee or PhD council	
Subtotal Personal Development	*3,6 credits**

TOTAL NUMBER OF CREDIT POINTS*	60.6

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

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

Printed by: Proefschriftmaken.nl || Uitgeverij BOXPress