

Propositions

1. A systems genetics study is only possible if knowledge-driven and data-driven approaches are combined. (this thesis)
2. Proper experimental design and phenotyping methods are the most crucial steps in genetic studies. (this thesis)
3. Without phenotypic data, next generation sequencing results do not increase insight into biological processes, but merely use hard-disk memory.
4. The shift from public research funding to public-private partnership funding will be at the expense of fundamental research.
5. Political stability in the current developing world is a pre-requisite for food security.
6. Unlike Dutch communication, Dutch language is very indirect and confusing.

Propositions belonging to the thesis, entitled

“A systems genetics study of seed quality and seedling vigour in *Brassica rapa*”

Ram Kumar Basnet

Wageningen, 24 August 2015

**A systems genetics study of seed quality and
seedling vigour in *Brassica rapa***

Ram Kumar Basnet

Thesis committee

Promotor

Prof. Dr R. G. F. Visser
Professor of Plant Breeding
Wageningen University

Co-promotors

Dr A. B. Bonnema
Associate professor, Laboratory of Plant Breeding
Wageningen University

Dr C. A. Maliepaard
Assistant professor, Laboratory of Plant Breeding
Wageningen University

Other members

Prof. Dr M. Koornneef, Wageningen University / Max Planck Institute for plant breeding research, Köln,
Germany

Prof. Dr C. Jung, University of Kiel, Germany

Dr L. Bentsink, Wageningen University

Dr M. Malosetti, Wageningen University

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

A systems genetics study of seed quality and seedling vigour in *Brassica rapa*

Ram Kumar Basnet

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr A. P. J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Monday 24 August 2015
at 11 a.m. in the Aula.

Ram Kumar Basnet

A systems genetics study of seed quality and seedling vigour in *Brassica rapa*

177 pages

PhD thesis, Wageningen University, Wageningen, NL (2015)

With references, with summary in English

ISBN 978-94-6257-425-0

Dedicated to my beloved parents

Contents

Chapter 1	9
General introduction	
Chapter 2	25
Comparative methods for association studies: a case study on metabolite variation in <i>Brassica rapa</i> core collection	
Chapter 3	43
Genome-wide analysis of coordinated transcript abundance during seed development in different <i>Brassica rapa</i> morphotypes	
Chapter 4	71
Quantitative trait locus analysis of seed germination and seedling vigour under non-stress and salt stress conditions in <i>Brassica rapa</i> : a possible role of <i>BrFLC2</i> and <i>BrFAD2</i>	
Chapter 5	105
A systems genetics approach identifies gene regulatory networks associated with fatty acid composition in <i>Brassica rapa</i> seed	
Chapter 6	135
General discussion	
References	147
Summary	165
Acknowledgements	169
About author	173
Publications	174
Education certificate	175

Chapter 1

General introduction

Brassicas and their economic importance

The mustard family (*Brassicaceae*) is a large angiosperm plant family with over 300 genera and over 3500 species, which are distributed worldwide (Al-Shehbaz et al., 2006). It includes several species with economic value, and accounts for approximately 10% of the world's vegetable crop production and 12% of the edible oil supplies, after sunflower and soybean (Economic Research Service, 2008). Apart from the commercial use of many *Brassica* species, their high genetic resemblance to the model species *Arabidopsis thaliana* has made them attractive model systems to study plant evolution and plant development.

The *Brassica* genus consists of six economically important species: the diploid species *B. rapa* ($2n = 20$, AA), *B. nigra* ($2n = 16$, BB) and *B. oleracea* ($2n = 18$, CC), and their natural inter-specific hybrids *B. juncea* ($2n = 36$, AABB), *B. napus* ($2n = 38$, AACC) and *B. carinata* ($2n = 34$, BBCC) (Lukens et al., 2004). *B. rapa* and *B. oleracea* are mainly grown as vegetable crops, *B. napus* and *B. juncea* as sources of vegetable oil and *B. nigra* for condiment (mustard) (Mun et al., 2011). The cytogenetic relationships between the genomes of *Brassica* crop species are referred to as U's triangle (U, 1935) (Figure 1).

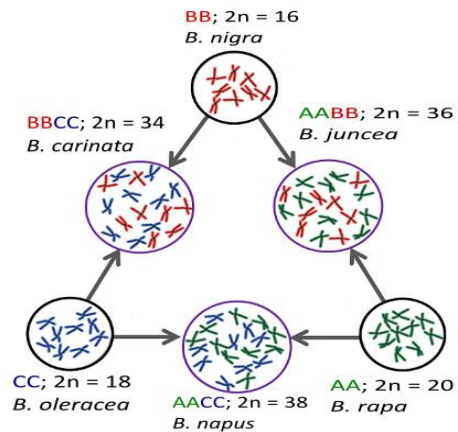
Brassicas are consumed for a wide range of plant parts. For example, *B. rapa* ssp. *pekinensis* (Chinese cabbage) and ssp. *chinensis* (pak choi) are consumed as leafy vegetables in Asia, while *B. rapa* ssp. *trilocularis* (Yellow sarson) and *B. rapa* ssp. *dichotoma* (brown sarson or toria) have been cultivated historically for seed oil in the Indian subcontinent, and the tubers of turnips (*B. rapa* ssp. *rapa*) are used as fodder and vegetables. From *B. oleracea*, the inflorescences are consumed in cauliflower and broccoli, leaves are consumed from non-heading kales and the heading cabbages, and the stem tuber of kohlrabi is consumed as a vegetable. Canola (*B. napus*), one of the descendants of *B. rapa*, is one of the most important sources of vegetable seed oil in the world.

The origin and diversity of *Brassica rapa*

B. rapa was the first domesticated *Brassica* species and has been cultivated for over 4,000 years from the highlands near the Mediterranean region to Scandinavia, Germany and Central Europe, and eventually to Central Asia (Bonnema et al., 2011). The Swedish botanist Linnaeus first described *B. rapa* as a turnip forming species and *B. campestris* as a wild weedy species, but in 1833, Metzger concluded that they represent the same species, which was then called *B. rapa* (CFIA, 2014). *B. rapa* has a long cultivation history and breeding for different consumable plant parts resulted in the selection of different morphotypes (Zhao et al., 2007). Based on earlier studies of morphological characteristics, geographic distribution and molecular variation, two independent centres of origin have been hypothesized (Denford and Vaughan, 1977; Song et al., 1988; Gómez-Campo, 1999). Europe has been considered as one of the centers of diversity for turnip rape, broccoletto and turnip types and it is generally assumed that from there the species spread to Russia, Central Asia and the Near East (Bonnema et al., 2011). Eastern Asia has been considered as the other centre of diversity and gave rise to Asian leafy vegetables. Turnip is believed to have originated

directly from the wild progenitor that came from the Iranian region to Europe (Reiner et al., 1995). De Candolle (1959) reported that turnip has been cultivated since 2500-2000 BC and its cultivation spread to Asia after 1000 BC. In Asia or India the oil types *B. rapa* ssp. *trilocularis*, known as “yellow sarson” and ssp. *dichotoma* known as “toria” and “brown sarson” diversified (Prakash and Hinata, 1980). In molecular analyses, these oil types form a separate subgroup suggesting the Indian subcontinent as a third centre of origin (Zhao et al., 2005; Warwick et al., 2008).

Figure 1: The “Triangle of U”, showing the genetic relationship between six *Brassica* species (U, 1935). Chromosomes from each of the genomes A, B and C are represented by different colours. Three diploid species (*B. rapa*, *B. nigra* and *B. oleracea*) are marked by black-circles and three tetraploid species (*B. carinata*, *B. juncea* and *B. napus*) are with purple circles. 2n indicate diploid chromosome number (U, 1935).



Polyploidization, genome duplication and comparative mapping

Brassica species and *A. thaliana* diverged from a common ancestor with karyotype ($n=8$) approximately 14.5-20.4 million years ago (Bowers et al., 2003; Yang et al., 2006; Hong et al., 2008; Cheng et al., 2013a). All *Brassica* species have complex genomes with whole genome duplications with intra- and inter- genomic conservation of chromosomal blocks (Prakash and Hinata, 1980; O'Neill and Bancroft, 2000; Town et al., 2006; Yang et al., 2006). Schranz et al. (2006) defined 24 conserved chromosomal blocks (coded as A-X) with naming, order and orientation between *A. thaliana* and *B. rapa* species. The ancestral genomic blocks of *A. thaliana* are generally replicated three times on different chromosomes of *B. rapa* (Schranz et al., 2006; Wang et al., 2011a). As the rate of gene loss in these three ancestral blocks differ, the *Brassica* genomic blocks are classified according to their rate of gene loss into the Least Fractionated (LF), medium fractionated (MF1) and Most Fractionated genomes (MF2) (Cheng et al., 2012; Liu et al., 2014; Parkin et al., 2014). Co-linearity of chromosomal segments between the two species allows for the possibility of comparative mapping, inferring knowledge from the model species *A. thaliana* and functional characterization of genes of interest based on annotation of genes in *A. thaliana* (Schmidt et al., 2001).

Seed development in *B. rapa*

Brassica seeds are non-endospermic, which means that the endosperm is not retained in the mature seeds and the embryo is enclosed by the seed coat. Seed development consists of embryogenesis followed by seed desiccation (Yu et al., 2010; Li et al., 2012a). In embryogenesis two overlapping phases are distinguished: morphogenesis and seed filling (Sabelli, 2012). Embryogenesis starts after the double fertilization process of fusion of two sperm nuclei with the egg cell and the central cell nuclei, respectively; the zygote then goes

through a series of cell divisions and differentiation events during morphogenesis; from a pre-globular and globular embryo stage, a heart stage, a torpedo stage, a bent-cotyledon stage to the mature embryo, with the basic body plan that perpetuates throughout the life of the plant (Angelovici et al., 2009; Le et al., 2010; Li et al., 2012a). During seed filling the seed accumulates storage compounds in the embryo. In this stage, the endospermic starch is consumed by the embryo and converted into proteins and oils (Basnet et al., 2013; Borisjuk et al., 2013a). Once the seed filling ends, the embryo stops growing and becomes metabolically quiescent. Seed desiccation, the final stage of seed development, is the bridge between maturation and germination (Angelovici et al., 2009). During this phase, the seed moisture content declines, this makes the seed able to withstand desiccation and to enter into a state of developmental and metabolic quiescence (Sabelli, 2012). The seed reserves are an essential source of energy during seed germination and early seedling growth. Therefore, unravelling the genetics of transcriptional regulation of seed metabolism and of metabolic switches in the seed is of great interest, not only for breeding oil content and improving seed quality but also for early seedling vigour.

Seed quality and seedling vigour traits

Seed is the basic and most critical input for most of the agricultural crops and seed quality determines plant establishment, growth and development in natural or agricultural ecosystems. In practice, definitions of seed quality vary according to the end users (Joosen, 2013) and there are different approaches to measure it. Seed quality attributes include seed germination, dormancy, seed viability, physical and genetic purity, seed size, seed weight, storability, being free of pathogens, normal embryo and seedling morphology and ability to develop into a normal plant under optimal and sub-optimal conditions (Ellis, 1992; El-Kassaby et al., 2008; Angelovici et al., 2009; Finch-Savage et al., 2010). The protrusion of the radicle from the seed is termed seed germination. Seedling vigour refers to the ability of a seed or seed lot to establish seedlings under a wide range of growing conditions (Foolad et al., 2007; Finch-Savage et al., 2010). In general, the higher the seed quality is, the higher the seedling vigour.

In the last few decades, the seed industry has played a vital role in agri-business and sustainable food production by ensuring efficient seed trade with continuous supply of high-yielding and disease resistant varieties suitable for different environmental conditions and consumers' demands. However, most of the seed companies and breeding centres hardly prioritize their breeding programs for high seed quality and seedling vigour related traits. Instead, seed size selection, disinfection, priming and coating have become the most common practices to ensure high and uniform seed germination and vigorous seedling growth of marketable seeds. The main bottleneck for studying the genetics of traits related to seed quality and seedling vigour is their complex genetic architecture and the large influence of the environment. Due to technological advances in high-throughput phenotyping and genomics, it is now better feasible to study the molecular aspects of seed

quality and seedling growth and it has become possible to incorporate tools like marker-assisted breeding in breeding programs. Many genetic studies of seed quality and seedling vigour have been conducted in *A. thaliana* (Koornneef et al., 2002; Fait et al., 2006; Joosen, 2013). Many transcription factors as well as structural genes were reported for their role in seed maturation and dormancy during seed development (reviewed by Koornneef et al., 2002; Bentsink and Koornneef, 2008).

In this thesis, five seed germination parameters were used to quantify different aspects of seed quality while root and shoot length and seedling biomass at multiple time points in different environments were used to assess seedling vigour. The seed germination parameters were T10 (time to reach 10% germination, onset of germination), T50 (time to reach 50% germination), U7525 (time interval between 25% and 75% germination, uniformity), Gmax (maximum germination percentage) and AUC (Area Under the germination Curve) (Joosen et al., 2010). Generally, large seed size and high seed weight supports seedling growth, especially during the heterotrophic stage (before photosynthesis starts) (Cheng et al., 2013b). High and uniform seed germination and good seedling vigour have great impact on crop establishment and therefore contribute directly to the economic success of commercial crops (Finch-Savage et al., 2010).

Seed germination and seedling vigour are governed by complex genetic architecture (Bettey et al., 2000; Koornneef et al., 2002; Finch-Savage et al., 2010; Kazmi et al., 2012; Joosen, 2013) and influenced by many non-genetic factors, such as the environmental conditions during seed production, the physiological stages of the seed at harvesting, processing, storage, germination and during early growth. Many efforts have been made to improve seed germination and seedling vigour by optimizing the non-genetic factors; however, the genetic factors to improve seed germination and seedling growth deserve attention as well. Genetic factors such as mutation, recombination and natural selection make plants adaptive to local conditions to continue their development and life cycle (Anderson et al., 2011). The ability of a plant (or any organism) to adjust its physiology and morphology in response to the biotic and abiotic environment is named phenotypic plasticity (Schlichting, 1986). Genotypes respond to environments differently; therefore, phenotypic plasticity has a genetic basis (Kooke, 2014). Salinity is a major abiotic stress that delays the onset of germination and decreases the rate of both seed germination, and seedling growth and establishment. Genetic aspects of salinity tolerance have been reported in different crops (Csanádi et al., 2001; Foolad et al., 2007; Kazmi et al., 2012; Wang et al., 2012; Joosen, 2013). One of the aims of this thesis was to identify genome regions and possibly causal genes involved in seed germination and seedling vigour traits under non-stress and salt-stress conditions.

Seed content determines seed quality and seedling vigour

The seed reserve consists of seed storage proteins (SSPs), carbohydrates (mostly starch) and storage lipids (mainly triacylglycerols (TAGs)) (Baud and Lepiniec, 2010). About 35-40% of

the total seed weight in *A. thaliana* and 35 to 52% in *B. napus* is composed of TAGs (Chia et al., 2005; Graham, 2008; Rahman et al., 2013). In germinating seeds, TAG lipases oxidize TAGs into free fatty acids (FAs) and glycerol, and, through a series of enzymatic reactions, free fatty acids are later converted into sugars. This breakdown process of TAGs results in energy and provides the carbon skeletons to support seed germination and seedling growth until photosynthesis becomes efficient (Miquel and Browse, 1995; Quettier and Eastmond, 2009; Baud and Lepiniec, 2010). Several studies in *A. thaliana* describe mutants with altered ability to degrade TAGs, and phenotypes range from increased dormancy, decreased germination and early seedling growth, up to effects on the later seedling stages (Hayashi et al., 2001; Baker et al., 2006; Fait et al., 2006; Quettier and Eastmond, 2009; Pracharoenwattana et al., 2010; Kelly et al., 2011). Elliott et al. (2007) observed large cotyledons, higher shoot dry weight, better seedling establishment and higher seedling vigour being associated with larger seed size and weight in summer turnip rape (*B. rapa*). Seed weight is highly correlated with seed size (Bagheri et al., 2013). The quantity and composition of seed reserves is regulated by biosynthetic processes during seed development (Baud et al., 2008), and subsequent mobilization of seed reserves during imbibition determines seed germination and the potential of seedling vigour (Fait et al., 2006; Cheng et al., 2013b). While many studies focus on seed development and germination, only few try to link metabolite variation in the mature seed with transcriptional regulation during the seed filling process and germination and seedling vigour. In this thesis, a systems genetics approach was used to integrate different ~omics datasets and to construct a gene regulatory network for fatty acids, that can be associated with genetic variation of seed germination and seedling vigour.

Natural variation in *B. rapa* and genetic studies of seed quality traits

Genetic variation is the basis for breeding improved cultivars for agricultural production. Major breakthroughs in molecular marker technologies and development of statistical methods enabled the construction of linkage maps and analysis of quantitative trait loci (QTL). A linkage map is a representation of the chromosomes with ordered genetic markers, based on recombination frequencies between pairs of markers. QTLs are the genomic regions that account for natural genetic variation of quantitative traits of interest. If QTLs for seed quality traits are identified, markers closely linked to such QTLs could potentially be used in marker-assisted selection (MAS) for seed quality traits in breeding programs. The advantage of MAS (relative to selection on phenotype) is that these markers themselves are not influenced by the environment, so that they can be more effective targets of selection. In addition, QTL regions can be the starting points for fine mapping and the identification of candidate genes and possibly for cloning of the causal genes of a trait. Basically, there are two approaches to identify and localize QTLs that might contain genes causing genetic variation of the traits: linkage mapping and association mapping.

Linkage mapping

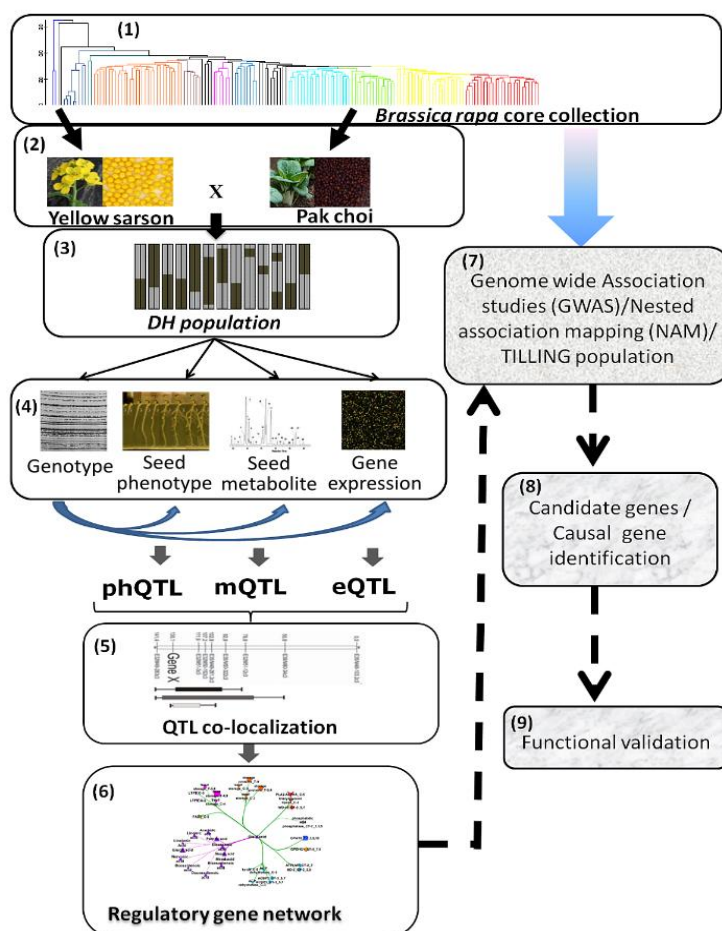
Linkage mapping is an approach for mapping QTLs in a bi-parental population, developed from a cross of two genotypes with contrasting phenotypes. In plants that allow inbreeding, the most commonly used bi-parental populations for QTL mapping are F_2 , doubled haploid (DH) and recombinant inbred line (RIL) populations. F_2 and DH populations have only one round of recombination (meiosis of the F_1), but take less time to create than a RIL population. A RIL population, on the other hand, has had several rounds of recombination, and therefore the mapping resolution is improved. In an F_2 population, both additive and dominance effects can be estimated while only additive effects can be estimated in DH and RIL populations with (almost) only homozygotes. DH and RIL populations are considered immortal populations, since they can be reproduced identically; this enables studying genotype-by-environment (GxE) and QTL-by-environment (QTLxE) interaction, which is not possible in an F_2 population, where the genotypes are mortal (unless F_2 plants are vegetatively propagated). In this thesis, a population of 175 DH lines was used for studying genetics of seed quality and seedling vigour under non-stress and salinity conditions. This population was developed from a cross of two contrasting *B. rapa* parents, an oil-type yellow sarson (YS143; accession number FIL500) and a vegetable-type pak choi (PC175; Accession number VO2B0226) and seeds harvested in two years from the DH population (Figure 2).

Association mapping (AM)

Association mapping (AM), also called linkage disequilibrium (LD) mapping, is another approach to find QTLs for traits. This method seeks to identify the genetic variants (i.e. loci and alleles at these loci) linked to phenotypic variation in a natural population or core collection of germplasm. Unlike bi-parental populations, a core collection can be used that represents the historic genetic variation and recombination events present in the germplasm collection. Therefore, it is not necessary to make a cross, create large segregating populations and produce a large number of progeny. In linkage mapping QTLs are mapped with low resolution, so that approximate confidence intervals are generally very large: sometimes even a whole linkage group. Markers that are far from QTLs are still strongly associated, especially when only one or just a few recombination rounds have taken place. Contrastingly, in association mapping, resolution of identified QTLs is much higher and, consequently, a much narrower window size of the genome will show the association with the phenotype. Ideally, association panels are composed of unrelated individuals that capture a wide genetic diversity, genotyped with a large number of markers; however, in such cases, the power to detect QTL(s) will be reduced if marker alleles are rare. One major drawback of an association panel is generally the presence of unequal relatedness (also called kinship) and population structure between individuals, which may lead to false positive marker-trait associations (Pino Del Carpio et al., 2011a). Statistical methods have been developed that take population structure (Q-matrix) and/or kinship (K-matrix) into

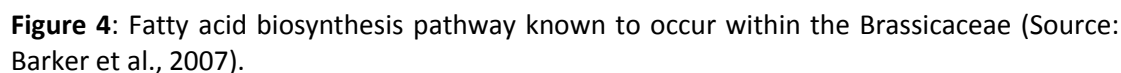
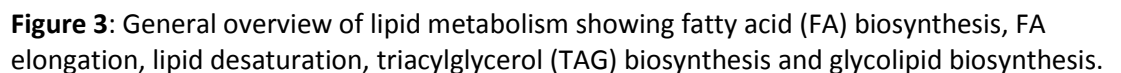
account that aim to optimize the balance between false-positives and false-negatives of marker-trait associations. In this study, an AM approach was followed for detection of QTLs for carotenoids and tocopherols and the power of QTL detection using this approach was assessed (Figure 2).

Figure 2: Strategy for systems genetics study of seed quality and seedling vigour related traits in *B. rapa*. Solid arrows indicate the methods implemented in this study and dotted arrows for future possibilities. Numbers on the left side of each box indicates different steps involved in the study.



The lipid metabolism and fatty acids biosynthesis pathway in *Brassica*

Brassica species are widely cultivated as oil crops and seed oil quality depends on the relative proportions of FAs, mainly erucic (C22:1), oleic (C18:1), linoleic (C18:2), linolenic (C18:3) acids (Jagannath et al., 2011). In oilseed breeding for human consumption, one of the important objectives is to increase the levels of C18:1 and C18:2 and decrease C22:1, while high C18:3 reduces the storability and causes rancidity upon deep frying (Töpfer et al., 1995). *Brassica* seed oil content and FA composition of seeds are quantitative in nature with a complex genetic basis (Barker et al., 2007). A general overview of lipid metabolism is depicted in Figure 3; it includes metabolic processes such as FA biosynthesis, FA elongation, FA desaturation, lipid degradation, triacylglycerol (TAG) synthesis, glycolipid biosynthesis and biosynthesis of storage proteins. In order to unravel the genetics of seed oil formation, a comprehensive study of lipid metabolism at both the genetic and biochemical levels is needed. In this thesis, the integration of information at the molecular, biochemical and transcriptomic levels is used as an approach to bridge the gap between gene function and biochemical processes.



In plants, the FA biosynthesis starts in the plastids, followed by FA elongation and TAG assembly in the endoplasmic reticulum (ER) (Figure 4; Barker et al., 2007). FAs are a group of compounds that consist of a long hydrocarbon chain and a terminal carboxylate group. They are symbolized by the number of carbons and double bonds present in the carbon chain: for

example, oleic acid is denoted as C18:1 indicating 18 carbons and one double bond in its carbon chain. The chain length and degree of saturation determine the properties. FAs without double bonds are called saturated FAs (SFAs), those with only one double bond are mono-unsaturated FAs (MUFAs) and those with more than two double bonds are poly-unsaturated FAs (PUFAs). MUFAs and PUFAs have a lower melting point than SFAs even when they have the same carbon chain length.

During seed development, sugar and other carbon sources obtained from photosynthesis and other processes are converted into acetyl-CoA in the plastids. Acetyl-CoA carboxylase (ACCase) catalyzes acetyl-CoA into malonyl CoA and the latter donates two carbons to growing FA chains, which are covalently attached to the acyl-carrier protein (ACP) of the FA synthase complex (Barker et al., 2007; Weselake et al., 2009). Acyl-ACP desaturase catalyzes the formation of C16:0 to C18:0 acyl-ACP pools by a series of keto-acyl ACP synthase (KAS) enzymes. Thioesterase (acyl-ACP hydrolase) hydrolyzes those ACP pools to produce FAs, SFAs palmitate (C16:0) and MUFA oleate (C18:1). The FAs move out from the plastidial envelope to the cytoplasm by two classes of fatty acyl-ACP thioesterases (FAT-A and FAT-B) and those FAs are reactivated as acyl-CoAs. FAT-A preferentially hydrolyses oleoyl-ACP whereas FAT-B hydrolyses saturated acyl-ACPs (Harwood, 2005). In the ER, elongation of FAs occurs using acyl-CoA substrates by a series of FA dehydrogenase and FA elongase (FAE) enzymes resulting in very long chain FAs (VLCFAs) and PUFAs. The pool of acyl-CoA FAs is then converted into diacylglycerols (DAGs) in the ER and, finally, DAGs and acyl-CoA FAs are condensed to triacylglycerol (TAG) (Barker et al., 2007; Tan et al., 2011). The glycerol backbone for TAG assembly is derived from *sn*-glycerol-3-phosphate (G3P). Schwender et al., (2004) reported that over 50% of the carbon flux in developing *B. napus* seed is consumed by TAG synthesis. The rate of FA synthesis depends on the carbon flux; therefore, biosynthesis of FAs and TAG is linked to the photosynthesis capacity and carbohydrate metabolism during the seed filling and maturation processes (Tan et al., 2011). TAGs are accumulated in the ER and stored in oil bodies in seed embryo cotyledons. Seed oil accumulates mainly in the form of TAG, which primarily serves as an energy reserve for germinating seeds (Sharma et al., 2008).

Linking a phenotypic trait to an ~omics data set

A systems genetics approach

The information obtained from QTL analysis or association mapping enables the detection of genomic regions and variants associated with traits. However, these provide little insight into the genes underlying this variation, nor in their regulatory mechanisms. To unravel the molecular basis of the genetic architecture of complex traits, systems genetic approaches are used, since they allow the integration of intermediate phenotypes at the metabolite, transcript and/or protein levels with the phenotypic trait(s) of interest. Molecular interactions of intermediate phenotypes can be studied with multiple genetic perturbations

already present in natural populations, rather than just an individual genetic perturbation as present in transgenic individuals (Civelek and Lusi, 2014).

Systems genetics analyses use different ways to integrate these different layers of information, such as simple correlation analyses, genetic mapping and network approaches (Civelek and Lusi, 2014). In this thesis, we applied all these different approaches to integrate genetic information about transcripts and fatty acids (Figure 2). Two traits can be correlated if they influence each other or if they are affected by a common factor or just by random association. Similarly, traits can be mapped to the same genomic region if one trait is the cause of the other, but also due to the influence of a common genetic factor or just because of linkage without any causal relationship. Network analysis can be used to construct regulatory gene networks, with genes as nodes and their expression networks, which can be inferred from high-throughput experimental data from a natural population. This approach allows the discovery of novel genes (genes not reported before) or the prediction of gene functions and gene interactions. Weighted gene co-expression network analysis (WGCNA) utilizes the Pearson correlation coefficients between transcript abundance of genes to identify gene modules (groups of co-expressed genes) that have similar expression patterns and possibly share common *cis*-regulatory elements. A *cis*-regulatory element is a short sequence of DNA, where a transcription factor binds the gene's regulatory elements to regulate its transcription. In addition, a network characteristic called "degree of connection" can be calculated for each gene, which refers to the possible interactions of genes in the network underlying a metabolic pathway or a biological system. Genes with a high degree of connection indicates essential genes with evolutionary conserved functions in the pathway, whereas genes with a low degree of connection are likely to be associated genes that can contribute to the expression of other genes and also to the regulation of phenotypic trait values (Khurana et al., 2013). Those associated genes could be interesting from a breeding perspective to improve a trait while essential genes could have a role to modify complete pathways.

Transcriptomics

The term transcriptome refers to the complete set of messenger RNA (mRNA) transcripts that are produced from genes of a genome in specific biological and/or environmental conditions (Feltus, 2014). The transcriptome is very dynamic, depending on specific circumstances: specific to cell, tissues, developmental stages or environmental conditions. Therefore, transcriptomics has been very popular in the last few decades to study the global expression of genes involved in biological processes and their functional annotation (Feltus, 2014). The mRNA transcripts can be quantified using high-throughput methods, such as microarrays and RNA-seq.

Microarray

Microarray analysis is a high-throughput technology to quantify the transcript abundance of genes as an indication of their relative expression. The probes on a microarray are designed from the coding sequence (CDS) of genes. There are broadly two types of microarrays – two-colour microarrays and one-colour microarrays. In two-colour microarrays, two samples are hybridized to one array using two dyes (Cy5 and Cy3), whereas only one sample is hybridized in a one-colour array. Microarrays are used to compare global gene expression levels across developmental stages, or between different genotypes or in different conditions such as different biotic or abiotic environments.

For microarrays probes need to be designed specifically for the target species of interest. When we started this study, no microarray was available for *B. rapa*. Due to the availability of the whole genome sequence of *B. rapa* Chinese cabbage var. Chiifu (Wang et al., 2011a), we could design a custom microarray (8 x 60K) containing 60,000 features in a two colour Agilent platform using 60-mer probes designed based on predicted gene models for *B. rapa*.

Experimental design for microarray gene expression

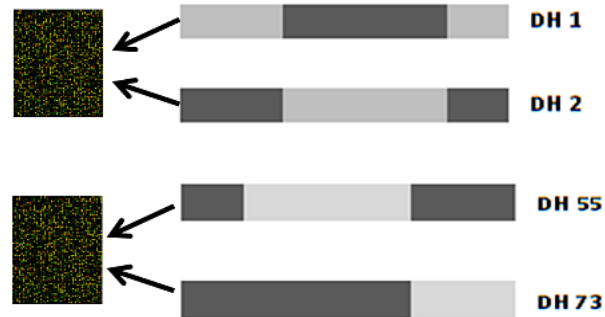
Microarray experiments are still costly and time consuming, which may limit sample sizes and as a result statistical power. Because of its high-throughput nature and high sensitivity (in the laboratory), it is crucial to optimize experimental design by carefully choosing the subjects to be assayed and cautiously controlling systematic factors affecting the experiment and the system. An optimal experimental design should allocate mRNA samples across the assay batches, slides and dye labelling so that effects of interest are not confounded with nuisance factors. Also, a good experimental design should be cost efficient while including biological replicates, technical replicates and spike-in controls. In this thesis, we used two types of experimental designs. Firstly, a double loop design was developed to study the global gene expression patterns across seed developmental stages in different genotypes: inner and outer loops were used for genotypes and developmental stages. The other design, used for an eQTL mapping study with a genetical genomics approach, was a distant-pair design (Fu and Jansen, 2006) with selected pairs of individuals hybridized in each array; in this design the pairs were chosen to maximize genetic dissimilarity and a higher number of recombination events (Figure 5).

Genetic architecture of gene expression variation: genetical genomics Cis- / trans-regulation

In genetical genomics studies, eQTLs can be categorized as local eQTLs (*cis*-acting) and distant eQTLs (*trans*-acting) depending on the distance between the physical position of a gene and its eQTL location. An eQTL mapped near the location or just upstream of a gene itself is called a *cis*-eQTL (Figure 6). In this case, sequence variation in the promoter of the gene may cause heritable variation in its expression. Alternatively, an eQTL might be mapped distantly from the gene, either in the same chromosome or in a different

chromosome; this is called a *trans*-eQTL (Figure 6). In case of a *trans*-eQTL, a polymorphism in a gene encoding a transcription factor may affect the expression of the target gene. In eQTL studies, the terms *cis*- and *trans*-regulation of eQTLs do not refer to the mechanism of gene action, but to the genomic position of the eQTL relative to the physical position of the gene of which the transcript is studied.

Figure 5: Distant-pair design applied for microarray hybridization for eQTL mapping. Graphical genotype of DH lines used as pair of samples in microarray hybridization.



In eQTL analysis, *cis*-eQTLs generally explain more variation than *trans*-eQTL, suggesting a stronger regulatory ability than *trans*-eQTL (Xiao et al., 2013). Mapping and positional cloning of a *cis*-eQTL may lead to the identification of structural genes, whereas *trans*-eQTLs may identify master regulators, such as transcription factors or small regulatory RNA (Holloway et al., 2011) that control the regulation of multiple genes in a pathway. Systems genetics or eQTL mapping studies allow the identification of novel functional pathways, genetic variants or gene-gene regulatory networks to better understand the underlying genetic regulation of the traits (Calabrese et al., 2012; Mäkinen et al., 2014).

Sequence variation in a master regulator, such as a transcription factor, usually regulates the expression of many genes. As a result a large number of *trans*-eQTLs co-localize at the position of a master regulator, yielding an eQTL hotspot. Genes from such an eQTL hotspot might be sharing their regulatory functions and be expressed simultaneously. The genetic study of co-expression patterns and their biological significance therefore is of great interest for functional studies and for the identification of genes involved in processes of interest.

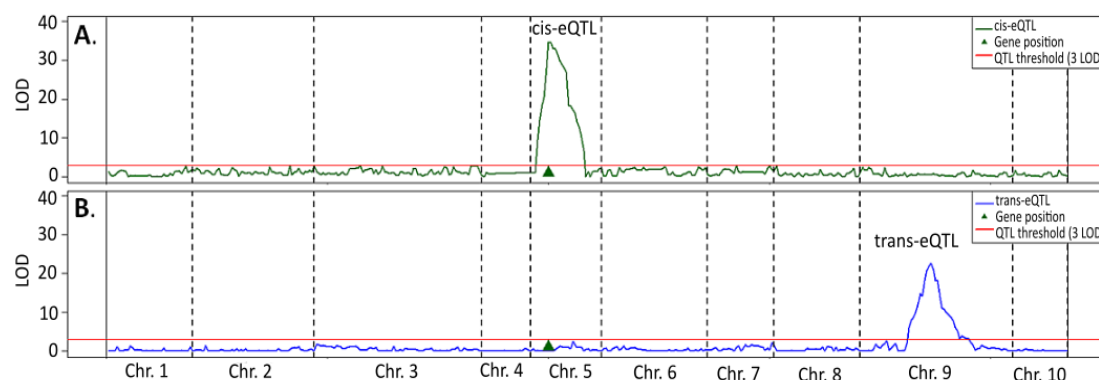


Figure 6: Types of eQTLs based on the physical position of gene and eQTL position (under study). **A.** *cis*-eQTL, result from allelic variation of the gene under study. **B.** distant *trans*-eQTL, gene is physically located in different location from its eQTL.

Scope of the thesis

The objective of this thesis is to unravel the genetics of seed quality and seedling vigour traits in *B. rapa* using a systems genetics approach. *Brassica* seeds are of high economic importance due to multiple uses as vegetable oil or condiments and as starting point of the crop's life cycle. *B. rapa* is an extremely diverse *Brassica* species which includes many vegetable and oilseed crops. At present, *B. napus* is the most important source of vegetable oil worldwide, but *B. rapa* is often introgressed to broaden its narrow genetic base resulting in genetic improvements. Therefore, the acquired knowledge is useful for the scientific community working in *B. napus* as well as other *Brassica* species.

Our hypothesis for this thesis is that transcriptional regulation of genes during seed development determines the composition and content of seed reserves, and that these seed reserves play a major role in seed germination and seedling growth, especially at the heterotrophic stage under optimal and sub-optimal conditions. Therefore, transcript profiling was carried out in the developing seeds, using genetical genomics. Fatty acids (FAs) were measured in ripe dry seeds of doubled haploid (DH) lines and integrated with transcriptome data using a systems genetic approach. The DH population was developed from a cross of two contrasting *B. rapa* parents, an oil-type yellow sarson (YS143; accession number FIL500) and a vegetable-type pak choi (PC175; Accession number VO2B0226). These two genotypes YS143 and PC175 are contrasting in their phylogenetic, morphological, metabolites and transcriptome characteristics (Zhao et al., 2005; Pino Del Carpio et al., 2011a; Pino Del Carpio et al., 2011b; Basnet et al., 2013). YS143 has yellow seed colour, large seed size, high oil content, faster germination rate, longer root and shoot length and is sensitive to salt stress, while PC175 has brown or black seeds, smaller seed size, slower germination rate, and shorter root and shoot length, but it is less sensitive to salt stress. QTLs were detected for traits related to seed germination and seedling vigour under different conditions.

Chapter 2 introduces the genetic diversity of *B. rapa* and identifies distinct groups in a core collection of 168 accessions. The genetic distances of accessions were compared with metabolic and morphological distances using multivariate statistical approaches. The geographical distribution of the accessions was very much congruent with genetic, metabolic and morphological diversity. Association studies were performed to find marker-metabolite associations for carotenoids and tocopherols. Knowledge of genetic relatedness allowed the choice of parents to create a segregating population for QTL analyses by maximizing genetic variation between the parents.

Chapter 3 describes the morphological development and the transcriptional signatures of developing seeds from yellow- and brown/black-seeded genotypes. Gene modules with different co-expression patterns, associated with temporal and/or genotypic contrasts, are shown. The variation in gene expression is shown to be due to developmental stages rather than morphotype differences. Finally, we focus on co-expression patterns of genes related to lipid metabolism and identify putative *cis*-regulatory elements (also called motifs).

Chapter 4 studies the genetic variation of traits related to seed germination and seedling vigour under non-stress and salinity stress in the doubled haploid population. We identify the QTL regions, where QTLs for seed germination and/or seedling vigour traits under both non-stress and salt stress conditions are co-located. Co-localization of eQTLs of two candidate genes, selected based on similar studies in the model species *A. thaliana* and closely related species *B. napus*, with QTL hotspots suggest a role of these genes in the trait variation.

Chapter 5 is a systems genetics study of fatty acid composition and transcriptional variation of genes related to lipid metabolism. Fatty acid QTLs (faQTLs) and eQTLs are mapped; for some FAs, major faQTLs co-localize with eQTL hotspots. Gene regulatory networks constructed for the economically most important fatty acids oleic-, erucic-, linoleic- and linolenic- acids, allow the identification of major hub genes possibly regulating lipid metabolism.

Chapter 6 summarizes and discusses the most important findings from this thesis. Finally, in this chapter we give suggestions for future studies of the genetic inheritance of complex traits, such as seed quality and seedling vigour.

Chapter 2

Comparative methods for association studies: a case study on metabolite variation in a *Brassica rapa* core collection

Dunia Pino Del Carpio^{1*}, Ram Kumar Basnet^{1,3*}, Ric C. H. De Vos^{2,3}, Chris Maliepaard¹, Maria João Paulo⁴, Guusje Bonnema^{1,3}.

¹ Laboratory of Plant Breeding, Wageningen University, Wageningen, The Netherlands,

² Plant Research International, Wageningen University and Research Centre (WUR), Wageningen, The Netherlands,

³ Centre for BioSystems Genomics, Wageningen, The Netherlands,

⁴ Biometris-Applied Statistics, Wageningen University and Research Center, Wageningen, The Netherlands

*equal contributors

Published in PLoS ONE (2011) 6(5): e19624.

Abstract

Association mapping is a statistical approach combining phenotypic traits and genetic diversity in natural populations with the goal of correlating the variation present at phenotypic and allelic levels. It is essential to separate the true effect of genetic variation from other confounding factors, such as adaptation to different uses and geographical locations. The rapid availability of large datasets makes it necessary to explore statistical methods that can be computationally less intensive and more flexible for data exploration. A core collection of 168 *Brassica rapa* accessions of different morphotypes and origins was explored to find genetic association between markers and metabolites: tocopherols, carotenoids, chlorophylls and folate. A widely used linear model with modifications to account for population structure and kinship was followed for association mapping. In addition, a machine learning algorithm called Random Forest (RF) was used as a comparison. Comparison of results across methods resulted in the selection of a set of significant markers as promising candidates for further work. This set of markers associated to the metabolites can potentially be applied for the selection of genotypes with elevated levels of these metabolites. The incorporation of the kinship correction into the association model did not reduce the number of significantly associated markers. However incorporation of the STRUCTURE correction (Q matrix) in the linear regression model greatly reduced the number of significantly associated markers. Additionally, our results demonstrate that RF is an interesting complementary method with added value in association studies in plants, which is illustrated by the overlap in markers, identified using RF and a linear mixed model with correction for kinship and population structure. Several markers that were selected in RF and in the models with correction for kinship, but not for population structure, were also identified as QTLs in two bi-parental DH populations.

Key words: *Brassica rapa*, core collection, metabolites, population structure, association studies, random forest.

Introduction

In plants association mapping has been developed as a tool to relate genetic diversity, expressed as allelic polymorphisms, to the observed phenotypic variation in complex traits without the need to develop mapping populations. Results obtained with association mapping methods in various crops indicate that this technique can be successful in the identification of markers linked to genes and/or genomic regions associated to a desirable trait (Remington et al., 2001; Simko et al., 2004; Thornsberry et al., 2001; Agrama et al., 2007; Kraakman et al., 2006; Zhao et al., 2007).

However, one of the most important constraints in the use of association mapping in crop plants is unidentified population sub-structure, which arises as a result of adaptation, genetic drift, domestication or selection (Thornsberry et al., 2001; Wright and Gaut, 2005). Spurious associations due to population structure may lead to false positive associations, if the cause of the correlation is not tight genetic linkage between polymorphic locus and the locus involved in the trait, but disproportional representation of the trait in one subpopulation (Brescaghello and Sorrells, 2006).

As a consequence, when association mapping is used to identify genes responsible for quantitative variation in a group of accessions, there is enough evidence that confounding will be a significant problem, especially if the trait varies geographically, as is the case for example of flowering time (Thornsberry et al., 2001; Aranzana et al., 2005; Yu et al., 2006).

Several methods can be used to infer multiple levels of relatedness in a population (Ritland et al., 1996; Yu et al., 2006). The STRUCTURE program uses a Bayesian approach to cluster accessions of a collection into subpopulations on the basis of multilocus genotype data (Pritchard et al., 2000; Falush et al., 2003, 2007). Designed statistical tests using PCA have also been used to check for the existence of population structure in a data set and monitor the number of significant principal component axes (Price et al., 2006; Reeves and Richards, 2009; Patterson et al., 2006). Similarly, kinship coefficients approximate identity by descent between pairs of accessions. In several association studies information about population structure and/or kinship has been included into the general linear regression and mixed linear models (Pritchard et al., 2000; Zhao et al., 2007; Yu et al., 2006; Malosetti et al., 2007). Results obtained in some studies suggest that the method that accounts both for subpopulations and kinship (also called the “QK method”) is the most appropriate for association mapping (Yu et al., 2006).

A different statistical approach, which carries one or more advantages above most other methods, is the Random Forest (Breiman, 2001). This is a tree-based method that has been used for marker trait associations with human disease data, because it allows the ranking and selection among very large sets of predictor variables (markers) that best explain the phenotype (Lunetta et al., 2004; Ye et al., 2005). This method is computationally very fast, scale-free and makes no strong assumptions about the distribution of the data. For emerging types of datasets like metabolite profiles, transcript profiles and the very large SNP datasets that emerge due to the rapid development of whole genome sequencing technology, it is

necessary to consider and validate association methods that can handle these high dimensional data sets.

Furthermore, the power to detect epistasis in moderately sized populations in general is low, while Random Forest can implicitly use interactions among regressor variables to predict the phenotype and can help identify multi-locus epistatic interactions (Jiang et al., 2009; Chen et al., 2007a).

For this study we choose to work with a core collection of 168 *Brassica rapa* accessions, representing the wide variation in crop types (hereafter called morphotypes) and geographical origins. *B. rapa* has been cultivated for many centuries in different parts of the world, increasing the variation within the species as a result of breeding. *B. rapa* is a diploid species which includes vegetable-, fodder- and oil crops. The leafy vegetables include both heading types (Chinese cabbage) and non-heading types (among others pak choi, mizuna, mibuna, komatsuna and broccoletti, consumed for its inflorescences), the turnips include vegetable and fodder turnips, and the oil crops include both annual and biannual crops. Most leafy vegetables, turnips and biannual oil types are self-incompatible and as a consequence the genebank accessions of this type are heterogeneous and plants are heterozygous. A smaller group of *B. rapa* is formed by the sarsons (brown sarson (*dichotoma*), toria (*dichotoma*) and yellow sarson (*trilocularis*)) characterized by very early flowering and self-compatibility of many accessions, which results in heterogeneous accessions with merely homozygous plants (Zhao et al., 2005). Modern cultivars and breeding lines from seed companies are homogenous heterozygous hybrids and homozygous inbred lines.

In a previous study the genotypic fingerprinting of a large collection of 160 accessions showed that there is considerable genotypic variation within the *B. rapa* gene pool (Zhao et al., 2005). The hierarchical cluster analysis revealed that accessions from the same geographical origin (Europe, Asia and India) are more related to each other genetically than accessions representing similar morphotypes from different geographical regions. These accessions from the same origin are genetically related possibly because they share part of their breeding history (Zhao et al., 2005).

Previously, in a collection of 160 *B. rapa* accessions association analysis with correction for population structure led to the identification of 27 AFLP markers, related to the variation in leaf and seed metabolites as well as morphological traits (Zhao et al., 2007). In the present study we consider the genetic association between markers and tocopherols, carotenoids, chlorophylls and folate in a core collection of 168 *B. rapa* accessions of different morphotypes and origin. We explore the results obtained with association methods that correct for kinship and population structure which mainly aim to reduce the rate of false-positive associations, and in addition we make use of Random Forest for comparison to the commonly used association methods.

Materials and methods

Plant material

The *B. rapa* core collection included a total of 168 accessions of diverse morphotype and origin (Supplementary Table S1). The leafy vegetables, (Chinese cabbage, pak choi and Japanese cultivars), neep greens, turnip rape, brocoletto (turnip tops) and turnip types are mainly self-incompatible and as a consequence the accessions are heterogeneous and heterozygous. The annual yellow sarson oil seed accessions are self-compatible, which results in homozygous plants. The modern cultivars and breeding lines from seed companies are homogeneous hybrids and inbred lines. 137 accessions were obtained from the Dutch Crop Genetic Resources Center (CGN) in Wageningen, the Chinese Academy of Agricultural Sciences (CAAS)-Institute for Vegetable and Flowers (IVF) and the CAAS Oil Crop Research Institute (OCRI) genebanks and the Osborn Lab, while six different breeding companies (Supplementary Table S1) provided 31 accessions. For the metabolite profiling two plants per accession were sown in the greenhouse (2006) under the following conditions: 16 hours light and temperature fluctuation between 18 and 21°C. The plants were distributed over two tables in a randomized design with one plant per accession on each table. In the 5th week after transplanting the leaf material (youngest expanded leaves) was harvested per plant. Upon harvesting, all plant materials were snap-frozen in liquid nitrogen and ground into a fine powder using an IKA A11 grinder cooled with liquid nitrogen. Frozen powders were stored at -70°C until analyses. DNA was extracted from the ground and frozen material with the DNAeasy kit (Qiagen, USA).

Metabolite analyses

Folate extraction and analysis

From each frozen powder, 0.15 g was weighed and 1.8 ml of Na-acetate buffer containing 1% ascorbic acid and 20 µM DTT, pH 4.7, was added. After sonication for 5 min and heating at 100°C for 10 min, total folate content of samples was quantified using a *Lactobacillus casei*-based microbiological assay, after enzymatic deconjugation for 4 hours at 37°C pH 4.8, with human plasma as a source of γ -glutamyl hydrolase activity (Sybesma et al., 2003). Each extract was assayed in 4-6 replicates using different dilutions. The total technical variation of this analysis was determined using 7 replicate extractions from the same frozen powder of two different randomly chosen genotypes, and was 5.5% and 6.9%, respectively.

HPLC analyses of lipid-soluble phytonutrients

Extraction and analyses of carotenoids, tocopherols and chlorophylls were performed as described in Bino et al. (2005). In short, 0.5 g of FW of frozen powder was taken and extracted with methanol-chloroform-Tris buffer twice, the chloroform fraction was dried using nitrogen gas and taken up in 1 ml of ethylacetate. The chromatographic system consisted of a W600 pump system, a 996 PDA detector and a 2475 fluorescence detector (Waters Chromatography), and an YMC-Pack reverse-phase C30 column (250 x 4.6 mm,

particle size 5 μ m) at 40°C was used to separate the compounds present in the extracts. Data were analyzed using Empower Pro software (Waters Chromatography). Quantification of compounds was based on calibration curves constructed from respective standards. The total technical variation was between 2 and 8 percent, depending on compound, as was established using 12 extractions of the same frozen powder from a randomly chosen genotype.

Genotypic data

The AFLP procedure was performed as described by Vos et al., (1995). Total genomic DNA (200 ng) was digested with two restriction enzymes *Pst* I and *Mse* I and ligated to adaptors. Pre amplifications were performed in 20 μ l volume of 1x PCR buffer, 0.2 mM dNTPs, 30 ng of adaptor primer, 0.4 Taq polymerase and 5 μ l of a 10x diluted restriction ligation mix, using 24 cycles of 94° C for 30 seconds, 56° C for 30 seconds and 72° C for 60 seconds. Pre-amplifications products were used as template for selective amplification with three primer combinations (P23M48, P23M50 and P21M47).

For the *Myb* family targeted profiling, total genomic DNA was digested using the following enzymes per reaction: Hae III, Rsa I, Alu I and Mse I and ligated to an adaptor. Pre amplifications with one primer directed to a common *myb* motif (Dr. Gerard van der Linden, Wageningen UR Plant Breeding, unpublished results) and one adaptor primer were performed in 25 μ l of 1X PCR buffer (with 15 Mm MgCl₂), 0.2 mM dNTPs, 0.8 pMol Gene specific primer, 0.8 pMol Adapter primer, U Hotstar Taq polymerase (Qiagen) and 5 μ l of a 10X diluted restriction ligation mix. Amplification products were used as template for selective amplification.

AFLP and *Myb* profiling images were analyzed using Quantar Pro™ software. This marker dataset (359 polymorphic bands) was scored as present (1) or absent (0) and treated as dominant markers. A map position could be assigned for 69 markers from this dataset; these markers were distributed over different positions in the linkage groups of a doubled haploid population (Pino Del Carpio, unpublished results).

For microsatellite (SSR) screening, 28 primers were selected for amplification in the accessions of the core collection. From the primers 10 were genomic and 18 were new Est based SSRs (Dr. Ma RongCai, Dr Tang Jifeng (WUR-PBR)). The primers were selected because of their map position in different maps of *B. rapa* and distribution over all the linkage groups (A01-A10) (Pino Del Carpio et al., 2011b). Microsatellites scores were converted to binary data per observed allele (194 fragments of defined size) as present (1) or absent (0) and were also treated as dominant markers.

Assessment of population structure

Marker data (AFLP, *Myb* and SSR) were used to identify the different subgroups and admixture within the accessions of the core collection through a model of Bayesian clustering for inferring population structure. For the SSRs only the most frequent SSR allele was taken into account to avoid over representation of the SSR loci.

A total of 539 markers was included in the analysis, and ploidy was set to one. The number of subpopulations was determined using the software STRUCTURE 2.2 (<http://pritch.bsd.uchicago.edu/software>), by varying the assumed number of subpopulations between one and ten, with a total of 300,000 iterations for Markov Chain Monte Carlo repetitions and 100,000 burns-in.

In addition, we also followed the procedure PCO-MC as described in Reeves and Richards (2009), to assess population structure. The method uses principal coordinate analysis (PCO) and clustering methods to infer subpopulations in a collection of accessions. We chose this method to complement the analysis performed by STRUCTURE because it is computationally efficient and model free and has been shown to be capable of capturing subtle population structure (Reeves and Richards, 2009). We used software NTSYS version 2.2 (Rohlf, 1998) to produce pairwise distances, among all accessions, based on the Jaccard measure. Principal coordinates were obtained based on the distance matrix as described by Reeves and Richards (2009). Then procedure PROC MODECLUS in SAS 9.1 software (SAS Institute, Cary, NC) was used to group the accessions into subpopulations according to kernel density estimates in the PCO space. Subpopulations were formed by decreasing order of the kernel densities, starting with the largest estimated kernel density (by setting method = 6 at PROC MODECLUS). We performed a test to determine which subpopulations were significantly distinct from the rest, using PROC MODECLUS, and estimated stability values for the subpopulations using the PCO-MC software (<http://lamar.colostate.edu/~reevesp/PCOMC/PCOMC.html>) (Reeves and Richards, 2009). The PCO plot of the first two components was drawn in DARwin software version 5.0.155 (Perrier and Jacquemoud-Collet, 2006).

Summary statistics of metabolite variation

Box plots were chosen as a tool to explore the variation of metabolite concentrations according to different STRUCTURE subpopulations. One-way ANOVA was performed for each metabolite to find the mean differences among the four STRUCTURE subpopulations. Least significant differences (LSD) were calculated to compare the differences of means of metabolite content between the four subpopulations obtained with STRUCTURE. Boxplots, ANOVA and LSD calculations were performed using R statistical software.

Association analysis

Association analysis was performed in several steps of increasing complexity; with and without correction for population structure (Yu et al. 2006) using TASSEL (www.maizegenetics.net). A total of 243 markers, with an allelic frequency higher than 10%, were included in the association analysis. Since AFLP and *Myb* markers gave dominant marker scores and TASSEL works with co-dominant data, within TASSEL we set the ploidy to one to work with dominant scores as we had done with STRUCTURE. In the case of the 28

microsatellites all alleles were included within TASSEL in a different run as codominant markers.

In the first step a “naïve” model was used to associate each marker to the trait,

$$\text{trait} = \text{marker} + \text{error} \quad (1)$$

This model was fitted by a least squares fixed effects linear model in TASSEL where the markers are considered as a factor taking the value 0 (fragment absent) or 1 (fragment present). In this case a t-test could also have been used to test association since we only have two classes for the marker. In this “naïve” model population substructure was not taken into account.

In the second step the vector of subpopulations memberships Q obtained from STRUCTURE was added as a fixed term to the previous model

$$\text{trait} = \text{marker} + Q + \text{error} \quad (2)$$

In the third step we corrected for kinship using a linear mixed model available in TASSEL. The model can be written as

$$\text{trait} = \text{genotype} + \text{marker} + \text{error} \quad (3)$$

where random terms are underlined. Genotype is a random factor with the different genotypes or accessions in the population. Kinship coefficients were calculated using SPAGeDi (Hardy and Vekemans, 2002). Like for the calculation of STRUCTURE, for the SSRs only the most frequent SSR allele was taken into account to avoid representation of the SSR loci. We have $V_G = \sigma^2 K$; V_G is the variance-covariance matrix of the random genotype effects, K is the matrix of kinship coefficients and σ^2 is the additive genetic variance.

In the fourth and final step we correct for kinship as well as population structure using a linear mixed model that combines the information contained in the two previous models. It is also known as the Q+K method (Yu et al., 2006).

$$\text{trait} = \text{genotype} + Q + \text{marker} + \text{error} \quad (4)$$

As before, genotype is a random factor, with covariances given by the kinship matrix K and Q is a fixed term containing the subpopulation memberships. The model is similar to those described by Yu et al., (2006) and Malosetti et al., (2007). Here we used the same set of AFLP, *Myb* and SSRs data to estimate both K and Q. The percentage of variation was also implemented in TASSEL and extracted from the output for further analysis and comparison.

Correction for multiple testing

The p-values resulting from all the models for association analysis were corrected for multiple testing using a resampling method as implemented in the R package “multtest” (Pollard et al., 2005).

Random Forests

Random Forests (RF) regression (Breiman, 2001) was used in this study to find markers (among the 243 AFLP and *Myb*, and 28 SSR marker set) associated to the tocopherol, carotenoids, flavonoids and folate metabolites. This method uses a bagging approach by

bootstrapping samples (Gislason et al., 2006) and gives the relative importance of each marker in the regression of metabolites. In this study, RF was performed using 5,000 regression trees for each analysis. Each tree is formed on a bootstrap sample of the individuals (the training dataset), while individuals that are not in the bootstrap sample (out-of-bag samples = OOB), are used for estimation of the mean squared error of prediction. Within each regression tree, at each split of the tree, a random subset of the markers is considered as a candidate set of markers for a binary split among the set of individuals. The partitioning of the samples is continued until homogeneous groups of small number of samples remain.

This procedure is fast and can handle high dimensional data (predictor variables >> number of samples). Each tree is fully grown (unpruned) to obtain low-bias, high variance (before averaging) and low correlation among trees. Finally, RF averages are calculated over all the trees and results in low-bias and low variance of predictions of the trait based on the markers used in the Random Forests (Svetnik et al., 2003). This method has an internal cross-validation (using the OOB samples) and has only a few tuning parameters which, if chosen reasonably, do not change results strongly (Gislason et al., 2006).

The parameter “mtry”, which indicates the number of random variables considered at each split node, was optimized by choosing the “mtry” with the highest percentage of explained variation among separate RF analyses done on “mtry” values 3, 6, 12, 24, 48 and 96 successively on the same dataset. The variance explained in RF is defined as $1 - (\text{Mean square error (MSE)} / \text{Variance of response})$, where MSE is the sum of squared residuals on the OOB samples divided by the OOB sample size (Pang et al., 2006). The “mean decrease in MSE” (InMSE) was considered to quantify the importance of each marker. The higher the “InMSE” value of the marker, the greater the increase in explained variation when it is included in the model.

In general, RF yields only the relative importance of markers that explain the variation present in metabolites, but does not give a significance threshold level to select a subset of associated markers. Therefore, a permutation method was used to calculate the significance of each marker association in this study (Wang et al., 2010a). All the observations of a metabolite (the response in the regression) were permuted to destroy the association between markers and metabolite, and RF analyses were repeatedly conducted on the permuted metabolite data 1000 times. For each metabolite, the “IncMSE” values of each marker from 1000 RF runs on permuted metabolites were stored, and used as a “null distribution” of the IncMSE value to assess the significance threshold of each marker. Then, the IncMSE values of each marker obtained from RF analysis on the original unpermuted metabolite dataset was compared to this “null distribution” at 0.05 level of significance to determine significantly associated markers.

RF regressions of metabolites on markers were conducted using the “RandomForest” package of the R-software (Liaw and Wiener, 2002).

Network visualization of metabolite and marker correlation

A network is an extended graph, which contains additional information on the vertices and edges of the graph (de Nooy et al., 2005). We used full-order partial correlation coefficients to construct correlation network of metabolites to remove the correlation between metabolites due to direct and indirect dependencies on the upstream metabolites in the pathway. We included in the network graph all the markers that were associated to the metabolites after correction for multiple testing ($\alpha = 0.05$). Since we are focusing on the tocopherols, carotenoids and folate pathway, correlation analysis can give spurious correlation between the metabolites due to the effect of upstream metabolites of the pathway. Partial correlation measures only the direct or unique parts of relation between metabolites controlling the effects of other metabolites of the pathway (Opgen-Rhein and Strimmer, 2007). The only significant non-zero pairwise partial correlation coefficients ($\alpha = 0.05$) between metabolites were shown in network. The vertices of the network are the metabolites, in this case tocopherols, carotenoids, chlorophylls and folate, and associated markers, whereas the edges correspond to metabolite-metabolite partial correlations and marker-metabolite association. For the visualization of the marker-metabolites association, the p-values obtained from model (4) were transformed into $-\log_{10}(\text{p-value})$. The network was constructed using the Pajek graph drawing software (Batagelj and Mrvar 2003).

Results***Principal coordinate analysis (PCO) and population structure of the core collection***

The genetic population structure of the core collection of 168 accessions was inferred using 553 markers (AFLP, *Myb* and SSR polymorphic bands). The Bayesian clustering method as implemented in STRUCTURE revealed four subpopulations. Subpopulation 1 included oil types of Indian origin, spring oil (SO), yellow sarson (YS) and rapid cycling (RC) (SO, YS and RC); subpopulation 2 included several types from Asian origin: pak choi (PC), winter oil, mizuna, mibuna, komasuna, turnip green, oil rape and Asian turnip (PC+T); subpopulation 3 included mainly accessions of Chinese cabbage (CC) and subpopulation 4 included mostly vegetable turnip (VT), fodder turnip (FT) and brocoletto accessions from European origin (VT+FT) (Figure 1B).

There was a high level of admixture between the different subpopulations. Of the 168 accessions, 109 were assigned to a subpopulation with a membership probability of $p > 0.70$. Fifty-nine accessions were assigned to more than one subpopulation and had membership probabilities below 0.7 corresponding to several subpopulations (Supplementary Table S1).

The PCO-MC method couples principal coordinate analysis to a clustering procedure for the inference of population structure from multi-locus genotype data. The PCO and STRUCTURE output produced comparable results. After the PCO analysis, in the second dimension one small distinct, statistically significant subpopulation, corresponding to oil types of Indian origin, could be distinguished. This subpopulation corresponds to subpopulation 1 (SO, YS and RC) as identified in STRUCTURE (Figure 1A). In the first dimension, the three

subpopulations as defined in STRUCTURE form three clusters with overlap. On the right, the cluster of yellow dots corresponds to accessions in subpopulation 4 (VT and FT from Europe) as defined in STRUCTURE, and on the left the blue dots represent the accessions corresponding to subpopulation 3 (CC), while the green dots represent accessions that correspond to subpopulation 2 (PC and T from Asia). When the top 5 components are calculated, they together account for 30% of the total variation present in the core collection. As many principal component loadings would have been needed to account for the variation within this collection, we decided to include STRUCTURE output into the association model to correct for population structure.

In figure 2 we show the frequencies of the different kinship coefficient classes. The highest frequency was found for values between 0–0.05 (79.47%) while the second highest frequency was found for values between 0.05–0.1 (11.21%). These values are similar to the ones obtained in *Brassica napus* (Jestin et al., 2011) in which the kinship calculation indicates a low level of relatedness between the accessions, with only few accessions being more related to each other.

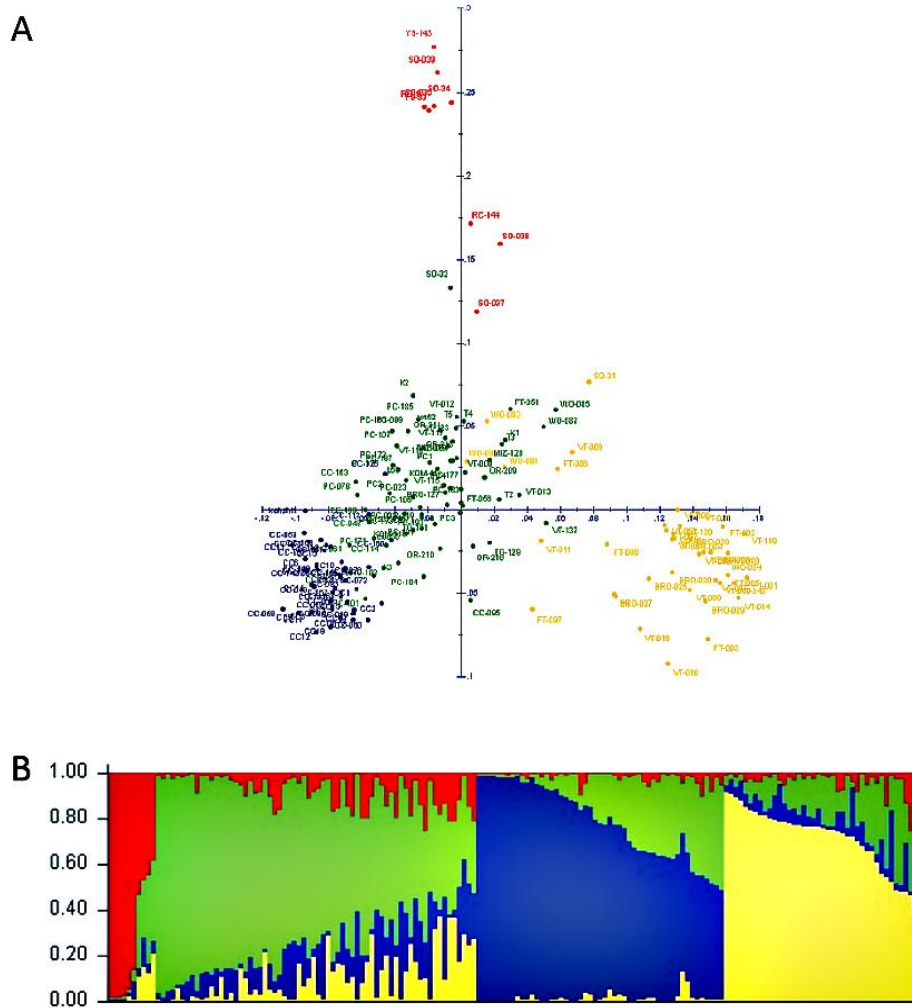


Figure 1: Principal co-ordinate analysis (PCO) - scatter plot (A) and STRUCTURE (B) results. Colors define subpopulations: red (oil: subpopulation 1), green (PC+T: subpopulation 2), blue (CC: subpopulation 3) and yellow (VT+FT: subpopulation 4).

Metabolite variation

To estimate the variation within and between the different *B. rapa* morphotypes, boxplots were constructed based on the total content value per metabolite in each subpopulation as defined by STRUCTURE (Figure 3). Visual inspection of the box plots and the least significant differences (LSD) in metabolite content between subpopulations showed variation in the amount of most of the carotenoids and folate between these subpopulations. Conversely, the content of chlorophyll *b* and lutein was significantly different between few subpopulations and the content of tocopherols was just significantly different between the Chinese cabbage (CC) subpopulation 3 compared to the other subpopulations (Supplementary Table S2).

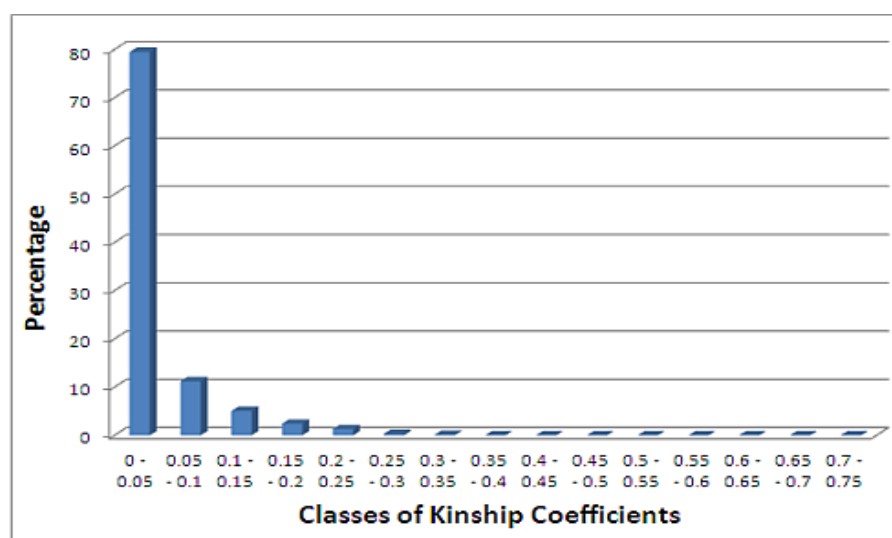


Figure 2: Distribution of kinship coefficients among 168 accessions of the *B. rapa* core collections.

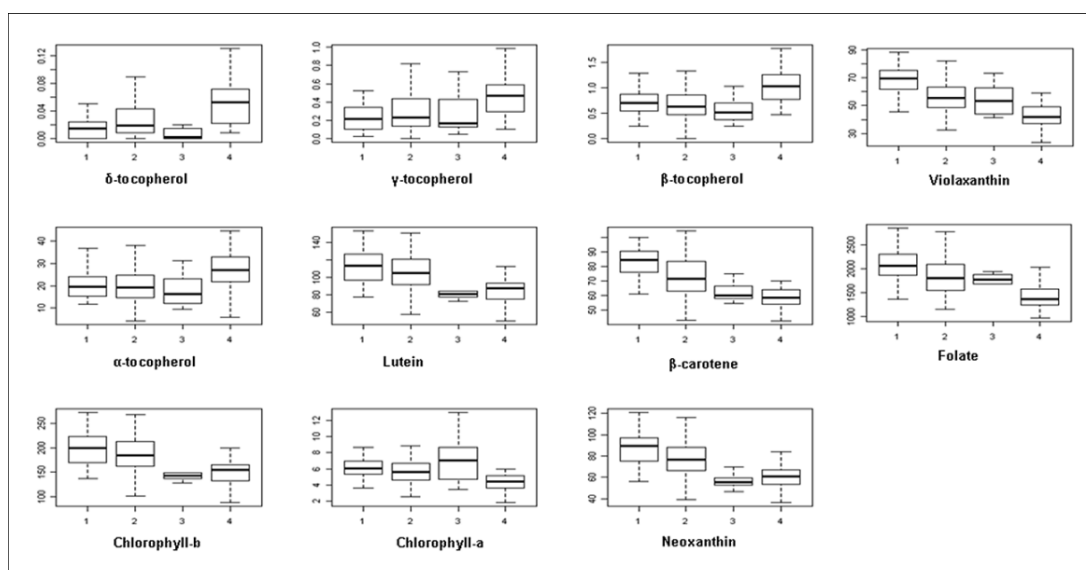


Figure 3: Boxplots of metabolite content variation present in sub-populations. The numbers indicate subpopulations as defined with STRUCTURE. Oil: subpopulation 1, PC+T: subpopulation 2, CC: subpopulation 3 and VT+FT: subpopulation 4.

Association analysis

Using linear and linear mixed models

Because many of the phenotypic trait values showed a distribution highly correlated to the underlying population structure it was expected that the number of significantly associated markers would differ to a large extent between different metabolites and between analysis methods, as shown in Table 1.

Table 1: Association mapping results from the different linear models and Random forests (RF) indicating the numbers of significant markers after correction for multiple testing (FDR p-value ≤ 0.05) for each metabolite.

	Tocopherols				Carotenoids				Chlorophylls		Folate
	δ -tocopherol	γ -tocopherol	β -tocopherol	α -tocopherol	Lutein	β -carotene	Neoxanthin	Violaxanthin	Chlorophyll <i>b</i>	Chlorophyll <i>a</i>	
Model (1)*	2	24	55	32	22	97	89	92	22	28	112
Model (2)	0	0	0	0	0	0	0	4	0	0	2
Model (3)	2	26	56	34	22	98	88	96	22	23	109
Model (4)	0	0	0	0	0	5	0	5	0	0	1
RF	16	24	12	8	16	36	39	34	32	17	28
RF-Model (4)**	0	0	0	0	0	3	0	3	0	0	1
RF-Model (1)**	1	11	4	2	6	31	26	30	9	7	21
RF-Model (3)**	1	11	4	2	6	32	26	30	9	7	21

Note: *Model (1): naïve model; model (2) correction for Q; model (3) correction for K; model (4) correction for K and Q; RF: Random Forest. ** - number of markers identified in both methods are listed.

To test for marker-trait associations we first applied an approach that did not include any correction for the level of relatedness or structure between accessions (model 1). As a result the number of significantly associated markers to a specific metabolite after multiple test correction was strongly inflated and ranged from 2 (for δ tocopherol) to 98 (for folate) per metabolite. The highest numbers of significant markers associated to a trait ($>.80$) were found for β -carotene, neoxanthin, violaxanthin and folate; these metabolites also showed the greatest variation in content between subpopulations.

To account for the level of relatedness between individuals we included the kinship correction (K matrix) in model (3). However, with the inclusion of this correction the number of significantly associated markers remained high (2-94). The results of these two models were highly similar not only in number but also in the identity of the significant markers for each metabolite.

In addition to the K matrix we introduced the STRUCTURE Q matrix as a correction. After accounting for population structure in model (2) the number of significant markers found per metabolite after a multiple correction step was dramatically reduced. Only for violaxanthin and folate few markers were identified. This drop down was as strong for the metabolites with subpopulation variation (carotenoids and folate) as for the tocopherols, which showed significant variation only between the CC subpopulation and the other subpopulations.

When we combined the information from the Q matrix and the K matrix in the full model (4), following the described approach (Yu et al., 2006), the performance is comparable to model (2), which includes the Q matrix only, in both the obtained number of associations and the identity of associated markers, except for β -carotene with five markers identified in model 4.

After correcting for multiple testing in the QK correction model, only ten markers remained significantly associated with metabolites: Alu_M476_0, pTAmCAC_148_3, Hae_M294_2, pGGmCAA_335_2 and pTAmCAT_312_3 for β -carotene; Alu_263_6, pTAmCAC_101_7 and pTAmCAC_270_9, Br13 and Br46 for violaxanthin and pGGmCAA_335_2 for folate.

To summarize the results obtained from the full model (4), we constructed a network with a total of three Myb, five AFLP and two microsatellite markers significantly associated to the metabolites ($p < 0.05$). The network allowed us to connect the metabolites of similar pathways through markers (Figure 4). The overlap of significant associated markers between all the pathways (carotenoids, tocopherols, chlorophylls and folate) was very limited as expected if we consider that biochemically different precursors are involved. We found only one marker (pGGmCAA_335_2) that was significantly associated to folate and β -carotene.

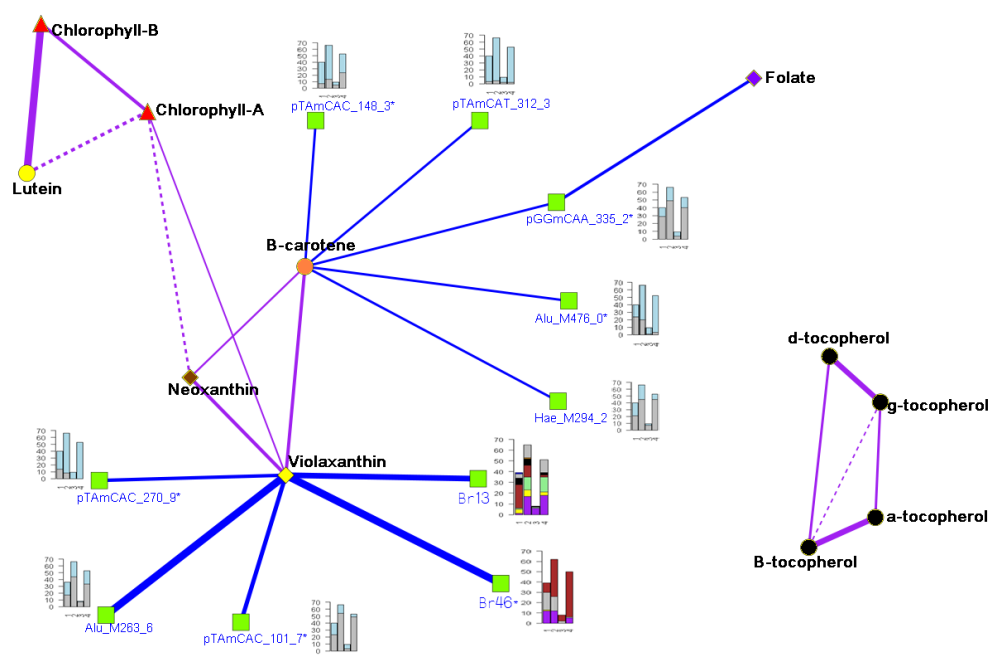


Figure 4: Network of partial correlation between metabolites, and marker-metabolite association under model 4 (QK correction). The thicker the line is the stronger the correlation and or association. Shape and color of vertices indicate metabolites and associated markers: β -carotenoids – round-orange; chlorophylls – triangle-red; folate - diamond-purple; lutein - round-yellow; neoxanthin - diamond-brown; tocopherols round - black; violaxanthin - diamond-yellow and associated markers (model 4)- square-green. The allele frequency distribution of each associated marker according to STRUCTURE sub-populations is illustrated with barplot. Colours in barplots represent different marker alleles. *- indicates markers that are common between model 4 (QK correction) and RF.

Random Forests (RF)

The number of significantly associated markers per metabolite ranged from eight for α -tocopherol to 39 for neoxanthin. Interestingly, when compared to the simple model (1), the number of significant markers obtained with the RF approach was much lower for all the metabolites except for δ -tocopherol.

Nonetheless, the overlap of significant markers between methods is large; many of the significant markers found with RF were also significant with the simple model (1) and with the model with correction for kinship (3). For example, an overlap between 20% and 30% of significant markers was observed for β -carotene, neoxanthin, violaxanthin, folate, α -tocopherol and β -tocopherol (Table 1; Supplementary Table S3).

In contrast, when the results obtained with Random Forest are compared to the results obtained with the full model (4) six out of the ten markers from this model are included in the Random Forest output. In the case of the microsatellite markers the overlap between significantly associated markers in Random Forest and in model (1) was high and almost complete for both markers except for those identified for δ -, β - and α -tocopherol. Additionally, one out of two significant SSR markers from model (4) was found also significant in the Random Forest output.

Discussion

An important consideration for the use of association mapping in crop plants is the presence of population structure. If a group of diverse accessions is chosen for this type of studies the risk exists that some of the accessions are more closely related to each other than the average pair of individuals taken at random in a population (Brescaghello and Sorrells, 2006). In our study we identified with STRUCTURE the presence of 4 subpopulations, which showed correlation with the origin and morphotypes of *B. rapa*. Results from both principal coordinate analysis (PCO) and STRUCTURE illustrate the highly admixed nature of the accessions within this collection. We decided to use membership probabilities obtained from STRUCTURE in the association mapping model to correct for populations structure, as it is widely used method. In addition, correction for kinship was included in other models.

In the four models we explored the impact from STRUCTURE (model 2), kinship coefficients (model 3) or both (model 4) in the association models.

Correcting for the level of relatedness using the Q matrix from the STRUCTURE output, resulted in a significant reduction of the number of marker-trait associations as shown by comparing model (1) and both models (2) and (4). Although there was always some overlap between the marker-trait associations identified by these models, new associations arose with models (2) and (4).

The inclusion of the kinship matrix in models (3) and (4) did not reduce the number of significant marker-trait associations. This was most likely due to the fact that kinship values were very low and the accessions of the core collection showed similar levels of relatedness. The results from STRUCTURE and the identical levels of relatedness as observed in K seem to contradict. Similarities based on the Jaccard measure were also tested in model (3), with the same results as obtained with the similarities obtained from SPAGeDi.

We tested thereafter both corrections in phenotypic models identical to models (2) and (3) but without the marker effect, and compared the resulting residual variance with the “empty” model: trait = error. We found that whereas Q explained the phenotypic variation

by as much as 60% for some traits, the K matrix did not seem to explain any part of the phenotypic variation, for all traits. This seems to support earlier evidence that K alone in some cases may not correct for population structure (Jestin et al., 2011). In terms of how these methods performed in reducing the false positive rate, we observed that metabolites with a distribution highly correlated to the underlying population structure, like for example the carotenoids, still retained the highest number of associated markers in all the statistical models. As a result, in spite of introducing a correcting term in our models we still expect some false positives within this list of significant markers. Even in association studies with *Arabidopsis* inbred lines it is difficult to distinguish true associations from false ones because of confounding by complex genetics and population structure (Atwell et al., 2010).

In the present study we considered the use of Random Forests (RF) as a complementary method to our association study. The performance of this method in association analysis has been recently tested in *Arabidopsis* (Nemri et al., 2010). Within that study the overlap of RF and Fisher's exact test was considerable.

In our study we evaluated the RF results in comparison to the results obtained with the already validated and widely used model (4) and the simple model (1). One striking result of the RF analysis is the small number of associated markers that are found for all the metabolites in comparison to model (1). Random Forests is rather robust to outliers, as opposed to linear models, making it an attractive alternative to the traditional linear models. We decided to evaluate the overlap of RF and the simple model (1), which does not include any correction, and the full model, which includes the Q and K matrix correction (4). Seven out of eleven marker-trait associations found significant after multiple test correction with model (4) were also found significant with RF, while also many Random Forest markers were identified with models 1 and 3 (K correction).

Several markers that are associated with the metabolites studied, were also identified in QTL studies for the same metabolites in DH populations derived from crosses between two accessions yellow sarson (YS143) and pak choi (PC175) and their reciprocal cross (Supplementary Table S3), or map to regions that harbour structural genes in the metabolic pathway based on *Arabidopsis*- *B. rapa* genome synteny (data not shown). This is a confirmation of the effect of the marker-trait association and makes these markers important candidate genes for further study.

For eight of the eleven metabolites analyzed, Random Forest selected at least one marker that mapped in the QTL interval for the respective metabolites in the biparental QTL studies. For several metabolites except the tocopherols, these markers were also identified in model 1 (no correction) and 3 (kinship correction) but not in models 2 and 4 (with Q correction).

In these same two doubled haploid populations QTL for lutein and chlorophyll-a and -b overlap in the region where the marker pTAmCAC_148_3 is located and identified as significant for β -carotene by all models. In this genomic region of linkage group A03 the genes ϵ -cyclase, β -carotene hydroxylase and carotenoid isomerase are predicted based on synteny with *Arabidopsis* (Schranz et al., 2006) and represent potential candidate genes for

β -carotene and lutein. In the case of violaxanthin the marker Alu_263_6 was identified as associated in model 4 (K and Q correction). Alu_263_6 is 5 cM apart from the structural gene *Phytoene desaturase* that we mapped in the biparental DH population. For most markers map positions are not available, however the linked microsatellite marker Br13 and marker Alu_263_6 on A08, were both associated to violaxanthin.

In this study we have identified several markers that can be applied to screen *B. rapa* collections or breeding populations to identify genotypes with elevated levels of important metabolites that are considered as healthy compounds. While further validation of these markers for marker assisted selection in *B. rapa* is needed, at least the eight *Myb* and AFLP markers and two microsatellites markers found significant with model (4), after multiple testing correction (Benjamini and Hochberg, 1995), and also with Random Forest, plus the markers identified using both Random Forest and the models (1) and (3) should be considered as likely candidates for further work.

At present we are in the process of expanding the core collection so that association mapping within the four subpopulations becomes feasible and to increase the power of the statistical analysis. In an attempt to separate true from spurious associations and/or false negatives in future association studies using the present core collection we will follow a similar approach, which takes into account the level of relatedness between individuals (K and Q) and the use of Random Forest.

Acknowledgements

We would like to thank Harry Jonker and Yvonne Birnbaum for their help in the isoprenoids and folate analyses, and Johan Bucher for his help in the molecular marker work and greenhouse experiments.

Supporting Information

Supplementary information files are available at:

<http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0019624#s5>

Supplementary Table S1: Membership probabilities and group assignment of all accessions used in this study based on STRUCTURE (.xlsx).

Supplementary Table S2: LSD result of metabolite variation based on STRUCTURE subpopulations (.xlsx).

Supplementary Table S3: Overview of associations between markers and metabolites for all metabolites investigated in this study across methods with p-values after multiple testing corrections. For mapped markers genetic map positions are listed. For RF only significant markers are indicated. In the last column QTL identified in DH populations from crosses between YS143 and PC175 and their reciprocal cross (map presented in Lou et al., (2008)) are listed (.xlsx).

Chapter 3

Genome-wide analysis of coordinated transcript abundance during seed development in different *Brassica rapa* morphotypes

Ram Kumar Basnet^{1,2}, Natalia Moreno-Pachon¹, Ke Lin^{1,3}, Johan Bucher¹, Richard G F Visser^{1,2}, Chris Maliepaard^{1,2}, Guusje Bonnema^{1,2}

¹Wageningen UR Plant Breeding, Wageningen University and Research Center, Wageningen, The Netherlands

²Centre for BioSystems Genomics, Wageningen, the Netherlands

³Wageningen UR Bioinformatics, Wageningen University and Research Center, Wageningen, The Netherlands

Abstract

Brassica seeds are important as basic units of plant growth and sources of vegetable oil. Seed development is regulated by many dynamic metabolic processes controlled by complex networks of spatially and temporally expressed genes. We conducted a global microarray gene co-expression analysis by measuring transcript abundance of developing seeds from two diverse *B. rapa* morphotypes: a pak choi (leafy-type) and a yellow sarson (oil-type), and two of their doubled haploid (DH) progenies, (1) to study the timing of metabolic processes in developing seeds, (2) to explore the major transcriptional differences in developing seeds of the two morphotypes, and (3) to identify the optimum stage for a genetical genomics study in *B. rapa* seed. Seed developmental stages were similar in developing seeds of pak choi and yellow sarson of *B. rapa*; however, the colour of embryo and seed coat differed among these two morphotypes. In this study, most transcriptional changes occurred between 25 and 35 DAP, which shows that the timing of seed developmental processes in *B. rapa* is at later developmental stages than in the related species *B. napus*. Using a Weighted Gene Co-expression Network Analysis (WGCNA), we identified 47 “gene modules”, of which 27 showed a significant association with temporal and/or genotypic variation. An additional hierarchical cluster analysis identified broad spectra of gene expression patterns during seed development. The predominant variation in gene expression was according to developmental stages rather than morphotype differences. Since lipids are the major storage compounds of *Brassica* seeds, we investigated in more detail the regulation of lipid metabolism. Four co-regulated gene clusters were identified with 17 putative *cis*-regulatory elements predicted in their 1000 bp upstream region, either specific or common to different lipid metabolic pathways. This is the first study of genome-wide profiling of transcript abundance during seed development in *B. rapa*. The identification of key physiological events, major expression patterns, and putative *cis*-regulatory elements provides useful information to construct gene regulatory networks in *B. rapa* developing seeds and provides a starting point for a genetical genomics study of seed quality traits.

Key words: *Brassica rapa*, Leafy vegetables, Oil-seed, Seed development, Microarray, Co-expression network analysis, Transcript abundance, *cis*-regulatory elements.

Introduction

Brassica rapa ($2n = 2x = 20$; AA) is an important crop that consists of diverse morphotypes (also called crop types), including oilseed (annual crops yellow sarson and brown sarson, and biannual winter oils), leafy vegetables (Chinese cabbage, pak choi and many non-heading leafy types), turnip (fodder and vegetable turnip) and broccoletto. It contributes the A-genome to the amphidiploid oil crop canola (*B. napus* L; $n = 19$; AACCC). Yellow sarson and brown sarson are grown for oil production in the Indian sub-continent, and in Canada, because of their early maturity and shatter resistance. *Brassica* seed is important for both plant propagation and oil production.

Brassica seed is non-endospermic, which means that the endosperm is not retained in mature seeds and only the embryo is enclosed by the seed coat (Sabelli, 2012). Seed development goes through basically three overlapping stages: morphogenesis, seed filling and seed desiccation (Li et al., 2012b; Yu et al., 2010). Embryo development, also known as embryogenesis, starts after the double fertilization process of fusion of two sperm nuclei with the egg cell and the central cell nuclei, respectively, and the zygote goes through a series of cell divisions and differentiation events from a pre-globular and globular embryo stage, a heart stage, a torpedo stage, a bent-cotyledon stage to the mature embryo (Li et al., 2012b; Le et al., 2010). Embryogenesis consists of two phases; morphogenesis and seed filling, as the seeds are non-endospermic.

Seed development goes through a complex network of many dynamic developmental, biochemical and metabolic processes such as cell division and differentiation, carbohydrate, protein, cell wall, lipid, amino acid, hormone and secondary metabolite biosynthesis (Baud et al., 2002). Several hundreds of genes are reported to be involved in spatial and temporal regulation of these metabolic processes. A systematic overview of metabolic processes and gene expression patterns during seed development has been well documented for the closely related model plant *Arabidopsis thaliana* (Baud et al., 2002; Girke et al., 2000; Peng and Weselake, 2011). In *B. napus*, transcript profiling was mainly reported in relation to oil biosynthesis and storage seed reserves (Yu et al., 2010; Jolivet et al., 2011). For oil biosynthesis, starch is synthesized at the early seed developmental stage, but after intermediate processes such as malonyl-CoA and fatty acid biosynthesis, converted into triacylglycerol (TAG), lipids and storage proteins during the seed-filling phase at a later stage of seed development in both *A. thaliana* and *B. napus* (Baud et al., 2002; Jiang et al., 2012; Niu et al., 2009). A starchless mutant contained up to 40% less lipids in mature *Arabidopsis* seed than the wild-type, while starch was undetectable (Andriotis et al., 2012). Starch turnover, breakdown of cytosolic and plastidic glycolytic pathways, malonyl-CoA and fatty acid (FA) synthesis, TAG assembly and oil body formation takes place during TAG synthesis in seed (Jiang et al., 2012). The plant hormones gibberellin, auxin, ethylene and abscisic acid (ABA) play key regulatory roles in seed development and growth (Bogatek and Gniazdowska, 2012; Xue et al., 2012) and changes in hormonal levels affect the seed size and seed number in *B. napus*, especially during the 10–20 days after pollination (DAP) period (Walton et al.,

2012). Transcription factors, for example, *ABI3* (Absciscic acid insensitive-3), *ABI4*, *ABI5*, *LEC1* (leafy cotyledon1), *LEC2* and *FUS3* (*FUSCA3*) are important regulators of the complex gene network during the process of seed development, maturation and germination (Wang et al., 2007; Santos-Mendoza et al., 2008).

Understanding the regulatory mechanisms of seed development is essential to identify the molecular basis of seed development. Transcript profiling of developing seeds has been a widely used strategy to identify functional genes and their regulatory elements for seed development that can be used as tools in breeding programs for seed quality traits. Transcriptomics provides a powerful tool and is widely used to examine the temporal and spatial changes in transcript abundance during seed development in *Arabidopsis* (Le et al., 2010; Peng and Weselake, 2011; Niu et al., 2009; Ruuska et al., 2002), *B. napus* (Yu et al., 2010; Dong et al., 2004; Beisson et al., 2003), wheat (Laudencia-Chingcuanco et al., 2007; Wan et al., 2008), maize (Lee et al., 2002; Liu et al., 2008), barley (Druka et al., 2006), rice (Xue et al., 2012; Zhu et al., 2003), soybean (Asakura et al., 2012), *Jatropha* (Jiang et al., 2012) and many other crops. So far, we are not aware of any studies connecting global gene expression profiles to seed developmental stages in the diploid *Brassica* species *B. rapa*. The release of the whole-genome sequence of *B. rapa* morphotype Chinese cabbage var. Chiifu (Wang et al., 2011a) facilitates genomic studies, such as gene expression analysis and genetical genomics studies (Jansen and Nap; 2001). The knowledge on changes in gene expression associated with specific stages of seed development is crucial to unravel the molecular and biochemical events that influence optimal seed metabolite composition (Hu et al., 2009). Timing of major transition stages differs between metabolic pathways (carbohydrates, fatty acids, storage proteins) and also between species. The higher number of differentially expressed sequence tags (ESTs) at 15 DAP than at 25 DAP in *B. napus* suggest that most developmental changes take place at 10–20 DAP (Dong et al., 2004). Major changes in gene expression profiles of genes involved in protein translation, starch metabolism and hormonal regulation were reported between 17–21 DAP in *B. napus*, whereas fatty acid synthesis related genes were highly expressed at 21 DAP as compared to earlier and later time points (Niu et al., 2009). In developing *B. napus* spring cultivar seeds, 20 DAP was the most active stage to measure variation in transcript abundance of genes related to the biosynthesis of starch, lipids, carotenoids, isoprenoids, proteins and storage reserves (Yu et al., 2010).

Recently, genetical genomics has become a powerful tool to find candidate genes for complex traits (Jansen and Nap, 2001; Gaffney et al., 2012), such as seed quality and seedling vigour traits (Jordan et al., 2007). In this approach variation in transcript abundance is considered as quantitative traits in quantitative trait loci (QTL) analyses per gene, resulting in identification of genomic regions regulating gene expression (called expression quantitative trait loci: eQTL). It is important to find an optimum stage during seed development for eQTL mapping studies, where large numbers of genes show differences in transcript abundance between genotypes in a segregating population. To obtain a comprehensive insight into transcriptional

changes during seed development in *B. rapa*, we carried out morphological characterization and global transcriptome analysis in a time range of developing seeds of a black/brown-seeded pak choi vegetable-type (PC175), a yellow-seeded oil-type yellow sarson (YS143) and both a yellow and a black/brown-seeded doubled haploid (DH) progeny line from their cross. In this study, we first describe embryo and seed morphological changes in time. Second, the differential expression profiles of genes from different metabolic pathways and transcription factors in developing seeds of the four genotypes are presented. Third, a window around the optimum seed development stage was defined based on genotypic and developmental transcriptomic profiles for more extended gene expression studies. Fourth, we investigated the regulation of lipid metabolism in more detail. Using a comparative analysis of gene expression networks among these four different genotypes, we explore the differential gene expression profiles and conserved regulatory mechanisms for seed development across these morphotypes of the diploid crop species *B. rapa*.

Material and methods

Plant materials and monitoring seed development

For this study two different *B. rapa* morphotypes were used; an oil-type yellow sarson (YS143) and a vegetable-type pak choi (PC175), as well as two DH lines (DH42 and DH78) from a cross of parental genotypes YS143 and PC175. These two parental morphotypes were selected based on their genetic distance, different plant phenology, flowering time and metabolite content in the seed (Supplementary Table S6). The two progeny DH lines, which also differ in morphological characteristics such as seed colour, flowering time and metabolite content were also included in this study (Supplementary Table S6). Three plants of parental genotypes and a single plant of each DH line was grown in a heated greenhouse under 16/8 hours light/dark from February to June, 2010 at Wageningen UR. Flowers were tagged the day they opened, assuming self-pollination on the day of flower opening. PC175 and other self-incompatible DH lines of the population were manually bud pollinated to get enough seed. For each genotype, siliques were harvested at 15 time points: 10, 15, 16, 17, 18, 20, 21, 25, 30, 35, 40, 45, 50, 55 and 60 DAP. About 100–150 seeds were excised from the seed pods, frozen in liquid nitrogen and used for RNA isolation. Randomly five seeds from each genotype at each time point (developmental stage) were dissected under the binocular stereo microscope at 1.6x magnification and pictures were taken using Axio Vision Rel. 4.8 software (Carl Zeiss Imaging Solutions, Wrek, Göttingen, Germany) to observe the morphological characteristics of embryos and seeds at each time point.

RNA isolation

Siliques harvested at defined stages were kept in liquid nitrogen (–196°C), and around 100–150 seeds were extracted under dry ice and ground in liquid nitrogen (–196°C). For real-time PCR, RNA was isolated using KingFisher Flex system (Thermo Scientific, Finland) and Ambion's MagMAXTM-96 Total RNA isolation kit according to the manufacturer's instruction

and RNA pellets were dissolved in nuclease-free water. For microarray, RNA isolation was done using Trizol reagent according to the manufacturer's instructions (Invitrogen, Burlington, ON, Canada) followed by DNase treatment (AmpGrade I, Invitrogen, Burlington, ON, Canada) and a purification step (RNeasy Mini Kit, Qiagen). The quantity of RNA was determined by NanoDrop ND-100 UV–VIS spectrophotometer and quality was assessed by A260/A280 and A260/A230 ratio (NanoDrop Technologies, Inc., Wilmington, DE, USA) as well as by 1% agarose gel.

Quantitative real-time PCR (qRT-PCR)

Ten genes involved in major metabolic processes of seed development according to the literature were selected to measure transcript abundance across seed development stages ranging from 10 to 60 DAP using real time-PCR (Supplementary Table S7). These candidate genes represent fatty acid biosynthesis (*DGAT1*, *DGAT2* and *FAE1*), carbohydrate metabolism (*GBSSI* and *SuSy3*), storage proteins (*12S-CRA1* and *LEA*), transcription factors (*LEC1* and *Glabra2*) and one CHD3-chromatine-remodeling factor (*PICKLE*). The detailed procedure of qRT-PCR and normalization is described in Supplementary Methods S1. The normalized transcript abundance ($\Delta\Delta CT$) of each gene for each sample was determined with respect to the reference gene β -actin. We use the term gene expression for this normalized transcript abundance in this paper. In order to identify common profiles of transcript abundance across the seed development stages, genes were grouped using hierarchical cluster analysis with Euclidean distance of normalized data ($\Delta\Delta CT$). Transcript abundance of ten genes obtained from real-time PCR were visualized using a heatmap tool in Supplementary Figure S1.

Microarray probe design

The whole genome sequence of *B. rapa* cv. Chiifu (a leafy vegetable inbred line) is publicly available (Wang et al., 2011a). We designed microarray probes for two-colour Agilent microarray platform based on the predicted gene models of the reference genome sequence. In this custom array, 61,654 probes were assembled, which represent 40,879 (99.74%) *B. rapa* gene IDs (Bra ID) and 108 (0.26%) scaffold IDs with no assignment of Bra ID (Supplementary Table S1). All the probes were annotated into 35 different functional categories or “BINS” as defined by MapMan software (Supplementary Methods S1). MapMan is an open source software tool to categorize and display functional genomics data (Usadel et al., 2005).

Experimental design for microarray hybridization

Microarray hybridization was done on developing seeds from four genotypes; the two parents (YS143 and PC175) and two DH lines (DH42 and DH78) at six time points: 18, 20, 25, 30, 35 and 40 DAP. Two independent experiments were done to compare two parental genotypes (hereafter, called experiment A) and two DH lines (hereafter, called experiment

B). Cy3 and Cy5 dyes were incorporated into cRNA samples according to the Agilent two-colour microarray based gene expression analysis (Low input quick Amp labelling G4140-90050) protocol (Agilent Technologies, Inc., Santa Clara, CA, USA) and hybridized on arrays following a double-loop design (Supplementary Figure S9A-B). In one array, two samples from the two consecutive time points of the same genotype or two genotypes from the same time point were hybridized. The same hybridization scheme was used for experiment B using the two DH lines. In both experiments A and B, each sample was hybridized four times generating four technical replicates. Loess was used for within-array normalization and quantile normalization for between-array normalization using the limma package in R (Smyth, 2005). The normalized Cy3 and Cy5 intensities were used as measures of transcript abundance and are sometimes referred to as gene expression in this paper.

Microarray data analysis

The aim of this study was to explore the effects of seed developmental stages, genotypic variation or both on transcript abundance of genes with special focus on important metabolic processes. Principal components analysis (PCA) was used to examine the global profiles of transcript abundance of the four *B. rapa* genotypes across six seed developmental stages.

For further analyses, we excluded probes with little variation in transcript abundance across seed development as well as between genotypes using a minimum two-fold change threshold (in absolute value). Fold change differences were calculated in contrasts between two consecutive time points (18 vs. 20, 20 vs. 25, 25 vs. 30 and 35 vs. 40 DAP) as well as between two pairs of genotypes (YS143 vs. PC175 and DH42 vs. DH78) per time point. In this study, we emphasized the metabolic processes that have either a high number of selected probes or apparent changes in the number of selected probes among time point or genotype contrasts for further analysis.

WGCNA is a widely used correlation-based network construction method to construct a scale-free network (Horvath and Dong, 2008). A signed WGCNA approach was applied in this study to find gene co-expression modules, so-called “gene modules” while keeping track of positive or negative correlation coefficients, where each gene module represents a group of genes having similar co-expression patterns across seed developmental stages or genotypes or their combinations. WGCNA first calculates Pearson’s correlation matrix of all genes, and transforms the correlation matrix into an adjacency matrix by raising all values to a soft threshold power β (default value 12) to emphasize strong correlations and penalize weaker correlations on an exponential scale. Then, the adjacency matrix is transformed into a topological overlap matrix (TOM), which summarizes the degree of shared connections between any two genes, and then converted into a dissimilarity matrix. A hierarchical cluster of genes is created based on a dissimilarity matrix and finally, gene co-expression modules were defined from the cluster dendrogram at a threshold of 0.2 dissimilarity value using the dynamic tree-cutting algorithm. Once gene modules were identified, the “Module

Eigengene" (ME; the first principal component of the expression values across subjects) was calculated using all probes in each gene module. The module eigengene represents the expression profiles of all probes from a gene module across subjects (i.e. genotypes at each time point), and high or low eigengene values of subjects correspond to over- or under expression in the corresponding subjects, respectively. The details of this method are described in Horvath and Dong (2008) and Mason et al., (2009), and the analysis was performed in R software using the WGCNA package (Langfelder and Horvath, 2008). The module eigengene of each subject was examined to determine the effects of time or genotype or both using an ANOVA test. In this case, genotype and time were two independent factors and a module's eigengene values as the response, consecutively for each module. The significance of the effects was determined at 0.001 FDR correction proposed by Benjamini and Hochberg (1995). The probes belonging to gene modules significant in ANOVA were grouped into three categories according to genotype or time or both genotype and time effect. Hierarchical clustering using Euclidean distance as a criterion for dissimilarity then was applied independently on the data sets of these three categories. From this hierarchical clustering, genes were broadly organized into clusters considering the height of the dendrogram, and each category was annotated with MapMan metabolic pathways. Fisher's exact test was used to test for over- and under-representation of metabolic pathways in a selected cluster of genes using R software. If a particular pathway was significantly over- or under-represented in the gene cluster that indicates a statistically significant number of probes from the pathway are present in the gene clusters with specific patterns of gene expression across seed development stages over four genotypes (Merico et al., 2010).

Motif analysis

We focused on discovering transcription factor binding sites or DNA motifs for the co-expressed genes of lipid metabolism. The 1000 bp upstream sequences of co-expressed *Brassica* genes from the transcription start site (TSS) were retrieved from *Brassica* database (<http://brassicadb.org/brad/>). Conserved DNA motifs were searched in the upstream regions using the expectation maximization algorithm implemented in MEME version 4.9.0 (Bailey et al., 2009). Motifs with 6–12 nucleotides length were searched on both strands of the input sequence using both "zero or one occurrence per sequence" and "any number of repetitions" options. Motifs with an E-value ≤ 1 were used to assess similarity to known motifs using TOMTOM (Gupta et al., 2007) in the JASPAR plant specific database (Portales-Casamar et al., 2010). This plant specific JASPAR database was considered because of the potential roles of these motifs in regulating lipid metabolism during seed development in higher plants.

Results

Morphology of developing seeds and embryos

Morphological changes in developing embryos and seeds were monitored from 10 DAP until 60 DAP. The images show that seed and embryo structure were visible at 10 and 15 DAP, respectively (Figure 1). The colour of the embryo in YS143 was green already at 15 DAP (torpedo stage) and changed from green to yellow at 55 DAP, while in PC175 the embryo turned green only around 25 DAP (bent-cotyledon) and changed from green to yellow at 40 DAP (embryo fully fills seed) (Figure 1). In the case of the seed coat, the colour gradually turned from pale yellow to greenish or green until 40 DAP, then turned to brown or black in pak choi; however, for YS143 the seed coat colour changed from green to yellow from 55 DAP (Figure 1). Different embryo developmental stages could be defined in time, such as: pre-globular, globular, heart shape (<15 DAP), torpedo (15–18 DAP), bent-cotyledon (20–30 DAP), embryo filling seed completely (30–40 DAP) (Figure 1).

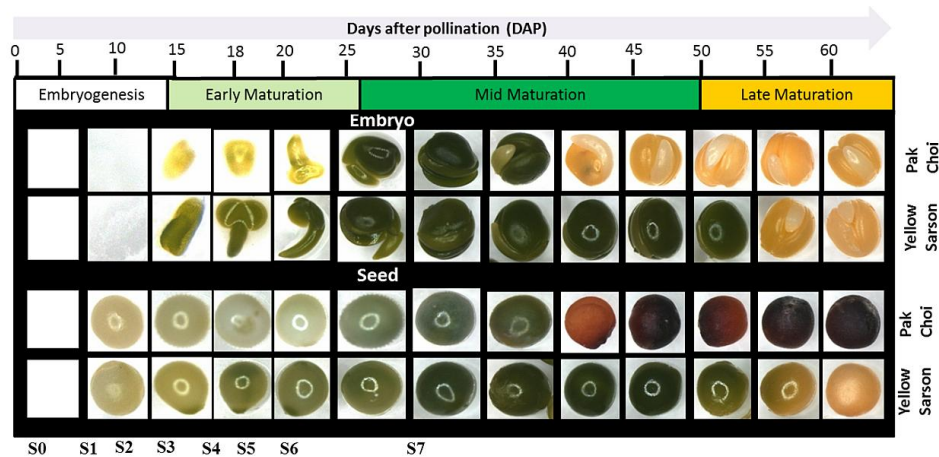


Figure 1: Morphological characterization of embryos and developing seeds of yellow-seeded oil-type genotype yellow sarson (YS143), and black/brown-seeded vegetable-type genotype pak choi (PC175) of *Brassica rapa*. Developing seeds represent different developmental stages after fertilization to seed maturity. Seed developmental stages are: S0: fertilization; S1: pre-globular; S2: globular; S3: heart; S4: torpedo; S5: linear cotyledon; S6: bent-cotyledon; S7: embryo fully fills seed.

Real-time gene expression profiling in developing seeds

The transcript abundance of 10 selected genes from a few key metabolic processes and transcription factors (Supplementary Figure S1) was measured from 10 to 60 DAP to obtain an overview of gene expression patterns during seed development. Three patterns were observed, with peak levels at 10–25 DAP, 25–40 DAP and 35–60 DAP which are defined as early-, mid- and later- stage, respectively (Supplementary Figure S1). These patterns were not very different among the four genotypes tested. Out of the ten genes, transcription factors *LEC1* and *Glabra2* and the starch gene *GBSSI* were expressed higher during earlier stages, lipid metabolism genes *DGAT2* and *FAE1*, and storage protein *12S-CRA1* were expressed higher at mid stages, while the lipid metabolism gene *DGAT1* (also called *TAG1*),

carbohydrate metabolism gene *SUS3*, and the storage protein *LEA* and CHD3-chromatine-remodeling factor *PICKLE* were expressed highest at late stages. For the whole genome microarray gene expression profiling, six time points were selected: (i) 18 DAP (torpedo), (ii) 20 DAP (bent-cotyledon), (iii) 25 DAP (transition bent-embryo fully fills seed), and the developmental stages where the embryo fully fills the seed, being (iv) 30 DAP (v) 35 DAP and (vi) 40 DAP. These time points captured transcriptional changes at early, mid and late stages of seed development.

Microarray hybridization and probe annotation

In a dedicated *B. rapa* Agilent array, 61,546 probes (99.7% of total 61,654 probes) represent 42,162 *Brassica rapa* gene ID (called Bra ID). Out of 42,162 Bra IDs, 30,363 Bra IDs (72%) were assigned to 34 MapMan functional annotation categories. The remaining 11,799 (28%) Bra IDs were not assigned to any functional category (Supplementary Table S1).

Pearson correlation coefficients were calculated to quantify how similar transcript abundance was between time-points and also between four replicates in each genotype (YS143, PC175, DH42 and DH78). All the replicates of each genotype from each time point had high correlations ($r > 0.95$) in all four genotypes (Supplementary Figure S2A-B). The correlation coefficients between time points decrease as the time points increase. Pearson correlation coefficients of transcript abundance between time-points were high ($r > 0.9$) from 18 to 25 DAP in PC175, and from 18 to 30 DAP in YS143, DH42 and DH78, but after those time points a transition from high ($r > 0.95$) to lower ($r < 0.85$) correlation coefficients occurs between early and later time points.

Correlation of transcript abundance of genes from real-time PCR and microarray analysis

Since transcript abundance was measured using two different techniques: qRT-PCR and microarray that might lead to a non-linear relationship, Spearman's rank correlation coefficients, which are free from parametric assumptions, were used to compare the outcome of these two techniques. The transcript abundance from qRT-PCR and microarray of 10 selected genes were significantly and positively correlated except for transcription factors *LEC1* and CHD3-chromatine-remodeling factor *PICKLE*. The rank correlation coefficients ranged from 0.43 for *DGAT2* to 0.94 for *LEA* protein (Supplementary Table S2).

Genome-wide variation in transcript abundance during seed development

Principal components analysis (PCA) on transcript abundance of the 61,654 probes showed a sequential distribution of the six time points according to seed developmental stages along the first principal component (PC1) and separation of the four *B. rapa* genotypes along PC2. PC1 explained 38.8% of total variation, and is associated mostly with variation in transcript abundance over the developmental stages, where 18 DAP and 20 DAP form a tight group, with 25 DAP more loosely grouped with these earlier stages. Similarly, 35 DAP and 40 DAP were grouped together (but distinct from the earlier time points) except in PC175. Major

changes in transcript abundance were observed between 25 and 35 DAP (Figure 2), which coincides with the period of transition from bent-cotyledon to the stage when the embryo fully fills the seed. PC2 explained 15.6% of the total variation, and reflects mostly genotypic differences. Interestingly, the two DH lines were grouped in between the two parental genotypes (Figure 2).

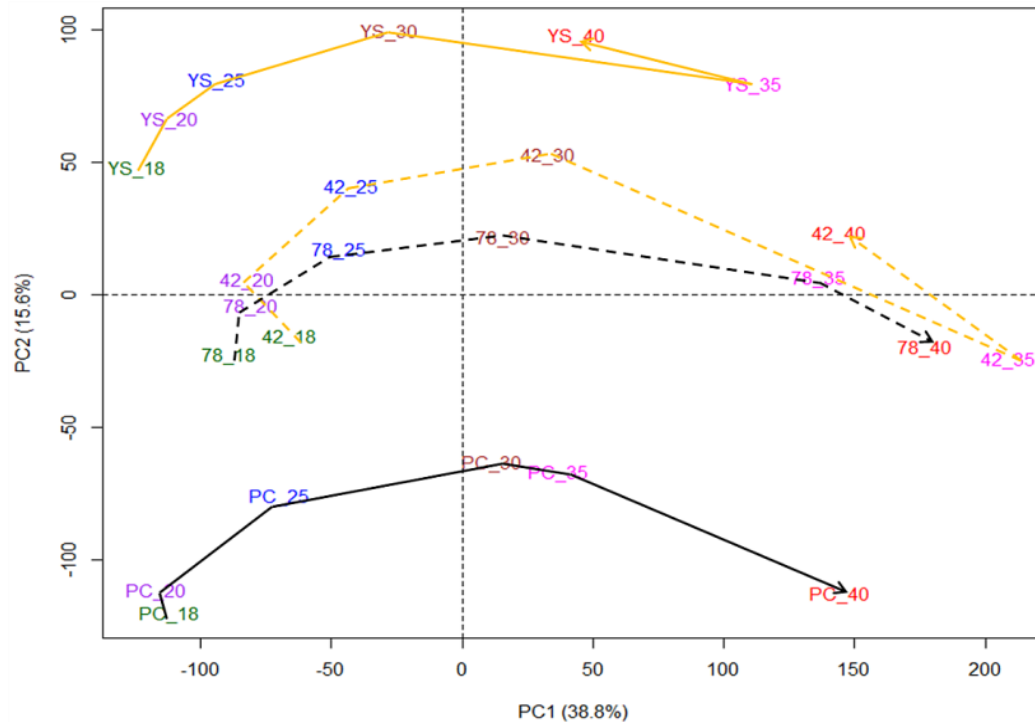


Figure 2: Principal components analysis (PCA) of two parental genotypes (YS143 and PC175) and two DH lines (DH42 and DH78) based on transcriptional profiles during seed development (18–40 DAP). Sample names are a combination of genotypes (YS = yellow sarson, PC = pak choi, 42 = DH line 42 and 78 = DH line 78) and time points in days after pollination (DAP). The yellow lines represent yellow-seeded genotypes YS143 and DH42, and black lines represent black/brown-seeded genotype PC175 and DH78. Parental genotypes are indicated with solid lines, and DH lines with dashed lines. Sample labels were coloured according to time points: 18 DAP - green, 20 DAP - purple, 25 DAP - blue, 30 DAP - brown, 35 DAP - pink and 40 DAP - red.

We investigated the loading values of probes on PC1, where probes with very low negative loadings were associated with the early stage of seed development (18–25 DAP) while the probes with very high positive loading were in response to later stages (35–40 DAP) (**Figure 2**). Among 34 MapMan functional categories, probes with high positive or low negative loadings mainly belong to metabolic pathways such as photosynthesis, cell wall metabolism, lipid metabolism, amino acid metabolism, protein metabolism, signalling, RNA (RNA processing, RNA binding and transcription factors), stress, transport, developmental processes, hormone metabolism, phosphate metabolism and secondary metabolism (Supplementary Figure S3).

Apparent changes in numbers of selected probes in contrasts between developmental stages or genotypes lead to selection of metabolic pathways

After excluding probes with rather constant transcript levels (< 2 -fold change) across seed development and between genotypes, 11,244 probes (18.2% of total 61,554 probes) were retained for further analysis (Supplementary Table S3). Based on either a high number of selected probes per pathway or apparent changes in the number of selected probes from contrasts between consecutive time points or between genotypes at each time point, the top thirteen metabolic pathways were emphasized in this study. These top thirteen metabolic pathways correspond to metabolic pathways highlighted based on higher PC1 and PC2 loadings in PCA analysis. Those top thirteen metabolic pathways are represented by 9606 probes (i.e. 5520 Bra ID) and used for network analysis to separate the gene clusters according to temporal (4178 probes) and/or genotypic variation (3169 probes) during seed development (Supplementary Table S4).

Signed weighted gene co-expression network analysis (WGCNA) identifies gene modules associated with temporal and or genotype effects

Signed WGCNA grouped the selected probes (> 2 fold-change) into 47 co-expression gene modules, each one containing probes with a similar transcript abundance across genotypes and seed developmental stages. In an analysis of variance (ANOVA) test, 17 gene modules (3169 probes) showed a genotype effect, 4 modules (4179 probes) a time effect, and 6 modules (555 probes) a genotype as well as a time effect at 0.001 significance level and the remaining 20 gene modules did not show any effect (Supplementary Table S5; Supplementary Figure S4A-C). Since some of the gene modules showed similar expression patterns with subtle differences, gene modules were combined according to the time or genotype or time and genotype effects, and subjected to hierarchical clustering to have a broader overview of the patterns of transcript abundance.

Temporal variation across seed development stages

Using hierarchical clustering, 4179 probes from the four gene modules (associated with differential expression in time) were classified into three clusters (Figure 3A-B). Cluster I (2043 probes corresponding to 1525 genes) represents genes with higher transcript levels at earlier stages (18–25 DAP) from linear cotyledon to bent-cotyledon. Both cluster II (837 probes or 655 genes) and cluster III (1298 probes or 977 genes) show increased transcript abundance in time, from 18 DAP for cluster II and from later stages after the embryo fills the seed at 30 DAP for cluster III (Figure 3A-B).

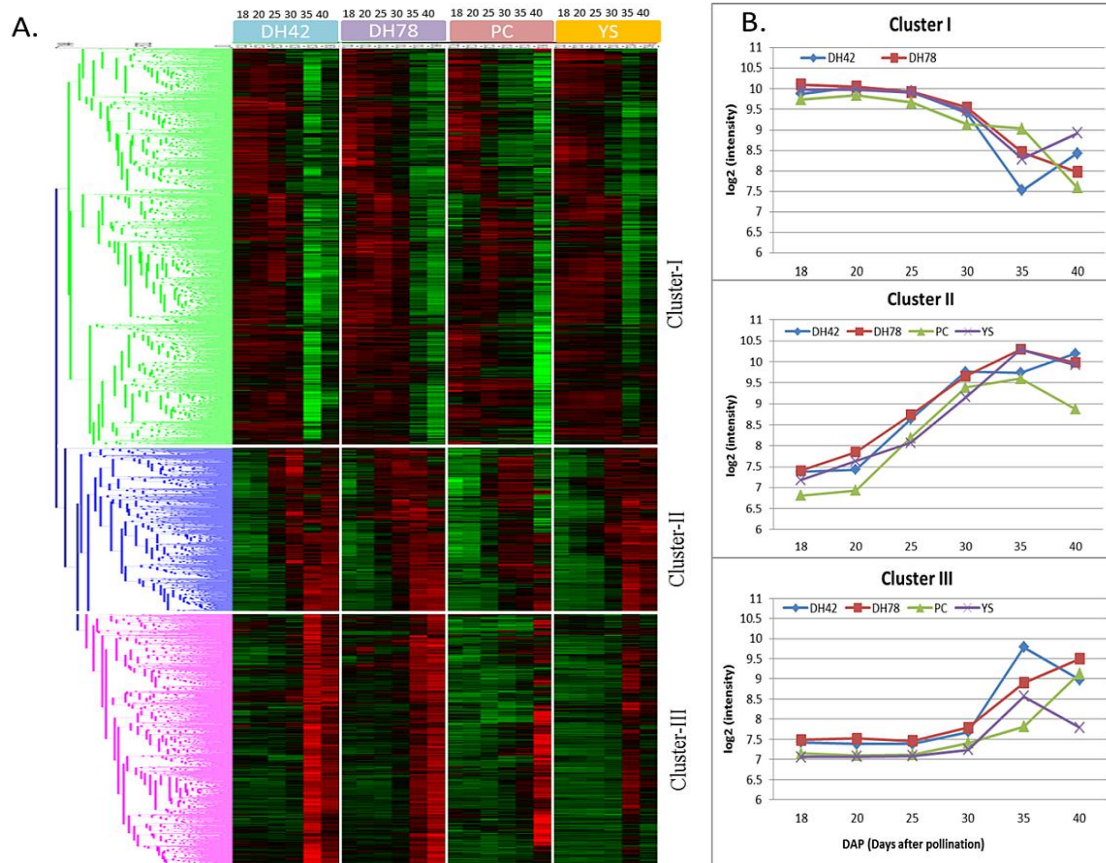


Figure 3: Temporal patterns of transcript abundance during seed development stages (18-40 DAP). **A.** Hierarchical cluster analysis using Euclidean distance and average linkage of all probes belonging to four WGCNA gene modules having a significant effect of developmental stages. Vertical white bars separate genotype and horizontal bars separate gene clusters. Red colour indicates a higher level of transcript abundance, green colour lower abundance and black an intermediate level. Colours of dendrogram branches indicate different gene clusters. **B.** Line graph that shows the expression level (\log_2 scale) of probes that belong to three clusters. The x-axis represents seed development time points (18, 20, 25, 30, 35 and 40 DAP: Days after pollination).

Genes associated with photosynthesis (Calvin cycle and photosystem-I and -II), Fatty acid (FA) synthesis, FA elongation and lipid degradation are over-represented only in cluster I, so, these genes are active early in seed development and down-regulated later (Figure 4; Supplementary Figure S5). Also genes from tocopherol biosynthesis, mevalonate and carotenoids in secondary metabolism, as well as from biosynthesis of serine, glycine, cysteine, glutamate, aspartate and alanine amino acids were only over-represented in cluster I. Transcription factors (TFs) were mostly under-represented in this cluster I. For example, *AP2/EREBP*, *bHLH*, *C2H2*, *Myb*, and *WRKY* TFs were under-represented in cluster I, and *bZIP* was overrepresented in cluster II. Genes involved in cell wall metabolism including precursor synthesis, cellulose synthesis, cell wall proteins and cell wall degradation, and genes in triacylglycerol synthesis (TAG) and FA desaturation were mainly over-represented in cluster II, which means that they continuously increase in abundance from 18 DAP till 35

DAP. Also, storage protein genes and genes related to the biosynthesis of auxin, brassinosteroid and gibberellin and branched-chain and aromatic amino acids (Supplementary Figure S5) were over-represented only in this cluster II. Similarly, metabolite transporter genes and major intrinsic protein genes from transport metabolism were mainly over-represented in both clusters I and II, but receptor kinases and G-proteins genes from signaling pathway, and genes involved in protein synthesis, protein posttranslational modification, protein degradation, RNA processing and RNA binding were under-represented in cluster I and or II. Genes related to cytochrome P450 and seed storage (lipid transfer protein, LTP) of phosphate metabolism, late embryogenesis abundant (*LEA*) proteins, and ethylene and abscisic acid from hormonal metabolism were over-represented in cluster II and or III, so their abundance increased during seed development. Biotic stress tolerance genes related to PR-proteins were underrepresented in cluster II and III, but genes related to heat shock proteins for abiotic stress tolerance were overrepresented in cluster III. Interestingly, cluster II and III had high transcript abundance during late stages of seed development with different patterns.

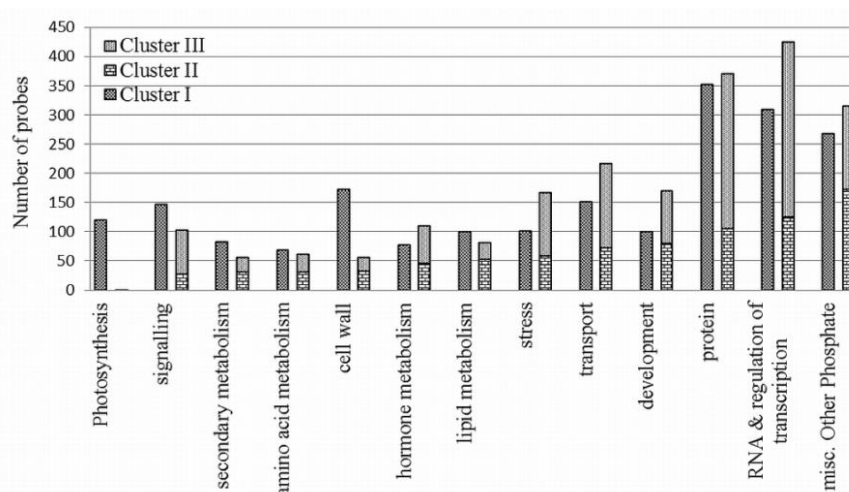


Figure 4: Comparison of numbers of probes belonging to MapMan functional categories in three clusters (Cluster I, cluster II and cluster III) showing temporal variation of transcript abundance. Fisher's exact test was carried out for over-representation against total numbers of probes annotated in each functional category. The significance level was determined at 0.01 p-value after FDR correction with the method of Benjamini-Hochberg (1995).

Putative cis-regulatory elements underlying co-expressed genes of lipid metabolism

We looked in more detail to changes in transcript abundance related to lipid metabolism because oil is the major storage compound of *Brassica* seeds. *B. rapa* and *B. napus* are widely grown for oil production, while *B. rapa* is also grown as vegetable crop. Therefore, it is interesting to know the variation in transcript abundance of genes related to oil biosynthesis during seed development in oil-type and non-oil type morphotypes. For this study, two genotypes: a yellow-seeded oil-type genotype YS143 and a black/brown-seeded vegetable-type genotype PC175 were chosen. In addition, two DH progeny lines, a yellow-seeded and a black/brown-seeded line, resembling the two parental lines were also used to

develop ideas on segregation of transcript abundance of oil biosynthesis genes. In Supplementary Figure S6, the pathway for oil biosynthesis is depicted, with acetyl-CoA as the main precursor for the synthesis of fatty acids (FA), triacylglycerol (TAG) and phospholipids. Transcript abundance was visualized separately for genes involved in FA synthesis, FA elongation, lipid degradation, FA desaturation, biosynthesis of TAG and phospholipids, and oleosin (oil bodies).

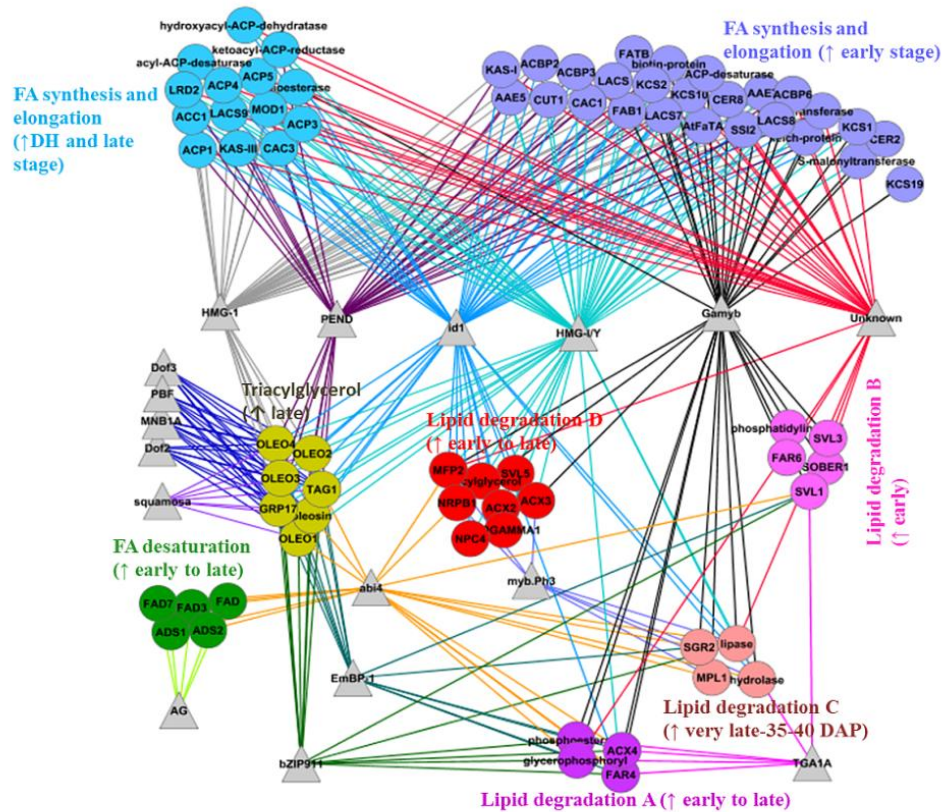


Figure 5: Graph showing motifs (TFBS: transcription factor binding sites) identified in sets of co-expressed genes from different metabolic processes of lipid metabolism. The elliptic shaped node represents genes, the triangular node represents conserved motifs, the edge between motif and gene represents the presence of a motif in a particular gene. The colour of nodes indicates co-expressed genes from different metabolic processes of lipid metabolism while the same colour of the edges indicates genes have same motif. An arrow-up symbol indicates high transcript abundance of a gene.

In the process of FA synthesis and elongation, transcript abundance of genes revealed patterns with either a clear temporal effect or with a clear genotype effect (Supplementary Figure S7A:I-II). Transcript abundance of 63% of FA synthesis and FA elongation related probes was high at early stages (18–30 DAP), followed by a gradual decrease, while other probes (37%) show clear genotype differences with higher transcript abundance in the two progeny lines (DH42 and DH78) as compared to the parental genotypes. FA desaturation genes such as ADS1, FAD6 and FAD7 were up-regulated before 30 DAP, but FAD3 and ADS2 genes including FAD6 and FAD7 paralogs were up-regulated after 25 DAP (Supplementary

Figure S7B). Triacylglycerides are the main constituents of vegetable oil and expressed at late stages of seed development. Genes involved in triacylglycerol biosynthesis, such as *DGAT-1* and *-2*, *GRP* (glycine rich protein) and oleosin (storage proteins) were mainly up-regulated after 25 DAP (Supplementary Figure S7D).







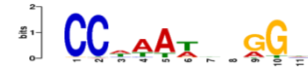


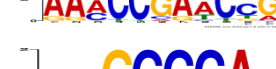






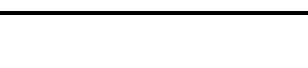
For lipid degradation, four different patterns of transcript abundance were observed. A set of probes (18.5%) had high transcript abundance at later stages of seed development (after 25 DAP) (Supplementary Figure S7C: I-IV), while a larger number of probes (40.7%) showed higher transcript abundance at earlier stages before 30 DAP (Supplementary Figure S7C: II). Supplementary Figure S7C: III consists of a set of probes (13%) with high transcript abundance only at 35 and 40 DAP. Probes (27.8%) from Supplementary Figure S7C: IV showed genotype differences in transcript abundance with lower levels in parental genotypes PC175 and YS143, than in the DH lines.

A set of the genes functionally related and/or co-expressed often share conserved regulatory motifs, which might be responsible for coordinated expression of the set of genes. In this study, genes related to lipid metabolism with different co-expression patterns (different clusters) were searched to computationally predict *cis*-acting regulatory elements for potential roles in regulating lipid metabolism during seed development in *B. rapa* species. For all the selected 194 *B. rapa* genes (> absolute 2-fold change), the 1000 bp upstream sequence from the gene start were retrieved.

In total, 17 regulatory motifs were predicted for FA synthesis and elongation (92 genes), lipid degradation (74 genes), lipid desaturation (12 genes) and triacylglycerol (16 genes) processes considering gene clusters with comparable patterns in transcript abundance (Table 1; Figure 5). Co-expressed gene clusters from the FA synthesis and elongation, and lipid degradation, and/or other lipid metabolic processes shared most of the motifs (Figure 5). Each TF (transcription factor) can have more than one putative binding site in each gene. The *DOF* motif family, including *DOF2*, *DOF3*, *PBF* and *MNB1A*, and MADS motif-squamosa were specific to the TAG biosynthesis process but another MADS motif – AG was specific to FA desaturation. TGA1A (leucine zipper family) and myb.Ph3 (*Myb* family) were shared among different co-expression groups of lipid degradation genes. The *ABI4* transcription factor binding site was present in genes involved in TAG biosynthesis, FA desaturation and different co-expression groups of lipid degradation, which had high transcript abundance at late stages (after 25 DAP) (Figure 5; Supplementary Figure S7A-D). We did not find any motif that is specific to the FA synthesis and elongation process. However, six motifs; *HMG-1*, *HMG-I/Y*, *PEND*, *id1*, *Gamyb* and four unknown motifs were shared between two co-expression groups of FA synthesis and elongation genes along with genes from other processes. Conserved motifs that were not significantly overrepresented in plant-specific TFs databases are here indicated as “unknown”. Motifs such as, *HMG-1* and *PEND* were specific to only genes involved in TAG biosynthesis, and FA synthesis and elongation process. Similarly, *Gamyb* (*Myb*-family) and unknown motifs were specific to only lipid degradation and FA synthesis and elongation process. Motifs- *bZIP911* and *EmBP-1* from the leucine

zipper family were shared among genes from TAG biosynthesis and lipid degradation (Table 1; Figure 5).

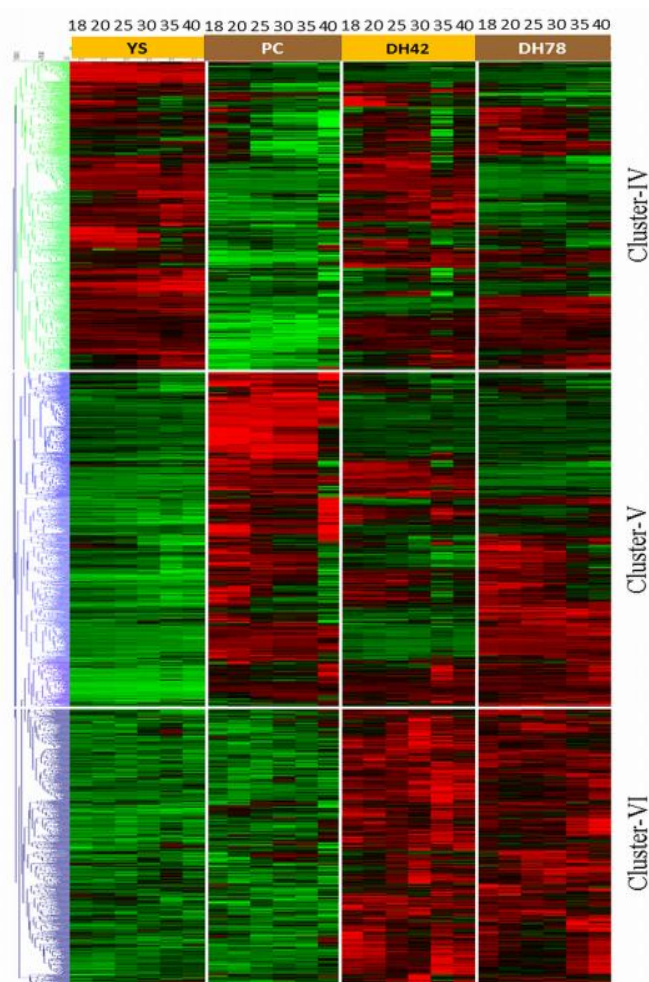
Table 1: List of overrepresented motifs identified in promoter regions (1000bp upstream) of genes involved in FA synthesis and elongation, FA desaturation, FA degradation and triacylglycerol (TAG) synthesis.

Sequence logo	Matrix ID	TFBS name	TFBS family
	MA0123.1	abi4	AP2 MBD-like
	MA0021.1	Dof2, Dof3, MNB1A, PBF	Dof
	MA0097.1	bZIP911	Leucine Zipper
	MA0129.1	TGA1A	Leucine Zipper
	MA0128.1	EmBP-1	Leucine Zipper
	MA0127.1	PEND	Leucine Zipper
	MA0005.1	AG	MADS
	MA0082.1	Squamosa	MADS
	Unknown	Unknown	Unknown
	Unknown	Unknown	Unknown
	Unknown	Unknown	Unknown
	Unknown	Unknown	Unknown
	MA0054.1	myb.Ph3	MYB
	MA0034.1	Gamyb	MYB
	MA0045.1	HMG-I/Y	High mobility group
	MA0044.1	HMG-1	High mobility group
	MA0120.1	id1	Zinc finger

Genotypic variation in overall metabolism

In total 17 modules (3169 probes) were divided into three clusters (cluster IV to VI) in hierarchical cluster based on clear contrasts in patterns of transcript abundance only between the two parental genotypes. Probes from cluster IV (1054 probes, 851 genes) were up-regulated in YS143 and down-regulated in PC175, while probes in cluster V (1149 probes, 951 genes) had higher transcript abundance in PC175 and lower in YS143 (Figure 6). These two clusters differentiate the transcript abundance between the two parental genotypes. However, the two DH lines had a mixture of levels of transcript abundance. In contrast, genes belonging to cluster VI (966 probes, 878 genes) had low transcript abundance in both parents but high in the two progeny DH lines. Genes mainly involved in the synthesis and degradation of amino acid, cell wall, hormones, lipids, isoprenoids and ion transport, and also different transcription factors were significantly over- or under- represented in those three clusters (Supplementary Figure S8).

Figure 6: Hierarchical cluster analysis (Euclidean distance; average linkage) on all probes from four WGCNA gene modules with significant genotypic effects. Vertical white lines separate genotypes and horizontal white lines separate gene clusters. The bright red to bright green colour represent high to low abundance levels, black for an intermediate level of abundance.



Genotypic as well as temporal variation in overall metabolism

Six WGCNA modules (555 probes) showed both significant genotypic and temporal variation in ANOVA (Supplementary Table S5) and four clusters of probes with different patterns of transcript abundance were observed in a hierarchical cluster analysis (cluster VII to X; Figure 7A-H). Genes in cluster VII (80 probes, 77 genes) reached their maximum at 18 DAP, gradually decreasing until 35 DAP (YS143 and DH42) or 40 DAP (PC175 and DH78). Transcript

abundance in PC175 was always higher than in other genotypes except at 40 DAP (Figure 7A and 7E). Transcript abundance of genes from cluster VIII (73 probes, 64 genes) gradually increased until 35 DAP and then started to decrease in all genotypes. Transcript abundance in DH78 was highest while it was lowest in YS143 across all the time points. Transcript abundance in PC175 was lower than in DH42 during 18–20 DAP but increased to the level of DH78 during 25 DAP to 35 DAP (Figure 7B and 7 F). Genes in cluster IX (85 probes, 76 genes) had similar transcript abundance compared to cluster VIII with a gradual increase across the developmental stages till 35 DAP which then remained constant. However, genes from cluster IX had a lower transcript abundance in PC175 across time (Figures 7C and 7G). A larger number of probes (317 probes, 267 genes) were grouped in cluster X, which showed a maximum at the earlier stages 18–20 DAP, and then a gradual decrease until 35 DAP from which time point it remained at a constant level (Figure 7D and 7H). This transcript abundance was similar to that of cluster VII except for PC175. Among the four genotypes across all the time points, cluster X genes had the lowest transcript abundance. These clusters indicate the occurrence of major changes in the transcription profiles between the bent-cotyledon to the fully-developed embryo stages of seed development (25–35 DAP).

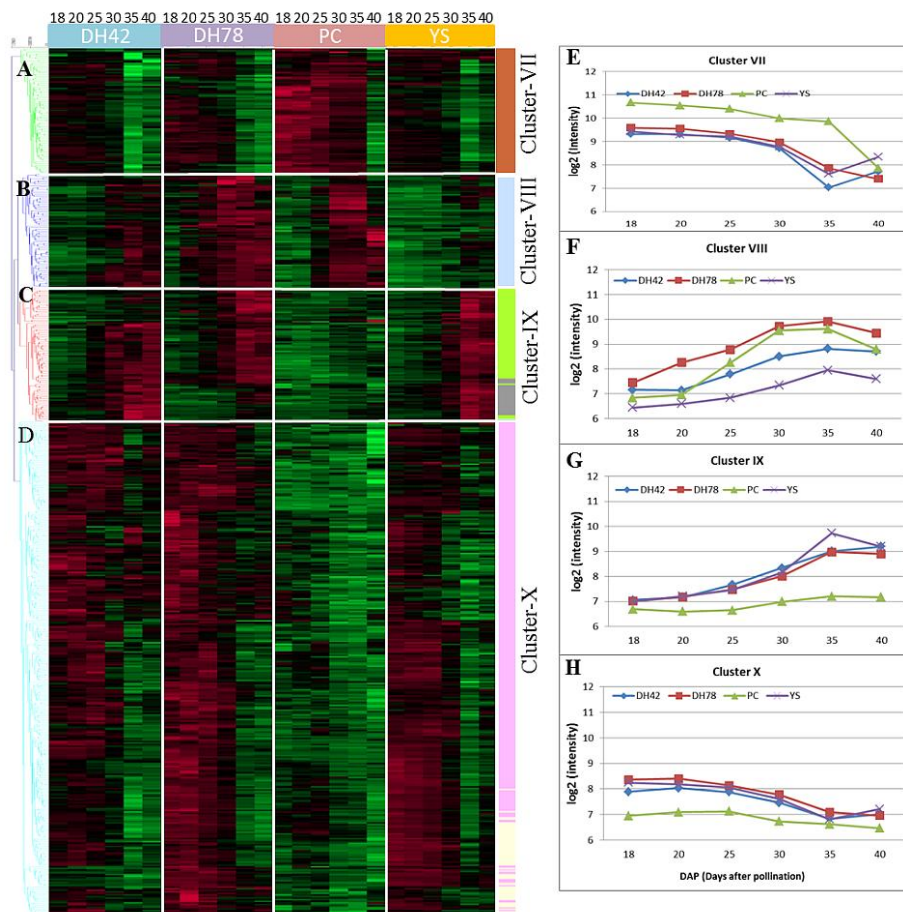


Figure 7: Characteristics of transcriptional patterns in seed development stages (18 – 40 DAP) showing genotypic and temporal variation. **A-D** Hierarchical clustering (Euclidean distance, average linkage) on 555 probes from six WGCNA gene modules with both a genotype and time effect. Vertical white bars separate genotype and horizontal white bars

separate cluster of genes. Colours of dendrogram branches indicate different clusters of genes while the colour bar on the right side indicates WGCNA gene modules. Red indicates high, green low, black an intermediate level of transcript abundance. **E-H** mean abundance of transcripts on four genotypes (YS143, PC175, DH42 and DH78) representing gene clusters **A-D** respectively.

Discussion

The understanding of morphological and transcriptional changes during seed development has fundamental applications in *Brassica* breeding, both for high quality vegetable oil content and for crop establishment. In this study, we focused on analysis of morphological characteristics and global transcriptome analysis in developing seeds of four genotypes including two diverse *B. rapa* morphotypes: a leafy-type pak choi and an annual yellow-seeded oil-type yellow sarson. We also predict putative regulatory elements for lipid metabolism to understand this complex regulatory network during seed development.

Seed morphology varies at the later stages of seed development

Seed developmental stages, which are defined based on the shape of the embryo, were similar in both YS143 and PC175, irrespective of apparent differences in phenological characteristics, such as flowering time or seed colour in the two distant morphotypes pak choi and yellow sarson (Figure 1). However, the colour of embryo differed among these two genotypes at early stages (in period 15–25 DAP, PC175 embryos are yellowish, while YS143 embryos are green); and at later stages (at 40 DAP PC175 embryo's turn from green to yellow, while YS143 embryo's turn yellow only at 55 DAP). Also, seed coat colour changes differed among these two morphotypes, as the seed coat of PC175 turns from green to brownish at 40 DAP, while the YS143 seed coat turns yellowish at 50 DAP (Figure 1). Also in the two DH lines, the black/brown-seeded line DH78 lost the green colour earlier than the yellow-seeded DH42. Yellow seed colour is a desired quality trait in breeding *Brassica* oilseed species, because of its association with higher oil content and more easily digestible seed meal as compared to dark coloured seeds. The accumulation of proanthocyanidins (PAs) in the seed coat of immature black/brown seeds (20 DAP) but not in yellow seed (Li et al., 2012b) might be an explanation for the earlier change in seed colour. In this study, we observed that the embryo completely filled the seed at the bent-cotyledon stage (30 DAP); also Li et al., (2012a) described that this stage was not yet reached at 25 DAP, but fully reached at 35 DAP in *B. campestris* (Synonymous: *B. rapa*). *Brassica* seed is non-endospermic, so, the endosperm is not retained in mature seeds, but only the embryo is enclosed by the seed coat (Sabelli, 2012). Evaluation of transcript abundance using real-time PCR was effective to define six time-points when abundance levels of a set of genes representative for the seed filling process varied with respect to their morphology: 18 DAP (torpedo), 20 DAP (bent- cotyledon), 25 DAP (transition bent-embryo fully fills seed) and 30 DAP, 35 DAP and 40 DAP (embryo fully fills the seed).

Seed developmental stages are the predominant cause for variation in transcript abundance

Genome-wide transcriptome analysis was used to explore global gene expression at six time points as representative stages for seed development in four genotypes of *B. rapa*. Despite the fact that *B. rapa* is an important vegetable and/or oil crop, this is the first study in which transcript abundance was profiled genome-wide during seed development in this species. The availability of whole genome sequence of *B. rapa* (Wang et al., 2011a) facilitated the design of a 60-mer oligonucleotide microarray platform (62,654 probes targeting 42,162 *Brassica* genes) based on predicted gene models from the genome sequence.

We used four approaches to define sets of genes with different transcript abundance during seed development in time (developmental stages) or between genotypes or both. First PCA was used to obtain an overview of variation in seed developmental stages and also between different genotypes using all the transcripts present in the microarray (Figure 2). The first principal component (PC1: 38.8% explained variance) captured mostly temporal variation in transcript abundance, supporting the earlier findings that seed developmental stages are major sources of transcriptional and metabolic variation in *Arabidopsis* (Peng and Weselake, 2011; Fait et al., 2006). A comparative study of the transcript and metabolite profiles in both wild-type and transgenic genotypes of *Arabidopsis* also showed more variation across seed developmental stages than changes due to genotypic differences (Angelovici et al., 2009). The genotypic variation was captured in PC2 (15.6% explained variance), which suggests that metabolic processes inside developing seed are largely conserved, even between yellow-seeded oil and black/brown-seeded genotypes. Secondly, we selected a subset of genes with variation in transcript abundance patterns between developmental stages as well as between genotypes based on PCA loadings with a minimum two fold change criterion for further analysis. These subsets of genes represent the most active metabolic processes occurring in *B. rapa* developing seeds, such as photosynthesis, hormonal regulation, stress tolerance, cell wall, lipid, phosphate, amino acid, protein, signal transduction, transport, secondary metabolites, developmental process, and RNA processing and regulation of transcription (Supplementary Table S4). Those selected metabolic processes were also reported as major metabolic processes during seed development in close relatives *A. thaliana* (Ruuska et al., 2002) and *B. napus* (Yu et al., 2010), but also in maize (Teoh et al., 2013). Thirdly, a WGCNA approach was used to discover possible modules consisting of groups of genes with similar transcript abundance, either across time or between genotypes of both, and 27 modules out of a total of 47 modules showed significant variation in transcript abundance across time points or genotypes or their combinations (Supplementary Table S5; Supplementary Figure S4A-C). Since WGCNA uses Pearson correlation coefficients to identify co-expressed modules, it could not group genes that have similar patterns of transcript abundance but different levels into separate modules. So, in addition a separate hierarchical clustering using Euclidean distance was done in all gene modules according to the type of effects. The combined analysis using both Pearson correlation coefficients with

WGCNA and hierarchical clustering with Euclidean distance resulted in clusters that are both similar in transcript abundance and level among genotypes across time points. Finally, we focused on transcriptional profiling related to lipid metabolism, in order to correlate co-expression patterns within pathways and to predict putative regulatory elements of lipid metabolism.

Global variation in transcript abundance: 25–35 DAP is a key period for major changes in *B. rapa* developing seed

In PCA, the early time points, before the embryo fills the seed (25 DAP), cluster tightly in PC1 but the later time points (35–40 DAP) cluster loosely, suggesting that physiological processes differentiate more at later stages. Higher correlations ($r > 0.9$) between the early time-points within genotypes and decreasing correlations between later stages also supports that there is more variation in transcript abundance at later stages (after 25 DAP) than at earlier stages (Supplementary Figure S2A-B). Variation in metabolite content, seed maturity, desiccation and dormancy induction occurred during the maturation phase (Sabelli, 2012), which corresponds to 25 DAP in this study. Interestingly, sequential changes in transcript abundance follow developmental changes in the black/brown-seeded genotypes (PC175 and DH78) but an extreme shift from 30 to 35 DAP and reversed at 40 DAP occurred in yellow-seeded genotypes (YS143 and DH42). This signifies the different transcriptome signatures of seed development in different genotypes, especially at the later stage. These findings are in agreement with a different timing of seed and embryo colour changes from 40 DAP onwards (Figure 1). The spatial position of the two DH lines between the two distant parental genotypes in the PC2 dimension points to variation in transcript abundance that can be used for genetic studies.

The largest changes in transcript abundance during seed development were observed during 25–35 DAP (bent-cotyledon to stage when embryo fully fills the seed), suggesting that this is the most optimal stage for genetical genomics studies for mapping eQTL in *B. rapa* developing seeds. In contrast, for *B. napus*, the major transcriptional transition was reported to be much earlier during heart-shaped to torpedo embryo stages i.e. 17–21 DAP, and for FA synthesis-related genes at 21 DAP in a spring and winter type *B. napus* L. cv HuYou15 (Hu et al., 2009).

Temporal changes in transcript abundance conserved across different morphotypes

The WGCNA method is a powerful and widely used tool to identify co-expressed gene clusters and to construct scale-free networks using topological properties of network construction (Horvath and Dong, 2008). Among 47 gene modules identified, four (4179 probes) show temporal variation in transcript abundance across seed development (Supplementary Table S5, Supplementary Figure S4B), and these were reduced to three clusters after hierarchical clustering using Euclidean distance (Figure 3). This result, like PCA, confirms that variation in transcript abundance during seed development is predominantly

conserved across genotypes in *B. rapa*. Similar observations were made for FA biosynthesis genes, which were conserved between *B. napus* and *A. thaliana* (Niu et al., 2009). The annotations of many genes belonging to these three clusters fitted what is known about different processes occurring during seed development. Among the three clusters, cluster I (48% genes) had high transcript abundance before 25 DAP with a gradual decrease till 35 DAP, with genes involved in photosynthesis, secondary metabolic pathways, and biosynthesis of tocopherols, mevalonate and carotenoids, and amino acids were over-represented. Amino acids are known as essential precursors for biosynthesis of secondary metabolites, proteins and other metabolic biosynthetic processes. Tocopherols are fat-soluble antioxidants and are one of the breeding goals to improve oil quality. Tocopherols accumulate slowly during 12–41 DAP and reach a maximum concentration during 41–53 DAP in developing seeds of *B. napus* (Goffman et al., 1999). It has been suggested that production of tocopherols during seed development might be needed for the protection of polyunsaturated fatty acids against peroxidation (Kamal-Eldin and Appelqvist, 1996). In cluster II (21% of genes with transcript abundance differences in time) and cluster III (31% genes) transcript abundance increased gradually or abruptly at 35–40 DAP, respectively (Figure 3). In these clusters, cytochrome P450, late embryogenesis abundant proteins (*LEA*), *LTP* (lipid transfer protein) and storage proteins, and abscisic acid and ethylene (hormone metabolism) were over-represented. This observation is in agreement with a number of other studies where storage proteins, abscisic acid and ethylene were highly expressed during late seed developmental stages because of their roles in growth and development of seed tissues, accumulation of seed reserves, maturation, desiccation tolerance, induction of seed dormancy and the utilization of storage reserves to support germination (Sabelli, 2012; Yu et al., 2010; Bogatek et al., 2012; Walton et al., 2012; Yang et al., 2011).

Gene co-expression patterns associated with genotypic differences, or genotype- and temporal differences

WGCNA analysis organized 3169 probes associated with genetic variation into 17 gene modules (3169 probes) (Supplementary Table S5; Supplementary Figure S4A), which could be represented by three gene clusters (cluster IV to VI) through hierarchical clustering (Figure 6). These clusters reveal genetic variation in patterns of transcript abundance during seed development, with distinct variation between the two parents with many genes showing transgressive segregation in DH lines.

Similarly, sets of genes (555 probes) displayed variation in transcript abundance due to both genotype and time contrasts in six gene modules (Supplementary Table S5; Supplementary Figure S4C). Four different patterns were identified in hierarchical clustering, mainly either with a gradual decrease in transcript abundance from early stages to late stages or a continuous increase across seed development (Figure 7). The leafy-type PC175 usually showed different patterns of transcript abundance compared to the other three genotypes (Figure 7A, 7C-E, 7G-H), while variation in transcript abundance of the two DH lines is more

similar to that of the maternal genotype YS143. This could be due to maternal effects on seed and seed characteristics, as reported before in another study (Andriotis et al., 2012).

Predicting cis-regulatory elements for co-expressed genes related to lipid metabolism

Brassica species are widely cultivated for seed oil, and seed oil is also a major source of energy during germination and seedling growth. Thus, we want to get an insight in the genetic regulation of lipid metabolism in both oil- and vegetable- morphotypes. First, we defined pathways, such as FA synthesis and elongation, FA desaturation, lipid degradation, triacylglycerol. The co-expression analysis identified clusters of genes in the respective pathways with different transcript abundance. For example, FA synthesis and elongation related genes shared a similar time-dependent (high at 18–25 DAP, decrease thereafter) and a genotype-dependent transcript abundance (Supplementary Figure S7A). Lipid degradation related genes showed four different patterns of transcript abundance. However, triacylglycerol and FA desaturation biosynthesis processes were highly conserved with similar transcript abundance, increasing during late stages or early to middle stages of development respectively, among all four studied genotypes (Supplementary Figure S7B, D). All these different sets of co-expressed genes in different pathways can be regulated by common or specific regulatory elements. The prediction of putative regulatory elements in co-regulated genes can increase our understanding of seed development and results in tools to breed for improved oil content. Transcription factors play regulatory roles not only in seed development but also in lipid metabolism (Deng et al., 2012) and transcription factor binding sites (or *cis*-regulating elements) are usually located in upstream regulatory regions of genes. The *ABI4* binding motif was shared by genes from the triacylglycerol biosynthesis pathway, FA desaturation and lipid degradation (Supplementary Figure S7C: III-IV), which were all up-regulated 25 DAP. Motif *ABI4* was reported as an important *cis*-regulator of the *DGAT* gene of triacylglycerol biosynthesis (Yang et al., 2011; Wind et al., 2013) and repressor of lipid degradation (Penfield et al., 2006), and is known for its role during seed maturation, seed size, seed germination and seedling growth. The AAAG binding domain was conserved in motifs *Dof2*, *Dof3*, *PBF* and *MNB1A* (*DOF* family) and was found specifically in triacylglycerol biosynthesis genes in our seed samples. The roles of *DOF* genes are in activating seed storage protein genes during seed development and germination in rice (Gaur et al., 2011), barley (Mena et al., 1998), maize (Vicente-Carbajosa et al., 1997), wheat (Mena et al., 1998) and *Arabidopsis* (Stamm et al., 2012). The interwoven connection of different regulatory motifs in Figure 5 supports the fact that target genes are regulated by multiple interacting TFs. The interaction between *Dof* proteins and *HMG* proteins was reviewed in maize seed (Yanagisawa, 2004). Similarly, the other identified motifs, in this study, that belong to the *bZIP*, *MADS*-box, *MYB* family, beta-beta-alpha zinc finger families, as well as unknown motifs, likely play roles in regulating gene expression during seed development and maturation in *B. rapa*. Some motifs reported in *Arabidopsis* seed that are similar to our findings, such as *AG*, *ABI4*, *squamosa*, *bZIP* and *PEND* for triacylglycerol biosynthesis genes,

and *HMG-1* and *Gamyb* for FA synthesis genes (Peng and Weselake, 2011). Moreover, they also reported many more motifs than our findings, and in addition, several motifs observed for triacylglycerol biosynthesis in our study were reported for FA synthesis in this study or vice versa. The possible explanations for finding different numbers of motifs with some disagreement could be (i) the sequence from 1000 bp upstream plus the UTR region was used by Peng and Weselake (2011), but we considered only 1000 bp upstream sequences because the majority of *cis*-regulatory elements are located in this region (Maeo et al., 2009), and (ii) the use of different motif finding tools; TFBS (Lenhard and Wasserman, 2002) and *fdrMotif* (Li et al., 2008) by Peng and Weselake (2011) but MEME tool (Bailey et al., 2009) in this study. The different tools use different algorithms and that could lead to some differences in finding motifs (Meireles-Filho and Stark, 2009). Besides the UTR region and the 1000 bp upstream region, *cis*-regulatory elements can also be located in the downstream sequence, in the gene's introns or in neighbouring genes' introns (Meireles-Filho and Stark, 2009) and consideration of these genomic regions can potentially improve in finding TFs binding motifs.

Acknowledgements

We thank the members of the Unifarm facility of Wageningen University and Research Centre (WUR) for taking care of the plants and all the necessary support. Authors highly appreciate the contributions of Aalt-Jan van Dijk, Plant Research International, Bioscience, Wageningen, The Netherlands for carefully evaluating motif prediction methods used in this study. Finally, the authors are thankful to the Centre for BioSystems Genomics (CBSG), Netherlands which is a part of the Netherlands Genomics Initiative (NGI) for partial financial support for this study.

Supplementary information

Supporting information are available at:

<http://www.biomedcentral.com/1471-2164/14/840>

Supplementary Figure S1: Transcript abundance profiles of ten genes used in real-time PCR gene expression. *FAE1*, *DGAT1*, *DGAT2* from lipid metabolism, *SUS3* and *GBSSI* from carbohydrate metabolism, *12S-CRA1* and *LEA* from storage proteins, *LEC1* and *Glabra2* are transcription factors and *PICKLE* as CHD3-chromatine-remodeling factor. The red colour indicates a high abundance level, green a low level, and grey for missing values. Vertical white lines separate genotypes, yellow-coloured square boxes mark three different groups with high abundance level. Purple coloured boxes at the top indicate that those time-points were selected for the later microarray experiments

Supplementary Figure S2: Pearson correlation coefficients between time points with four replicates per time point within each genotype using all 61654 microarray probes. **A.** Upper triangle: PC175, lower triangle: YS143. **B.** upper triangle: DH78, lower triangle: DH42.

Supplementary Figure S3: Number of probes associated with early stages 18–25 DAP (< -0.01 PC1 loadings) and late stages 35–40 DAP (> 0.01 PC1 loadings) in principal components analysis (PCA). The probes were classified according to MapMan functional categories.

Supplementary Figure S4: Representative abundance levels of gene transcripts belonging to 27 WGCNA gene modules that are significantly associated with **A.** Genotypic differences **B.** temporal differences (time points) **C.** both genotypic differences and temporal differences. Horizontal solid lines separate gene modules and vertical dashed lines separate genotypes. Time points are in ascending order in all genotypes. Numbers in the left corner represent gene modules.

Supplementary Figure S5: Over- and under- representation analysis of time dependent clusters (Cluster I, II and III) into MapMan functional categories using Fisher's exact test. Pink to red colour indicates increasing significance levels for overrepresentation and purple to blue colour increasing significance levels for under-representation. The darker the colour intensity, the more significant. Only significance levels with $p < 0.05$ after FDR correction with the Benjamini-Hochberg method are highlighted. The horizontal green lines separate different pathways.

Supplementary Figure S6: General overview of lipid metabolism showing fatty acid (FA) biosynthesis, FA elongation, lipid desaturation, TAG biosynthesis and glycolipid biosynthesis.

Supplementary Figure S7: Heatmap of gene expression values with hierarchical clustering (Euclidean distance) of all the selected probes ($>$ absolute 2-fold change) belonging to **A.** Fatty acid (FA) synthesis and elongation **B.** FA desaturation **C.** FA degradation **D.** Triacylglycerol (TAG) biosynthesis pathways of lipid metabolism. Vertical white bars separate genotypes (YS: yellow sarson, PC: pak choi, DH42: DH line 42 and DH78: DH line 78). Time points are arranged in ascending order from 18 to 40 DAP within each genotype.

Supplementary Figure S8: Over- and under- representation analysis of gene clusters IV, V and VI, that showed genotypic differences in expression patterns, into MapMan functional categories using Fisher's exact test. Pink to red colour indicates increasing significance levels of overrepresentation and purple to blue colour for increasing significance levels for underrepresentation. The darker the colour intensity, the more significant. Only significance levels with $p < 0.05$ after FDR correction with the Benjamini-Hochberg method are highlighted. The horizontal green lines separate different pathways.

Supplementary Figure S9: Double loop design for hybridization of samples on two-colour Agilent microarrays. Sample names are a combination of genotypes (YS = yellow sarson, PC = pak choi, 42 = DH line 42 and 78 = DH line 78) and time points (18, 20, 25, 30, 35 and 40 Days after pollination). The colours of the arrows in the loop indicate Cy3 (green) and Cy5 (red) dyes in this microarray experiment. **A.** Experiment A represents the design for hybridization of the parental genotypes (yellow sarson and pak choi). **B.** experiment B for the two DH lines (DH42 and DH78).

Supplementary Table S1: Number and percentage of *Brassica* ID (Bra ID) represented in the microarrays, annotated according to MapMan defined metabolic processes.

Supplementary Table S2: Spearman correlation coefficients between real-time PCR and microarray transcript abundance profiles across genotype and seed developmental stages.

Supplementary Table S3: List of selected probes with genotype contrasts in two experiments A and B, as well as time point contrasts in all four genotypes using a minimum 2-fold change criterion.

Supplementary Table S4: Number of selected probes (> absolute 2 fold-change criteria) from temporal contrasts and genotype contrasts into MapMan functional categories. Metabolic processes in highlighted cells are used for further analysis because of apparent changes in the number of selected probes.

Supplementary Table S5: WGCNA gene modules with a significant association with genotypes or time or both genotype and time in ANOVA analyses. The threshold for the level of significance was set at the 0.001 FDR level (Benjamini and Hochberg method). The highlighted cells indicate significant gene modules selected for further analysis.

Supplementary Table S6: Morphological and metabolic descriptions of two parental and two doubled haploid genotypes.

Supplementary Table S7: List of genes used for real-time PCR with their gene name, primer sequence (forward and reverse primers), melting temperature (T_m), GC content percentage, metabolic process and gene ontology biological process (BP).

Supplementary Method S1: Methods used for quantitative real-time PCR.

Supplementary Method S2: Method used for annotation of microarray probes into MapMan functional categories.

Chapter 4

Quantitative trait locus analysis of seed germination and seedling vigour under non-stress and salt stress conditions in *Brassica rapa*: a possible role of *BrFLC2* and *BrFAD2*

Ram Kumar Basnet^{1,2}, Anita Duwal¹, Dev Nidhi Tiwari¹, Dong Xiao^{1,3}, Sokrat Monakhos^{1,4}, Johan Bucher¹, Richard G. F. Visser^{1,2}, Steven P. C. Groot⁵, Guusje Bonnema^{1,2} and Chris Maliepaard¹

- 1- Wageningen UR Plant Breeding, Wageningen University and Research Center, Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands
- 2- Centre for BioSystems Genomics, PO Box 98, 6700 AB, Wageningen, The Netherlands.
- 3- Current address: State Key Laboratory of Crop Genetics and Germplasm Enhancement, Horticultural College, Nanjing Agricultural University, Nanjing, 210095 Jiangsu, China
- 4- Current address: Russian State Agrarian University, Moscow Timiryazev Agricultural Academy, Timiryazevskaya str. 49, 127550 Moscow, Russia
- 5- Plant Research International, Wageningen UR, PO Box 619, 6700 AP, Wageningen, The Netherlands.

Abstract

The genetic basis of seed germination and seedling vigour is largely unknown in *Brassica* species. We performed a study to evaluate the genetic basis of these important traits. Eight quantitative trait loci (QTL) hotspots were identified for seed weight, seed germination, and root and shoot lengths under both non-stress and salt stress conditions in a *B. rapa* doubled haploid population from a cross of a yellow-seeded oil-type yellow sarson and a black-seeded vegetable-type pak choi. A QTL hotspot for seed germination on A02 co-located with a homologue of the *FLOWERING LOCUS C* (*BrFLC2*) gene and its *cis*-acting expression QTL (*cis*-eQTL). *FLC2* is an important repressor of flowering in both *A. thaliana* and *B. rapa* and recently, *FLC2* was reported as a regulator of seed germination in *A. thaliana*. A hotspot on A05 with salt stress specific QTLs co-located with the gene and *cis*-eQTL of *FATTY ACID DESATURASE 2* (*BrFAD2*), which corroborates a reported role of *FAD2* in seed germination and hypocotyl elongation under salinity in *B. napus*. The co-localization of *cis*-eQTLs with these two QTL hotspots suggests these genes as possible candidate genes for the evaluated traits in *Brassica* species. Under salt stress, an epistatic interaction of the QTL hotspots at the *BrFLC2* and *BrFAD2* loci was observed. These results contribute to the understanding of the genetics of seed quality and seedling vigour in *B. rapa* and can offer tools for *Brassica* breeding.

Key words (max. 8): *Brassica rapa*, *BrFAD2*, *BrFLC2*, QTL mapping, eQTL, salt stress, seed germination, seedling vigour.

Introduction

Brassica rapa (A genome, $2n = 20$) consists of several economically important morphotypes, such as leafy vegetables, oilseed types and turnips, with huge morphological and genetic diversity. In recent years, the allopolyploid species *B. napus* (A and B genomes, $2n=38$) has replaced *B. rapa* as the main oilseed crop within the *Brassica* species. *B. rapa* annual oil seed crops like yellow sarson and brown sarson are still grown in regions with short seasons, but *B. rapa* is also used as an important source of genetic variation in *B. napus* improvement, especially in China and Australia (Chen *et al.*, 2007; Rygulla *et al.*, 2007; Chen *et al.*, 2010). Besides the oil content of seeds for especially oil crops, good quality seed and vigorous seedling growth are important traits for crop establishment and higher yield in any crop (Kazmi *et al.*, 2012; Khan *et al.*, 2012). The protrusion of the radicle from the seed is termed seed germination, while seedling vigour refers to the ability of a seed lot to establish seedlings after seed germination under a wide range of environmental conditions (Foolad *et al.*, 2007; Finch-Savage *et al.*, 2010). Seed germination and seedling vigour are very complex traits influenced by different factors, such as the size and composition of the seed, physiological state of the seed, environmental effects during seed production, harvesting, processing and storage, and conditions during germination and early growth. Many efforts have been made to improve seed germination and seedling vigour by optimizing the non-genetic factors; however, the paradigm has shifted to investigate also the genetic factors and to use these to improve crop performance. In several species studies were done to identify quantitative trait loci (QTLs) for seed germination and seedling vigour traits under non-stress and abiotic stress conditions, *e.g.* in tomato (Foolad *et al.*, 2007; Kazmi *et al.*, 2012; Khan *et al.*, 2012), rice (Wang *et al.*, 2011; Wang *et al.*, 2012), soy bean (Csanádi *et al.*, 2001), wheat (Bai *et al.*, 2013), barley (Mano and Takeda, 1997), *Arabidopsis* (Galpaz and Reymond, 2010; DeRose-Wilson and Gaut, 2011; Bouteillé *et al.*, 2012) and *B. napus* (Zhang and Zhou, 2006; Yang *et al.*, 2012). These studies have reported that seed germination and seedling vigour traits are governed by many genes and are strongly affected by environmental conditions (Bettey *et al.*, 2000; Koornneef *et al.*, 2002; Finch-Savage *et al.*, 2010).

Environmental conditions will vary in the presence and level of abiotic and biotic stresses that the seeds and seedlings have to cope with. Therefore, studies of seeds and seedlings need to be carried out under more than only optimal conditions, if they are to be relevant for practical growing situations. Salinity stress is becoming one of the most important abiotic stresses affecting crop growth and yield (DeRose-Wilson and Gaut, 2011; Zhang *et al.*, 2012). About 20% of agricultural land and 50% of irrigated land are affected by salinity (Ren *et al.*, 2010; Su *et al.*, 2013; Dang *et al.*, 2013). Salinity stress reduces the plant's ability for water uptake, causing osmotic stress. At the same time, the accumulation of ions leads to the disturbance of ion homeostasis of plant cells (Wang *et al.*, 2012). Since germinating seeds and establishing seedlings are also vulnerable to salinity stress (Ashraf and McNeilly, 2004), crop establishment and yield can be greatly affected. Salinity tolerance is related to genetic variation (DeRose-Wilson and Gaut, 2011) and thought to be a complex phenomenon controlled by many genes (Ouyang *et al.*, 2007; Galpaz and Reymond, 2010; Joosen *et al.*, 2010; DeRose-Wilson and Gaut, 2011). For a number of crops, it

has been established that larger seed size and higher seed weight indicate more reserve food and contribute positively to seedling establishment (Khan *et al.*, 2012; Ellis, 1992). For the *Brassica* genus, traits such as seed and seedling weight, and seed germination were primarily studied in *B. napus* rather than *B. rapa* (Zhang and Zhou, 2006; Yang *et al.*, 2012). For *B. rapa*, knowledge about genomic regions responsible for seed germination and seedling vigour is largely lacking as are publications describing the molecular basis of seed germination and seedling vigour and response of germinating seed and seedlings to salinity. In this study, a *B. rapa* doubled haploid (DH) population from a cross of an oilseed yellow sarson and a vegetable pak choi was used to study the genetics of seed weight, seed germination and seedling vigour.

We identified 26 QTL regions for traits related to seed weight, seed germination and seedling vigour under non-stress and stress conditions and QTLs for multiple traits co-localized. We identified the candidate genes *B. rapa Flowering Locus C (BrFLC2)* and *B. rapa Fatty acid desaturase2 (BraFAD2)*, homologues of the *A. thaliana FLC* and *FAD2* genes, based on co-location of their expression QTLs (eQTLs) with germination and seedling vigour QTL hotspots and supported by their described functions in related species.

Materials and Methods

Plant material and growing conditions

A *B. rapa* progeny of 170 DH lines (DH68) was developed from three F₁ plants of a cross of a yellow sarson (YS143; accession number: FIL500) and a pak choi (PC175 cultivar: Nai Bai Cai; accession number: VO2B0226) (Xiao *et al.*, 2013). Yellow sarson is a self-compatible annual oil crop with yellow seed colour, while pak choi is a self-incompatible leafy vegetable with black seed colour.

This DH68 population was sown in the greenhouse on a single day on 25th January 2010 (18°C/16°C day/night temperature, 80% humidity and 16 hrs day light). The DH lines varied in time to flowering (43 to 99 days after sowing; DAS) and thus seed maturation was non-synchronous. In 2011, the DH68 population was sown again, this time however, at five different dates from the second week of January to the last week of March to have flowering of all the lines in the same period in order to avoid different environmental conditions during seed development. As a result all the DH lines started flowering during the first two weeks of April, 2011 (31 to 76 DAS). The harvested seeds were stored at 13°C temperature and 30% relative humidity. Germination and seedling vigour experiments were carried out with seeds of 120 DH lines for which enough seeds were available. In addition, thousand-seed weight, which reflects seed content and seed size, was measured.

Pilot study to select NaCl concentrations for salt stress experiments

A pilot study was conducted to determine the optimum level of NaCl concentration for the evaluation of salt stress. The NaCl levels were chosen in such a way that the seedlings could still survive. For two parental lines and a small subset (5-7 DH lines) of the DH population, seed germination and root- and shoot- lengths were initially screened by germinating 30 seeds per genotype in petridishes with two layers of filter papers soaked in seven different NaCl

concentrations: 10, 15, 25, 50, 75, 100 and 150 mM NaCl. In case of seedling vigour assay, root and shoot lengths were measured at 1, 3, 5, 7 and 9 days after germination (DAGs). The materials and methods used for media preparation, seed germination and seedling growth are described in the following sections.

Germination conditions and seed sterilization

Seeds harvested in 2010 and 2011 were used to assess seed germination of the DH lines. Seed germination experiments were conducted in petridishes on two layers of filter papers soaked with agar (non-stress; 0 mM NaCl) or 50 mM NaCl solution (salt stress). The solutions were autoclaved at 120°C and 1.5 bar for 18 minutes. Seeds were sterilized by keeping the seeds overnight in a closed container in chlorine gas fumes of a solution of 20 ml demi-water, 3 ml of 37% fuming HCl and 80 ml of 12% NaOCl. Per treatment per DH line, one set of 30 sterilized seeds was transferred to a petridish. Seeds were all placed between 16:00 and 18:00 hour, so that radicle protrusion would start in the morning of the next day. The petridishes were placed in a climate room (21°C) with 16/8 hours light/dark conditions. Seeds were considered to have germinated when the radicle protrusion had occurred. Starting the next day, the number of germinated seeds in each petridish was counted five times per day in three-hour intervals from 9:00 to 21:00 until all seeds had germinated.

Seedling vigour assay

Seeds harvested in 2010 and 2011 were used to assess seedling vigour of the DH lines. Seedling vigour was measured by placing germinated seeds on vertical plates with 0.8% agar-medium without NaCl (non-stress) or with 50 mM NaCl (salt stress). About 80-90 ml of the agar medium was poured into rectangular plates (12 x 12 x 1.7 cm) in a laminar flow-cabinet. The top one-third portion of agar was removed to leave space for shoot growth. Germinated seeds from the DH lines and parental accessions were transferred from petridishes of the germination assay onto the agar edges of the vertical plates so that all the seedlings in a plate were in the same phase of germination. In total, fifteen seeds per DH line (five seeds per plate, three replicate plates per DH line) were transferred and spaced equally. The plates were sealed with plastic foil and placed in a slanting position at a 60° angle to keep plants growing vertically and to avoid covering of the plates with transpired moisture. All the plates were placed in a climate chamber (21°C temperature, 16/8 hours of light/dark) according to a randomized complete block design (replicates as blocks). Since the study focused on seed germination and early stages of seedling establishment, the seedlings were grown for only the first 10 DAGs. Seedling vigour was quantified by measuring the lengths of shoot and root at different DAG (during first 10 DAGs) and weighing fresh and dry weight of root and shoot per DH line at 10 DAG under both non-stress and salt stress conditions.

In the 2010 assay, the seedlings were grown for ten days, and root length was measured at 3, 5, 7 and 9 DAGs while the shoot length was measured at 3 and 5 DAGs. The root and shoot lengths were measured manually at 3 and 5 DAGs, while image analysis was also done to measure root

length at 3 and 5 DAGs for calibration against the manual measurements and then continued to 7 and 9 DAGs. For image analysis, photos were taken with a digital camera (Nikon D80) as described in Joosen *et al.*, (2010). The root length from the digital image was analyzed using the EZ-Rhizo software package following the procedure described by Armengaud *et al.*, (2009). In the assay of 2011, root and shoot length were measured only manually with a ruler at 3, 5, 7 and 9 DAGs.

Seedling dry weight and fresh weight measurements

At 10 DAG, seedlings were taken out from the agarose-gel and rinsed with water to remove the agar from the roots. Root and shoot were separated and wrapped in white tissue paper for two hours to absorb adhering water before determination of fresh weights. Root or shoot samples of DH lines were pooled over all seedlings of three replicates before taking the weight in order to avoid measurement error due to a too low weight of the samples. For each DH line, roots and shoots were dried overnight at 105⁰C, then dry weights were measured.

Calculation of seed germination parameters

A non-linear germination curve was fitted for each DH line using the Hill function (El-Kassaby *et al.*, 2008) in the software package Germinator (Joosen *et al.*, 2010); growth curves were not fitted for root and shoot length because of the limited number of time points (only 4 time points: 3, 5, 7, and 9 DAGs). Five germination parameters were estimated from the non-linear germination curves: the onset of germination (T10: time to reach 10% germination, in hr), the rate of germination (T50: time to reach 50% germination, hr), uniformity of germination (U7525: time between 25 and 75% germination, hr), maximum germination (Gmax: maximum germination, %) and area under the germination curve (AUC: area under curve) between time zero and 68 h, the latest time point in this study; higher values for AUC correspond to earlier germination, higher germination rate and more uniform germination.

Calculation of salt tolerance parameters

In order to assess the performance of the DH lines for root or shoot length under non-stress and salt stress conditions, two different parameters were used: relative salt tolerance (RelST) and a salt tolerance index (STI) (Saad *et al.*, 2014). RelST is the ratio of a trait value under salt stress versus non-stress conditions (see the formula below), and indicates the relative performance of genotypes for their root, or shoot growth across conditions. A genotype with RelST greater than one for root length has a longer root under salt stress than under non-stress; a genotype with RelST lower than one is sensitive to salt as illustrated by reduced root length under salt stress. A value of one for RelST indicates that the root length of the genotype is not affected by the stress. The other parameter, Salt Tolerance Index (STI), was calculated by comparing the shoot or root length under stress and non-stress conditions, but now relative to the average length under the non-stress condition over the whole population using the formula below, as described by Fernandez, (1992). An STI equal to one indicates that root or shoot length of a specific genotype under stress/non stress is equal to the average length under the non-stress condition over all DH

lines. An STI greater than one indicates that the root/shoot length of a DH line is higher in one condition, or in both conditions relative to the mean of the population under non-stress conditions. If the STI is lower than one, there is lower root/shoot length in one or both conditions as compared to the average population under non-stress conditions.

$$\text{RelST}_{ij} = \frac{X_{ij} \text{ at stress}}{X_{ij} \text{ at non-stress}} \quad \text{STI}_{ij} = \frac{(X_{ij} \text{ at non-stress} * X_{ij} \text{ at stress})}{(X_{\text{average (j)}} \text{ at non-stress})^2}$$

where X_{ij} = root or shoot length of genotype i at j days after germination (DAG)

Summary statistics, graphical representation and heritability

Descriptive statistics were calculated for all traits. Box plots were made to visualize the distributions of seed germination parameters across the experiments, and shoot and root length across growing days and experiments.

Separate heatmaps were generated to visualize the Pearson correlation coefficients among seed germination parameters or shoot and root lengths at different DAGs for two treatments and using seed batches of two years. The heatmaps of the correlations were combined with hierarchical clustering using Euclidean distances and complete linkage after scaling the traits to zero mean and standard deviation one (this is equivalent to clustering on the Pearson correlations).

As DH lines are genetically fixed homozygotes, there is no variation due to dominance and therefore, narrow-sense heritability was estimated as: $h^2 = \sigma_a^2 / (\sigma_a^2 + \sigma_e^2)$, where h^2 is narrow-sense heritability, σ_a^2 is additive genetic variance and σ_e^2 = environmental variance (Bernardo, 2002). The variance components (σ_{DH}^2 and σ_e^2) were estimated using a linear model (one-way ANOVA): trait = DH line + replication + error. The additive genetic variance (σ_a^2) was estimated as $\sigma_a^2 = \sigma_{DH}^2 / 2$ (Bernardo, 2002). The computation of Pearson correlation coefficients, the hierarchical clustering, heritability calculation and heatmap visualization were performed using R statistical software (R Core Team, 2012).

Genetic map construction

Linkage analysis and map construction were performed with JoinMap 4.0 (Van Ooijen, 2006) using a regression approach and the Kosambi map function. In total 435 markers mainly AFLPs, SSRs, *Myb* targeted markers and gene-targeted markers were mapped in an integrated map of 10 linkage groups. This integrated map is a slightly modified version of the linkage map presented in Xiao *et al.*, (2013), as additional DH lines from two different F_1 plants of the same cross were used in the present paper. As the parents were not homozygous, these F_1 plants were not identical, which resulted in minor changes in the genetic maps (Supplementary Figure S4).

QTL analysis

Single trait QTL analysis was performed to identify genomic regions controlling a trait, using interval mapping (IM), and restricted and full multiple QTL model mapping (rMQM and MQM) in MapQTL 6.0 (van Ooijen, 2009). Initially peak markers from a map region with LOD score > 2 were

used as cofactors and a final list of cofactors was selected using the automatic cofactor selection procedure, which uses a backward elimination approach. The cofactor selection process was repeated with different sets of cofactors until the QTL profile was stable.

QTL mapping was carried out for all the germination parameters, root and shoot length, and fresh and dry weight of both the root and the shoot. Similarly, QTL mapping was done for the salt tolerance index (STI) and for relative salt tolerance (RelST) of root and shoot length at different DAGs. In this study, a genome-wide significance LOD score threshold of 3.0 was derived at the 95 percentile of 10,000 permutations of each trait. There was hardly any variation in the threshold between traits so, this threshold was used for all traits to declare a QTL as significant. QTLs with a LOD score between 2 and 3 were considered as putative QTLs. Finally, 1-LOD support intervals were determined for the assigned QTLs. The cut-off value for declaring a number of co-locating QTLs as a hotspot was calculated using the package “hotspots” in R (Darrouzet-Nardi, 2010). In this study, co-localized QTLs were coded as “Co-QTL k - m ”, where k indicates for linkage group and m for QTL number.

Based on the observed main effects of significant or putative QTLs, epistatic interactions were tested for all possible pairs of two QTLs using the following ANOVA model: $\text{trait}_i = \text{QTL}_1 + \text{QTL}_2 + \text{QTL}_1 * \text{QTL}_2 + \text{error}$. First, an ANOVA model consisting of only the main effects of the QTLs was fitted. Then, the QTL*QTL interaction term was added, and this change to the model was tested for significance; if significant, the contribution of this epistatic interaction to the phenotypic variance was quantified. ANOVA were performed in R.

Quantitative real-time PCR (RT-qPCR) and eQTL analysis

Transcript abundance of candidate genes *BrFLC2* and *BrFAD2*, was determined in developing seeds of the DH population using RT-qPCR. Transcripts of the genes were profiled with two technical replicates using RNA samples of seeds harvested 28 days after pollination of the 120 DH lines. RNA isolation and purification were done following the same protocol used by Basnet *et al.*, (2013). Transcript abundance of these genes was measured in RT-qPCR in 96-well optical reaction plates using the iQ™ SYBR® Green Supermix (Bio-Rad, www.Bio-rad.com) according to Xiao *et al.*, (2013), but actin was used as reference gene to calculate the cycle threshold (C_t) values and $\Delta\Delta C_t$ values. eQTL analyses were done using interval mapping (IM), and restricted and full multiple QTL model mapping (rMQM and MQM) in MapQTL 6.0 (van Ooijen, 2009). Molecular markers specific for *BrFLC2* and *BrFAD2* genes were mapped in this population; an eQTL was defined as a *cis*-eQTL (local eQTL) if the edge of a 2-LOD support interval of an eQTL was within 10 cM of the genetic map position of the gene, otherwise the eQTL was defined as a *trans*-QTL (distant eQTL).

Results

Salt stress conditions and observation time points

A pilot study showed that seed germination and root and shoot length under salt stress conditions at 10, 15 and 25 mM NaCl were comparable to that under non-stress (0 mM NaCl), indicating that these concentrations were too low to induce visible symptoms of salt stress. At the concentration

of 100 mM NaCl, seeds hardly germinated; at 75 mM NaCl there was germination, but the seedlings did not grow out enough to be able to measure root and shoot length. Therefore, in this study, we used 0 mM NaCl (non-stress) and 50 mM NaCl for phenotyping the DH population. As roots and shoots had hardly grown at one DAG and showed very little variation among DH lines, it was decided to measure these traits at 3, 5, 7 and 9 DAGs under both non-stress and salt stress (50 mM NaCl).

Seed weight and seed germination

Yellow sarson had larger and heavier seeds, which germinated earlier than pak choi seeds; generally, germination under non-stress was earlier and more uniform than under salt stress. Thousand-seed weight was almost three times higher for yellow sarson than for pak choi: 6.6 g and 5.8 g, versus 1.9 g and 1.4 g for seed batches of 2010 and 2011, respectively. As germination tests were performed without replicates, differences in the germination parameters between the two parents were not tested statistically. The Gmax of two parental genotypes and DH lines varied from 26.7% to 100% across conditions (Table 1). Germination of yellow sarson seed was marginally more uniform than that of pak choi under non-stress (in 2010 and 2011) and salt stress (in 2010), while pak choi germinated more uniformly than yellow sarson under salt stress in 2011 (Table 1). Lower T10 and T50 for yellow sarson under both conditions and in both years indicates earlier onset as well as faster rate of germination for yellow sarson than for pak choi (Table 1).

In the DH population, the average T10 and U7525 were lower under non-stress than salt stress, which indicates that seed germination started earlier and was more uniform under non-stress conditions. T50 and uniformity (U7525) were positively correlated to each other, and negatively with Gmax and AUC (Supplementary Figure S1). T10 had a positive correlation with T50, and a negative correlation with AUC and Gmax, and no correlation with U7525. Pearson correlation coefficients of the same parameter were higher between two seed batches (two growing years) than between stress levels (Supplementary Figure S1). Thousand-seed weight was positively correlated with AUC ($r = 0.25$ to 0.37), and Gmax ($r = 0.16$ to 0.24) and negatively correlated with T10 ($r = -0.18$ to -0.36), T50 ($r = -0.25$ to -0.31) and U7525 ($r = -0.11$ to -0.24) under non-stress and salt stress conditions across two years' seed batches.

Root and shoot length of seedlings

The roots of yellow sarson were longer than roots of pak choi ($p \leq 0.05$) on all DAGs and the differences in root length between the parents increased over time. The variance of root - and shoot- length over the DH lines was increased with time (Figure 1; Supplementary Figure S2). Under salt stress, yellow sarson had longer roots than pak choi between 3 and 7 DAGs, while at 9 DAG root lengths were similar. Root length was reduced under salt stress (Figure 1; Supplementary Figure S2). Similar to root length, also shoot length was larger in yellow sarson than pak choi but for shoot length the difference between the two parents was smaller under salt stress than under non-stress conditions. Large variation in shoot length was observed across the

DH lines. In both conditions in both years, a large number of transgressive segregants were observed across the DH population for all the seedling traits (Figure 1; Supplementary Figure S2). Both fresh and dry weight of root and shoot were higher in yellow sarson than in pak choi, except for root fresh and dry weight of the 2011 seed batch at 50 mM NaCl (Table 2). Fresh and dry weight of root and shoot decreased under salt stress for the parents as well as the DH population, but the decrease was stronger for yellow sarson than for pak choi.

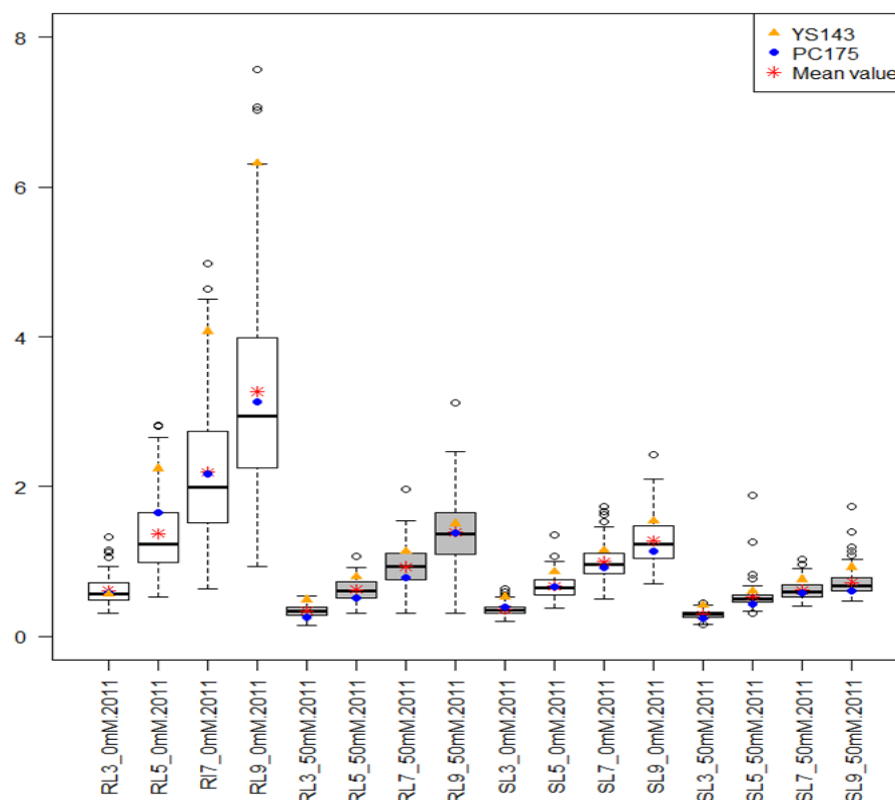


Figure 1: Box plots showing the distributions of root length (RL) and shoot length (SL) at different days after germination (DAG) under non-stress (0 mM NaCl) and salt stress (50 mM NaCl) for the 2011 seed batch. The shaded colour of the boxes indicates the treatments: white for non-stress and grey for salt stress. The y-axis indicates root and shoot length (in cm). The x-axis label is the combination of RL or SL at 3, 5, 7 and 9 DAG at non-stress and salt stress conditions. Box plots showing the distributions of RL and SL for the 2010 seed batch are shown in Supplementary Figure S2.

Cluster analysis of root and shoot traits

In a hierarchical cluster analysis using Pearson correlation coefficients among the traits, two clusters were observed, one for root traits, the other for shoot traits, with generally low correlation coefficients between the two clusters (Supplementary Figure S3). Within both the root and shoot trait clusters, sub-clusters with stronger correlations were observed for the treatments followed by years. So, in the cluster analysis first the tissues are separated (root versus shoot), then treatments, then seed batches/years.

Table 1: Summary statistics of seed germination parameters under non-stress and salt stress conditions of 2010 and 2011 seed batches.

Traits	Treatment	2010				2011			
		Parents		DH lines		Parents		DH lines	
		PC175	YS143	Mean \pm SD	Range	PC175	YS143	Mean \pm SD	Range
Gmax (%)	Control	100.0	100.0	97.6 \pm 10.3	26.7-100	96.5	100.0	96.4 \pm 7.7	45.7-100
	50mM	100.0	100.0	96.5 \pm 9.0	50.0-100	100.0	100.0	95.4 \pm 10.4	32.3-100
U7525 (hr)	Control	1.5	1.0	4.2 \pm 3.6	0.4-22.7	2.5	2.2	7.8 \pm 5.1	1.4-30.9
	50mM	1.3	0.7	4.4 \pm 3.2	0.7-16.8	4.0	7.9	9.2 \pm 5.7	1.8-33.2
T50 (hr)	Control	19.0	15.3	21.9 \pm 5.1	11.1-44.7	19.3	15.1	21.4 \pm 6.9	10.0-48.6
	50mM	17.8	17.0	24.4 \pm 6.3	15.1-52.5	22.1	13.9	26.1 \pm 10.4	11.4-79.2
T10 (hr)	Control	17.6	14.3	18.5 \pm 4.3	7.0-40.5	16.9	13.0	14.8 \pm 4.3	4.7-26.3
	50mM	16.5	16.3	20.5 \pm 5.4	11.3-41.2	18.5	8.0	17.9 \pm 6.0	7.8-42.3
AUC	Control	81.0	84.7	75.9 \pm 10.1	13.7-88.1	75.6	84.8	73.5 \pm 11.0	26.2-88.6
	50mM	82.2	83.1	72.7 \pm 10.9	29.9-84.9	76.5	84.5	68.9 \pm 13.0	15.5-88.0
Seed weight*	-	1.9	6.6	2.5 \pm 0.9	0.5 - 5.9	1.4	5.8	2.2 \pm 0.8	0.2 - 5.3

* - 1000 seed weight in gram (g).

Table 2: Summary statistics of root and shoot length, and their fresh and dry weight under non-stress and salt stress conditions of 2010 and 2011 seed batches.

Trait	Year	Day	Treatment	Root				Shoot			
				Parents		DH lines		Parents		DH lines	
				PC175	YS143	Mean \pm SD	Range	PC175	YS143	Mean \pm SD	Range
Length (cm)	2010	3	Control	0.6 \pm 0.1	1.6 \pm 0.1	1.3 \pm 0.6	0.2-3.5	0.5 \pm 0.1	0.6 \pm 0.1	0.6 \pm 0.2	0.3-1.5
		3	50mM	0.5 \pm 0.1	0.7 \pm 0.1	0.5 \pm 0.2	0.1-1.2	0.4 \pm 0.1	0.5 \pm 0.1	0.4 \pm 0.1	0.2-0.9
		5	Control	1.3 \pm 0.9	5.6 \pm 0.1	2.3 \pm 1.0	0.3-6.7	0.9 \pm 0.2	1.5 \pm 0.2	1.0 \pm 0.4	0.5-2.7
		5	50mM	1.1 \pm 0.2	1.3 \pm 0.4	0.9 \pm 0.3	0.2-1.8	0.6 \pm 0.1	0.7 \pm 0.1	0.5 \pm 0.1	0.2-1.4
		7	Control	1.9 \pm 0.3	7.0 \pm 0.1	2.8 \pm 1.2	0.7-7.3	-	-	-	-
		7	50mM	1.8 \pm 0.2	1.8 \pm 0.7	1.2 \pm 0.4	0.3-2.6	-	-	-	-
		9	Control	1.9 \pm 0.3	7.2 \pm 0.2	3.3 \pm 1.4	0.8-8.6	-	-	-	-
		9	50mM	2.4 \pm 0.2	2.5 \pm 1.6	1.5 \pm 0.6	0.3-3.9	-	-	-	-
	2011	3	Control	0.6	0.6	0.6 \pm 0.3	0.1-2.1	0.4	0.5	0.4 \pm 0.1	0.1-0.9
		3	50mM	0.3	0.5	0.3 \pm 0.1	0.1-0.7	0.2	0.4	0.3 \pm 0.1	0.1-0.5
		5	Control	1.7	2.2	1.4 \pm 0.8	0.1-6.2	0.7	0.9	0.7 \pm 0.2	0.1-2.0
		5	50mM	0.5	0.8	0.6 \pm 0.2	0.1-2.0	0.4	0.6	0.5 \pm 0.1	0.1-1.0
		7	Control	2.2	4.1	2.2 \pm 1.3	0.2-9.0	0.9	1.2	1.0 \pm 0.4	0.1-3.0
		7	50mM	0.8	1.1	0.9 \pm 0.4	0.2-3.5	0.6	0.8	0.6 \pm 0.2	0.1-1.7
		9	Control	3.1	6.3	3.3 \pm 1.9	0.3-12.0	1.1	1.5	1.3 \pm 0.5	0.4-3.0
		9	50mM	1.4	1.5	1.4 \pm 0.7	0.2-0.4	0.6	0.9	0.7 \pm 0.2	0.3-2.5
Weight (g)	2010	Fresh	Control	70.0	110.0	69.8 \pm 46.2	11.4 - 276.6	357.3	770.0	440.7 \pm 133.7	158.6 - 813.3
			50 mM	25.2	40.6	13.5 \pm 11.0	1.5 - 72.0	311.2	327.6	226.5 \pm 88.8	57.0 - 503.0
		Dry	Control	9.8	21.0	16.0 \pm 15.3	2.1 - 124.8	29.2	72.2	43.0 \pm 19.6	15.5 - 114.8
			50 mM	5.8	7.2	4.2 \pm 1.9	0.5 - 10.3	24.6	45.7	25.7 \pm 10.7	3.0 - 51.5
	2011	Fresh	Control	90.0	130.0	65.5 \pm 33.2	13.9 - 199.3	400.0	540.0	428.0 \pm 98.5	151.1 - 687.5
			50 mM	17.1	7.2	29.1 \pm 21.9	2.5 - 141.4	246.0	353.6	325.7 \pm 89.7	90.0 - 560.5
		Dry	Control	7.7	17.4	8.2 \pm 0.3	2.9 - 17.4	26.7	80.9	34.4 \pm 12.2	5.3 - 86.6
			50 mM	5.6	4.3	4.5 \pm 0.2	0.1 - 11.7	19.0	60.7	26.9 \pm 9.2	11.4 - 60.7

Note: - not available

Heritability

Root and shoot length under non-stress and stress conditions had low to moderately high narrow-sense heritabilities (0.2-0.7) (Table 3). Under salt stress, heritabilities were generally lower than under control conditions. The heritabilities were generally higher for seed batch 2011, where flowering and seed ripening were synchronised, than for the seed batch of 2010. The salt tolerance parameters STI and RelST both for root and shoot length also had low to moderately high heritability estimates (0.2 to 0.7); RelST had a lower heritability than STI for both root and shoot length (Table 3).

Table 3: Narrow-sense heritabilities of root length (RL), shoot length (SL) and salt tolerance parameters under non-stress and salt stress conditions of 2010 and 2011 seed batches for different days after germination (DAG). RelST indicates Relative Salt Tolerance, STI indicates the Salt Tolerance Index.

	DAG	Control		50 mM NaCl		2010		2011	
		2010	2011	2010	2011	RelST	STI	RelST	STI
Root length	3	0.6	0.6	0.5	0.6	0.3	0.5	0.2	0.6
	5	0.6	0.6	0.6	0.6	0.5	0.5	0.2	0.6
	7	0.6	0.7	0.6	0.6	0.5	0.5	0.2	0.6
	9	0.6	0.7	0.5	0.6	0.4	0.4	0.4	0.6
Shoot length	3	0.6	0.6	0.4	0.5	0.3	0.6	0.2	0.6
	5	0.5	0.6	0.4	0.6	0.3	0.5	0.3	0.6
	7	-	0.7	-	0.6	-	-	0.5	0.6
	9	-	0.7	-	0.7	-	-	0.5	0.7

QTLs for seed germination and seed weight

Over the two seed batches (2010 and 2011), two conditions (non-stress and 50 mM salt) and five germination parameters, in total, 25 QTLs and 20 putative QTLs (LOD score between 2 and 3) were found (Figure 2B; Supplementary Table S2). For the significant QTLs, the explained variances ranged from 7.5% to 27.2%. No QTLs were found for Gmax under control conditions. On A02, QTLs were detected mainly for T10 and T50 with explained variance ranging from 9.3 % to 27.2 % (Supplementary Table S2), with the favourable effect coming from the yellow sarson allele (*i.e.* the pak choi allele increases T10 and T50). And, putative QTLs were detected mainly for AUC on A02 under salt stress in 2010 and non-stress condition in 2011, with a favourable effect from the yellow sarson allele (Figure 2B). However, putative QTLs for T10 and T50 were detected on A01 with a favourable allelic effect of the yellow sarson allele. On A05, one QTL and three putative QTLs with explained variance from 6.5 % to 11.1% were mapped for mainly uniformity under salt stress. For thousand-seed weight in 2010 and 2011 a single QTL was found on A05, with explained variance ranging from 8.3% to 16.1% (Figure 2B; Supplementary Table S2).

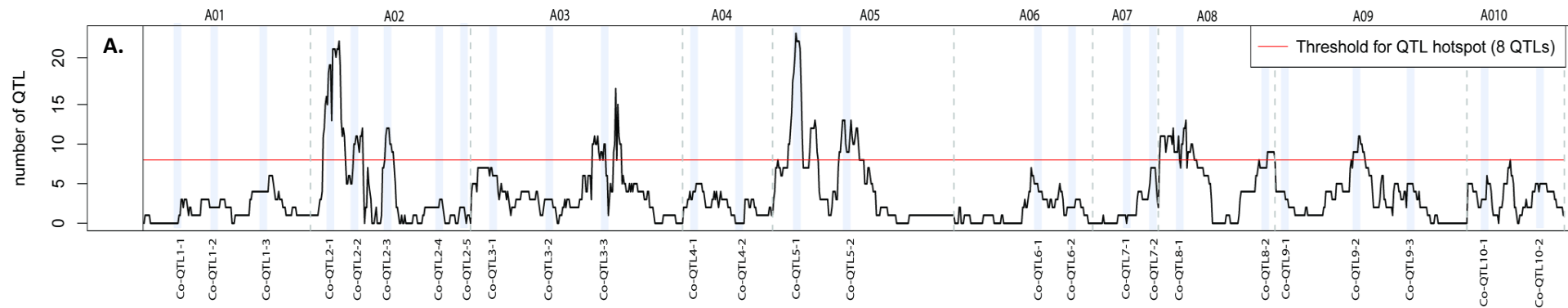
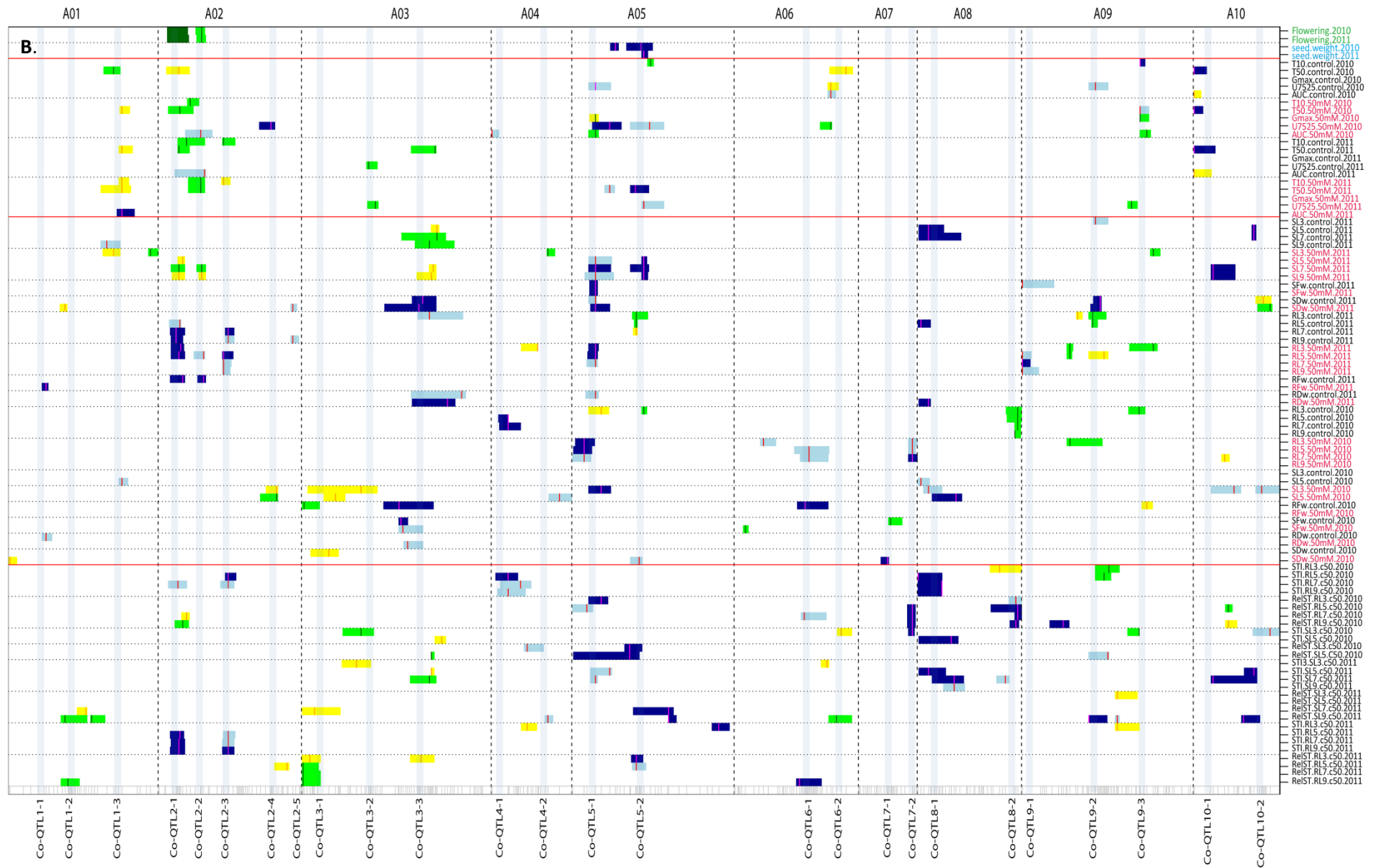


Figure 2: An overview of single trait QTL profiles of seed germination, seedling vigour and salt tolerance parameters under non-stress and 50 mM NaCl salt stress conditions of 2010 and 2011 seed batches. **A.** QTL hotspots (>8 significant QTLs) indicated by the number of significant QTLs plotted for 10 linkage groups (“A0 to A10”). The dotted vertical lines separate the ten linkage groups. **B.** Single trait QTL profiles of flowering time, seed weight, seed germination, seedling vigour and salt tolerance parameters. Seed germination includes six parameters: T10, T50, U7525, Gmax and AUC; seedling vigour includes root and shoot length at different DAGs, and root and shoot biomass measured under non-stress and 50 mM NaCl salt stress conditions. The different traits are indicated along the y-axis, the 10 linkage groups on the x-axis, separated by dotted lines. QTLs that had a high phenotypic value for the YS143 allele are in light blue (for QTLs with LOD 2-3) and blue (for QTL with LOD >3) while QTLs with a high phenotypic value for the PC175 allele are indicated in yellow (for QTLs with LOD 2-3) and green (for QTLs with LOD > 3). The colour streak on the QTL profile indicates the QTL peak position. The 26 QTL co-localization regions are indicated as Co-QTL followed by the number of the linkage group and a serial number within a linkage group, for example, Co-QTL1-2 indicates the second co-localization QTL region on A01. The red colour of a trait label indicates the trait under 50 mM NaCl salt stress condition. The trait labels are described in Supplementary Table S1.



QTLs for seedling vigour

For the 24 seedling vigour traits measured in 2010 and 2011 seed batches, 69 QTLs were identified, distributed over 10 linkage groups (Figure 2B; Supplementary Table S3). The explained variances ranged from 7.1% to 24.3%. For the 2010 seed batch, 22 QTLs were identified for 13 traits on different linkage groups with at least one QTL per trait and an additional 24 putative QTLs for 13 different traits (Figure 2B; Supplementary Table S3). For the 2011 seed batch, 47 QTLs were identified for 20 traits with at least one QTL per trait and all those traits also had putative QTLs. For two traits, many QTLs were detected: for root length at 5 DAG under 50 mM salt in the 2011 seed batch, four QTLs with explained variance ranging from 8.3 to 13.9%, and for shoot length at 7 DAG under 50 mM salt for the 2011 seed batch, five QTLs on with explained variance ranging from 10.0 to 14.2% (Figure 2B; Supplementary Table S3).

QTLs for seedling salt tolerance parameters

Two salt tolerance parameters, salt tolerance index (STI) and relative salt tolerance ratio (RelST) were calculated for root and shoot lengths. In total, 47 QTLs were identified for 24 traits for root and shoot length at different DAGs from two seed batches with explained variance ranging from 7.8% to 22.2% (Figure 2B; Supplementary Table S4). The trait RelST for shoot length at 9 DAG in 2011 had the largest number of QTLs (six) with explained variance from 11.7% to 17.7% while the other traits for STI and RelST had 1 to 4 QTLs. Among the 47 QTLs, 8 QTLs had < 10 % explained variance, 30 QTLs had 10-15% and 9 QTLs had > 15% explained variance (Figure 2B; Supplementary Table S4). Since root and shoot length traits were measured repeatedly at 3, 5, 7 and 9 DAGs, many QTLs for the same trait at different DAGs probably represents the same QTL.

Co-localization of QTLs

QTLs co-localized on 26 unique genomic regions across ten linkage groups; however 9 significant QTL hotspots (with ≥ 8 QTLs) were detected (Figure 2A). Hotspots *Co-QTL2-1*, -2 and -3 on A02, *Co-QTL3-3* at the middle of A03, *Co-QTL5-1* and -2 on A05, *Co-QTL8-1* and -2 at the top of A08 and *Co-QTL9-2* at the middle of A09 were considered major QTL hotspots (≥ 8 QTLs). QTLs co-localized on these hotspots are often for the same traits measured at different DAGs, at different treatments or in different years (Figure 2A-B; Supplementary Table S2-S4). We also considered whether the QTL alleles from parents yellow sarson or pak choi could be of importance to breeders. QTL hotspots, such as *Co-QTL1-3* on A01, *Co-QTL2-1* and -2 on A02 and *Co-QTL10-1* on A10 mainly included QTLs for T10 and T50 seed germination parameters, with the yellow sarson allele on A01 and A02 and the pak choi allele on A10 associated with earlier onset and faster germination (Figure 2B). On hotspots *Co-QTL2-1* and -2 on A02, QTLs for root and shoot traits from the 2011 seed batch were also co-localized; here, the yellow sarson allele is associated with an increase in root length and the pak choi allele with increased shoot length. Hotspots *Co-QTL9-2* and -3 contain QTLs associated with T50, Gmax and AUC under salt stress in 2010 seeds and U7525 under salt stress in 2011 seeds; these hotspots also harbour a major QTL for seed colour with 32.7% explained variation (data not shown). In addition, these hotspots *Co-QTL9-2* and -3 contain QTLs

for shoot and root length and shoot weight under both conditions and also for salt tolerance parameters. At this locus, the pak choi allele increased the G_{max}, AUC, root- and shoot- lengths and shoot weight. On *Co-QTL8-1*, QTLs for root and shoot lengths, their weights and salt tolerance parameter STI co-localize. The yellow sarson allele for that QTL increased root and shoot length and weight and also improved seedling performance under salt stress (Figure 2B).

Stress treatment specific QTLs

The two QTL hotspots on A05 *Co-QTL5-1* and -2 harbour QTLs for AUC and G_{max} (in 2010), T50 (in 2011) and U7525 (in 2010 and 2011) under salt stress and RelST of root and shoot length in both years (Figure 2B). The pak choi allele increased the germination parameters G_{max} and AUC from the 2010 seed batch, but the same allele is associated with a lower germination rate (T50 in 2011) and with a decreased uniformity of germination (U7525 in 2010 and 2011). The values of both salt tolerance parameters were higher in yellow sarson; consistent with that, it was the presence of the yellow sarson allele in the DH progeny that gave higher RelST and STI of root and shoot lengths (Figure 2B).

Epistatic interactions between QTLs

Among the 64 traits related to germination and seedling vigour measured under two treatments from seeds harvested in two years, epistatic interactions were observed for 16 traits: 9 for germination and 7 for seedling vigour. Most had only a single epistatic interaction, with an explained variance between 5 and 10%; however, four traits had 2 or 3 epistatic interactions (Figure 3). For T50 (50 mM salt, 2011) there were two epistatic interactions between *Co-QTL2-2* and *Co-QTL5-1*, and between *Co-QTL2-2* and *Co-QTL5-2* with 10.54% and 17.58% explained variation, respectively. Similarly, shoot length under salt stress at 3 and 9 DAGs in the 2011 seed batch and root fresh weight under non-stress in the 2010 seed batch had two or three epistatic interactions. QTL hotspots *Co-QTL10-1*, *Co-QTL2-2* and *Co-QTL6-2* were the main QTL regions that had the largest number of interactions, with 8, 6 and 5 other QTL regions, respectively (Figure 3).

Co-localization of phenotypic QTLs with candidate genes BrFLC2 and BrFAD2

Two genes *BrFLC2* and *BrFAD2* were reported for their roles in seed germination and seedling vigour traits in *A. thaliana* and *B. napus* (Chiang *et al.*, 2009; Wang *et al.*, 2010; Zhang *et al.*, 2012). Based on the co-location of *co-QTL2-1* and -2 on A02 with *BrFLC2* and of *co-QTL5-1* and -2 on A05 with *BrFAD2*, we analyzed transcript abundance of these genes in the DH population and mapped eQTLs for these genes to find out whether their eQTLs co-locate with the phenotypic QTL hotspots and with the physical position of the genes themselves.

For *BrFLC2*, a *cis*-eQTL was mapped over *Co-QTL2-1* and -2 on A02 with LOD scores 5.4 and 3.7 that explained 20.8% and 14.0% of the total variation in transcript abundance, respectively. For *BrFAD2*, a *cis*-eQTL co-located over the *Co-QTL5-1* and -2 regions with LOD score 7.1 and 22.0% explained variance while another *trans*-eQTL mapped at the *Co-QTL9-2* region on A09 with LOD score 3.6 and 10.5% explained variance (Figure 4; Table 4).

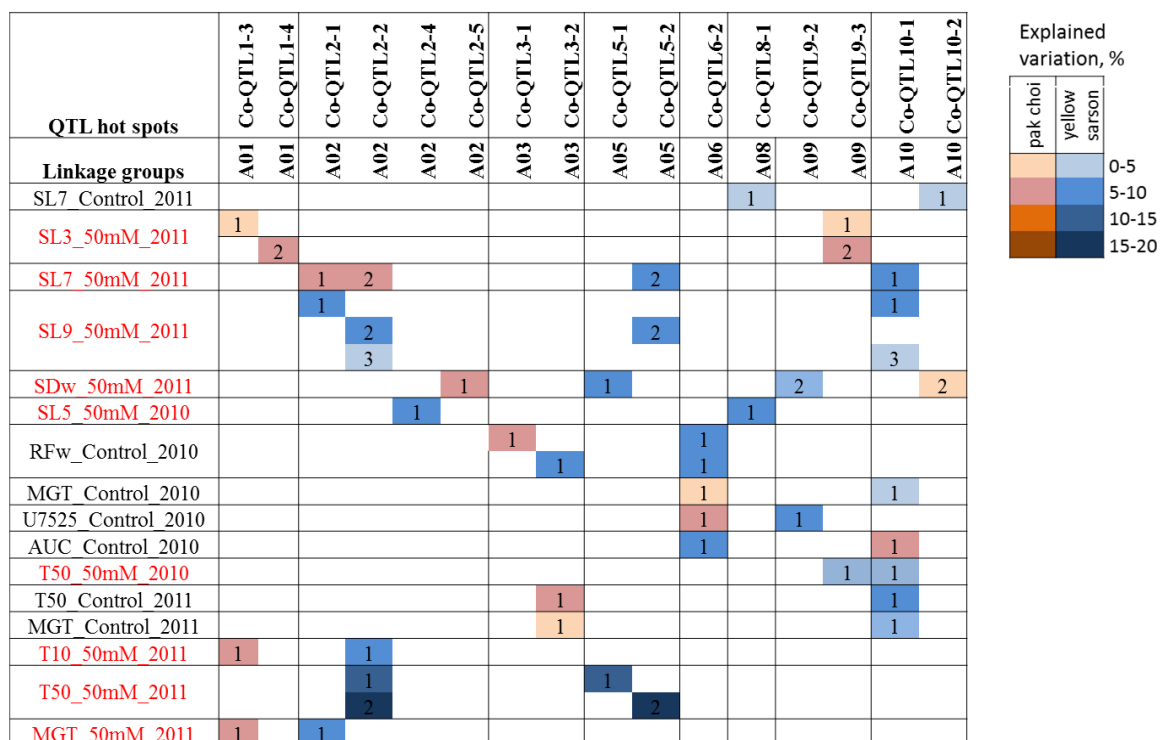


Figure 3: Epistatic interactions of QTL regions for seed germination and seedling vigour traits. QTLs identified for different traits were tested in a pair of two QTLs for each trait. Identical numbers indicate a pair of QTLs with significant epistatic interaction ($\alpha = 0.05$). The colour intensity increases with higher explained variance (%) of the interaction.

Discussion

Good seed germination and high seedling vigour under diverse conditions are essential for the establishment of a crop. In this study, germination characteristics were studied and seedling vigour was evaluated during the early growth stages of *B. rapa* seedlings, both under non-stress and salt stress conditions. Vigorous seeds have fast and uniform germination, a high germination rate and high seedling vigour at the early stages of growth under diverse conditions; this is the main focus of any crop breeding program for good crop establishment.

Salinity is one of the major limiting abiotic stresses for high crop production, affecting nearly 40% of agricultural lands in the world (Mittler, 2006). High levels of salt especially during seed germination and early plant growth, directly affects the crop establishment, in severe cases leading to complete crop failure or strongly reduced yields (Mano and Takeda, 1997; Ashraf and McNeilly, 2004; Su *et al.*, 2013). In this study, 120 genotypes from a DH population from a wide cross of *B. rapa* yellow sarson and pak choi were used to identify genomic regions associated with seed germination characteristics and seedling vigour under non-stress and salt stress conditions. Seed germination and shoot and root related traits during the first 10 days after germination were assessed to quantify seed- and seedling- vigour. Two major QTL hotspots co-locate with candidate genes *BrFLC2* and *BrFAD2* for seed germination and seedling vigour. Even though many genes map under these QTL hotspots, we did a follow-up investigation of these two genes in particular, since these genes were also reported for roles in the studied traits in *A. thaliana* and *B. napus* (Chiang *et*

al., 2009; Wang *et al.*, 2010; Zhang *et al.*, 2012); in this follow-up study we found that their *cis*-eQTLs co-located with the phenotypic QTL hotspots.

Early seedling growth is more affected by salt stress than seed germination

Under natural conditions, plants are exposed to different levels of salt stress; in this study 50 mM NaCl was chosen to mimic the salinity stress that affects seed germination and seedling vigour in a field situation. Maximum seed germination was not drastically affected at 50 mM NaCl, being still at a level of 95-100%; however, in general, other germination parameters, relating to rate and uniformity of germination were negatively affected. Root and shoot growth were also reduced.

Oil-type yellow sarson has improved seed germination and seedling vigour compared to vegetable-type pak choi under non-stress and salt stress conditions but is more sensitive to salt stress

The yellow sarson parent had larger seed size and higher thousand-seed weight than the pak choi parent, and displayed earlier onset, more uniform and faster germination under both non-stress and salt stress conditions (Table 1). This parent also had a higher root- and shoot- length and biomass (Table 2). The positive correlations of thousand-seed weight with AUC and Gmax, and negative correlations with T10, T50 and uniformity (U7525) of germination in the DH population supports that larger seeds germinate earlier, faster, more uniformly and to a higher maximum germination than smaller seeds. Thus, we conclude that yellow sarson had higher seed quality and seedling vigour than pak choi. The explanation for the larger seeds of yellow sarson (than those of pak choi) could be that yellow sarson was selected for high oilseed yield since it is an oil-type crop, while pak choi was selected for high vegetative mass. The increased seedling vigour of yellow sarson might then be a result of the seeds containing more nutrients (Ambika *et al.*, 2014). Susko and Lovett-Doust (2000) reported a positive effect of seed mass (weight) on higher and faster seed germination and seedling growth in *Alliaria petiolata* (a *Brassicaceae*) and Khan *et al.*, (2012) also observed a positive effect of higher seed weight and larger seed size on seedling vigour traits in tomato. However, during seedling growth yellow sarson was more severely affected by salt stress than pak choi, with more strongly reduced shoot and root length and biomass under salt stress (Table 2; Figure 1; Supplementary Figure S2). One possible explanation of this result could be the thinner seed coat in yellow-seeded genotypes than that of brown/black-seeded genotypes (Xiao *et al.*, 2012). The thickness of the seed coat, the colour itself, composed of pro-anthocyanin (an antioxidant), and antioxidants such as anthocyanin and flavonoids are reported to protect germination under salt stress (Umnajkitikorn *et al.*, 2013). The thinner the seed coat, the higher the permeability, which can lead to higher cumulative absorption of salt solutes over time and this could possibly cause yellow sarson to be more affected by salt stress.

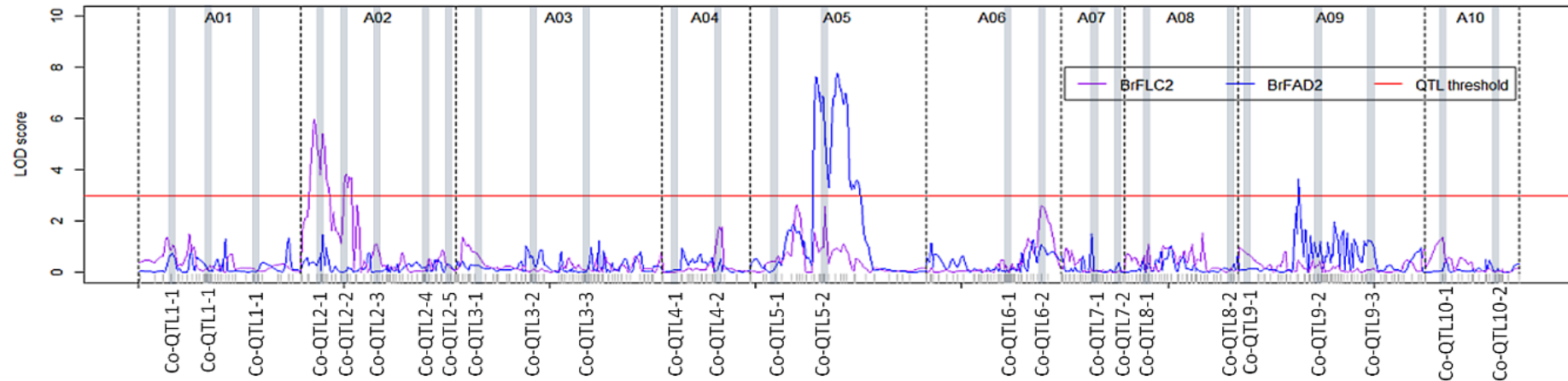


Figure 4: eQTL profiles for *BrFLC2* (a *Brassica* homologue of *FLC2* of *A. thaliana*) and *BrFAD2* (a *Brassica* homologue of *FAD2* of *A. thaliana*) measured in developing *B. rapa* seeds (28 days after pollination) across ten linkage groups. The y-axis represents the LOD score, the x-axis represents the 10 linkage groups separated by dotted lines; the QTL significance threshold is indicated by a red coloured solid line at LOD score 3.

Table 4: Summary of expression QTLs (eQTLs) of *BrFLC2* and *BrFAD2* genes identified using interval mapping (IM) and multiple QTL mapping (MQM) in this DH population.

Gene Name	<i>A. thaliana</i> orthologs	Peak marker	Linkage group	Peak marker		2-LOD support interval		% Explained variation	Total variation explained
				Position (cM)	LOD	Lower position	Upper position		
<i>BrFLC2</i>	<i>FLC2</i>	BRH04D11flc2	2	17.4	5.41	6.5	20.8	19.8	33.8
		BrPIP1b	2	37.9	3.69	33.1	42.7	14.0	
<i>BrFAD2</i>	<i>FAD2</i>	Myb2HaeIIIM-605.3	5	78.2	7.1	56.1	89.1	22.0	32.5
		BrFRY1P1b	9	51.5	3.64	48.1	57.6	10.5	

Phenotypic variation, correlation and heritabilities of the traits

The variability of root and shoot length in this population increases over the growing days under both conditions (Figure 1; Table 2; Supplementary Figure S2); however, the heritability remained similar (Table 3). The transgressive segregation observed in all traits suggests a quantitative and polygenic inheritance of these traits, requiring a QTL mapping approach to characterize the genetics of seed quality and seedling vigour in *B. rapa*. Root and shoot lengths form two separate clusters (negative correlation), over differences in treatments and seed batches from different years (Supplementary Figure S3) suggesting differences in regulation of root and shoot growth. However, high negative correlations between root and shoot lengths suggest partly shared regulation (Supplementary Figure S3).

Narrow-sense heritabilities were calculated for seedling vigour traits. In general, the heritability was lower for seedling vigour traits using 2010 seed batches (0.4 to 0.6) than for seeds of the 2011 batches (0.5 to 0.7) (Table 3). In 2011, the DH lines were sown in staggered fashion to synchronize flowering as much as possible and to minimize differing environmental influences during seed development. It is likely that this caused the higher heritabilities of most traits in 2011 than in 2010.

Major QTL hotspots for seed germination and seedling vigour

We identified major hotspots for seed germination and seedling vigour related traits on A02, A03, A05, A08 and A09. QTLs for T10 and T50 were found on hotspot regions on A02 (*Co-QTL2-1* and -2); at these loci, the pak choi allele is associated with later germination onset as well as a decrease in the germination speed (Figure 2B). It cannot be excluded that *Co-QTL2-1* and -2 are in fact a single QTL. Additional QTL regions on A01 (*Co-QTL1-3*) and A10 (*Co-QTL10-1*) were mainly associated with rate of germination (T50) and AUC; at *Co-QTL1-3* the pak choi allele was associated with higher T50 (time to reach 50% germination) and lower AUC, while at *Co-QTL10-1* the yellow sarson allele associated with higher T50 and the pak choi allele with lower AUC. The fact that at different loci both parents contribute positive alleles is another illustration of the polygenic transgressive nature of the inheritance of these traits.

The combination of QTLs from A01, A02 and A10 can increase the onset (T10) and rate or speed (T50) of seed germination in *B. rapa*. Interestingly, QTLs for flowering time, both in 2010 (peak LOD score 13.4, explained variance 38.0%) and 2011 (LOD 14.8, explained variance 40.9%) mapped to the *Co-QTL2-1* and -2 regions on A02, which could point to pleiotropy or linkage of QTLs for flowering time and seed germination. In a recent paper, it has been described that flowering time regulatory genes can pleiotropically or directly influence multiple agronomic traits, like the number and size of seeds, seedling vigour, biomass and resistance/tolerance to biotic or abiotic stress (Quijada *et al.*, 2006; Chen *et al.*, 2007; Ni *et al.*, 2009; Chianga *et al.*, 2009; Basunanda *et al.*, 2010; Li *et al.*, 2010), which likely put these genes under selection during crop breeding. The gene *BrFLC2* also maps to this *Co-QTL2-1* region (Xiao *et al.*, 2013) and the expression QTL for the *BrFLC2* gene both in leaves of six week old plants and in 28 days developing seeds co-localizes with this region (Figure 2B; Table 4). Xiao *et al.*, (2013) identified *BrFLC2* as a major regulator of

flowering time, using the same DH population, and reported the allelic variation between the *BrFLC2* alleles of the two parents yellow sarson and pak choi; a deletion of 56 bp at the exon 4 (12 bp) and intron 4 (44 bp) junction in yellow sarson rendered the gene non-functional; the pak choi allele does not have this deletion. In the related species *A. thaliana*, Chiang *et al.*, (2009) reported a pleiotropic effect of *FLC* (a homologue of *BrFLC2*) on temperature-dependent germination through additional genes *FT*, *SOC1* and *AP1* in the flowering time pathway in *A. thaliana*. They also reported the sharing of pathways by flowering time and seed germination, and showed that *FLC* regulates the germination through the abscisic acid catabolic pathway (ABA degradation) and gibberellin biosynthetic pathway in seeds.

An alternative explanation for the co-localization of *BrFLC2* with QTLs for seed germination could be a major regulatory role of this *FLC2* earliness gene in developmental processes. The possibility of confounding effects of two major loci involved in earliness was reported in *A. thaliana* and potato. In the Landsberg erecta x Cape Verde Islands a RIL population of *A. thaliana*, QTLs for many developmental traits were co-located on the *ERECTA* locus (Stinchcombe *et al.*, 2009). Similarly, in a diploid population of potato, the many QTLs were co-located on the *EARLINESS* locus (Hurtado-Lopez, 2012; Kloosterman *et al.*, 2012). Further study is needed to deconfound the causal relationships of the *BrFLC2* with seed germination parameters in *B. rapa*.

Across 24 traits, many QTLs co-localized on *Co-QTL3-3* on A03, *Co-QTL5-1* and *-2* on A05, *Co-QTL8-1* on A08 and *Co-QTL9-2* and *-3* on A09 (Figure 2B). Corroborating the finding that there are high correlations between the time points, we found co-localized QTLs for root and shoot lengths measured repeatedly in time. For most traits, two or more than two QTLs were detected and most of the QTLs for the 2011 seed batch had higher explained variance (range: 7.1 to 24.3%; mode value: 14.2%) rather than for the 2010 seed batch (range: 7.8 to 22.6%; mode value: 11.1%) (Supplementary Table S3). This again reflects the higher explained genetic variation when seeds ripened synchronously, as was the case in 2011. Several putative QTLs co-localized with significant QTLs for correlated traits. QTLs found at multiple time points increase the reliability of these QTLs. Increasing the power of QTL detection by either enlarging the population size or increasing the precision of phenotyping, possibly, could be used to confirm additional candidate QTLs reported in this study.

QTLs specific to salt stress conditions

QTLs for uniformity (U7525 in 2010 and 2011), AUC, Gmax (in 2010) and rate (in 2011) of germination under salt stress co-localized on *Co-QTL5-1* and *Co-QTL5-2* on A05 and for both loci the yellow sarson allele has a positive effect on maximum germination (Gmax) potential, AUC and rate of germination, but a negative effect on uniformity under salt stress. On hotspots *Co-QTL5-1* and *-2* on A05, also QTLs for root and shoot lengths and shoot weight under salt stress and thousand-seed weight were mapped. QTLs for relative salt tolerance (RelST) parameters also mainly mapped to *Co-QTL5-1* and *-2* (Figure 2B). Finally, yellow sarson alleles at these salt stress specific QTL regions contribute to a larger seed size and higher thousand-seed weight than pak choi alleles, which could be in support of a higher maximum germination and faster germination

rate in yellow sarson. As marker density was rather low with on average 7-10 cM distance between two markers, more markers and more recombinants are needed to conclude whether the regions actually represent a single or two closely linked QTLs. QTLs at these two hotspots were in coupling phase for all the traits (with the yellow sarson allele having a favourable effect) supporting that this is a single QTL hotspot.

The *BrFAD2* gene, a key gene responsible for biosynthesis of poly-unsaturated fatty acids, is located inside this *Co-QTL5-2* region. Two eQTLs were mapped for *BrFAD2*: a *cis*-eQTL across the region of *Co-QTL5-1* and -2 on A05 and a *trans*-eQTL co-locating with *Co-QTL9-2* on A09 (Figure 2; Figure 4). This QTL region on A05 is the major locus with QTLs for seed germination and seedling vigour under salt stress (Figure 2B), suggesting a role of *BrFAD2* in regulating germination and seedling vigour under salt stress. Besides the role of *FAD2* gene in fatty acid desaturation, Wang *et al.*, (2010) reported that the up-regulation of the *FAD2* gene enhanced seed germination and hypocotyl length in their study on *FAD2*-transgenic and non-transgenic lines of the closely related species *B. napus*. In another study on the comparison of a *fad2* mutant of *A. thaliana* with the wild type, a functional role of *FAD2* was reported in increasing salt tolerance during seed germination and early seedling growth (Zhang *et al.*, 2012). High homology of coding sequences (86%) was found among the homologs of a *FAD2* gene in *A. thaliana*, *B. rapa* and *B. napus*. *FAD2* extrudes Na^+ out of the cell and compartmentalizes it into the vacuolar membrane using Na^+/H^+ antiporters (NHXs) and thus maintains ion homeostasis. Thus, our results in *B. rapa* are in good agreement with the findings on the roles of *FAD2* in *A. thaliana* and *B. napus* that the *BrFAD2* gene is a candidate gene in *B. rapa* to improve of seed germination and early seedling vigour under salt stress.

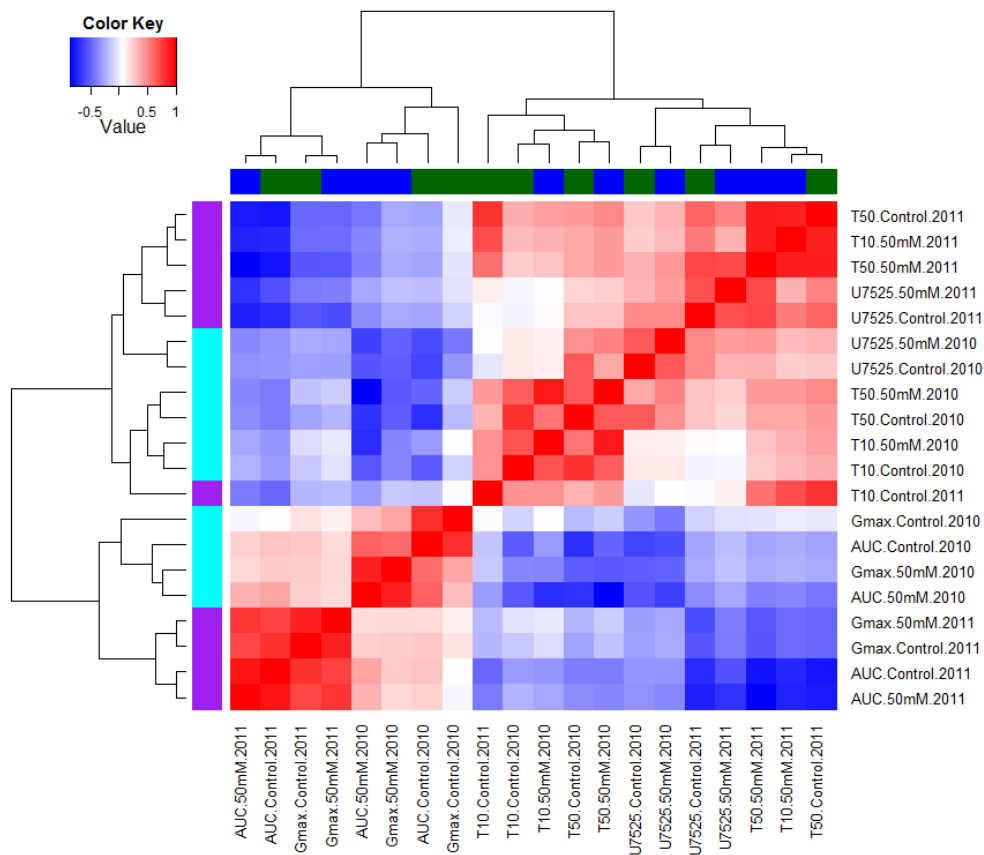
Epistatic interaction between QTLs at *BrFLC2* and *BrFAD2*

Epistatic interactions between individual genes are prevalent to account for variation in quantitative traits (Jannink and Jansen, 2001). For seed germination and seedling vigour related traits, several studies reported epistatic interactions between genes in *Arabidopsis* (Galpaz and Reymond, 2010; Bouteillé *et al.*, 2012), tomato (Kazmi *et al.*, 2012; Khan *et al.*, 2012), rice (Wang *et al.*, 2012), *B. napus* (Yang *et al.*, 2012) as well as other crops. Interactions with other QTL regions were observed for *Co-QTL2-2*, *Co-QTL6-2* and *Co-QTL10-1* indicating that not only main effects of these QTLs but also their epistatic interactions are important for fitness traits like seed germination and seedling vigour (Figure 3). The *Co-QTL2-2* locus showed clear interactions with *Co-QTL5-1* and -2, which likely represent a single QTL hotspot. The *Co-QTL2-2* region co-locates with *BrFLC2*, the *Co-QTL5-2* locus co-locates with *BrFAD2*. This suggests that, in addition to their main effects, an epistatic interaction of these two loci may play an important role (explained up to 17.6% of total phenotypic variation) in the genetic regulatory network of seed germination and seedling vigour in *B. rapa* under salt stress. A further understanding of interactions with other QTL regions will help to explore the complex genetic architecture of seed germination and seedling vigour in *B. rapa*.

Acknowledgements

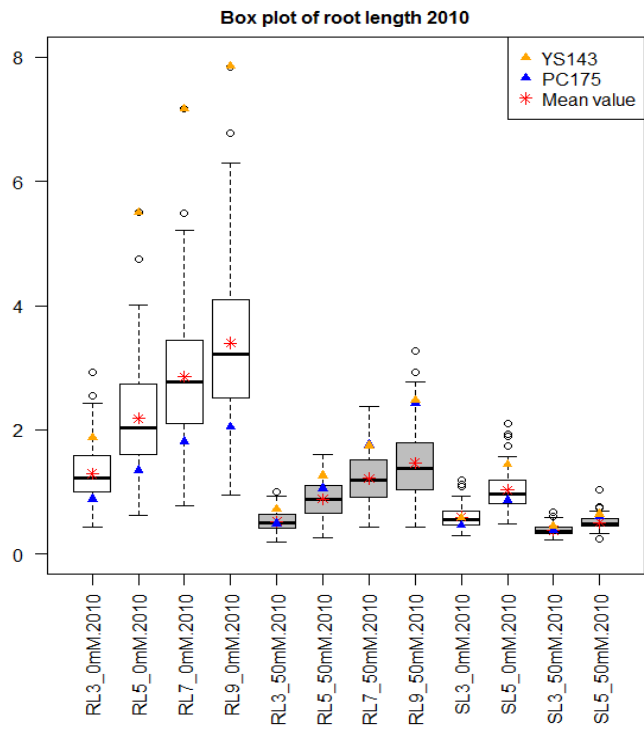
We thank the members of the Unifarm facility of Wageningen University and Research Centre (WUR) for taking care of the plants and all the necessary support. Authors highly appreciate the cooperations from Dr. Ronny Joosen from the SeedLab, Laboratory of Plant Physiology, Wageningen UR, The Netherlands in fixing camera setup for seedling assay experiments. We also would like to thank Natalia Carreno Quintero, Laboratory of Plant Physiology and Plant Breeding for her kind help with the figures. Finally, the authors are thankful to the Centre for BioSystems Genomics (CBSG), Netherlands, which is a part of the Netherlands Genomics Initiative (NGI) for partial financial support for this study.

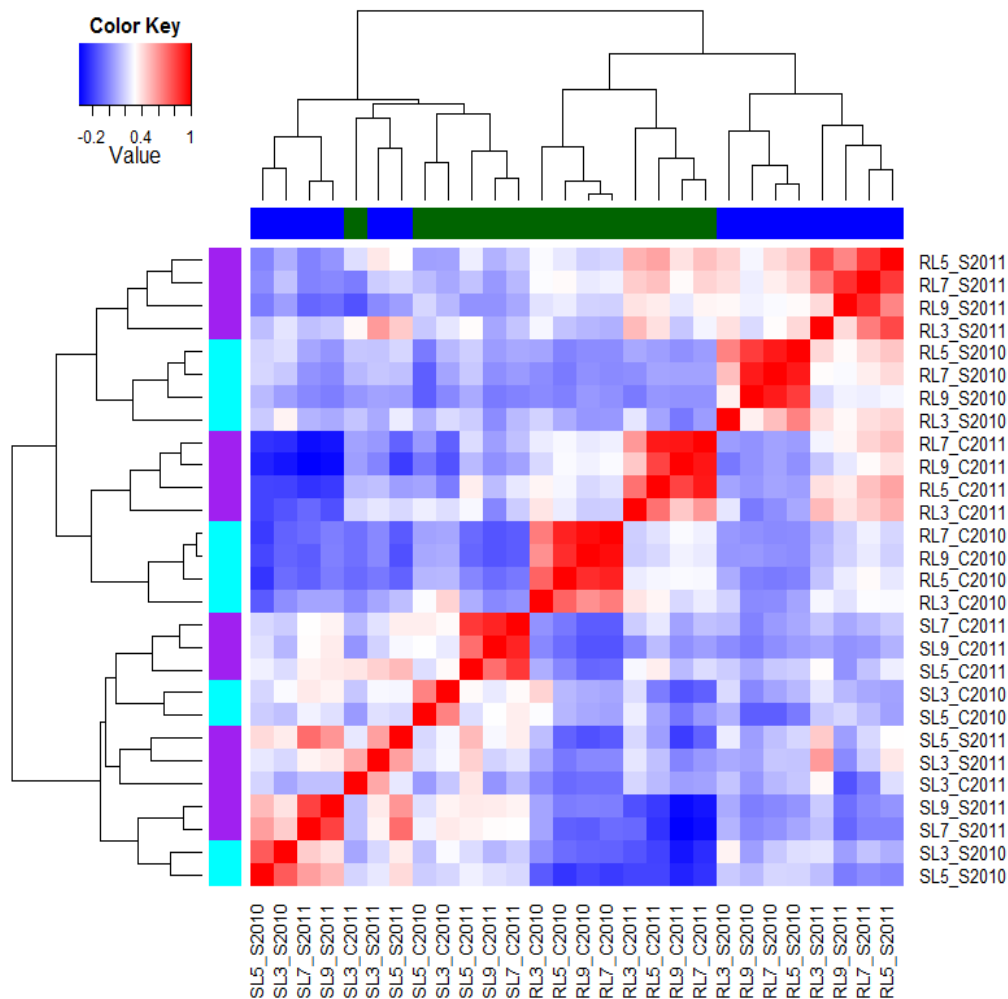
Supplemental Data



Supplementary Figure S1: Heatmap of Pearson correlation coefficients and hierarchical cluster analysis of seed germination parameters under non-stress (control) and 50 mM NaCl salt stress conditions. The colour bar on the left side indicates the two different seed batches: 2010 and 2011, and the colour bars on the top the two different treatments: non-stress (control) and 50 mM NaCl salt. The gradient from red to blue colour indicates the degree of positive or negative correlation, respectively.

Supplementary Figure S2: Box plot showing the distribution of root length (RL) and shoot length (SL) at different days after germination (DAG) under non-stress (0 mM NaCl) and salt stress (50 mM NaCl “S”) for the 2010 seed batch. The shaded colour of the boxes indicates the treatments: white for non-stress and grey for salt stress. The y-axis indicates root- and shoot length (in cm). The x-axis label is the combination of RL at 1, 3, 5, 7 and 9 DAG or SL at 1, 3 and 5 DAG under non-stress and salt stress conditions of the 2010 seed batch.





Supplementary Figure S3: Heatmap of Pearson correlation coefficients and hierarchical cluster analysis of root length (RL) and shoot length (SL) measured at 3, 5, 7 and 9 days after germination (DAG) at non-stress (control; coded as “C”) and 50 mM NaCl salt stress (“S”) conditions. The colour bar on the left side indicates the two different seed batches: 2010 and 2011, the colour bars on the top the two different treatments: non-stress (control) and 50 mM NaCl salt. The gradient from red to blue colour indicates the degree of positive or negative correlation, respectively.

Supplementary Table S1: List of traits related to seed weight, seed germination, seedling growth, seedling weight and salt tolerance parameters with their codes in combination with treatment, DAG (days after germination) and year of the seed batch.

SN	Year	Treatment	DAG	traits	Trait code	Measurement unit
A. Thousand seed weight						
1	2009	-	-	thousand seed	seed.weight.2009	g
2	2010	-	-	weight	seed.weight.2010	g
3	2011	-	-		seed.weight.2011	g
B. Flowering Time						
1	2010	-	-	Flowering time	flowering.2010	days
2	2011	-	-		flowering.2011	days
C. Seed germination						
1	2010	Control		T10	T10_Control_2010	hr
2				T50	T50_Control_2010	hr
3				Gmax	Gmax_Control_2010	%

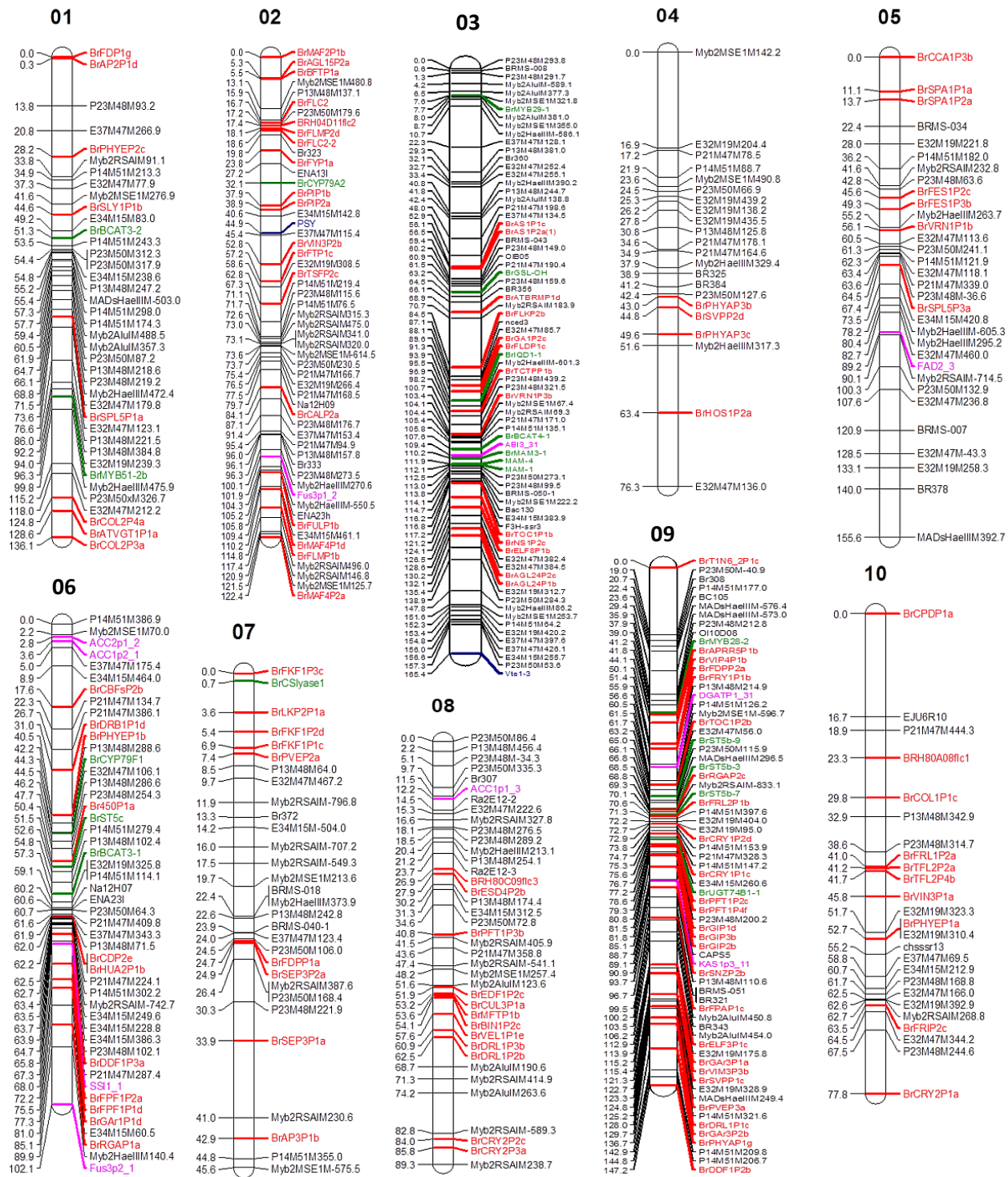
4				U7525	U7525_Control_2010	hr
5				AUC	AUC_Control_2010	-
6		50mM NaCl (salt stress)		T10	T10_50mM_2010	hr
7				T50	T50_50mM_2010	hr
8				Gmax	Gmax_50mM_2010	%
9				U7525	U7525_50mM_2010	hr
10				AUC	AUC_50mM_2010	-
11	2011	Control		T10	T10_Control_2011	hr
12				T50	T50_Control_2011	hr
13				Gmax	Gmax_Control_2011	%
14				U7525	U7525_Control_2011	hr
15				AUC	AUC_Control_2011	-
16		50mM NaCl		T10	T10_50mM_2011	hr
17				T50	T50_50mM_2011	hr
18				Gmax	Gmax_50mM_2011	%
19				U7525	U7525_50mM_2011	hr
20				AUC	AUC_50mM_2011	-

D. Seedling growth (Root length, shoot length, root weight, shoot weight)

1	2010	Control	1	root length	RL1_Control.2010	cm
2			3		RL3_Control.2010	cm
3			5		RL5_Control.2010	cm
4			7		RL7_Control.2010	cm
5			9		RL9_Control.2010	cm
6		50mM NaCl	1		RL1_50mM.2010	cm
7			3		RL3_50mM.2010	cm
8			5		RL5_50mM.2010	cm
9			7		RL7_50mM.2010	cm
10			9		RL9_50mM.2010	cm
11		Control	1	shoot length	SL1_Control.2010	cm
12			3		SL3_Control.2010	cm
13			5		SL5_Control.2010	cm
14		50mM NaCl	1		SL1_50mM.2010	cm
15			3		SL3_50mM.2010	cm
16			5		SL5_50mM.2010	cm
17		Control		Root fresh weight	RFw_Control.2010	mg
18				Root dry weight	RDw_Control.2010	mg
19		50mM NaCl		Root fresh weight	RFw_50mM.2010	mg
20				Root dry weight	RDw_50mM.2010	mg
21		Control		Shoot fresh weight	SFw_Control.2010	mg
22				Shoot dry weight	SDw_Control.2010	mg
23		50mM NaCl		Shoot fresh weight	SFw_50mM.2010	mg
24				Shoot dry weight	SDw_50mM.2010	mg
25	2011	Control	3	root length	RL3_Control.2011	cm
26			5		RL5_Control.2011	cm
27			7		RL7_Control.2011	cm
28			9		RL9_Control.2011	cm
29		50mM NaCl	3		RL3_50mM.2011	cm
30			5		RL5_50mM.2011	cm
31			7		RL7_50mM.2011	cm
32			9		RL9_50mM.2011	cm

33			3		SL3_Control.2011	cm
34			5		SL5_Control.2011	cm
35		Control	7		SL7_Control.2011	cm
36			9		SL9_Control.2011	cm
37			3	shoot length	SL3_50mM.2011	cm
38		50mM	5		SL5_50mM.2011	cm
39		NaCl	7		SL7_50mM.2011	cm
40			9		SL9_50mM.2011	cm
41		Control		Root fresh weight	RFw_Control.2011	mg
42				Root dry weight	RDw_Control.2011	mg
43		50mM		Root fresh weight	RFw_50mM.2011	mg
44		NaCl		Root dry weight	RDw_50mM.2011	mg
45		Control		Shoot fresh weight	SFw_Control.2011	mg
46				Shoot dry weight	SDw_Control.2011	mg
47		50mM		Shoot fresh weight	SFw_50mM.2011	mg
48		NaCl		Shoot dry weight	SDw_50mM.2011	mg
E. Relative Satl Tolerance (RelST) and Salt Tolerance Index (STI)						
1			1		STI1.RL_C50.2010	-
2			3		STI3.RL_C50.2010	-
3			5	STI of root length	STI5.RL_C50.2010	-
4			7		STI7.RL_C50.2010	-
5			9		STI9.RL_C50.2010	-
6			1		RelST1.RL_C50.2010	-
7			3	Relative salt	RelST3.RL_C50.2010	-
8	2010		5	tolerance of root	RelST5.RL_C50.2010	-
9			7	length	RelST7.RL_C50.2010	-
10			9		RelST9.RL_C50.2010	-
11			1		STI1_SL.C50.2010	-
12			3	STI of shoot	STI3_SL.C50.2010	-
13			5	length	STI5_SL.C50.2010	-
14		Control	1	Relative salt	RelST1_SL.C50.2010	-
15		versus	3	tolerance of shoot	RelST3_SL.C50.2010	-
16		50mM	5	length	RelST5_SL.C50.2010	-
17		NaCl	3		STI3_RL.C50.2011	-
18			5	STI of root length	STI5_RL.C50.2011	-
19			7		STI7_RL.C50.2011	-
20			9		STI9_RL.C50.2011	-
21			3		RelST3_RL.C50.2011	-
22			5	Relative salt	RelST5_RL.C50.2011	-
23			7	tolerance of root	RelST7_RL.C50.2011	-
24			9	length	RelST9_RL.C50.2011	-
25			3		STI3_SL_C50.2011	-
26			5	STI of shoot	STI5_SL_C50.2011	-
27			7	length	STI7_SL.C50.2011	-
28			9		STI9_SL.C50.2011	-
29			3		RelST3_SL.C50.2011	-
30			5	Relative salt	RelST7_SL.C50.2011	-
31			7	tolerance of shoot	RelST5_SL.C50.2011	-
32			9	length	RelST9_SL.C50.2011	-

Note: hr – hour; g – gram; cm – centimetre; mg – milligram.



Supplementary Figure S4: An integrated linkage map (a slightly modified version of Xiao et al., 2013) with map positions of 435 molecular markers on ten linkage groups for this *B. rapa* doubled haploid population. Gene-targeted markers were highlighted with different colours: red colour indicates markers for flowering time pathway genes, green colour for glucosinolate pathway genes, magenta for fatty acid biosynthesis pathway genes and blue for tocopherol and carotenoids pathway genes. Numbers at the top of each map indicate linkage groups.

Supplementary Table S2: Summary of significant QTLs identified for seed germination parameters and seed weight (thousand seed weight) in the *B. rapa* DH population from a cross of yellow sarson (YS143) and pak choi (PC175).

SN	Traits	QTL name	Linkage group	Peak marker	QTL position	LOD score	Confidence interval		QTL explained variance	Total explained variance ¹	Allele from	Additive effect
1	T10_Control_2010	q-0mM-2010-T10.1	5	E34M15M420.8	73.5	3.1	71.4	76.5	7.6	21.6	PC	1.2
		q-0mM-2010-T10.2	9	Myb2AluIM450.8	100.2	5.0	100.2	104.5	14.0		YS	1.8
2	T50_Control_2010	q-0mM-2010-T50.1	1	P13M48M384.8	92.2	3.2	83.6	98.3	9.5	38.9	PC	1.7
		q-0mM-2010-T50.2	10	BrCPDP1a	0.0	4.9	0.0	13.0	15.4		YS	2.1
		q-0mM-2010-T50.3	2	BRH04D11flc2	17.4	2.3	6.5	26.8	6.7		YS	1.4
		q-0mM-2010-T50.4	6	Myb2HaeIIIM140.4	89.9	2.5	75.2	96.9	7.3		PC	1.5
3	Gmax_Control_2010	-	-	-	-	-	-	-	-	-	-	-
4	U7525_Control_2010	q-0mM-2010-U7525.1	5	BRMS-034	22.4	2.3	15.7	37.2	6.5	22.2	YS	1.0
		q-0mM-2010-U7525.2	6	BrFPF1P1d	75.5	2.9	73.2	83.0	8.2		PC	1.1
		q-0mM-2010-U7525.3	9	MADsHaeIIIM296.5	66.8	2.6	61.7	75.3	7.5		YS	1.0
5	AUC_Control_2010	q-0mM-2010-AUC.1	6	BrFPF1P1d	75.5	2.9	73.2	80.3	9.1	18.4	YS	3.1
		q-0mM-2010-AUC.2	10	BrCPDP1a	0.0	2.8	0.0	7.0	9.3		PC	3.2
6	T10_50mM_2010	q-50mM-2010-T10.1	2	ENA13l	27.2	5.3	24.8	36.1	15.7	20.3	YS	2.1
7	T50_50mM_2010	q-50mM-2010-T50.1	2	BrFLMP2d	18.1	3.0	8.5	30.2	9.3	29.1	YS	1.9
		q-50mM-2010-T50.2	10	BrCPDP1a	0.0	3.9	0.0	9.0	12.3		YS	2.2
		q-50mM-2010-T50.3	1	Myb2HaeIIIM475.9	99.8	2.5	98.3	107.8	7.5		PC	1.8
8	Gmax_50mM_2010	q-50mM-2010-Gmax.1	9	Myb2AluIM450.8	100.2	3.1	100.2	108.2	10.7	19.7	PC	3.3
		q-50mM-2010-Gmax.2	5	BRMS-034	22.4	2.7	16.7	25.4	9.0		PC	2.8
9	U7525_50mM_2010	q-50mM-2010-U7525.1	2	P13M48M157.8	96.0	4.3	87.1	98.3	11.0	34.2	PC	1.2
		q-50mM-2010-U7525.2	5	P14M51M182.0	36.2	3.1	19.7	46.6	7.5		YS	0.9
		q-50mM-2010-U7525.3	6	BrFPF1P1d	75.5	3.6	66.8	76.5	8.9		PC	1.0
		q-50mM-2010-U7525.4	5	E34M15M420.8	73.5	2.8	56.1	85.7	6.8		YS	0.9
10	AUC_50mM_2010	q-50mM-2010-AUC.1	5	BRMS-034	22.4	4.2	15.7	25.4	11.8	36.9	PC	3.8
		q-50mM-2010-AUC.2	9	Myb2AluIM454.0	106.2	3.9	100.2	110.2	11.0		PC	4.0
		q-50mM-2010-AUC.3	2	BrPIP1b	37.9	2.3	22.8	48.4	6.1		PC	2.7
		q-50mM-2010-AUC.4	4	Myb2MSE1M142.2	0.0	3.0	0.0	7.0	8.0		YS	3.2
11	T10_Control_2011	q-0mM-2011-T10.1	2	BrFYP1a	23.8	5.9	17.2	41.6	21.9	39.4	YS	2.0
		q-0mM-2011-T10.2	2	E32M19M308.5	58.6	4.5	58.2	69.3	17.5		YS	1.8
12	T50_Control_2011	q-0mM-2011-T50.1	2	BRH04D11flc2	17.4	3.4	17.2	26.8	11.3	40.3	YS	2.4
		q-0mM-2011-T50.2	3	E34M15M383.9	116.2	3.1	97.9	116.8	10.4		PC	2.3
		q-0mM-2011-T10.3	10	BrCPDP1a	0.0	3.1	0.0	20.9	10.3		YS	2.4
		q-0mM-2011-T50.3	1	Myb2HaeIIIM475.9	99.8	2.5	97.3	110.8	8.3		PC	2.1
13	Gmax_Control_2011	-	-	-	-	-	-	-	-	-	-	-
14	U7525_Control_2011	q-0mM-2011-U7525.1	3	BRMS-043	59.4	4.2	57.5	66.1	17.1	17.1	PC	2.2
15	AUC_Control_2011	q-0mM-2011-AUC.1	2	E34M15M142.8	40.6	2.0	15.1	42.6	7.5	13.6	PC	3.2
		q-0mM-2011-AUC.2	10	BrCPDP1a	0.0	2.0	0.0	17.7	6.1		PC	3.0
16	T10_50mM_2011	q-50mM-2011-T10.1	2	BrPIP1b	37.9	7.1	25.8	41.6	27.2	45.5	YS	3.0
		q-50mM-2011-T10.2	1	Myb2HaeIIIM475.9	99.8	2.2	97.3	106.8	7.4		PC	1.5
		q-50mM-2011-T10.3	2	E32M19M308.5	58.6	2.6	57.2	64.8	10.9		YS	1.9
17	T50_50mM_2011	q-50mM-2011-T50.1	2	BrPIP1b	37.9	6.9	25.8	41.6	26.1	56.2	YS	4.4
		q-50mM-2011-T50.2	1	Myb2HaeIIIM475.9	99.8	2.3	80.6	108.8	7.8		PC	2.3
		q-50mM-2011-T50.3	5	P14M51M182.0	36.2	2.7	32.0	41.2	9.8		YS	2.6
		q-50mM-2011-T50.4	5	E32M47M113.6	60.5	3.6	56.1	72.4	12.5		YS	3.1
18	Gmax_50mM_2011	-	-	-	-	-	-	-	-	-	-	-
19	U7525_50mM_2011	q-50mM-2011-U7525.1	3	P23M48M159.6	64.5	3.3	58.5	67.1	12.7	36.6	PC	2.2
		q-50mM-2011-U7525.2	9	P13M48M110.6	93.7	3.3	90.1	97.7	12.8		PC	2.1
		q-50mM-2011-U7525.3	5	BrSPL5P3a	67.4	2.9	65.5	85.7	11.1		YS	2.1
20	AUC_50mM_2011	q-50mM-2011-AUC.1	1	Myb2HaeIIIM475.9	99.8	3.3	96.0	112.8	13.0	13	YS	5.0
21	Thousand seed weight_2010	q-testWt-2010.1	5	Myb2RSAlM232.8	41.6	4.1	37.2	43.8	12.9	29	YS	0.3
		q-testWt-2010.2	5	P23M48M-36.6	64.5	5.2	52.3	75.5	16.1		YS	0.3
22	Thousand seed weight_2011	q-testWt-2011.1	5	BrSPL5P3a	67.4	3.6	65.5	71.4	8.3	8.3	YS	0.3
23	Flowering time_2010	q-floweringTime-2010.1	2	BrFLMP2d	18.1	13.4	6.5	24.8	38.0	51.2	YS	4.3
		q-floweringTime-2010.2	2	BrPIP2a	38.9	4.0	33.1	41.6	13.2		YS	2.6
24	Flowering time_2011	q-floweringTime-2011.1	2	P13M48M137.1	15.9	14.8	7.5	25.8	40.9	54.3	YS	5.2
		q-floweringTime-2011.2	2	BrPIP2a	38.9	4.1	34.1	42.6	13.4		YS	3.1

Note: 1 - Total explained variance of significant QTL (LOD >3) and candidate QTL (LOD >2), LL - lower limit and UL - upper limit.

Supplementary Table S3: Summary of significant QTLs identified for root length (RL), shoot length (SL), root weight and shoot weight in the *B. rapa* DH population from a cross of yellow sarson (YS143) and pak choi (PC175).

SN	Year	Treat- ments	Trait	QTL name	Linkage group	Peak marker	QTL position	LOD score	Confidence interval		QTL explained variance	Total explained variance ¹	Herita- bility	Allele from	Allelic effect
									LL	UL					
1	2010	Control	RL3_Control.2010	q-0mM-2010-RL3.2	5	BrSPL5P3a	67.4	3.3	65.5	70.4	9.8	41.8	0.7	PC	0.14
	2010	Control		q-0mM-2010-RL3.3	8	BrCRY2P3a	85.8	3.7	75.2	89.3	11.3			PC	0.16
	2010	Control		q-0mM-2010-RL3.4	9	BrFPAP1c	99.5	4.5	90.9	104.5	13			PC	0.18
	2010	Control		q-0mM-2010-RL3.1	5	E32M19M221.8	28.0	2.5	15.7	36.0	7.7			PC	0.13
2	2010	Control	RL5_Control.2010	q-0mM-2010-RL5.1	4	E32M19M204.4	16.9	3.4	7.0	16.9	10.6	23.1	0.7	YS	0.26
	2010	Control		q-0mM-2010-RL5.2	8	BrCRY2P3a	85.8	4.0	76.2	89.3	12.5			PC	0.28
3	2010	Control	RL7_Control.2010	q-0mM-2010-RL7.1	4	E32M19M204.4	16.9	4.0	8.0	27.8	13.4	23.4	0.7	YS	0.37
	2010	Control		q-0mM-2010-RL7.2	8	BrCRY2P3a	85.8	3.0	83.8	88.8	10			PC	0.36
4	2010	Control	RL9_Control.2010	q-0mM-2010-RL9.1	8	BrCRY2P3a	85.8	4.4	83.8	88.8	15.8	15.8	0.7	PC	0.54
5	2010	Salt	RL3_50mM.2010	q-50mM-2010-RL3.1	5	BrSPA1P1a	11.1	3.1	3.0	21.7	8.7	35.7	0.6	YS	0.05
	2010	Salt		q-50mM-2010-RL3.4	9	BrVIP4P1b	44.1	4.9	41.8	71.3	14.3			PC	0.06
	2010	Salt		q-50mM-2010-RL3.2	6	P21M47M386.1	26.7	2.3	24.3	39.0	6.2			YS	0.04
	2010	Salt		q-50mM-2010-RL3.3	7	Myb2RSAIM230.6	41.0	2.4	37.9	45.6	6.5			YS	0.04
6	2010	Salt	RL5_50mM.2010	q-50mM-2010-RL5.1	5	BrSPA1P1a	11.1	3.2	1.0	18.7	9.7	24.4	0.7	YS	0.09
	2010	Salt		q-50mM-2010-RL5.2	6	P13M48M71.5	62.0	2.3	53.6	74.2	7			YS	0.08
	2010	Salt		q-50mM-2010-RL5.3	7	Myb2RSAIM230.6	41.0	2.5	37.9	43.9	7.7			YS	0.09
7	2010	Salt	RL7_50mM.2010	q-50mM-2010-RL7.3	7	Myb2RSAIM230.6	41.0	3.6	37.9	45.6	10.8	30.5	0.7	YS	0.16
	2010	Salt		q-50mM-2010-RL7.1	5	BrSPA1P1a	11.1	2.2	0.0	17.7	6.5			YS	0.11
	2010	Salt		q-50mM-2010-RL7.2	6	P13M48M71.5	62.0	2.8	58.3	73.2	8.2			YS	0.12
	2010	Salt		q-50mM-2010-RL7.4	10	BrCOL1P1c	29.8	2.4	27.3	33.9	5			PC	0.11
8	2010	Control	SL5_Control.2010	q-0mM-2010-SL5.1	1	Myb2HaeIIIM475.9	99.8	2.3	97.3	105.8	5.7	12.9	0.7	YS	0.07
	2010	Control		q-0mM-2010-SL5.2	8	P13M48M456.4	2.2	2.9	0.0	10.7	7.2			YS	0.08
9	2010	Salt	SL3_50mM.2010	q-50mM-2010-SL3.3	5	E32M19M221.8	28.0	3.4	15.7	37.2	8.8	39.3	0.6	YS	0.02
	2010	Salt		q-50mM-2010-SL3.1	2	Myb2HaeIIIM270.6	100.1	2.1	92.4	101.1	5.4			YS	0.02
	2010	Salt		q-50mM-2010-SL3.2	3	E37M47M134.5	52.9	2.1	5.2	66.1	5.3			PC	0.02
	2010	Salt		q-50mM-2010-SL3.4	8	P23M50M335.3	9.7	3.0	5.1	20.4	7.7			YS	0.02
	2010	Salt		q-50mM-2010-SL3.5	10	P23M48M314.7	38.6	2.2	17.7	43.7	5.6			YS	0.02
	2010	Salt		q-50mM-2010-SL3.6	10	P23M48M168.8	61.7	2.5	57.2	77.8	6.5			YS	0.02
10	2010	Salt	SL5_50mM.2010	q-50mM-2010-SL5.1	2	Myb2HaeIIIM270.6	100.1	3.1	87.1	101.1	8.1	30.1	0.6	YS	0.03
	2010	Salt		q-50mM-2010-SL5.4	8	E34M15M312.5	31.3	4.1	13.2	36.6	11.1			YS	0.04
	2010	Salt		q-50mM-2010-SL5.2	3	Br360	32.1	2.1	17.7	38.4	5.4			PC	0.03
	2010	Salt		q-50mM-2010-SL5.3	4	BrHOS1P2a	63.4	2.1	52.6	76.3	5.5			YS	0.03
11	2010	Control	RFw_Control.2010	q-0mM-2010-RF.1	3	P23M48M291.7	1.3	3.4	0.0	12.7	11.1	46.3	-	PC	16.44
	2010	Control		q-0mM-2010-RF.2	3	E32M47M85.7	88.1	5.0	72.7	114.7	15.6			YS	20.72
	2010	Control		q-0mM-2010-RF.3	6	ENA23I	60.6	4.5	55.8	73.2	12.7			YS	17.62
	2010	Control		q-0mM-2010-RF.4	9	Myb2AluIM454.0	106.2	2.6	102.2	112.2	6.9			PC	14.12
12	2010	Control	SFw_Control.2010	q-0mM-2010-SF.1	3	E32M47M85.7	88.1	4.7	88.1	95.5	16.3	25.7	-	YS	56.75
	2010	Control		q-0mM-2010-SF.2	7	BRMS-040-1	23.9	3.1	22.6	31.3	9.4			PC	41.92
13	2010	Salt	SFw_50mM.2010	q-50mM-2010-SF.2	6	E34M15M464.0	8.9	3.0	7.0	11.9	9.8	17.1	-	PC	29.88
	2010	Salt		q-50mM-2010-SF.1	3	BrFLDP1c	91.3	2.3	88.1	107.6	7.3			YS	26.41
14	2010	Control	RDw_Control.2010	q-0mM-2010-RD.1	1	P14M51M213.3	34.9	2.2	32.2	40.3	7.7	7.7	-	YS	7.30
15	2010	Salt	RDw_50mM.2010	q-50mM-2010-RD.1	3	Myb2HaeIIIM-601.3	95.5	2.2	92.3	107.6	9.5	9.5	-	YS	0.65
16	2010	Control	SDw_Control.2010	q-0mM-2010-SD.1	3	E37M47M128.1	22.3	2.2	7.5	32.1	7.4	7.4	-	PC	5.98
17	2010	Salt	SDw_50mM.2010	q-50mM-2010-SD.3	7	Myb2HaeIIIM373.9	22.4	4.1	17.5	22.6	12.7	27.1	-	YS	4.29
	2010	Salt		q-50mM-2010-SD.1	1	BrAP2P1d	0.3	2.5	0.0	7.3	7.3			PC	3.30
	2010	Salt		q-50mM-2010-SD.2	5	E32M47M118.1	63.4	2.4	56.1	65.5	7.1			YS	2.95
18	2010	Salt	RL9_50mM.2010	-	-	-	-	-	-	-	-	-	0.7	-	-
19	2010	Control	SL3_Control.2010	-	-	-	-	-	-	-	-	-	0.7	-	-
20	2010	Salt	RFw_50mM.2010	-	-	-	-	-	-	-	-	-	-	-	-
21	2011	Control	SL3_Control.2011	q-0mM-2011-SL3.1	9	BrRGAP2c	68.8	2.8	66.0	75.3	14.7	14.7	0.8	YS	0.03
22	2011	Control	SL5_Control.2011	q-0mM-2011-SL5.2	8	P23M50M335.3	9.7	4.9	1.0	22.2	17.1	38.5	0.7	YS	0.06
	2011	Control		q-0mM-2011-SL5.3	10	chsssr13	55.2	3.7	53.7	57.2	12.6			YS	0.05
	2011	Control		q-0mM-2011-SL5.1	3	BrTOC1P1b	117.2	2.6	113.8	119.2	8.8			PC	0.04
23	2011	Control	SL7_Control.2011	q-0mM-2011-SL7.1	3	BrTOC1P1b	117.2	3.1	90.6	125.1	10.5	32.7	0.8	PC	0.08
	2011	Control		q-0mM-2011-SL7.2	8	P23M50M335.3	9.7	3.1	1.0	35.6	9.5			YS	0.07
	2011	Control		q-0mM-2011-SL7.3	10	chsssr13	55.2	4.0	53.7	57.2	12.7			YS	0.09
24	2011	Control	SL9_Control.2011	q-0mM-2011-SL9.2	3	P23M50M273.1	112.5	3.4	101.7	132.1	15	24.8	0.8	PC	0.13
	2011	Control		q-0mM-2011-SL9.1	1	P13M48M221.5	86.0	2.3	80.6	98.3	9.8			YS	0.11
25	2011	Salt	SL3_50mM.2011	q-50mM-2011-SL3.2	1	BrATVGT1P1a	128.6	5.8	126.8	136.1	14.4	36.9	0.7	PC	0.02
	2011	Salt		q-50mM-2011-SL3.3	4	Myb2HaeIIIM317.3	51.6	3.2	51.6	58.6	7.3			PC	0.02
	2011	Salt		q-50mM-2011-SL3.4	9	BrELF3P1c	112.9	3.0	110.2	117.4	8.9			PC	0.02
	2011	Salt		q-50mM-2011-SL3.1	1	P13M48M384.8	92.2	2.4	82.6	98.3	6.3			PC	0.02

Supplementary Table S3 (Continue): Summary of significant QTLs identified for root length (RL), shoot length (SL), root weight and shoot weight in the *B. rapa* DH population from a cross of yellow sarson (YS143) and pak choi (PC175).

26	2011	Salt	SL5_50mM.2011	q-50mM-2011-SL5.3	5	BrSPL5P3a	67.4	4.6	65.5	70.4	16	33.7	0.6	YS	0.03
	2011	Salt		q-50mM-2011-SL5.1	2	Br323	19.8	2.5	17.2	21.8	8.3			YS	0.02
	2011	Salt		q-50mM-2011-SL5.2	5	BRMS-034	22.4	2.6	15.7	38.2	9.4			YS	0.02
27	2011	Salt	SL7_50mM.2011	q-50mM-2011-SL7.1	2	BRH04D11flc2	17.4	3.1	11.5	21.8	10	66.1	0.8	YS	0.04
	2011	Salt		q-50mM-2011-SL7.2	2	BrPIP2a	38.9	3.4	34.1	42.6	11.1			YS	0.04
	2011	Salt		q-50mM-2011-SL7.4	5	BRMS-034	22.4	3.9	15.7	37.2	14.2			YS	0.04
	2011	Salt		q-50mM-2011-SL7.5	5	BrSPL5P3a	67.4	3.8	56.1	72.4	13.8			YS	0.05
	2011	Salt		q-50mM-2011-SL7.6	10	P21M47M444.3	18.9	3.3	17.7	39.6	10.7			YS	0.04
	2011	Salt		q-50mM-2011-SL7.3	3	Bac130	114.7	2.6	112.5	116.8	6.3			PC	0.03
28	2011	Salt	SL9_50mM.2011	q-50mM-2011-SL9.5	5	BrSPL5P3a	67.4	3.6	65.5	71.4	12.7	56.8	0.8	YS	0.05
	2011	Salt		q-50mM-2011-SL9.6	10	P21M47M444.3	18.9	3.0	17.7	39.6	10.5			YS	0.05
	2011	Salt		q-50mM-2011-SL9.1	2	BRH04D11flc2	17.4	2.9	12.5	21.8	10.1			YS	0.04
	2011	Salt		q-50mM-2011-SL9.2	2	BrPIP2a	38.9	2.2	36.1	42.6	7.7			YS	0.04
	2011	Salt		q-50mM-2011-SL9.3	3	BRMS-050-1	113.8	2.4	103.4	116.8	8.3			PC	0.05
	2011	Salt		q-50mM-2011-SL9.4	5	BRMS-034	22.4	2.0	12.1	40.2	7.5			YS	0.04
29	2011	Control	SFw_Control.2011	q-0mM-2011-SF.1	5	BRMS-034	22.4	3.5	16.7	24.4	11.9	20.6	-	YS	34.52
	2011	Control		q-0mM-2011-SF.2	9	BrT1N6_2P1c	0.0	2.6	0.0	31.4	8.7			YS	28.43
30	2011	Salt	SFw_50mM.2011	q-50mM-2011-SF.1	5	BRMS-034	22.4	3.4	16.7	24.4	12.5	12.5	-	YS	31.20
31	2011	Control	SDw_Control.2011	q-0mM-2011-SD.1	3	BrBCAT4-1	107.6	3.5	99.2	116.8	10.9	36.6	-	YS	3.98
	2011	Control		q-0mM-2011-SD.3	9	BrST5b-7	70.1	3.7	66.0	70.6	11.5			YS	4.04
	2011	Control		q-0mM-2011-SD.4	10	E32M19M392.9	62.6	2.4	57.2	68.5	7.2			PC	3.48
	2011	Control		q-0mM-2011-SD.2	5	BRMS-034	22.4	2.3	15.7	23.4	7			YS	3.29
32	2011	Salt	SDw_50mM.2011	q-50mM-2011-SD.3	3	Myb2RSAIM69.3	104.4	4.0	97.9	116.8	7.1	51.3	-	YS	2.44
	2011	Salt		q-50mM-2011-SD.4	5	BRMS-034	22.4	6.9	17.7	36.2	13.4			YS	3.46
	2011	Salt		q-50mM-2011-SD.5	9	BrST5b-7	70.1	7.2	63.2	70.6	14.1			YS	3.74
	2011	Salt		q-50mM-2011-SD.6	10	P23M48M244.6	67.5	4.3	58.8	69.5	7.8			PC	2.71
	2011	Salt		q-50mM-2011-SD.1	1	BrBCAT3-2	51.3	2.4	47.6	53.3	4			PC	2.28
	2011	Salt		q-50mM-2011-SD.2	2	BrFLMP1b	114.8	2.8	113.2	118.4	4.9			PC	2.14
33	2011	Control	RL3_Control.2011	q-0mM-2011-RL3.2	5	P23M50M241.1	61.3	4.9	58.1	71.4	17.7	49.8	0.7	PC	0.08
	2011	Control		q-0mM-2011-RL3.4	9	BrST5b-9	65.0	4.1	61.7	73.8	14.4			PC	0.07
	2011	Control		q-0mM-2011-RL3.3	9	BrFRY1P1b	51.4	2.6	51.1	55.9	9.4			PC	0.06
	2011	Control		q-0mM-2011-RL3.1	3	P23M50M273.1	112.5	2.5	104.1	139.9	8.3			YS	0.06
34	2011	Control	RL5_Control.2011	q-0mM-2011-RL5.2	5	P23M50M241.1	61.3	3.2	60.1	62.3	12.2	42.4	0.8	PC	0.18
	2011	Control		q-0mM-2011-RL5.3	8	P13M48M456.4	2.2	3.8	0.0	11.5	11.5			YS	0.18
	2011	Control		q-0mM-2011-RL5.4	9	BrST5b-9	65.0	3.5	64.2	68.5	10.7			PC	0.17
	2011	Control		q-0mM-2011-RL5.1	2	BrFLMP2d	18.1	2.7	9.5	18.6	8			PC	0.15
35	2011	Control	RL7_Control.2011	q-0mM-2011-RL7.1	2	P13M48M137.1	15.9	5.3	10.5	21.8	18.5	43.5	0.8	PC	0.39
	2011	Control		q-0mM-2011-RL7.2	2	BrTSFP2c	62.8	4.3	60.6	68.3	15.4			PC	0.38
	2011	Control		q-0mM-2011-RL7.3	5	P23M50M241.1	61.3	2.9	59.1	62.3	9.6			PC	0.28
36	2011	Control	RL9_Control.2011	q-0mM-2011-RL9.1	2	P13M48M137.1	15.9	5.6	11.5	19.8	23.4	44.8	0.8	PC	0.64
	2011	Control		q-0mM-2011-RL9.2	2	BrTSFP2c	62.8	2.2	60.6	68.3	10.1			PC	0.43
	2011	Control		q-0mM-2011-RL9.3	2	BrFLMP1b	114.8	2.5	113.2	120.4	11.3			PC	0.47
37	2011	Salt	RL3_50mM.2011	q-50mM-2011-RL3.1	2	BrFLC2-2	18.6	3.7	11.5	20.8	7.4	55.3	0.7	PC	0.02
	2011	Salt		q-50mM-2011-RL3.3	5	BRMS-034	22.4	6.4	15.7	25.4	13.4			YS	0.04
	2011	Salt		q-50mM-2011-RL3.4	9	BrVIP4P1b	44.1	3.1	41.8	47.1	15.4			PC	0.03
	2011	Salt		q-50mM-2011-RL3.5	9	BrELF3P1c	112.9	3.1	91.9	115.2	14.2			PC	0.03
	2011	Salt		q-50mM-2011-RL3.2	4	P23M50M127.6	42.4	2.5	28.8	42.4	4.9			PC	0.02
38	2011	Salt	RL5_50mM.2011	q-50mM-2011-RL5.1	2	BRH04D11flc2	17.4	3.5	11.5	21.8	9.1	69.9	0.7	PC	0.05
	2011	Salt		q-50mM-2011-RL5.3	2	E32M19M308.5	58.6	3.1	58.2	67.3	8.3			PC	0.05
	2011	Salt		q-50mM-2011-RL5.4	5	BRMS-034	22.4	4.6	14.7	24.4	12.4			YS	0.06
	2011	Salt		q-50mM-2011-RL5.6	9	BrVIP4P1b	44.1	3.3	41.8	46.1	13.9			PC	0.06
	2011	Salt		q-50mM-2011-RL5.2	2	E34M15M142.8	40.6	2.3	31.2	41.6	6.1			YS	0.04
	2011	Salt		q-50mM-2011-RL5.5	9	BrT1N6_2P1c	0.0	2.3	0.0	9.0	10			PC	0.05
	2011	Salt		q-50mM-2011-RL5.7	9	E32M19M95.0	72.7	2.3	61.7	75.3	10.1			PC	0.05
39	2011	Salt	RL7_50mM.2011	q-50mM-2011-RL7.3	9	BrT1N6_2P1c	0.0	3.5	0.0	8.0	13.8	28.6	0.8	YS	0.10
	2011	Salt		q-50mM-2011-RL7.1	2	E32M19M308.5	58.6	2.8	58.2	65.8	8.6			PC	0.08
	2011	Salt		q-50mM-2011-RL7.2	5	BRMS-034	22.4	2.1	13.7	24.4	6.2			YS	0.07
40	2011	Salt	RL9_50mM.2011	q-50mM-2011-RL9.1	2	E32M19M308.5	58.6	2.2	58.2	64.8	7	16.4	0.8	PC	0.13
	2011	Salt		q-50mM-2011-RL9.2	9	BrT1N6_2P1c	0.0	2.9	0.0	17.0	9.4			YS	0.15
41	2011	Control	RFw_Control.2011	q-0mM-2011-RF.1	2	Br323	19.8	5.3	10.5	21.8	21.9	37.9	-	PC	16.70
	2011	Control		q-0mM-2011-RF.2	2	E34M15M142.8	40.6	3.7	35.1	42.6	16			PC	13.74
42	2011	Control	RDw_Control.2011	q-0mM-2011-RD.1	3	P23M50M284.3	138.9	2.8	98.2	142.9	9.7	16.9	-	YS	0.10
	2011	Control		q-0mM-2011-RD.2	5	BRMS-034	22.4	2.1	13.1	25.4	7.2			YS	0.08
43	2011	Salt	RFw_50mM.2011	q-50mM-2011-RF.1	1	P14M51M213.3	34.9	4.0	32.2	36.9	13.9	13.9	-	YS	14.15
44	2011	Salt	RDw_50mM.2011	q-50mM-2011-RD.1	3	E32M47M382.4	126.5	3.8	99.2	133.1	16.6	40.9	-	PC	2.38
	2011	Salt		q-50mM-2011-RD.2	8	P23M50M335.3	9.7	5.3	1.0	11.5	24.3			PC	0.26

Note: **1** - Total explained variance of significant QTL (LOD >3) and candidate QTL (LOD >2), **UL** - lower limit and **UL** - upper limit. PC-pak choi and YS - Yellow sarson

Supplementary Table S4: Summary of significant QTLs identified for salt tolerance parameters in the *B. rapa* DH population from a cross of yellow sarson (YS143) and pak choi (PC175).

SN	Traits	QTL name	Linkage group	Peak marker	QTL position	LOD score	Confidence interval		QTL explained variance	Total explained variance ¹	Heritability	Allele from	Allelic effect
							LL	UL					
1	STI.RL3_C50.2010	q-STI0.50-2010-RL3.2	9	BrCRY1P1c	75.6	4.2	66.8	82.8	13.6	22.4	0.62	PC	0.08
		q-STI0.50-2010-RL3.1	8	Myb2AluIM190.6	68.7	2.4	60.6	89.3	8.8			PC	0.07
2	STI.RL5_C50.2010	q-STI0.50-2010-RL5.1	2	BrTSFP2c	62.8	4.0	60.6	70.3	11.9	48.5	0.66	PC	0.07
		q-STI0.50-2010-RL5.2	4	E32M19M204.4	16.9	4.5	4.0	25.3	13.6			YS	0.07
		q-STI0.50-2010-RL5.3	8	P23M50M86.4	0.0	3.3	0.0	20.4	9.9			YS	0.07
		q-STI0.50-2010-RL5.4	9	E32M19M95.0	72.7	4.3	66.8	76.7	13.1			PC	0.07
		q-STI0.50-2010-RL7.4	8	Myb2HaeIIIM213.1	20.4	4.3	0.0	20.4	15.5			YS	0.09
3	STI.RL7_C50.2010	q-STI0.50-2010-RL7.1	2	P23M50M179.6	17.2	2.4	8.5	23.8	8	40.8	0.66	PC	0.06
		q-STI0.50-2010-RL7.2	2	BrTSFP2c	62.8	2.6	56.8	68.3	9			PC	0.07
		q-STI0.50-2010-RL7.3	4	E32M19M435.5	27.8	2.6	9.0	37.9	8.3			YS	0.06
		q-STI0.50-2010-RL9.1	8	Myb2HaeIIIM213.1	20.4	4.2	0.0	20.4	14.2			YS	0.10
4	STI.RL9_C50.2010	q-STI0.50-2010-RL3.1	5	E32M19M221.8	28.0	4.7	15.7	35.0	14.2	22	0.49	YS	0.07
		q-ratio.0/50-2010-RL3.2	8	BrCRY2P2c	84.0	2.7	78.2	89.3	7.8			YS	0.06
6	RelST.RL5_C50.2010	q-ratio.0/50-2010-RL5.2	7	Myb2RSAIM230.6	41.0	5.5	36.9	43.9	16	46.9	0.62	YS	0.11
		q-ratio.0/50-2010-RL5.3	8	BrCRY2P3a	85.8	5.3	60.9	89.3	15.4			YS	0.09
		q-ratio.0/50-2010-RL5.4	10	P13M48M342.9	32.9	3.3	30.8	36.9	9			PC	0.07
		q-ratio.0/50-2010-RL5.1	5	BrSPA1P2a	13.7	2.4	0.0	19.7	6.5			YS	0.06
		q-ratio.0/50-2010-RL7.3	7	Myb2RSAIM230.6	41.0	7.1	36.9	43.9	19.9			YS	0.13
7	RelST.RL7_C50.2010	q-ratio.0/50-2010-RL7.4	8	BrCRY2P3a	85.8	3.4	83.8	89.3	8.6	39.1	0.61	YS	0.07
		q-ratio.0/50-2010-RL7.1	2	BrFYP1a	23.8	2.1	19.6	26.8	5.5			YS	0.06
		q-ratio.0/50-2010-RL7.2	6	Na12H07	60.2	2.1	59.1	72.0	5.1			YS	0.06
		q-ratio.0/50-2010-RL9.1	2	Br323	19.8	3.5	15.1	25.8	8.9			YS	0.08
		q-ratio.0/50-2010-RL9.2	7	Myb2RSAIM230.6	41.0	7.9	36.9	43.9	22.2			YS	0.14
8	RelST.RL9_C50.2010	q-ratio.0/50-2010-RL9.3	8	BrCRY2P2c	84.0	4.3	79.2	86.8	11.1	57.2	0.57	YS	0.09
		q-ratio.0/50-2010-RL9.4	9	O10D08	39.0	3.1	27.6	43.8	7.8			YS	0.07
		q-ratio.0/50-2010-RL9.5	10	P13M48M342.9	32.9	2.9	30.8	41.0	7.2			PC	0.07
		q-STI0-50-2010-SL3.1	3	E37M47M134.5	52.9	3.2	36.4	63.2	9.1			PC	0.09
		q-STI0-50-2010-SL3.3	7	Myb2RSAIM230.6	41.0	3.2	37.9	42.9	8.9			YS	0.10
9	STI_SL3.C50.2010	q-STI0-50-2010-SL3.4	9	BrFPAP1c	99.5	3.6	90.1	99.5	10.3	42.7	0.71	PC	0.10
		q-STI0-50-2010-SL3.2	6	BrRGAP1a	85.1	2.5	82.0	95.9	7			PC	0.08
		q-STI0-50-2010-SL3.5	10	P23M48M244.6	67.5	2.6	54.7	77.8	7.4			YS	0.08
		q-STI0-50-2010-SL5.2	8	BrESD4P2b	27.9	4.0	1.0	33.3	12.5			YS	0.07
		q-STI0-50-2010-SL5.1	3	BrNS1P2c	121.2	2.1	116.2	125.1	6			PC	0.05
11	RelST_SL3.C50.2010	q-ratio.0/50-2010-SL3.2	5	Myb2HaeIIIM263.7	55.2	3.0	50.3	65.5	11.4	19.6	0.49	YS	0.06
		q-ratio.0/50-2010-SL3.1	4	P21M47M178.1	34.6	2.4	31.8	47.8	8.2			PC	0.05
		q-ratio.0/50-2010-SL5.1	3	BRMS-050-1	113.8	4.8	113.8	115.7	13.4			PC	0.06
12	RelST_SL5.C50.2010	q-ratio.0/50-2010-SL5.2	5	Myb2HaeIIIM263.7	55.2	3.9	1.0	63.4	9.9	30.5	0.48	YS	0.05
		q-ratio.0/50-2010-SL5.3	9	P14M51M147.2	75.3	2.9	61.7	75.6	7.2			YS	0.04
		q-STI0-50-2011-SL3.1	3	P21M47M198.6	48.0	2.3	35.4	60.9	7	15.4	0.75	PC	0.08
13	STI_SL3_C50.2011	q-STI0-50-2011-SL3.2	6	BrFPF1P2a	72.2	2.8	67.3	74.2	8.4			PC	0.09
14	STI_SL5_C50.2011	q-STI0-50-2011-SL5.3	8	P23M50M335.3	9.7	4.5	1.0	23.7	15	46.2	0.72	YS	0.10
		q-STI0-50-2011-SL5.4	10	chsssr13	55.2	4.0	46.8	58.2	13.1			YS	0.10
		q-STI0-50-2011-SL5.1	3	BRMS-050-1	113.8	2.8	113.8	115.7	8.9			PC	0.09
		q-STI0-50-2011-SL5.2	5	P14M51M182.0	36.2	2.7	17.7	38.2	9.2			PC	0.08
		q-STI0-50-2011-SL7.1	3	P23M50M273.1	112.5	3.1	97.9	116.8	10			PC	0.07
15	STI_SL7.C50.2011	q-STI0-50-2011-SL7.5	10	P21M47M444.3	18.9	3.6	17.7	58.2	11.6	48	0.73	YS	0.09
		q-STI0-50-2011-SL7.3	8	P13M48M174.4	30.2	3.1	13.2	38.6	10.1			YS	0.07
		q-STI0-50-2011-SL7.2	5	BRMS-034	22.4	2.9	17.7	24.4	9.3			YS	0.07
		q-STI0-50-2011-SL7.4	8	Myb2AluIM263.6	74.2	2.1	66.5	78.2	7			YS	0.07
		q-STI0-50-2011-SL9.1	8	P13M48M174.4	30.2	2.2	22.2	39.6	9.7			YS	0.07
17	RelST_SL3.C50.2011	q-ratio.0/50-2011-SL3.1	9	BrGIP1d	81.5	2.1	80.3	97.7	8.1	8.1	0.33	PC	0.05
18	RelST_SL7.C50.2011	q-ratio.0/50-2011-SL7.3	5	Myb2RSAIM-714.5	90.1	3.7	59.1	95.1	14.1	32.7	0.68	YS	0.06
		q-ratio.0/50-2011-SL7.1	1	P23M48M219.2	66.1	2.9	59.4	67.1	10.6			PC	0.06
		q-ratio.0/50-2011-SL7.2	3	Myb2MSE1M321.8	7.6	2.2	0.0	33.4	8			PC	0.04
19	RelST_SL9.C50.2011	q-ratio.0/50-2011-SL9.1	1	BrBCAT3-2	51.3	6.2	48.6	67.1	17.7	93.2	0.67	PC	0.08
		q-ratio.0/50-2011-SL9.2	1	E32M47M179.8	71.5	4.5	70.8	84.6	13.4			PC	0.06
		q-ratio.0/50-2011-SL9.4	5	Myb2RSAIM-714.5	90.1	6.1	90.1	98.1	12.2			YS	0.05
		q-ratio.0/50-2011-SL9.5	6	E34M15M60.5	81.0	5.9	74.2	95.9	12.9			PC	0.05
		q-ratio.0/50-2011-SL9.6	9	BrCRY1P1c	75.6	3.9	61.7	75.6	11.7			YS	0.05
		q-ratio.0/50-2011-SL9.8	10	BrVIN3P1a	45.8	6.9	44.7	60.7	14.2			YS	0.06
		q-ratio.0/50-2011-SL9.7	9	BrGIP1d	81.5	2.9	80.3	82.8	7.2			YS	0.04
		q-ratio.0/50-2011-SL9.3	4	Myb2HaeIIIM317.3	51.6	2.2	49.6	56.6	3.9			YS	0.03
		q-STI0-50-2011-RL3.2	5	BR378	140.0	4.6	133.1	151.0	15.7	28.9	0.74	YS	0.14
20	STI_RL3.C50.2011	q-STI0-50-2011-RL3.1	4	P21M47M178.1	34.6	2.5	28.8	42.2	7			PC	0.08
		q-STI0-50-2011-RL3.3	9	BrGIP1d	81.5	2.2	80.3	99.5	6.2			PC	0.07

Supplementary Table S4 (continued): Summary of significant QTLs identified for salt tolerance parameters in the *B. rapa* DH population from a cross of yellow sarson (YS143) and pak choi (PC175).

21	STI_RL5.C50.2011	q-STI0-50-2011-RL5.1	2	BrFLMP2d	18.1	3.3	10.5	20.8	11.3	19	0.74	PC	0.10
		q-STI0-50-2011-RL5.2	2	BrTSFP2c	62.8	2.2	58.6	69.3	7.7			PC	0.08
22	STI_RL7.C50.2011	q-STI0-50-2011-RL7.1	2	BRH04D11flc2	17.4	4.5	11.5	21.8	17.2	25.4	0.76	PC	0.12
		q-STI0-50-2011-RL7.2	2	BrTSFP2c	62.8	2.0	58.2	68.3	8.2			PC	0.09
23	STI_RL9.C50.2011	q-STI0-50-2011-RL9.1	2	BRH04D11flc2	17.4	3.2	10.5	21.8	12.5	25.7	0.77	PC	0.11
		q-STI0-50-2011-RL9.2	2	BrTSFP2c	62.8	3.4	58.2	68.3	13.2			PC	0.11
24	RelST_RL3.C50.2011	q-ratio.0/50-2011-RL3.3	5	P23M50M241.1	61.3	5.5	57.1	66.5	17.5	36.8	0.39	YS	0.06
		q-ratio.0/50-2011-RL3.1	3	Myb2AluIM377.3	6.5	2.3	0.0	13.7	9.6			PC	0.04
		q-ratio.0/50-2011-RL3.2	3	P14M51M135.1	105.8	2.4	97.9	115.7	9.7			PC	0.04
25	RelST_RL5.C50.2011	q-ratio.0/50-2011-RL5.2	3	BRMS-008	0.6	3.1	0.0	11.7	11	28.6	0.34	PC	0.04
		q-ratio.0/50-2011-RL5.1	2	E34M15M461.1	109.4	2.3	98.3	111.2	8.1			YS	0.04
		q-ratio.0/50-2011-RL5.3	5	P23M50M241.1	61.3	2.7	58.1	69.4	9.5			YS	0.04
26	RelST_RL7.C50.2011	q-ratio.0/50-2011-RL7.1	3	BRMS-008	0.6	3.8	0.0	13.7	13.5	13.5	0.39	PC	0.06
27	RelST_RL9.C50.2011	q-ratio.0/50-2011-RL9.1	1	P14M51M243.3	53.5	6.4	48.6	60.5	14.7	40.7	0.54	PC	0.10
		q-ratio.0/50-2011-RL9.2	3	BRMS-008	0.6	5.1	0.0	13.7	14.7			PC	0.07
		q-ratio.0/50-2011-RL9.3	6	BrBCAT3-1	57.3	3.5	54.8	67.3	11.3			YS	0.06
28	RelST_SL5.C50.2011	-	-	-	-	-	-	-	-	-	0.49	-	-

Note: 1 Total variance explained by significant (LOD >3) and candidate QTLs (LOD >2), **LL** - lower limit & **UL** - upper limit; PC-pak choi, YS-yellow sarson

Chapter 5

A systems genetics approach identifies gene regulatory networks associated with fatty acid composition in *Brassica rapa* seed

Ram Kumar Basnet^{1,2}, Dunia Pino Del Carpio³, Dong Xiao⁴, Johan Bucher¹, Mina Jin⁵, Kerry Boyle⁶, Pierre Fobert⁶, Richard G. F. Visser^{1,2}, Chris Maliepaard^{1,2}, Guusje Bonnema^{1,3}

¹Laboratory of Plant Breeding, Wageningen University, Wageningen, The Netherlands,

²Centre for BioSystems Genomics, Wageningen, The Netherlands,

³Current address: Department of Plant Breeding and Genetics, Cornell University, Ithaca, New York, USA,

⁴Current address: State Key Laboratory of Crop Genetics and Germplasm Enhancement, Horticultural College, Nanjing Agricultural University, Nanjing, China

⁵Department of Agricultural Biotechnology, National Academy of Agricultural Science, Rural Development Administration, 150 Suin-ro, Gwonseon-gu, Suwon 441-707, Korea

⁶Plant Biotechnology Institute, National Research Council of Canada, Saskatoon, Canada

Abstract

Fatty acids in seed affect seed germination and seedling vigour and fatty acid composition determines the quality of seed oil. In this study, quantitative trait locus (QTL) mapping of fatty acids and for transcript abundance was integrated with gene network analysis to unravel the genetic regulation of seed fatty acid composition in a *Brassica rapa* doubled haploid population from a cross between a yellow sarson oil-type and a black seeded pak choi. The distribution of major QTLs for fatty acids showed a relationship with the fatty acid types: linkage group A03 for mono-unsaturated fatty acids (MUFAs), A04 for saturated fatty acids (SFAs) and A05 for poly-unsaturated fatty acids (PUFAs). Using a genetical genomics approach, expression QTL (eQTL) hotspots were found at major fatty acid QTLs on linkage groups A03, A04, A05 and A09. An eQTL-guided gene co-expression network of lipid metabolism related genes showed major hubs at the genes *BrPLA2-ALPHA*, *BrWD-40*, a number of seed storage protein genes and a transcription factor *BrMD-2*, suggesting essential roles for these genes in lipid metabolism. Three sub-networks were extracted for the economically important and most abundant fatty acids erucic-, oleic-, and linoleic- and linolenic- acids. Network analysis, combined with comparison of genome positions of *cis*- or *trans*-eQTLs with fatty acid QTLs, allowed identification of candidate genes for genetic regulation of these fatty acids. The generated insights in the genetic architecture of fatty acid composition and underlying complex gene regulatory networks in *B. rapa* seeds are discussed.

Key words: Systems genetics, eQTL mapping, fatty acids, confounding factor, network analysis, major hub gene

Introduction

The species *Brassica rapa* displays enormous morphological variation as illustrated by the diversity of crops, including leafy vegetables, turnips and oil types (Zhao et al., 2005). *B. rapa* ssp. *oliferous* (oil-type rape) consists of the annual oil crops yellow sarson and brown sarson with high seed oil content (> 42%) (Kumar et al., 2011; Lühs et al. 1999). In the past, both sarsons were preferred as oil crops over *B. napus* in Asia, Canada and other parts of the world as they mature earlier and have a higher level of shattering resistance and spring frost tolerance (Kadkol et al., 1986; Karim et al., 2014). However, *B. rapa* rape seed was gradually replaced by *B. napus*, mainly because of the latter's higher oilseed yield and the availability of double-low genotypes, which are low in glucosinolate content as well as erucic acid content (Rahman et al., 2001; Karim et al., 2014). Still, *B. rapa* has been used to widen the genetic variation for improvement of *B. napus* (Qian et al., 2006; Karim et al., 2014).

Lipids are a group of naturally occurring molecules that include fats, glycerolipids, fatty acids, glycerophospholipids, sphingolipids, waxes and others. *De novo* synthesized fatty acids are modified by desaturation and elongation reactions and form triacylglycerols, which are the major storage form of seed oil in plants (Guschina and Harwood, 2007). For both nutritional and industrial purposes, the composition of fatty acids (FAs) determines the economic value of seed oil (Yan et al., 2011; Sanyal and Randal Linder, 2012). For food or feed, oil that is high in the level of human health beneficial oleic acid (C18:1) is preferred. This in turn can be easily desaturated to linoleic- and linolenic- acids or elongated to erucic acid. Oil with high erucic acid (C22:1) has a health risk, but can be used for industrial purposes, while high linolenic acid (C18:3) negatively affects oil storability (Yan et al., 2011). To breed for optimal fatty acid (FA) composition and high oil yield, understanding the genetic regulation of FA composition and the FA regulatory network is crucial.

QTL analysis and gene functional studies have been performed to unravel the genetics of FA composition in *Arabidopsis thaliana* (Sanyal and Randal Linder, 2012) soybean (Wang et al., 2014b), *Jatropha* (Liu et al., 2011) and *B. napus* (Peng et al., 2010; Yan et al., 2011). In *A. thaliana*, almost all genes and transcription factors involved in lipid metabolism and storage oils have been identified (Beisson et al., 2003; Le et al., 2010; Peng and Weselake, 2011). The *B. rapa* genome (A genome), like the *B. oleracea* genome (C genome) is syntenic to *A. thaliana*, but underwent a genome triplication (Wang et al., 2011a; Liu et al., 2014). *B. napus* is an amphidiploid resulting from natural hybridization between *B. rapa* (A genome) and *B. oleracea* (C genome). These genome triplications, resulting in many genes with paralogues, add another level of complexity to the genetic regulation of fatty acid composition in Brassicas. Across a number of studies, a large number of QTLs for fatty acids and oil content have been reported, suggesting a complex genetic architecture (Sanyal and Randal Linder, 2012; Yan et al., 2011). Genetic studies have identified the fatty acid desaturase genes *BnaFAD2* and *BnaFAD3* as the major genes for regulation of C18:1 (oleic acid) and C18:3 (linolenic acid) content in *B. napus* (Peng et al., 2010; Yang et al., 2012b; Lee et al., 2013). *BnaFAE1* is a candidate gene for erucic acid and total oil content in *B. napus* seed (Peng et al., 2010), while *BrFAD3* is a candidate gene for the synthesis of linolenic acid in seed

triacylglycerols in *B. rapa* ssp. *oleifera* (Tanhuanpää and Schulman, 2002). The present study is the first genome-wide genetic study for FA composition and transcriptional regulation of developing seeds of *B. rapa*.

Many genes involved in different metabolic processes are regulated in a coordinated fashion during seed development in *Arabidopsis* (Ruuska et al., 2002), *B. napus* (Yu et al., 2010) and *B. rapa* (Basnet et al., 2013). The combined study of phenotypic QTLs and expression QTLs (eQTLs) provides a basis to investigate the molecular mechanism and to understand the regulatory networks of genes involved in pathways of specific phenotypic traits in different organisms (Civelek and Lusi, 2014). Genome-wide mapping of gene transcripts in a segregating population was first proposed by Jansen and Nap (2001) and was named genetical genomics. Using genetical genomics, candidate genes were identified in *B. rapa* for flowering time and leaf development (Xiao et al., 2013; Xiao et al., 2014), phytonutrient content (Pino Del Carpio et al., 2014) and phosphorus use efficiency (Hammond et al., 2011).

The aim of the present study is to identify QTLs for FA content and composition in *B. rapa* seeds using a doubled haploid (DH) population from a cross between an oil-type yellow sarson and a vegetable pak choi. In order to understand the lipid gene regulatory network in *B. rapa*, we followed a genetical genomics approach combining QTLs for fatty acids in mature seeds with eQTLs for genes related to the lipid metabolism in developing seeds: FA biosynthesis and elongation, triacylglycerol biosynthesis, glycerol synthesis and lipid degradation. Based on gene expression variation, a gene co-expression network was constructed for genes involved in lipid metabolism and relative content of FAs. Because of the economic importance of erucic acid, oleic acid, linoleic acid and linolenic acid, individual sub-networks of those FAs were also derived. Finally, eQTL results were integrated with known FA pathways to unravel the regulation of genes involved in the composition of fatty acids in *B. rapa* seeds. This resulted in the identification of a number of QTL hotspots and key regulatory genes that are of importance for breeding purposes.

Materials and Methods

Plant materials and growth conditions

A *B. rapa* DH population of 163 DH lines developed from a cross of a yellow sarson (YS143; accession number: FIL500) as a female parent and a pak choi (PC175 cultivar: Nai Bai Cai; accession number: VO2B0226) as a male parent was used in this study (Basnet et al., 2013; Xiao et al., 2013). YS143 is a self-compatible annual oil crop with yellow seed colour, while PC175 is a self-incompatible leafy vegetable with brown/black seed colour. These two parents differ in seed size, seed colour, oil and fatty acid content and in many morphological and developmental traits (Zhao et al., 2005; Pino Del Carpio et al., 2011b; Basnet et al., 2013).

The experiments were carried out in two years, 2009 and 2011, using two different experimental designs, without and with synchronization of time of flowering of the DH lines, respectively. In both years, seeds of DH lines and parents were sown in pressed soil cubes of a standard soil mixture of 85% peat and 15% clay (Lentse Potgrond no. 4; Lentse Potgrond Lent, The Netherlands) for germination in a greenhouse at the Unifarm facility of Wageningen University. Two plants per

DH plus parents were transplanted to plastic pots (diameter 17 cm) filled with the same standard soil mixture in the greenhouse and later only one plant was kept for harvesting developing seeds. In 2009, seeds were sown on 27th March and DH lines flowered over a period from 1st week of May to 2nd week of June. In 2011, the population was evaluated again, but this time, the lines were sown staggered at different dates from the second week of January to the last week of February to synchronize the flowering of the lines. The aim of synchronizing flowering time is to avoid different environmental conditions during seed development. All lines flowered during the first two weeks of April. The ripe seeds were harvested per plant, dried and stored in a certified manner (ISO certified method 9001:2008) at 13°C temperature and 30% relative humidity, and later used for fatty acid measurements. Transcript abundance measurements were done in the developing seeds of the DH lines from 2011.

Fatty acid measurements

About 0.2-2.0 g of seeds were used to determine oil content using near-infrared reflectance spectroscopy (Foss NIRS system; Tillmann, 1997), calibrated with oil seed extracted with hexane following the standard protocol described by Raney et al. (1987). The oil content was calculated as a mass percentage of whole seed dry matter (zero moisture). Seed oil was analyzed for FA composition using gas chromatography (GC) following the preparation of fatty acid methyl esters by base-catalysed methanolysis (Thies, 1971). The individual FAs were reported as mass percentages of total fatty acids.

These relative contents of 18 fatty acids were studied in mature-dry seeds from 2009 and 2011 seed lots. In 2009, FAs were measured in 135 DH lines but, due to lack of availability of seeds, in 2011 only of a subset of 92 DH lines and 2-3 biological replicates of each of the two parents were studied. The 18 FAs consisted of 7 saturated FAs (SFAs), 5 monounsaturated FA (MUFAs) and 6 polyunsaturated FA (PUFAs) (Supplementary Table S1). Unidentified FAs were categorized as “Other FAs”. Lauric acid (C12:0) was excluded from data analysis because its concentration was below the detection limit in almost all DH lines including the parents.

RNA isolation for gene expression studies

In earlier studies we observed that at 28 days after pollination (DAP), a large subset of genes related to lipid metabolism was differentially expressed between the two parents as well as between two selected DH lines (Basnet et al., 2013). Therefore, siliques from a subset of 118 DH lines from the 2011 experiment, selected on the basis of genotypic contrast, were harvested at 28 DAP and kept in liquid nitrogen (-196°C); seeds were taken from the siliques under dry ice and around 150-200 seeds were ground in liquid nitrogen (-196°C) using RNase-free mortar and pestle. Seeds and RNA samples were stored at -80°C. Since *Brassica* seeds have high concentrations of oils, organic acids and proteins, 5% (w/v) polyvinylpyrrolidone (PVP-40) (Sigma) was added to RLC lysis buffer (Qiagen) and kept overnight at 65°C to dissolve properly. After adding RLC lysis buffer in each tube, the powdered seed materials were incubated for 30 minutes at 65°C in a water bath. The total RNA was extracted with RNeasy Plant Mini Kit (Qiagen) following the manufacturer's

instructions and purified using the RNA Clean-up protocol for RNeasy columns (RNeasy Mini Kit, Qiagen) with on-column digestion with DNase Kit (Qiagen) to remove residual genomic DNA. The quantity of RNA samples was measured by using a NanoDrop ND-100 UV-VIS spectrophotometer (NanoDrop, Technologies Inc., Wilmington, DE, USA) and quality was assessed by A260/A280 and A260/A230 ratio (NanoDrop, Technologies Inc.) and by 1% agarose gel.

Distant pair design and microarray hybridization

A distant pair design was used to find an optimal combination of pairs of DH lines with maximum genetic contrast for the microarray analysis (Fu and Jansen, 2006). Missing marker data was imputed based on the genetic positions of flanking markers using the “R/qrtl” package (Broman et al., 2003). Using the marker information, the pairs of DH lines to hybridize on an array were designed using the R package “designGG” (Li et al., 2009). The cRNA samples were labelled with Cy3 (green) and Cy5 (red) dyes using the QuickAmp Labeling Kit (Agilent Technologies, Inc., Santa Clara, CA, USA) and hybridized on our 8x60 K custom-made *B. rapa* array in a two-colour Agilent platform as described in Basnet et al. (2013). Arrays were then washed and scanned on an Agilent scanner, according to the manufacturer’s instructions. Data files were generated using the Agilent Feature Extraction Software (version 10.10.1.1). In total, 59 arrays for 118 DH lines and 4 arrays for two parents and their biological replicates were used. The raw data was normalized without background correction, using loess for within-array normalization and quantile normalization between arrays using the “limma” package in R (Smyth, 2005; R Core Team, 2012).

Gene expression measurements in RT-qPCR

The transcript abundance of 21 genes including genes for FA synthesis and FA elongation, FA desaturation, lipid degradation, seed storage proteins and lipid degradation was measured using RT-qPCR to validate microarray transcript abundance using cDNA from 28 days developing seeds collected in 2011. Genes in each pathway were selected based on the literature, and primers for each gene and for the reference gene are listed in Supplementary table S2. The genes *flowering locus C* (*BrFLC2*) and *transparent testa 8* (*BrTT8*) were also included because this population segregates for flowering time and seed colour which both have confounding effects on QTL mapping for fatty acids and transcript abundance. RT-qPCR was performed with paralogue-specific primers for the genes that have paralogues. The detailed procedure was as described in Xiao et al. (2013); we used the β -actin gene as reference gene to estimate the normalized gene expression ($\Delta\Delta CT$) of each gene and each sample.

QTL mapping of FAs (faQTL) and transcript abundance (eQTL)

In this study, an integrated genetic map was constructed (this thesis, Chapter 4) and used for QTL mapping of FAs and transcript abundance in developing seeds. This integrated map comprised 435 molecular markers: AFLP, myb-targeted, microsatellite (SSR) and gene targeted markers from the flowering time, FA and glucosinolate pathways. QTL interval mapping was performed for 17 FAs from the two years’ seed lots with the R/qrtl package (Broman et al., 2003), followed by multiple-

QTL mapping (MQM) with marker co-factors. Initially, cofactors were selected from the peak markers of significant QTLs in IM mapping. After that, a backward elimination was performed to select the final set of cofactors. LOD thresholds for significance of QTLs were determined at the 95 percentile of 10,000 permutations of each of the FAs. For QTL analysis, FA abundance was transformed using either a reciprocal or a log transformation, depending on the observed distribution of FA abundance values. In the 2009 seed lot, a log transformation was done for stearic acid while a reciprocal transformation was used for arachidic acid, behenic acid, lignoceric acid and palmitoleic acid. For the 2011 seed lot, a reciprocal transformation was done for palmitic acid, stearic acid, arachidic acid and behenic acid, and a log transformation for lignoceric acid. The same procedure as described above for mapping QTLs for fatty acids was followed to map eQTLs for transcript abundance of genes measured using RT-qPCR.

eQTL analysis was performed using single marker regression analysis as proposed by Fu and Jansen (2006). This method relates the log-ratios of each probe (transcript abundance contrast of a pair of genotypes) to DNA markers on the linkage map. The following regression model was used to regress the log-ratios of each probe against each marker as:

$$y_{ij} = \alpha_{ik} + \beta_{ik}x_{jk} + e_{ijk}$$

where, y_{ij} is the log-ratio of transcript abundance of pair j for gene i and x_{jk} denotes the marker allele contrast for the pair j at marker k , with the following marker values: 1 for yellow sarson / pak choi, -1 for pak choi / yellow sarson and 0 for yellow sarson / yellow sarson and pak choi / pak choi. The regression coefficient β_{ik} represents the allele substitution effect at marker k for probe i . The intercept α_{ik} is also estimated in the regression approach, but should be close to zero unless there is a dye bias; e_{ijk} is the residual error.

In total, 61,551 probes (representing 40,904 for *B. rapa* gene models, called “BralD”) for the two-colour Agilent microarray platform were designed using gene models predicted based on the reference genome sequence of *B. rapa* cv. Chiifu (a vegetable type inbred line), published by Wang et al. (2011a). A slightly modified version of the array designed for a microarray study in Basnet et al. (2013) was used in this study. All probes were annotated into 35 functional categories or “BINS” as defined by MapMan software. However, only 1568 probes related to lipid metabolism, lipid signaling, lipid storage proteins and lipid transfer proteins were extracted and subjected to eQTL analysis in this study. Significant eQTLs were declared using a genome-wide threshold of $\alpha = 0.001$ (or $-\log_{10}(0.001) = 3$). The $-\log_{10}(\text{p-value})$ are here denoted as LOP values (to increase readability); the interpretation of the LOP score is different from a LOD score as used in the faQTL analysis since the LOD is obtained by likelihood estimation whereas the LOP is from least-squares estimation using per-marker regression. The estimated regression coefficients of markers (β) represent the estimated additive effects of hypothetical QTLs at the marker position; the sign of β gives the direction of the effect of a parental allele (a positive value indicating a higher mean for the yellow sarson allele, a negative value indicating a higher mean for the pak choi allele).

Correction for the effect of seed colour on eQTL mapping

The two parents were contrasting in seed colour, YS143 being yellow-seeded and PC175 being black-seeded. Yellow-coloured seeds are associated with a high content of oil and protein and low fiber in the meal of *Brassica* oil seeds (Chen and Heneen, 1992; Rahman et al., 2001). Since a large number of eQTLs was mapped on A09 in the vicinity of a major QTL for seed colour, we were also interested in the eQTL results after correcting for a possibly confounding effect of seed colour in addition to the results without such a correction. The correction for seed colour was done by regressing transcript abundance (per probe) on image pixel size of colour intensity (a quantitative score of seed colour) from image analysis. A higher intensity indicates yellow seed coat colour, a lower intensity a black seed coat colour. The residuals of transcript abundance of each probe were then used for eQTL mapping, instead of the original values.

Cis- and trans- eQTL

In this study, an eQTL was defined as *cis*-acting if the eQTL was observed in the same linkage group of the physical position of the probe. An eQTL that was mapped to another linkage group was defined as a *trans*-eQTL. This broad definition is likely to result in an overestimation of the number of *cis*-eQTLs since also distant eQTLs on the same chromosome are now considered to be *cis*-regulated. On the other hand, it could result in a possible underestimation of the total number of eQTLs and *trans*-eQTLs if more eQTLs were on the same chromosome.

Significance of eQTL hotspots

The significance of the presence of eQTL hotspots at each genetic marker was tested using the “hotspots” package in R software (Darrouzet-Nardi, 2012). First, the number of eQTLs ($LOP > 3$) was counted at each genetic marker. The “hotspots” package then uses this eQTL number as its input variable. This package calculates a hotspot cut-off for the eQTL number distribution across the genome based on deviance from the normal distribution of a variable, the number of eQTLs in this case. Statistical as well as computational details of this package are as described by Darrouzet-Nardi (2012).

Construction of co-expression networks of genes and fatty acids

All the probes that were annotated as being involved in lipid metabolic processes in the MapMan annotation (Usadel et al., 2005; Basnet et al., 2013) were selected for eQTL mapping, and among those selected probes, only the probes with an eQTL ($LOP > 3$) were used to calculate Pearson correlation coefficients in R software (R Core Team, 2012) and for construction of a correlation network using Cytoscape (Shannon et al., 2003). For correlation based co-expression analysis, gene expression measured by RT-qPCR as well as relative abundance of 17 different FAs was included. FAs from the 2011 experiment were used for construction of this genes-FA co-expression network analysis because the RNA transcripts were also measured from the 2011 seed lot. For network visualization, correlation coefficients were considered to be significant at absolute value of 0.3 at $p < 0.05$. In the network, nodes represent probes or metabolites and edges represent

significant correlation coefficients. Since all the genes are from the lipid metabolism, most were significantly correlated to each other, which makes it difficult to visualize the network due to the large number of edges. To increase the visibility of the network, only absolute correlation coefficients > 0.5 were shown. A “NetworkAnalyzer” plug-in available in Cytoscape was used to calculate relevant network parameters, such as the degree of connection (Assenov et al., 2008). The degree of connection measures the number of incoming and outgoing edges of a node. A higher degree of connection indicates a node with a higher number of edges, which identifies a gene as a major hub of the network; these genes may represent major regulating genes of the pathway.

Weighted gene co-expression network (WGCNA) to correlate gene expression to FA abundance

A weighted correlation based network was used to complement the correlation co-expression network. Unlike the correlation co-expression network, in the WGCNA approach, not only the probes with an eQTL were used, but all the genes related to lipid metabolism. WGCNA constructs a network based on correlation coefficients between genes (expression values after correction) and further classified gene modules (groups of highly correlated genes). Finally, this method calculates the significance of association of gene modules and each FA. A detailed description of the WGCNA approach was given by Horvath and Dong (2008), and the analysis was performed in R software using the WGCNA package (Langfelder and Horvath, 2008). In this WGCNA approach, the network was constructed based on significant associations of genes with FAs. The number of times that a probe was related to different FAs was calculated and compared with the degree of connection as calculated from the correlation co-expression network.

Results

Variation of fatty acids in seed

The most abundant fatty acids in the seed lots were three MUFAs: erucic acid (C22:1), oleic acid (C18:1) and eicosenoic acid (C20:1) and two PUFAs: linoleic acid (C18:2) and linolenic acid (C18:3). Together, these accounted for 71.6-74.2% (MUFA's) and 16.9-19.2% (PUFA's) of the total oil concentration (Figure 1; Supplementary Table S1). Erucic acid (C22:1) was the most predominant FA in both years' seed lots and had a higher level in YS143 (55.8%) than in pak choi (47.0%). Oleic acid (C18:1) was the second most abundant FA, but was higher in pak choi (17.3%) than in yellow sarson (13.7%). Linoleic- and linolenic acids had comparable levels in both parents: 8.2-11.0% in pak choi and 6.8-10.8% in yellow sarson. The content of the FAs was very similar in the two different years. Total oil concentration was higher in yellow sarson (44.2%) than in pak choi (29.3%). For most of the FAs and total oil, a number of DH lines had higher or lower content than the two parents, indicating transgressive segregation in this population.

Correlations between years and among fatty acids

For most FAs there were high positive Pearson correlation coefficients ($r = 0.5-0.9$) between 2009 and 2011 (Figure 2). A notable exception was mead acid (C20:3) with a much lower positive

correlation ($r = 0.20$) between years. However, its abundance was near the detection limit for both the years. A combined analysis from both seed lots shows high positive correlations among SFAs, PUFAs and MUFAs, with the exception of the two MUFAs nervonic acid (C24:1) and the predominant FA erucic acid (C22:1). Nervonic acid and erucic acid were both negatively correlated with SFAs and MUFAs and positive correlated with several PUFAs, but they were not significantly correlated with each other (Figure 2).

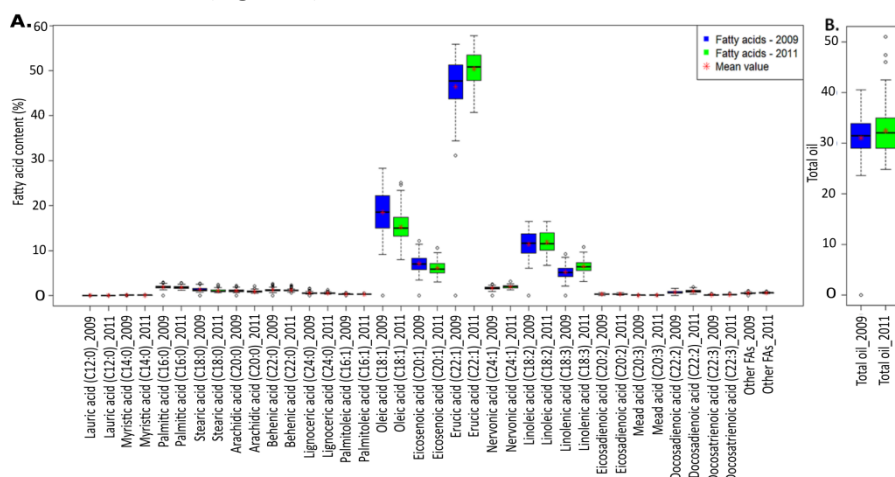


Figure 1: Boxplots showing the distribution of fatty acids (A) and total oil content (B) of ripe seeds of the *B. rapa* DH population from the 2009 and 2011 seed lots. Fatty acids were measured in mass percentage of the total oil content and the total oil content in mass percentage on the basis of whole seed dry matter (zero moisture).

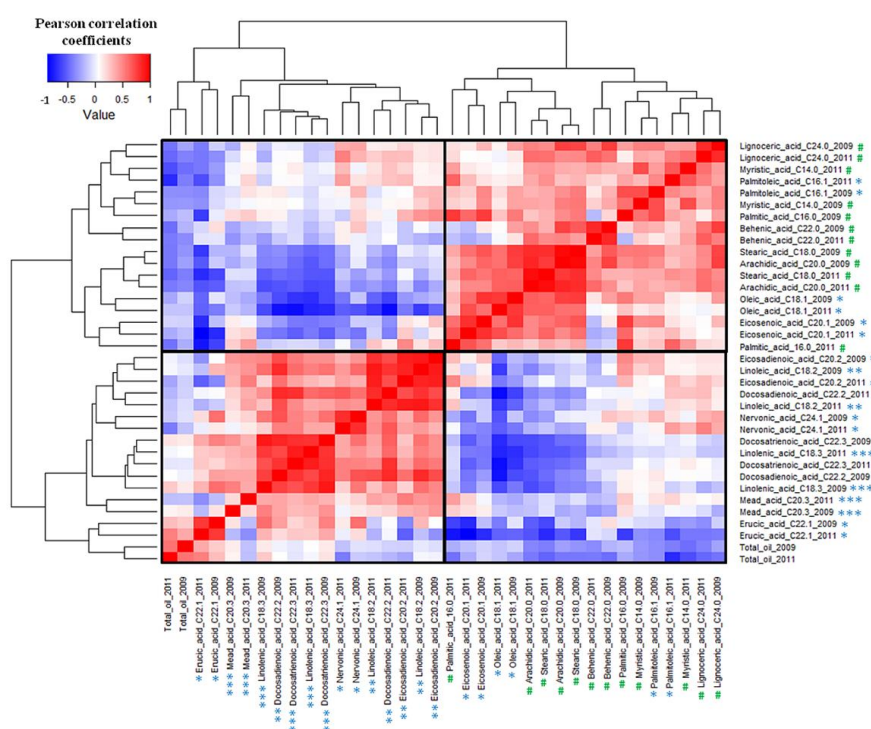


Figure 2: Heatmap of Pearson correlation coefficients of FAs and total oil content from the 2009 and 2011 seed lots. The name of a FA, its molecular structure and the year were concatenated with a “_” symbol. “#” indicates saturated fatty acids (SFAs), * monounsaturated fatty acids (MUFAs) and ** and *** indicate polyunsaturated fatty acids (PUFAs) with 2 and 3 double bonds, respectively.

QTL mapping of fatty acids

We performed QTL analyses for 17 different FAs from the 2009 and 2011 seed lots. We observed more QTLs for FAs (hereafter called faQTLs) from the 2009 seed lot when plants flowered without synchronization than from the 2011 seed lot, harvested from plants that flowered synchronously. In the 2009 seed lot, 56 faQTLs were detected and 24 of these faQTLs (43%) had at least 10% explained variance (mean: 11%; maximum: 36%). For the 2011 seed lot, only 32 faQTLs were detected, but for 2011 a much higher percentage, 24 faQTLs (75% of the faQTLs) had an explained variance of at least 10% (mean: 15.8%; maximum: 46%) (Supplementary Table S3 and S4).

Major faQTLs (LOD scores > 10, explained variances > 15%) were observed on linkage groups A03, A04 and A05 (Figure 3). faQTLs were observed across all ten linkage groups; faQTLs for 11 FAs co-located with a major flowering QTL at the genomic region (16.7 cM) of the *BrFLC2* gene-targeted marker on A02 in the 2009 seed lot (Figure 3; Supplementary Table S3). For the 2009 and 2011 seed lot, faQTLs were mainly mapped on A03, A04, A05 and A07 (Figure 3). In both seed lots, at major faQTLs detected for SFAs (myristic acid and behenic acid) on A04 and PUFAs (linoleic acid, eicosadienoic acid and docosadienoic acid) on A05, the yellow sarson allele was associated with a higher concentration (Figure 3). For the erucic acid QTL on A03 and A05, the pak choi allele was associated with higher concentrations, while on A02, A07 and A09, the yellow sarson allele was associated with higher concentrations (Figure 3).

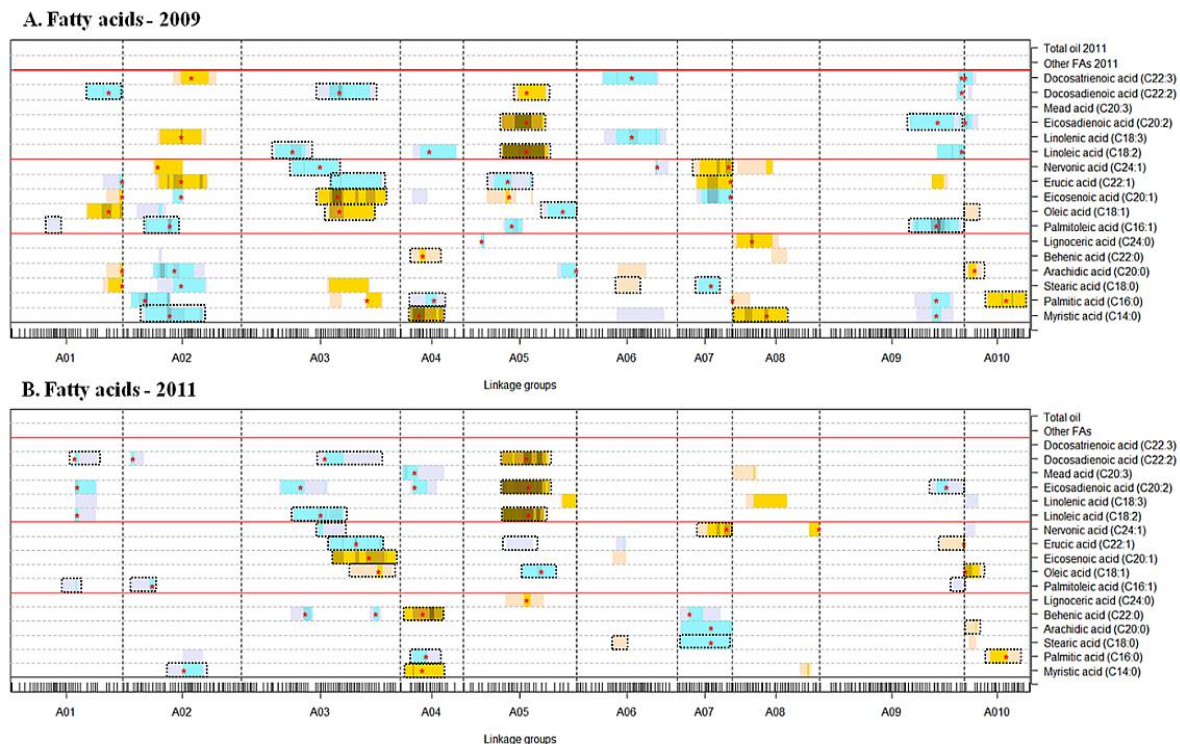


Figure 3: Heatmap of QTL profiles for FAs in the 2009 and 2011 seed lots. The darker the colour intensity, the higher the LOD score. Yellow indicates a QTL effect where the yellow sarson allele is associated with higher abundance while the blue colour indicates an effect where the pak choi allele is associated with higher abundance. The red triangles indicate the positions of QTL peak markers. The black coloured tick marks indicate marker positions; the vertical dashed lines separate the linkage groups. The horizontal dotted black lines separate traits and red lines separate SFAs, MUFAs and PUFAs. Dotted boxes indicate QTLs present in both years' seed lots.

eQTL mapping of transcript abundance of lipid related genes

1568 probes (of 921 BraIDs) for genes related to lipid metabolism, lipid signaling, lipid storage proteins and lipid transfer proteins were used for eQTL analyses. Those lipid related genes were selected according to MapMan annotation. Out of the 1568 probes, 760 probes (representing 537 BraIDs) had at least one eQTL at LOP value ($LOP = -\log_{10}$ of pvalue) > 3 and 270 probes (194 BraIDs) had at least one eQTL at LOP value > 5 . In total, 1118 eQTLs were detected for 760 probes (537 BraIDs), including 304 *cis*-eQTLs (27%) and 814 *trans*-eQTLs (73%) (Table 1). 146 probes (115 BraIDs) had both *cis*- and *trans*-eQTLs. The majority of probes (745 probes) had 1-3 eQTLs, while only 12 probes had four eQTLs and three had a maximum of five eQTLs per probe. Five linkage groups, A03, A04, A05, A07 and A09 had significant eQTL hotspots (> 42 eQTLs) (Figure 4A and B – left panel).

Most of the *cis*-eQTLs had a higher significance (maximum LOP 29 and mean LOP 7) than the *trans*-eQTLs (maximum LOP 18 and mean LOP 4). The largest *trans*-eQTL hotspots were on A05 (19% of total *trans*-eQTL) and A09 (25% *trans*-eQTL) (Figure 4B-left panel). The *trans*-eQTL hotspot at A09 co-locates with a major QTL for seed coat colour, which explains 33% of the colour variation (data not shown).

QTL mapping after correcting for seed colour differences

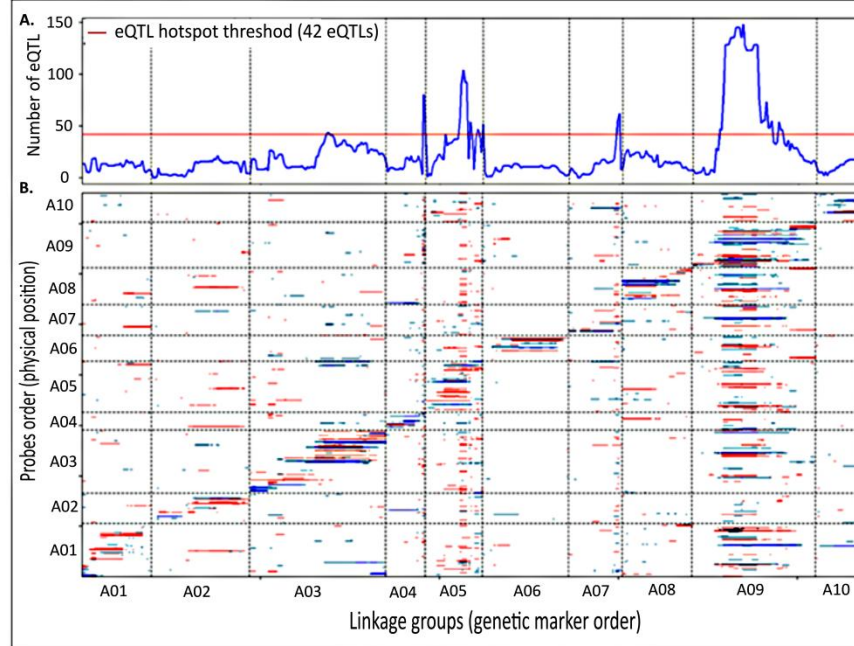
A possible confounding effect of seed colour on transcript abundance was corrected for using a simple linear regression model as described in the M&M section. After correction for seed colour, 946 eQTLs were observed for 662 probes (488 BraIDs) across the genome, 273 probes (194 BraIDs) had *cis*-eQTLs (29%) and 513 probes (397 BraIDs) had 673 *trans*-eQTLs (71%) (Table 1). Most of the probes (641 probes) had 1-3 eQTLs, while seven probes had four eQTLs and three probes had a maximum of five eQTLs per probe.

eQTL hotspots were now detected on A03 (153 eQTLs, 16% of total eQTLs), A04 (115 eQTLs, 12%), A05 (172 eQTLs, 18%), A07 (68 eQTLs, 7 %) and A09 (125 eQTLs, 13%) (Table 1; Figure 4A and B - right panel). Like in the analysis before correction, *cis*-eQTLs had higher significance than *trans*-eQTLs.

Table 1: Numbers and percentages of *cis*- and *trans*- eQTLs detected in each linkage group before and after correction for seed colour.

Linkage group	Before correction				After correction			
	<i>cis</i> -eQTLs	<i>trans</i> -eQTLs	Total	eQTL (%)	<i>cis</i> -eQTLs	<i>trans</i> -eQTLs	Total	eQTL (%)
A01	36	48	84	7.5	31	46	77	8.1
A02	19	43	62	5.6	20	48	68	7.2
A03	64	89	153	13.7	55	98	153	16.2
A04	17	103	120	10.7	19	96	115	12.2
A05	43	154	197	17.6	39	133	172	18.2
A06	16	23	39	3.5	15	25	40	4.2
A07	14	76	90	8.1	12	56	68	7.2
A08	26	54	80	7.2	27	61	88	9.3
A09	53	199	252	22.6	36	89	125	13.2
A10	16	24	40	3.6	19	21	40	4.2
Total	304 (27.2%)	813 (72.8%)	1117		273 (28.9%)	673 (71.1%)	946	

Before correction



After correction

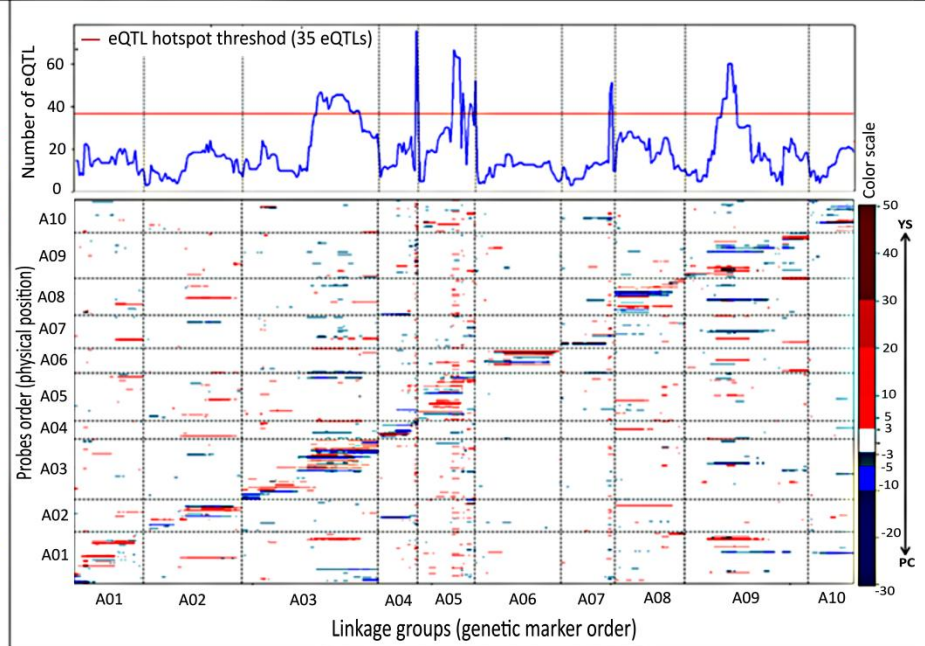


Figure 4: Genome-wide distribution of expression quantitative trait loci (eQTLs) in the developing seeds (28 DAP: days after pollination) of a *B. rapa* DH population: before (left panel) and after correction for seed colour (right panel). **A.** The frequency of eQTLs at each genetic marker along the 10 linkage groups (LGs), separated by dashed lines. The y-axes represent the number of eQTLs. The red line indicates the threshold for declaring a significant eQTL hotspot. **B.** Scatter plots of *cis*-/*trans*-eQTLs of probes related to lipid metabolism before (left panel) and after correction (right panel). The y-axes represent the order of probes according to physical positions in the genome, the x-axes the order of genetic markers in the genetic map. eQTLs on the diagonal represent *cis*-eQTLs and off-diagonal eQTLs are *trans*-eQTLs. Significant eQTLs associated with higher transcript abundance from the yellow sarson allele or from the pak choi allele are shown in red and blue colour gradients, respectively. Significance of eQTLs was determined at LOP 3 ($-\log_{10}(\text{P-value})$). Vertical dashed lines separate the LGs in the genetic map, horizontal dashed lines separate the LGs in the physical map.

Comparison of eQTLs before and after correcting for seed colour

A large number of probes (630 probes for 461 BralDs) with significant eQTLs were in common between both analyses. After correction, an additional 32 probes (28 BralDs), belonging to 8 different pathways, such as lipid degradation, FA synthesis and elongation, seed storage proteins, FA desaturation, had eQTLs. eQTLs of the 130 probes (110 BralDs) that were lost after correction, could be either false positives in the analysis before correction or false negatives after correction due to overcorrection for seed colour because their expression variation correlated with seed colour variation (for example, due to linkage or a pleiotropic effect). Chi-square tests were performed for each linkage group to test the significance of the differences in number of eQTLs before and after correction. There were no significant changes in the number of *cis*-eQTLs ($p > 0.05$) but the number of *trans*-eQTLs was significantly changed on A03 (89 to 98 *trans*-eQTL, $p = 0.03$) and A09 (199 to 89 *trans*-eQTLs, $p < 0.0001$) (Table 1; Figure 4).

eQTLs from the RT-qPCR gene expression studies

The expression values of 23 genes obtained by RT-qPCR were also subjected to correction for seed colour and then eQTL analysis to validate microarray eQTL results. For 16 out of 23 genes, at least one of the eQTLs was detected at the same position in both microarray and RT-qPCR experiments (Supplementary Figure S1). In case of *BrCER8*, *BrLACS2*, *BrCRU3* (Bra011036) and Br006444, eQTL profiles did not correspond between RT-qPCR and microarrays. The expression of *flowering locus C* (*BrFLC2*, a major regulator of flowering time in *B. rapa*) and the *transparent testa 8* gene (*BrTT8*, a major regulator of seed colour in *B. rapa*), that map under the faQTL hotspot on A02 in 2009 and eQTL hotspot on A09, respectively, were also measured with RT-qPCR. For *BrFLC2*, a *cis*-eQTL on A02 was confirmed in both RT-qPCR and microarray experiments, but an additional *trans*-eQTL was detected on A05 only in the RT-qPCR experiment. For *BrTT8* gene, a *cis*-eQTL was detected on A09 under the *trans*-eQTL hotspot, only for RT-qPCR (Supplementary Figure S1). For almost all genes, eQTLs detected for expression profiles measured in RT-qPCR were stronger (higher explained variance) than in the microarray.

Co-localization of mQTLs and eQTLs

Major faQTLs detected in both seed lots were compared with eQTL hotspots observed after correction for seed colour (Figure 3; Figure 4A and B – right panel). On A03, faQTLs of the MUFAs oleic acid, eicosenoic acid and erucic acid from both seed lots, co-localized with an eQTL hotspot. On A04, major faQTLs of the SFAs behenic acid and myristic acid and a minor faQTL for palmitic acid, also from both seed lots, co-localized with an eQTL hotspot. Major faQTLs for the PUFAs linoleic acid and eicosadienoic acid, and minor faQTLs for docosadienoic acid for both seed lots were detected at the same region on A05 where a gene-targeted marker for the *BrFAD2* gene mapped with its *cis*-eQTL and where also an eQTL hotspot was located (Figure 4A and B – right panel; Supplementary Table S3 and S4). If we also consider minor faQTLs, the association of A03, A04 and A05 with MUFAs, SFAs and PUFAs, respectively, is not perfect anymore.

The eQTL hotspots on A03, A04 and A05 are interesting for further investigation towards candidate genes for lipid metabolism and *Brassica* oil crop improvement. At the major eQTL hotspot on A09 (89 *trans*-eQTLs), where the *cis*-regulated *transparent testa 8 (BrTT8)* gene for seed colour is located we did not detect major faQTL.

We looked for co-location of eQTLs of the genes *FAE1*, *TAG1* (also called *DGAT1*), *FAD2* and other *FAD* genes with faQTLs, as these genes were reported as candidate genes for the synthesis of linoleic acid, linolenic acid, erucic acid, oleic acid or total oil content in *A. thaliana* and *B. napus* (Peng et al., 2010; Yang et al., 2012; Lee et al., 2013). *BrFAE1* has two paralogs on A01 and A03, which only have *trans*-eQTLs on A02, A03, A05 and A07, and *BrTAG1* on A07 had a *trans*-eQTL on A05 (Supplementary Figure S1). *Trans*-eQTLs of *BrFAE1* gene co-located mainly with faQTLs of MUFAs on A03, PUFA linolenic acid on A05, and SFAs and MUFAs on A07, while *trans*-eQTL of *BrTAG1* was co-located with eQTL hotspot on A05, mainly with faQTLs of PUFAs (Figure 3). The *BrFAD* series genes (*BrFAD2*, *BraFAD5* and *BrFAD7*) from the FA desaturation pathway that are co-located within a physical range of 19.5-21.6 Mb on A05 all had a *cis*-eQTL on A05 at the map position (89.1 cM) of a *BrFAD2* gene-targeted marker, co-locating with faQTLs for SFA lignoceric acid, MUFAs palmitoleic, oleic, eicosenoic and erucic acids, and PUFAs linoleic, eicosadienoic and docosadienoic acids (Figure 3).

Co-expression network of lipid related genes and fatty acids

Pearson correlation coefficients were calculated based on transcript abundance (after correction for seed colour) of 662 probes (488 BraIDs), with at least one eQTL, and 17 FAs (2011 seed lot) to construct a co-expression network (Supplementary Figure S2). This figure shows that many genes from the lipid metabolic pathways are involved in the regulation of FAs. In most previous QTL studies in related species *A. thaliana* and *B. napus*, mostly *BrFAD* series genes, *BrFAE1*, *BrTAG1* and a small subset of other genes are usually reported as being responsible for genetic regulation of fatty acids. We here not only show the interactions between genes but also whether they are *cis*- or *trans*-regulated. The co-expression network was constructed using two approaches: in the first one, correlations among all probes and FAs were considered (Supplementary Figure S2), while in the second approach, we considered only the probes associated with fatty acids, and using WGCNA (we call this a FAs-centered network). In the first approach, a higher degree of connection among genes and fatty acids indicates a major hub gene that could potentially be a major regulator of lipid metabolism. In the second method, the degree of connection of genes indicates the numbers of significant associations with only FAs, where a gene with a high degree of connection could be potentially a major gene involved in the regulation of those FAs.

In both analyses, a very similar set of genes with a high degree of connection was observed. The gene *BrPLA2-ALPHA* (*phospholipase A2*, BraID: Bra038125) with a *cis*-eQTL on A05 had the highest degree of connection in both network analyses (Figure 5; Table 2), suggesting that this is an essential gene for lipid metabolism. Since *BrPLA2-ALPHA* is one of the major hub genes, a *BrPLA2-ALPHA* centered sub-network (Figure 5) was extracted from the whole genes-FAs co-expression network (Supplementary Figure S2). Figure 5 shows that genes from the lipid metabolic pathways

and seed storage are interacting with this major hub gene. From the top 25 genes, based on their degree of connection from both analyses, 16 were selected in both (Table 2). Among those 16 genes, three genes had a *cis*-eQTL on A05 and two genes had a *cis*-eQTL on A09, while 11 genes had a *trans*-eQTL on A09 and four genes had a *trans*-eQTL on A03 and only four genes from single analyses had a *trans*-eQTL on A04 (Table 2).

Sub-networks were extracted for economically important FAs: erucic acid, oleic acid, linoleic acid and linolenic acid, which were also the most predominant FAs in this population as shown in Figure 1. All the FAs and gene nodes that were directly linked with each of these four metabolites were included in the sub-networks.

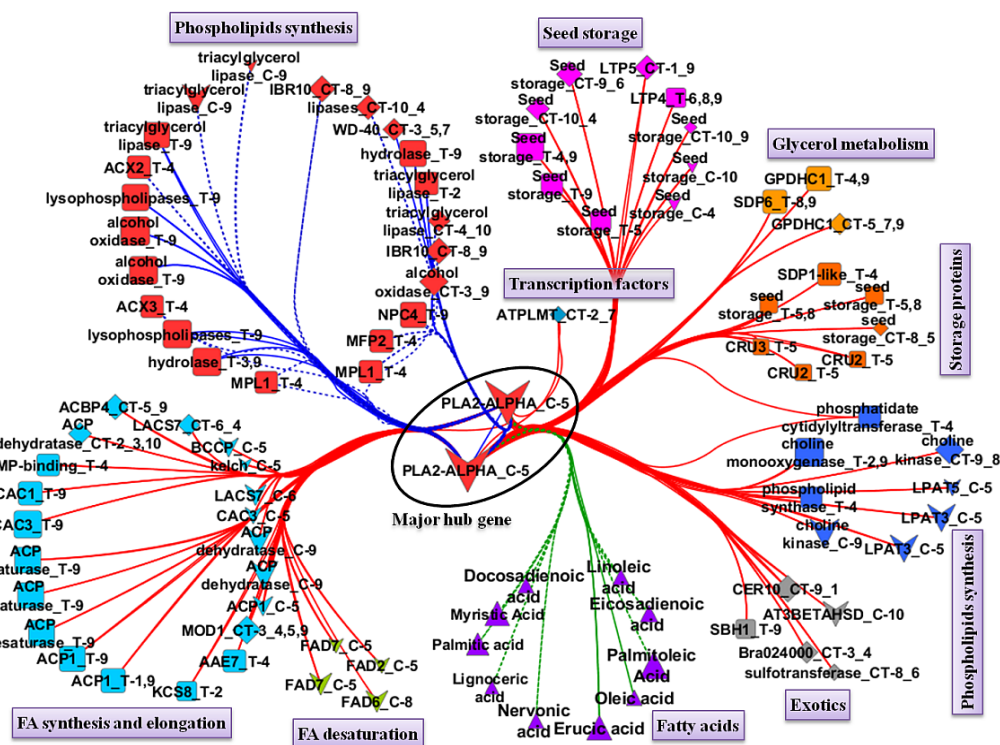


Figure 5: Major hub gene *BrPLA2-ALPHA* centered gene-coexpression network using probes with at least one eQTL (after correction). Only genes that are connected with *BrPLA2-ALPHA* gene in the whole gene-metabolite co-expression network (Supplementary Figure S2) are shown. Nodes represent genes or fatty acids (FAs). Genes from different pathways are shown in different node colours, while metabolites are shown in triangle-shaped nodes in purple colour. Edges represent high absolute correlations ($|r| > 0.5$). Edges connecting the major hub gene and genes from the same pathway (lipid degradation) are in blue, edges between the major hub gene and genes from the other pathways are in red, and edges connecting genes and FAs are in green. Edges connecting genes (other than with the major hub gene) of different pathways have been left out to improve the visibility of the network. The shapes of gene nodes indicate *cis*-eQTLs (V-shape), *trans*-eQTLs (square) and *cis*-/*trans*-eQTLs (diamond). Node names are coded by concatenating gene name, *cis*-/*trans*-regulation (separated by “_”) and linkage group (separated by “-”). For example, node “FAD2_C-5” indicates gene “FAD2”, “C” for *cis*-eQTL and “5” for linkage group A05, where the *cis*-eQTL was detected. In the case of a *cis*-/*trans*-eQTL, “CER10_CT-9_1” indicates gene “CER10” and “CT-9_1” indicates a *cis*-eQTL on A09 and *trans*-eQTLs on A01. Solid lines indicate positive correlations and dotted lines indicate negative correlations. All the gene names are prefixed with “Br” because of *Brassica rapa* gene nomenclature. Multiple occurrence of the same gene names represents genes with multiple paralogues or probes.

Table 2: List of the top 25 genes based on degree of connection in each of two network approaches, NetworkAnalyzer (Cytoscape) and WGCNA (FAs-centered). Sixteen genes are in common between the two lists, resulting in a total of 34 genes over the two approaches.

SN	Gene Symbol	BraID	At-locus	Network analysis method	Degree of connection		eQTL		Saturated Fatty Acids (SFAs)					Monounsaturated Fatty Acids			Polyunsaturated Fatty Acids			MapMan pathway	
					NetworkAnalyzer	WGCNA			Myristic	Stearic	Palmitic	Arachidic	Behenic	Palmitoleic	Oleic	Eicosenoic	Erucic	Docosadienoic	Linoleic		Eicosadienoic
1	PLA2-ALPHA	Bra038125	At2g06925	Both	91	5	A05	-	✓	-	-	-	-	✓	-	-	-	✓	✓	✓	Lipid degradation
2	alcohol oxidase	Bra013391	At4g19380	Both	90	2	-	A09	✓	-	-	-	-	✓	-	-	-	-	-	-	Lipid degradation
3	choline monooxygenase	Bra024118	At4g29890	Both	88	3	-	A02	✓	✓	-	-	-	✓	-	-	-	-	-	-	Phospholipid synthesis
4	lysophospholipases	Bra003814	At1g74210	Both	87	4	-	A09	-	✓	✓	-	-	✓	-	-	✓	-	-	-	Lipid degradation
5	choline kinase	Bra028032	At4g09760	Both	87	4	A09	A03, A08	✓	-	✓	-	-	✓	-	-	✓	-	-	-	Phospholipid synthesis
6	hydrolase	Bra016558	At1g18360	Both	87	2	-	A03, A09	✓	-	-	-	-	✓	-	-	-	-	-	-	Lipid degradation
7	alcohol oxidase	Bra012548	At4g19380	Both	87	2	A03	A09	-	-	✓	-	-	✓	-	-	-	-	-	-	Lipid degradation
8	LPAT3	Bra030448	At1g51260	Both	84	6	A05	A09	✓	✓	✓	-	-	✓	-	✓	✓	-	-	-	Phospholipid synthesis
9	CER10	Bra007154	At3g55360	Both	81	3	A09	A01	-	-	✓	-	-	✓	-	-	✓	-	-	-	Exotics
10	IBR10	Bra039860	At4g14430	Both	79	2	A08	A09	✓	-	-	-	-	✓	-	-	-	-	-	-	Lipid degradation
11	NPC4	Bra021355	At3g03530	Both	78	4	-	A03, A09	✓	-	✓	-	-	✓	-	-	✓	-	-	-	Lipid degradation
12	stearoyl-ACP desaturase	Bra021427	At3g02630	Both	78	2	-	A09	✓	-	-	-	-	✓	-	-	-	-	-	-	FA synthesis and FA elongation
13	lipases	Bra035263	At4g10955	Both	77	3	-	A03, A07, A09	-	-	✓	-	-	✓	-	-	✓	-	-	-	Lipid degradation
14	AT3BETAHSD	Bra015621	At1g47290	Both	74	2	A10	A05	-	-	✓	-	-	✓	-	-	-	-	-	-	Exotics
15	GPDHC1	Bra029669	At2g41540	Both	73	2	A05	A07, A09	✓	-	-	-	-	✓	-	-	-	-	-	-	Glyceral metabolism
16	SDP6	Bra035180	At3g10370	Both	69	3	A07	A05, A08, A09	✓	-	✓	-	-	✓	-	-	-	-	-	-	Glyceral metabolism
17	MOD1	Bra013159	At2g05990	NetworkAnalyzer	86	1	A03	A04, A05, A09	-	-	-	-	-	✓	-	-	-	-	-	-	FA synthesis and FA elongation
18	triacylglycerol lipase	Bra007686	At3g62590	NetworkAnalyzer	86	1	A09	-	-	-	-	-	-	✓	-	-	-	-	-	-	Lipid degradation
19	CAC3	Bra000037	At2g38040	NetworkAnalyzer	85	1	A03	-	-	-	-	-	-	✓	-	-	-	-	-	-	FA synthesis and FA elongation
20	ACP dehydratase	Bra038539	At2g22230	NetworkAnalyzer	84	1	A09	A02	-	-	-	-	-	✓	-	-	-	-	-	-	FA synthesis and FA elongation
21	stearoyl-ACP desaturase	Bra008631	At3g02630	NetworkAnalyzer	77	1	-	A09	-	-	-	-	-	✓	-	-	-	-	-	-	FA synthesis and FA elongation
22	ACP1	Bra039471	At3g05020	NetworkAnalyzer	77	1	-	A09	-	-	-	-	-	✓	-	-	-	-	-	-	FA synthesis and FA elongation
23	ACBP4	Bra039439	At3g05420	NetworkAnalyzer	75	1	A05	A01, A09	-	-	-	-	-	✓	-	-	-	-	-	-	FA synthesis and FA elongation
24	lipases	Bra002174	At5g18630	NetworkAnalyzer	73	1	-	A04	-	-	-	-	-	✓	-	-	-	-	-	-	Lipid degradation
25	LTP5	Bra038908	At3g51600	NetworkAnalyzer	71	1	A01	A06, A09	-	-	-	-	-	✓	-	-	-	-	-	-	Lipid transfer proteins
26	WD-40	Bra001726	At3g18860	WGCNA	62	5	A03	A05	-	✓	-	✓	-	-	✓	✓	✓	-	-	-	Lipid degradation
27	LTP4	Bra020323	At5g59310	WGCNA	52	3	-	A06, A08, A09	-	✓	-	-	-	-	-	✓	✓	-	-	-	Lipid transfer proteins
28	ACBP3	Bra019240	At4g24230	WGCNA	51	3	A03	A05, A07	-	✓	-	-	-	-	-	✓	✓	-	-	-	FA synthesis and FA elongation
29	LTP5	Bra012847	At3g51600	WGCNA	31	3	A03	-	-	✓	-	-	-	-	-	✓	✓	-	-	-	Lipid transfer proteins
30	Protein kinase	Bra034040	At1g66980	WGCNA	13	3	-	A04, A10	-	-	✓	-	✓	-	-	✓	-	-	-	-	Lipid degradation
31	mtACP2	Bra035355	At1g65290	WGCNA	-	3	-	A04	-	-	✓	-	-	-	-	✓	✓	-	-	-	FA synthesis and FA elongation
32	phosphoethanolamine NMT2	Bra018740	At1g48600	WGCNA	-	5	No eQTL	No eQTL	-	✓	✓	-	-	✓	-	✓	✓	-	-	-	Phospholipid synthesis
33	Unknown	Bra001486	At3g13062	WGCNA	-	3	No eQTL	No eQTL	-	-	✓	-	-	-	-	✓	✓	-	-	-	Lipid signalling
34	ATVPS34	Bra027152	At1g60490	WGCNA	-	3	No eQTL	No eQTL	-	✓	-	-	-	-	-	✓	✓	-	-	-	Lipid signalling
Note: ✓ indicates a significant association of a gene with FAs in WGCNA																					

Note: ✓ indicates a significant association of a gene with FAs in WGCNA.

Oleic acid (C18:1)

The MUFA oleic acid had faQTLs on A03, A05 and A10 in both years' seed lots with explained variance ranging from 8% to 18%, plus additional QTLs on A01 and A02 in 2009 (Figure 3; Supplementary Table S3 and S4). In the oleic acid centered sub-network, ~19 genes were connected to oleic acid, with genes from the lipid metabolism pathways, with *cis*- and or *trans*-acting regulation, were involved in oleic acid biosynthesis (Figure 6A). The eQTLs of these genes co-localized with several oleic acid faQTLs. Overall, seven genes had a *cis*-eQTL and three genes had a *trans*-eQTL on A03 (Figure 6A). This included two lipid degradation genes, two FA synthesis genes (among a total of three genes), a FA elongation gene and three out of seven seed storage protein genes (Figure 6A). Four genes had a *cis*-eQTL and four had a *trans*-eQTL on A05. This included two lipid degradation genes: *BrPLA2-ALPHA* and *BrWD-40*, one FA desaturation gene *BrFAD7*, and the glycerol metabolism gene *BrGPDHC1*. Only one gene *BrGPAT6* had a *trans*-eQTL on A10 and also on A01, A03 and A09 (Figure 6A). For some of the genes, a *cis*- or *trans*-eQTL was detected also on A01, A04, A07, A09 and A10 (Figure 6A).

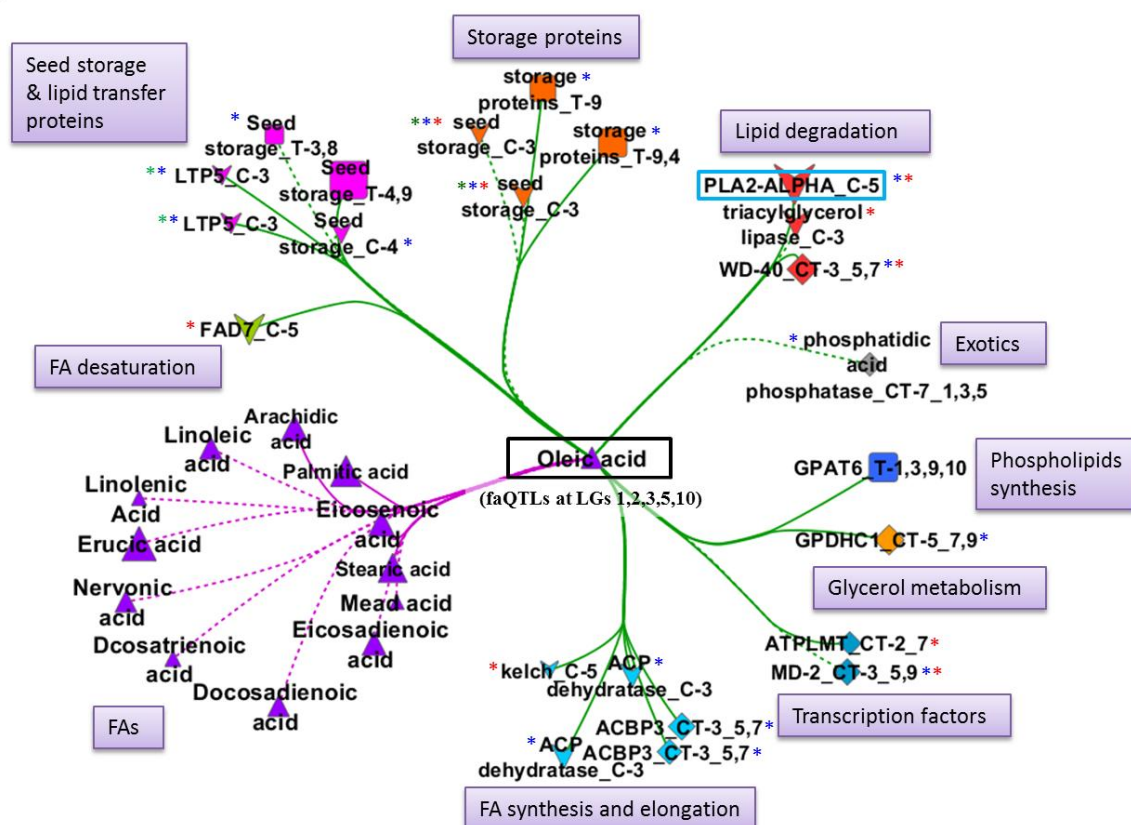


Figure 6A: Oleic acid centered gene-metabolite co-expression sub-network. All the nodes that have a connection with oleic acid were extracted from the main network shown in Supplementary Figure S2. Multiple occurrence of the same gene names represents genes with multiple paralogues or probes.

Erucic acid (C22:1)

The main fatty acid, the MUFA erucic acid, had faQTLs on A03 and A09 for the 2011 seed lot with explained variance 16% and 14%, respectively (Figure 3; Supplementary Table S4). For the 2009

seed lot, putative faQTLs (LOD scores between 2 and 3) were detected on A03 and A09 as well, with additional faQTLs on A01, A02, A05, and A07 (Figure 3; Supplementary Table S3). In the erucic acid-centered sub-network, ~16 out of the more than 50 genes had a *cis*- or *trans*-eQTL on A03 (Figure 6B), and, in general, genes related to seed storage proteins, FA synthesis and elongation and lipid degradation had *cis*-eQTLs while genes related to phospholipid synthesis, lipid binding and transcription factor had *trans*-eQTLs (Figure 6B). On A09, 6 phospholipid synthesis and seed storage protein genes and *BrCER10* (very long chain fatty acid elongation gene) had a *cis*-eQTL. Twenty-four genes, of which the majority of genes are involved in lipid degradation, had a *trans*-eQTL on A09 (Figure 6B).

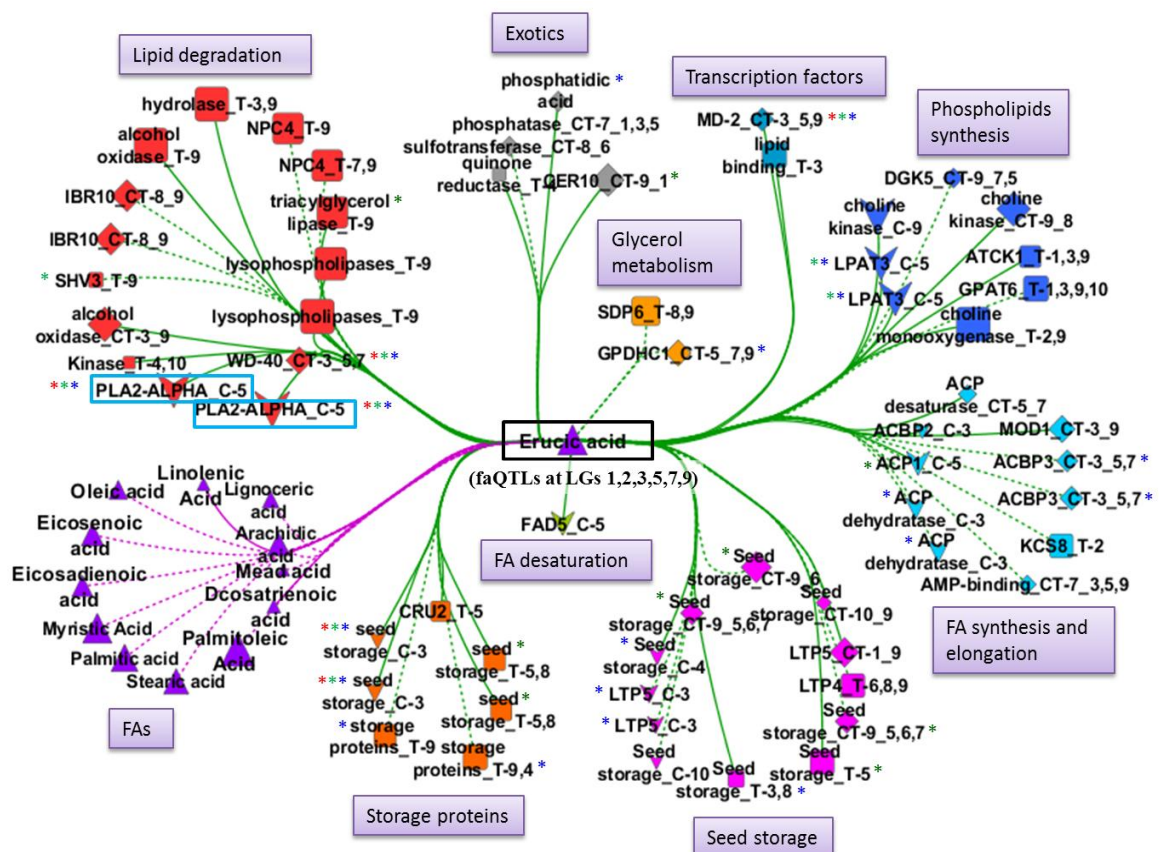


Figure 6B: Erucic acid centered gene-metabolite co-expression sub-network. All the nodes that have a connection with erucic acid were extracted from the main network shown in Supplementary Figure S2. Multiple occurrence of the same gene names represents genes with multiple paralogues or probes.

Linoleic acid (C18:2) and linolenic acid (C18:3)

FaQTLs were mapped for the PUFAs linoleic and linolenic acids. A major faQTL of explained variance 36-45% was detected for linoleic acid in both experiments on A05, where we also mapped a gene-targeted marker for *BrFAD2* and its *cis*-eQTL (Supplementary table S3 and S4). This faQTL overlapped with the eQTL hotspot on A05 (Figure 3; Figure 4A and B – right panel). In the linoleic and linolenic acids centered sub-network, only three genes coding for seed storage proteins were correlated with linolenic acid, while more than 25 genes were associated with linoleic acid (Figure 6C). Among these linoleic associated genes, nine genes had a *cis*-eQTL and five

genes had a *trans*-eQTL on A05. In general, the genes related to FA synthesis and elongation, phospholipids, lipid degradation (including *BrPLA2-ALPHA*) and FA desaturation (*BrFAD7* gene) had *cis*-eQTLs while seed storage genes and a transcription factor (*BrMD-2*) had *trans*-eQTLs on A05 (Figure 6C). Genes such as *BrPLA2-ALPHA* had a high degree of connection (> 69 edges) based on their co-expression ($|r| \geq 0.5$) with other genes and FAs (Table 2).

In case of linolenic acid, only three seed protein encoding genes were connected: two with a *cis*-eQTL on A03 and one with *trans*-eQTLs on A04 and A09 (Figure 6C), while its minor faQTLs were detected on A05, A08 and A10, so, none of them co-localized (Figure 3).

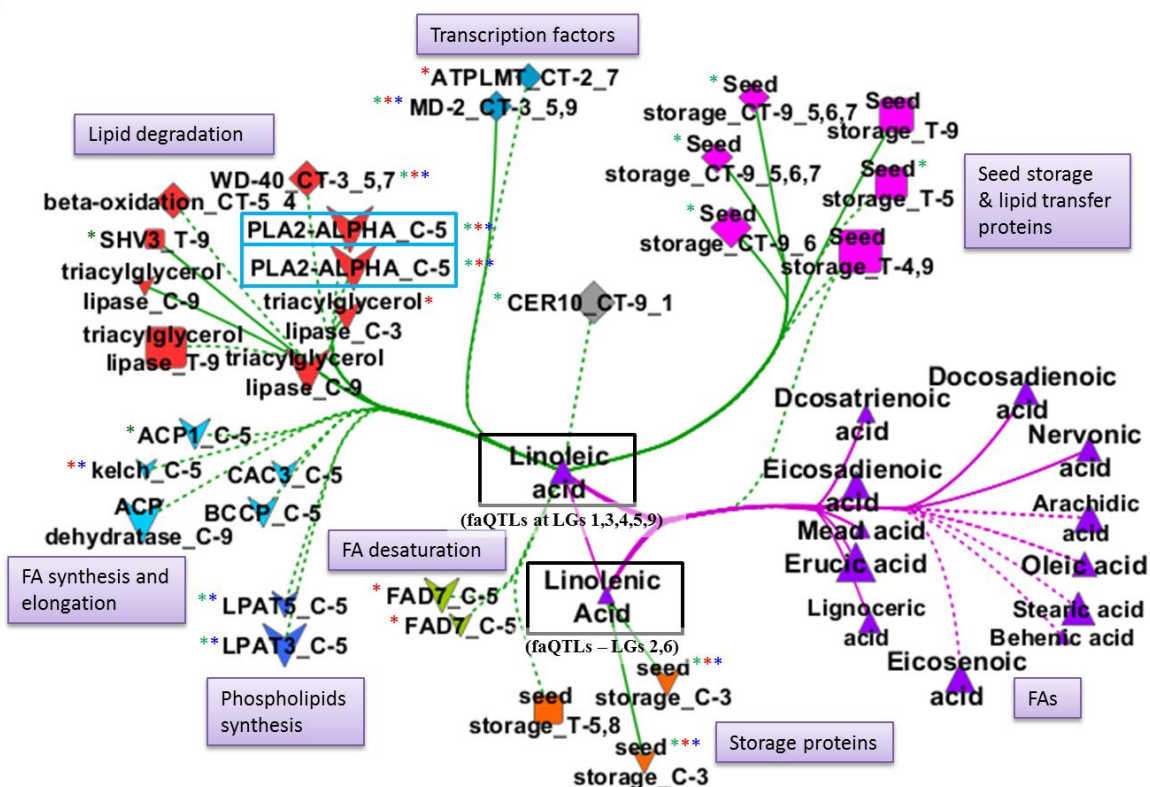


Figure 6C: Linoleic and linolenic acids centered gene-metabolite co-expression sub-network. All the nodes that have connections with either linoleic acid or linolenic acid were extracted from the main network shown in Supplementary Figure S2. Multiple occurrence of the same gene names represents genes with multiple paralogues or probes.

Genetic co-regulation of erucic acid, oleic acid and linoleic and linolenic acids

On A05, a major faQTL for linoleic acid and minor faQTLs for erucic and oleic acid were detected, together with major QTLs for the two PUFAs eicosadienoic acid and docosadienoic acid, co-locating with an eQTL hotspot (Figure 3; Figure 4A and B – right panel). In the network analyses for each of these FAs, genes having either *cis*- or *trans*-acting eQTLs on A05 had a high degree of connection and also high correlations with these FAs (Table 2; Figure 5; Supplementary Figure S2). These results suggest that these genes are likely major hub genes and have an essential role in the metabolic pathway of these FAs.

Two lipid degradation genes *BrPLA2-ALPHA* and *BrWD-40*, one seed storage protein gene (*Bra019067*) and the transcription factor *BrMD-2* were present in all three sub-networks of erucic-,

oleic- and linoleic- and linolenic- acid (Figure 6). *BrPLA2-ALPHA* had a *cis*-eQTL on A05, and had the highest degree of connection with 91 nodes in the whole network in Cytoscape, while WGCNA network analysis (FAs-centered network) also identified this gene as being associated with five FAs: the SFA myristic acid, the MUFA palmitoleic acid, and the PUFAs linoleic acid, eicosadienoic acid and docosadienoic acid, each with a faQTL on A05.

On A03, QTLs for the three FAs erucic acid, oleic acid, and linoleic acid, plus additional QTLs for another 4-5 FAs were mapped (Figure 3). The seed storage protein encoding gene (*Bra019067*), *BrWD-40* and a transcription factor *BrMD2* had a *cis*-eQTL on A03, in addition *BrWD-40* and *BrMD2* genes had *trans*-eQTLs on A05 and A07 or A09 (Figure 6), where additional faQTLs for one or more of these four FAs mapped. These genes had a high degree of connection (Table 2). In WGCNA analysis, the phospholipid synthesis gene *BrLPAT3* (1-acylglycerol-3-phosphate O-acyltransferase; *Bra030448*) had the highest degree of connection and this gene had a *cis*-eQTL on A05, overlapping with *cis*-eQTL of the *BrPLA2-ALPHA* gene and major faQTLs of the PUFAs linoleic acid (C18:2), eicosadienoic acid (C20:2), docosadienoic acid (C22:2) and minor QTLs of MUFAs palmitoleic acid (C16:1), oleic acid (C18:1), eicosenoic acid (C20:1) and erucic acid (C22:1) (Table 2; Figure 3).

Discussion

Seed FA composition and content per se are important for seed oil crops, but also as a source of energy for the emerging seedling (Wang et al., 2010b; Zhang et al., 2012). In this paper we describe the FA composition in seeds of a DH progeny from a cross between an oil type and a vegetable type *B. rapa*. We combined the QTLs for FA composition in seeds with eQTL analysis followed by gene co-expression network analysis with the aim to identify major regulatory genes. Systems genetics has been widely used as an approach to integrate data at metabolic and gene expression levels in segregating populations. Interestingly, major faQTLs for MUFAs, SFAs and PUFAs mapped on different linkage groups, respectively on A03, A04 and A05 (Figure 3), which may suggest some level of genome organization according to the fatty acid types (MUFAs, SFAs and PUFAs). However, the biosynthetic processes of these FAs share part of their biochemical pathways; therefore, there could be common regulatory genes or genetic interactions in the regulation of these different FAs. This could be indicated by the fact that in some cases minor faQTLs of one type co-locate with a major faQTL of a different FA type. In this study, we observed that SFAs in general were positively correlated with MUFAs, apart from erucic and nervonic acids, but PUFAs had a low correlation with SFAs and MUFAs (Figure 2).

In genome-wide genetic studies, the presence of hidden confounding factors, such as unobserved covariates or unknown subtle environmental perturbations can lead to spurious marker-trait associations or mask real genetic association signals. These confounding factors can be introduced or are inherent to the data at different steps while conducting experiments (Fusi et al., 2012). In this study, we observed the effects of two such confounding factors. The first was flowering time, and, related to that, timing of seed set and seed maturation on faQTLs (metabolite level) at A02 (co-localizing gene-targeted marker for *BrFLC2* and a major flowering QTL (Figure 3; Figure 1 in

Chapter 6). Flowering time variation is very obvious in this DH population and the *BrFLC2* gene at A02 (16.7 cM) is the major regulator of flowering time, with a non-functional allele in the yellow sarson parent (Wu et al., 2012; Xiao et al., 2013). In the 2009 seed lot, when flowering was asynchronous, many faQTLs co-located with the *BrFLC2* gene, which can point to pleiotropy or linkage of QTLs for flowering time and FAs. However, the synchronization of flowering time in the 2011 experiment removed this confounding effect on faQTL detection. The synchronization of flowering time of all DH lines resulted in similar environmental conditions during seed development, which is important to study the genetic variation of seed metabolites and seed quality related traits. Other studies also reported the possibility of such confounding or pleiotropic effects of major genes on many developmental traits, for example at the *ERECTA* gene in *A. thaliana* (Stinchcombe et al., 2009) and the *EARLINESS* locus in potato (Hurtado-Lopez, 2012). Despite the differences during seed development in 2009 and 2011 (asynchronous versus synchronous), very high correlations between the two seed lots (2009 and 2011) for each FA across the DH genotypes were observed (Figure 2). Major faQTLs were also always detected in both years, while minor faQTLs varied between years.

The second confounding factor was the effect of seed colour on eQTLs (transcript level) at A09 (Figure 4A and B – left panel). In addition to variation in morphotypes, and, as a result, in many other morphological and biochemical traits (Zhao et al., 2005; Pino Del Carpio et al., 2011b; Xiao et al., 2013; Xiao et al., 2014), the yellow sarson and pak choi parents have contrasting seed coat colour (Basnet et al., 2013), which also introduced a confounding effect for eQTL mapping. A strong seed colour QTL with 33% explained variance was mapped on A09 (data not shown); the causal gene, the *bHLH* transcription factor *BrTT8*, was cloned and its role in seed color was functionally validated in *B. rapa* (Li et al., 2012) and in other species (Padmaja et al., 2014). At this *BrTT8* position, a large *trans*-eQTL hotspot was mapped, even after correction for seed colour (Figure 4A and B – left panel). In contrast with the eQTL hotspot, only minor faQTLs for erucic acid and eicosadienoic acid and a few additional minor faQTLs co-localized with this seed colour QTL (Figure 3). Possible explanations for the confounding effect of seed colour only at the transcript level but not at the metabolite level might be the fact that genetic regulation of the metabolic process is not always completely hierarchical in translating gene expression variation to metabolite regulation. Ter Kuile and Westerhoff (2001) reported a lack of hierarchical genetic control over metabolic flux in their study on the regulation of the glycolytic pathway. Additionally, the absence of a strong one-transcript one-metabolite relationship is quite common in the regulation of biological processes due to the complexity involved in the extrapolation of gene expression variation to changes in metabolite content, such as post-transcriptional modification and epigenetic regulation. Another explanation could be that FA metabolites and transcripts were measured in different developmental stages: in mature ripe seeds and in developing seeds (28 DAP), respectively, which could lead to different levels of interactions at different stages of seed development.

In this genetical genomics study, we were able to subset to only those genes that had eQTLs, detect eQTL hotspots and identify *cis*- and *trans*- acting eQTLs (Figure 4). Following this genetical

genomics approach, an eQTL-guided gene co-expression network was constructed that allowed us to identify (possible) candidate genes and their regulatory interactions for lipid metabolism.

Genes with a high degree of connection in a network could possibly be major regulators of a pathway. Those genes could be essential genes in the sense that variation in these genes could change the pathway. In contrast, genes with a lower degree of connection could indicate genes that play a role in modifying FA content or composition. From the two types of network analyses carried out in this study (using Cytoscape and WGCNA), the top 25 genes with a high degree of connection were selected, which are likely to be key drivers in lipid metabolism; 16 genes were in common between these two lists, illustrating that these approaches were quite effective in selecting the most essential genes. These 16 genes belong to pathways such as lipid degradation, FA synthesis and elongation, phospholipid synthesis, glycerol metabolism and lipid transfer proteins (Table 2). Interestingly those top ranking genes were from different pathways, inferring an extensive coordination among biosynthetic pathways in lipid metabolism in *B. rapa* seed. Those top 16 genes had *cis*- and *trans*-eQTLs mainly co-localized with major faQTLs on A03, A05 and A09 (Table 2), suggesting that those regions harbour the possible key regulator genes. Among the top selected genes, the lipid degradation pathway gene *BrPLA2-ALPHA* (phospholipase A2-alpha, Bra038125) had the highest degree of connection and a *cis*-eQTL on A05, co-locating with major PUFA faQTLs for linoleic acid (C18:2) and eicosadienoic acid (C20:2) and minor faQTLs for other FAs (Table 2). Ryu et al. (2005) functionally characterized *BrPLA2-ALPHA* gene and concluded that it has an acyl preference for linoleic acid over palmitic acid in phospholipid hydrolysis in *A. thaliana*. They also reported a role of this gene in the release of free fatty acids and lysophospholipids from membrane phospholipids. However, we have not found any other study reporting this gene as a potential regulator of seed FA composition in *A. thaliana*, *B. napus* or other oil crops. Many studies did report however that *cis*-eQTL for master regulator genes, mapped under *trans*-eQTL hotspots (Civelek and Lusi, 2014; Wang et al., 2014a), similar to what we found for *BrPLA2-ALPHA* on A05.

The explanation that such key regulator genes, like *BrPLA2-ALPHA* in this study, are generally not reported as potential regulators of, in our case, seed FA composition, is that these genes are often highly conserved in regulatory networks during evolution (Khurana et al., 2013), and are less likely to be genetically perturbed in mapping populations (Mäkinen et al., 2014). In our study, we still found a *cis*-eQTL for the highly connected gene *BrPLA2-ALPHA*, which might be due to the different selection history of oil and vegetable types.

Additional genes from the top 25 were genes in the FA synthesis and FA elongation pathway, e.g. *BrCAC3*, *BrMOD1*, ACP dehydratase, two paralogs of ACP desaturase, *BrACP1* and *BrACBP4* (Table 2), whose functional roles are described for converting acetyl-CoA to malonyl-CoA chain at the beginning of the pathway in the Kyoto Encyclopedia of Genes and Genome (KEGG) pathway database.

The MUFA erucic acid was the most abundant fatty acid (47-55.8% of total dry weight) in both parents (yellow sarson and pak choi) and their DH progenies (Figure 1; Supplementary Table S1). Lühs et al. (1999) also reported high erucic acid content (54.8%) in yellow sarson seeds. Breeding

for low erucic acid can cause a decrease in the total oil content if low erucic acid is not compensated for by an increase of other FAs. Even though the MUFA oleic acid is a substrate for both erucic acid and for the PUFA linoleic acid, oleic acid was strongly negatively correlated only with linoleic acid but not with erucic acid (Figure 2). Linoleic acid shares the genomic region of its major faQTL on A05 (36-46% explained variance; yellow sarson effect) with a QTL for its precursor oleic acid (9-13% explained variation) and the chain elongated erucic acid (9% explained variation) with opposite allelic effects (Figure 3; Supplementary Table S3 and S4), suggesting that these FAs share regulatory elements. Erucic acid, oleic acid, linoleic acid and linolenic acid are the most predominant FAs (Figure 1) and economically important for oil quality. Therefore, we further looked into detail to eQTL-guided co-expression networks particularly associated with those four FAs (called sub-networks), with the aim to unravel the underlying gene regulatory networks. The six genes *BrPLA2-ALPHA*, *BrWD-40*, three seed storage genes (*Bra024983*, *Bra019067* and *Bra024983*) and one transcription factor *BrMD-2* were in common among all those three sub-networks (Figure 6), inferring essential roles of these genes in the biosynthesis of these FAs. *BrWD-40* has a protein domain that regulates chromatin dynamics and transcription of the evolutionary well conserved gene *Adipose* (*ADP*). The loss of function allele of *ADP* promotes TAG (triacylglycerol) storage in *Drosophila* flies (Häder et al., 2003). *MD-2* also has a conserved domain present across plants, animals and bacteria and is involved in lipopolysaccharides binding (Inohara and Nuñez, 2002). Most of the genes are shared by at least two of the three sub-networks (Figure 6) and might have pleiotropic effects on lipid metabolism. There were also many genes only present in the erucic acid sub-network (Figure 6B). Erucic acid has a larger gene metabolite co-expression network (> 50 genes) than oleic acid (~ 20 genes) and linoleic- and linolenic acids (> 25 genes), and also has many minor faQTLs (Figure 6), implying a polygenic inheritance and more complex gene network (Figure 6B). For an eQTL based gene network, the genetic composition of the mapping population could pose a limiting factor in selecting genes for network construction; therefore, QTL mapping for fatty acids and gene expression in multiple populations from genetically diverse parents could confirm our network components.

Well studied genes such as *FAE1*, *TAG1* and *FAD2* in *A. thaliana*, *B. napus* and other oilseed crops were reported in the literature, from QTL analyses, as candidate genes for the synthesis of linoleic acid, linolenic acid, erucic acid, oleic acid or total oil content (Peng et al., 2010; Yang et al., 2012b; Lee et al., 2013). *Cis*-eQTL of *BrFAD2* and other *BrFAD* genes (*BrFAD5* and *BrFAD7*), and *trans*-eQTL of *BrFAE1* and *BrTAG1* genes co-localized with faQTLs for oleic acid (C18:2), linoleic acid (C18:2), erucic acid (C22:1) and other fatty acids (Supplementary Figure S1; Figure 3). It is likely that those genes are regulated by some of the key regulators present on A05; those genes were not highlighted in the network analyses due to their low degrees of connection, but could play a role in modifying FA content or composition. The expression profile of *BrFAD7* was correlated with linoleic acid and oleic acid content, while *BrFAD5* was correlated with erucic acid content (Supplementary Figure 2). *BrFAD2* had only a weak correlation ($r < 0.5$) with any of the FAs, but its expression was correlated with that of *BrFAD5* and *BrFAD7* ($r > 0.5$). Several studies in *B. napus* and *Arabidopsis* reported that *BrFAD2* regulates the conversion of oleic acid to linoleic acid in the

endoplasmic reticulum (ER), while *BrFAD5* desaturates C18:0 acyl carrier protein to oleic acid, and *BrFAD7* desaturates linoleic acid to linolenic acid in plastids (Zhang et al., 2012). Therefore, there could be interactive roles among *BrFAD* genes.

In conclusion, we were able to identify major regulatory genes involved in the genetic regulation of lipid metabolism and those genes belonging to the different lipid metabolic pathways: lipid degradation, FA synthesis and elongation, phospholipid synthesis, glycerol metabolism, transfer protein, signaling and very long chain elongation (Table 2). Those results suggest the need of a global study of lipid metabolism, rather than a strict focus on the FA biosynthesis pathway per se. This study gives a starting point for understanding the genetic regulation of lipid metabolism, by identification of a number of key regulatory genes, identified as major hub genes, and candidate genes for faQTLs. Finally, the data generated in this study will be valuable in *Brassica* breeding as it offers tools to breed for yield and optimal oil composition.

Acknowledgment

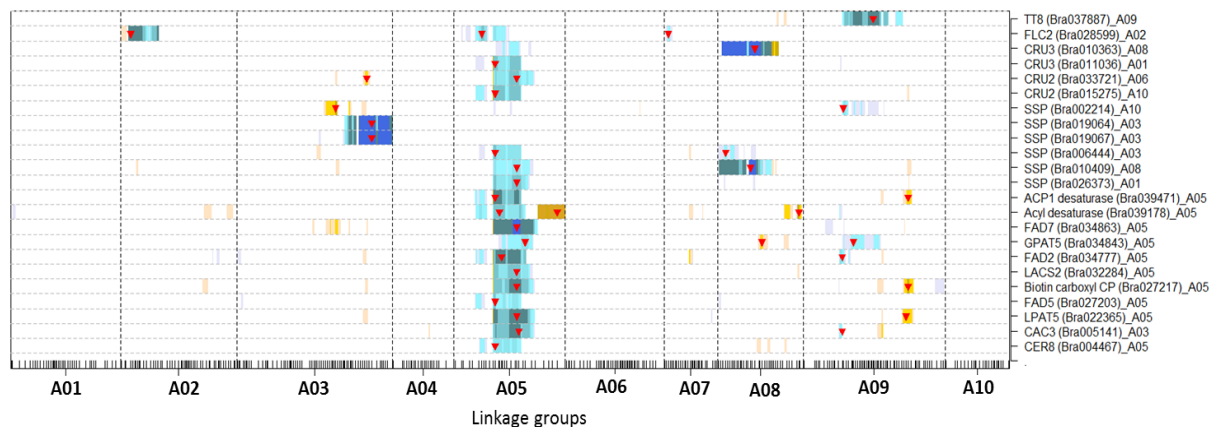
We thank the members of the Unifarm facility of Wageningen University and Research Centre (WUR) for taking care of the plants and all the necessary support, and for students and colleagues for their assistance during pollination. This work was supported by the funding from the Centre for BioSystems Genomics (CBSG), The Netherlands, which is a part of the Netherlands Genomics Initiative (NGI).

Supplementary Information

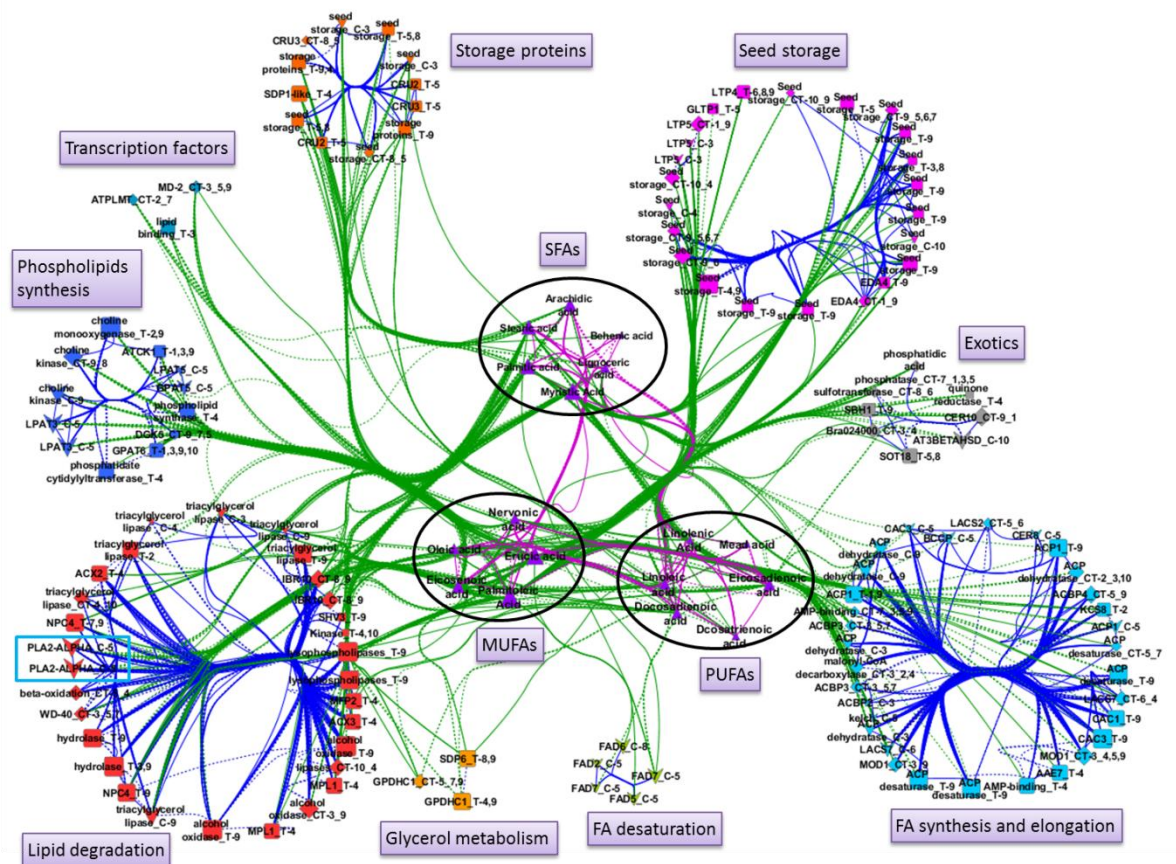
Supplementary Table S1: Summary statistics of fatty acids (FAs) measured in ripe seeds of the *Brassica rapa* DH population, including the parents yellow sarson and pak choi in the 2009 and 2011 seed lots.

Fatty acids	Formula	Types	DH lines, 2009					Yellow sarson, 2011					Pak choi, 2011					DH lines, 2011				
			Mean	Min.	Max.	Std. dev.	CV, %	Mean	Min.	Max.	Std. dev.	CV, %	Mean	Min.	Max.	Std. dev.	CV, %	Mean	Min.	Max.	Std. dev.	CV, %
Myristic acid	C14:0	SFA	0.0	0.0	0.1	0.0	31.6	0.04	0.03	0.04	0.001	3.4	0.04	0.04	0.05	0.01	13.1	0.04	0.02	0.1	0.01	24.2
Palmitic acid	C16:0	SFA	1.9	1.2	3.0	0.3	17.7	1.4	1.2	1.5	0.1	6.8	1.9	1.7	2.2	0.2	10.1	1.8	1.2	2.9	0.3	18.2
Palmitoleic acid	C16:1	MUFA	0.3	0.2	0.6	0.1	29.3	0.2	0.2	0.2	0.01	2.4	0.3	0.3	0.3	0.04	12.1	0.3	0.2	0.6	0.1	22.9
Stearic acid	C18:0	SFA	1.4	0.7	2.7	0.4	32.8	0.8	0.6	1	0.1	16.3	1	0.8	1.2	0.2	15.4	1.1	0.6	2.4	0.3	31.1
Oleic acid	C18:1	MUFA	18.8	9.1	28.3	4.3	22.8	13.7	11.4	15	1.5	11.0	17.3	16.6	18.0	0.6	3.6	15.2	7.9	25	3.8	24.7
Linoleic acid	C18:2	PUFA	11.6	6.1	16.4	2.5	21.4	10.8	10.2	11.2	0.4	3.4	11.0	9.8	13.7	1.8	16.7	11.9	6.7	16.5	2.4	19.7
Linolenic acid	C18:3	PUFA	5.3	2.1	9.2	1.3	24.7	6.8	5.8	8	0.8	11.5	8.2	7.5	9.1	0.7	8.6	6.4	3.1	10.7	1.3	21.0
Arachidic acid	C20:0	SFA	1.1	0.6	2.2	0.3	31.6	0.7	0.6	0.8	0.1	11.2	0.8	0.6	0.9	0.1	17.2	0.9	0.5	2.1	0.3	29.2
Eicosenoic acid	C20:1	MUFA	7.2	3.5	12.1	1.7	23.8	4.7	3.7	5.4	0.6	13.8	7.3	6.7	8.4	0.8	10.6	6.2	3.1	10.6	1.5	23.7
Eicosadienoic acid	C20:2	PUFA	0.3	0.1	0.6	0.1	33.4	0.2	0.2	0.3	0.03	13.2	0.4	0.3	0.5	0.1	24.2	0.3	0.1	0.6	0.1	34.6
Mead acid	C20:3	PUFA	0.0	0.0	0.1	0.0	58.5	0.04	0.03	0.04	0.005	12.4	0.1	0.1	0.1	0.01	16.7	0.1	0.02	0.1	0.02	33.7
Behenic acid	C22:0	SFA	1.3	0.6	2.6	0.3	26.6	1.1	0.9	1.2	0.1	8.6	0.8	0.6	1.1	0.2	26.0	1.2	0.8	2.3	0.3	23.6
Erucic acid	C22:1	MUFA	47.1	31.1	55.9	5.0	10.6	55.8	53.5	57.7	1.5	2.8	47.0	44.4	49.3	2.0	4.3	50.2	40.6	56.3	4	7.9
Docosadienoic acid	C22:2	PUFA	0.7	0.2	1.5	0.3	38.8	0.8	0.7	1.0	0.1	15.3	0.7	0.7	0.8	0.03	4.4	1.0	0.4	1.9	0.4	38.2
Docosatrienoic acid	C22:3	PUFA	0.1	0.0	0.4	0.1	41.7	0.2	0.2	0.3	0.04	18.4	0.2	0.2	0.2	0.03	14.0	0.2	0.1	0.5	0.1	40.1
Lignoceric acid	C24:0	SFA	0.6	0.3	1.6	0.2	38.7	0.5	0.4	0.5	0.05	10.3	0.4	0.3	0.6	0.1	29.2	0.5	0.3	1.3	0.2	29.3
Nervonic acid	C24:1	MUFA	1.7	1.0	2.6	0.3	17.0	1.8	1.6	1.9	0.1	5.1	2.0	1.5	2.2	0.3	17.3	2.0	1.3	3.2	0.3	17.6
Total oil	-	-	31.0	0.0	40.5	5.2	16.8	44.2	38.8	51	4.7	10.5	29.3	24.8	36.2	4.9	16.7	31.8	25.0	42.5	3.8	11.9

Note: SFA = Saturated fatty acid; MUFA = Monounsaturated fatty acid; PUFA = Polyunsaturated fatty acid; Min. = minimum; Max. = maximum; Std. dev = standard deviation; CV = coefficient of variation (%). FAs were measured in mass percentage of total oil content. Total oil content was measured in mass percentage of whole seed dry matter (zero moisture basis).



Supplementary Figure S1: Heatmap showing the eQTL LOD scores for gene expression of the probes measured in RT-qPCR. eQTL LOD profiles were calculated after correction for seed colour. *Brassica rapa* gene IDs (Bra IDs) are indicated between parentheses and the gene locations in the genome are preceded by “_” in the gene name. The darker the intensity, the higher the LOD score. Yellow indicates a QTL effect with high transcript abundance being associated with a yellow sarson allele while blue indicates a QTL effect with high transcript abundance being associated with the pak choi allele. Red triangles indicate the positions of QTL peak markers. The black coloured tick marks at the bottom indicate the markers in the linkage map, vertical dashed lines separate linkage groups. The horizontal dotted lines separate genes.



Supplementary Figure S2: Gene-metabolite co-expression network using probes with at least one eQTL (after correction for seed color). Nodes represent genes or fatty acids (FAs). Genes from different pathways are shown in different node colours, while metabolites are shown in triangle-shaped nodes in purple colour. Edges represent high absolute correlations ($|r| > 0.5$). Edges between genes within the same pathway are in blue, edges connecting genes and FAs are in green, edges connecting FAs are in purple. Edges connecting genes of different pathways have been left out to improve the visibility of the network. The shapes of gene nodes indicate *cis*-eQTLs (V-shape), *trans*-eQTLs (square) and *cis/trans*-eQTLs (diamond). Node names are coded by concatenating gene name, *cis/trans*-regulation (separated by “_”) and linkage group (separated by “-”). For example, node “FAD2_C-5” indicates gene “FAD2”, “C” for *cis*-eQTL and “5” for linkage group A05, where the *cis*-eQTL was detected. In the case of a *cis/trans*-eQTL, “LACS2_CT-5_6” indicates gene “LACS2” and “CT-5_6” indicates a *cis*-eQTL on A05 and *trans*-eQTLs on A06. Solid lines indicate positive correlations and dotted lines indicate negative correlations. All the gene names are prefixed with “Br” because of *Brassica rapa* gene nomenclature. Multiple occurrence of the same gene names represents genes with multiple paralogues or probes.

Supplemental Table S2: List of genes with their primer sequences used in RT-qPCR.

Gene	Bra ID	Chr.	Forward primer sequence (5'-3')	Reverse primer sequence (5'-3')	<i>Arabidopsis</i> gene ID	Pathway
<i>BrCER8</i>	Bra004467	A05	AAACCCTGAAGTCGTGAACC	TTCTTCTCTGCCGTGGACTT	At2g47240	FA synthesis and elongation
<i>BrCAC3</i>	Bra005141	A03	GGCTCAAGAAAGGCAAGAAG	CTCGAAATCCAAAGTGACAGG	At2g38040	FA synthesis and elongation
<i>BrLPAT5</i>	Bra022365	A05	ACTGCGTAAAGGTCAGATTGG	CTGCCTCAAGTTTGCTTCATC	At3g18850	Phospholipid synthesis
<i>BrFAD5</i>	Bra027203	A05	CATGCTTTCGAGTTCTCTGCT	CACTGGTAAGTGCCATTCGTT	At3g15850	FA desaturation
Biotin carboxyl carrier protein	Bra027217	A05	GGCAACTCTTGGTTCTGTCC	TGGCTTCGTCGTCTCAGTATC	At3g15690	FA synthesis and elongation
<i>BrLACS2</i>	Bra032284	A05	GGTTTCGCTGGTATTCGTTC	CATTGGTTCTCAGTTCTTCCT	At1g49430	FA synthesis and elongation
<i>BrFAD2</i>	Bra034777	A05	TCGGAGAACTCAAGAAAGCAA	AGTGGTGGCGACGTAGTAGAA	At3g12120	FA desaturation
<i>BrGPAT5</i>	Bra034843	A05	TCTTATCTCCCATCCCAACC	GCGCTAAACCTCAGCAAGAA	At3g11430	Phospholipid synthesis
<i>BrFAD7</i>	Bra034863	A05	CAAGAAGTCCAGGGAAGAAGG	CCAACAAGCGGTAGAAGTGAG	At3g11170	FA desaturation
Acyl desaturase	Bra039178	A05	GGCTTCATTACACGTCTTTCC	ATCCGTGTGTAAGCTGTCTCGT	At3g02610	FA synthesis and elongation
<i>BrACP1</i>	Bra039471	A05	CCAGAGACGGTTGAGAAAGTG	ATCTGCTCCAAGATCAGCAA	At3g05020	FA synthesis and elongation
Seed storage protein	Bra026373	A01	GACAACAAGGAGGACAACAGG	TCTTCTGGAAAGGACAAATGCT	At4g27140	Storage proteins
Seed storage protein	Bra010409	A08	TCTTACTCACCAACGCCTCC	TCCATTGCTGACATGCTCTT	At4g27140	Storage proteins
Seed storage protein	Bra006444	A03	AACAGGTTGACAACAACAGG	TCTGGAAGGGACAAATGCTAA	At4g27170	Storage proteins
Seed storage protein	Bra019067	A03	GGACAACCCCAAGTAGTGAGA	TAGTAGTGGGGAGGCTGGAA	At4g27170	Storage proteins
Seed storage protein	Bra019064	A03	CACCAACGCTTCCATCTACC	AACTCCCTCTGGCATTCTGT	At4g27150	Storage proteins
Aspartic	Bra002214	A10	ATCTCTCTCCCTCGCAACTC	CGAAAACATCTTCAACCGAAC	At5g19120	Storage proteins
<i>BrCRU2</i>	Bra015275	A10	CCGACAGCAACAAAAACAAA	GTTGACTTGCTGGTTTTGGAG	At1g03880	Storage proteins
<i>BrCRU2</i>	Bra033721	A06	GACACCATCGCTACACATCC	TTCTGGTGGCTGGCTAAATC	At1g03880	Storage proteins
<i>BrCRU3</i>	Bra011036	A01	CAGCAGCAACAACAACAGAAC	ACGATGTTTCCTCTGCTGTCTT	At4g28520	Storage proteins
<i>BrCRU3</i>	Bra010363	A08	ACAACAACCGCAACAGAACAA	AGGTCCCCTAACACGAACAA	At4g28520	Storage proteins
<i>BrFLC2</i>	Bra028599	A02	AAGTATGGTTCACACCATGAG	GAGTCGACGCTTACATCAGA	At5g10140	MADS box transcription factor
<i>BrTT8</i>	Bra037887	A09	AGAATGTCAAAGAGCATCAGCA	TTTAGTGTTGTCGTGGAGGAA	At4g09820	bHLH transcription factor

Supplementary Table S3: Summary of QTLs detected for the relative abundance of fatty acids (FAs) from the 2009 seed lot.

Fatty acid	QTL no	LG	Peak marker	cM	LOD	Expl. var	Allelic effect	Flanking markers				Transformation
								UP	LP	Upper marker	Lower marker	
Myristic acid (C14:0)	Q1	2	BrVIN3P2b	53	7.3	14	PC	17	84	BrFLC2	BrCALP2a	None
	Q2	4	E32M19M439.2	25	12.9	22	YS	17	31	P21M47M78.5	P13M48M125.8	
	Q3	8	E34M15M312.5	31	5.5	9	YS	0	89	P23M50M86.4	Myb2RSAIM238.7	
	Q4	9	BrGAR3P1a	115	4.1	7	PC	61	137	Myb2MSE1M-596.7	BrPHYAP1g	
Palmitic acid (C16:0)	Q1	2	BrFYP1a	24	9.9	17	PC	20	27	Br323	ENA13l	None
	Q2	3	E32M47M382.4	127	3.1	6	YS	87	165	nced3	Vte1-3	
	Q3	4	BR325	39	3.7	7	PC	22	43	P14M51M88.7	BrPHYAP3b	
	Q4	8	P23M50M86.4	0	3.3	5	YS	0	27	P23M50M86.4	BRH80C09flc3	
	Q5	9	BrGAR3P1a	115	4.8	7	PC	69	147	BrRGAP2c	BrDDF1P2b	
	Q6	10	E32M19M310.4	53	4.1	9	YS	33	63	P13M48M342.9	BrFRIP2c	
Stearic acid (C18:0)	Q1	1	BrCOL2P3a	136	4.2	8	YS	115	136	P23M50M326.7	BrCOL2P3a	Log
	Q2	2	P14M51M219.4	67	4.7	10	PC	17	71	BrFLC2	P23M48M115.6	
	Q3	7	P23M50M106.0	25	3.6	8	PC	5	46	BrFKF1P2d	Myb2MSE1M-575.5	
Arachidic acid (C20:0)	Q1	1	BrCOL2P3a	136	3.5	8	YS	14	136	P23M48M93.2	BrCOL2P3a	Reciprocal
	Q2	2	E32M19M308.5	59	5.9	16	PC	17	67	BrFLC2	P14M51M219.4	
	Q3	5	MADsHaeIIIM392.7	156	3.4	7	PC	0	156	BrCCA1P3b	MADsHaeIIIM392.7	
	Q4	10	EJU6R10	17	3.4	7	YS	0	30	BrCPDP1a	BrCOL1P1c	
Behenic acid (C22:0)	Q1	4	E32M19M435.5	28	3.9	11	YS	17	50	P21M47M78.5	BrPHYAP3c	Reciprocal
Lignoceric acid (C24:0)	Q1	5	E32M19M221.8	28	4.8	13	PC	22	36	BRMS-034	P14M51M182.0	Reciprocal
	Q2	8	P23M48M276.5	18	5.9	14	YS	0	41	P23M50M86.4	BrPFT1P3b	
Palmitoleic acid (C16:1)	Q1	2	BrVIN3P2b	53	7.9	16	PC	20	57	Br323	BrFTP1c	Reciprocal
	Q2	5	E32M47M118.1	63	4.7	10	PC	56	74	BrVRN1P1b	E34M15M420.8	
	Q3	9	BrGAR3P1a	115	9.9	17	PC	106	121	Myb2AluIM454.0	BrSVPP1c	
Oleic acid (C18:1)	Q1	1	E32M47M212.2	118	6.1	12	YS	100	136	Myb2HaeIIIM475.9	BrCOL2P3a	None
	Q2	3	P14M51M135.1	106	9.5	18	YS	98	109	P23M48M439.2	ABI3	
	Q3	5	E32M19M258.3	133	4.7	9	PC	100	156	P23M50M132.9	MADsHaeIIIM392.7	
Eicosenoic acid (C20:1)	Q1	1	BrCOL2P3a	136	4	7	YS	34	136	Myb2RSAIM91.1	BrCOL2P3a	None
	Q2	2	P14M51M219.4	67	4.8	8	PC	57	71	BrFTP1c	P23M48M115.6	
	Q3	3	Myb2MSE1M67.4	104	10.6	17	YS	98	152	P23M48M439.2	Myb2MSE1M253.7	
	Q4	5	P23M50M241.1	61	3	6	YS	36	67	P14M51M182.0	BrSPL5P3a	
	Q5	7	P14M51M355.0	45	6.3	14	PC	43	46	BrAP3P1b	Myb2MSE1M-575.5	
Erucic acid (C22:1)	Q1	1	BrCOL2P3a	136	3.7	8	PC	86	136	P13M48M221.5	BrCOL2P3a	None
	Q2	2	P14M51M219.4	67	6.8	14	YS	57	71	BrFTP1c	P23M48M115.6	
	Q3	5	E32M47M113.6	61	4	9	PC	55	63	Myb2HaeIIIM263.7	E32M47M118.1	
	Q4	7	P14M51M355.0	45	3.5	11	YS	7	46	BrPVEP2a	Myb2MSE1M-575.5	
Nervonic acid (C24:1)	Q1	2	BrPIP2a	39	3.3	7	YS	32	84	BrCYP79A2	BrCALP2a	None
	Q2	3	E32M47M85.7	88	3.7	8	PC	57	90	BrAS1P2a	BrGA1P2c	
	Q3	6	BrFPF1P1d	75	3.1	5	PC	57	85	BrBCAT3-1MiAo7	BrRGAP1a	
	Q4	7	BrAP3P1b	43	6.4	14	YS	34	45	BrSEP3P1a	P14M51M355.0	
Linoleic acid (C18:2)	Q1	3	BRMS-043	59	5.6	9	PC	57	113	BrAS1P2a	P23M48M99.5	None
	Q2	4	P21M47M178.1	35	4	6	PC	25	76	E32M19M439.2	E32M47M136.0	
	Q3	5	Myb2HaeIIIM295.2	80	23	36	YS	74	89	E34M15M420.8	FAD2	
	Q4	9	P14M51M206.7	145	5.1	8	PC	106	147	Myb2AluIM454.0	BrDDF1P2b	
Linolenic acid (C18:3)	Q1	2	P14M51M219.4	67	5.2	12	YS	13	80	Myb2MSE1M480.8	Na12H09	None
	Q2	6	Na12H07	60	5.9	15	PC	57	75	BrBCAT3-1	BrFPF1P1d	
Eicosadienoic acid (C20:2)	Q2	5	Myb2HaeIIIM295.2	80	15.8	25	YS	67	89	BrSPL5P3a	FAD2	None
	Q3	9	BrVIM3P3b	115	4.8	8	PC	89	147	CAPS5	BrDDF1P2b	
	Q4	10	BrCPDP1a	0	5.3	8	PC	0	17	BrCPDP1a	EJU6R10	
Docosadienoic acid (C22:2)	Q1	1	E32M47M212.2	118	4.9	8	PC	71	136	E32M47M179.8	BrCOL2P3a	None
	Q2	3	P14M51M135.1	106	7.6	14	PC	98	113	P23M48M439.2	P23M48M99.5	
	Q3	5	Myb2HaeIIIM295.2	80	4.8	9	YS	65	140	P23M48M-36.6	BR378	
	Q4	9	P14M51M206.7	145	3.6	7	PC	82	147	BrGIP3b	BrDDF1P2b	
Docosatrienoic acid (C22:3)	Q1	2	Myb2MSE1M-614.5	74	4.1	9	YS	17	121	BrFLC2	Myb2MSE1M125.7	None
	Q2	6	Na12H07	60	4.7	11	PC	31	75	BrDRB1P1d	BrFPF1P1d	
	Q3	9	P14M51M206.7	145	3.1	9	PC	137	147	BrPHYAP1g	BrDDF1P2b	
	Q4	10	BrCPDP1a	0	3	7	PC	0	19	BrCPDP1a	P21M47M444.3	

Note: QTL no - number of QTLs; LG - linkage group; cM - Peak marker position, cM; Expl. var. - explained variance, %; allelic effect - orientation of parental allele for positive effect on relative FAs content; YS - yellow sarson allele; PC - pak choi allele; UP - Upper flanking marker position; LP - Lower flanking marker position.

Supplementary Table S4: Summary of QTLs detected for the relative abundance of fatty acids (FAs) from the 2011 seed lot.

Fatty acid	QTL no	LG	Peak marker	cM	LOD	Expl. Var.	Allelic effect	Flanking markers				Transformation
								UP	LP	Upper marker	Lower marker	
Myristic acid (C14:0)	Q1	2	P23M48M115.6	71	3.5	12	PC	13	121	Myb2MSE1M480.8	Myb2RSAIM146.8	None
	Q2	4	E32M19M138.2	26	4.5	16	YS	0	50	Myb2MSE1M142.2	BrPHYAP3c	
Palmitic acid (C16:0)	Q1	4	P13M48M125.8	31	4.4	13	PC	0	45	Myb2MSE1M142.2	BrSVPP2d	Reciprocal
	Q2	10	E32M19M310.4	53	3.7	11	YS	33	78	P13M48M342.9	BrCRY2P1a	
Stearic acid (C18:0)	Q1	7	P23M50M106.0	25	4.6	18	PC	1	46	BrCSlyase1	Myb2MSE1M-575.5	Reciprocal
Arachidic acid (C20:0)	Q1	7	P23M50M106.0	25	4.3	19	PC	1	46	BrCSlyase1	Myb2MSE1M-575.5	Reciprocal
Behenic acid (C22:0)	Q1	3	BrATBRMP1d	69	5.5	14	PC	66	84	BR356	BrFLKP2b	Reciprocal
	Q2	3	E32M19M312.7	135	4.4	8	PC	129	148	E32M47M384.5	Myb2HaeIIIM86.2	
	Q3	4	E32M19M435.5	28	8.9	27	YS	17	50	P21M47M78.5	BrPHYAP3c	
	Q4	7	E32M47M467.2	10	3.7	8	PC	0	34	BrFKF1P3c	BrSEP3P1a	
Lignoceric acid (C24:0)	Q1	5	Myb2HaeIIIM295.2	80	3.6	15	YS	14	108	BrSPA1P2a	E32M47M236.8	Log
Palmitoleic acid (C16:1)	Q1	2	BrCYP79A2	32	3.7	15	PC	13	117	Myb2MSE1M480.8	Myb2RSAIM496.0	None
Oleic acid (C18:1)	Q1	3	P23M50M284.3	139	3.6	8	YS	0	165	P23M48M293.8	Vte1-3	None
	Q2	5	P23M50M132.9	100	5.3	13	PC	11	156	BrSPA1P1a	MADsHaeIIIM392.7	
	Q3	10	BrCPDP1a	0	6.8	16	YS	0	23	BrCPDP1a	BRH80A08flc1	
Eicosenoic acid (C20:1)	Q1	3	E32M47M384.5	129	7.7	25	YS	112	148	MAM-4	Myb2HaeIIIM86.2	None
Erucic acid (C22:1)	Q1	3	E34M15M383.9	116	6.7	16	PC	71	152	Myb2RSAIM183.9	Myb2MSE1M253.7	None
	Q2	9	BrDDF1P2b	147	5.2	14	YS	115	147	BrVIM3P3b	BrDDF1P2b	
Nervonic acid (C24:1)	Q1	7	Myb2RSAIM230.6	41	9.8	26	YS	34	45	BrSEP3P1a	P14M51M355.0	None
	Q2	8	Myb2RSAIM238.7	89	3.7	10	YS	5	89	P23M48M-34.3	Myb2RSAIM238.7	
Linoleic acid (C18:2)	Q1	1	BrSPL5P1a	74	4.1	9	PC	69	125	Myb2HaeIIIM472.4	BrCOL2P4a	None
	Q2	3	BrGA1P2c	90	6.3	11	PC	87	109	nced3	ABI3	
	Q3	5	E32M47M460.0	83	22.3	45	YS	74	89	E34M15M420.8	FAD2	
Eicosadienoic acid (C20:2)	Q1	1	BrSPL5P1a	74	5.5	8	PC	69	86	Myb2HaeIIIM472.4	P13M48M221.5	None
	Q3	4	P14M51M88.7	22	4	5	PC	17	50	E32M19M204.4	BrPHYAP3c	
	Q4	5	E32M47M460.0	83	34	46	YS	78	89	Myb2HaeIIIM-605.3	FAD2	
	Q5	9	BrPVEP3a	125	4.1	6	PC	62	147	BrTOC1P2b	BrDDF1P2b	
Mead acid (C20:3)	Q1	4	P14M51M88.7	22	3.5	13	PC	0	63	Myb2MSE1M142.2	BrHOS1P2a	None
Docosadienoic acid (C22:2)	Q1	1	E32M47M179.8	71	3.7	11	PC	69	136	Myb2HaeIIIM472.4	BrCOL2P3a	None
	Q2	2	Myb2MSE1M480.8	13	3.4	8	PC	6	115	BrBFTP1a	BrFLMP1b	
	Q3	3	BrIQD1-1	94	4.8	12	PC	87	148	nced3	Myb2HaeIIIM86.2	
	Q4	5	Myb2HaeIIIM295.2	80	12.5	27	YS	74	100	E34M15M420.8	P23M50M132.9	

Note: QTL no - number of QTLs; LG - linkage group; cM - Peak marker position, cM; Expl. var. - explained variance, %; allelic effect - orientation of parental allele for positive effect on relative FAs content; YS - yellow sarson allele; PC- pak choi allele; UP - Upper flanking marker position; LP - Lower flanking marker position.

Chapter 6

General Discussion

The aim of this thesis is to unravel the genetics of seed quality and seedling vigour in *Brassica rapa* using a systems genetics approach. In order to achieve this, we first studied phenotypic and genetic variation in a *B. rapa* core collection to gain insights into population structure and in the genetic regulation of different morphological traits and a wide range of metabolites. Furthermore, QTL studies were done in a bi-parental doubled haploid (DH) population for seed germination and seedling vigour related traits and for oil content and fatty acid composition in ripe seed. QTL results were then combined with expression QTLs using systems genetics and gene co-expression network analysis (Steps 1-6 in Figure 2 – Chapter 1). In this general discussion, the relationships among phenotypic traits, metabolites and expression variation as well as their QTL map locations are discussed.

Hypothesis of the thesis

The hypothesis of this PhD study was that expression of genes during the seed-filling stages regulates seed metabolite content and composition in the ripe seed and that seed metabolites determine ultimate seed quality and seedling vigour. Therefore, in this study, both seed germination and root- (RL) and shoot- (SL) length and weight at the seedling stage were considered as parameters of seed quality to investigate the genetics of seed quality and seedling vigour.

Better understanding of the process of seed development in *B. rapa* is a prerequisite for genetical genomics studies

Good quality seed accompanied by high seedling vigour is imperative to improve crop establishment and increase agricultural production. Studies in many crops have indicated that seed quality and seedling vigour related traits are of complex genetic architecture and are influenced by interactions of multiple genetic and non-genetic factors (Chapter 4; Bettey et al., 2000; Koornneef et al., 2002; Finch-Savage et al., 2010; Joosen, 2013). Modern molecular technologies, ~omics data, in combination with classical genetics will allow unravelling the complexities of seed development, seed quality and seedling vigour.

In Chapter 3, we studied the temporal as well as genotypic variation of seed and embryo morphological characteristics and global transcriptional changes during seed development. For this purpose we used two black/brown- and two yellow-seeded genotypes (a vegetable type pak choi, an oil-type yellow sarson, and two of their progeny DH lines), which were early, mid- and late flowering. The changes in shape, size and colour of embryo and seed were compared with developing seeds of *B. napus* that have been studied intensively because of its importance as an oilseed crop. Our first hypothesis was that seed development processes in time would be comparable between these two *Brassica* species, for accessions with comparable flowering time. Also we assumed that the relationship between embryo developmental stages and metabolite profiles (oil, protein, sugar and starch) would be similar in *B. rapa* and the well-studied model plant *A. thaliana*. Because of these assumptions and the lack of data on the main metabolic pathways (oil, protein, sugar and starch profiles) in *B. rapa*, in the first year, we collected the

developing seeds for the genetical genomics study at 15-20 days after pollination (DAP), assuming that at this stage many fatty acid biosynthesis genes were differentially regulated. In *A. thaliana*, lipid biosynthesis is initiated in the torpedo stage (Baud et al., 2002), and this corresponded to seed development stages between 15 and 20 DAP in *B. napus* (Fernandez et al., 1991; Ilić-Grubor et al., 1998). Niu et al. (2009) reported that major changes in expression profiles of genes involved in protein translation, starch metabolism and hormonal regulation were between 17 and 21 DAP in *B. napus*, whereas fatty acid biosynthesis-related genes were highly expressed at 21 DAP as compared to earlier and later time points. However, data from Chapter 3 revealed that the timing of several seed developmental processes was later in *B. rapa* than in *B. napus*. The period from 25 to 35 DAP (between the bent-cotyledon stage and the stage when the embryo fully fills the seed) was the key period for major changes in transcript abundance in *B. rapa* developing seeds. This detailed insight into the timing of the most active stages of transcriptional regulation was useful to choose the optimum time point (28 DAP) for collecting RNA samples for the genetical genomics experiment to unravel regulation of seed fatty acid profiles, described in Chapter 5. Previous studies indicated that differential expression of genes involved in the regulation of phenotypic traits or metabolic pathways vary largely depending on the developmental stages or tissue samples chosen (Fu et al., 2012). Therefore, in genetical genomics, the choices of the optimum stage and the right tissue to isolate RNA are crucial if the aim is to use expression profiles to identify genes underlying traits of interest. Metabolite profiling during seed development could have complemented our findings and would have been helpful to understand the seed development process at the metabolite level too. However, within the period of this PhD study we were not able to obtain funds for such a study.

In Chapter 3, we identified groups of genes (called gene modules) co-expressed during seed development and those co-expressed genes might be involved in similar biological processes, or play roles in the same cellular processes. One reason why such a group of genes might be co-regulated is that the genes in the module share *cis*-regulatory elements (motifs) in their promoter regions. In Chapter 3, we identified motifs for groups of co-expressed genes related to lipid metabolism. A subset of those motifs was also reported in other plant species for their roles in seed development (Chapter 3).

Systems genetics and co-expression network analysis assist in the identification of candidate genes for lipid metabolism

Systems genetics is widely used to understand the underlying biological processes and molecular mechanisms of complex traits (Civelek and Lusis, 2014; Van der Sijde et al., 2014). In this approach, intermediate phenotypes, such as transcript and metabolite levels, as well as physiological traits, obtained from a range of experiments are integrated using statistical genetics and network analyses. Initially, expectations of systems genetics were very high in that one could identify candidate genes, pathways and regulatory networks underlying the traits of interest. Genetical genomics has become an important tool in systems genetics studies to map eQTLs in genetically perturbed natural populations (Civelek and Lusis, 2014; Feltus, 2014). One of the important

advantages is the ability to detect *cis*- and or *trans*-acting eQTLs, which provide information about the gene regulation networks. The overlap of QTL regions for phenotypic traits (phQTLs), metabolite levels (mQTLs) and transcript level (eQTLs) highlights the genomic regions that contain possible candidate genes involved in the regulation of these biological processes (Civelek and Lusis, 2014; Feltus, 2014). In Figure 1, we observed the co-localization of QTLs detected for different levels of *omics* traits: phenomics (seed germination, seedling vigour, flowering time, seed weight, seed colour), metabolites (fatty acids, fibre content, total protein content, total glucosinolate, total oil) and eQTL profiles (genes related to lipid metabolism, and *BrFLC2* and *BrTT8* genes), mainly on linkage groups A02, A03, A05, A08 and A09. All co-localization in these regions can be due to causal relationships among different levels of *omics* traits, but can also be coincidental. Additional analyses are necessary to validate hypothesized relationships among seed germination and seedling vigour parameters, seed metabolites and gene expression, which is discussed below under “future perspectives”.

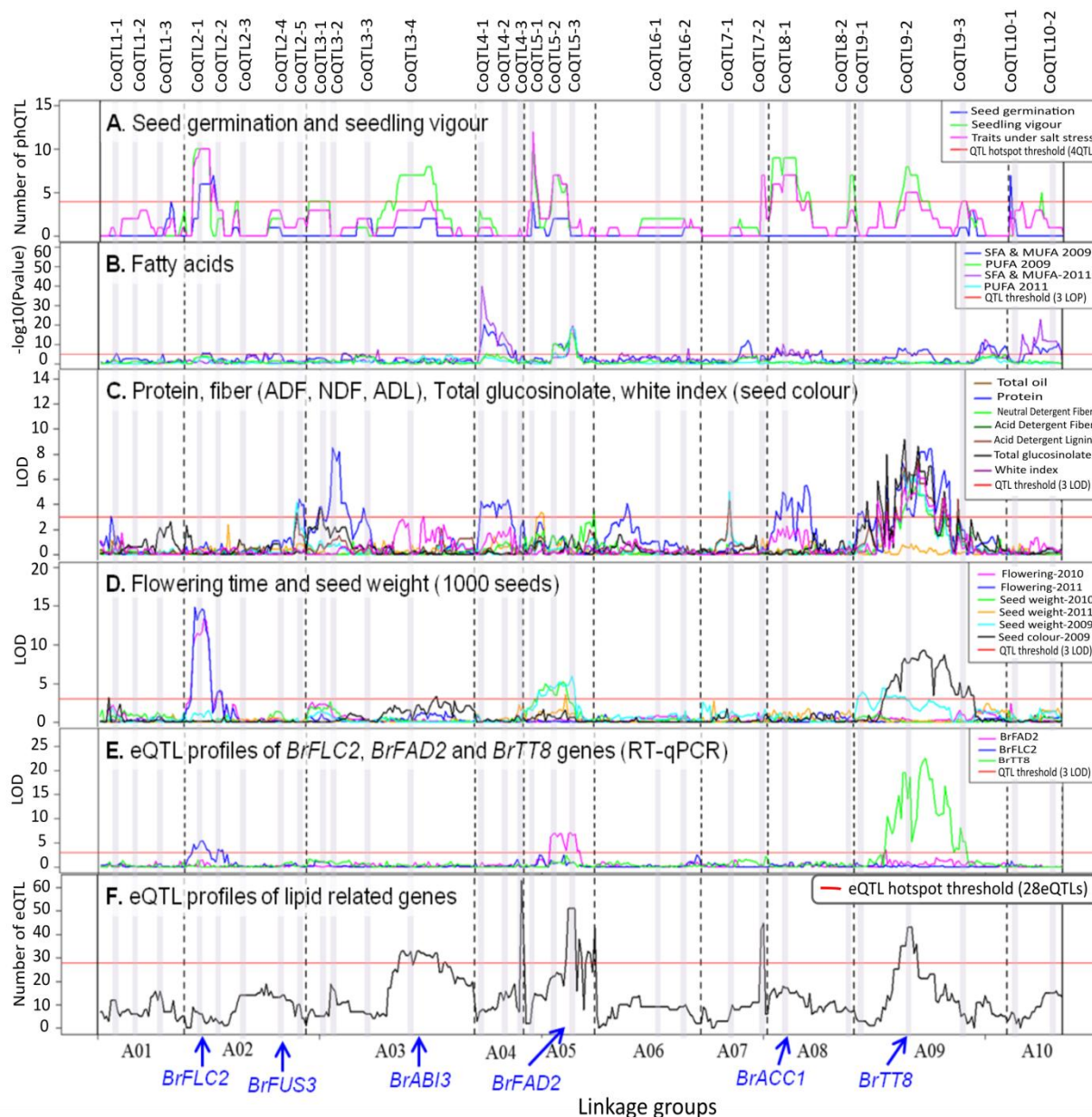


Figure 1 (above): Co-localization of QTLs detected for different data sets: **A.** Seed germination and seedling vigour. **B.** Fatty acids. **C.** Total protein content, fibre (ADF, NDF, ADL), total glucosinolates, white index (~ seed colour). **D.** Flowering time and thousand-seed weight. **E.** eQTL profiles of *BrFLC2*, *BrFAD2* and *BrTT8* genes (real time-qPCR expression analysis) and **F.** eQTL profiles of genes related to lipid metabolism (after correction for seed colour). QTL profiles of fatty acids are from multi-trait QTL analysis in Genstat. Gene-targeted markers in the QTL co-localized regions are highlighted in blue on the x-axis.

In Chapter 5, eQTL hotspots for genes related to lipid metabolism were identified on A03, A04, A05 and A09. Major fatty acid QTLs (faQTLs) for monounsaturated fatty acids (MUFAs) were also mapped on A03, for saturated fatty acids (SFAs) on A04 and for poly-unsaturated fatty acids (PUFAs) on A05, but not on A09. These results could guide us to find the master regulatory genes for these hotspots, or important *cis* regulated genes as candidate genes for specific fatty acids. Like in Chapter 3, a search for *cis*-regulatory elements in the upstream regions of genes from those eQTL hotspots might help in the future to find the underlying master regulator genes.

If one looks at co-localization of eQTLs for individual transcripts with a phenotypic trait, the information from gene-gene interrelations is not taken into account, which are likely important for complex traits (Civelek and Lusi, 2014; Feltus, 2014). To explore these gene-gene relationships, gene regulatory networks need to be constructed using network analysis. We used eQTL-guided correlation network and weighted gene co-expression network analysis (WGCNA) to consider gene expression variation as well as gene-gene relationships; both approaches are discussed in Chapter 5. The parameter 'degree of connection' was used to identify genes with a high number of connections in these networks. In general, genes with a high degree of connection are most likely essential genes to knock out a pathway, whereas genes with lower degree of connection could be interesting to modify metabolite content (Khurana et al., 2013; Mäkinen et al., 2014). For example, a strong QTL with large explained variance (27-46.3%) was detected for linoleic acid on A05, where a *cis*-eQTL of the gene *BrPLA2-ALPHA* was co-mapped. The gene *BrPLA2-ALPHA* had the highest degree of connection and, interestingly, was reported as a key regulator of linoleic acid metabolism (Ryu et al., 2005). This gene has an acyl preference for linoleoyl over palmitoyl. Even though the genetic regulation of lipid metabolism has been extensively studied in *Arabidopsis*, Brassicas and other oil crops, this gene has hardly been reported for its molecular function. Instead, genes such as *fatty acid desaturase 2* (*BnFAD2*), *fatty acid desaturase 3* (*BnFAD3*), *diacylglycerol acyltransferase* (*BnDGAT1*) and *fatty acid elongase 1* (*BnFAE1*) were reported in *B. napus* (Peng et al., 2010; Yang et al., 2012; Lee et al., 2013, Tanhuanpää and Schulman, 2002). One of the possible reasons could be that such key regulator genes are often highly conserved in regulatory networks during evolution (Khurana et al., 2013), and are less likely to be genetically perturbed in mapping populations (Mäkinen et al., 2014). The fact that we still found a *cis*-eQTL for the highly connected gene *BrPLA2-ALPHA*, and consider it a candidate gene for linoleic acid variation, might be due to the different selection history of oil and vegetable types. In our experience, systems genetics should be considered as a hypothesis generating approach for identifying candidate genes for complex traits. Using the combined approach of eQTL mapping and

network analysis, we were able to subset the list of important genes that are likely involved in the genetic regulation of lipid metabolism and fatty acids biosynthesis in *B. rapa* developing seeds (Chapter 5). Thus, systems genetics seems to be a promising approach as an alternative to gene identification by fine mapping, especially in *B. rapa* because of hurdles in creating large mapping populations due to self-incompatibility. For functional validation, gene transformation is commonly used; however, also transformation in *B. rapa* is very challenging and would require further optimization. Instead, genetic transformation in *B. napus*, close relative of *B. rapa*, would be an alternative to perform a functional validation.

Another issue while dealing with high-dimensional data sets (e.g. *omics* data) is the computational limitation for data analysis, especially for network construction. In this study, microarray experiments produced transcriptomics data for 61,557 probes in 120 DH lines. To analyse all the genes in order to explore the transcriptional cross-talk among pathways is computationally challenging within the capacity of a PhD project. Therefore, we focused on the probes related to lipid metabolism, since the main focus of this PhD study was on understanding the genetic regulation of lipid metabolism and fatty acid biosynthesis. Furthermore, we constructed eQTL-guided gene correlation networks of only those probes that had at least one eQTL. These criteria reduced the computational load for network analysis considerably. This type of reduction is not possible, however, for less well studied traits, such as turnip formation, where pre-selection of a subset of genes is less obvious.

Genetics of seed quality and seedling vigour related traits

In addition to significant developments in seed technology to improve seed quality in the last decades, many public and private breeding institutions have been focusing on the genetic improvement of seed quality and seedling vigour. As abiotic factors are the major challenge for crop establishment and salinity is one of the major abiotic stresses limiting high crop production (Mittler, 2006), these stresses need to be taken into account when studying seed quality traits.

Chapter 4 focused on genetic analysis of seed germination dynamics and seedling vigour over time (during the first 10 days after germination, DAG) under non-stress and salt stress conditions in the *B. rapa* DH mapping population. We identified QTLs for seed germination parameters, root- and shoot- length, seed weight and flowering time, which were mainly co-localized at eight hotspots. Interestingly, seed germination QTLs on A02 co-localized with a flowering QTL in the region of *BrFLC2* and a QTL for seedling vigour under salt stress on A05 co-localized with major QTLs of seed weight and PUFAs (linoleic, eicosadienoic, and docosadienoic acid) in the region of *BrFAD2*. Many QTLs were confirmed across two years' trials with a different experimental set up: with and without synchronization of flowering time. The putative roles of *FLC2* and *FAD2* in seed quality traits as hypothesized in this study correspond with studies in *A. thaliana* showing that *FLC2* has a pleiotropic effect on seed germination (Chiang et al., 2009) and in *B. napus* pointing to a role of *FAD2* gene on seedling growth under saline conditions (Wang et al., 2010b; Zhang et al., 2012).

We carried out a follow-up experiment to identify genomic regions associated with expression variation of these two genes. This showed that *cis*-eQTLs for both *BrFLC2* and *BrFAD2* genes were

co-localized with the QTL hotspots on A02 and A05, respectively. However, we are aware of the fact that those QTL intervals contain many genes and, based on the literature, we had selected only two genes *BrFLC2* and *BrFAD2* for mapping expression variation (Chiang et al., 2009; Wang et al., 2010b; Zhang et al., 2012). As both fatty acid measurements and gene expression (genetical genomics) studies were carried out under non-stress, we suggest carrying out metabolomics and genetical genomics experiments also under salinity stress to study the role of *BrFAD2* in seed quality traits under salinity at these different *omics* levels.

Confounding effects in genome-wide genetic studies

The population used in this study is segregating for many traits (Chapter 2, Chapter 3, Chapter 5) (Zhao, 2007; Pino Del Carpio, 2010; Basnet et al., 2013; Xiao et al., 2013; Xiao et al., 2014), and as such is appropriate for genetic studies of these traits. However, this variation possibly also introduces confounding effects on traits: for example, the effects from flowering time and seed colour variation in this study. A major challenge in genome-wide analysis is the presence of hidden confounding factors, such as unobserved covariates or unknown subtle environmental perturbations (Kang et al., 2008a; Fusi et al., 2012). These factors can induce a pronounced artefactual correlation structure in the variation of phenotypic traits, which may create spurious false marker-trait associations or mask real genetic association signals (Kang et al., 2008b; Fusi et al., 2012). The use of proper experimental design and statistical models might be able to exclude or correct for such confounding effects in the data.

As an example of this, we think that the variation in flowering time has led to such artefacts in the sense that spurious fatty acid QTLs were detected on A02, where a major flowering time QTL and the *BrFLC2* gene are located (Figure 1). In our study, the synchronization of flowering of the DH lines in the year 2011 seemed an effective approach to unmask the confounding effect of the flowering locus (*BrFLC2*) on detecting those possibly spurious fatty acid QTLs on A02 (Figure 3 in Chapter 5). When flowering time is synchronized, environmental conditions during seed ripening are more uniform. However, a drawback is that the environmental conditions during early plant growth differed very much, and that late flowering plants had to be sown very early, in the winter, when growing conditions are suboptimal due to insufficient light. In the 2009 experiment, flowering was asynchronous, resulting in considerable variation in environmental conditions during seed ripening; however conditions during early plant growth were very similar and in the optimal spring season.

Similarly, variation in seed colour has probably led to a large number of false *trans*-eQTLs on A09, where a major seed colour QTL (explained variance 32%) was detected. The gene responsible for seed colour has been cloned in yellow sarson, and is a *bHLH* transcription factor, *BrTT8* (Li et al., 2012b). In Chapter 3, we also observed differences in transcriptional signature in the yellow-seeded YS143 as compared to the brown/black-seeded PC175. In eQTL analysis, seed colour therefore was used as a covariate in our statistical model to avoid spurious eQTLs on A09; however, this could also lead to false negatives. Therefore, it would be more optimal to have a population that is segregating only for fatty acids, seed germination or seedling growth related

traits with uniform flowering time and uniform seed colour. However, as there is a strong association between seed colour and oil content, with yellow seed colour as important preferred trait in oilseed breeding (Chen and Heneen, 1992; Rahman et al., 2001), a pure yellow seeded population would probably have much more limited variation.

Future perspectives

Data integration including complete transcriptome, seed metabolite and phenomics data sets

In this PhD study, we focused on transcript profiles of genes related to lipid metabolism (Chapter 3 and 5) and fatty acid metabolites (Chapter 5). We developed methodological tools to integrate the transcriptomics and metabolomics data sets and to construct regulatory networks related to major fatty acids. Finally, this resulted in a set of (possible) candidate genes involved in lipid metabolism (Chapter 5). Besides, we also identified major QTLs for other important seed constituents (*e. g.* Total protein content, fibres, glucosinolates; Figure 1), and those constituents also play roles in determining seed quality and seedling vigour. In the future, it would be highly interesting to integrate the genome-wide transcriptome data set with all major seed metabolites and phenotypic data. This integrated analysis could directly link all three components: transcriptome, metabolome and phenotypic traits, and ultimately could expand the knowledge on the genetic regulation of seed metabolites, seed quality and seedling vigour in *B. rapa* as well as other *Brassica* species. Data-driven integrated analyses could also be helpful for functional annotation of genes with unknown function and opens the opportunity to discover novel gene functions.

Use of automated high-throughput phenotyping platforms

Plant survival and fitness related traits, such as seed quality and seedling vigour traits are under historical and natural selection. Complex networks of many genes usually regulate those life history traits. Phenotyping of seed germination and seedling growth related traits is labour-intensive, time-consuming, tedious and also frequently destructive, and might lead to high experimental as well as measurement errors. Those errors will reduce the heritability of the traits and also negatively influence QTL detection and any further results. Automated high-throughput phenotyping platforms are efficient, fast, precise and able to measure many traits in large numbers of plants at the same time, reducing the experimental errors (Furbank and Tester; Chen et al., 2014). In addition, the availability of such phenotyping platforms enables the quantification of traits that are difficult to measure, such as number of lateral roots, total root length, root hairs and/or non-visible traits, such as carbon flow, photosynthesis efficiency, imaging plant canopy and chlorophyll fluorescence. Measurement of a wide range of traits, including such difficult-to-measure and non-visible traits can be useful to better depict the dynamic range of biological processes (also called holophenotype – the ultimate phenotypic reality we attempt to measure; Chitwood and Topp, 2015), such as seed quality. For example, germination parameters like earliness, speed, uniformity and maximum germination derived from a seed germination curve over time help to understand the holophenotype of the seed germination process (Chapter 3). In this study, manual counting of germinated seeds of the entire DH population was limited to day-

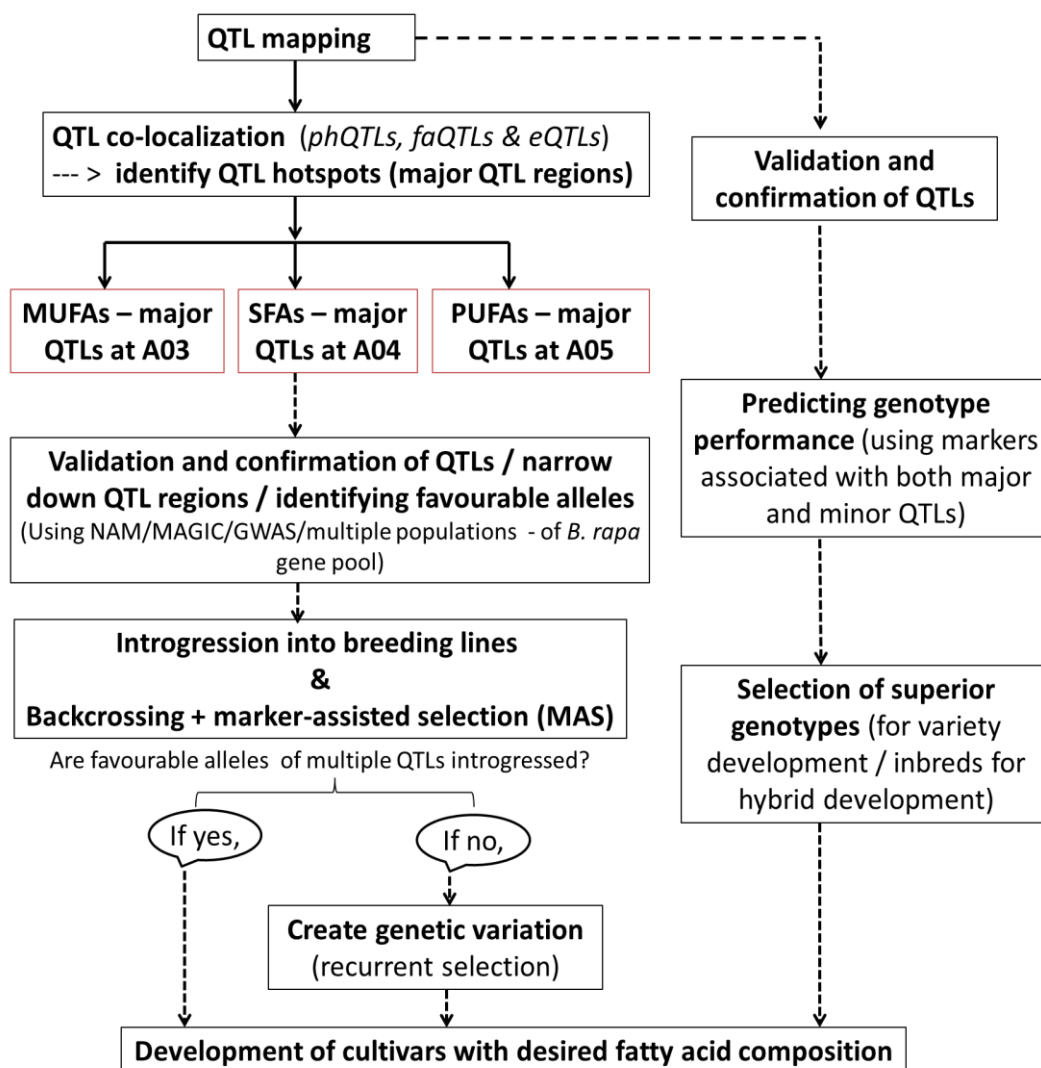
time (Chapter 3). Similarly, root and shoot lengths were measured at only four time points: 3, 5, 7, 9 DAGs (Chapter 3). With automated high-throughput phenotyping platforms it would be possible to comprehensively collect digital and time-resolved data that describes seed germination and seedling growth in greater detail. Such dynamic phenotypic observations also allow fitting growth models and to predict growth parameters, and unravel the so-called cryptotype (a hidden or latent combination of traits that maximize the separation of a *priori* known or defined groups; Chitwood and Topp, 2015).

Beyond QTL mapping: QTL validation, fine mapping and allelic variation

In this study, we identified many QTLs for fatty acids, seed germination and seedling vigour traits. These QTLs need to be confirmed in association panels or mapping populations with different genetic backgrounds (Step 7 in Figure 2 – Chapter 1). Genome-wide association studies (GWAS), taking advantage of historic recombination, can be used for high resolution QTL mapping (narrowing QTL intervals), discriminating pleiotropy or linkage of genes and allelic variation (Varshney et al., 2014; Zhao et al., 2007). This approach has been successfully applied in several crop plants (Varshney et al., 2014; Zhao et al., 2007). Lately, the nested association mapping (NAM), the *Arabidopsis* multi-parent RIL (AMPRIL) and the multi-parent advanced generation inter-cross (MAGIC) populations, which take advantage of both historic and recent recombination events, have been used for studying trait genetics in maize, *A. thaliana* and wheat, respectively (Cavanagh et al., 2008; Yu et al., 2008; McMullen et al., 2009; Huang et al., 2011; Huang et al., 2012). These types of populations could be used for validating, fine mapping and studying allelic variation of QTLs that were identified in our study (Step 7 in Figure 2 – Chapter 1). These populations aim to combine the advantages of association mapping populations with those of bi-parental mapping populations: a high recombination rate, presence of more than two alleles, equal kinship relatedness and lack of population structure. Professor R. Amasino and co-workers (University of Wisconsin-Madison, USA) are constructing a NAM/AIRIL (nested association mapping/advanced inter-cross RIL) population from crosses of R500 (YS143) and DH lines of many morphotypes supplied by our group (Wageningen UR plant breeding), after inter-crossing F2s for two generations followed by several generations of inbreeding to make RILs (Dr. A. B. Bonnema, Wageningen UR, personal communication). This population offers excellent possibilities to validate our QTLs and candidate genes, as it has the same YS143 genotype as central parent, crossed to seven diverse genotypes, with increased recombination frequency, and increased allelic variation.

Joint efforts in the labs of Wageningen UR Plant Breeding and IVF CAAS resulted in the resequencing of 125 *B. rapa* genotypes. These data could be mined to reveal haplotype variation or structural and/or allelic variation within the candidate regions identified in this study. After phenotyping of the studied traits, the re-sequencing data could be used for association mapping. Alternatively, reverse genetics approaches, such as in targeting-induced local lesions in genomes (TILLING) populations could be used for screening genes of interest to identify useful alleles, novel rare mutants, causal mutations and gene functions for phenotypic traits (Stephenson et al., 2010;

Gilchrist et al., 2013; Graham et al., 2014; Varshney et al., 2014). Graham et al., (2014) showed a link between eQTL and functional validation of the Bra017134 gene for an altered Calcium (Ca) concentration in shoots using a TILLING population in *Arabidopsis* and a *B. rapa* yellow sarson (RO18) (*BraA.CAX1a* mutant). As RO18 is a yellow sarson background, is self-compatible and has a good seed yield for phenotyping, this TILLING population could be very relevant to us as well to



further validate candidate genes and to identify allelic variation and even rare variants.

Figure 2: Proposed possible breeding strategy to develop varieties with desired fatty acid composition, using fatty acid QTLs identified in this study. Solid lines indicate the activities carried out in this study and dotted lines for the future activities that can be used in practical breeding to obtain desired fatty acid composition.

Breeding perspectives

Vegetable type *B. rapa*'s are mainly field crops, but seedlings are raised inside greenhouses by plant raisers, whereas oil-types and turnips are directly sown field crops. Seed quality and seedling establishment, especially under sub-optimal conditions, are more crucial for the latter ones. For the quality of seed oil in oilseed breeding, the optimum fatty acid composition is an important parameter. In this study, QTL hotspots for seed quality, seedling vigour and fatty acid composition

were detected across years and growing conditions, and those QTLs could be selected for a breeding program. We have provided a general scheme that can be applicable for implementing our QTL results in, for instance, marker-assisted selection (MAS). For example developing cultivars with different fatty acid composition would mean that in a practical breeding program one has to concentrate on linkage group A03 if MUFAs are to be changed while changing SFAs would mean to focus on linkage group A04 (Figure 2). In this schematic flow chart, we emphasize that the QTLs ideally first would need to be validated using several populations with different genetic backgrounds. In this PhD study, a doubled haploid population was used, which generally results in wide QTL regions. Therefore, it is important to narrow down the QTL regions before attempting to introgress genes (regions) of interest into breeding germplasm. In practical breeding, the identification of the actual responsible genes for phenotypic traits does not matter so much, as long as the favourable allele (or, haplotype alleles) can be selected for in marker-assisted selection (MAS) using different populations or breeding germplasm. After introgression of favourable alleles (*e. g.* through backcrossing), followed by MAS, the suitable breeding lines with desired fatty acid composition can be identified for variety development. In case of multiple QTLs with both major and minor effects, we suggest another route for breeding superior varieties, in which markers associated with major as well as minor QTLs (after QTL validation) can be used to predict genotype performances to select superior genotypes for variety development or inbreds for hybrid development (Figure 2). The advantage of MAS is that it allows the elimination of undesirable genotypes at the seedling stage, even for traits that are expressed at the later developmental stages. For very complex traits, genomic selection (GS) has nowadays received high importance to increase genetic gain in breeding per unit time and cost. In GS, a prediction model of breeding values is estimated using a training data set, which consists of both genotype (marker) and phenotype information of a large number of individuals. Using this model, genomic breeding values of novel breeding lines are estimated based on genotype data only. GS is based on simultaneous estimation of effects on phenotypes of all loci or markers available across the genome. Metabolites are more closely linked to the phenotype than genes, thus, may be used as predictive biomarkers for phenotypic traits as well. For example, the nutritional or industrial value of crop plants is ultimately dependent on their metabolic composition and some of these metabolites have been successfully employed for improving quality traits like fatty acid composition, oil content and seed meal in oilseed crops, protein, oil and provitamin A content in maize, starch content in potato and rice, carotenoid content in tomato, cold-sweetening in potato (reviewed by Fernie and Schauer, 2008). The measurement of hundreds of metabolites could lead to a better understanding of metabolism itself and, when used as biomarkers, also to predict agronomic traits or resistance to stresses. As systems biology approaches reveal genes that underlay phenotypic variation, this will result in tools for the breeders.

Spatial and temporal mapping of metabolites or gene expression

In this PhD project, we primarily studied global transcriptional variation during seed development (Chapter 3) as well as eQTL mapping in genetical genomics (Chapter 5). As mentioned earlier, we

aimed to measure, additionally, metabolic profiles of developing seeds to monitor metabolic switches during seed development and to associate these with transcript profiles. Unfortunately, this metabolomics project was not funded. We believe that understanding the seed development process at morphological, transcriptional and metabolomic levels would give detailed understanding of the metabolic processes during seed development. Furthermore, the *in vivo* spatial and temporal mapping (distributions and quantifications) of metabolite deposition at cellular levels in developing seeds until germination would add an additional level of information to this study. This could assist in relating the transcriptome variation, metabolite patterns and phenotypic observations with functional information and transport processes within the plant at anatomical level. Imaging technologies, such as matrix-assisted laser desorption/ionization (MALDI), electrospray ionization (ESI), nuclear magnetic resonance (NMR) and magnetic resonance imaging (MRI) have been successfully used in many plant species for accurate *in vivo* tissue-specific localization and quantification of metabolites (*e. g.* lipid deposition) and expression of related genes in intact seeds (Borisjuk et al., 2013b). Therefore, such spatial and temporal mapping technologies would be helpful to integrate the current data sets as well as to validate the function of our candidate genes at the tissue-specific level. All that information could be an important contribution for subsequent breeding or genetic research purposes.

Conclusions and final remarks

Analyses of fatty acids, transcripts and phenomics of *B. rapa* seeds using systems genetics has increased our understanding of the genetics of seed quality and seedling vigour in *B. rapa*. Based on the discussion of co-localization of QTLs in Figure 1 and network analysis in Chapter 5, we were able to support our hypothesis that expression of lipid metabolism related genes is involved in the regulation of fatty acid biosynthesis, and that seed metabolites most likely are involved in the regulation of seed quality and seedling vigour related traits. The knowledge on the genetic regulation of fatty acids at the individual gene level may provide new opportunities for optimizing oil quality for different purposes. We expect that the findings of this thesis will have important contributions in seed quality research in *B. rapa* and ultimately in *Brassica* oilseed breeding.

References

- Agrama H, Eizenga G, Yan W** (2007) Association mapping of yield and its components in rice cultivars. *Molecular Breeding* 19(4): 341-356.
- Al-Shehbaz IA, Beilstein MA, Kellogg EA** (2006) Systematics and phylogeny of the *Brassicaceae* (*Cruciferae*): an overview. *Plant Systematics and Evolution* 259(2-4): 89-120.
- Ambika S, Manonmani V, Somasundaram G** (2014) Review on effect of seed size on seedling vigour and seed yield. *Research Journal of Seed Science* 7(2): 31-38.
- Anderson JT, Willis JH, Mitchell-Olds T** (2011) Evolutionary genetics of plant adaptation. *Trends in genetics* 27(7): 258-266.
- Andriotis VME, Pike MJ, Schwarz SL, Rawsthorne S, Wang TL, Smith AM** (2012) Altered starch turnover in the maternal plant has major effects on *Arabidopsis* fruit growth and seed composition. *Plant Physiology* 160(3): 1175-1186.
- Angelovici R, Fait A, Zhu X, Szymanski J, Feldmesser E, Fernie AR, Galili G** (2009) Deciphering transcriptional and metabolic networks associated with lysine metabolism during *Arabidopsis* seed development. *Plant Physiology* 151(4): 2058-2072.
- Aranzana MaJ, Kim S, Zhao K, Bakker E, Horton M, Jakob K, Lister C, Molitor J, Shindo C, Tang C, Toomajian C, Traw B, Zheng H, Bergelson J, Dean C, Marjoram P, Nordborg M** (2005) Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. *PLoS Genetics* 1(5): e60.
- Armengaud P, Zambaux K, Hills A, Sulpice R, Pattison RJ, Blatt MR, Amtmann A** (2009) EZ-Rhizo: integrated software for the fast and accurate measurement of root system architecture. *The Plant Journal* 57 (5): 945-956.
- Asakura T, Tamura T, Terauchi K, Narikawa T, Yagasaki K, Ishimaru Y, Abe K** (2012) Global gene expression profiles in developing soybean seeds. *Plant Physiology and Biochemistry* 52: 147-153.
- Ashraf M, McNeilly T** (2004) Salinity tolerance in *Brassica* oilseeds. *Critical Reviews in Plant Sciences* 23(2): 157-174.
- Assenov Y, Ramírez F, Schelhorn SE, Lengauer T, Albrecht M** (2008) Computing topological parameters of biological networks. *Bioinformatics* 24(2): 282-284.
- Atwell S, Huang YS, Vilhjalmsón BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, Jiang R, Muliyati NW, Zhang X, Amer MA, Baxter I, Brachi B, Chory J, Dean C, Debieu M, de Meaux J, Ecker JR, Faure N, Kniskern JM, Jones JDG, Michael T, Nemri A, Roux F, Salt DE, Tang C, Todesco M, Traw MB, Weigel D, Marjoram P, Borevitz JO, Bergelson J, Nordborg M** (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* 465(7298): 627-631.
- Bagheri H, Pino Del Carpio D, Hanhart C, Bonnema G, Keurentjes J, Aarts MGM** (2013) Identification of seed-related QTL in *Brassica rapa*. *Spanish Journal of Agricultural Research* 11:1085-1093.
- Bai C, Liang Y, Hawkesford MJ** (2013) Identification of QTLs associated with seedling root traits and their correlation with plant height in wheat. *Journal of Experimental Botany* 64(6): 1745-1753.
- Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS** (2009) MEME Suite: tools for motif discovery and searching. *Nucleic Acids Research* 37: W202-W208.

- Baker A, Graham IA, Holdsworth M, Smith SM, Theodoulou FL** (2006) Chewing the fat: β -oxidation in signalling and development. *Trends in Plant Science* 11(3): 124-132.
- Barker GC, Larson TR, Graham IA, Lynn JR, King GJ** (2007) Novel insights into seed fatty acid synthesis and modification pathways from genetic diversity and quantitative trait loci analysis of the *Brassica* C genome. *Plant Physiology* 144(4): 1827-1842.
- Basnet RK, Moreno-Pachon N, Lin K, Bucher J, Visser RGF, Maliepaard C, Bonnema G** (2013) Genome-wide analysis of coordinated transcript abundance during seed development in different *Brassica rapa* morphotypes. *BMC Genomics* 14: 840.
- Basunanda P, Radoev M, Ecke W, Friedt W, Becker HC, Snowdon RJ** (2010) Comparative mapping of quantitative trait loci involved in heterosis for seedling and yield traits in oilseed rape (*Brassica napus* L.). *Theoretical and Applied Genetics* 120(2): 271-281.
- Batagelj V, Mrvar A** (2003) Pajek - analysis and visualization of large networks. In: Juenger M, Mutzel P (eds.). *Graph Drawing Software*, Springer, Berlin, pp: 77-103.
- Baud S, Boutin J-P, Miquel M, Lepiniec L, Rochat C** (2002) An integrated overview of seed development in *Arabidopsis thaliana* ecotype WS. *Plant Physiology and Biochemistry* 40(2): 151-160.
- Baud S, Dubreucq B, Miquel M, Rochat C, Lepiniec L** (2008) Storage reserve accumulation in *Arabidopsis*: metabolic and developmental control of seed filling. *The Arabidopsis Book* / American Society of Plant Biologists 6: e0113.
- Baud S, Lepiniec L** (2010) Physiological and developmental regulation of seed oil production. *Progress in Lipid Research* 49(3): 235-249.
- Beisson F, Koo AJK, Ruuska S, Schwender J, Pollard M, Thelen JJ, Paddock T, Salas JJ, Savage L, Milcamps A, Mhaske VB, Cho Y, Ohlrogge JB** (2003) *Arabidopsis* genes involved in acyl lipid metabolism. A 2003 census of the candidates, a study of the distribution of expressed sequence tags in organs, and a web-based database. *Plant Physiology* 132(2): 681-697.
- Benjamini Y, Hochberg Y** (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57(1): 289-300.
- Bentsink L, Koornneef M** (2008) Seed dormancy and germination. *The Arabidopsis Book* / American Society of Plant Biologists 6: e0119.
- Bernardo R** (2002) *Breeding for Quantitative traits in plants*. Woodbury, Minnesota, the United States of America: Stemma Press.
- Betty M, Finch-Savage WE, King GJ, Lynn JR** (2000) Quantitative genetic analysis of seed vigour and pre-emergence seedling growth traits in *Brassica oleracea*. *New Phytologist* 148(2): 277-286.
- Bino RJ, De Vos CHR, Lieberman M, Hall RD, Bovy A, Jonker HH, Tikunov Y, Lommen A, Moco S, Levin I** (2005) The light-hyperresponsive high pigment-2dg mutation of tomato: alterations in the fruit metabolome. *New Phytologist* 166(2): 427-438.
- Bogatek R, Gniazdowska A** (2012) Ethylene in seed development, dormancy and germination. In: McManus MT (ed.) *The plant hormone ethylene*. *Annual Plant Reviews* 44: 189-218.
- Bonnema G, Carpio DD, Zhao J** (2011) Diversity analysis and molecular taxonomy of *Brassica* vegetable crops. In: Sadowski J, Kole C (eds.) *Genetics, genomics and breeding of vegetable Brassicas*. Science Publishers, pp: 81-124.
- Borisjuk L, Neuberger T, Schwender J, Heinzl N, Sunderhaus S, Fuchs J, Hay JO, Tschiersch H, Braun HP, Denolf P, Lambert B, Jakob PM, Rolletschek H** (2013a) Seed architecture shapes embryo metabolism in oilseed rape. *Plant Cell* 25(5): 1625-1640.

- Borisjuk L, Rolletschek H, Neuberger T** (2013b) Nuclear magnetic resonance imaging of lipid in living plants. *Progress in Lipid Research* 52(4): 465-487.
- Bouteillé M, Rolland G, Balsera C, Loudet O, Muller B** (2012) Disentangling the intertwined genetic bases of root and shoot growth in *Arabidopsis*. *PLoS ONE* 7(2): e32319.
- Bowers JE, Chapman BA, Rong J, Paterson AH** (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422(6930): 433-438.
- Breiman L** (2001) Random Forests. *Machine Learning* 45(1): 5-32.
- Breseghello F, Sorrells ME** (2006) Association analysis as a strategy for improvement of quantitative traits in plants. *Crop Science* 46(3): 1323-1330.
- Broman KW, Wu H, Sen S, Churchill GA** (2003) R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19(7): 889-890.
- Calabrese G, Bennett BJ, Orozco L, Kang HM, Eskin E, Dombret C, De Backer O, Lusi AJ, Farber CR** (2012) Systems genetic analysis of osteoblast-lineage cells. *PLoS Genetics* 8(12): e1003150.
- Cavanagh C, Morell M, Mackay I, Powell W** (2008) From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. *Current Opinion in Plant Biology* 11(2): 215-221.
- CFIA** (2014) The Biology of *Brassica rapa* L. Canadian Food Inspection Agency (CFIA). The Plant Biosafety Office, Ottawa, Canada. <http://www.inspection.gc.ca/plants/plants-with-novel-traits/applicants/directive-94-08/biology-documents/brassica-rapa-l-eng/1330965093062/1330987674945>.
- Chen B, Heneen W** (1992) Inheritance of seed colour in *Brassica campestris* L. and breeding for yellow-seeded *B. napus* L. *Euphytica* 59(2-3): 157-163.
- Chen D, Neumann K, Friedel S, Kilian B, Chen M, Altmann T, Klukas C** (2014) Dissecting the phenotypic components of crop plant growth and drought responses based on high-throughput image analysis. *Plant Cell* 26(12): 4636-4655.
- Chen S, Nelson MN, Ghamkhar K, Fu T, Cowling WA** (2007b) Divergent patterns of allelic diversity from similar origins: the case of oilseed rape (*Brassica napus* L.) in China and Australia. *Genome* 51(1): 1-10.
- Chen S, Zou J, Cowling WA, Meng J** (2010) Allelic diversity in a novel gene pool of canola-quality *Brassica napus* enriched with alleles from *B. rapa* and *B. carinata*. *Crop and Pasture Science* 61(6): 483-492.
- Chen X, Liu CT, Zhang M, Zhang H** (2007a) A forest-based approach to identifying gene and gene-gene interactions. *Proceedings of the National Academy of Sciences* 104(49): 19199-19203.
- Cheng F, Mandáková T, Wu J, Xie Q, Lysak MA, Wang X** (2013a) Deciphering the diploid ancestral genome of the mesohexaploid *Brassica rapa*. *Plant Cell* 25(5): 1541-1554.
- Cheng F, Wu J, Fang L, Sun S, Liu B, Lin K, Bonnema G, Wang X** (2012) Biased gene fractionation and dominant gene expression among the subgenomes of *Brassica rapa*. *PLoS ONE* 7(5): e36442.
- Cheng X, Cheng J, Huang X, Lai Y, Wang L, Du W, Wang Z, Zhang H** (2013b) Dynamic quantitative trait loci analysis of seed reserve utilization during three germination stages in rice. *PLoS ONE* 8(11): e80002.
- Chia TYP, Pike MJ, Rawsthorne S** (2005) Storage oil breakdown during embryo development of *Brassica napus* (L.). *Journal of Experimental Botany* 56(415): 1285-1296.

- Chiang GCK, Barua D, Kramer EM, Amasino RM, Donohue K** (2009) Major flowering time gene, *FLOWERING LOCUS C*, regulates seed germination in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences* 106(28): 11661-11666.
- Chitwood DH, Topp CN** (2015) Revealing plant cryptotypes: defining meaningful phenotypes among infinite traits. *Current Opinion in Plant Biology* 24: 54-60.
- Civelek M, Lusi AJ** (2014) Systems genetics approaches to understand complex traits. *Nature Reviews Genetics* 15(1): 34-48.
- Csanádi G, Vollmann J, Stift G, Lelley T** (2001) Seed quality QTLs identified in a molecular map of early maturing soybean. *Theoretical and Applied Genetics* 103(6-7): 912-919.
- Dang Z, Zheng L, Wang J, Gao Z, Wu S, Qi Z, Wang Y** (2013) Transcriptomic profiling of the salt-stress response in the wild recretahalophyte *Reaumuria trigyna*. *BMC Genomics* 14: 29.
- Darrouzet-Nardi A** (2012) hotspots: an R package version 1.0.2. <http://CRAN.R-project.org/package=hotspots>.
- de Candolle A** (1959) *Origin of cultivated plants*. Hafner, New York.
- de Nooy W, Mrvar A, Batagelj V** (2005) *Exploratory social network analysis with Pajek*. Cambridge: Cambridge University Press.
- DeRose-Wilson L, Gaut BS** (2011) Mapping salinity tolerance during *Arabidopsis thaliana* germination and seedling growth. *PLoS ONE* 6(8): e22832.
- Denford K, Vaughan J** (1977) A comparative study of certain seed isoenzymes in the ten chromosome complex of *Brassica Campestris* and its allies. *Annals of Botany* 41(2): 411-418.
- Deng W, Chen G, Peng F, Truksa M, Snyder CL, Weselake RJ** (2012) Transparent *Testa 16* plays multiple roles in plant development and is involved in lipid synthesis and embryo development in Canola. *Plant Physiology* 160(2): 978-989.
- Dong J, Keller W, Yan W, Georges F** (2004) Gene expression at early stages of *Brassica napus* seed development as revealed by transcript profiling of seed-abundant cDNAs. *Planta* 218(3): 483-491.
- Druka A, Muehlbauer G, Druka I, Caldo R, Baumann U, Rostoks N, Schreiber A, Wise R, Close T, Kleinhofs A, Graner A, Schulman A, Langridge P, Sato K, Hayes P, McNicol J, Marshall D, Waugh R** (2006) An atlas of gene expression from seed to seed through barley development. *Functional & Integrative Genomics* 6(3): 202-211.
- Economic Research Service** (2008) *Oil crops outlook*. USDA Economic Research Service. U.S. Government Priong office, Washington, DC.
- El-Kassaby Y, Moss I, Kolotelo D, Stoehr M** (2008) Seed germination: mathematical representation and parameters extraction. *Forest Science* 54(2): 220-227.
- Elliott RH, Mann LW, Olfert OO** (2007) Effects of seed size and seed weight on seedling establishment, seedling vigour and tolerance of summer turnip rape (*Brassica rapa*) to flea beetles, *Phyllotreta spp.* *Canadian Journal of Plant Science* 87(2): 385-393.
- Ellis R** (1992) Seed and seedling vigour in relation to crop growth and yield. *Plant Growth Regulation* 11: 249-255.
- Fait A, Angelovici R, Less H, Ohad I, Urbanczyk-Wochniak E, Fernie AR, Galili G** (2006) *Arabidopsis* seed development and germination is associated with temporally distinct metabolic switches. *Plant Physiology* 142(3): 839-854.
- Falush D, Stephens M, Pritchard JK** (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164(4): 1567-1587.

- Falush D, Stephens M, Pritchard JK** (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes* 7(4): 574-578.
- Feltus FA** (2014) Systems genetics: a paradigm to improve discovery of candidate genes and mechanisms underlying complex traits. *Plant Science* 223: 45-48.
- Fernandez DE, Turner FR, Crouch ML** (1991) *In situ* localization of storage protein mRNAs in developing meristems of *Brassica napus* embryos. *Development* 111(2): 299-313.
- Fernandez GCJ** (1992). Effective selection criteria for assessing plant stress tolerance. In: Kuo CG (ed.) *Proceedings of the International Symposium on Adaptation of vegetables and other food crops in temperature and water stress: Asian Vegetable Research and Development Center (AVRDC)*, pp: 257-270.
- Fernie AR and Schauer N** (2008) Metabolomics-assisted breeding: a viable option for crop improvement? *Trends in Genetics* 25(1): 39-48.
- Finch-Savage WE, Clay HA, Lynn JR, Morris K** (2010) Towards a genetic understanding of seed vigour in small-seeded crops using natural variation in *Brassica oleracea*. *Plant Science* 179(6): 582-589.
- Foolad MR, Subbiah P, Zhang L** (2007) Common QTL affect the rate of tomato seed germination under different stress and nonstress conditions. *International Journal of Plant Genomics* 2007: 97386.
- Fu J, Jansen RC** (2006) Optimal design and analysis of genetic studies on gene expression. *Genetics* 172(3): 1993-1999.
- Fu J, Wolfs MGM, Deelen P, Westra HJ, Fehrmann RS, te Meerman GJ, Buurman WA, Rensen SS, Groen HJ, Weersma RK, van den Berg LH, Veldink J, Ophoff RA, Snieder H, van Heel D, Jansen RC, Hofker MH, Wijmenga C, Franke L** (2012) Unravelling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genetics* 8(1): e1002431.
- Furbank RT, Tester M** (2011) Phenomics – technologies to relieve the phenotyping bottleneck. *Trends in Plant Science* 16(12): 635-644.
- Fusi N, Stegle O, Lawrence ND** (2012) Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Computational Biology* 8(1): e1002330.
- Gaffney D, Veyrieras JB, Degner J, Pique-Regi R, Pai A, Crawford G, Stephens M, Gilad Y, Pritchard J** (2012) Dissecting the regulatory architecture of gene expression QTLs. *Genome Biology* 13(1): R7.
- Galpaz N, Reymond M** (2010) Natural variation in *Arabidopsis thaliana* revealed a genetic network controlling germination under salt stress. *PLoS ONE* 5(12):e15198.
- Gaur V, Singh US, Kumar A** (2011) Transcriptional profiling and *in silico* analysis of Dof transcription factor gene family for understanding their regulation during seed development of rice *Oryza sativa* L. *Molecular Biology Reports* 38(4): 2827-2848.
- Gilchrist EJ, Sidebottom CHD, Koh CS, MacInnes T, Sharpe AG, Haughn GW** (2013) A mutant *Brassica napus* (Canola) population for the identification of new genetic diversity via TILLING and next generation sequencing. *PLoS ONE* 8(12): e84303.
- Girke T, Todd J, Ruuska S, White J, Benning C, Ohlrogge J** (2000) Microarray analysis of developing *Arabidopsis* seeds. *Plant Physiology* 124(4): 1570-1581.
- Gislason PO, Benediktsson JA, Sveinsson JR** (2006) Random Forests for land cover classification. *Pattern Recognition Letters* 27(4): 294-300.

- Goffman FD, Velasco L, Becker HC** (1999) Tocopherols accumulation in developing seeds and pods of rapeseed (*Brassica napus* L.). *Lipid* 101(10): 400-403.
- Gómez-Campo C** (1999) Biology of *Brassica* Coenospecies. In: Gomez-Campo C (ed.) *Developments in Plant Genetics and Breeding*. Elsevier, Amsterdam, The Netherlands.
- Graham IA** (2008) Seed storage oil mobilization. *Annual Review of Plant Biology* 59: 115-142.
- Graham NS, Hammond JP, Lysenko A, Mayes S, Ó Lochlainn S, Blasco B, Bowen HC, Rawlings CJ, Rios JJ, Welham S, Carion PWC, Dupuy LX, King GJ, White PJ, Broadley MR** (2014) Genetical and comparative genomics of *Brassica* under altered Ca supply identifies *Arabidopsis* Ca-transporter orthologs. *Plant Cell* 26(7): 2818-2830.
- Gupta S, Stamatoyannopoulos J, Bailey T, Noble W** (2007) Quantifying similarity between motifs. *Genome Biology* 8(2): R24.
- Guschina IA, Harwood JL** (2007) Complex lipid biosynthesis and its manipulation in plants. In: Ranalli P (ed.) *Improvement of crop plants for industrial end uses*. Springer Dordrecht, The Netherlands, pp: 253-279.
- Häder T, Müller S, Aguilera M, Eulenberg KG, Steuernagel A, Ciossek T, Kühnlein RP, Lemaire L, Fritsch R, Dohrmann C, Vetter IR, Jäckle H, Doane WW, Brönnner G** (2003) Control of triglyceride storage by a WD40/TPR-domain protein. *EMBO Reports* 4(5): 511-516.
- Hammond JP, Mayes S, Bowen HC, Graham NS, Hayden RM, Love CG, Spracklen WP, Wang J, Welham SJ, White PJ, King GJ, Broadley MR** (2011) Regulatory hotspots are associated with plant gene expression under varying soil phosphorus supply in *Brassica rapa*. *Plant Physiology* 156(3): 1230-1241.
- Hardy OJ, Vekemans X** (2002) SPAGeDi: a versatile computer program to analyze spatial genetic structure at the individual or population levels. *Molecular Ecology Notes* 2: 618–620.
- Harwood J** (2005) Fatty acid biosynthesis. In: Murphy DJ (ed.) *Plant Lipids: biology, utilisation and manipulation*. Blackwell Publishing, Oxford, pp: 27-66.
- Hayashi Y, Hayashi M, Hayashi H, Hara-Nishimura I, Nishimura M** (2001) Direct interaction between glyoxysomes and lipid bodies in cotyledons of the *Arabidopsis thaliana ped1* mutant. *Protoplasma* 218(1): 83-94.
- Holloway B, Luck S, Beatty M, Rafalski JA, Li B** (2011) Genome-wide expression quantitative trait loci (eQTL) analysis in maize. *BMC Genomics* 12: 336.
- Hong CP, Kwon SJ, Kim JS, Yang TJ, Park BS, Lim YP** (2008) Progress in understanding and sequencing the genome of *Brassica rapa*. *International Journal of Plant Genomics* 2008.
- Horvath S, Dong J** (2008) Geometric interpretation of gene coexpression network analysis. *PLoS Computational Biology* 4(8): e1000117.
- Hu Y, Wu G, Cao Y, Wu Y, Xiao L, Li X, Lu C** (2009) Breeding response of transcript profiling in developing seeds of *Brassica napus*. *BMC Molecular Biology* 10: 49.
- Huang BE, George AW, Forrest KL, Kilian A, Hayden MJ, Morell MK, Cavanagh CR** (2012) A multiparent advanced generation inter-cross population for genetic analysis in wheat. *Plant Biotechnology Journal* 10(7): 826-839.
- Huang X, Paulo MJ, Boer M, Effgen S, Keizer P, Koornneef M, van Eeuwijk FA** (2011) Analysis of natural allelic variation in *Arabidopsis* using a multiparent recombinant inbred line population. *Proceedings of the National Academy of Sciences* 108(11): 4488-4493.
- Hurtado-Lopez P** (2012) Investigating genotype by environment and QTL by environment interactions for developmental traits in Potato. PhD thesis. Wageningen University, Wageningen, The Netherlands.

- Ilić-Grubor K, Attree SM, Fowke LC** (1998) Comparative morphological study of zygotic and microspore-derived embryos of *Brassica napus* L. as revealed by scanning electron microscopy. *Annals of Botany* 82: 157-165.
- Inohara N, Nuñez G** (2002) ML – a conserved domain involved in innate immunity and lipid metabolism. *Trends in Biochemical Sciences* 27(5): 219-221.
- Jagannath A, Sodhi Y, Gupta V, Mukhopadhyay A, Arumugam N, Singh I, Rohatgi S, Burma P, Pradhan A, Pental D** (2011) Eliminating expression of erucic acid-encoding loci allows the identification of “hidden” QTL contributing to oil quality fractions and oil content in *Brassica juncea* (Indian mustard). *Theoretical and Applied Genetics* 122(6): 1091-1103.
- Jannink JL, Jansen R** (2001) Mapping epistatic quantitative trait loci with one-dimensional genome searches. *Genetics* 157(1): 445-454.
- Jansen RC, Nap JP** (2001) Genetical genomics: the added value from segregation. *Trends in Genetics* 17(7): 388-391.
- Jestin C, Lodé M, Vallée P, Domin C, Falentin C, Horvais R, Coedel S, Manzanares-Dauleux M, Delourme R** (2011) Association mapping of quantitative resistance for *Leptosphaeria maculans* in oilseed rape (*Brassica napus* L.). *Molecular Breeding* 27(3): 271–287.
- Jiang H, Wu P, Zhang S, Song C, Chen Y, Li M, Jia Y, Fang X, Chen F, Wu G** (2012) Global analysis of gene expression profiles in developing physic nut (*Jatropha curcas* L.) seeds. *PLoS ONE* 7(5): e36522.
- Jiang R, Tang W, Wu X, Fu W** (2009) A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinformatics* 10: S65.
- Jolivet P, Boulard C, Bellamy A, Valot B, d’Andréa S, Zivy M, Nesi N, Chardot T** (2011) Oil body proteins sequentially accumulate throughout seed development in *Brassica napus*. *Journal of Plant Physiology* 168(17): 2015-2020.
- Joosen RVL** (2013) Imaging genetics of seed performance. PhD thesis. Wageningen University, Wageningen, The Netherlands.
- Joosen RVL, Kodde J, Willems LAJ, Ligterink W, van der Plas LHW, Hilhorst HWM** (2010) Germinator: a software package for high-throughput scoring and curve fitting of *Arabidopsis* seed germination. *The Plant Journal* 62(1): 148-159.
- Jordan MC, Somers DJ, Banks TW** (2007) Identifying regions of the wheat genome controlling seed development by mapping expression quantitative trait loci. *Plant Biotechnology Journal* 5(3): 442-453.
- Kadkol G, Beilharz V, Halloran G, Macmillan R** (1986) Anatomical basis of shatter-resistance in the oilseed Brassicas. *Australian Journal of Botany* 34(5): 595-601.
- Kamal-Eldin A, Appelqvist LÅ** (1996) The chemistry and antioxidant properties of tocopherols and tocotrienols. *Lipids* 31(7): 671-701.
- Kang HM, Ye C, Eskin E** (2008a) Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics* 180(4): 1909-1925.
- Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E** (2008b) Efficient control of population structure in model organism association mapping. *Genetics* 178(3): 1709-1723.
- Karim MM, Siddika A, Tonu NN, Hossain DM, Meah MB, Kawanabe T, Fujimoto R, Okazaki K** (2014) Production of high yield short duration *Brassica napus* by interspecific hybridization between *B. oleracea* and *B. rapa*. *Breeding Science* 63(5): 495-502.
- Kazmi RH, Khan N, Willems LAJ, van Heusden AW, Ligterink W, Hilhorst HWM** (2012) Complex genetics controls natural variation among seed quality phenotypes in a recombinant

- inbred population of an interspecific cross between *Solanum lycopersicum* × *Solanum pimpinellifolium*. *Plant, Cell & Environment* 35(5): 929-951.
- Khan N, Kazmi RH, Willems LAJ, van Heusden AW, Ligterink W, Hilhorst HWM** (2012) Exploring the natural variation for seedling traits and their link with seed dimensions in tomato. *PLoS ONE* 7(8): e43991.
- Kelly AA, Quettier AL, Shaw E, Eastmond PJ** (2011) Seed storage oil mobilization is important but not essential for germination or seedling establishment in *Arabidopsis*. *Plant Physiology* 157(2): 866-875.
- Khurana E, Fu Y, Chen J, Gerstein M** (2013) Interpretation of genomic variants using a unified biological network approach. *PLoS Computational Biology* 9(3): e1002886.
- Kloosterman B, Anithakumari AM, Chibon PY, Oortwijn M, van der Linden GC, Visser RGF, Bachem CWB** (2012) Organ specificity and transcriptional control of metabolic routes revealed by expression QTL profiling of source--sink tissues in a segregating potato population. *BMC Plant Biology* 12: 17.
- Kooke R** (2014) Missing heritability and soft inheritance of morphology and metabolism in *Arabidopsis*. PhD thesis. Wageningen University, Wageningen, The Netherlands.
- Koornneef M, Bentsink L, Hilhorst H** (2002) Seed dormancy and germination. *Current Opinion in Plant Biology* 5(1): 33-36.
- Kraakman A, Martínez F, Mussiraliev B, van Eeuwijk F, Niks R** (2006) Linkage disequilibrium mapping of morphological, resistance, and other agronomically relevant traits in modern spring barley cultivars. *Molecular Breeding* 17(1): 41-58.
- Kumar H, Anubha, Vishwakarma M, Lal J** (2011) Morphological and molecular characterization of *Brassica rapa* ssp yellow sarson mutants. *Journal of Oilseed Brassica* 2: 1-6.
- Langfelder P, Horvath S** (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9: 559.
- Laudencia-Chingcuanco D, Stamova B, You F, Lazo G, Beckles D, Anderson O** (2007) Transcriptional profiling of wheat caryopsis development using cDNA microarrays. *Plant Molecular Biology* 63(5): 651-668.
- Le BH, Cheng C, Bui AQ, Wagmaister JA, Henry KF, Pelletier J, Kwong L, Belmonte M, Kirkbride R, Horvath S, Drews GN, Fischer RL, Okamuro JK, Harada JJ, Goldberg RB** (2010) Global analysis of gene activity during *Arabidopsis* seed development and identification of seed-specific transcription factors. *Proceedings of the National Academy of Sciences* 107(18): 8063-8070.
- Lee JM, Williams M, Tingey S, Rafalski A** (2002) DNA array profiling of gene expression changes during maize embryo development. *Functional & Integrative Genomics* 2(1-2): 13-27.
- Lee KR, In Sohn S, Jung JH, Kim SH, Roh KH, Kim JB, Suh MC, Kim HU** (2013) Functional analysis and tissue-differential expression of four *FAD2* genes in amphidiploid *Brassica napus* derived from *Brassica rapa* and *Brassica oleracea*. *Gene* 531(2): 253-262.
- Lenhard B, Wasserman WW** (2002) TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics* 18(8): 1135-1136.
- Li L, Bass RL, Liang Y** (2008) fdrMotif: identifying *cis*-elements by an EM algorithm coupled with false discovery rate control. *Bioinformatics* 24(5): 629-636.
- Li W, Gao Y, Xu H, Zhang Y, Wang J** (2012a) A proteomic analysis of seed development in *Brassica campestris* L. *PLoS ONE* 7(11): e50290.

- Li X, Chen L, Hong M, Zhang Y, Zu F, Wen J, Yi B, Ma C, Shen J, Tu J, Fu T (2012b) A large insertion in *bHLH* transcription factor *BrTT8* resulting in yellow seed coat in *Brassica rapa*. PLoS ONE 7(9): e44145.
- Li Y, Huang Y, Bergelson J, Nordborg M, Borevitz JO (2010) Association mapping of local climate-sensitive quantitative trait loci in *Arabidopsis thaliana*. Proceedings of the National Academy of Sciences 107(49), 21199–21204.
- Li Y, Swertz M, Vera G, Fu J, Breitling R, Jansen R (2009) designGG: an R-package and web tool for the optimal design of genetical genomics experiments. BMC Bioinformatics 10: 188.
- Liaw A, Wiener M (2002) Classification and regression by randomForest. R News 2(3): 18-22.
- Liu P, Wang C, Li L, Sun F, Liu P, Yue G (2011) Mapping QTLs for oil traits and eQTLs for oleosin genes in *Jatropha*. BMC Plant Biology 11: 132.
- Liu S, Liu Y, Yang X, Tong C, Edwards D, Parkin IAP, Zhao M, Ma J, Yu J, Huang S, Wang X, Wang J, Lu K, Fang Z, Bancroft I, Yang TJ, Hu Q, Wang X, Yue Z, Li H, Yang L, Wu J, Zhou Q, Wang W, King GJ, Pires JC, Lu C, Wu Z, Sampath P, Wang Z, Guo H, Pan S, Yang L, Min J, Zhang D, Jin D, Li W, Belcram H, Tu J, Guan M, Qi C, Du D, Li J, Jiang L, Batley J, Sharpe AG, Park B-S, Ruperao P, Cheng F, Waminal NE, Huang Y, Dong C, Wang L, Li J, Hu Z, Zhuang M, Huang Y, Huang J, Shi J, Mei D, Liu J, Lee T-H, Wang J, Jin H, Li Z, Li X, Zhang J, Xiao L, Zhou Y, Liu Z, Liu X, Qin R, Tang X, Liu W, Wang Y, Zhang Y, Lee J, Kim HH, Denoeud F, Xu X, Liang X, Hua W, Wang X, Wang J, Chalhou B, Paterson AH (2014) The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. Nature Communications 5:3930.
- Liu X, Fu J, Gu D, Liu W, Liu T, Peng Y, Wang J, Wang G (2008) Genome-wide analysis of gene expression profiles during the kernel development of maize (*Zea mays* L.). Genomics 91(4): 378-387.
- Lou P, Zhao JJ, He HJ, Hanhart C, Pino Del Carpio D, Verkerk R, Custers J, Koornneef M, Bonnema G (2008) Quantitative trait loci for glucosinolate accumulation in *Brassica rapa* leaves. New phytologist 179 (4): 1017–1032.
- Lukens LN, Quijada PA, Udall J, Pires JC, Schranz ME, Osborn TC (2004) Genome redundancy and plasticity within ancient and recent *Brassica* crop species. Biological Journal of the Linnean Society 82: 665-674.
- Lühs WW, Voss A, Seyis F, Friedt W (1999) Molecular genetics of erucic acid content in the genus *Brassica*. In: Wratten N, Salisbury P (eds.) New horizons for an old crop. Proceedings of the 10th International Rapeseed Congress, Canberra, Australia.
- Lunetta K, Hayward LB, Segal J, Van Eerdewegh P (2004) Screening large-scale association study data: exploiting interactions using random forests. BMC Genetics 5: 32.
- Maeo K, Tokuda T, Ayame A, Mitsui N, Kawai T, Tsukagoshi H, Ishiguro S, Nakamura K (2009) An AP2-type transcription factor, *WRINKLED1*, of *Arabidopsis thaliana* binds to the AW-box sequence conserved among proximal upstream regions of genes involved in fatty acid synthesis. The Plant Journal 60(3): 476-487.
- Mäkinen V-P, Civelek M, Meng Q, Zhang B, Zhu J, Levian C, Huan T, Segrè AV, Ghosh S, Vivar J, Nikpay M, Stewart AFR, Nelson CP, Willenborg C, Erdmann J, Blakenberg S, O'Donnell CJ, März W, Laaksonen R, Epstein SE, Kathiresan S, Shah SH, Hazen SL, Reilly MP, Lusi AJ, Samani NJ, Schunkert H, Quertermous T, McPherson R, Yang X, Assimes TL, the Coronary ADG-WR, Meta-Analysis C (2014) Integrative genomics reveals novel molecular pathways and gene networks for coronary artery disease. PLoS Genetics 10(7): e1004502.

- Malosetti M, van der Linden CG, Vosman B, van Eeuwijk FA** (2007) A mixed-model approach to association mapping using pedigree information with an illustration of resistance to *Phytophthora infestans* in potato. *Genetics* 175(2): 879-889.
- Mano Y, Takeda K** (1997) Mapping quantitative trait loci for salt tolerance at germination and the seedling stage in barley (*Hordeum vulgare* L.). *Euphytica* 94(3): 263-272.
- Mason M, Fan G, Plath K, Zhou Q, Horvath S** (2009) Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells. *BMC Genomics* 10: 327.
- McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q, Flint-Garcia S, Thornsberry J, Acharya C, Bottoms C, Brown P, Browne C, Eller M, Guill K, Harjes C, Kroon D, Lepak N, Mitchell SE, Peterson B, Pressoir G, Romero S, Rosas MO, Salvo S, Yates H, Hanson M, Jones E, Smith S, Glaubitz JC, Goodman M, Ware D, Holland JB, Buckler ES** (2009) Genetic properties of the maize nested association mapping population. *Science* 325(5941): 737-740.
- Meireles-Filho ACA, Stark A** (2009) Comparative genomics of gene regulation-conservation and divergence of *cis*-regulatory information. *Current Opinion in Genetics & Development* 19(6): 565-570.
- Mena M, Vicente-Carbajosa J, Schmidt Robert J, Carbonero P** (1998) An endosperm-specific *DOF* protein from barley, highly conserved in wheat, binds to and activates transcription from the prolamin-box of a native B-hordein promoter in barley endosperm. *The Plant Journal* 16(1): 53-62.
- Merico D, Isserlin R, Stueker O, Emili A, Bader GD** (2010) Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation. *PLoS ONE* 5(11): e13984.
- Miquel M, Browse J** (1995) Lipid biosynthesis in developing seeds. *In*: Kigel J, Galili G (eds.) *Seed development and germination*. Marcel Dekker, New York, pp: 169-193.
- Mittler R** (2006) Abiotic stress, the field environment and stress combination. *Trends in Plant Science* 11(1): 15-19.
- Mun J, Yang T, Kwon S, Park B** (2011) *Brassica rapa* genome sequencing project: strategies and current status. *In*: Sadowski J, Kole C (eds.) *Genetics, genomics and breeding of vegetable Brassicas*. Science Publishers, Inc, Lebanon, pp: 304-327.
- Nemri A, Atwell S, Tarone AM, Huang YS, Zhao K, Studholme DJ, Nordborg M, Jones JDG** (2010) Genome-wide survey of *Arabidopsis* natural variation in downy mildew resistance using combined association and linkage mapping. *Proceedings of the National Academy of Sciences* 107(22): 10302–10307.
- Ni Z, Kim ED, Ha M, Lackey E, Liu J, Zhang Y, Sun Q, Chen ZJ** (2009) Altered circadian rhythms regulate growth vigour in hybrids and allopolyploids. *Nature* 457(7227): 327-331.
- Niu Y, Wu GZ, Ye R, Lin WH, Shi QM, Xue LJ, Xu XD, Li Y, Du YG, Xue HW** (2009) Global analysis of gene expression profiles in *Brassica napus* developing seeds reveals a conserved lipid metabolism regulation with *Arabidopsis thaliana*. *Molecular Plant* 2(5): 1107-1122.
- O'Neill CM, Bancroft I** (2000) Comparative physical mapping of segments of the genome of *Brassica oleracea* var. *alboglabra* that are homoeologous to sequenced regions of chromosomes 4 and 5 of *Arabidopsis thaliana*. *The Plant Journal* 23(2): 233-243.
- Opgen-Rhein R, Strimmer K** (2007) From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Systems Biology* 1: 37.

- Ouyang B, Yang T, Li H, Zhang L, Zhang Y, Zhang J *et al.*, (2007) Identification of early salt stress response genes in tomato root by suppression subtractive hybridization and microarray analysis. *Journal of Experimental Botany* 58(3): 507-520.
- Padmaja L, Agarwal P, Gupta V, Mukhopadhyay A, Sodhi Y, Pental D, Pradhan A (2014) Natural mutations in two homoeologous *TT8* genes control yellow seed coat trait in allotetraploid *Brassica juncea* (AABB). *Theoretical and Applied Genetics* 127(2): 339-347.
- Pang H, Lin A, Holford M, Enerson BE, Lu B, Lawton MP, Floyd E, Zhao H (2006) Pathway analysis using random forests classification and regression. *Bioinformatics* 22(16): 2028-2036.
- Parkin I, Koh C, Tang H, Robinson S, Kagale S, Clarke W, Town C, Nixon J, Krishnakumar V, Bidwell S, Denoeud F, Belcram H, Links M, Just J, Clarke C, Bender T, Huebert T, Mason A, Pires J, Barker G, Moore J, Walley P, Manoli S, Batley J, Edwards D, Nelson M, Wang X, Paterson A, King G, Bancroft I, Chalhoub B, Sharpe A (2014) Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea*. *Genome Biology* 15(6): R77.
- Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genetics* 2(12): e190.
- Penfield S, Li Y, Gilday AD, Graham S, Graham IA (2006) *Arabidopsis* ABA *INSENSITIVE4* regulates lipid mobilization in the embryo and reveals repression of seed germination by the endosperm. *The Plant Cell* 18(8): 1887-1899.
- Peng Q, Hu Y, Wei R, Zhang Y, Guan C, Ruan Y, Liu C (2010) Simultaneous silencing of *FAD2* and *FAE1* genes affects both oleic acid and erucic acid contents in *Brassica napus* seeds. *Plant Cell Reports* 29(4): 317-325.
- Peng F, Weselake R (2011) Gene coexpression clusters and putative regulatory elements underlying seed storage reserve accumulation in *Arabidopsis*. *BMC Genomics* 12: 286.
- Perrier X, Jacquemoud-Collet JP (2006) DARwin software <http://darwin.cirad.fr/darwin>
- Pino Del Carpio D (2010) The genetics of the metabome in *Brassica rapa*. PhD thesis. Wageningen University, Wageningen.
- Pino Del Carpio D, Basnet RK, Arends D, Lin K, De Vos RCH, Muth D, Kodde J, Boutilier K, Bucher J, Wang X, Jansen R, Bonnema G (2014) Regulatory network of secondary metabolism in *Brassica rapa*: insight into the glucosinolate pathway. *PLoS ONE* 9(9): e107123.
- Pino Del Carpio D, Basnet RK, De Vos RCH, Maliepaard C, Paulo MJ, Bonnema G (2011a) Comparative methods for association studies: a case study on metabolite variation in a *Brassica rapa* core collection. *PLoS ONE* 6(5): e19624.
- Pino Del Carpio D, Basnet RK, De Vos RCH, Maliepaard C, Visser RGF, Bonnema G (2011b) The patterns of population differentiation in a *Brassica rapa* core collection. *Theoretical and Applied Genetics* 122(6): 1105-1118.
- Pollard KS, Dudoit S, van der Laan MJ (2005) Multiple Testing Procedures: the multtest package and applications to genomics. In: *Bioinformatics and computational biology solutions using R and Bioconductor*, Springer, pp: 249-271.
- Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D, Lenhard B, Wasserman WW, Sandelin A (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Research* 38 (Database issue): D105-D110.
- Pracharoenwattana I, Zhou W, Smith S (2010) Fatty acid beta-oxidation in germinating *Arabidopsis* seeds is supported by peroxisomal hydroxypyruvate reductase when malate dehydrogenase is absent. *Plant Molecular Biology* 72(1-2): 101-109.

- Prakash S, Hinata K** (1980) Taxonomy, cytogenetics and origin of crop Brassicas, a review. *Opera Botanica* 55-57.
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D** (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38(8): 904-909.
- Pritchard JK, Przeworski M** (2001) Linkage disequilibrium in humans: models and data. *The American Journal of Human Genetics* 69(1): 1-14.
- Pritchard JK, Stephens M, Donnelly P** (2000) Inference of population structure using multilocus genotype data. *Genetics* 155(2): 945-959.
- Qian W, Meng J, Li M, Frauen M, Sass O, Noack J, Jung C** (2006) Introgression of genomic components from Chinese *Brassica rapa* contributes to widening the genetic diversity in rapeseed (*B. napus* L.), with emphasis on the evolution of Chinese rapeseed. *Theoretical and Applied Genetics* 113(1): 49-54.
- Quettier AL, Eastmond PJ** (2009) Storage oil hydrolysis during early seedling growth. *Plant Physiology Biochemistry* 47(6): 485-490.
- Quijada PA, Udall JA, Lambert B, Osborn TC** (2006) Quantitative trait analysis of seed yield and other complex traits in hybrid spring rapeseed (*Brassica napus* L.): 1. Identification of genomic regions from winter germplasm. *Theoretical and Applied Genetics* 113(3), 549–561.
- Rahman H, Harwood J, Weselake R** (2013) Increasing seed oil content in *Brassica* species through breeding and biotechnology. *Lipid Technology* 25(8): 182-185.
- Rahman MH, Joersbo M, Poulsen MH** (2001) Development of yellow-seeded *Brassica napus* of double low quality. *Plant Breeding* 120(6): 473-478.
- Raney JP, Love HK, Rakow GFW, Downey RK** (1987) An apparatus for rapid preparation of oil and oil-free meal from *Brassica* seed. *Lipid* 89(6): 235-237.
- R Core Team** (2012) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>
- Reeves PA, Richards CM** (2009) Accurate inference of subtle population structure (and other genetic discontinuities) using principal coordinates. *PLoS ONE* 4(1): e4269.
- Reiner H, Holzner W, Ebermann R** (1995) The development of turnip-type and oilseed-type *Brassica rapa* crops from the wild type in Europe. - An overview of botanical, historical and linguistic facts. In: Rapeseed today and tomorrow, Vol 4, pp: 1066-1069. 9th International Rapeseed Congress, Cambridge, UK 4-7 July 1995, pp: 1066-1069.
- Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES** (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences* 98(20):11479-11489.
- Ren Z, Zheng Z, Chinnusamy V, Zhu J, Cui X, Iida K, Zhu JK** (2010) RAS1, a quantitative trait locus for salt tolerance and ABA sensitivity in *Arabidopsis*. *Proceedings of the National Academy of Sciences* 107(12): 5669-5674.
- Ritland K** (1996) Estimators for pairwise relatedness and individual inbreeding coefficients. *Genetics Research* 67(02): 175-185.
- Rohlf FJ** (1998) NTSYS-pc: numerical taxonomy and multivariate analysis system, version 3.2. 1st edn., Exeter software, New York.
- Ruuska SA, Girke T, Benning C, Ohlrogge JB** (2002) Contrapuntal networks of gene expression during *Arabidopsis* seed filling. *The Plant Cell* 14(6): 1191-1206.

- Rygulla W, Snowdon RJ, Eynck C, Koopmann B, von Tiedemann A, Lühs W Friedt W** (2007) Broadening the Genetic Basis of *Verticillium longisporum* Resistance in *Brassica napus* by Interspecific Hybridization. *Phytopathology* 97(11): 1391-1396.
- Ryu SB, Lee HY, Doelling JH, Palta JP** (2005) Characterization of a cDNA encoding *Arabidopsis* secretory phospholipase A₂-α, an enzyme that generates bioactive lysophospholipids and free fatty acids. *Biochimica et Biophysica Acta (BBA)-Molecular and cell biology of lipids* 1736(2): 144-151.
- Saad FF, El-Mohsen AAA, El-Shafi MAA, Al-Soudan IH** (2014) Effective selection criteria for evaluating some Barley crosses for water stress tolerance. *Advance in Agriculture and Biology* 2(3): 112-123.
- Sabelli PA** (2012) Seed Development: a comparative overview on biology of morphology, physiology, and biochemistry between monocot and dicot plants. *In: Agrawal GK, Rakwal R, Sabelli P (eds.) Seed development: OMICS technologies toward improvement of seed quality and crop yield.* Springer Netherlands, pp: 3-25.
- Santos-Mendoza M, Dubreucq B, Baud S, Parcy F, Caboche M, Lepiniec L** (2008) Deciphering gene regulatory networks that control seed development and maturation in *Arabidopsis*. *The Plant Journal* 54(4): 608-620.
- Sanyal A, Randal Linder C** (2012) Quantitative trait loci involved in regulating seed oil composition in *Arabidopsis thaliana* and their evolutionary implications. *Theoretical and Applied Genetics* 124(4): 723-738.
- Schlichting CD** (1986) The evolution of phenotypic plasticity in plants. *Annual Review of Ecology and Systematics* 17: 667-693.
- Schmidt R, Acarkan A, Boivin K** (2001) Comparative structural genomics in the *Brassicaceae* family. *Plant Physiology and Biochemistry* 39(3-4): 253-262.
- Schranz ME, Lysak MA, Mitchell SE** (2006) The ABC's of comparative genomics in the *Brassicaceae*: building blocks of crucifer genomes. *Trends in Plant Science* 11(11): 535-542.
- Schwender J, Ohlrogge J, Shachar-Hill Y** (2004) Understanding flux in plant metabolic networks. *Current Opinion in Plant Biology* 7(3): 309-317.
- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T** (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* 13(11): 2498-2504.
- Sharma N, Anderson M, Kumar A, Zhang Y, Giblin EM, Abrams S, Zaharia LI, Taylor D, Fobert P** (2008) Transgenic increases in seed oil content are associated with the differential expression of novel *Brassica*-specific transcripts. *BMC Genomics* 9: 619.
- Simko I** (2004) One potato, two potato: haplotype association mapping in autotetraploids. *Trends in Plant Science* 9(9): 441-448.
- Smyth GK** (2005) limma: linear models for microarray data. *In: Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S (eds.) Bioinformatics and computational biology solutions using R and Bioconductor.* Springer, New York, pp: 397-420.
- Song KM, Osborn TC, Williams PH** (1988) *Brassica* taxonomy based on nuclear restriction fragment length polymorphisms (RFLPs): 2. Preliminary analysis of subspecies within *B. rapa* (syn. *campestris*) and *B. oleracea*. *Theoretical and Applied Genetics* 76(4): 593-600.
- Stamm P, Ravindran P, Mohanty B, Tan E, Yu H, Kumar P** (2012) Insights into the molecular mechanism of RGL2-mediated inhibition of seed germination in *Arabidopsis thaliana*. *BMC Plant Biology* 12: 179.

- Stephenson P, Baker D, Girin T, Perez A, Amoah S, King G, Østergaard L** (2010) A rich TILLING resource for studying gene function in *Brassica rapa*. *BMC Plant Biology* 10: 62.
- Stich B** (2009) Comparison of mating designs for establishing nested association mapping populations in maize and *Arabidopsis thaliana*. *Genetics* 183(4): 1525-1534.
- Stinchcombe JR, Weinig C, Heath KD, Brock MT, Schmitt J** (2009) Polymorphic genes of major effect: consequences for variation, selection and evolution in *Arabidopsis thaliana*. *Genetics* 182(3): 911-922.
- Storey JD, Tibshirani R** (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* 100(16): 9440-9445.
- Su J, Wu S, Xu Z, Qiu S, Luo T, Yang Y, Chen Q, Xia Y, Zou S, Huang BL, Huang B** (2013) Comparison of Salt Tolerance in *Brassicaceae* and some related species. *American Journal of Plant Sciences* 4(10): 1911-1917.
- Susko DJ, Lovett-Doust L** (2000) Patterns of seed mass variation and their effects on seedling traits in *Alliaria petiolata* (*Brassicaceae*). *American Journal of Botany* 87(1): 56-66.
- Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP** (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences* 43(6): 1947-1958.
- Sybesma W, Starrenburg M, Tijsseling L, Hoefnagel MHN, Hugenholtz J** (2003) Effects of cultivation conditions on folate production by lactic acid bacteria. *Applied and Environmental Microbiology* 69(8): 4542-4548.
- Tan H, Yang X, Zhang F, Zheng X, Qu C, Mu J, Fu F, Li J, Guan R, Zhang H, Wang G, Zuo J** (2011) Enhanced seed oil production in Canola by conditional expression of *Brassica napus* *LEAFY COTYLEDON1* and *LEC1-LIKE* in developing seeds. *Plant Physiology* 156(3): 1577-1588.
- Tanhuanpää P, Schulman A** (2002) Mapping of genes affecting linolenic acid content in *Brassica rapa* ssp. *oleifera*. *Molecular Breeding* 10: 51-62.
- Teoh KT, Requesens DV, Devaiah S, Johnson D, Huang X, Howard J, Hood E** (2013) Transcriptome analysis of embryo maturation in maize. *BMC Plant Biology* 13: 19.
- ter Kuile BH, Westerhoff HV** (2001) Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Letters* 500(3): 169-171.
- Thies W** (1971) Schnelle und einfache Analysen der Fettsäurezusammensetzung in einzelnen Raps-Kotyledonen. 1. Gaschromatographische und papierchromatographische Methoden. *Z Pflanzenzüchtg* 65: 181-202.
- Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES** (2001) Dwarf8 polymorphisms associate with variation in flowering time. *Nature Genetics* 28(3): 286-289.
- Tillmann P** (1997) Recent experiences with NIRS analysis in rapeseed. *GCIRC Bull* 13: 84-87.
- Töpfer R, Martini N, Schell J** (1995) Modification of plant lipid synthesis. *Science* 268(5211): 681-687.
- Town CD, Cheung F, Maiti R, Crabtree J, Haas BJ, Wortman JR, Hine EE, Althoff R, Arbogast TS, Tallon LJ, Vigouroux M, Trick M, Bancroft I** (2006) Comparative genomics of *Brassica oleracea* and *Arabidopsis thaliana* reveal gene loss, fragmentation, and dispersal after polyploidy. *The Plant Cell* 18(6): 1348-1359.
- U N** (1935) Genome analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. *Japan Journal of Botany* 7: 389-452.
- Umnajkitikorn K, Faiyue B, Saengnil K** (2013) Enhancing antioxidant properties of germinated Thai rice (*Oryza sativa*) cv Ku Doi Saket with salinity. *Journal of Rice Research* 1:103.

- Usadel B, Nagel A, Thimm O, Redestig H, Blaesing OE, Palacios-Rojas N, Selbig J, Hannemann J, Piques MC, Steinhauser D, Scheible WR, Gibon Y, Morcuende R, Weicht D, Meyer S, Stitt M** (2005) Extension of the visualization tool MapMan to allow statistical analysis of arrays, display of corresponding genes, and comparison with known responses. *Plant Physiology* 138(3): 1195-1204.
- van der Sijde MR, Ng A, Fu J** (2014) Systems genetics: From GWAS to disease pathways. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 1842(10): 1903-1909.
- Van Ooijen JW** (2006) JoinMap 4, Software for the calculation of genetic linkage maps in experimental populations. Kyazma B.V., Wageningen, Netherlands.
- Van Ooijen JW** (2009) MapQTL 6, Software for the mapping of quantitative trait loci in experimental populations of diploid species. Kyazma B.V., Wageningen, Netherlands.
- Varshney RK, Terauchi R, McCouch SR** (2014) Harvesting the promising fruits of genomics: applying genome sequencing technologies to crop breeding. *PLoS Biology* 12(6): e1001883.
- Vicente-Carbajosa J, Moose SP, Parsons RL, Schmidt RJ** (1997) A maize zinc-finger protein binds the prolamin box in zein gene promoters and interacts with the basic leucine zipper transcriptional activator *Opaque2*. *Proceedings of the National Academy of Sciences* 94(14): 7685-7690.
- Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Friters A, Pot J, Paleman J, Kuiper M, Zabeau M** (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acid Research* 23(21): 8.
- Walton LJ, Kurepin LV, Yeung EC, Shah S, Emery RJN, Reid DM, Pharis RP** (2012) Ethylene involvement in silique and seed development of canola, *Brassica napus* L. *Plant Physiology and Biochemistry* 58: 142-150.
- Wan Y, Poole R, Huttly A, Toscano-Underwood C, Feeney K, Welham S, Gooding M, Mills C, Edwards K, Shewry P, Mitchell R** (2008) Transcriptome analysis of grain development in hexaploid wheat. *BMC Genomics* 9: 121.
- Wang H, Guo J, Lambert KN, Lin Y** (2007) Developmental control of *Arabidopsis* seed oil biosynthesis. *Planta* 226(3): 773-783.
- Wang J, Yu H, Weng X, Xie W, Xu C, Li X, Xiao J, Zhang Q** (2014a) An expression quantitative trait loci-guided co-expression analysis for constructing regulatory network using a rice recombinant inbred line population. *Journal of Experimental Botany* 65(4): 1069-1079.
- Wang M, Chen X, Zhang H** (2010a) Maximal conditional chi-square importance in random forests. *Bioinformatics* 26(6): 831-837.
- Wang M, Liu M, Li D, Wu J, Li X, Yang Y** (2010b) Overexpression of *FAD2* promotes seed germination and hypocotyl elongation in *Brassica napus*. *Plant Cell Tissue and Organ Culture* 102(2): 205-211.
- Wang X, Jiang GL, Green M, Scott RA, Hyten DL, Cregan PB** (2014b) Quantitative trait locus analysis of unsaturated fatty acids in a recombinant inbred population of soybean. *Molecular Breeding* 33(2): 281-296.
- Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun JH, Bancroft I, Cheng F, Huang S, Li X, Hua W, Wang J, Wang X, Freeling M, Pires JC, Paterson AH, Chalhoub B, Wang B, Hayward A, Sharpe AG, Park BS, Weisshaar B, Liu B, Li B, Liu B, Tong C, Song C, Duran C, Peng C, Geng C, Koh C, Lin C, Edwards D, Mu D, Shen D, Soumpourou E, Li F, Fraser F, Conant G, Lassalle G, King GJ, Bonnema G, Tang H, Wang H, Belcram H, Zhou H, Hirakawa H, Abe H, Guo H, Wang H, Jin H, Parkin IAP, Batley J, Kim JS, Just J, Li J, Xu J, Deng J, Kim**

- JA, Li J, Yu J, Meng J, Wang J, Min J, Poulain J, Wang J, Hatakeyama K, Wu K, Wang L, Fang L, Trick M, Links MG, Zhao M, Jin M, Ramchiary N, Drou N, Berkman PJ, Cai Q, Huang Q, Li R, Tabata S, Cheng S, Zhang S, Zhang S, Huang S, Sato S, Sun S, Kwon SJ, Choi SR, Lee TH, Fan W, Zhao X, Tan X, Xu X, Wang Y, Qiu Y, Yin Y, Li Y, Du Y, Liao Y, Lim Y, Narusaka Y, Wang Y, Wang Z, Li Z, Wang Z, Xiong Z, Zhang Z (2011a) The genome of the mesopolyploid crop species *Brassica rapa*. *Nature Genetics* 43(10): 1035-1039.
- Wang Z, Chen Z, Cheng J, Lai Y, Wang J, Bao Y, Huang J, Zhang H (2012) QTL analysis of Na⁺ and K⁺ Concentrations in roots and shoots under different levels of NaCl stress in rice (*Oryza sativa* L.). *PLoS ONE* 7(12): e51202.
- Wang Z, Wang J, Bao Y, Wu Y, Zhang H (2011b) Quantitative trait loci controlling rice seed germination under salt stress. *Euphytica* 178(3): 297-307.
- Warwick SI, James T, Falk KC (2008) AFLP-based molecular characterization of *Brassica rapa* and diversity in Canadian spring turnip rape cultivars. *Plant Genetic Resources* 6(01): 11-21.
- Weselake RJ, Taylor DC, Rahman MH, Shah S, Laroche A, McVetty PBE, Harwood JL (2009) Increasing the flow of carbon into seed oil. *Biotechnology Advances* 27(6): 866-878.
- Wind JJ, Peviani A, Snel B, Hanson J, Smeekens SC (2013) *ABI4*: versatile activator and repressor. *Trends in plant science* 18(3): 125-132.
- Wright SI, Gaut BS (2005) Molecular population genetics and the search for adaptive evolution in plants. *Molecular Biology and Evolution* 22(3): 506-519.
- Wu J, Wei K, Cheng F, Li S, Wang Q, Zhao J, Bonnema G, Wang X (2012) A naturally occurring InDel variation in *BraA.FLC.b* (*BrFLC2*) associated with flowering time variation in *Brassica rapa*. *BMC Plant Biology* 12: 151.
- Xiao D, Wang H, Basnet RK, Zhao J, Lin K, Hou X, Bonnema G (2014) Genetic dissection of leaf development in *Brassica rapa* using a genetical genomics approach. *Plant Physiology* 164(3): 1309-1325.
- Xiao D, Zhao JJ, Hou XL, Basnet RK, Carpio DPD, Zhang NW, Bucher J, Lin K, Cheng F, Wang XW, Bonnema G (2013) The *Brassica rapa* *FLC* homologue *FLC2* is a key regulator of flowering time, identified through transcriptional co-expression networks. *Journal of Experimental Botany* 64(14): 4503-4516.
- Xiao L, Zhao Z, Du D, Yao Y, Xu L, Tang G (2012) Genetic characterization and fine mapping of a yellow-seeded gene in Dahuang (a *Brassica rapa* landrace). *Theoretical and Applied Genetics* 124(5): 903-909.
- Xue LJ, Zhang JJ, Xue HW (2012) Genome-wide analysis of the complex transcriptional networks of rice developing seeds. *PLoS ONE* 7(2): e31081.
- Yan X, Li J, Wang R, Jin M, Chen L, Qian W, Wang X, Liu L (2011) Mapping of QTLs controlling content of fatty acid composition in rapeseed (*Brassica napus*). *Genes & Genomics* 33(4): 365-371.
- Yanagisawa S (2004) *Dof* domain proteins: plant-specific transcription factors associated with diverse phenomena unique to plants. *Plant and Cell Physiology* 45(4): 386-391.
- Yang P, Shu C, Chen L, Xu J, Wu J, Liu K (2012a) Identification of a major QTL for silique length and seed weight in oilseed rape (*Brassica napus* L.). *Theoretical and Applied Genetics* 125(2): 285-296.
- Yang Q, Fan C, Guo Z, Qin J, Wu J, Li Q, Fu T, Zhou Y (2012b) Identification of *FAD2* and *FAD3* genes in *Brassica napus* genome and development of allele-specific markers for high oleic and low linolenic acid contents. *Theoretical and Applied Genetics* 125(4): 715-729.

- Yang TJ, Kim JS, Kwon SJ, Lim KB, Choi BS, Kim JA, Jin M, Park JY, Lim MH, Kim HI, Lim YP, Kang JJ, Hong JH, Kim CB, Bhak J, Bancroft I, Park BS** (2006) Sequence-level analysis of the diploidization process in the triplicated *FLOWERING LOCUS C* region of *Brassica rapa*. *The Plant Cell* 18(6): 1339-1347.
- Yang Y, Yu X, Song L, An C** (2011) *ABI4* Activates *DGAT1* Expression in Arabidopsis Seedlings during Nitrogen Deficiency. *Plant Physiology* 156(2): 873-883.
- Ye Y, Zhong X, Zhang H** (2005) A genome-wide tree- and forest-based association analysis of comorbidity of alcoholism and smoking. *BMC Genetics* 6: S135.
- Yu B, Gruber M, Khachatourians GG, Hegedus DD, Hannoufa A** (2010) Gene expression profiling of developing *Brassica napus* seed in relation to changes in major storage compounds. *Plant Science* 178(4): 381-389.
- Yu J, Holland JB, McMullen MD, Buckler ES** (2008) Genetic design and statistical power of nested association mapping in maize. *Genetics* 178(1): 539-551.
- Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES** (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* 38(2): 203-208.
- Zhang G, Zhou W** (2006) Genetic analyses of agronomic and seed quality traits of synthetic oilseed *Brassica napus* produced from interspecific hybridization of *B. campestris* and *B. oleracea*. *Journal of Genetics* 85(1): 45-51.
- Zhang J, Liu H, Sun J, Li B, Zhu Q, Chen S, Zhang H** (2012) *Arabidopsis* fatty acid desaturase *FAD2* is required for salt tolerance during seed germination and early seedling growth. *PLoS ONE* 7(1): e30355.
- Zhao J** (2007) The genetics of phytate content and morphological traits in *Brassica rapa*. PhD thesis. Wageningen University, Wageningen.
- Zhao J, Paulo M-J, Jamar D, Lou P, van Eeuwijk F, Bonnema G, Vreugdenhil D, Koornneef M** (2007) Association mapping of leaf traits, flowering time, and phytate content in *Brassica rapa*. *Genome* 50(10): 963-973.
- Zhao J, Wang X, Deng B, Lou P, Wu J, Sun R, Xu Z, Vromans J, Koornneef M, Bonnema G** (2005) Genetic relationships within *Brassica rapa* as inferred from AFLP fingerprints. *Theoretical and Applied Genetics* 110(7): 1301-1314.
- Zhu T, Budworth P, Chen W, Provart N, Chang H-S, Guimil S, Su W, Estes B, Zou G, Wang X** (2003) Transcriptional control of nutrient partitioning during rice grain filling. *Plant Biotechnology Journal* 1(1): 59-70.

Summary

Seed is the basic and most critical input for seed propagated agricultural crops: seed quality and seedling vigour determine plant establishment, growth and development in both natural and agricultural ecosystems. Seed quality and seedling vigour are mainly determined by the interactions of the following three components: genetic background, physiological quality and the environmental conditions during seed set, seed ripening, storage, seed germination and early seedling development. In the past, many efforts have been made to improve seed germination and seedling vigour by optimizing physiological and environmental factors (non-genetic factors); however, the paradigm has shifted to investigate genetic factors and to use these to improve crop performance by plant breeding. The aim of this thesis is to unravel the genetics of seed germination and seedling vigour under different conditions in *Brassica rapa*, using a systems genetics approach. Studies in many crop species have reported that seed germination and seedling vigour traits are governed by many genes and are strongly affected by environmental conditions. As salinity stress is becoming one of the most important abiotic stresses affecting crop growth and yield, we studied the genetics of seed germination and seedling vigour under neutral and salt stress conditions. For a number of crops, it has been established that larger seed size and higher seed weight indicate more reserve food and contribute positively to seedling establishment. Therefore, our hypothesis for this thesis is that transcriptional regulation of genes during seed development determines the composition and content of seed reserves, and that these seed reserves play a major role in seed germination and seedling growth, especially at the heterotrophic stage under optimal and sub-optimal conditions.

B. rapa is an extremely diverse *Brassica* species which includes, besides many diverse leafy vegetable types and turnips, also oilseed crops. *Brassica* seeds are of high economic importance for several reasons. They are the starting point of the life cycle of the crop, but also they are directly used as sources of vegetable oil or condiments. At present, *B. napus* is the most important source of vegetable oil worldwide, but *B. rapa* is often used for introgression breeding to broaden its narrow genetic base resulting in genetic improvements. Therefore, the acquired knowledge is also useful for the scientific community and plant breeders working in *B. napus* and other *Brassica* species.

In **Chapter 2** we evaluated the genetic diversity of a *B. rapa* core collection of 168 accessions representing different crop types and geographic origins. Using the Bayesian cluster analysis software STRUCTURE, we identified four subpopulations: subpopulation 1 with accessions of Indian origin, spring oil, yellow sarson and rapid cycling; subpopulation 2 consisting of several types from Asian origins: pak choi, winter oil, mizuna, mibuna, komasuna, turnip green, oil rape and Asian turnip; subpopulation 3, which included mainly accessions of Chinese cabbage and subpopulation 4 with mostly vegetable turnip, fodder turnip and brocoletto accessions from European origin. The geographical distribution of the accessions was very much congruent with genetic, metabolic and morphological diversity. This initial study was followed by association studies for secondary metabolites from the tocopherol and carotenoids pathways, using the population structure of these four subpopulations as a correction term to control for spurious

marker-trait associations (**Chapter 2**). Additionally, we used a machine learning approach, Random Forest (RF) regression, to find marker-trait associations. We chose the RF approach as it can handle large numbers of variables (markers, metabolites, transcript abundance) in combination with relatively small sample sets of accessions, to show its perspectives for application to the increasing amounts of data available through the different *~omics* technologies. In our analysis, the markers showing significant association with metabolites identified by the RF approach overlapped with markers obtained from association mapping. Those markers could potentially be used for marker-assisted selection (MAS) in breeding for these secondary metabolites in different morphotypes or sub-populations. Knowledge of genetic distance as evaluated in this chapter allowed the choice of parents to create a segregating population for QTL analyses by maximizing genetic variation between the parents.

In **Chapter 4**, a doubled haploid (DH) population from a cross of genetically diverse morphotypes of *B. rapa*, an oil-type yellow sarson (YS143) and a vegetable pak choi (PC175) (**Chapter 2**), was used to evaluate the genetic basis of seed germination and seedling vigour traits under both non-stress and salt stress conditions. The yellow sarson parent had larger seed size and higher thousand-seed weight than the pak choi parent, and displayed earlier onset, higher uniformity in germination, faster germination and maximum germination, and higher root- and shoot- lengths and biomass under both non-stress and salt stress conditions. Positive correlations of thousand-seed weight with earliness, speed and uniformity of germination and maximum germination percentage, supports that larger seeds germinate earlier, faster, more uniformly and to a higher maximum germination percentage than smaller seeds. Thus, we conclude that yellow sarson had higher seed quality and seedling vigour than pak choi. However yellow sarson also contributed negative alleles to seed germination, as illustrated by its allele of the QTL at A05 which decreases the uniformity of seed germination. In addition we also observed that yellow sarson seedling growth was more affected by salt stress than pak choi. All traits were scored over the DH population, and this clearly showed transgressive variation for most traits. Eight QTL hotspots were identified for seed weight, seed germination, and root and shoot lengths. A QTL hotspot for seed germination on A02 co-located with a homologue of the *FLOWERING LOCUS C* (*BrFLC2*) genes and its *cis*-acting expression QTL (*cis*-eQTL). *FLC2* (*BrFLC2* in *B. rapa*) is an important repressor of flowering time in both *A. thaliana* and *B. rapa* and recently, *FLC2* was reported for its pleiotropic effect on seed germination in *A. thaliana*. A QTL hotspot on A05 with salt stress specific QTL co-located with the *FATTY ACID DESATURASE 2* (*BrFAD2*) gene and its *cis*-eQTL. Besides the role of *FAD2* in fatty acid desaturation, the up-regulation of this gene was associated with enhanced seed germination and hypocotyl elongation under salinity in *B. napus* (*BnFAD2*) and *A. thaliana* (*FAD2*). We observed epistatic interactions between the QTL hotspots at the *BrFLC2* and *BrFAD2* loci, and between other QTL hotspots.

Seed development is regulated by many dynamic metabolic processes controlled by complex networks of spatially and temporally expressed genes. Therefore, morphological characteristics and the transcriptional signatures of developing seeds from yellow- and brown/black-seeded genotypes were studied to get to know the timing of key metabolic processes, to explore the

major transcriptional differences and to identify the optimum stage for a genetical genomics study for *B. rapa* seed traits (**Chapter 3**). This is the first study of genome-wide profiling of transcript abundance during seed development in *B. rapa*. Most transcriptional changes occurred between 25 and 35 days after pollination (between the bent-cotyledon stage and the stage when the embryo fully fills the seed), which is later than in the related species *B. napus*. A weighted gene co-expression network analysis (WGCNA) identified 47 gene modules with different co-expression patterns, of which 17 showed a genotype effect, 4 modules a time effect during seed development and 6 modules both genotype and time effects. Based on the number of genes in gene modules, the predominant variation in gene expression was according to developmental stages rather than morphotype differences. We identified 17 putative *cis*-regulatory elements (motifs) for four co-regulated gene clusters of genes related to lipid metabolism. The identification of key physiological events, major expression patterns, and putative *cis*-regulatory elements provides useful information to construct gene regulatory networks in *B. rapa* developing seeds and provides a starting point for a genetical genomics study of fatty acid composition and additional seed traits in **Chapter 5**.

Since *Brassica* seeds are sources of vegetable oil, genetic studies of the gene regulatory mechanisms underlying lipid metabolism is of high importance, not only in relation to seed and seedling vigour, but also for *Brassica* oilseed breeding. In **Chapter 5**, an integrative approach of QTL mapping for fatty acids composition and for transcript abundance (eQTL) of genes related to lipid metabolism, together with gene co-expression networks was used to unravel the genetic regulation of seed fatty acid composition in the DH population of *B. rapa*. In this study, a confounding effect of flowering time variation was observed on fatty acid QTLs (metabolite level) at linkage group A02 and of seed colour variation on eQTLs (transcript level) at linkage group A09. At A02, fatty acid QTLs from 2009 seeds co-locate with the genetic position of a gene-targeted marker for *BrFLC2*, its *cis*-QTL, and a major flowering time QTL. Flowering time variation is very obvious in this DH population and the *BrFLC2* gene at A02 (16.7 cM) is the major regulator of flowering time, with a non-functional allele in the yellow sarson parent. When QTL analysis was performed on seeds from 2011, from DH lines that flowered synchronously due to staggered sowing, this fatty acid QTL hotspot disappeared. The 2011 seed lot was used for further analysis combining fatty acid QTLs with eQTLs in this study. On A09, a large *trans*-eQTL hotspot was co-localized with a major seed colour QTL, in the region where the causal gene, the *bHLH* transcription factor *BrTT8*, was cloned. The role of this gene in seed colour development was functionally proven in *B. rapa*. As the yellow sarson and pak choi parents of this population have contrasting seed coat colour (**Chapter 3**) the DH lines segregated for seed colour. When seed colour variation was used as a co-variate in our statistical model, we could exclude its confounding effect on eQTL mapping. We compared the fatty acid QTL and eQTL results from the analyses before and after seed colour correction and later discuss the results from the analysis after correction. The distribution of major QTLs for fatty acids showed a relationship with the types of fatty acids: linkage group A03 contained major QTLs for monounsaturated fatty acids (MUFAs), A04 for saturated fatty acids (SFAs) and A05 for polyunsaturated fatty acids (PUFAs). Using a

genetical genomics approach, eQTL hotspots were found at major fatty acid QTLs on A03, A04 and A05 and on A09. Finally, an eQTL-guided gene co-expression network of lipid metabolism related genes showed major hubs at the genes *BrPLA2-ALPHA*, *BrWD-40*, a number of seed storage protein genes and a transcription factor *BrMD-2*, suggesting essential roles of these genes in lipid metabolism. Several genes, such as *BrFAE1*, *BrTAG1*, *BrFAD2*, *BrFAD5*, *BrFAD7*, which were reported as important genes for fatty acid composition in seeds in other studies of related species, had relatively lower degrees of connection in the networks. However their *cis*-eQTLs co-localized with specific fatty acid QTLs, making them candidate genes for the observed variation. We hypothesize that these play a role in modifying fatty acid content or composition across genotypes, rather than playing essential roles in the pathway itself. These results suggest the need of a global study of lipid metabolism rather than a strict focus on the fatty acid biosynthesis pathway per se. This study gives a starting point for understanding the genetic regulation of lipid metabolism, by identification of a number of key regulatory genes, identified as major hub genes, and candidate genes for fatty acid QTLs.

In the final chapter (**Chapter 6**) we summarize and critically discuss the relationships among phenotypic traits, metabolites and expression variation as well as the co-localization of QTLs from these different levels. In this thesis, we developed methodology to integrate transcriptomics and metabolomics data sets and to construct gene regulatory networks related to major fatty acids, and found a set of (possible) candidate genes involved in lipid metabolism. In the future, we recommend to integrate the genome-wide transcriptome data set with all major seed metabolites and phenotypic data on seed and seedling vigour to directly link all three components: transcriptome, metabolome and phenotypic traits, and ultimately expand the knowledge on the genetic regulation of seed metabolites, seed quality and seedling vigour in *B. rapa* to other *Brassica* species.

Acknowledgements

I never dreamed that I would achieve a PhD degree and a Doctor's title, but I am finally at that stage. It was beyond my imagination. At first, I came to Wageningen University as an MSc student in plant sciences, specialization plant breeding and genetic resources with a NUFFIC fellowship. I had high ambitions and motivation to learn all about plant breeding and quantitative genetics, but I did not know what to expect. In Nepal, I had heard about molecular aspects of plant breeding and their role in modern plant breeding. Therefore, I had a very strong motivation to work hard, but I was also nervous to meet these new challenges due to my presumed lack of background knowledge. Luckily, as things progressed, my learning slowly turned in the desired direction, which kept me motivated. During both my major and minor theses with Dr Guusje Bonnema (Brassica group) and Dr Chris Maliepaard (Quantitative genetics group), I was fortunate to work on new and diverse fields in statistical genetics, genetic diversity, QTL mapping, association studies, molecular biology, network analysis and many other issues. Working in both groups was the perfect combination for me, which gave me lots of experience, exposure and insight into the latest developments in the fields of plant breeding, quantitative genetics and statistical genetics. Later, I got the golden opportunity to continue with a PhD study and it was one of the happiest moments in my life. Guusje and Chris, you were not only my supervisors, but were also very good guardians. At the beginning, it was quite difficult to follow discussions in both aspects; biological (with Guusje) and statistical (with Chris) perspectives, but you both were so kind that I was slowly able to catch the discussions. I enjoyed these discussions a lot and had amazing experiences in working with both of you, on topics ranging from field to greenhouse, molecular lab to microarray wet lab, and statistical analyses and integration of high-dimensional omics datasets. I think very few students can have this chance to acquire experience and develop expertise in these diverse fields within the four years' duration of a PhD project. Guusje and Chris, while working with you I learnt many things not only in subject matter, but also scientific skills. I am heartily thankful for everything you did for me; developing research skills and helping me to grow as a scientist.

I am very grateful to Prof. Dr Richard GF Visser, who accepted me as his PhD student. It took more time than expected to finish my thesis, especially on the writing part. Richard, you always were very critical in the discussions, which helped to reshape my way of looking at things. I am very thankful to your invaluable support to complete this thesis.

I would like to give special thanks to my dear friend Johan Bucher for helping me to set up and carry out experiments successfully. Johan, you are always more than a technician, who likes to participate in all kinds of scientific discussion. I always admire your very friendly, supportive and creative attitude. Also, I am very thankful to you and Dennis van Muijen (PhD student and my colleague at Rijk Zwaan) for being my paranymphs and making my graduation a very special event. My deepest thanks also go to Dunia Pino Del Carpio. I am very happy to have had you as my MSc thesis supervisor. You are also a very good friend and I have learnt so much from you that was helpful to develop myself into the academic field. After you left the department, I really missed our scientific discussions, especially at the coffee corner.

Dr Steven Groot, I am very glad for your support and critical remarks as a seed physiologist especially on seed germination and seedling vigour experiments. I would also like to thank Natalia Moreno Panchon, who did an MSc thesis in the Brassica group. Your thesis led to very important results, which were later used to publish a joint paper.

Many BSc and MSc students have contributed to my research. I am very thankful to Dev Nidhi Tiwari for generating phenotypic data of seed germination, which was very useful for chapter 4. I would also like to extend my thanks to Andre Meijaard, Punam, Umer, Frank, Cassie and other students for their help at different stages of the experiments.

We had a very nice, creative and cooperative Brassica group, which I miss a lot since I left for my new job. I have very special memories of working together as a family while conducting experiments in the field and greenhouse, and during other wet lab experiments. I am very thankful to Kristin Hnning, Mina Jin and all the Brassica members for their kind cooperation, especially when there was lots of work during pollination, tagging developing seeds and harvesting seeds, and for social gatherings. Jianjun Zhao and Ningwen Zhang, you were always very friendly, team players and passionate for all kinds of discussions, which I really appreciate. Dong Xiao, you are an amazing person, very hard-working and with persistent nature. Your skills in the molecular lab always surprise me, and your contributions help me a lot in my research, which resulted in the publishing of our collaborative works; thank you very much. I would also like to thank Lin Ke (Ke Lin) for your contribution as a bio-informatician, for being very supportive, especially for your efforts in designing the custom microarray for the gene expression study.

Thanks to Gerrit Polder for providing tools to quantify seed colour using image analysis. I also like to thank the staff at the greenhouses of Unifarm (Nergena) and the molecular labs for their assistance. I especially would like to thank the Center for BioSystems Genomics (CBSG) for funding this research. My project was in collaboration with the National Research Council (NRC) in Saskatoon, Canada. I like to thank very much Dr Pierre Fobert and Kerry Boyle for their contributions in primary metabolite measurements in seeds.

My thanks also go to thank Ronny Joosen (who later became my colleague at Rijk Zwaan), Rashid Kazmi, Narullah Khan and Hanzi He from the Seed Lab (Laboratory of Plant Physiology, WU) for the useful discussions with them. Ronny, you helped us to set up an image analysis platform for my experiments on *B. rapa* seed germination *at plant breeding*. Later, when we became colleagues at Rijk Zwaan, I realized you are an incredibly nice person, always eager to help, smiling, willing to share experiences, open to discuss and facilitating a lot to settle down in a new work place. You are an amazing person.

When I joined Rijk Zwaan in December 2013, I was not completely done with my thesis. I was very worried to complete my thesis as well as to get acquainted with the new job. But I am very impressed with the environment at Rijk Zwaan, especially with my team leader Dr Evert Gutteling, who always encouraged me to complete my thesis. Your regular interest to ask about my progress of the thesis was really useful to speed up thesis writing. Hats off to you Evert and also to Rijk Zwaan for giving me the flexibility needed to complete this thesis. I would also like to thank Paula Hurtado Lopez, first at the university and later at Rijk Zwaan, for giving all kinds of support,

creating a very friendly environment and stimulating me to complete my thesis. Paula, but also Cesar and Maria (my little and cute friend), thank you very much for providing a space to stay when I was in Wageningen to work on the thesis.

During my thesis, I have made many friends who helped me to create a friendly environment not only at the department, but also at off-hours. I would like to thank all of my friends, mainly Yusuf, Xi (Chen), Arwa, Peter Dingh, Maria, Suxian, Animesh Acharjee, Natalia Carreno, Pierre, Thijs, Alessandro, Anitha, Madhuri, Cesar, Louis, Christos, Freddy, Saurav and others. Animesh bro, you always treat me like your brother and I learnt a lot from you, not only about science but also about societal aspects. Thank you very much for our discussions on statistics and systems biology.

Many thanks go to the secretariat of Plant Breeding (Letty, Janneke and Nicole) for their help with all the administrative works.

I would like to thank Peter Kruyssen, Titus and Okko, and Joost van Langen for organizing gatherings and chit-chats.

I am very thankful to everyone from the Nepalese Student Association – Wageningen, which gave me a homely environment and made my stay lively. I am very glad to be associated with this organization and to participate in cultural events and social gatherings. My special thanks to Suresh Baral, Key Shore and Shailendra, who always allowed me to stay at their place during a difficult transition time and whenever I came back to work on my thesis. Also, thanks to Hiranya, Tulsi jee, Khagendra, Ekaraj, Saroj Yakami, Hom dai, Rajendra Uprety dai, Yogesh, Salyan, Subash, Anukram, Arun Thapa, Arun and Puja jee, Deepak (KC and Sapkota), Udit, Dharma, Tilak dai, Pudasainy dai, Shiva dai, Ishwor, Sheeva and all other Nepalese friends for giving me such a nice company.

Lunish (Yakami), thanks for your patience and creativity and the valuable time that you have put into designing a thesis book cover, which provides a wonderful impression of my thesis work. I like it very much.

I would like to take this opportunity to thank also my previous employer, the Nepal Agricultural Research Council (NARC), especially senior scientists/breeders Ashok Mudwari, Dr Dhruva Thapa, Narayan Dhami, Dr. Jwala Bajracharya and Madan Raj Bhatta for motivating me to develop myself as a plant breeder and for other supports.

Since this PhD degree is my special achievement in my career, I would like to take this opportunity to thank Puspa Thapa, Shyam Basnet, Yogendra Basnet and Raj Kumar Niroula for guiding me into the right direction at different time points of my life, which helped me to arrive at this level.

I am very grateful to my family and all relatives, including my beloved parents, wife, brother and his wife Anjana, Sister Kabita Basnet and her husband Shree Prasad Vista for their encouragement. My very special thanks to father (Krishna Bahadur Basnet), mother (Ganga Devi Basnet) and my twin brother (Shyam Kumar Basnet) for their continuous support, good wishes and encouragements for achieving this success, not only during this study but also throughout my life. You are all truly my inspiration; I can't express my thanks with words. I am proud of you all and always grateful. It is my great pleasure to thank my parents-in-law and family, including Sunita sister and Hari brother.

Finally, last but not least, I would like to give a very special thanks to my wife Anita Duwal, who shared both my happiness as well as my difficulties. Anita, you not only encouraged me to complete this thesis, but also gave very remarkable help to complete one of my experiments. At a certain moment, I was in a very difficult situation that I had to repeat a phenotyping experiment at the last moment of the thesis. But you carried out the whole experiment unconditionally full of motivation. So, I would like to dedicate this achievement to you too. I am always proud to have you and feel loved.

Thank you all!

Bedankt iedereen!

Ram Kumar Basnet

About the author

Ram Kumar Basnet was born on October 9, 1980, in a small village (Rajghat) at Morang district in Eastern Nepal. After completing high school, he joined the Institute of Agriculture and Animal Science (IAAS), Tribhuvan University, Nepal and completed Intermediate of Science (I. Sc. Agriculture) in 1998 and the Bachelor degree (B. Sc. Agriculture) in 2003. After his Bachelor degree, he started his career at Nepal Agricultural Research Council (NARC) in the seed technology unit as a technical officer and later as an assistant breeder in the wheat breeding program at Khumaltar. During this work, he was involved in developing superior wheat varieties for mid- and high- hill conditions of Nepal and also was involved in participatory plant breeding (PPB) and participatory varietal selection (PVS) programs in wheat funded by CIMMYT-Nepal. He continued his education with a Master degree in plant breeding and genetic resources at Wageningen University, The Netherlands in 2007, under the Netherland Fellowship Programme (NFP). His major MSc thesis was on multivariate statistical analysis and association studies and his minor thesis was on QTL mapping and molecular marker development in *Brassica rapa*. In September, 2009, he started his PhD study entitled “A systems genetic study of seed germination and seedling vigour in *Brassica rapa*” in two research groups (quality and development, and quantitative genetics) at Wageningen UR Plant Breeding, Wageningen University & Research Centre. In December 2013, he started working as a Quantitative Geneticist at the vegetable breeding company Rijk Zwaan at Fijnaart, The Netherlands.

List of publications

Related to this thesis

Basnet RK, Pino Del Carpio D, Xiao D, Bucher J, Jin M, Lin K, Boyle K, Fobert P, Visser RGF, Maliepaard C, Bonnema G. A systems genetics approach identifies gene regulatory networks associated with fatty acid composition in *Brassica rapa* seed. (**Accepted after revisions – Plant Physiology**).

Basnet RK, Duwal A, Tiwari DN, Xiao D, Monakhos S, Bucher J, Visser RGF, Groot SPC, Bonnema G, Maliepaard C. Quantitative trait loci analysis of seed germination and seedling vigour under non-stress and salt stress conditions in *Brassica rapa*: a possible role of *BrFLC2* and *BrFAD2*. (**under review**)

Pino Del Carpio D, **Basnet RK**, Arends D, Lin K, De Vos RCH, Muth D, Kodde J, Boutilier K, Bucher J, Wang X, Jansen X, Bonnema G (2014) Regulatory network of secondary metabolism in *Brassica rapa*: insight into the glucosinolate pathway. PLoS ONE 9(9): e107123.

Xiao D, Wang H, **Basnet RK**, Zhao J, Lin K, Hou X, Bonnema G (2014) Genetic dissection of leaf development in *Brassica rapa* using a genetical genomics approach. Plant Physiology 164(3): 1309-25.

Zhang N, Zhao J, Lens F, de Visser J, Menamo T, Fang W, Xiao D, Bucher J, **Basnet RK**, Lin K, Cheng F, Wang X, Bonnema G (2014) Morphology, carbohydrate composition and vernalization response in a genetically diverse collection of Asian and European turnips (*Brassica rapa* subsp. *rapa*). PLoS ONE 9(12): e114241.

Basnet RK, Moreno-Pachon N, Lin K, Bucher J, Visser RGF, Maliepaard C, Bonnema G (2013) Genome-wide analysis of coordinated transcript abundance during seed development in different *Brassica rapa* morphotypes. BMC Genomics 14:840.

Xiao D, Zhao JJ, Hou XL, **Basnet RK**, Carpio DP, Zhang NW, Bucher J, Lin K, Cheng F, Wang XW, Bonnema G (2013) The *Brassica rapa* *FLC* homolog *BrFLC2* is a key regulator of flowering time, identified through transcriptional co-expression networks. Journal of Experimental Botany 64(14):4503-16.

Pino Del Carpio D*, **Basnet RK***, De Vos RCH, Maliepaard C, Paulo MJ, Bonnema G (2011) Comparative methods for association studies: a case study on metabolite variation in a *Brassica rapa* core collection. PLoS One 6(5): e19624. doi:10.1371/journal.pone.0019624.

*-authors are equally contributed.

Pino Del Carpio D, **Basnet RK**, De Vos RCH, Maliepaard C, Visser RGF, Bonnema G (2011) The patterns of population differentiation in a *Brassica rapa* core collection. Theoretical and Applied Genetics 122(6):1105–1118.

Lin K, Kools H, de Groot PJ, Gavai AK, **Basnet RK**, Cheng F, Wu J, Wang X, Lommen A, Hooiveld GJ, Bonnema G, Visser RGF, Muller MR, Leunissen JA (2011) MADMAX – management and analysis database for multiple *omics* experiments. Journal of Integrative Bioinformatics 8(2):160.

Zhao J, Artemyeva A, Del Carpio DP, **Basnet RK**, Zhang N, Gao J, Li F, Bucher J, Wang X, Visser RGF, Bonnema G (2010) Design of a *Brassica rapa* core collection for association mapping studies. Genome 53(11): 884-898

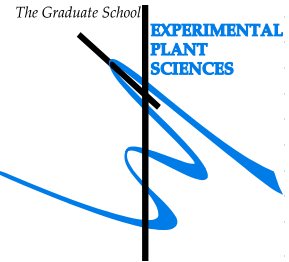
Other publications

Thapa DB, Mudwari A, **Basnet RK**, Sharma S, Ortiz-Ferrara G, Sharma B, Murphy K (2009) Participatory varietal selection of wheat for micro-niches of Kathmandu valley. Journal of Sustainable Agriculture 33 (7): 745-756.

Thapa DB, Sharma RC, Mudwari A, Ortiz-Ferrara G, Sharma S, **Basnet RK**, Witcombe JR, Virk DS, Joshi KD (2009) Identifying superior wheat cultivars in participatory research on resource poor farms. Field Crops Research.12 (2-3): 124-130.

Education Statement of the Graduate School

Experimental Plant Sciences



Issued to: Ram Kumar Basnet
Date: 24 August 2015
Group: Laboratory of Plant Breeding
University: Wageningen University & Research Centre

1) Start-up phase	
<ul style="list-style-type: none"> ► First presentation of your project Genetic dissection of seed quality and seedling vigour in <i>Brassica rapa</i>: systems genetics approach ► Writing or rewriting a project proposal Profiling of primary and secondary metabolites during seed development in selected <i>B. rapa</i> genotypes ► Writing a review or book chapter ► MSc courses MIB 21306 Bioinformation Technology ► Laboratory use of isotopes 	<u>date</u> Oct 01, 2009 Aug 2012 Feb 2010
Subtotal Start-up Phase	
8.5 credits*	
2) Scientific Exposure	
<ul style="list-style-type: none"> ► EPS PhD student days EPS PhD student day, Leiden University EPS PhD student day, Utrecht University EPS PhD student day, Wageningen University EPS PhD student day, University of Amsterdam EPS PhD student day, Leiden University ► EPS theme symposia EPS theme 1 'Developmental Biology of Plants', Wageningen University EPS theme 4 "Genome Biology", Wageningen University EPS theme 1 Developmental Biology of Plants", Wageningen UR EPS theme 3 'Metabolism and Adaptation', Utrecht University ► NWO Lunteren days and other National Platforms ALW meeting 'Experimental Plant Sciences', Lunteren ALW meeting 'Experimental Plant Sciences', Lunteren ALW meeting 'Experimental Plant Sciences', Lunteren ALW meeting 'Experimental Plant Sciences', Lunteren ALW meeting 'Experimental Plant Sciences', Lunteren ALW meeting 'Experimental Plant Sciences', Lunteren ► Seminars (series), workshops and symposia CBSG <i>Brassica</i> cluster meeting 2011 CBSG Cluster meeting Arabidopsis 2012 CBSG summit 2012 Invited seminar Veronica Grieneisen (John Innes Center, Norwich): understanding spatial regulation of intracellular cell polarity, cell shape changes, and intercellular polarity coupling and signaling during tissue morphogenesis Invited seminar Régine Delourme "Genetic and functional analysis of disease resistance in <i>Brassica</i>" Invited seminar Ian Henderson (University of Cambridge): "Genetics and epigenetics" Invited seminar Guilherme Rosa (University of Wisconsin-Madison): "Inferring Causal Phenotype Networks Using Structural Equation Models". Invited seminar ME (Eric) Schranz (University of Amsterdam): Genome duplications, species radiations and trait diversification in the Brassicales Mini-symposia on "Plant Breeding in the Genomics Era", Wageningen Invited seminar Salvatore Ceccarelli (ICARDA): Participatory Plant Breeding - a response to the problems of hunger, biodiversity and climate changes ► Seminar plus ► International symposia and congresses 17th Crucifer Genetics Workshop, Canada Symposium "Improving yield prediction by combining statistics, genetics, physiology and phenotyping: the EU SPICY project in pepper" 4th International Conference on Quantitative Genetics (ICQG), Edinburgh, Scotland Next Generation Plant Breeding Conference, Ede, The Netherlands Challenges for integrated analysis of omics datasets, LUMC, Leiden All-inclusive breeding: Integrating high throughput science, Wageningen 	<u>date</u> Feb 29, 2009 Jun 01, 2010 May 20, 2011 Nov 30, 2012 Nov 29, 2013 Jan 28, 2010 Dec 09, 2011 Jan 19, 2012 Apr 26, 2012 Apr 19-20, 2010 Apr 04-05, 2011 Apr 02-03, 2012 Apr 22-23, 2013 Apr 14-15, 2014 Apr 13-14, 2015 Oct 06, 2011 Oct 23, 2012 Feb 29-Mar 01, 2012 Nov 17, 2011 Oct 06, 2011 Dec 13, 2010 Apr 29, 2011 Nov 23, 2011 Nov 25, 2011 May 29, 2012 Sep 05-08, 2010 Mar 07-09, 2012 Jun 17-22, 2012 Nov 11-14, 2012 Apr 27, 2013 Oct 16, 2014

<p>► Presentations</p> <p>CBSG summit, 2011: "Genetic analysis of seed and seedling vigour under salt stress conditions in <i>Brassica rapa</i>", Wageningen (Poster)</p> <p>Plant Breeding Research Day (Talk)</p> <p>CBSG mid-term evaluation (Talk)</p> <p>CBSG Brassica-Arabidopsis cluster meeting (Talk)</p> <p>CBSG summit, 2012: "Exploring growth models for genetic analysis of early seedling growth in <i>Brassica rapa</i>", Wageningen Netherlands (Poster)</p> <p>"Genetic analysis of fatty acid biosynthesis for seed and seedling vigour in <i>Brassica rapa</i>" in 4th International conference on Quantitative Genetics (ICQG), Edinburgh, Scotland (Poster)</p> <p>"Genome-wide eQTL analysis and the discovery of gene expression networks associated with seed quality and seedling vigour in <i>Brassica rapa</i>" in Next Generation Plant Breeding Conference, Ede, The Netherlands</p> <p>CBSG summit, 2013: "Genome-wide eQTL analyses of seed quality and seedling vigour in <i>Brassica rapa</i>; a case study on fatty acid biosynthesis", Wageningen Netherlands (Poster)</p> <p>Lunteren meeting, 2013: "Genome-wide eQTL analyses of seed quality and seedling vigour in <i>Brassica rapa</i>; a case study on fatty acid biosynthesis" (Talk)</p> <p>► IAB interview</p> <p>Meeting with a member of the International Advisory Board of EPS (Prof. Dr. Ted Farmer)</p> <p>► Excursions</p> <p>Excursion to Key Gene company</p>	<p>Feb 2011</p> <p>Mar 08, 2011</p> <p>Mar 24, 2011</p> <p>Oct 06, 2011</p> <p>Feb 29-01 Mar, 2012</p> <p>Jun 17-22, 2012</p> <p>Nov 11-14, 2012</p> <p>Feb 11-12, 2013</p> <p>April 22-23, 2013</p> <p>Nov 15, 2012</p> <p>Jan 26, 2012</p>
<i>Subtotal Scientific Exposure</i>	
	<i>23.7 credits*</i>
<p>3) In-Depth Studies</p> <p>► EPS courses or other PhD courses</p> <p>Master class on seed technology</p> <p>Basic Data Analysis on Gene Expression Arrays</p> <p>Analysis of Microarray Gene Expression Data using R/BioC and web tools, Rotterdam</p> <p>Statistical learning methods for DNA-based prediction of complex traits</p> <p>Mixed model based QTL mapping in GenStat</p> <p>Systems Biology: statistical analysis of -omics data</p> <p>Methods and models for plant genomic prediction and selection in plant breeding, SLU and Plant Link, Alnarp, Sweden</p> <p>► Journal club</p> <p>Literature group discussion: Laboratory of plant breeding</p> <p>► Individual research training</p>	<p><u>date</u></p> <p>Oct 26-29, 2009</p> <p>Nov 09-11, 2010</p> <p>Jun 20-24, 2011</p> <p>Oct 17-21, 2011</p> <p>May 14-16, 2012</p> <p>Dec 08-11, 2008</p> <p>Sep 08-11, 2014</p> <p>2009-2013</p>
<i>Subtotal In-Depth Studies</i>	
	<i>11.7 credits*</i>
<p>4) Personal development</p> <p>► Skill training courses</p> <p>PCDI career perspectives course</p> <p>PhD Competence Assessment</p> <p>ExPectationS Career Day, Wageningen University</p> <p>Reviewing a Scientific Paper</p> <p>Enza Zaden student experience</p> <p>Personal and professional assessment, The Assessment Centre Berenschot, Utrecht</p> <p>Project management multidisciplinary team</p> <p>Dutch language course</p> <p>► Organisation of PhD students day, course or conference</p> <p>► Membership of Board, Committee or PhD council</p>	<p><u>date</u></p> <p>Mar 13-15, 2013</p> <p>Aug 30 & Sep 28, 2011</p> <p>Nov 18, 2011</p> <p>Dec 20, 2011</p> <p>May 15, 2015</p> <p>Aug 08, 2013</p> <p>Apr 23, May 21, Jun 11, 2015</p> <p>Oct-Dec 2015</p>
<i>Subtotal Personal Development</i>	
	<i>4.0 credits*</i>
TOTAL NUMBER OF CREDIT POINTS*	
	47.9

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

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

Cover design was done by Lunish Yakami (yakami.lunish@gmail.com) and the thesis layout by the author.

Printed by: CPI –Wöhrmann print service, Zutphen