The genetics of the metabolome in *Brassica rapa*

Dunia Pino Del Carpio

Thesis committee

Thesis supervisor
Prof. dr. R.G.F. Visser
Professor of Plant Breeding
Wageningen University

Thesis co-supervisor
Dr. ir. A.B. Bonnema
Assistant professor, Laboratory of Plant Breeding
Wageningen University

Other members
Prof. dr. ir. M. Koornneef, Wageningen University
Dr. M.J. Regine Delourme, INRA, France
Prof. dr. H.J. Bouwmeester, Wageningen University
Dr. W. Brigs, Syngenta, Enkhuizen

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

The genetics of the metabolome in *Brassica rapa*

Dunia Pino Del Carpio

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. dr. M.J. Kropff,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Monday 4 October 2010
at 1.30 p.m. in the Aula.

Dunia Pino Del Carpio,
The genetics of the metabolome in *Brassica rapa*,
169 pages

Thesis, Wageningen University, Wageningen, NL (2009)
With references, with summaries in Dutch and English

ISBN 978-90-8585-721-1

Contents

Chapter 1

Introduction

1. Origin and history of Brassicas

1.1 Brassica taxonomy

The Brassicaceae is a large plant family and comprises 338 genera and 3,700 species of major scientific and economic importance. The taxonomy of this group is difficult, which is partly due to convergent evolution in nearly every morphological feature used to define tribes and genera. Three family wide molecular phylogenetic analyses lead to the proposal of a new classification scheme to organize genera into tribes which greatly improved the understanding of the evolutionary relationships in the Brassicaceae, compared to earlier attempts that were mainly based on traits like fruit morphological characteristics (Al-Shebaz et al. 2006; Bailey et al. 2006; Beilstein et al. 2006). Most of the 338 genera were placed into 25 tribes and support the lineages I and II as firstly defined by Beilstein et al. (2006). Lineage I contains the Camelineae (to which Arabidopsis belongs) and lineage II contains the agronomically important Brassiceae tribe. The tribe Brassiceae is a monophyletic group and comprises around 240 species and 49-54 genera. This tribe includes the economically important Brassica crops and radish (*Raphanus*).

Brassicas are an important vegetable species, which provide a large proportion of the daily food intake in many regions of the world. Among the Brassica species, *B. rapa* (AA,2n=10), *B. nigra* (BB,2n=16) and *B. oleracea* (CC,2n=18) are diploid and *B. juncea* (AABB,2n=36), *B. napus* (AACC,2n=38) and *B. carinata* (BBCC, 2n=34) are amphidiploids, which result from a combination of the three different diploid genomes.

The morphological variation present within Brassica species is enormous and is the result of local selection and breeding. This variation in appearance includes the leaves in crops like heading cabbages and the leafy types that do not form heads (pak choi, komatsuna etc.), the terminal and axillary buds in cauliflower, broccoli, broccoletto, the seedpods in seed stalk mustard, the swollen stems in taitsai and kohlrabi, the swollen roots in turnips and swede and the seeds in oil crops.

1.2 The origin of Brassica rapa

Brassica rapa has been cultivated for many centuries (over 4000 years) from the Mediterranean region to Scandinavia, to Germany and into central Europe, and eventually to Central Asia (Gomez-Campo 1999; Dixon 2006). Earlier studies based on morphology, geographic distribution, isozymes and molecular data indicate that *B. rapa* originates from two independent centers (Denford and Vaughan 1977; Song et al. 1988; Gomez-Campo 1999).

Europe has been proposed as the centre of origin for turnip types such as turnip rape, which were further developed in Russia, Central Asia and the Near East. Turnip is an old *B. rapa* sub-species and was probably domesticated from the wild progenitor which was transferred from the Iranian region into Europe (Reiner et al. 1995). The broccoletto's originated from Italy, and form a clearly separate group somewhat related to European turnip and oil types (Zhao et al. 2005).

Eastern Asia is proposed as the centre of origin for Asian leafy vegetables. Chinese cabbage is native to China and one hypothesis suggests that Chinese cabbage originated from hybridization between turnip (or turnip rape) and pak choi (Li 1981).

Other cultivar groups of *B. rapa* most likely originated from different morphotypes within the two origin centers and subsequently evolved separately. Japanese vegetables are likely to be derived directly or indirectly from different types of pak choi, but have diverged through geographic isolation and intensive selection (Song et al. 1988, 1990). Song et al. (1988) considered that sarson and toria types in India were derived from European turnip rape and evolved separately. In studies of Zhao et al. (2005) and Warwick et al. (2008), the spring oilseed types including the yellow sarson types from India formed a subgroup which was clearly separating from European and Asian groups, suggesting the Indian subcontinent as a third center of origin at which a separate breeding tradition led to the development of the sarson types.

1.3 *Brassica rapa* crop types

Based on the organs used for consumption and their morphological appearance, a number of major cultivar groups (Figure 1), can be distinguished in *Brassica rapa* (Specht and Diederichsen 2001; http://www.plantnames.unimelb.edu.au/Sorting/Brassica_rapa.html) with different sub-species names.

a) Chinese cabbage: *B. rapa* L. subsp. *pekinensis* (Lour.) Hanelt.

Chinese cabbage is native to China and is characterized by larger leaves and heads of different shape with winged petioles. It is mainly cultivated north of the Yangze river of China, in Korea and in Japan. In Korea, Chinese cabbage is used as the major component of "kim-chi", the traditional preserved side dish and salad. At present the Chinese cabbage is commonly found in markets throughout the world.

b) Pak choi: *B. rapa* L. subsp. *chinensis* (L.) Hanelt. Pak choi unlike the Chinese cabbage does not form a head, is characterized by green-white, enlarged midribs and it is widely cultivated in southern and central China.

c) Wutacai: *B. rapa* L. subsp. *narinosa* (L.H. Bailey) Hanelt; *Brassica* chinensis L. var. *rosularis* Tsen & Lee.

Wutacai (flat Chinese cabbage) forms a subgroup of pak choi-like cultivars that differ from typical pak choi types by their flat rosettes and many dark-green leaves. This crop is mainly cultivated in southeastern China, and is more cold tolerant and resistant to bolting.

d) Caixin (or Caitai): *B. rapa* L. var. *parachinensis* (L.H. Bailey) Hanelt.

Caixin is an early flowering non-heading vegetable with leafy features similar to pak choi. It is mainly cultivated in southern and central China, and distributed in southeastern Asian countries nowadays. The edible parts of this crop are the young inflorescences and stems that can be harvested 40-80 days after sowing.

e) Zicaitai: *B. rapa* L. var. *purpuraria* (L.H. Bailey) Kitam.

Zicaitai is characterized by the purple red stem and non-heading phenotype, and is mainly cultivated in southern and central China. This flowering purple-stemmed Chinese cabbage has tender early inflorescences, stems and shoots, which are edible. This vegetable is tolerant to low temperatures and the purple color intensifies as the temperature decreases.

f) Taicai (or Tai tsai): *B. rapa* L. ssp. *chinensis* Makino var. tai-tsai Hort.

Taicai's are non-heading cabbage cultivars with irregularly notched leaves of different blade shapes. The tender leaves, stems, and even the conical-shaped succulent taproots are edible. These types are mainly distributed throughout eastern China and are widely cultivated in the Shandong and Jiangsu provinces.

g) Mizuna and mibuna: *B. rapa* L. subsp. *nipposinica* (L.H. Bailey) Hanelt; *B. rapa* L. var. *Japonica.*

Mizuna and mibuna types are a small group of Japanese leafy vegetables with numerous serrated leaves or long narrow leaves. This crop is mainly cultivated and consumed in Japan.

h) Komatsuna: *B. rapa* L. subsp. *perviridis* Bailey.

Komatsuna is a type of neep greens, which is consumed for its young leaves, stalks and flower shoots. It is mainly grown in Japan and is also known as Japanese mustard spinach. It is most often grown in the spring and autumn, as it cannot endure extreme heat or cold for more than a short time.

i) Turnip: *B. rapa* L. em. Metzg. subsp. *rapa.*.

The turnip types are a group of cultivars grown for their enlarged hypocotyl and taproot, which can be subdivided in vegetable and fodder turnips. Manifold shapes and colors are typical characteristics of turnips. The turnip preferably grows in misty and cold regions.

j) Broccoletto, Broccoli raab, Cima di rapa: *Broccoletto* group; *B. ruvo* L.H. Bailey.

Broccoletto is a main Italian group of vegetable *B. rapa* of which the young compact inflorescences are consumed. Broccoletto has a strong stem and short internode length. The edible parts of this type are the small flower heads that appear when the plants are about 20 cm tall.

k) Turnip rape: *B. rapa* L. subsp. *oleifera* (DC.) Metzg.

Turnip rape is an oil type of *B. rapa*, which is mainly cultivated in Europe, China, India, Pakistan, Bangladesh and Canada. Summer oil types (summer or spring turnip rape) are mainly cultivated in Canada, northern Europe and Bangladesh. Dissected leaves and rosette seedlings characterize the Pakistan winter turnip rape Cultivars of winter oil type are still cultivated for oil and biomass production in Scandinavian countries, they are more cold tolerant than oilseed rape and have high growth rate under low temperature. In China, three main oleiferous *B. rapa* ecotypes, viz. spring, winter and semi-winter turnip rape, were developed in adaptation to different climates, soil conditions, cultivation methods and farmer preferences (He et al. 2003).

l) Sarsons: *B. rapa* L. subsp. *dichotoma* and *trilocularis* (Roxb.) Hanelt.

Indian oleiferous *B. rapa* includes three ecotypes, viz. brown sarson (*Dichotoma*), toria (*Dichotoma*) and yellow sarson (*trilocularis*). Brown sarson has long roots, with a limited lateral spread, enabling its successful cultivation under drier conditions. Toria has similar traits to brown sarson in morphology, and it is believed that it was selected from this type (Gomez Campos 1999). Very early flowering, self-compatibility and yellow seeds characterize the yellow sarson.

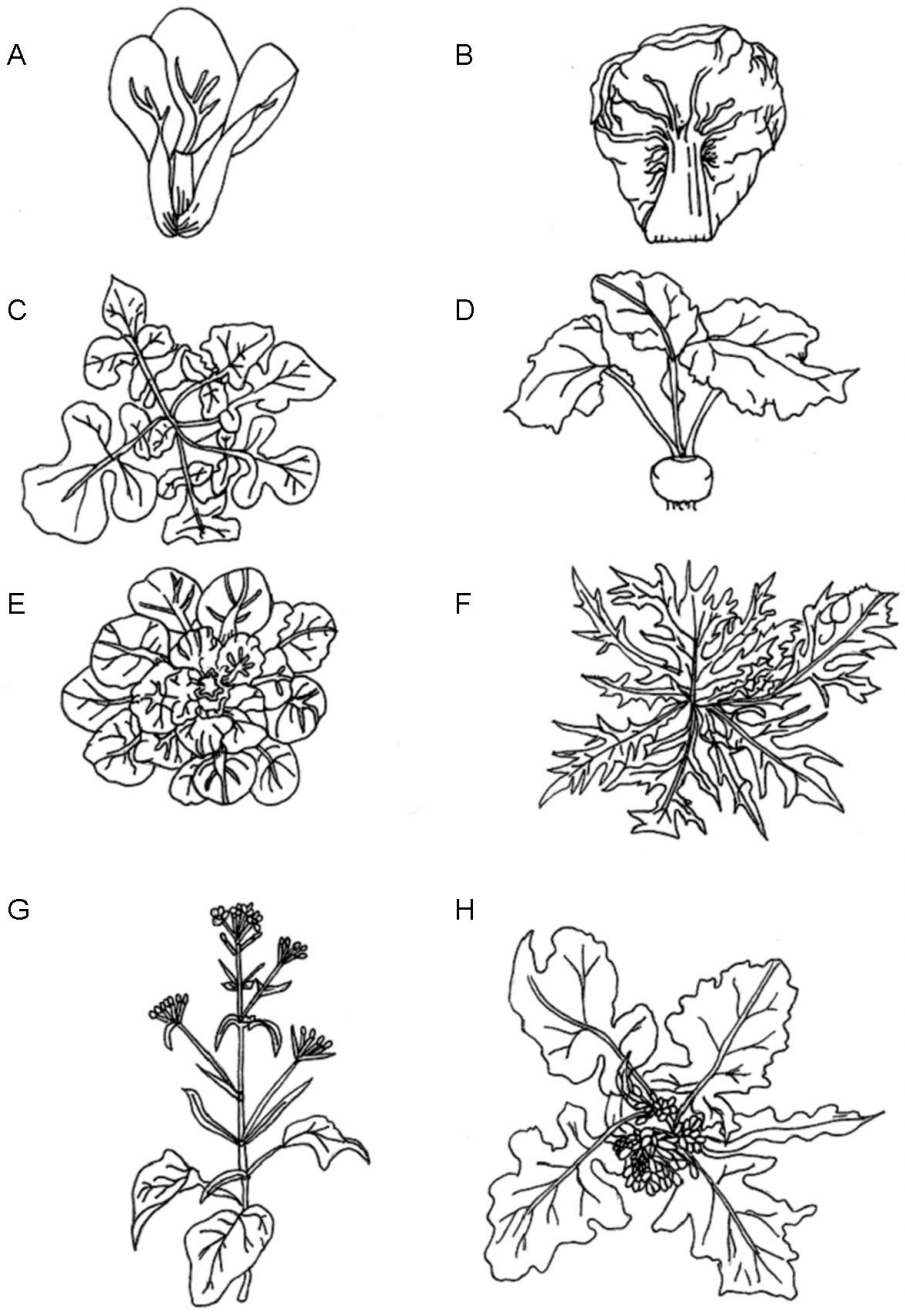


Fig 1. Morphological types of Brassica rapa. (A) Pak choi, (B) Chinese cabbage, (C) Winter oil, (D) Turnip, (E) Wucatai, (F) Mizuna, (G) oil, (H) Brocoletto.

2. Metabolite profiling of Brassicas

Metabolomics is the comprehensive analysis in which all the metabolites of an organism are identified and quantified (Fiehn et al. 2001). The components of the metabolome can be viewed as the end products of gene expression that define the biochemical phenotype of a cell or tissue. Currently, several analytical methods need to be applied in order to achieve the qualitative and quantitative analysis of all the metabolites in an organism. Among the technologies available at present for analyzing a metabolome, mass spectrometry (MS) and nuclear magnetic resonance (NMR) are considered to be the most universal approaches (Verpoorte et al. 2008). For example, liquid Chromatography-Mass Spectrometry (LC-MS) is used to detect highly rich polar or semi polar and thermo-labile positively or negatively charged compounds (Weckwerth and Morgenthal, 2005). Mass spectrometry only provides information on the mass of the detected metabolite for its indirect identification through the molecular formula (Moco et al. 2007).

Nuclear Magnetic resonance (NMR) is a powerful tool to identify wide-spectrum structural groups of complex mixtures of compounds from biological samples (Liang et al 2006, Ward et al 2003).1H NMR can detect all the proton bearing (1H) compounds including most of the non-polar, "organic" compounds such as carbohydrates, aminoacids, organic and fatty acids, amines, esters, ethers and lipids present in a sample (Ward et al. 2003)

One major limitation of metabolomics is the identification of the detected metabolites. One way to overcome this difficulty has been the use of reference compounds for well-known primary metabolites. However, for secondary metabolites, which are plant specific, many of these references are not available.

Brassica is an important food source and it is considered to have beneficial nutritional properties such as antitumoral activities (Leoni et al. 1997,Cohen et al. 2000, Podsędek 2007). The healthy components such as phenylpropanoids, phenolics, flavonoids and glucosinolates have been widely characterized in Brassica (Liang et al.2006, Onyilagha et al. 2003, Vallejo et al. 2004). For example, Brassica leaves have been found to accumulate flavonols (quercetin, kaempferol and isorhamnetin) and flavones (apigenin and luteolin) (Onyilagha et al.2003). The metabolomic analysis of various cultivars of *Brassica rapa* performed by NMR spectroscopy

showed that within the most important metabolites that contribute to the differentiation between cultivars and developmental stages in *Brassica rapa* we can find aminoacids, carbohydrates, adenine, indole acetic acid (AA), phenylpropanoids, flavonoids, and glucosinolates (Abdel-Farid et al. 2007).

For breeding purposes the diversity in total glucosinolate content and glucosinolate profile was profiled in leaves among varieties of *Brassica rapa*. Additionally their sensory attributes were evaluated in relation to glucosinolate content (Padilla et al. 2007).

Another area with extensive research is the patterns of changes in metabolite profiling in relationship to biotic stress. The metabolic alterations of *Brassica rapa* leaves attacked by larvae of the specialist and generalist insects have been investigated with nuclear magnetic resonance (NMR) spectroscopy. Within this study the major signals contributing to the discrimination between the biotic responses were alanine, threonine, glucose, sucrose, feruloyl malate, sinapoyl malate, and gluconapin (Widarto et al.2006)

3. Genetic dissection of phenotypic traits

3.1.1 The candidate gene approach

The candidate gene approach has been widely used in plant genetics to identify important genes controlling quantitative traits. This approach applies a priori knowledge of annotated genes, which have a known function and possible co-segregation with detected quantitative trait loci. When the biochemical and/or physiological pathways related to a trait of interest are well known, the candidate genes may be chosen from among the genes which have been proved by different validation methods like transformation or cloning to be involved in this pathway or phenotypic regulation. These candidate genes may play a role as regulatory or structural genes for metabolic or growth and developmental pathways (Pfliger et al. 2001).

As depicted in Figure 2, a general approach to identify genomic regions that can contain candidate genes is first to evaluate the existence of genetic and phenotypic variation in a plant system. Secondly, a mapping population can be developed from accessions that show genetic and phenotypic variation for selected traits like for example, a double haploid or a recombinant inbred population. The genotypic

variation can be visualized as molecular polymorphisms that can be used to construct a genetic map and to localize candidate genes. Statistical correlations between the genetic polymorphisms and the phenotypic variation can be revealed through QTL analysis in these mapping populations. Alternatively, association studies can also be followed in a core collection of accessions.

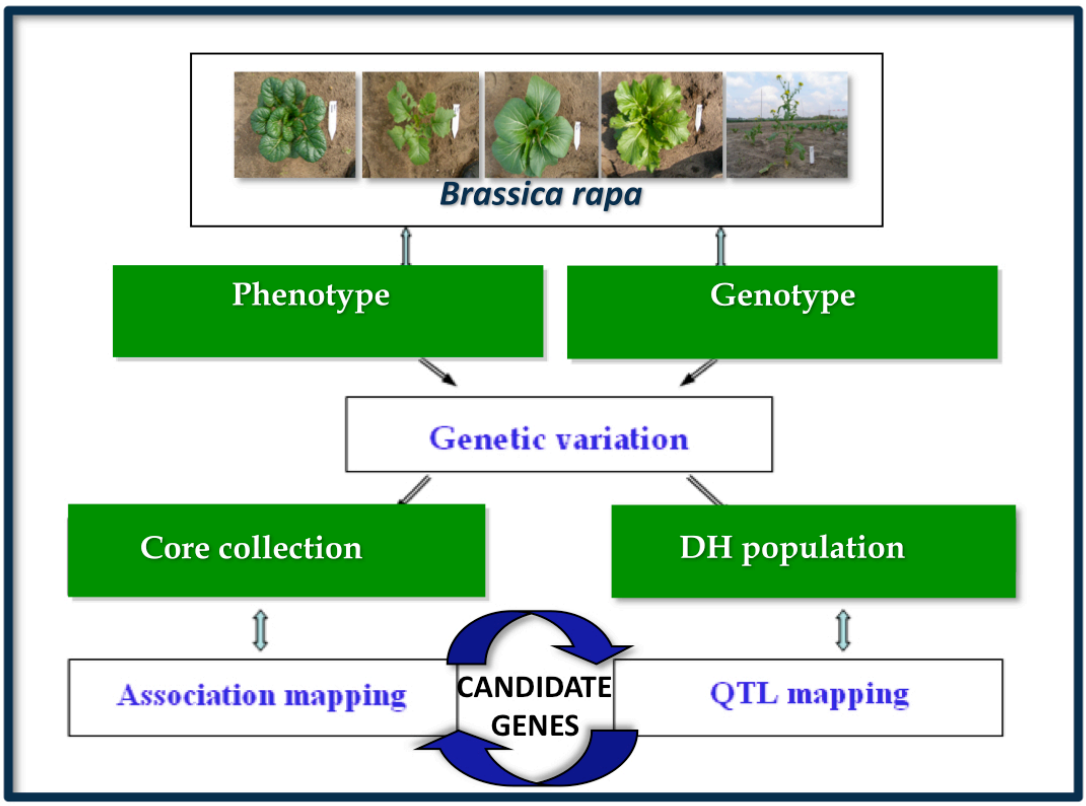


Figure 2.Candidate gene approach development for core collections and mapping populations

3.2 Comparative mapping

In *Brassica rapa* the candidate gene approach is complemented with a comparative map analysis, which takes advantage of the knowledge acquired from *Arabidopsis thaliana* in which the genome sequence and functions of many genes have been characterized.

Brassica and Arabidopsis diverged 14.5 to 20.4 MYA from a common ancestor (Blanc et al. 2003; Bowers et al. 2003) and belong to the same Brassicaceae family.

In the tetraploid *B. napus* linkage groups N01 to N10 represent the *Brassica rapa* A genome, whereas linkage groups N11-N19 represent the *Brassica oleracea* C genome (Parkin et al. 2005, Sharpe et al.1995).

In comparative genome analysis between *B. napus* and Arabidopsis 21 conserved genomic blocks were identified with RFLP probes which in most cases were triplicated within both the A and C genomes (Parkin et al. 2005). The homeologous regions to the Arabidopsis genome, which are triplicated within Brassica genomes result from an ancestral triplication event (Langercrantz, 1998). Further studies showed that although *A. thaliana* and Brassica differ by genomic rearrangements, the comparison of collinear regions identified between *B. napus* and *A. thaliana* (Parkin et al. 2005) resulted in the identification of 24 genomic blocks (A-X) (Schrantz et al. 2006) largely corresponding to the 21 blocks identified by Parkin et al. (2005). These genomic blocks represent conserved segments identified among the ancestral karyotype *A. lyrata* and the A genome of *Brassica napus* (from *B. rapa*).

For the construction of a reference linkage map for the multinational *Brassica rapa* genome sequencing project, 116 SSR markers were assigned to the linkage groups of the A genome of *B. napus*. The marker order was conserved between *B. rapa* and *B. napus* with only few observed minor rearrangements. Sequence-tagged markers were used to test also the homology between Brassica sequences and Arabidopsis. The five chromosomes of Arabidopsis were represented by homologous sequences distributed throughout the ten *B. rapa* linkage groups (Choi et al. 2006).

The consequences of the diploidization process from the hexaploid ancestor have been described for example for the FLOWERING LOCUS C (FLC) genes in *B. rapa*, which are found in Arabidopsis syntenic regions (Yang et al. 2006). In the same study, it was observed that at the microsynteny level more than 50% were not conserved within syntenic regions.

Synteny analysis of QTL regions of *B. rapa* have been of aid for example in the identification of the clubroot resistance gene (Crr3) (Saito et al. 2006). EST-based SNP markers have also been used to construct a *B. rapa* linkage map, which was compared with the Arabidopsis linkage map. In this study based on homologous markers a fine synteny relationship was revealed and QTL loci were identified for flowering time and leaf morphological characteristics. Furthermore, based on the syntenic regions candidate genes were inferred for the control of developmental traits (Li et al. 2009).

3.3 Association mapping

When studying complex traits in plants, association mapping is a tool to relate the genetic diversity expressed as allelic polymorphisms to the observed phenotypic variation. Results obtained with association mapping 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 2001, Simko et al 2004, Thornberry 2001, Wilson et al .2004, Agrama et al. 2007, Kraakman et al. 2006, Zhao et al. 2007).

For example in potato, marker-trait associations were investigated by fitting single marker regression models; this association mapping approach identified AFLP marker loci for agro-morphological and quality traits (D'Hoop et al. 2008). In a collection of 220 spring barley accessions, the association between flowering time and the variation in three genes known to play an essential role in the regulation of flowering time was found to be significant (Stracke et al. 2009). In *Gossypium arboretum*, fifty-six germplasm accessions from regions of Africa, Asia and Europe were evaluated for eight fiber characters and genotyped with 98 SSR markers. In this study the general linear model method was used to disclose marker-trait associations (Kantartzi et al. 2008).

One of the most important constraints for the use of association mapping in crop plants is the unidentified population sub structuring and admixture. Population structure is the occurrence of subpopulations in the sample in which individuals are more closely related to each other than the average pair of individuals taken at random in a population (Breseghello 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 to believe that confounding will be a significant problem, especially if the trait varies geographically, as does height, skin color or flowering time (Thornsberry 2001, Aranzana et al. 2005, Yu et al.2006).

In *Brassica rapa* association mapping has been used to unravel the genetic variation of leaf traits, flowering time and phytate content (Zhao et al. 2007). This study followed a whole genome approach with random markers, which introduced population structure into the statistical model for confounding correction. Although most of the markers did not have a map position, 3 markers were confirmed in

additional QTL studies showing the potential of this approach for identification of markers for breeding purposes.

Additionally in genetically diverse *B. napus* genotypes a candidate gene approach via structure-based allele-trait association studies was followed to identify important seed glucosinolate loci (Hasan et al.2008). In this study the Arabidopsis-Brassica comparative genome analysis proofed to be relevant for the synteny-based identification of gene-linked SSR markers.

In Brassicas several core collections have been well established and various genetic resources for association mapping are actually under development (Zhao et al unpublished results, www.**brassica.info**). In the near future the association mapping methodology will extend to the dissection of additional complex traits which combined with genomics tools will have more practical applications in Brassica breeding programs.

Scope of the thesis

The main objective of this thesis is to unravel the genetics of the *Brassica rapa* metabolome. To achieve this goal a core collection was designed in order to contain a diverse group of morphotypes with different geographic origin. This group of accessions was profiled with available molecular markers with unknown map positions (AFLP and motif targeted markers) and known map positions (microsatellites). Association studies were performed in order to take advantage of the genetic diversity and metabolite variation evaluated through metabolite profiling technology. Additionally, the results obtained in previous QTL studies are compared to the association mapping results for glucosinolate variation.

Furthermore, in order to identify candidate genes for metabolic regulation a double haploid population was profiled for metabolite content and expression abundance. The genetical genomics approach was followed for a selected group of biochemical pathways. The results will indicate how feasible this approach is in *B. rapa*, whether this could lead us faster to the identification of candidate genes and whether loci affecting metabolic traits can be identified for future marker assisted selection in breeding programs.

In **Chapter 2** we investigated the genetic diversity of a core collection of 168 *B. rapa* accessions. The number of groups and composition were defined through cluster analysis using molecular marker and morphological data. Additionally, we evaluated the use of metabolic profiling data for group classification in comparison to the genetic and morphological results.

In **Chapter 3** we evaluated the results obtained with different methodologies for association mapping studies in a group of selected targeted metabolites. Statistical analysis compared widely used methods, with and without correction for population structure, to Random Forests results.

In **Chapter 4** we followed an association mapping approach to dissect the genetics of a major QTL previously found in the A3 linkage group for glucosinolate variation. We applied a candidate gene approach with the profiling of microsatellites with known map position and physically linked to genes of the glucosinolate pathway.

In **Chapter 5** we combine information from metabolic variation and transcript abundance in a doubled haploid population developed from a cross between a yellow sarson and a pak choi. Using a genetical genomics approach we identify candidate

genes for selected biochemical pathways. Additionally, the use of a newly designed microarray based on EST sequences from different Brassicas was evaluated for this type of studies.

In **Chapter 6** all the findings are discussed and suggestions for future approaches for the dissection of traits in core collections and mapping populations are provided.

Chapter 2

The Patterns of Population Differentiation in a *Brassica rapa* Core Collection

Dunia Pino Del Carpio Ram Kumar Basnet, Ric CH.De Vos, Chris Maliepaard, Richard G.F. Visser ,Guusje Bonnema

Abstract

With the recent advances in high throughput profiling techniques the amount of genetic and phenotypic data available has increased dramatically. Although many genetic diversity studies combine morphological and genetic data, metabolite profiling has yet to be integrated into these studies. For our study we selected 168 accessions representing the different morphotypes and geographic origins of *Brassica rapa*. Metabolite profiling was performed on all plants of this collection in the youngest expanded leaves, five weeks after transplanting and the same material was used for molecular marker profiling. During the same season a year later, twenty-six morphological characteristics were measured on plants that had been vernalized in the seedling stage.The number of groups and composition following a hierarchical clustering with molecular markers was highly correlated to the groups based on morphological traits (r=0.420) and metabolic profiles (r=0.476). To reveal the admixture levels in *B. rapa*, comparison with the results of the programme STRUCTURE was needed to obtain information on population substructure.To analyze 5546 metabolite (LCMS) signals the groups identified with STRUCTURE were used for Random Forests classification. When comparing the random forest and STRUCTURE membership probabilities 86% of the accessions were allocated into the same subgroup.Our findings indicate that if extensive phenotypic data (metabolites) is available classification based on this type of data is very comparable to genetic classification. These multivariate types of data and methodological approaches are valuable for the selection of accessions to study the genetics of selected traits and for genetic improvement programs, and additionally provide information on the evolution of the different morphotypes in *B. rapa.*

Introduction

In order to assess the levels and patterns of genetic diversity in a core collection of accessions one of the most important considerations is the choice of informative datasets. The ultimate goal is to examine simultaneously loci or traits to obtain a genome-wide pattern of genetic variation and to be able to classify the accessions. Morphological traits and molecular marker data have been widely used for this purpose and for the selection of lines with maximum variation for plant breeding programmes (Liu et al. 2007; Hartings et al. 2008; Zhang et al. 2008 ; Smykal et al. 2008). . Recently, high throughput metabolomics data have emerged as a valuable resource to measure phenotypic variation in natural populations on a large scale (Keurentjes et al. 2006). Because of its rapid development metabolomics are expected to be adopted as an essential component in plant breeding programs (Fernie et al. 2008; Verpoorte et al 2008). Furthermore, because metabolites are the result of the interaction of genes from several pathways, the integration with molecular marker and morphological trait data could increase our understanding of natural variation and can facilitate the identification of valuable genetic resources.

Unravelling the complexity of large datasets presents a challenge in the selection of appropriate multivariate statistical approaches for the identification of subgroups within a core collection. This issue is particularly important in association studies where the relatedness among accessions is complex and the prevalence of a trait of interest in one subpopulation compared to the other subpopulations is known to be an important constraint (Yu et al. 2006; Aranzana et al. 2005).

Brassica rapa is one of the most important members of the Brassica genus and has been cultivated for many centuries across Europe expanding eventually to Central and East Asia (Gomez-Campo 1999; Dixon 2006). Because in *B. rapa* subgroups representing the different morphotypes have arisen as a result of selection by plant breeders and adaptation to different geographic regions its phenotypic diversity can be related to population structure (Zhao et al. 2007) .

In addition to the assessment of the diversity based on morphological characteristics and molecular markers Brassica morphotypes have been extensively screened for metabolite composition using GC-MS, LC-MS and/or NMR techniques (Abdel-Farid et al. 2007; Chen et al. 2008; Liang et al. 2006; Padilla et al. 2007; Rochfort et al. 2006; Romani et al. 2006). The most frequently studied compounds are glucosinolates, flavonoids, carotenoids, tocopherols and phenylpropanoids. Generally, these studies report an

overview of compounds found in different organs or under different cultivation methods. However, in these studies the phytonutrient composition is neither included in a context of correlation to morphological characteristics nor related to molecular genetic diversity and it is restricted to the screening of a limited number of accessions within morphotypes and to known reference compounds.

In the present study a *Brassica rapa* core collection of 168 accessions, representing different types and origins, was screened and classified using a genetic approach with two types of molecular markers, (AFLP and SSRs), a comprehensive metabolomics approach using LCMS-based untargeted profiling (De Vos et al. 2007), and a phenotypic approach based on 26 different morphological characteristics of vernalized plants.

Population structure was calculated with data of AFLP markers (random, dominant) (Vos et al. 1995) and a set of non-linked multi-allelic SSRs (EST-based and genomic, co-dominant). Accessions were assigned to subgroups in STRUCTURE and the amount of genetic differentiation between and within populations was calculated using F-statistics (Wright 1951; Cockerham 1969,1973). In addition for the analysis of metabolite data we used GeneSrF, a web-based tool and R package to analyze the 5546 individual LC-MS signals from the 168 accessions that implements individual LCMS signals selection and classification using Random Forests (Breiman 2001,Diaz-Uriarte et al. 2006,2007; Lunetta et al. 2004).

The large and diverse types of datasets obtained in this study allowed us to (1) Use clustering approaches to reveal the diversity of the sample set based on morphological characteristics and metabolite data in correlation to population structure subgroups (2) Reveal the relationship between *B. rapa* accessions and evolution of *B. rapa* types and forms based on genetic diversity data and F-statistics, and (3) Make use of a large data set of LC-MS untargeted metabolite profiling to classify and sub select a small number of metabolites specific for subgroups from the core collection using Random Forests.

Materials and Methods

Selection of plant materials and experimental design

The *Brassica rapa* core collection included a total of 168 accessions representing the different morphotypes and geographic origin (Table S1). The core collection included 132 accessions that were part of the study of Zhao et al. (2005). From the 168 accessions, 137 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, Beijing) and the Oil Crop Research Institute (OCRI, China)- and the Osborn Lab (Madison, Wisconsin USA), while Dutch breeding companies provided 31 accessions (hybrid varieties and breeding lines). Firstly, uniformity in growth and appearance was checked in the field. Ten plants per accession were distributed over two blocks in groups of five plants and morphological traits were measured for comparison of plants within accessions (data not shown).

For the metabolite profiling two plants per accession were sown in the greenhouse in September 2006 and September 2007 under the following conditions: 16 hrs light and temperature between 18 and 21C°. 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 fully expanded leaves) was harvested from one plant per accession from one table and directly frozen in liquid nitrogen, ground and stored at -70 C°.

DNA was extracted from the ground and frozen material, from the same plant selected for metabolite profiling, with the DNAeasy kit (Qiagen, USA).

Morphological data collection

For the assessment of morphological traits, 10 seeds per accession were sown on filter paper on a Petri dish. After germination, the plates with seedlings were transferred to a dark cold room (5C°) for 4 weeks. After vernalization treatment, four germinated seeds were transferred to pots in the greenhouse and distributed over four tables, with one plant per table in a completely randomized design.

Twenty-six morphological characteristics were measured from each plant per table at the time when the first flower opened, these included leaf, flower and plant architecture traits.For further calculations the final values were averaged over all the observations (Table 2).Photos of the different plant organs, one flower and one fully developed leaf (third leaf) from 4 plants were analysed using ImageJ software **(http://rsb.info.nih.gov/ij/).**The different morphological data values (continuous and categorical variables) were averaged over the four plants per accession and autoscaled within each variable using the formula z= x-mean/sd (x: variable to be standardized and sd: standard deviation). Data analysis of the autoscaled data, correlations between morphological variables and UPGMA hierarchical clustering of the accessions was performed using Genemaths XT (Applied Maths, Belgium). The dissimilarity matrix was calculated based on Euclidean distances between the morphological variables.

Table 2. Descriptors of morphological traits

A. Leaf traits			
	Leaf length	LL	Length from base of petiole to tip of lamina (cm)
	Lamina blade length	Lbl	Distance from the tip lamina to the first lobe (cm)
	Lamina width	LW	Lamina width at the widest point (cm)
	Leaf index	LI	Ratio of Lbl/LW
	Leaf area	LA	The whole surface of full leaf including lobes (cm2)
	Leaf perimeter	LP	The edge of full leaf (cm)
	Petiole length	LPL	Distance from the base of the petiole to button of lamina (cm)
	Leaf lobes	LB	Number of lobs on the leaf
	Leaf color	LC	Visual score (1= dark, 2= high green, 3= medium green, 4= light green, 5= green-yellow, 6= yellow)
	Leaf edge shape	LES	Score (1= Entire, 2= Slightly serrated, 3= Intermediate serrated, 4= Very serrated)
	Presence of petiole	LPP	Score (0= absent, 1= present)
	SPAD	SPAD	Chlorophyll content
B. Flower traits			
	Corolla length	CL	Symmetric length between petals (mm)
	Corolla width	CW	Symmetric width between petals (mm)
	Petal length	pL	Distance from base to the top of the petal (mm)
	Petal width	pW	Petal width at the widest point (mm)
	Petal index	pI	Ratio of pL/pW
	Petal area	pA	The whole surface of petal (mm2)
	Petal perimeter	pP	The edge of petal (mm)
	Petal shape	pS	Scored (1=round, 2= oval, 3= elongate)
	Petal color	PC	Visual screening of petal color (1=orange, 2= high yellow, 3=Yellow, 4= medium yellow, 5=light yellow)
	Flowering in time	DTF	Number of days from transplant till the appearance of the first open flower (days)
C. Plant Architecture trait			
	Leaf number	LN	Number of the leaves when the first flower opens
	Plant branch	PB	Number of the branches at flowering time
	Plant height	PH	Distance from the cotyledons to the top of the plant at pre-mature stage (cm)
	Plant final height	PfH	Distance from the cotyledons to the top of the plant at mature stage (cm)

LC-MS metabolic profiling

Brassica rapa leaf samples were analyzed for variation in semi-polar metabolite composition using LC-QTOF MS, essentially as described in De Vos et al. (2007). In short, 0.5 g FW of frozen leaf powder, from an individual plant per accession, was weighed in 10 ml glass tubes and extracted with 1.5 ml of methanol containing 0.1% formic acid. Samples were sonicated and then filtered (Captiva 0.45 µm PTFE filter plate, Ansys Technologies) into 96-well plates with 700µl glass inserts (Waters) using a

TECAN Genesis Workstation equipped with a 4-channel pipetting robot and a TeVacS 96-wells filtration unit. Samples were injected (5 µl) using an Alliance 2795 HT instrument (Waters), separated on a Phenomenex Luna C18 (2) column (2.0x 150 mm, 3 mm particle size) using a 5-35% acetonitrile gradient in water (acidified with 0.1% formic acid) and then detected on-line firstly by a Waters photodiode array detector (wavelength 220-600nm (Waters) and secondly by a Water-Micromass QTOF Ultima MS with negative electrospray ionization (m/z 80-1500).

Metalign software (www.metalign.nl) was used to automatically extract and align all relevant mass signals (signal to local noise ratio > 3) from the raw data files. The total of 46,788 signals was filtered for signals present in at least 10 samples and having amplitudes of at least 200 (about 8 times the noise value) in at least one of the samples. Then, all signals eluting within 3 min of retention time (i.e. the injection peak, mostly consisting of signals from non-retained highly polar compounds) were removed from the dataset.

A total of 5,546 LCMS peaks defined by mass and scan number from 168 accessions were included in the subsequent data analysis. Hierarchical cluster analysis of the accessions was done in GeneMaths XT (Applied Maths, Belgium). The similarity matrix was calculated on the log2 transformed 5,546 LCMS data with Pearson's correlation and UPGMA clustering in the program DARwin (Perrier et al. 2006) and drawn with TreeDyn (Chevenet et al. 2006).

Because it is known that many metabolites can be influenced by environmental conditions, we tested the repeatability of the metabolic profiles. For this purpose we selected homogeneous accessions, which correspond to the ones obtained from Dutch seed companies. The 17 accessions were grown in the greenhouse during the same season of two consecutive years, each year at two separate locations in the greenhouse (blocks). We thus obtained 4 biological repetitions from two consecutive years. For each replicate, leaf material from two plants was collected at the same plant age (5 weeks after sowing). Upon harvest, the leaves were immediately frozen in liquid nitrogen, ground into a fine powder in liquid nitrogen, and stored at -80°C until use. The 4 replicates of the 17 accessions had a unique profiling but were simultaneously extracted and analyzed for variation in metabolic composition, using the untargeted LC-MS profiling approach described above (De Vos et al. 2007). We subsequently analyzed by multiple linear regressions the correlation (R^2) between metabolite signals within each year (by comparing the different blocks) and between years.

Genotypic datasets

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 µl volume of 1x PCR buffer, 0.2mM dNTPs, 30ng of adaptor primer, 0.4 Taq polymerase and 5 µl of a 10x diluted restriction ligation mix, using 24 cycles of 94° C for 30s, 56° C for 30 s and 72° C for 60s. Pre-amplifications products were used as template for selective amplification with three primers combinations (P23M48, P23M50 and P21M47).

For The MYB 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 µl of 1X PCR buffer (with 15Mm MgCl2), 0.2 mM dNTPs, 0.8 pMol Gene specific primer, 0.8 pMol Adapter primer, U Hotstar Taq polymerase (Qiagen) and 5 µl of a 10X diluted restriction ligation mix. Amplification products were used as template for selective amplification.

For microsatellite (SSR) screening twenty-eight primers were selected for amplification in the core collection accessions. From the primers 10 were genomic and 18 were new EST- SSRs (Dr. Marongcai, Dr. Jifeng Tang, which institute, place personal communication). The primers were selected because of their map position in different maps of *Brassica rapa* and distribution over all the linkage groups (A1-A10).

AFLP and MYB profiling images were analyzed using Quantar PRO software (Keygene, The Netherlands); marker data were scored as present (1) or absent (0) and treated as dominant markers. Microsatellites scores were converted to binary data per observed allele (fragment of defined size) as present (1) or absent (0) and were also treated as dominant markers.

The genetic distance values were calculated using Jaccard's coefficient for 412 polymorphic AFLP and SSR fragments.

Marker data was used for cluster analysis using the unweighted pair group method with arithmetic averages (UPGMA) clustering in MEGA 4.0 (Kumar et al. 2008).

Assessment of genetic diversity

The genetic diversity was assessed in 166 accessions (excluding accession numbers 164 and 165 because of large numbers of missing values) with 23 SSR primer pairs, which represented loci from different linkage groups. These twenty-three SSRs were subselected from 28 SSRs scored over the core collection not to overrepresent any of the linkage groups in the analysis (Table 1).

The SPAGeDi 1.2 g program (from whom and where) was used to calculate Dst, the average gene diversity between subpopulations, the allele frequency per locus and the genetic diversity corrected for sample size (H_e) [Hardy et al. 2002; Nei 1978). A hierarchical analysis of molecular variance (AMOVA) was performed on the SSR data with Arlequin 3.11 software (Excoffier et al. 2005). Data conversion was done in SPAGeDi and Genepop web application (http://genepop.curtin.edu.au/). F statistics (Fst) values were computed according to Weir and Cockerham 1984, to quantify the extent of between-within population differentiation. (FCT: between populations, FSC: within population).

Fst values equal 0 when subpopulations are identical in allele frequencies and 1 when they have different alleles. Populations with little divergence have Fst values less than 0.05, moderately differentiated populations have values between 0.05 and 0.15, greatly differentiated populations have values between 0.15 and 0.25 and very greatly differentiated populations have values greater than 0.25 (Hartl et al. 1997; Mohammadi et al. 2003).

Locus#	name	LG	pop 1	pop 2	pop 3	pop 4	mean	s.d	number	H_e	Forward primer	Reverse primer
1	Br46	R1	3	2	2	2	2.25	0.5	3	0.4145	AGGTTTCGAG GTTTGTGGCT TCT	CTAAACTCATCGCTT CCGTAAACA
2	br333	R1	5	4	3	3	3.75	0.957	6	0.5833	AGTTGGCCCC ATTTCATTGTT AT	CATCTTGACGGCCTC CATCTCCA
3	KS50420	R10	17	15	8	11	12.75	4.031	21	0.9112	TTCACACAAG GTTTGTGCC	CGTAAAGGCATCAAG GAAAA
4	Br27	R2	5	4	4	4	4.25	0.5	5	0.526	AAGTACATGG TCATCCAAGG	AAGGATCCATCACAT GGTAA
5	Br48	R2	4	5	4	3	4	0.816	5	0.6257	GGTGGTGGGC TGGGGAGTA	CGTCGATCGATTCAT AACCGTAGA
6	br323	R2	8	4	4	6	5.5	1.915	8	0.7442	GTGGTGAACG TGCTTAAGAT	ACGAGCTGGTTGAAA GTTTA
7	F3H-SSR2	R3	4	4	3	4	3.75	0.5	4	0.7432	GTCATCTCCA GGTAAATCCA	TCTTGAACAACCTCT CCCTA
8	fito63	R3	6	5	4	6	5.25	0.957	9	0.6985	GTTCAGTTCC CAGATTCCTA A	TTTCCTCTTCCTTCTC TCTTC
9	br356	R3	3	3	3	2	2.75	0.5	4	0.4978	GCATCTCAGC CTTACAACTT	AGCAAGAACCCAGAA ACATA
10	br377	R4	4	2	2	2	2.5	1	4	0.352	GAAATGAGCG ACAGTGTGAT	ACAAACGACCAGTTC ATAGG
11	Br65	R4	3	4	2	3	3	0.816	4	0.3567	TTCCGTCCCTT CCCTAAACAA	TGAACACTACTGCCC AGAGAACAC
12	Na10D09	R4	8	7	5	4	6	1.826	8	0.7452	Lowe et al (2003)	
13	br384	R4	6	6	5	4	5.25	0.957	6	0.7669	TTCAATCACT TCTTCGTTTG	GAAGTAGCAGAAACA GCACC
14	brms34	R5	8	8	6	7	7.25	0.957	9	0.7947	GATCAAATAA CGAACGGAGA GA	GAGCCAAGAAAGGA CCTAAGAT
15	br378	R5	8	5	5	3	5.25	2.062	8	0.6027	TTCATCCATC CATCTTTCTC	ATGATTCCTCCATGTT CATC
16	Br51	R6	3	3	2	3	2.75	0.5	3	0.4109	CCGAGGAAGA AAGCTGTTGA GTTG	ATCGCTTCCGTAGAC ACCTTCGTT
17	Na12h07	R6	6	5	4	3	4.5	1.291	6	0.5812	Lowe et al (2003)	
18	Br89	R6	6	6	4	6	5.5	1	9	0.6583	CGTCCGTAGC GCTATTTTTCA GA3	ACGTTGTCGATCGCC CAGTTC
19	BR372-WU	R7	2	2	2	2	2	0	2	0.4491	AACGTAGTCA CCAACGAAAC	TCTGAGAAAAGAAGG AGCTG
20	brms36	R7	6	5	4	4	4.75	0.957	7	0.7416	Suwabe et al (2002)	
21	Ra2A01	R7	8	16	7	10	10.25	4.031	17	0.8576	Lowe et al (2003)	
22	br319	R8	4	3	3	3	3.25	0.5	4	0.671	TCTATGATCA TGGCTTCCTC	TCTCCGGTGTAGAGT TTGTT
23	br321	R9	3	3	2	2	2.5	0.577	3	0.3669	CCTTATCCCA TCTCTCCTCT	GAGATCAAAGTCGTA GTGGC
24	Br360*	R3									CATCGTCGTC TCCAATACTA	GAGTTGAGATCGTTC CTCTG
25	Br63*	R3									TTCCGTCCCTT CCCTAAACA	GAACACTACTGCCCA GAGAACAC
26	brms43*	R3									Suwabe et al (2002)	
27	ol11b05*	R3									Lowe et al (2003)	
28	brms50*	R3									Suwabe et al (2002)	
Mean			5.652	5.261	3.826	4.217	4.739	0.859	6.739	14.0992		
s.d			3.142	3.583	1.642	2.449	2.704	0.848	4.366	0.166630		

Table 1.Number of SSR alleles found in the core collection and per population as defined by STRUCTURE

He: gene diversity corrected for sample size (Nei 1978); LG:chromosome location; Br:prefix for an EST SSR.*not considered for Fst statistics analysis.

Assessment of population structure

Marker data (AFLP 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.
To be included in this analysis the SSR alleles were scored as dominant data making a total of 412 markers for the analysis. The number of subpopulations was determined using the software STRUCTURE 2.2 (http://pritch.bsd.uchicago.edu/software), assuming a model for *Brassica rapa* of K between 1 and 10 subpopulations, with a total of 300,000 iterations for Markov Chain Monte Carlo repetitions and burn in of 100,000.

Principal components analysis

The autoscaled data from the 26 morphological traits were used for principal components analysis (PCA). PCA is the most commonly used visualization technique in multivariate statistics, which also identifies 'Eigen'vectors and amounts of variance and cumulative explained variances per component. The PCA analysis was conducted by using the "FactoMineR" package in R-software (Husson et al. 2008).

Mantel Test

To test the correlation between the clusters calculated with the phenotypic and genotypic marker data, a Mantel test was applied in R packages ape4 and ecodist (Goslee et al. 2007; Mantel 1967).

Random Forests classification of LC-MS data

We used GeneSrF, http://genesrf.bioinfo.cnio.es (Diaz-Uriarte et al.2007) a web-based tool originally implemented for microarray data to select very small sets of genes that preserve classification accuracy. The output includes bootstrapped estimates of prediction error rate and assessment of the prediction error. Based on the allelic frequency classification, STRUCTURE assigns the 168 accessions to subgroups. In the GeneSrF web-based application this subgroup classification was inserted as the class file and the 5,546 LC-MS mass scan data were input as the equivalent of the expression data file. We considered the mean class membership probabilities obtained from the random forests output for comparison with the probabilities of inferred ancestry of individuals (membership probabilities) from AFLP and SSR markers obtained with STRUCTURE.

Results

Population structure

The genetic structure of 168 accessions was inferred using 412 markers (AFLP and SSR polymorphic bands). The Bayesian clustering implemented in the STRUCTURE software revealed 4 subpopulations. The selection of the subgroups (K=4) was done after the average likelihood value of runs for a given K value increased gradually until K=4. The selection of a K>4 would result in groups with no relationship to morphotype and/or origin of the accessions.

Population 1, includes mostly vegetable turnip (VT) and fodder turnip (FT) from European origin and broccoletto accessions: (VT+FT); population 2, includes several types: pak choi, (PC) winter oil, mizuna, mibuna, komatsuna, turnip green, oil rape and Asian turnip (T): (PC+T); population 3, includes annual oil type accessions,spring oil (SO), yellow sarson(YS) and rapid cycling (RC): (SO, YS and RC) and population 4 includes, mainly accessions of Chinese cabbage (CC) (Table S1).

Of the 168 accessions, 112 were assigned to a group with a probability value of p>0.70. Fifty-six accessions have p< 0.7 probability values with different levels of admixture between subgroups (Table S1). The morphotypes with highest levels of admixture were the winter oils accessions from Pakistan with 6 out of 7 accessions having probability values of p<0.7 The highest membership probability values of the six winter oil accessions corresponded to population 1 (VT+FT) and population 2 (PC+T).

Multivariate analyses

Morphological traits

The hierarchical cluster analysis of 26 morphological traits using the UPGMA method showed three distinct groups and few small groups of accessions (Figure 1). The first group (I) consisted of 28 accessions mainly European vegetable and fodder turnips; a second group (II) consisted of 62 accessions, including 51 Chinese cabbage accessions, landraces/cultivars and modern breeding lines from companies and 11 accessions corresponding to different types; the third (III) group consisted of 50 accessions, composed by a combination of pak choi, broccoletto, Asian turnips and few annual oil accessions**.** The turnip cluster (I) is formed mostly by accessions of European origin in correspondence to the STRUCTURE subgroup (population 1) without the broccoletto's. The Chinese cabbage cluster (II) is very similar to the STRUCTURE subgroup (population 4) and independent of cultivated form (landrace or modern breeding lines

from companies) plus several accessions from STRUCTURE population 2. The admixed cluster (III) groups accessions independent of cultivated form, similar to the STRUCTURE population 2 with several Asian types (pak choi, Chinese Cabbage, Winter oils, Asian turnips etc) plus the European broccoletto's from STRUCTURE population 1.

The results of the PCA indicated that more than 50 % of the total variance was explained by the first two PCs (31.22 % by PC1 and 23.87 % by PC2).

The most relevant loadings for the PC1 were mostly leaf characteristics and flowering time and for PC2 the most important loadings were a combination of leaf and flower characteristics (Table 3).

Correlation analysis of the 26 morphological traits showed significant and high correlations within leaf, flower and plant architecture traits, high values were also found between leaf (area and width) and flower traits (area and width) in correspondence to a possible modularity and genetic co-regulation of these developmental traits (data not shown).

Table 3. Variance and Loading values of the first three Principal Components using morphological data.

	PC 1	PC 2	PC 3
Variance	31.22	23.87	9.58
LPL	0.95	0.07	0.1
LP	0.9	0.25	0.14
LL	0.88	0.31	0.22
DTF	0.86	0.04	-0.07
LB	0.82	-0.13	0.11
LA	0.58	0.68	0.05
LES	0.46	0.09	0.22
LW	0.43	0.74	-0.05
SPAD	0.42	-0.27	-0.19
LPP	0.41	-0.46	0.013
Lbl	0.39	0.65	0.36
pl	0.06	-0.47	0.74
pS	0.06	-0.46	0.58
pL	-0.02	0.48	0.47
LI	-0.04	-0.06	0.71
pP	-0.06	0.74	0.08
pW	-0.07	0.85	-0.37
pA	-0.08	0.87	-0.14
CW	-0.11	0.76	-0.2
CL	-0.11	0.76	0.09
LC	-0.42	0.53	0.25
LN	-0.47	0.22	0.53
PfH	-0.66	0.44	0.11
PB	-0.66	0.07	0.18
pC	-0.76	-0.1	0.05
PH	-0.8	0.33	0.15

*Abbreviation description in Table 1

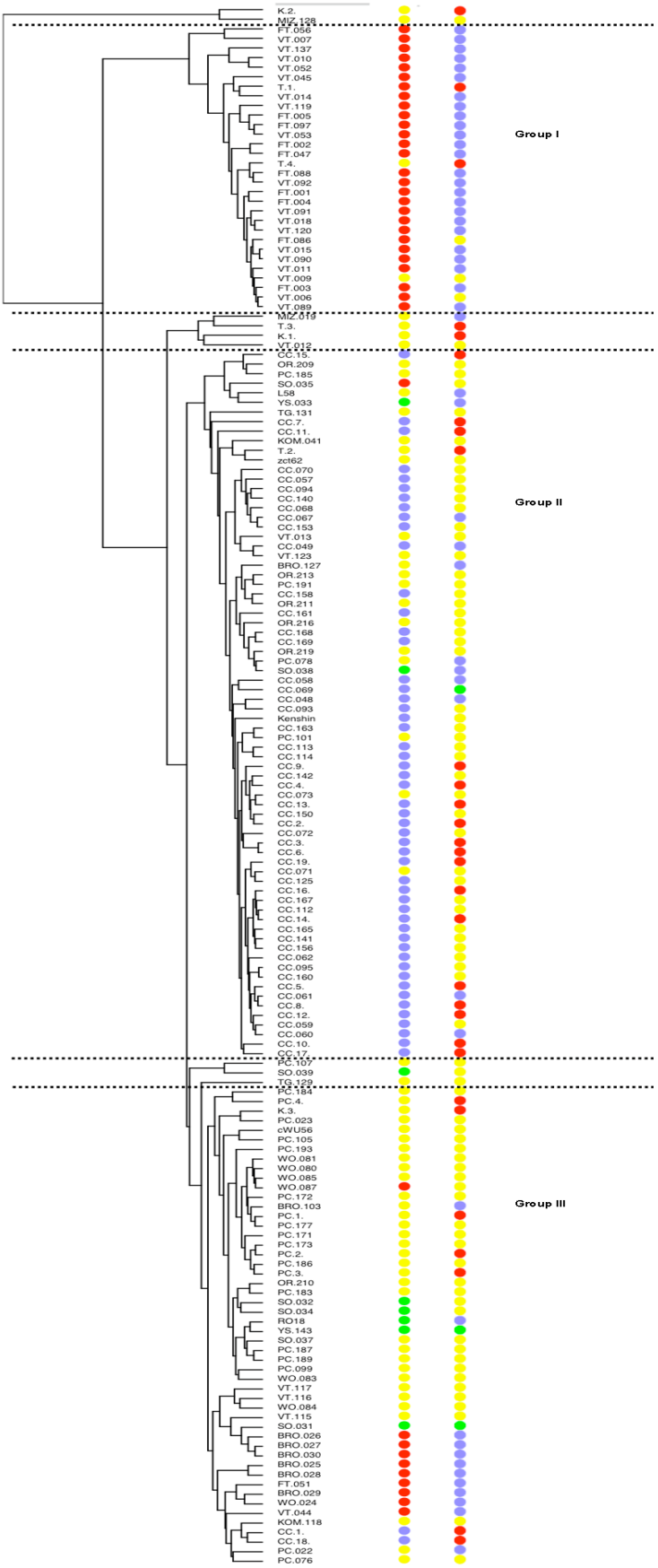


Figure 1.Hierarchical cluster UPGMA obtained with 26 morphological traits..Colours in the first column indicate STRUCTURE subgroups; red: population 1, yellow, population 2, green: population 3 and blue: population 4.Colors in the second column indicate geographic origin; red: company line, yellow: Asia, blue: Europe, green: America.

Metabolite LC-MS data

The clustering based on the log_2 transformed LC-MS data from the core collection (5,546 mass-scan numbers) showed 4 main groups and few small mixed groups (Figure 2B). Group I consists of 7 accessions, including spring oil (YS and SO) and one winter oil accession; group II includes 33 accessions of the following types: broccoletto, European vegetable and fodder turnips; group III, the most admixed group in terms of morphotypes and its relationship to the STRUCTURE subgroups, is composed of 81 accessions of the following types: pak choi, Asian turnips, turnip rape, winter oil, mizuna and few Chinese cabbages; and group IV consists of 47, mainly, Chinese cabbage accessions and modern company accessions. The composition of each group showed a high correspondence with the four subpopulations found with molecular marker information in STRUCTURE, especially in the case of groups I, II and IV and to a lesser extend for group III.

In the hierarchical clustering based on LC-MS data, 31 accessions were differently classified with respect to the subpopulations found with STRUCTURE. From these 31 accessions, 12 were also differently classified when molecular marker data were used for the hierarchical clustering (Figure 2A and 2B). These accessions have an admixed genetic nature, which allows for flexibility in terms of group assignment (Table S1). In the case of the remainder 19 accessions the genetic admixture is not significant, but still they are assigned to a different group based on this high number of metabolite mass scan signals (5,546) compared to their assignment to STRUCTURE subpopulations.

Data of 3,564 mass scan signals from the 4 biological repetitions of the 17 homogenous company lines were evaluated through a linear regression on each individual mass scan signal. Comparison of blocks from the same year, by means of linear correlation regression analysis, resulted in average r^2 values of 0.84 for year 1 and 0.86 for year 2. Comparison of results between years gave an average r^2 value of 0.79. These results indicate that the LC-MS profiles of mass scan signals obtained from each accession are highly repeatable within and between years. Based on this biological repetition experiment on 17 homogeneous lines, we can assume that the metabolite profiles observed within the *B. rapa* core collection are consistent for most metabolites over different years and biological repeats.

Molecular marker data

The clustering based on 412 AFLP and SSR markers gave as a result the separation of the collection into three groups. The three groups were composed as follows: I; a small group of oil types, II; a group of European turnips (vegetable and fodder turnip) and broccoletto and III; a group of morphotypes of Asian origin (Chinese cabbage, Pak choi and few Asian turnips). Comparison of the three groups identified with hierarchical clustering with the 4 subpopulations found in STRUCTURE indicated the subdivision of cluster/group III with the morphotypes of Asian origin into two different groups with also mainly accessions of Asian origin (II and IV) in STRUCTURE (Figure 2A). When the four groups found in STRUCTURE are compared with the oil group I, the European turnips group II and the Pak choi and Chinese cabbage group III found with the hierarchical clustering, the classification is different for 19 accessions. However, as mentioned above, these accessions also group in a different way compared to STRUCTURE when metabolite data is used for clustering and reflects the admixed nature of these 19 accessions (Figure 2A and 2B, table S1).

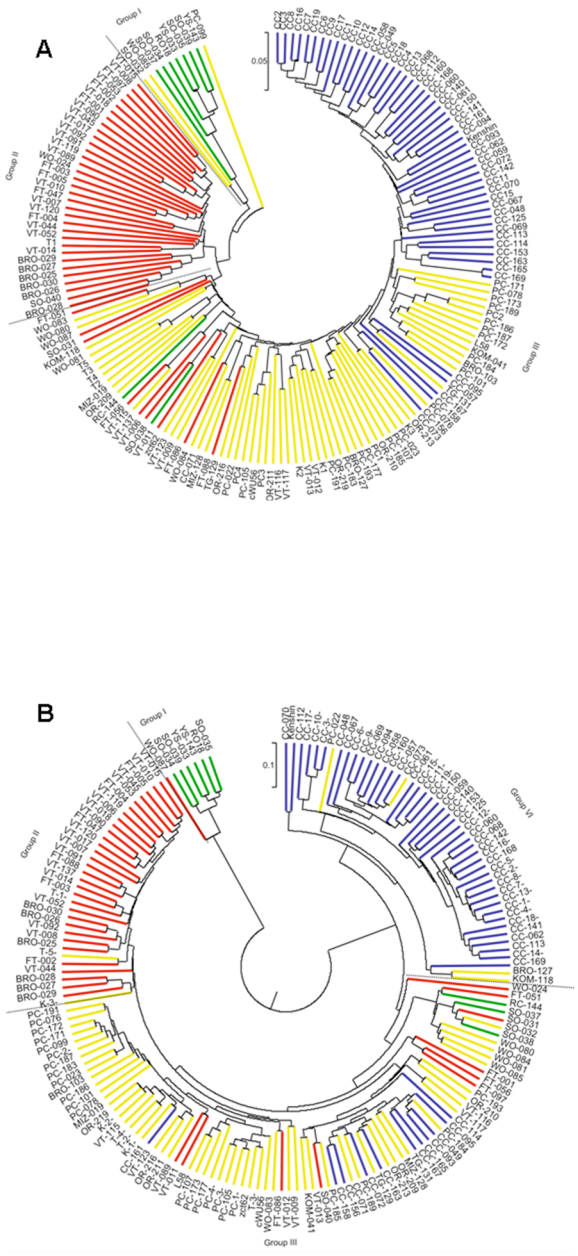


Figure 2. Hierarchical cluster UPGMA obtained with (A) molecular markers and (B) 5,546 LC-MS mass-scan signals. The colors indicate STRUCTURE subgroups red: population 1, yellow, population 2, green: population 3 and blue: population 4.

Comparison of morphological, metabolic and molecular group classification

The Mantel test between morphological and molecular distances revealed a strong and significant correlation value of r= 0.420 ($p<0.01$). In the comparison between metabolite and molecular distances the result is also strong and significant with a value of r=0.476 ($p<0.01$). The congruence between metabolite and morphological data resulted in a weak but significant correlation with a value of r=0.174 ($p<0.01$).

Population differentiation with microsatellite marker data

The molecular marker data set of 23 SSRs and 166 genotypes was used for calculation of between and within population differentiation. In the core collection the allele number ranged from 2 to 21 amplified fragments (alleles) per SSR. The mean allele number per SSR over all loci in the four sub populations is 4.7. The highest mean value for alleles was found in population 1(VT+FT) (European turnip and broccoletto group) and population 2 (PC+T) with a value of 5.6 and 5.3 respectively (Table 1).

The genetic diversity corrected for sample size (H_e) calculated over the four populations ranged from 0.91 for KS50420 (A10) to 0.352 for br377 (A04). Pairwise Ds (Nei's 1978 standard distance) between populations, as defined in STRUCTURE, indicated a high genetic distance value between population 1 (VT+FT) (European turnip and broccoletto group) and population 3 (SO, YS and RC) (oil group) of 0.4267 and a low value between population 1 (VT+FT) (European turnip and broccoletto group) and population 2 (PC+T) (pak choi and turnip group) of 0.1539 (Table 4). Global F-statistics results indicate the presence of moderately differentiated populations (Fst=0.1534), which points towards the occurrence of population structure. The Fst values of differentiation between populations ranged from 0.09 (population 1(VT+FT)-population 2(PC+T)) to 0.23 (population 3(SO, YS and RC)-population 4 (CC)) (Table 5). AMOVA results indicate that the highest percentage of variation is found within populations (84.66%) compared to among populations (15.34%). (Table 6)

Table 4.Pairwise Ds (Nei's 1978 standard distance) for all 23 loci profiled with SSR markers in the populations as defined by STRUCTURE.

	pop 1	pop 2	pop 3	pop 4
pop 1		0.1539	0.4267	0.3151
pop 2	0.1539		0.2637	0.1797
pop 3	0.4267	0.2637		0.3245
pop 4	0.3151	0.1797	0.3245	

Table 5. F statistics values of differentiation between populations as defined in STRUCTURE.

	pop 1	pop 2	pop 3	pop 4
pop 1				
pop 2	0.08961			
pop 3	0.19046	0.14982		
pop 4	0.19843	0.14113	0.23374	

Table 6. AMOVA result of Global F statistics

Source of Variation*	d.f	Sum of Squares	Variance components	Percentage of Variation
Among populations	3	251.524	1.0058	15.34
Within populations	328	1820.609	5.55064	84.66
Total	331	2072.133	6.55643	
	Fixation Index FST: 0.15341			

*Source of variation within and among populations as defined in STRUCTURE

Random Forests classification and identification of variables

Genotype information of the STRUCTURE subgroup classification was used in the random forest web based application to account for the estimated relatedness between accessions. The number of variables (metabolite: mass scan number) from the original dataset that can summarize, in a small subset, the differences between the classes (4 population subgroups) was estimated. The mean class membership probabilities based on metabolite data obtained for each accession were plotted in a similar way to the STRUCTURE membership probabilities (Figure 3). Based on the Random Forests probabilities using metabolite data, 145 out of the 168 accessions were allocated into the same class compared to the one assigned using the molecular markers with a Bayesian clustering algorithm in STRUCTURE (Table S1).

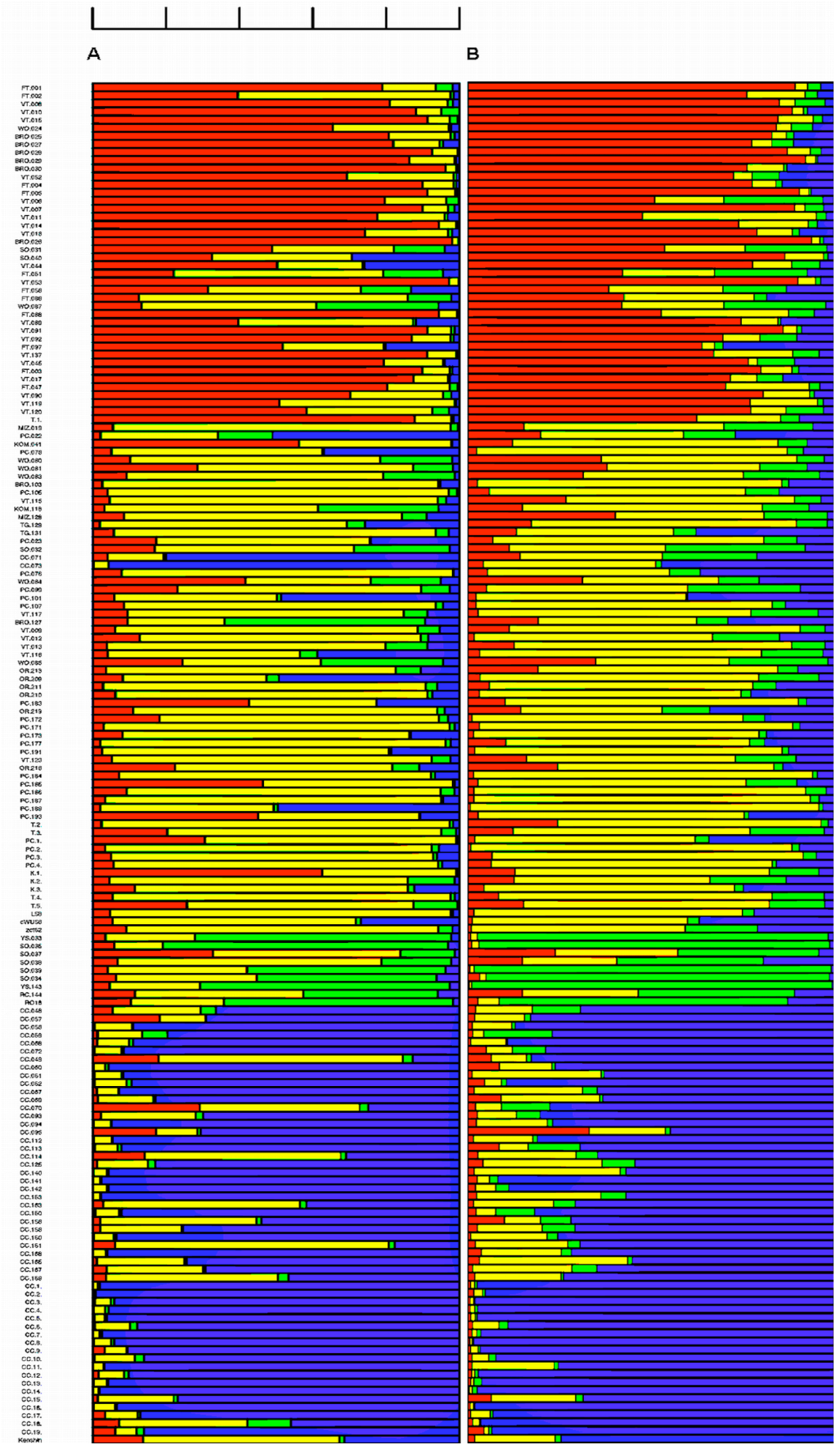


Figure 3. Membership probabilities output for each accession represented as a barplot: (A) Random Forests, (B) STRUCTURE.

In the variable selection using all data a number of 100 variables was found to be the set with a good predictive accuracy for the classes (4 population subgroups). For the variable selection the bootstrap estimate of prediction error was 13.3% (using 200 bootstrap iterations).

Correlation analysis between these 100 individual mass signals indicated that these might represent 35 different metabolites (Dr. Ric de Vos, PRI, unpublished results). Between the metabolites identified are isopropyl glucosinolate, methylpropyl glucosinolate, hexyl glucosinolate 2, caffeoylquinic acid, chlorogenic acid, coumaroylquinic acid, quercetin3-(2-feruloylsophoroside) 7-diglucoside and kaempferol coffeoyl tetraglucoside. The mass signals representing these thirty-five compounds were further used to construct a UPGMA dendrogram. Two biological replicates were included in these cluster analyses, an oil type RO18 and a caixin type L58 , which showed consistency in the clustering. In the hierarchical clustering three very well defined groups are separated in accordance to the STRUCTURE groups: Group I (VT+FT: population 1), Group 2 (CC: population 4) and group 3 (PC+T: population 2) and few admixed subgroups. The oil group (SO, YS and RC: population 3) was found to be subdivided, the summer and spring oil accessions with STRUCTURE membership probability 0.79-0.98 were grouped based on metabolites in the sub-cluster oil I, while the other summer and spring oil accessions with STRUCTURE membership probabilities 0.38-0.43 were grouped in the sub-cluster oil II (Figure S1).

Discussion

With the recent advances in high throughput profiling techniques the amount of genetic and phenotypic data has increased dramatically. Although many studies combine morphological and genetic data, metabolite profiling has yet to be integrated into diversity studies. More importantly, to establish the existence of relationships between accessions with such a multivariate approach could lead us to a better understanding of the processes that shape natural variation.

For our research we composed a *Brassica rapa* core collection with accessions widely diverse in geographical origin, morphotype and phenotypic characteristics. We consider *Brassica rapa* as suitable for genetic diversity studies because of the selection and

adaptation this crop has undergone during centuries of cultivation. Hence, the effect of these processes can be measured on the evolution of morphotypes and in the interaction of metabolite composition and developmental traits.

B. rapa is mainly an out-crosser and self incompatible, except for some annual oil types, and as a result landraces are heterogeneous. In contrast, modern cultivars and breeding lines from seed companies are homogeneous hybrids with homozygous inbred lines.

Hierarchical cluster analysis based on molecular markers identified three groups, very similar to the groups found in previous studies in *B. rapa* (Zhao et al. 2005 and 2007). Using hierarchical clustering the group number and composition identified with molecular markers is highly correlated to the groups based on morphological traits of vernalized accessions (r=0.420) and based on metabolic profiling of leaves(r=0.476). However, the number of groups separated with STRUCTURE is four, since group III (Asian crop types) defined using hierarchical clustering is represented by two structured groups (II and IV, comprising Chinese cabbages, and Asian turnips and Pak choi respectively); the resulting four subpopulations show correspondence to morphotype (turnip, oil, leafy types) and geographic origin (Asian and European). Because this output reflects the value of allele exchange (admixture) and relatedness among individuals we decided to use the structured sub-groups as a priori reference to define populations/classes for the random forests and genetic diversity estimation with microsatellites.

A constraint in the use of AFLP profiling in a core collection is that the chromosomal map position of the AFLP markers is mostly unknown and that the markers are scored dominantly. In our study we made use of SSRs with known map position, which allowed us to overcome this issue and to confirm with AMOVA the presence of moderately to strong population differentiation (Fst=0.15341). The distribution of variation in the group of accessions was found to be larger at the within population level (84.66%) compared to the variation between populations (15.34%). This suggests that subpopulations harbor enough genetic variation and could be enlarged to study morphotype specific characteristics in association studies avoiding in this way the effect of population confounding.

Although we collected data from leaf material at one developmental stage only, metabolite data proved to be very valuable in the classification of morphotypes, comparable to the genetic profiling and reflected the level of admixture found with STRUCTURE (figure 3). Random Forests provided a valuable tool to select a small group of variables from the unidentified LC-MS that represent/define these sub populations. For

future research high throughput metabolite profiling in combination with Random Forests and clustering analysis can be worthy for the identification of lines that carry interesting metabolites for crop improvement and/or for the selection of parental lines to create populations for metabolic QTL studies or to define subgroups with wide variation for association studies.

Besides the classification purposes of this study, the value of multivariate data analysis to reveal morphotypes ancestry and evolutionary processes should be acknowledged. It has been suggested that in *Brassica rapa* similar morphotypes have an independent origin and/or a long and separate domestication and breeding history in Asia and Europe (Zhao et al.2005). In the study of genetic relationships among cultivated types of *Brassica rapa* with AFLP markers it is considered that turnip was the primitive type of cultivated *B.rapa* which originated in Central Asia or in Europe and spread both to East Asia and to Europe and India (Takuno et al 2007).

Our results indicate that genetically the winter oil accessions from Pakistan show higher levels of population admixture, which indicate the presence of a genetic background shared both with European turnips and Asian pak choi types (STRUCTURE and F statistics). This ancestry of the winter oil type is reflected for example in morphological characteristics like the weedy-type look, the hairy, lobulated and rosette arrangement of the leaves, which resemble the European turnip type. If indeed winter oils are the ancient crop types that further developed into turnips and broccoletto in Europe and into turnips and pak choi in Asia then both turnip and pak choi types are a consequence of adaptation and selection on each continent. Further studies that will include sequencing of genes both affected and not affected by selection, will clarify the evolution of the *B. rapa* forms in different geographical regions.

In this study we have estimated the genetic diversity in a group of 168 accessions with different types of data. Our statistical approaches have been successful to unravel the complexity of the data and to establish relationships between accessions using genetic markers, morphological data and metabolite profiles.

The phenotypic diversity, represented by the evaluated traits (metabolite and morphology) from a large group of accessions, in this *Brassica rapa* core collection showed generally a high correlation with the genetic classification. However, hierarchical clustering of molecular data was not sufficient to reveal the level of admixture in *Brassica rapa* and comparison of clustering results with STRUCTURE is needed, especially when

association studies in *B. rapa* are the goal, since it gives additional information on the level of population substructure.

Additionally this type of multivariate type of data and methodological approach is valuable for the selection of accessions to study the genetics of selected traits and for genetic improvement programs.

Acknowledgements

We thank Noortje Bas from the Centre for Genetic Resources the Netherlands (CGN) for kindly supplying accessions used in this study and Cristina Requena Robles and Johan Bucher for their assistance in the greenhouse work and data collection. This research was funded by the IOP Genomics project "Brassica vegetable nutrigenomics" IGE 05010.

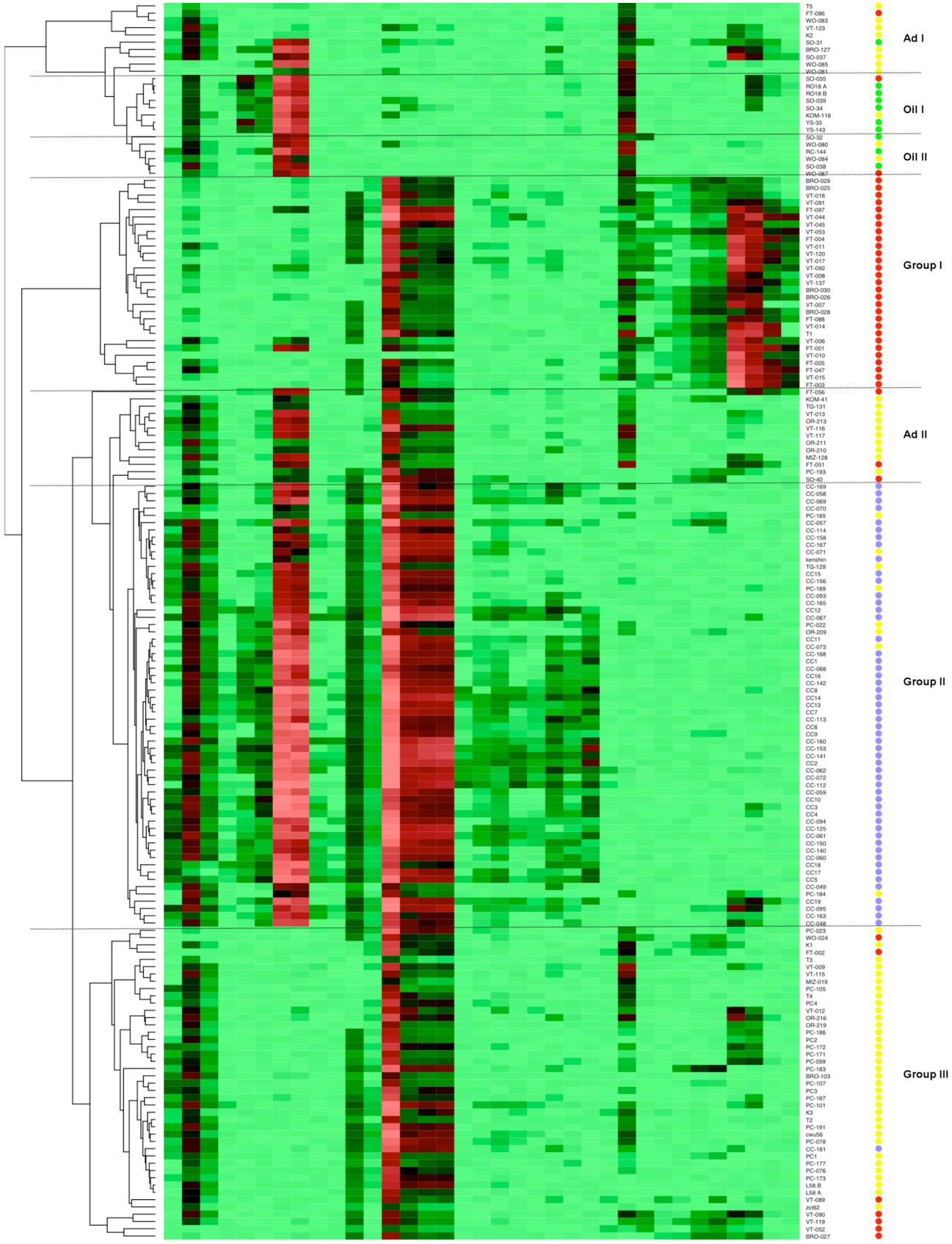


Figure S1. Hierarchical cluster analysis of 35 LC-MS mass scan signals.The signals were selected variables of the Random Forests output.

Table S1.Membership probabilities and group assignment based on metabolic and genetic data: Random Forests (RF) and STRUCTURE.Values p<0.70 are indicated with an asterisk and differences between RF and STRUCTURE are indicated in bold and italics in the group column.

Accession	origin	cultivar name		genebank	order
FT.001	Netherlands	Halflange Witte Blauwkop Ingesneden Blad-Barenza	Turnip	CGN06669	1
FT.002	United Kingdom	De Norfolk a Collet Rouge	Turnip	CGN06673	2
VT.008	Pusa Chandrina	India	Turnip	CGN06711	3
VT.010	Hungary	Platte Ronde Blauwkop Ingesneden Blad-Lila Ker	Turnip	CGN06718	4
VT.015	Italy	Bianca Lodigiana; Italiaanse Witte	Turnip	CGN06724	5
MIZ.019	Japan	Bladmoes Geslu	Mizuna	CGN06790	6
PC.022	Netherlands		Pak choi	CGN06816	7
WO.024	Sweden	Svalof 0308	Turnip rape	CGN06818	8
BRO.025	Italy	Natalino	Broccoleto	CGN06823	9
BRO.027	Italy	Quarantina	Broccoleto	CGN06825	10
BRO.028	Italy	Tardivo	Broccoleto	CGN06827	11
BRO.029	Italy	Norantino	Broccoleto	CGN06828	12
BRO.030	Italy	Sessantina	Broccoleto	CGN06829	13
YS.033	Germany	Dys 1	Yellow sarson	CGN06835	14
SO.035	Bangladesh	Somali Sarisa	Turnip rape	CGN06837	15
SO.037	Bangladesh	Kalyania	Turnip rape	CGN06839	16
SO.038	Germany	Toria	Turnip rape	CGN06840	17
SO.039	Bangladesh	Sampad	Turnip rape	CGN06841	18
KOM.041	Japan	Komatsuna	Neep greens	CGN06843	19
CC.048	Soviet Union		Chinese cabbage	CGN06867	20
VT.052	Netherlands	Hilversumse; Marteau	Turnip	CGN07166	21
CC.057	China		Chinese cabbage	CGN07182	22
CC.058	Czech Republic		Chinese cabbage	CGN07183	23
CC.059	Korea		Chinese cabbage	CGN07184	24
CC.068	Bulgaria		Chinese cabbage	CGN07196	25
CC.072	China	BRA 207/70	Chinese cabbage	CGN07201	26
PC.078	Netherlands	Choy Sam	caixin	CGN07211	27
WO.080	Pakistan		Turnip rape	CGN07216	28
WO.081	Pakistan		Turnip rape	CGN07217	29
WO.083	Pakistan		Turnip rape	CGN07220	30
BRO.103	Indonesia	Tsja Sim; No. P1R5T5	caixin	CGN15158	31
PC.105	China	BRA 77/72	Wutacai	CGN15171	32
VT.115	Japan	Kairyou Hakata	Turnip	CGN15199	33
KOM.118	Japan	Komatsuna	Neep greens	CGN15202	34
MIZ.128	Japan	Round Leaved Mibuna	Mizuna	CGN17279	35
TG.129	Japan	Vitamin Na	Neep greens	CGN17280	36
TG.131	Japan	Maruba Santo Sai	Neep greens	CGN17282	37
FT.004	Denmark	Lange Gele Bortfelder	Turnip	CGN06678	38
FT.005	Germany	Ochsenhorner	Turnip	CGN06688	39
VT.006	India	Pusa Chandrina	Turnip	CGN06709	40
VT.007	Soviet Union	Maiskaja	Turnip	CGN06710	41
VT.011	Soviet Union	Platte Witte Blauwkop Ingesneden Blad-Siniaja	Turnip	CGN06719	42
VT.014	Italy	Platte Witte Blauwkop Heelblad-Milan	Turnip	CGN06722	43
VT.018	Netherlands	Goudbal; Golden Ball	Turnip	CGN06774	44
PC.023	China	Si Yue Man	Pak choi	CGN06817	45
BRO.026	Italy		Broccoleto	CGN06824	46
SO.031	USA		Turnip rape	CGN06832	47
SO.032	India	Pusa Kalyani	Turnip rape	CGN06834	48
SO.034	Bangladesh	Australian RARS	Turnip rape	CGN06836	49
SO.040	Canada	Candle	Turnip rape	CGN06842	50
VT.044	Soviet Union	Soloveckaja	Turnip	CGN06859	51
CC.049	Netherlands	Granaat	Chinese cabbage	CGN07143	52
FT.051	Soviet Union	Krasnaja	Turnip	CGN07164	53
VT.053	Germany	Teltower Kleine	Turnip	CGN07167	54
FT.056	France	Daisy; Bladraap	Turnip	CGN07179	55
CC.060	China		Chinese cabbage	CGN07185	56
CC.061	Yugoslavia		Chinese cabbage	CGN07188	57
CC.062	Germany		Chinese cabbage	CGN07189	58
CC.067	Japan		Chinese cabbage	CGN07195	59
CC.069	USA		Chinese cabbage	CGN07198	60
CC.070	Korea	BRA 47/22	Chinese cabbage	CGN07199	61
CC.071	Japan	BRA 211/69	Chinese cabbage	CGN07200	62
CC.073	China	BRA127/67	Chinese cabbage	CGN07202	63
PC.076	China		Pak choi	CGN07205	64
WO.084	Pakistan		Turnip rape	CGN07221	65
FT.086	Pakistan		Turnip	CGN07223	66
WO.087	Pakistan		Turnip rape	CGN07226	67
FT.088	Netherlands	Blauwkop Heelblad-Oliekannetjes	Turnip	CGN10985	68
VT.089	France	D'Auvergne Hative	Turnip	CGN10995	69
VT.091	United Kingdom	Snowball; Blanc Rond de Jersey	Turnip	CGN10999	70
VT.092	Netherlands	Amerikaanse Witte Roodkop Heelbad	Turnip	CGN11000	71
CC.093	China		Chinese cabbage	CGN11002	72
CC.094	Japan		Chinese cabbage	CGN11003	73
CC.095	China		Chinese cabbage	CGN11005	74
FT.097	Germany	Buko; Bladraap	Turnip	CGN11010	75
PC.099	China	Chinese Bai Cai	Pak choi	CGN13924	76
PC.101	China	Tientsin; Celery, Shantung, Peking	Pak choi	CGN13926	77
PC.107	Hong Kong	Dwarf	Pak choi	CGN15184	78
CC.112	China	Bao Tou Qing	Chinese cabbage	CGN15194	79
CC.113	China	Bei jing 106	Chinese cabbage	CGN15195	80

Table S1.Membership probabilities and group assignment based on metabolic and genetic data: Random Forests (RF) and STRUCTURE.Values p<0.70 are indicated with an asterisk and differences between RF and STRUCTURE are indicated in bold and italics in the group column.

CC.114	China	Xiao Qing Kou	Chinese cabbage	CGN15196	81
VT.117	Japan	Toya	Turnip	CGN15201	82
CC.125	Korea		Chinese cabbage	CGN15222	83
BRO.127	Japan	Edible Flower	Broccoleto	CGN17278	84
VT.137	Uzbekistan		Turnip	CGN20735	85
CC.140	Japan	Kashin	Chinese cabbage	CGN20771	86
CC.141	Japan	Kyoto Sang	Chinese cabbage	CGN21731	87
CC.142	Japan	Matsushima Jun Sang	Chinese cabbage	CGN21732	88
VT.009	Japan	Ronde Rode-Tsutsui	Turnip	CGN06717	89
VT.012	Japan	Ronde Rode Heelblad-Yurugu Red	Turnip	CGN06720	90
VT.013	Japan	Ronde Rode Heelblad-Scarlet Ball	Turnip	CGN06721	91
VT.045	Italy	Milanskaja; Italiaanse Witte	Turnip	CGN06860	92
VT.116	Japan	Nagasaki Aka	Turnip	CGN15200	93
YS.143	USA	R500	Yellow sarson	FIL500	94
RC.144	USA	Rapid cycling	Turnip rape	FIL501	95
WO.085	Pakistan		Turnip rape	CGN07222	96
OR.213	China	Huang Po Tian You Cai	Turnip rape	OCRI0235	97
OR.209	China	Huang Gang Bai You Cai	Turnip rape	OCRI1752	98
OR.211	China	Yi Chang Xiao You Cai	Turnip rape	OCRI1771	99
OR.210	China	Lou Tian You Bai Cai	Turnip rape	OCRI1757	100
PC.183	China	Ai Kuang Qing	Pak choi	OCRI3742	101
OR.219	China	Ping Ba Bai You Cai	Turnip rape	OCRI3801	102
CC.153	China	Bao Tou Bai Cai	Chinese cabbage	VO2A0020	103
CC.163	China	Tian jing Bai Cai	Chinese cabbage	VO2A0049	104
CC.150	China	Yu Quan Bao Tou Qing	Chinese cabbage	VO2A0012	105
CC.156	China	Huang Yang Bai	Chinese cabbage	VO2A0030	106
CC.158	China	Gao Zhuang Huang Yang Bai	Chinese cabbage	VO2A0034	107
CC.160	China	Qing Kou Bai Cai	Chinese cabbage	VO2A0044	108
CC.161	China	Huang Yang Bai	Chinese cabbage	VO2A0046	109
CC.168	China	Lou Yang Da Bai Cai	Chinese cabbage	VO2A0068	110
PC.172	China	No 17 Bai Cai	Pak choi	VO2B0207	111
PC.171	China	B139 Xiao Bai Cai	Pak choi	VO2B0206	112
PC.173	China	Kui Shan Li Ye Bai Cai	Pak choi	VO2B0223	113
PC.177	China	Ai Jiao Huang	Pak choi	VO2B0396	114
PC.191	China	Wuhan Ai Jiao Huang	Pak choi	VO2B0988	115
FT.003	Netherlands	Lange Witte Roodkop	Turnip	CGN06675	116
VT.017	Netherlands	Platte Witte Meirapen	Turnip	CGN06732	117
FT.047	Soviet Union	Moskovskij	Turnip	CGN06866	118
VT.090	France	De Croissy	Turnip	CGN10996	119
VT.119	Netherlands	Roodkop-Pfalzer	Turnip	CGN15209	120
VT.120	Platte Gele Boterknol	Netherlands	Turnip	CGN15210	121
VT.123	Terauchi-Kabu	Japan	Turnip	CGN15220	122
CC.165	China	Tian jing Bai Cai	Chinese cabbage	VO2A0054	123
CC.167	China	Lou Yang Large Bai Cai	Chinese cabbage	VO2A0062	124
CC.169	China	Huang Yang Bai	Chinese cabbage	VO2A0069	125
OR.216	China	Xi Qiu Bai Cai	Turnip rape	VO2B0655	126
PC.184	China	Ai Jiao Bai	Pak choi	VO2B0656	127
PC.185	China	Qing Ken Bai Cai	Pak choi	VO2B0691	128
PC.186	China	D94 Bai Cai	Pak choi	VO2B0694	129
PC.187	China	Ai Hei Ye Kui Shan Bai Cai	Pak choi	VO2B0695	130
PC.189	China	Ai Hei Ye Kui Shan Bai Cai	Pak choi	VO2B0715	131
PC.193	China	CII	Pak choi	VO2B1263	132
T.1.	Company line		Turnip		133
T.2.	Company line		Turnip		134
T.3.	Company line		Turnip		135
PC.1.	Company line		Pak choi		136
PC.2.	Company line		Pak choi		137
PC.3.	Company line		Pak choi		138
PC.4.	Company line		Pak choi		139
K.1.	Company line		Komatsuna		140
K.2.	Company line		Komatsuna		141
K.3.	Company line		Komatsuna		142
CC.1.	Company line		Chinese cabbage		143
CC.2.	Company line		Chinese cabbage		144
CC.3.	Company line		Chinese cabbage		145
CC.4.	Company line		Chinese cabbage		146
CC.5.	Company line		Chinese cabbage		147
CC.6.	Company line		Chinese cabbage		148
CC.7.	Company line		Chinese cabbage		149
CC.8.	Company line		Chinese cabbage		150
T.4.	Company line		Turnip		151
T.5.	Company line		Turnip		152
CC.9.	Company line		Chinese cabbage		153
CC.10.	Company line		Chinese cabbage		154
CC.11.	Company line		Chinese cabbage		155
CC.12.	Company line		Chinese cabbage		156
CC.13.	Company line		Chinese cabbage		157
CC.14.	Company line		Chinese cabbage		158
CC.15.	Company line		Chinese cabbage		159
CC.16.	Company line		Chinese cabbage		160
CC.17.	Company line		Chinese cabbage		161
CC.18.	Company line		Chinese cabbage		162
CC.19.	Company line		Chinese cabbage		163
RO18	India		Yellow sarson		164
L58	United Kingdom		Pak choi		165
Kenshin	Korea		Chinese cabbage		166
cWU56	China		wutacai		167
zct62	China		zicaitai		168

Table S1.(continued)

	RF						*structure*					
order	membership probability				group	<0.70	membership probability				group	<0.70
	1	2	3	4			1	2	3	4		
1	0.79	0.15	0.05	0.02	1		0.89	0.03	0.04	0.04	1	
2	0.40	0.58	0.01	0.02	***2***	*	0.76	0.16	0.03	0.05	***1***	
3	0.81	0.16	0.01	0.02	1		0.85	0.04	0.08	0.03	1	
4	0.88	0.07	0.05	0.00	1		0.89	0.03	0.01	0.07	1	
5	0.91	0.06	0.02	0.00	1		0.85	0.10	0.03	0.03	1	
6	0.06	0.92	0.02	0.00	2		0.15	0.55	0.24	0.06	2	*
7	0.02	0.32	0.15	0.51	***4***	*	0.20	0.39	0.14	0.27	***2***	*
8	0.65	0.32	0.01	0.02	1	*	0.84	0.04	0.06	0.06	1	
9	0.81	0.17	0.01	0.01	1		0.83	0.04	0.02	0.11	1	
10	0.82	0.13	0.01	0.04	1		0.78	0.06	0.06	0.11	1	
11	0.93	0.07	0.00	0.00	1		0.87	0.06	0.03	0.04	1	
12	0.86	0.12	0.00	0.01	1		0.92	0.03	0.01	0.05	1	
13	0.96	0.03	0.01	0.00	1		0.76	0.10	0.01	0.13	1	
14	0.04	0.24	0.70	0.02	3		0.01	0.02	0.96	0.02	3	
15	0.06	0.13	0.78	0.03	3		0.01	0.02	0.96	0.01	3	
16	0.33	0.51	0.15	0.01	***2***	*	0.24	0.34	0.39	0.04	***3***	*
17	0.07	0.72	0.19	0.02	***2***		0.15	0.26	0.40	0.19	***3***	*
18	0.04	0.38	0.54	0.04	3	*	0.01	0.01	0.98	0.01	3	
19	0.56	0.42	0.00	0.02	***1***	*	0.12	0.73	0.08	0.07	***2***	
20	0.06	0.24	0.04	0.66	4	*	0.02	0.16	0.13	0.69	4	*
21	0.69	0.29	0.01	0.01	1	*	0.73	0.05	0.07	0.15	1	
22	0.18	0.12	0.00	0.69	4	*	0.02	0.14	0.02	0.83	4	
23	0.01	0.10	0.00	0.89	4		0.01	0.11	0.01	0.87	4	
24	0.01	0.12	0.07	0.80	4		0.01	0.03	0.19	0.77	4	
25	0.01	0.09	0.01	0.89	4		0.01	0.10	0.00	0.89	4	
26	0.01	0.07	0.01	0.92	4		0.05	0.07	0.09	0.79	4	
27	0.05	0.58	0.00	0.37	2	*	0.02	0.85	0.04	0.08	2	
28	0.10	0.68	0.20	0.02	2	*	0.36	0.53	0.08	0.03	2	*
29	0.29	0.59	0.11	0.02	2	*	0.38	0.42	0.13	0.07	2	*
30	0.09	0.70	0.20	0.01	2		0.32	0.44	0.21	0.04	2	*
31	0.03	0.91	0.01	0.05	2		0.03	0.83	0.02	0.12	2	
32	0.04	0.93	0.02	0.01	2		0.06	0.70	0.09	0.15	2	
33	0.05	0.89	0.02	0.04	2		0.27	0.62	0.07	0.04	2	*
34	0.03	0.58	0.33	0.06	2	*	0.15	0.66	0.13	0.07	2	*
35	0.09	0.76	0.07	0.09	2		0.40	0.41	0.17	0.02	2	*
36	0.02	0.67	0.05	0.26	2	*	0.17	0.72	0.08	0.02	2	
37	0.06	0.88	0.04	0.03	2		0.13	0.43	0.06	0.38	2	*
38	0.90	0.08	0.01	0.01	1		0.78	0.07	0.02	0.14	1	
39	0.91	0.08	0.00	0.01	1		0.91	0.03	0.01	0.06	1	
40	0.80	0.17	0.03	0.00	1		0.51	0.19	0.27	0.03	1	*
41	0.90	0.07	0.02	0.01	1		0.89	0.03	0.05	0.03	1	
42	0.78	0.18	0.01	0.03	1	*	0.48	0.47	0.03	0.02	1	*
43	0.94	0.05	0.00	0.01	1		0.74	0.19	0.03	0.05	1	
44	0.74	0.23	0.01	0.02	1		0.79	0.10	0.02	0.09	1	
45	0.17	0.58	0.00	0.24	2	*	0.07	0.74	0.09	0.10	2	
46	0.98	0.02	0.00	0.00	1		0.94	0.02	0.01	0.03	1	
47	0.49	0.33	0.14	0.04	1	*	0.54	0.22	0.22	0.02	1	*
48	0.17	0.54	0.26	0.03	2	*	0.11	0.43	0.38	0.08	2	*
49	0.06	0.38	0.49	0.06	3	*	0.03	0.02	0.94	0.01	3	
50	0.33	0.38	0.00	0.29	***2***	*	0.87	0.11	0.01	0.02	***1***	
51	0.50	0.23	0.00	0.26	1	*	0.78	0.11	0.08	0.04	1	
52	0.18	0.67	0.03	0.13	***2***	*	0.06	0.10	0.01	0.83	***4***	
53	0.22	0.57	0.16	0.05	***2***	*	0.42	0.25	0.25	0.07	***1***	*
54	0.97	0.03	0.00	0.00	1		0.90	0.05	0.01	0.04	1	
55	0.32	0.42	0.14	0.13	***2***	*	0.39	0.31	0.17	0.13	***1***	*
56	0.00	0.03	0.01	0.96	4		0.09	0.14	0.01	0.76	4	
57	0.01	0.07	0.00	0.92	4		0.01	0.35	0.01	0.63	4	*
58	0.01	0.09	0.01	0.89	4		0.05	0.05	0.01	0.90	4	
59	0.01	0.06	0.00	0.93	4		0.02	0.30	0.04	0.65	4	*
60	0.01	0.15	0.01	0.83	4		0.09	0.27	0.01	0.63	4	*
61	0.29	0.44	0.02	0.25	***2***	*	0.02	0.07	0.13	0.78	***4***	
62	0.04	0.15	0.01	0.80	***4***		0.14	0.39	0.26	0.21	***2***	*
63	0.00	0.04	0.00	0.96	***4***		0.04	0.47	0.02	0.47	***2***	*
64	0.08	0.90	0.00	0.02	2		0.05	0.50	0.08	0.37	2	*
65	0.42	0.34	0.19	0.05	***1***	*	0.31	0.37	0.11	0.20	***2***	*
66	0.13	0.73	0.12	0.02	2		0.43	0.36	0.04	0.18	1	*
67	0.13	0.48	0.33	0.06	2	*	0.42	0.28	0.28	0.02	1	*
68	0.94	0.05	0.01	0.00	1		0.53	0.35	0.07	0.05	1	*
69	0.40	0.48	0.01	0.12	***2***	*	0.75	0.10	0.01	0.15	***1***	
70	0.91	0.06	0.01	0.01	1		0.86	0.04	0.01	0.09	1	
71	0.87	0.10	0.01	0.02	1		0.70	0.10	0.10	0.10	1	
72	0.02	0.26	0.02	0.70	4		0.03	0.11	0.07	0.80	4	
73	0.00	0.05	0.00	0.95	4		0.02	0.19	0.01	0.77	4	
74	0.17	0.11	0.01	0.71	4		0.33	0.21	0.01	0.45	4	*
75	0.52	0.28	0.00	0.20	1	*	0.64	0.03	0.02	0.30	1	*
76	0.23	0.66	0.08	0.03	2	*	0.05	0.55	0.31	0.09	2	*
77	0.06	0.44	0.01	0.49	***4***	*	0.02	0.57	0.01	0.40	***2***	*
78	0.09	0.85	0.02	0.05	2		0.02	0.81	0.04	0.13	2	
79	0.00	0.05	0.00	0.94	4		0.02	0.16	0.01	0.81	4	
80	0.01	0.06	0.01	0.92	4		0.09	0.08	0.14	0.69	4	

Table S1.. (continued)

81	0.14	0.53	0.01	0.31	***2***	*	0.03	0.27	0.06	0.65	***4***	*
82	0.10	0.75	0.06	0.09	2		0.03	0.75	0.18	0.04	2	
83	0.01	0.14	0.02	0.83	4		0.04	0.33	0.09	0.54	4	*
84	0.10	0.26	0.55	0.09	***3***	*	0.19	0.43	0.09	0.29	***2***	*
85	0.91	0.08	0.01	0.00	1		0.67	0.22	0.07	0.04	1	*
86	0.00	0.04	0.00	0.96	4		0.02	0.40	0.02	0.57	4	*
87	0.00	0.02	0.00	0.98	4		0.03	0.03	0.02	0.92	4	
88	0.00	0.04	0.00	0.96	4		0.02	0.06	0.04	0.89	4	
89	0.06	0.82	0.06	0.05	2		0.11	0.79	0.07	0.03	2	
90	0.13	0.77	0.02	0.09	2		0.02	0.65	0.18	0.15	2	*
91	0.04	0.76	0.11	0.09	2		0.13	0.76	0.09	0.03	2	
92	0.79	0.16	0.01	0.04	1		0.77	0.03	0.10	0.11	1	
93	0.04	0.52	0.05	0.39	2	*	0.03	0.77	0.17	0.03	2	
94	0.05	0.25	0.68	0.03	3	*	0.00	0.01	0.98	0.01	3	
95	0.12	0.46	0.37	0.06	***2***	*	0.15	0.32	0.43	0.11	***3***	*
96	0.25	0.38	0.33	0.05	2	*	0.35	0.36	0.27	0.02	2	*
97	0.04	0.79	0.07	0.11	2		0.19	0.66	0.06	0.10	2	*
98	0.08	0.39	0.03	0.49	***4***	*	0.03	0.76	0.16	0.04	***2***	
99	0.03	0.88	0.03	0.06	2		0.06	0.72	0.06	0.16	2	
100	0.06	0.85	0.01	0.07	2		0.03	0.72	0.03	0.23	2	
101	0.43	0.35	0.00	0.23	***1***	*	0.10	0.80	0.02	0.08	***2***	
102	0.11	0.83	0.02	0.04	2		0.14	0.39	0.13	0.34	2	*
103	0.00	0.02	0.00	0.98	4		0.02	0.34	0.07	0.57	4	*
104	0.03	0.54	0.02	0.42	***2***	*	0.02	0.22	0.06	0.71	***4***	
105	0.01	0.06	0.01	0.92	4		0.02	0.05	0.11	0.82	4	
106	0.02	0.43	0.01	0.54	4	*	0.10	0.10	0.08	0.72	4	
107	0.02	0.22	0.00	0.75	4		0.03	0.18	0.09	0.71	4	
108	0.00	0.05	0.01	0.94	4		0.01	0.21	0.03	0.76	4	
109	0.06	0.75	0.01	0.18	***2***		0.02	0.25	0.03	0.69	***4***	*
110	0.00	0.03	0.00	0.96	4		0.04	0.22	0.03	0.72	4	
111	0.18	0.76	0.02	0.03	2		0.01	0.74	0.06	0.19	2	
112	0.03	0.94	0.01	0.01	2		0.03	0.86	0.02	0.09	2	
113	0.08	0.78	0.00	0.13	2		0.02	0.78	0.02	0.19	2	
114	0.02	0.94	0.01	0.02	2		0.10	0.67	0.04	0.19	2	
115	0.03	0.78	0.01	0.19	2		0.02	0.84	0.02	0.13	2	
116	0.90	0.07	0.01	0.02	1		0.80	0.09	0.02	0.10	1	
117	0.87	0.09	0.01	0.03	1		0.72	0.07	0.07	0.14	1	
118	0.80	0.17	0.02	0.01	1		0.70	0.23	0.03	0.04	1	
119	0.70	0.25	0.02	0.02	1		0.81	0.13	0.01	0.05	1	
120	0.51	0.48	0.00	0.01	1	*	0.77	0.18	0.02	0.03	1	
121	0.58	0.34	0.05	0.03	1	*	0.77	0.10	0.10	0.03	1	
122	0.05	0.87	0.05	0.03	2		0.16	0.57	0.17	0.10	2	*
123	0.01	0.24	0.01	0.74	4		0.03	0.41	0.01	0.55	4	*
124	0.04	0.26	0.01	0.69	4	*	0.01	0.27	0.07	0.65	4	*
125	0.04	0.47	0.03	0.47	***2***	*	0.02	0.24	0.01	0.74	***4***	
126	0.22	0.59	0.07	0.11	2	*	0.24	0.59	0.02	0.14	2	*
127	0.07	0.85	0.01	0.07	2		0.02	0.93	0.01	0.05	2	
128	0.46	0.52	0.01	0.01	2	*	0.03	0.73	0.14	0.10	2	
129	0.09	0.85	0.04	0.01	2		0.02	0.91	0.06	0.02	2	
130	0.03	0.92	0.01	0.04	2		0.02	0.92	0.02	0.04	2	
131	0.02	0.47	0.01	0.50	***4***	*	0.01	0.95	0.01	0.03	***2***	
132	0.45	0.44	0.00	0.11	***1***	*	0.03	0.73	0.01	0.23	***2***	
133	0.88	0.10	0.01	0.02	1		0.63	0.23	0.05	0.10	1	*
134	0.02	0.95	0.01	0.02	2		0.24	0.72	0.02	0.02	2	
135	0.20	0.75	0.04	0.01	2		0.12	0.65	0.20	0.03	2	*
136	0.31	0.68	0.00	0.01	2	*	0.02	0.60	0.04	0.34	2	*
137	0.03	0.89	0.02	0.06	2		0.01	0.88	0.02	0.09	2	
138	0.05	0.88	0.01	0.06	2		0.07	0.85	0.04	0.05	2	
139	0.06	0.88	0.01	0.05	2		0.06	0.77	0.01	0.16	2	
140	0.63	0.37	0.00	0.01	***1***	*	0.13	0.70	0.14	0.04	***2***	
141	0.05	0.81	0.13	0.01	2		0.13	0.54	0.21	0.13	2	*
142	0.11	0.74	0.02	0.12	2		0.04	0.68	0.02	0.25	2	*
143	0.00	0.01	0.01	0.98	4		0.01	0.01	0.01	0.98	4	
144	0.00	0.01	0.00	0.99	4		0.02	0.02	0.01	0.95	4	
145	0.01	0.04	0.01	0.94	4		0.01	0.01	0.00	0.98	4	
146	0.00	0.03	0.01	0.96	4		0.01	0.01	0.01	0.97	4	
147	0.01	0.03	0.00	0.96	4		0.01	0.01	0.01	0.98	4	
148	0.01	0.09	0.02	0.88	4		0.01	0.07	0.02	0.89	4	
149	0.00	0.02	0.01	0.97	4		0.01	0.01	0.01	0.97	4	
150	0.00	0.05	0.01	0.94	4		0.01	0.01	0.00	0.99	4	
151	0.06	0.92	0.01	0.01	2		0.08	0.75	0.14	0.03	2	
152	0.26	0.62	0.12	0.01	2	*	0.15	0.67	0.15	0.03	2	*
153	0.03	0.06	0.00	0.91	4		0.01	0.02	0.01	0.97	4	
154	0.01	0.11	0.03	0.86	4		0.01	0.05	0.02	0.92	4	
155	0.00	0.03	0.00	0.97	4		0.01	0.22	0.01	0.75	4	
156	0.02	0.07	0.01	0.90	4		0.01	0.01	0.01	0.97	4	
157	0.00	0.04	0.00	0.96	4		0.01	0.01	0.02	0.96	4	
158	0.00	0.02	0.00	0.98	4		0.00	0.01	0.01	0.98	4	
159	0.01	0.21	0.01	0.77	4		0.06	0.23	0.02	0.69	4	
160	0.00	0.06	0.00	0.93	4		0.01	0.01	0.00	0.98	4	
161	0.03	0.09	0.01	0.87	4		0.01	0.05	0.01	0.93	4	
162	0.07	0.35	0.12	0.46	4	*	0.01	0.02	0.00	0.97	4	
163	0.06	0.06	0.02	0.86	4		0.04	0.01	0.01	0.94	4	
164	0.11	0.25	0.62	0.02	3	*	0.03	0.06	0.79	0.13	3	
165	0.05	0.93	0.01	0.02	2		0.02	0.73	0.03	0.22	2	
166	0.14	0.53	0.01	0.31	***2***	*	0.02	0.22	0.02	0.75	***4***	
167	0.05	0.66	0.01	0.27	2	*	0.01	0.59	0.03	0.37	2	*
168	0.09	0.85	0.04	0.02	2		0.01	0.58	0.20	0.21	2	*

Chapter 3

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

Dunia Pino Del Carpio*, Ram Kumar Basnet*, Ric CH.De Vos, Chris Maliepaard, Maria João Paulo, Guusje Bonnema.
*equal contributors

Abstract

Since an association mapping approach combines the observed phenotypic variation and genetic diversity through statistical analyses, with the final goal of correlating trait levels and alleles, it is important to separate the true effect of genetic variation from other confounding factors. In crop plants, for example these factors are related to adaptation to different uses and geographical locations. An additional consideration in this type of studies is the rapid availability of large datasets, which makes it necessary to explore statistical methods that can be computationally less intensive and more flexible for data exploration.

In the present study we consider the genetic association between markers and tocopherols, carotenoids, chlorophylls and folate in a core collection of 168 *Brassica rapa* accessions of different morphotypes and origin. We followed widely used linear model association methods but in addition, we include Random Forests results for comparison. When the results across methods were compared we were able to successfully select a set of 16 significant markers. This set includes at least one marker associated per metabolite that can potentially be applied for the selection of genotypes with elevated levels of important metabolites. We showed that in this core collection of *B. rapa* confounding effects are present and that the incorporation of the STRUCTURE correction (Q matrix) in the linear regression model greatly reduces the number of significant associated markers. Additionally, our results demonstrate that Random Forests is an interesting complementary method with added value in association studies in plants.

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 a, b, Thornberry et al. 2001, Wilson et al. 2004, 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 the unidentified population sub structuring, which arises as a result of adaptation, genetic drift, domestication or selection (Thornberry 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. (Breseghello 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 to acknowledge that confounding will be a significant problem, especially if the trait varies geographically, as is the case for example for 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 populations 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/monitor for the existence of population structure in a data set and for 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, 2000b, 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 sub-

populations and kinship (also called the "QK method") is the most appropriate for association mapping (Yu et al. 2006).

In the near future the rapid development of whole genome sequencing technology will present challenges in the statistical analysis of marker-trait associations of extremely large datasets with sequence data of core collections with hundreds of individuals. Under these conditions it is necessary to consider and validate association methods that can handle such aspects as the size of the experimental population and the quality and quantity of the phenotypic data.

A very different statistical approach, which carries one or more advantages above most other methods, is the Random Forests (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, Yuanqing Ye et al. 2004). 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 i.e. metabolite profiling these issues are of particular relevance.

Furthermore, the power to detect epistasis in moderately sized populations in general is low, while Random Forests 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. 2007).

Structure correction cannot be included in association studies with Random Forests, which could be a disadvantage for its use in plant systems. However, to avoid population structure specialized populations from multiple intercrosses are being developed in maize and Arabidopsis (Stich et al. 2009); in these cases and when population structure is not present the use of Random Forests as a marker-trait association approach is suitable.

Brassica rapa is an important member of the Brassica genus and has been cultivated for many centuries across Europe expanding eventually to Central and East Asia. Subgroups like the leafy vegetables; turnips and oil types have arisen as a result of selection by plant breeders and adaptation to different geographic regions. Previously, in a collection of 160 *B. rapa* accessions association analysis with correction for population structure led to the identification of 27 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 Forests for comparison and as a complementary method to the commonly used association methods.

Materials and methods

Plant material

The *Brassica rapa* core collection included a total of 168 accessions of diverse morphotype and origin: 137 accessions were obtained from CGN, CAAS-IVF and CAAS-OCRI genebanks and the Osborn Lab while 31 accessions were provided by six different breeding companies (Supplementary Table 1). For the metabolite profiling two plants per accession were sown in the greenhouse 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 h 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 by 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 µm) 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 µl volume of 1x PCR buffer, 0.2mM dNTPs, 30ng of adaptor primer, 0.4 Taq polymerase and 5 µl of a 10x diluted restriction ligation mix, using 24 cycles of 94° C for 30s, 56° C for 30 s and 72° C for 60s. 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 µl of 1X PCR buffer (with 15Mm MgCl2), 0.2 mM dNTPs, 0.8 pMol Gene specific primer, 0.8 pMol Adapter primer, U Hotstar

Taq polymerase (Qiagen) and 5 µl of a 10X diluted restriction ligation mix. Amplification products were used as template for selective amplification.

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) (data not shown).

AFLP and *Myb* profiling images were analyzed using Quantar Pro™ software; marker data were scored as present (1) or absent (0) and treated as dominant markers. Microsatellites scores were converted to binary data per observed allele (fragment of defined size) as present (1) or absent (0) and were also treated as dominant markers.

Assessment of population structure

Marker data (AFLP, *Myb*, 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.

To be included in this analysis the SSR alleles were scored as dominant markers yielding a total of 539 markers for 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 burn in.

In addition, we also followed the procedure PCO-MC as described by Reeves et al. (2009), to assess population structure. The method uses principal coordinate analysis (PCO) and clustering methods to infer subgroups in the population. 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 et al. 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 et al. (2009). Then procedure PROC MODECLUS in SAS (SAS 1997) was used to group the accessions into clusters according to kernel density estimates in the PCO space. Clusters 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 clusters were significantly distinct from the rest, using PROC MODECLUS, and estimated stability values for the clusters using the PCO-MC software (http://lamar.colostate.edu/~reevesp/PCOMC/PCOMC.html). Both are described in Reeves et al (2009). The PCO plot of the first two components was drawn in DARwin software version 5.0.155 (Perrier et al. 2006).

Summary statistics of metabolite variation

Box plots were chosen as a tool to explore the variation of metabolite concentrations according to different STRUCTURE classes (Fig1). One-way ANOVA was performed for each metabolite to find the mean differences among the four STRUCTURE classes. The least significant differences (LSD) was calculated to compare the differences of means of metabolite content in all possible combinations of the four STRUCTURE classes. Box plots, 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 first step a "naïve" model was used to associate each marker to the trait,

trait = marker + error (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 cluster memberships Q obtained from Structure was added as a fixed term to the previous model

trait = marker + Q + error (2)

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

trait = genotype + marker + error (3)

Where random terms are underlined. Genotype is a random factor with the different genotypes or accessions in the population. Tassel calculates kinship coefficients which are used to model the covariance between the different accessions. We have $V_G=\sigma^2K$; 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 in the terminology of Yu et al. (2006).

trait = genotype + Q + marker + error (4)

As before, genotype is a random factor, with covariances given by the kinship matrix K and Q is a fixed term containing the cluster memberships. The model is similar to those described in Yu et al. (2006) and in 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 the model that included kinship and population structure matrix association analysis were corrected for multiple testing using Storey and Tibshirani's method (Storey and Tibshirani, 2003) as implemented in the R package "qvalue".

Random Forests

Random Forests (RF) regression (Breiman, 2001) was used in this study to find the associated markers among the 243 marker set to the tocopherol, carotenoids, flavonoids and folate metabolites. This method uses both a boosting and bagging approach (Gislason et al. 2006) and yields importance measures for each marker in the regression of metabolites on the multivariate marker data. 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 set, 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. In each split the samples are subdivided into two groups according to the marker that generates the 'purest' set of two groups according to the ratio of between and within group variance for the metabolite values. This procedure is fast and can handle high dimensional data ($p >> n$). 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 this 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 data set. The variance explained in RF is defined as 1-(Mean square error (MSE) / 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.

However, RF yields only the relative importance of markers that explain the variation present in metabolites, but does not give a significance threshold level to choose a possible subset of associated markers. Therefore, a permutation method was used to calculate the significance of each marker association in this study. All the observations of a metabolite (the response in the regression) were permuted 1000 times. For each metabolite, 1000 values for the increase in MSE of every marker were stored from RF regression analyses conducted for each permuted data set for a metabolite. The stored values for increase in MSE were ranked in ascending order. The ranks from 1 to 1000 for the observed incMSE values then can be used as the quantiles (0.001 to 1) of a "null distribution" for the incMSE values of each separate

marker For each marker the 0.95-quantile of the incMSE values of that marker from the permutations per metabolite was used as the threshold for significance.
RF regressions of metabolites on markers were conducted by using the "Random Forest" package of the R-software (Breiman et al. 2005).

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). In the marker-metabolite network, the vertices are the metabolites, in this case tocopherols, carotenoids, chlorophylls and folate, and associated markers. The edges correspond to metabolite-metabolite correlations and marker-metabolite correlations based on a predefined significance threshold $p<0.05$.In the present study we used correlation networks for the visualization of within and between pathway interactions through unique and shared significant markers .The network was constructed using the Pajek graph drawing software (Batagelj and Mrvar 2003)

Results

Principal coordinates analysis (PCO) and population structure of the core collection

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

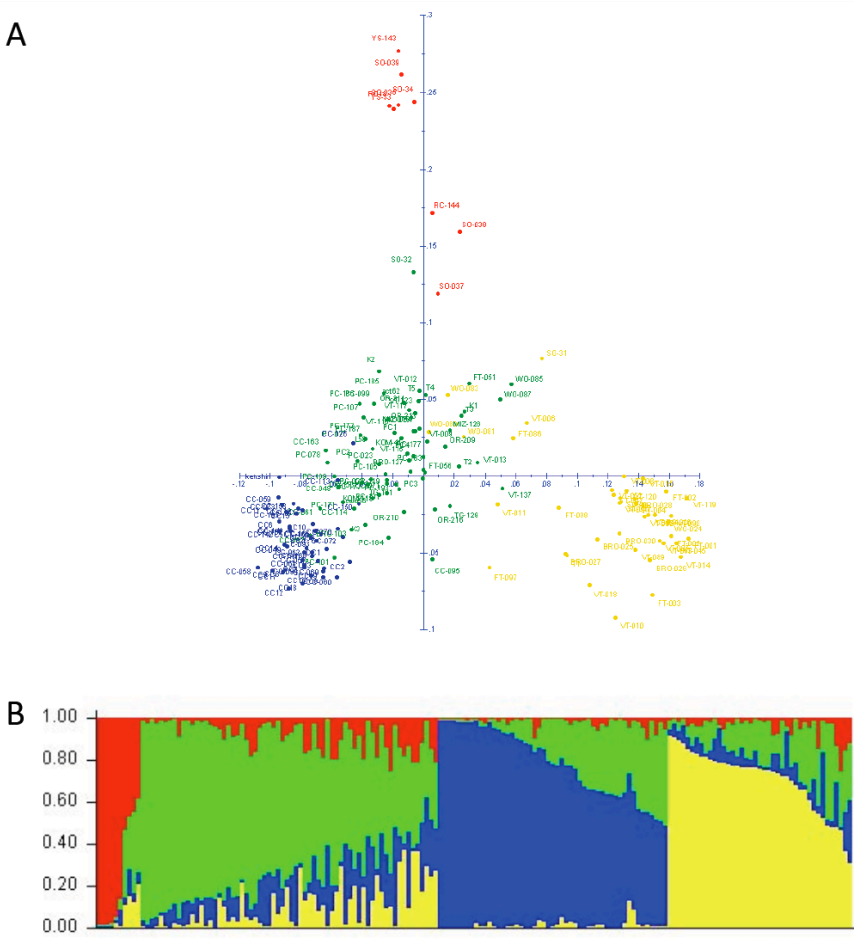


Figure 1.Principal components analysis(A) and STRUCTURE(B) results. Colors define supopulations: red (oil: Population 1), green(PC+T:population 2),blue(CC: population 3) and yellow(VT+FT: population 4)

Of the 168 accessions, 109 were assigned to a subgroup with a membership probability of p>0.70. Fifty-nine accessions were assigned to more that one subgroup and had membership probabilities below 0.7 corresponding to several subgroups (Supplementary Table 1).

The PCO-MC method, which couples principal coordinate analysis to a clustering procedure for the inference of population structure from multi-locus genotype data, only showed one small distinct, statistically significant cluster, corresponding to oil types of Indian origin (Figure 1A). The first two axes accounted for 12.27% and 8.21% of the total variation in the marker data.

Metabolite variation

To estimate the variation within and between the different Brassica morphotypes, boxplots were constructed based on the total content value per metabolite for each subgroup as defined by STRUCTURE (Figure 2). Visual inspection of the box plots and the analysis of the significant differences between groups (LSD) showed a significant variation in the amount of most of the carotenoids and folate between the four population subgroups. Conversely, the content of chorophyll *b* and lutein was significantly different between few subgroups and the content of tocopherols was just significantly different between the Chinese cabbage (CC) subgroup 3 compared to the other subgroups (Supplementary Table 2).

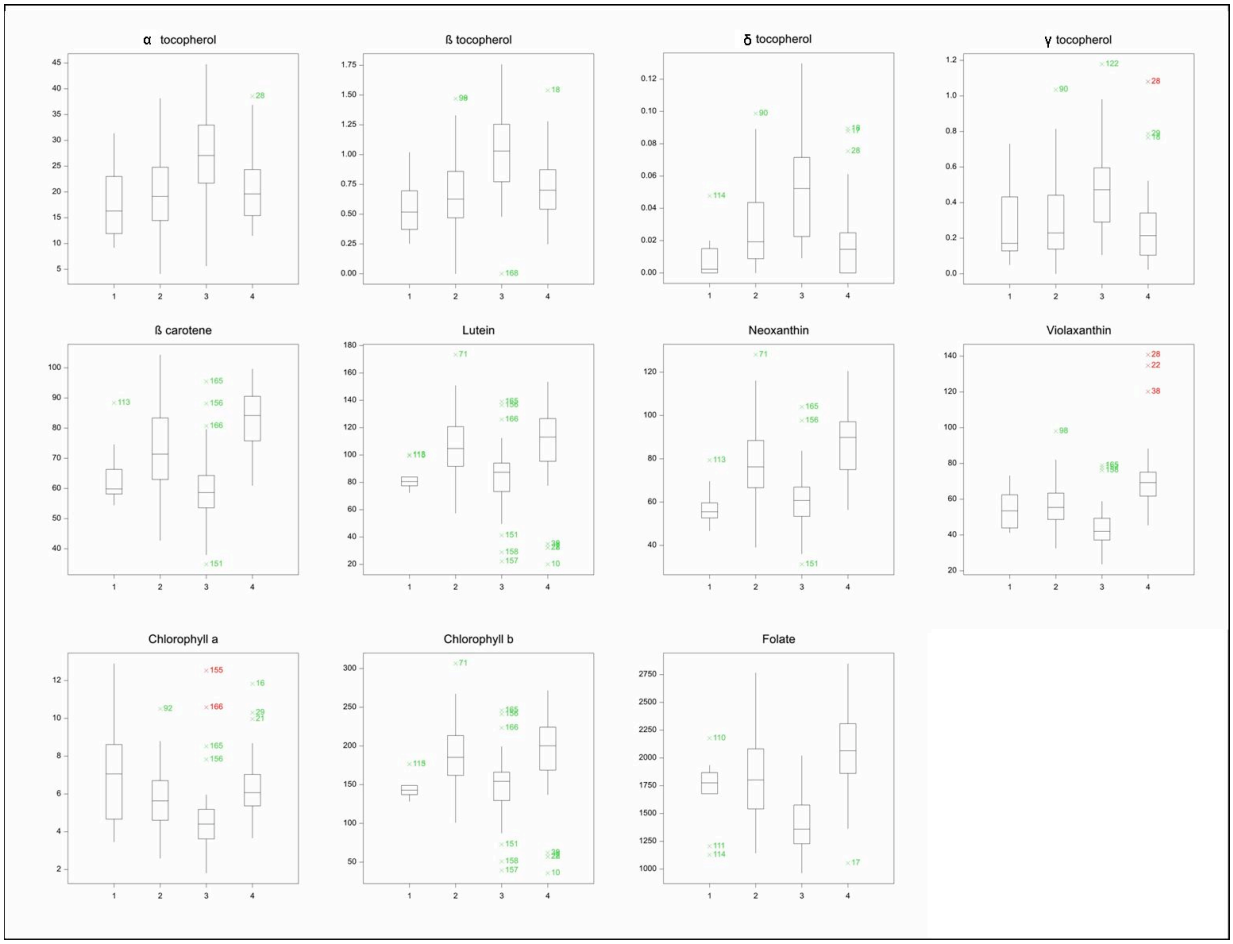


Figure 2. Boxplots of metabolite content variation.The numbers indicate subpopulation as defined with STRUCTURE.Oil:Population 1, PC+T:population 2,CC: population 3 and VT+FT: population 4 .

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 significant markers associated varied greatly between the different metabolites and analysis methods as shown in Table 1.

	TOCOPHEROLS				CAROTENOIDS				CHLOROPHYLLS		
	δ tocopherol	γ tocopherol	β tocopherol	α tocopherol	lutein	βcarotene	neoxanthin	violaxanthin	chlorophyll *b*	chlorophyll *a*	folate
Model (1)	39	56	91	70	58	108	108	104	58	66	115
Model (2)	9	12	13	16	24	32	32	30	23	19	16
Model (3)	41	57	89	71	59	108	110	105	59	65	103
Model (4)	9	12	6	16	24	32	32	29	24	20	19
RF	17	13	16	21	8	15	12	14	10	8	17
RF-Model(4)*	1	2	3	5	3	7	5	2	4	2	4
RF-Model(1)*	12	11	15	17	6	15	12	14	8	6	14

Table 1.Association mapping result for the different linear models and Random forests (RF).Numbers indicate significant markers (P<0.05) found per metabolite.

To test for marker-trait associations in our data 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 significant markers found as associated to a specific metabolite was strongly inflated and ranged from 39 to 115 per metabolite. The highest number of significant markers (>100) was found for β-carotene, neoxanthin, violaxanthin and folate; these metabolites also showed the greatest variation in content between subgroups.

To account for the level of relatedness between individuals and to reduce confounding by population structure, 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 (41-110). Interestingly, the results of these two models are 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 was substantially reduced. Although in many cases the number of significant markers was reduced by more than 50%, 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 subgroup and the other subgroups.

Furthermore, when we combined the information from the Q matrix and the K matrix in the full model 4, following the approach of 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.

After correcting for multiple testing only eight markers remained significantly associated withmetabolites: Alu_M476_0, pTAmCAC_148_3, Hae_M294_2, pGGmCAA_335_2 and pTAmCAT_312_3 for β-carotene; Alu_M476_0 for neoxanthin; pTAmCAC_101_7 and pTAmCAC_270_9 for violaxanthin and pGGmCAA_335_2 and pGGmCAA_386_7 for folate. Each of these markers explained between 4.3% and 6.3 % of the metabolites phenotypic variation

To summarize the results obtained from the full model 4, we constructed a network with a total of 102 significant associated markers associated to the metabolites ($P<0.05$). This network allowed us to connect the metabolites of similar pathways through markers (Figure 3).

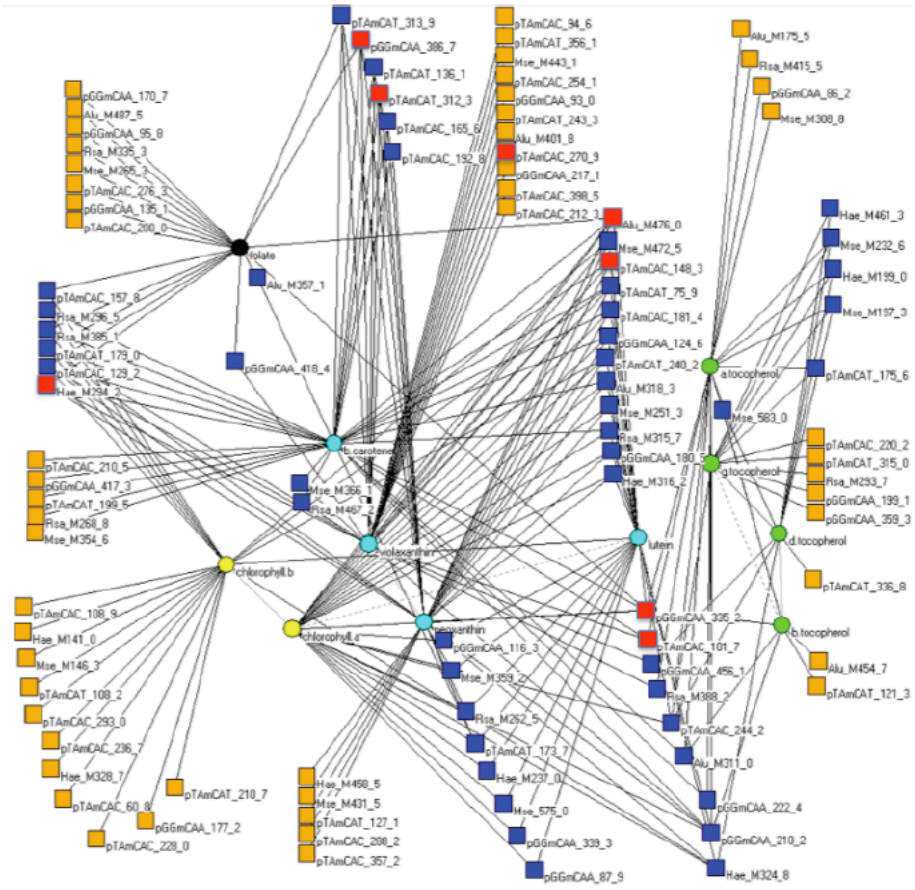


Figure 3. Network of metabolite and marker Association for model 4. Carotenoids (light blue), tocopherols (green), chlorophylls (yellow) and folate (black).Markers box color indicate association to one metabolite (orange), to more than one metabolite(blue) and significant q-value(red).

The overlap of 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 that was significantly associated to all the four pathways. The overlap between pathways was restricted to nine markers associated with both tocopherols and carotenoids, five with tocopherols and chlorophylls, nine with chlorophylls and folate, ten with both carotenoids and folate and one with both tocopherols and folate. However, the largest overlap was found between markers associated with both carotenoids and the chlorophylls with a total of 37 markers found as significant in both pathways. Within pathways the overlap between individual metabolites of the pathway varied as follows: in the carotenoids the largest number of common markers (18) was found between lutein and neoxanthin, in the tocopherols the largest number (six) was found between α- and γ tocopherols and the chlorophylls *a* and *b* share four markers (data not shown).

Random Forests

In spite of not considering any correction for population structure in the Random Forests association approach we decided to evaluate its performance in comparison to the simple model (1), which does not include any correction, and the full model, which includes the Q and K matrix correction (4). The number of associated markers with significance level of $p < 0.05$ per metabolite ranged between metabolites from eight for lutein and chlorophyll *a* to 21 for α-tocopherol. Interestingly, the results showed that the number of significant markers obtained with the RF approach was much lower for all the metabolites if compared to the simple model (1).

However, in the Random Forests we used a rather conservative approach to calculate the markers p-values, which in turn resulted in a low number of significant markers for all the metabolites (see material and methods). Nonetheless, the overlap of significant markers between methods is large; with few exceptions the significant markers found with RF were also significant in the simple model (1). For example, a complete overlap was found for β-carotene, neoxanthin and violaxanthin, while a

lower overlap was found for δ tocopherol; 12 markers were identified with RF out of the 17 that were significant in the simple model (1) (Table 1).

In contrast, when the results obtained with Random Forests are compared to the results obtained with the full model (4) the overlap between associated markers obtained through both methods was very low and ranged from one out of nine for δ-tocopherol to seven out of 15 for β-carotene.

Marker-trait Associations across methods

To compile a list of plausible strong associations for future research we compared the markers found as significant $p<0.05$ with RF to the markers found as significant with all the linear models (1-4) (Supplementary Table 3). If a marker was found as significant across methods it was considered as a strong metabolite-marker association; in total we found 34 associations, which represent 16 markers that met this criteria. Interestingly, we found at least one marker per metabolite in common across all the association mapping methods, from one for δ tocopherol to seven for β-carotene. Furthermore, from these markers many were also common between metabolites from the same pathway but not between metabolites from different pathways. For example, marker pTAmCAC_244_2 was found significant for all the tocopherols except β-tocopherol but was not found significant associated with the carotenoids, chlorophylls or folate, while marker pTAmCAC_148_3 was found as significant for all the carotenoids and chlorophylls except chlorophyll a but was not found as significant for the tocopherols or folate.

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 (Breseghello and Sorrells, 2006).

B. rapa crop types are the result of different breeding histories around the world. In our study we identified with STRUCTURE the presence of 4 subpopulations, which showed correlation with the origin and morphotypes of *B. rapa*. These groups possibly arose as a result of different patterns of adaptation, domestication and

background selection (Thornberry et al. 2001, Wright and Gaut 2005, Zhao et al. 2005).

An alternative method to the STRUCTURE Q matrix, which has been used to capture the genome wide patterns of relatedness between accessions, is the principal coordinates analysis (PCO) loading results. In our study, after PCO the evidence for the presence of subgroups was not as clear when compared to the results obtained with STRUCTURE. Furthermore, when the first two components are plotted, the PCO only captured 20% of the variation. Results from both the STRUCTURE membership probabilities and PCO analysis illustrate the highly admixed nature of the accessions in this *B. rapa* core collection.

In practice, the PCO analysis as reported in Reeves et al. (2009) was computationally much more laborious and less successful to assign the highly heterogeneous *B. rapa* accessions into subgroups. Based on the above, the inclusion of the principal components loadings into the association model was not considered within the scope of this research.

To reveal the marker-metabolite associations we applied four linear models and explored the impact of the different levels of relatedness between accessions on the results. We included either the results from STRUCTURE (model 2), kinship coefficients (model 3) or both (model 4) in the association models.

Correcting for the level of relatedness (Kinship and/or Q matrix) resulted in a significant reduction in the number of marker-trait associations as shown in the number differences between model (1) and model (2) and (4). Although there was always overlap between the markers identified in these models, comparisons revealed that new associations arise when the Q matrix (model 2) and the Q matrix together with the K matrix (model 4) were introduced in the model.

After correction with the inclusion of the kinship matrix in model (3) the reduction in the number of markers was not evident. Explanation can be found in the fact that the K matrix does not capture large differences between subpopulations (as defined by STRUCTURE) because the within population variation is larger than the between population variation (Pino Del Carpio et al. chapter 2, Zhao et al. 2005).

In terms of how these methods performed reducing the false positive rates to distinguish causal from spurious signals, 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 overall the

statistical models. As a result, in spite of introducing a correcting term in our models we still expect a rate of 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).

Within our results we decided to analyze further the significant markers obtained with the full model (4), which included the K and Q matrix in order to reveal the existence of different levels of genetic co-regulation of the metabolites.

For example we observed that common markers could be found between and within pathways (carotenoids, tocopherols, chlorophylls and folate). The high number of common markers that were found between carotenoids and chlorophylls confirmed biological regulatory and biochemical expectations. Both are derivatives of the central intermediate geranylgeranyl diphosphate (GGDP) used for the synthesis of phytoene and chlorophylls. We hypothesize that these compounds could be under pleiotropic effect or under the control of genes in close linkage because of their association to the same marker Alu_476_0, a marker that stays significant after multiple testing correction (q-value).

In the present study we considered the use of Random Forests (RF) as a complementary method to our association study. Because the performance of this method in association analysis has not been previously tested in plant systems, we first evaluated our RF results in comparison to the ones obtained with the already validated and widely used model (4) and the simple model (1). One reason to include the Random Forests approach in the present study is that although population structure methods correct for confounding these are underpowered because of the small sample size. One striking result of the RF analysis is the small number of associated markers that are found for all the metabolites. 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, to be more accurate in our selection and to be fair in our comparison, we introduced a permutation method to calculate the significance of each marker association in this study. However, this calculation proved to be very stringent in the number of selected significant markers if compared to the numbers obtained with the linear models (1-4).

For the selection of true strong associations we compiled markers, which are consistently found as significant across all methods, including Random Forest. (Supplementary Table 3).

The importance of these results is confirmed for example in the case of the marker pTAmCAC_244_2 which was found as associated to δ-, γ- and α-tocopherol, pTAmCAC_148_3 found as associated to lutein, β-carotene, neoxanthin, violaxanthin and chlorophyll b; and pTAmCAC_101_7 found as associated to violaxanthin. The QTL mapping results from a double haploid population show the existence of overlapping QTLs for β-carotene, lutein, chlorophyll b and neoxanthin in the region where the marker pTAmCAC_148_3 is located. In this genomic region of chromosome A03 the genes ε-cyclase, β-carotene hydroxylase and carotenoid isomerase can be named as potential candidates as identified based on synteny with Arabidopsis (Schranz et al 2006). For markers pTAmCAC_244_2 located on chromosome A10 and pTAmCAC_101_7 located on chromosome A06, no QTL have been reported for the metabolites included in our study. This result is not unexpected since with QTL mapping we only measure the effects of parental alleles, while in a core collection many alleles are represented. However, their genomic location points towards syntenic regions for the candidate genes Phytoene desaturase 1 for the tocopherols and Zeaxanthin epoxidase for violaxanthin respectively (data not shown).

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 markers that are kept as significant after multiple testing correction (q-value) of the model (4) results and the 16 markers selected across methods should be considered as the most promising 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. We consider that the results obtained with Random Forests when compared to model (1) can be promising for its use in association studies to identify significant markers within subpopulations or marker-associations for traits with a distribution that is not correlated to the underlying population structure.

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 (Q) and the use of Random Forests.

Acknowledgements

We would like to thank Harry Jonker and Yvonne Birnbaum for their help in the isoprenoids and folates analyses. This research was funded by the IOP Genomics project "Brassica vegetable nutrigenomics" IGE 05010.

Table S1.Membership probabilities and group assignment based on STRUCTURE

Accession	origin	cultivar name		genebank	order
FT.001	Netherlands	Halflange Witte Blauwkop Ingesneden Blad-Barenza	Turnip	CGN06669	1
FT.002	United Kingdom	De Norfolk a Collet Rouge	Turnip	CGN06673	2
VT.008	Pusa Chandrina	India	Turnip	CGN06711	3
VT.010	Hungary	Platte Ronde Blauwkop Ingesneden Blad-Lila Ker	Turnip	CGN06718	4
VT.015	Italy	Bianca Lodigiana; Italiaanse Witte	Turnip	CGN06724	5
MIZ.019	Japan	Bladmoes Geslu	Mizuna	CGN06790	6
PC.022	Netherlands		Pak choi	CGN06816	7
WO.024	Sweden	Svalof 0308	Turnip rape	CGN06818	8
BRO.025	Italy	Natalino	Broccoleto	CGN06823	9
BRO.027	Italy	Quarantina	Broccoleto	CGN06825	10
BRO.028	Italy	Tardivo	Broccoleto	CGN06827	11
BRO.029	Italy	Norantino	Broccoleto	CGN06828	12
BRO.030	Italy	Sessantina	Broccoleto	CGN06829	13
YS.033	Germany	Dys 1	Yellow sarson	CGN06835	14
SO.035	Bangladesh	Somali Sarisa	Turnip rape	CGN06837	15
SO.037	Bangladesh	Kalyania	Turnip rape	CGN06839	16
SO.038	Germany	Toria	Turnip rape	CGN06840	17
SO.039	Bangladesh	Sampad	Turnip rape	CGN06841	18
KOM.041	Japan	Komatsuna	Neep greens	CGN06843	19
CC.048	Soviet Union		Chinese cabbage	CGN06867	20
VT.052	Netherlands	Hilversumse; Marteau	Turnip	CGN07166	21
CC.057	China		Chinese cabbage	CGN07182	22
CC.058	Czech Republic		Chinese cabbage	CGN07183	23
CC.059	Korea		Chinese cabbage	CGN07184	24
CC.068	Bulgaria		Chinese cabbage	CGN07196	25
CC.072	China	BRA 207/70	Chinese cabbage	CGN07201	26
PC.078	Netherlands	Choy Sam	caixin	CGN07211	27
WO.080	Pakistan		Turnip rape	CGN07216	28
WO.081	Pakistan		Turnip rape	CGN07217	29
WO.083	Pakistan		Turnip rape	CGN07220	30
BRO.103	Indonesia	Tsja Sim; No. P1R5T5	caixin	CGN15158	31
PC.105	China	BRA 77/72	Wutacai	CGN15171	32
VT.115	Japan	Kairyou Hakata	Turnip	CGN15199	33
KOM.118	Japan	Komatsuna	Neep greens	CGN15202	34
MIZ.128	Japan	Round Leaved Mibuna	Mizuna	CGN17279	35
TG.129	Japan	Vitamin Na	Neep greens	CGN17280	36
TG.131	Japan	Maruba Santo Sai	Neep greens	CGN17282	37
FT.004	Denmark	Lange Gele Bortfelder	Turnip	CGN06678	38
FT.005	Germany	Ochsenhorner	Turnip	CGN06688	39
VT.006	India	Pusa Chandrina	Turnip	CGN06709	40
VT.007	Soviet Union	Maiskaja	Turnip	CGN06710	41
VT.011	Soviet Union	Platte Witte Blauwkop Ingesneden Blad-Siniaja	Turnip	CGN06719	42
VT.014	Italy	Platte Witte Blauwkop Heelblad-Milan	Turnip	CGN06722	43
VT.018	Netherlands	Goudbal; Golden Ball	Turnip	CGN06774	44
PC.023	China	Si Yue Man	Pak choi	CGN06817	45
BRO.026	Italy		Broccoleto	CGN06824	46
SO.031	USA		Turnip rape	CGN06832	47
SO.032	India	Pusa Kalyani	Turnip rape	CGN06834	48
SO.034	Bangladesh	Australian RARS	Turnip rape	CGN06836	49
SO.040	Canada	Candle	Turnip rape	CGN06842	50
VT.044	Soviet Union	Soloveckaja	Turnip	CGN06859	51
CC.049	Netherlands	Granaat	Chinese cabbage	CGN07143	52
FT.051	Soviet Union	Krasnaja	Turnip	CGN07164	53
VT.053	Germany	Teltower Kleine	Turnip	CGN07167	54
FT.056	France	Daisy; Bladraap	Turnip	CGN07179	55
CC.060	China		Chinese cabbage	CGN07185	56
CC.061	Yugoslavia		Chinese cabbage	CGN07188	57
CC.062	Germany		Chinese cabbage	CGN07189	58
CC.067	Japan		Chinese cabbage	CGN07195	59
CC.069	USA		Chinese cabbage	CGN07198	60
CC.070	Korea	BRA 47/22	Chinese cabbage	CGN07199	61
CC.071	Japan	BRA 211/69	Chinese cabbage	CGN07200	62
CC.073	China	BRA127/67	Chinese cabbage	CGN07202	63
PC.076	China		Pak choi	CGN07205	64
WO.084	Pakistan		Turnip rape	CGN07221	65
FT.086	Pakistan		Turnip	CGN07223	66
WO.087	Pakistan		Turnip rape	CGN07226	67
FT.088	Netherlands	Blauwkop Heelblad-Oliekannetjes	Turnip	CGN10985	68
VT.089	France	D'Auvergne Hative	Turnip	CGN10995	69
VT.091	United Kingdom	Snowball; Blanc Rond de Jersey	Turnip	CGN10999	70
VT.092	Netherlands	Amerikaanse Witte Roodkop Heelbad	Turnip	CGN11000	71
CC.093	China		Chinese cabbage	CGN11002	72
CC.094	Japan		Chinese cabbage	CGN11003	73
CC.095	China		Chinese cabbage	CGN11005	74
FT.097	Germany	Buko; Bladraap	Turnip	CGN11010	75
PC.099	China	Chinese Bai Cai	Pak choi	CGN13924	76
PC.101	China	Tientsin; Celery, Shantung, Peking	Pak choi	CGN13926	77
PC.107	Hong Kong	Dwarf	Pak choi	CGN15184	78
CC.112	China	Bao Tou Qing	Chinese cabbage	CGN15194	79
CC.113	China	Bei jing 106	Chinese cabbage	CGN15195	80

Table S1.Membership probabilities and group assignment based on STRUCTURE (continued)

CC.114	China	Xiao Qing Kou	Chinese cabbage	CGN15196	81
VT.117	Japan	Toya	Turnip	CGN15201	82
CC.125	Korea		Chinese cabbage	CGN15222	83
BRO.127	Japan	Edible Flower	Broccoleto	CGN17278	84
VT.137	Uzbekistan		Turnip	CGN20735	85
CC.140	Japan	Kashin	Chinese cabbage	CGN20771	86
CC.141	Japan	Kyoto Sang	Chinese cabbage	CGN21731	87
CC.142	Japan	Matsushima Jun Sang	Chinese cabbage	CGN21732	88
VT.009	Japan	Ronde Rode-Tsutsui	Turnip	CGN06717	89
VT.012	Japan	Ronde Rode Heelblad-Yurugu Red	Turnip	CGN06720	90
VT.013	Japan	Ronde Rode Heelblad-Scarlet Ball	Turnip	CGN06721	91
VT.045	Italy	Milanskaja; Italiaanse Witte	Turnip	CGN06860	92
VT.116	Japan	Nagasaki Aka	Turnip	CGN15200	93
YS.143	USA	R500	Yellow sarson	FIL500	94
RC.144	USA	Rapid cycling	Turnip rape	FIL501	95
WO.085	Pakistan		Turnip rape	CGN07222	96
OR.213	China	Huang Po Tian You Cai	Turnip rape	OCRI0235	97
OR.209	China	Huang Gang Bai You Cai	Turnip rape	OCRI1752	98
OR.211	China	Yi Chang Xiao You Cai	Turnip rape	OCRI1771	99
OR.210	China	Lou Tian You Bai Cai	Turnip rape	OCRI1757	100
PC.183	China	Ai Kuang Qing	Pak choi	OCRI3742	101
OR.219	China	Ping Ba Bai You Cai	Turnip rape	OCRI3801	102
CC.153	China	Bao Tou Bai Cai	Chinese cabbage	VO2A0020	103
CC.163	China	Tian jing Bai Cai	Chinese cabbage	VO2A0049	104
CC.150	China	Yu Quan Bao Tou Qing	Chinese cabbage	VO2A0012	105
CC.156	China	Huang Yang Bai	Chinese cabbage	VO2A0030	106
CC.158	China	Gao Zhuang Huang Yang Bai	Chinese cabbage	VO2A0034	107
CC.160	China	Qing Kou Bai Cai	Chinese cabbage	VO2A0044	108
CC.161	China	Huang Yang Bai	Chinese cabbage	VO2A0046	109
CC.168	China	Lou Yang Da Bai Cai	Chinese cabbage	VO2A0068	110
PC.172	China	No 17 Bai Cai	Pak choi	VO2B0207	111
PC.171	China	B139 Xiao Bai Cai	Pak choi	VO2B0206	112
PC.173	China	Kui Shan Li Ye Bai Cai	Pak choi	VO2B0223	113
PC.177	China	Ai Jiao Huang	Pak choi	VO2B0396	114
PC.191	China	Wuhan Ai Jiao Huang	Pak choi	VO2B0988	115
FT.003	Netherlands	Lange Witte Roodkop	Turnip	CGN06675	116
VT.017	Netherlands	Platte Witte Meirapen	Turnip	CGN06732	117
FT.047	Soviet Union	Moskovskij	Turnip	CGN06866	118
VT.090	France	De Croissy	Turnip	CGN10996	119
VT.119	Netherlands	Roodkop-Pfalzer	Turnip	CGN15209	120
VT.120	Platte Gele Boterknol	Netherlands	Turnip	CGN15210	121
VT.123	Terauchi-Kabu	Japan	Turnip	CGN15220	122
CC.165	China	Tian jing Bai Cai	Chinese cabbage	VO2A0054	123
CC.167	China	Lou Yang Large Bai Cai	Chinese cabbage	VO2A0062	124
CC.169	China	Huang Yang Bai	Chinese cabbage	VO2A0069	125
OR.216	China	Xi Qiu Bai Cai	Turnip rape	VO2B0655	126
PC.184	China	Ai Jiao Bai	Pak choi	VO2B0656	127
PC.185	China	Qing Ken Bai Cai	Pak choi	VO2B0691	128
PC.186	China	D94 Bai Cai	Pak choi	VO2B0694	129
PC.187	China	Ai Hei Ye Kui Shan Bai Cai	Pak choi	VO2B0695	130
PC.189	China	Ai Hei Ye Kui Shan Bai Cai	Pak choi	VO2B0715	131
PC.193	China	CII	Pak choi	VO2B1263	132
T.1.	Company line		Turnip		133
T.2.	Company line		Turnip		134
T.3.	Company line		Turnip		135
PC.1.	Company line		Pak choi		136
PC.2.	Company line		Pak choi		137
PC.3.	Company line		Pak choi		138
PC.4.	Company line		Pak choi		139
K.1.	Company line		Komatsuna		140
K.2.	Company line		Komatsuna		141
K.3.	Company line		Komatsuna		142
CC.1.	Company line		Chinese cabbage		143
CC.2.	Company line		Chinese cabbage		144
CC.3.	Company line		Chinese cabbage		145
CC.4.	Company line		Chinese cabbage		146
CC.5.	Company line		Chinese cabbage		147
CC.6.	Company line		Chinese cabbage		148
CC.7.	Company line		Chinese cabbage		149
CC.8.	Company line		Chinese cabbage		150
T.4.	Company line		Turnip		151
T.5.	Company line		Turnip		152
CC.9.	Company line		Chinese cabbage		153
CC.10.	Company line		Chinese cabbage		154
CC.11.	Company line		Chinese cabbage		155
CC.12.	Company line		Chinese cabbage		156
CC.13.	Company line		Chinese cabbage		157
CC.14.	Company line		Chinese cabbage		158
CC.15.	Company line		Chinese cabbage		159
CC.16.	Company line		Chinese cabbage		160
CC.17.	Company line		Chinese cabbage		161
CC.18.	Company line		Chinese cabbage		162
CC.19.	Company line		Chinese cabbage		163
RO18	India		Yellow sarson		164
L58	United Kingdom		Pak choi		165
Kenshin	Korea		Chinese cabbage		166
cWU56	China		wutacai		167
zct62	China		zicaitai		168

Table S1.Membership probabilities and group assignment based on STRUCTURE (continued)

	structure					
order	membership probability				group	<0.70
	1	2	3	4		
1	0.89	0.03	0.04	0.04	2	
2	0.76	0.16	0.03	0.05	2	
3	0.85	0.04	0.08	0.03	2	
4	0.89	0.03	0.01	0.07	2	
5	0.85	0.10	0.03	0.03	2	
6	0.15	0.55	0.24	0.06	2	*
7	0.20	0.39	0.14	0.27	2	*
8	0.84	0.04	0.06	0.06	2	
9	0.83	0.04	0.02	0.11	2	
10	0.78	0.06	0.06	0.11	2	
11	0.87	0.06	0.03	0.04	2	
12	0.92	0.03	0.01	0.05	2	
13	0.76	0.10	0.01	0.13	2	
14	0.01	0.02	0.96	0.02	3	
15	0.01	0.02	0.96	0.01	3	
16	0.24	0.34	0.39	0.04	3	*
17	0.15	0.26	0.40	0.19	3	*
18	0.01	0.01	0.98	0.01	3	
19	0.12	0.73	0.08	0.07	2	
20	0.02	0.16	0.13	0.69	3	*
21	0.73	0.05	0.07	0.15	2	
22	0.02	0.14	0.02	0.83	3	
23	0.01	0.11	0.01	0.87	3	
24	0.01	0.03	0.19	0.77	3	
25	0.01	0.10	0.00	0.89	3	
26	0.05	0.07	0.09	0.79	3	
27	0.02	0.85	0.04	0.08	2	
28	0.36	0.53	0.08	0.03	2	*
29	0.38	0.42	0.13	0.07	2	*
30	0.32	0.44	0.21	0.04	2	*
31	0.03	0.83	0.02	0.12	2	
32	0.06	0.70	0.09	0.15	2	
33	0.27	0.62	0.07	0.04	2	*
34	0.15	0.66	0.13	0.07	2	*
35	0.40	0.41	0.17	0.02	2	*
36	0.17	0.72	0.08	0.02	2	
37	0.13	0.43	0.06	0.38	2	*
38	0.78	0.07	0.02	0.14	2	
39	0.91	0.03	0.01	0.06	2	
40	0.51	0.19	0.27	0.03	2	*
41	0.89	0.03	0.05	0.03	2	
42	0.48	0.47	0.03	0.02	2	*
43	0.74	0.19	0.03	0.05	2	
44	0.79	0.10	0.02	0.09	2	
45	0.07	0.74	0.09	0.10	2	
46	0.94	0.02	0.01	0.03	2	
47	0.54	0.22	0.22	0.02	2	*
48	0.11	0.43	0.38	0.08	2	*
49	0.03	0.02	0.94	0.01	3	
50	0.87	0.11	0.01	0.02	2	
51	0.78	0.11	0.08	0.04	2	
52	0.06	0.10	0.01	0.83	3	
53	0.42	0.25	0.25	0.07	2	*
54	0.90	0.05	0.01	0.04	2	
55	0.39	0.31	0.17	0.13	2	*
56	0.09	0.14	0.01	0.76	3	
57	0.01	0.35	0.01	0.63	3	*
58	0.05	0.05	0.01	0.90	3	
59	0.02	0.30	0.04	0.65	3	*
60	0.09	0.27	0.01	0.63	3	*
61	0.02	0.07	0.13	0.78	3	
62	0.14	0.39	0.26	0.21	2	*
63	0.04	0.47	0.02	0.47	2	*
64	0.05	0.50	0.08	0.37	2	*
65	0.31	0.37	0.11	0.20	2	*
66	0.43	0.36	0.04	0.18	2	*
67	0.42	0.28	0.28	0.02	2	*
68	0.53	0.35	0.07	0.05	2	*
69	0.75	0.10	0.01	0.15	2	
70	0.86	0.04	0.01	0.09	2	
71	0.70	0.10	0.10	0.10	2	
72	0.03	0.11	0.07	0.80	3	
73	0.02	0.19	0.01	0.77	3	
74	0.33	0.21	0.01	0.45	3	*
75	0.64	0.03	0.02	0.30	2	*
76	0.05	0.55	0.31	0.09	2	*
77	0.02	0.57	0.01	0.40	2	*
78	0.02	0.81	0.04	0.13	2	
79	0.02	0.16	0.01	0.81	3	
80	0.09	0.08	0.14	0.69	3	

Table S1.Membership probabilities and group assignment based on STRUCTURE (continued)

81	0.03	0.27	0.06	0.65	3	*
82	0.03	0.75	0.18	0.04	2	
83	0.04	0.33	0.09	0.54	3	*
84	0.19	0.43	0.09	0.29	2	*
85	0.67	0.22	0.07	0.04	2	*
86	0.02	0.40	0.02	0.57	3	*
87	0.03	0.03	0.02	0.92	3	
88	0.02	0.06	0.04	0.89	3	
89	0.11	0.79	0.07	0.03	2	
90	0.02	0.65	0.18	0.15	2	*
91	0.13	0.76	0.09	0.03	2	
92	0.77	0.03	0.10	0.11	2	
93	0.03	0.77	0.17	0.03	2	
94	0.00	0.01	0.98	0.01	3	
95	0.15	0.32	0.43	0.11	3	*
96	0.35	0.36	0.27	0.02	2	*
97	0.19	0.66	0.06	0.10	2	*
98	0.03	0.76	0.16	0.04	2	
99	0.06	0.72	0.06	0.16	2	
100	0.03	0.72	0.03	0.23	2	
101	0.10	0.80	0.02	0.08	2	
102	0.14	0.39	0.13	0.34	2	*
103	0.02	0.34	0.07	0.57	3	*
104	0.02	0.22	0.06	0.71	3	
105	0.02	0.05	0.11	0.82	3	
106	0.10	0.10	0.08	0.72	3	
107	0.03	0.18	0.09	0.71	3	
108	0.01	0.21	0.03	0.76	3	
109	0.02	0.25	0.03	0.69	3	*
110	0.04	0.22	0.03	0.72	3	
111	0.01	0.74	0.06	0.19	2	
112	0.03	0.86	0.02	0.09	2	
113	0.02	0.78	0.02	0.19	2	
114	0.10	0.67	0.04	0.19	2	
115	0.02	0.84	0.02	0.13	2	
116	0.80	0.09	0.02	0.10	2	
117	0.72	0.07	0.07	0.14	2	
118	0.70	0.23	0.03	0.04	2	
119	0.81	0.13	0.01	0.05	2	
120	0.77	0.18	0.02	0.03	2	
121	0.77	0.10	0.10	0.03	2	
122	0.16	0.57	0.17	0.10	2	*
123	0.03	0.41	0.01	0.55	3	*
124	0.01	0.27	0.07	0.65	3	*
125	0.02	0.24	0.01	0.74	3	
126	0.24	0.59	0.02	0.14	2	*
127	0.02	0.93	0.01	0.05	2	
128	0.03	0.73	0.14	0.10	2	
129	0.02	0.91	0.06	0.02	2	
130	0.02	0.92	0.02	0.04	2	
131	0.01	0.95	0.01	0.03	2	
132	0.03	0.73	0.01	0.23	2	
133	0.63	0.23	0.05	0.10	2	*
134	0.24	0.72	0.02	0.02	2	
135	0.12	0.65	0.20	0.03	2	*
136	0.02	0.60	0.04	0.34	2	*
137	0.01	0.88	0.02	0.09	2	
138	0.07	0.85	0.04	0.05	2	
139	0.06	0.77	0.01	0.16	2	
140	0.13	0.70	0.14	0.04	2	
141	0.13	0.54	0.21	0.13	2	*
142	0.04	0.68	0.02	0.25	2	*
143	0.01	0.01	0.01	0.98	3	
144	0.02	0.02	0.01	0.95	3	
145	0.01	0.01	0.00	0.98	3	
146	0.01	0.01	0.01	0.97	3	
147	0.01	0.01	0.01	0.98	3	
148	0.01	0.07	0.02	0.89	3	
149	0.01	0.01	0.01	0.97	3	
150	0.01	0.01	0.00	0.99	3	
151	0.08	0.75	0.14	0.03	2	
152	0.15	0.67	0.15	0.03	2	*
153	0.01	0.02	0.01	0.97	3	
154	0.01	0.05	0.02	0.92	3	
155	0.01	0.22	0.01	0.75	3	
156	0.01	0.01	0.01	0.97	3	
157	0.01	0.01	0.02	0.96	3	
158	0.00	0.01	0.01	0.98	3	
159	0.06	0.23	0.02	0.69	3	
160	0.01	0.01	0.00	0.98	3	
161	0.01	0.05	0.01	0.93	3	
162	0.01	0.02	0.00	0.97	3	
163	0.04	0.01	0.01	0.94	3	
164	0.03	0.06	0.79	0.13	3	
165	0.02	0.73	0.03	0.22	2	
166	0.02	0.22	0.02	0.75	3	
167	0.01	0.59	0.03	0.37	2	*
168	0.01	0.58	0.20	0.21	2	*

		TOCOPHEROLS				CAROTENOIDS				CHLOROPHYLLS		
population		α tocopherol	β tocopherol	δ tocopherol	γ tocopherol	lutein	βcarotene	neoxanthin	violaxanthin	chlorophyll *a*	chlorophyll *b*	folate
1	Oil	a	a	a	a	a	a	a	ac	a	a	a
2	PC	a	a	a	a	b	b	b	ab	b	b	a
3	CC	b	b	b	b	a	a	a	c	c	a	c
4	FT VT	a	a	a	a	b	c	c	d	ab	b	b

Table S2. LSD result of metabolite variation based on STRUCTURE subpopulations

Marker	Model(1)	Model(2)	Model (3)	Model(4)	RF
pTAmCAC_244_2	0,01	0,04	0	0,03	<0.024
pTAmCAC_244_2	9,99E-04	0,04	9,03E-04	0,04	<0.04
pGGmCAA_210_2	9,99E-04	0	0	0	<0.01
Mse_583_0	9,99E-04	0,01	1,39E-04	0,01	<0.014
pGGmCAA_210_2	9,99E-04	0	2,78E-05	0	<0.007
Mse_583_0	9,99E-04	0,01	2,15E-04	0,01	<0.006
Mse_M308_8	9,99E-04	0,03	6,62E-09	0,03	<0.005
Alu_M311_0	9,99E-04	0,01	3,62E-05	0,01	<0.031
pTAmCAC_244_2	9,99E-04	0,02	9,53E-04	0,02	<0.026
pGGmCAA_86_2	0	0,02	2,47E-04	0,03	<0.003
pGGmCAA_210_2	0	0	8,40E-05	0	<0.004
pTAmCAC_148_3	0,01	0,03	0	0,04	<0.001
pTAmCAT_75_9	0,01	0,01	0	0,01	<0.004
Alu_M476_0	9,99E-04	9,99E-04	1,25E-10	0.00050152*	<0.006
pTAmCAC_148_3	9,99E-04	0	4,46E-06	0.00065427*	<0.01
pTAmCAC_192_8	9,99E-04	0,01	5,31E-07	0	<0.038
pTAmCAT_173_7	9,99E-04	0,02	1,33E-09	0,01	<0.017
pTAmCAT_199_5	9,99E-04	0,03	5,87E-08	0,03	<0.011
pTAmCAT_75_9	9,99E-04	0	2,80E-04	0	<0.019
pGGmCAA_335_2	0	0	0	0.0009223*	<0.024
Alu_M476_0	9,99E-04	0,01	3,74E-05	0,01	<0.049
pTAmCAC_148_3	0	0,04	0	0,04	<0.002
pTAmCAT_75_9	0	0,02	0	0,01	<0.008
pTAmCAT_75_9	9,99E-04	0,02	0	0,01	<0.002
Hae_M328_7	0	0,02	0	0,01	<0.024
Alu_M476_0	9,99E-04	9,99E-04	1,58E-10	0.00022003*	<0.009
pTAmCAC_357_2	9,99E-04	0,03	2,60E-04	0,03	<0.024
pTAmCAC_148_3	9,99E-04	0	7,97E-06	0	<0.007
pTAmCAT_173_7	9,99E-04	0	1,64E-09	0	<0.004
pGGmCAA_335_2	0	0	0	0	<0.03
pTAmCAC_101_7	9,99E-04	0	1,60E-07	0.000090982*	<0.012
pTAmCAC_148_3	0	0,01	3,14E-04	0,01	<0.034
Alu_M476_0	9,99E-04	0,03	5,12E-07	0,04	<0.037
pGGmCAA_335_2	9,99E-04	9,99E-04	8,12E-04	0.00009688*	<0.017

Table S3. Metabolite-marker associations .Listed are the P values of the linear models (1-4) and Random forests (RF).

Chapter 4

Association mapping reveals a role for *MAM*-genes and *Myb28* on A03 in the regulation of aliphatic glucosinolate levels in leaves of *Brassica rapa*

Dunia Pino Del Carpio, Mina Jin, Xiaowu Wang, Maria Joao Paulo, Richard GF Visser, Guusje Bonnema

Abstract

The present study was set up to determine the role of *MAM* and *Myb28* in the regulation of aliphatic glucosinolate levels and composition in *Brassica rapa*. These genes were identified as the candidate genes mapping under a previously detected major QTL for glucosinolates. Multi allelic microsatellites markers linked to these genes were developed to sample the allelic variation around these loci. In addition, SSRs were developed to span the whole A03 chromosome to identify additional regions that could be involved in glucosinolate biosynthesis. The SSR markers were profiled over a core collection of 168 accessions, and glucosinolate composition of 6 weeks old leaf material was measured. Association mapping was conducted taking into account the relatedness among accessions because of the presence of population structure in this collection and since the glucosinolate levels and profile varied between subpopulations. Interestingly, not only *MAM* and *Myb28,* but also additional genes (*AOP* and *GS-OH*) involved in side chain modification and transcriptional regulation (*Myb29*) were associated with glucosinolates levels. This illustrates the power of combining QTL and association mapping, with the latter revealing additional allelic variation that segregates in the core collection. Furthermore, a very important observation was the significant reductions in allelic variation around genes associated with glucosinolates regulation in comparison with other positions in A03, illustrating that plant breeding changed genomic patterns of linkage through selection.

Introduction

Glucosinolates are secondary metabolites, which are limited to species of the order Brassicales, which include Brassicas of economic and nutritional importance and the model plant *Arabidopsis thaliana* (Fahey et al 2001, Wittstock et al 2002). Although certain glucosinolate derivatives have antinutritional properties (Fahey et al. 2001, Mithen et al. 2000), several studies have shown that others, like methionine-derived isothiocyanates can offer protection against cancer (Talalay & Fahey, 2001, Zickute et al. 2005, Moore et al. 2007, Traka et al 2008). Additionally, studies have shown the existence of a relationship between glucosinolates and flavour and to insect resistance (Baik et al. 2003, Fenwick et al. 1983, Poelman et al. 2009)

The biosynthesis of glucosinolates proceeds in three stages: side-chain elongation for the methionine and phenylalanine derived glucosinolates, development of the core structure from these glucosinolates, and secondary side-chain modification.

In *Brassica rapa* the study of the variation on the glucosinolate content has largely been focused on the profiling of the diversity of these compounds in specific morphological types or regional varieties but not with the goal to identify genes with regulatory functions (Padilla et al. 2007).

In a previous QTL study in *B. rapa* for glucosinolate variation in leaves, Lou et al. (2008) showed the presence of a major QTL for content of a number of aliphatic glucosinolates on linkage group A03 in a double haploid population of a cross between a yellow sarson and a pak choi accession. A major QTL at this position was also reported by the group of Dr. Wang Xiaowu (IVF-CAAS, Beijing, personal information).

Based on Arabidopsis-Brassica synteny and knowledge of genes involved in aliphatic glucosinolate biosynthesis in Arabidopsis, several candidate genes explaining the QTL can be proposed, like *MAM* and *Myb28* for the major QTL and *AOP* for the minor QTL also on A03.

In Arabidopsis three partially redundant *MAM* (methylthioalkylmalate) genes control the variation in side chain length of methionine-derived glucosinolates (Field et al. 2004,Kroymann et al.2003, 2001). Two α-ketoglutarate-dependent dioxygenases encoded by the tightly linked and duplicated *AOP2* and *AOP3* genes, control production of alkenyl and hydroxyalkyl glucosinolates, respectively (Kliebenstein et al. 2001b). In addition, transcription factors like *HAG1/MYB28, HAG2/MYB76* and

HAG3/MYB29 have been shown to be positive regulators of aliphatic glucosinolate biosynthesis (Gigolashvili et al. 2007, 2008).

In general, the number of QTL one can map for a given trait is limited to allelic differences between parents of a cross. However, the number of potential QTL underlying the trait depends on the degree to which the mapping population captures the total genetic diversity available in nature (Zhu et al. 2008). Population or association mapping exploits both the allelic diversity and all of the recombination events that have occurred in the evolutionary history of a collection of individuals, such as those derived from wild populations, germplasm collections or subsets of breeding lines. As a consequence for identification of genes within a large genomic region, higher mapping resolution could be obtained in studies with natural populations than with bi-parental experimental crosses (Flint Garcia et al. 2005).

In the present study we investigated the variation in the content and structure of glucosinolates (GLS) in a core collection of 168 accessions of diverse morphotype and origin with the ultimate goal to unravel the genetics of the glucosinolate biosynthetic pathway in *B. rapa*. SSR markers with defined physical distance from the candidate genes *MAM* and *Myb28* for the major QTL on A03 were developed using sequences from BACs and contigs containing these genes. Comparative genomic information with Arabidopsis and sequence information of *B. rapa* revealed the presence of additional genes on A03 involved in glucosinolate biosynthesis, like *AOP, GS-OH* and *Myb29*. The SSRs linked to *MAM* and *Myb28* plus SSR markers covering different positions across the A03 genetic map were profiled over the core collection. We followed an association mapping approach to verify the role of *MAM* and *Myb28* in the regulation of glucosinolate composition and to identify additional genes/loci that regulate this pathway in *B.rapa*. In the analyses, the level of relatedness of the accessions from the core collection was taken into account with the inclusion of the Q matrix into the statistical model (Yu et al. 2006, Zhao et al. 2007). Furthermore, we investigated the level of variability of the SSR markers across linkage group A03 in comparison to the whole genome in the different sub populations to reveal whether breeding did change linkage patterns because of selection.

Materials and Methods

Selection of plant material and experimental design

The core collection included a total of 168 accessions representing the different morphotypes and geographic origin of *Brassica rapa* (Table S1). The core collection included 132 accessions that were part of the study of Zhao et al. (2005). From the 168 accessions, 137 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 Oil Crop Research Institute (OCRI) and the Osborn Lab, while breeding companies provided 31 accessions (hybrid varieties and breeding lines). For the metabolite profiling two plants per accession were sown in the greenhouse in September 2006 and September 2007 under the following conditions: 16 hrs light and temperature between 18 and 21C°. 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 from one plant per accession and directly frozen in liquid nitrogen, ground and stored at -70 C°.

DNA was extracted from the ground and frozen material, from the same plant selected for metabolite profiling, with the DNAeasy kit (Qiagen, USA).

LC-MS profiling of glucosinolates

Brassica leaf samples were analyzed for glucosinolates using accurate mass LC-QTOF MS, based on the protocol described in De Vos et al. 2007. In short, 500 mg FW of frozen leaf powder was weighed and extracted with 1500 µl of 99.875% methanol containing 0.125% formic acid. Samples were sonicated for 15 min and then filtered (Captiva 0.45 µm PTFE filter plate, Ansys Technologies) into 96-well plates with 700µl glass inserts (Waters) using a TECAN Genesis Workstation equipped with a 4-channel pipetting robot and a TeVacS 96-wells filtration unit. Per 96-wells plate, 6 quality control samples, consisting of repeated extractions from Chinese Cabbage-068 were included to check for total technical variation, including variation due to extraction, metabolite stability and MS sensitivity. Extracts (5 µl) were injected using an Alliance 2795 HT instrument (Waters), separated on a Phenomenex Luna C18 (2) column (2.0x 150 mm, 3 mm particle size) using a 45 minutes 5-35% acetonitrile gradient in water (both acidified with 0.1% formic acid)

and then detected firstly by a photodiode array detector (Waters 2996) at a wavelength range of 220-600nm and secondly by a Waters-Micromass QTOF Ultima MS with negative electrospray ionization at a mass range of m/z 80-1500. Leucine enkaphalin was used as lock mass for on-line mass calibration. Masslynx software (Waters) was used to control instruments and to process raw data. Peak areas of exact masses corresponding within 5 ppm to known Brassica glucosinolates (http://kanaya.naist.jp/knapsack_jsp/top.html) or as identified in previous LC-QTOF MS Arabidopsis metabolomics experiments (Beekwilder et al. 2008) were integrated using the Quantlyx tool of Masslynx, and used directly in the subsequent data analyses. Variation in relative glucosinolate level between technical replicates varied between 7 and 50%, depending upon peak intensity of the compound.

For the association analysis we included individual values of the glucosinolates and total values for glucosinolate content based on the side chain length: tot3C, tot4C, tot5C, tot6C, tot7C and tot8C. Moreover, ratios comparing the amount of glucosinolates with different chain lengths: R3=(C4+C5+C6+C7+C8)/C3, R4=(C5+C6+C7+C8)/C4, R5=(C6+C7+C8)/C5, R6=(C7+C8)/C6, R7=C8/C7 were calculated in an attempt to find association between these values and markers linked to side chain elongation. Additionally, to identify markers with association to the values of alkenyl glucosinolates and side chain modification, we calculated the values for total alkenyl glucosinolates (totALK= 2Pr+3B+4P); and the ratios for hydroxylation: rat1=2H3B/2H3B+3B, rat2=H4P/H4P+4P and for the ratio of S-oxiginated to alkenyl glucosinolates: rat3=3B/3B+4MSB and rat4=4B/4B+MSP.

Microsatellite profiling

We selected BACs along linkage group A03 to be screened for the presence of microsatellites (SSRs). Out of the total 37 SSR 15 were provided by Mina Jin (Korea) (KS 1-15) which have a random position on A03. Eleven microsatellites were designed because they are linked to candidate genes for glucosinolate biosynthesis (*MAM, Myb28, Myb29*) and 11 newly designed primers were also randomly distributed along the linkage group (WUR 1-11) The BAC sequences targeting *Myb28* and *Myb29* were obtained from Korea (Mina Jin) and the contig sequence containing the triplicated *MAM* gene was obtained from China (Dr. Xiaowu Wang, IVF Caas,

Beijing; personal communication). Physical and genetic position (if known) of most of the markers is shown in Table 1.

PCR amplifications were carried out as follows: in a total volume of 10 ul, the mix included 1 unit of Taq DNA polymerase, 5mM of dNTP, 2.5 ul 10x supertaq buffer and 50 ng of each primer.Final concentration of DNA in the PCR reaction was of 2ng.The PCR was performed on a GeneAmp PCR system 9700 (Applied Bio-system with the following program : 94 C for 2 min,35 cycles with 94 C denaturation for 30 sec, 56 C annealing for 1 min and 72 C elongation for 1 min each step, and then a final elongation step 0f 5 min.

				position map	
Order	BAC	Korea SSR	NAME	vcs3m	jwf
1	KBrB043B23		WUR1	2.5	30.5
2	KBrB069M23	KS40670	KS1	4.8	
3	KBrB021P17		WUR2	5.3	
4	KBrS011B08	KS40510	KS2	13	
5	KBrB091M11	B091	KS3	13.2	
6	KBrH006A08		WUR3	16.9	
7	KBrB080C12		WUR4		
8	KBrB015D21		WUR5		
9	KBrS001M03		WUR6	24	
10	KBrB002E24		WUR7		48
11	KBrB013J16		WUR8		
18	KBrH108E21		BAC-MYB28-7	*	
17	KBrH108E21		BAC-MYB28-6	*	
16	KBrB034G03		BAC-MYB28-5	30.1	
15	KBrB034G03		BAC-MYB28-4	*	
14	KBrB034G03		BAC-MYB28-3	*	
13	KBrB034G03		BAC-MYB28-2	*	
12	KBrB034G03		BAC-MYB28-1	*	
19	*		BAC-MAM3-1	*	
20	*		BAC-MAM3-2	*	
21	KBrB009B09	KS30530	KS4	43.1	
22	KBrH122D17	KS30251	KS5	46.2	73.8
23	KBrB086B23	KS50800	KS6	64.6	
24	KBrH034P23	H034P23-S3	KS7		91.1
25	KBrB030G12	B030G12	KS8		92.5
26	KBrB086M08		WUR9	70.6	
27	KBrB054N05	KS50870	KS9	87.4	126
28	KBrB068H20		WUR10	88.5	
29	KBrH006B04	H006B04-s1	KS10	96.5	
30	KBrH048N09		WUR11	97.8	
31	KBrB016B21	KS50300	KS11	105.5	
32	KBrH013I08	KS50280	KS12	105.6	131
33	KBrB010H02	B010H02-1	KS13		142
34	KBrH006C14	KS50140	KS14	112.5	152
36	KBrB001B07		BAC-MYB29-1	113.1	
37	KBrB001B07		BAC-MYB29-2	*	
35	KBrB055N13	KS50090	KS15	113.7	155

Table 1.List of microsatellites profiled over the core collection. Indicated are the corresponding BAC name and marker KS (korea marker information),WUR (newly designed with sequence information),BAC (marker linked to candidate gene).The asterisk (*) indicates markers with known physical position at a BAC or contig.Map position correspond to genetic maps developed in Korea as reference the sequencing project.

Assessment of population structure

Marker data (AFLP, *Myb*, 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. The Bayesian clustering method as implemented in STRUCTURE revealed 4 subpopulations. Population 1, includes mostly vegetable turnip (VT) and fodder turnip (FT) from European origin and broccoletto accessions: (VT+FT); population 2, includes several types: pak choi, (PC) winter oil, mizuna, mibuna, komatsuna, turnip green, oil rape and Asian turnip (T): (PC+T); population 3, includes annual oil type accessions,spring oil (SO), yellow sarson(YS) and rapid cycling (RC): (SO, YS and RC) and population 4 includes, mainly accessions of Chinese cabbage (CC)

Assessment of genetic diversity

Power Marker V3.0 software was used to estimate polymorphism information content (PIC) according to the following equation:(Botsein et al.1980):

$$PIC=1-\sum_{i=1}^{n}p_i^2-2\left[\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}p_i^2p_j^2\right]$$

where p_i is the frequency of the ith allele, and n is the number of alleles (Botsein et al. 1980). According to Botstein et al. (1980) a marker is highly informative if its PIC is greater than 0.5. Additionally the allelic frequencies were estimated using Power Marker V3.0 software. All the genetic diversity calculations were performed with defined subpopulations as described under population structure: population 1 (VT+FT), population 2: (PC+T), population 3: (YS+SO+RC) and population 4: (CC).

Arlequin 3.11 software was used to perform a hierarchical analysis of molecular variance (AMOVA) (Excoffier et al. 2005) locus by locus, using 1000 permutations and the number of different alleles (FST-like). A distance matrix was also computed. Data conversion for Arlequin was done in SPAGeDi and Genepop web application (http://genepop.curtin.edu.au/). F statistics values were computed according to Weir and Cockerham 1984, to quantify the extent of between-within population differentiation. (FCT: between populations, FSC: within population). Fst values equal 0 when subpopulations are identical in allele frequencies and 1 when they have different alleles. Populations with little divergence have Fst values less than 0.05, moderately differentiated populations have values between 0.05 and 0.15, greatly differentiated populations have values between 0.15 and 0.25 and very greatly

differentiated populations have values greater than 0.25 (Hartl et al. 1997; Mohammadi et al. 2003).

Association mapping

For the association analysis we included microsatellite markers with alleles that showed a frequency higher than 10% and lower than 90% over the 168 accessions. The selection of markers based on these criteria resulted in 95 alleles (Table 4), corresponding to 33 microsatellite marker loci that were tested for marker-trait associations.

A General Linear Model (GLM) as implemented in TASSEL v2.01 software was used for the marker-trait associations. The GLM performs association analysis by a least squares fixed effects linear model (Searle 1987), while also accounting for population structure.

The membership probabilility of each accession was calculated in STRUCTURE as reported previously (Chapter 3). The vector of cluster memberships Q obtained from STRUCTURE was added as a fixed term to the model:

trait = marker + Q + error

A value of 5,000 was set in the parameters for the permutation test (Churchill and Doerge 1994) in order to provide a test of significance that corresponds to the experiment-wise error in order to correct for the fact that multiple comparisons are being made. Marker-trait association was considered significant when the marker allele had a main effect P-value <0.05.

Results

Glucosinolate composition between morpho-groups

A total of 22 glucosinolates could be identified from a whole LC-MS profiling data set of *B.rapa* leaves. The results included twelve aliphatic glucosinolates of 3C (2-propenyl), 4C(3-butenyl, 2-hydroxybut-3-enyl, 4-methylsulfinylbutyl, 4-methylthiobutyl,

methsulphonylbutyl), 5C (4-pentenyl, hydroxypentenyl, methylsulfinyl pentyl), 6C (hexyl), 7C (heptyl) and 8C (8-methylsulfinyloctyl); three Indolic glucosinolates (4-methoxyindol-3-ylmethyl, 1-methoxyindol-3-ylmethyl and indol-3-ylmethy) and one

aromatic glucosinolate (2-phenethyl). The common names, abbreviations, scientific names and descriptions are shown in Supplementary Table 1.

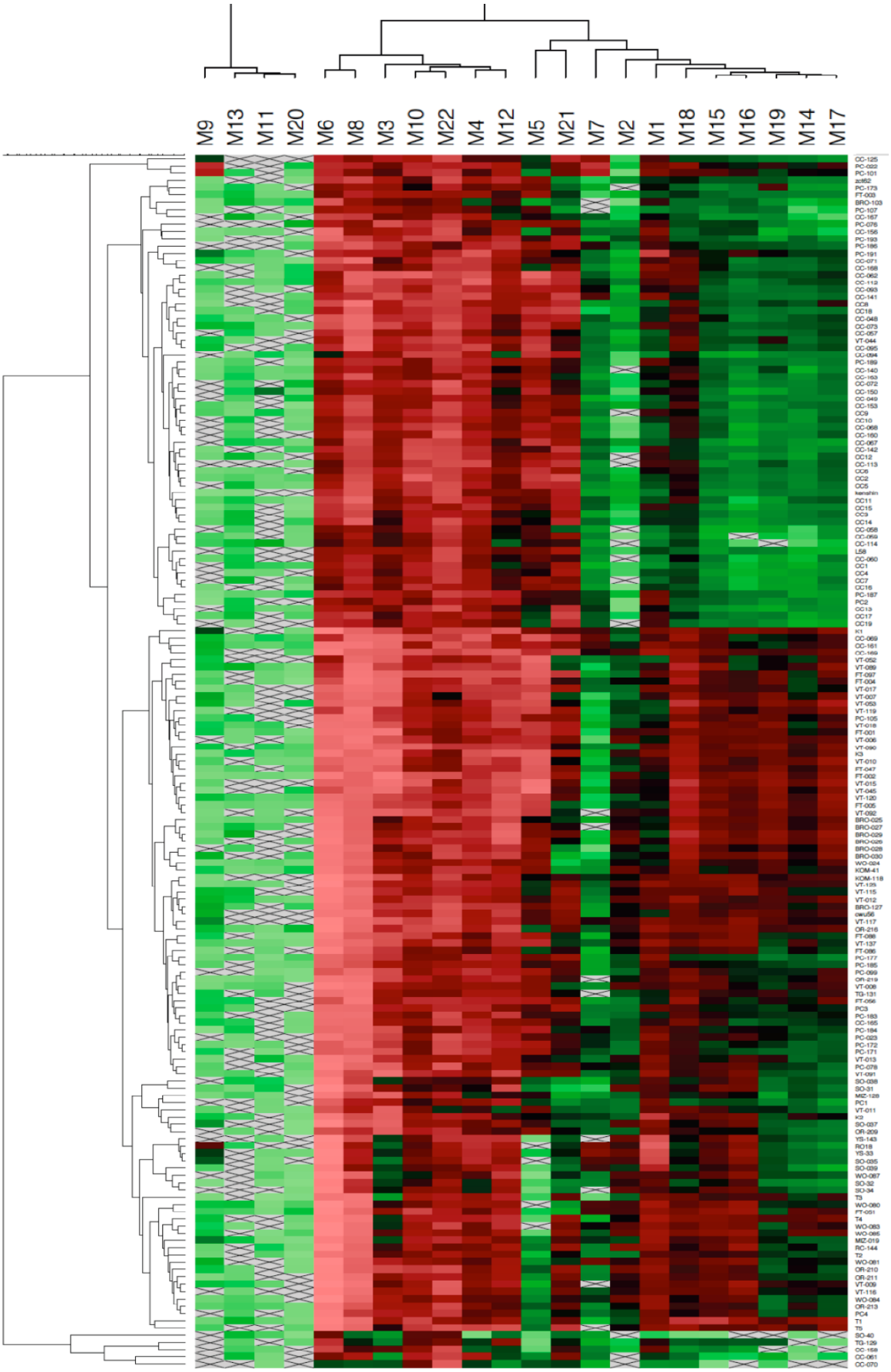
The identification was based on information obtained from Brassica glucosinolates (http://kanaya.naist.jp/knapsack_jsp/top.html) and from glucosinolates that were identified in previous LC-QTOF MS Arabidopsis metabolomics experiments (Beekwilder et al. 2008). We can however not exclude the possibility that we missed glucosinolates, since this analysis was limited to the ones that could be identified.

The highest detection signal over all the glucosinolates was found for 3-butenyl (gluconapin), and this was so across morphotypes.

We used the variation found in the core collection for classification of the accessions. Three different clusters of glucosinolates can be identified based on the variation pattern of the glucosinolate content among accessions, using the log2-transformed values of the LC-MS signals of the identified glucosinolates (Figure 1).

The first group contains the aliphatic glucosinolates 8-methylsulfinyloctyl (glucohirsutin I,II,III) isomers and 4-methylthiobutyl (glucoerucin) , the second group contains the aliphatic glucosinolates : 3-butenyl (gluconapin) , 4-pentenyl (glucobrassicanapin), 2-hydroxybut-3-enyl (progoitrin) and methylsulfinyl pentyl (glucoalyssin), the indole glucosinolates 1-methoxyindol-3-ylmethyl (neoglucobrassicin) and indol-3-ylmethyl (glucobrassicin) and the aromatic glucosinolate 2-phenylethyl (gluconasturtiin); the third group contains the aliphatic glucosinolates hydroxypentenyl, methsulphonylbutyl, 2-propenyl (sinigrin), 4-methylsulfinyl butyl (glucoraphanin) and isomers of hexyl (I,II,III) and heptyl (I,II,III) and the indole glucosinolate 4-Methoxyindol-3-ylmethyl (methglucobrassicin).

Figure 1. Hierarchical cluster of glucosinolate variation in a core collection of 168 accessions.Table indicate corresponding names and abbreviations, * isomere.(M1) 4-methylsulfinyl butyl, (M2) prop-2-enyl , (M3) 2-hydroxybut-3-enyl, (M4) methylsulfinyl pentyl, (M5) hydroxypentenyl, (M6) but-3-enyl, (M7) methsulphonylbutyl, (M8) pent-4-enyl, (M9) 4-methylthiobutyl,(M10)indol-3-ylmethyl, (M11) 8-methylsulfinyloctyl *, (M12) 2-phenylethyl, (M13) 8-methylsulfinyloctyl *,(M14) hexyl*, (M15) hexyl*, (M16) hexyl*, (M17) heptyl*,(M18) heptyl*,(M19) heptyl*, (M20) 8-methylsulfinyloctyl*, (M21) 4-Methoxyindol-3-ylmethyl, (M22) 1-methoxyindol-3-ylmethyl.


M9
M13
M11
M20
M6
M8
M3
M10
M22
M4
M12
M5
M21
M7
M2
M1
M18
M15
M16
M19
M14
M17

The largest differences among accessions were observed for the isomers of the long chained aliphatic glucosinolates, heptyl (I,II,III) and hexyl (I,II,III). The different glucohirsutin isomers and glucoerucin had the highest number of observations with values below the LC-MS detection signal (0.05).

Based on the variation observed among accessions three clusters could be separated: group I included in its majority the Chinese cabbage morphotype, group II included vegetable and fodder turnips, brocolettos and pak choi types; and group III included the oil types (winter and spring oils, turnip rape and yellow sarson accessions) and turnip lines provided by the breeding companies. Separation between the Chinese cabbage (group I) and the groups II and III was determined mostly by the difference in long chain aliphatic glucosinolates heptyl and hexyl, but not glucohirsutin, which was not detectable in many accessions. On the other hand the separation between the group III oil types (winter and spring oils, turnip rape, yellow sarson and turnips of breeding companies) and groups II and I was mostly determined by the difference in the hydroxypentenyl composition. With the exception of the heptyl, hexyl and hydroxypentenyl, the glucosinolates showed a variation among accessions that is not morphotype specific or determined by the geographic origin.

Markers and allelic variation

For the association study we designed primers targeting microsatellites within sequenced BACs and contigs along the linkage group A03. Depending on the type of marker (random or linked to glucosinolate related genes) we had different objectives (Table 1). The group of random markers (KS and WUR) did not aim for the identification of any specific locus. These markers were selected because they are located in different genetic positions across the linkage group A03 and could lead to the identification of additional genomic regions of relevance for the regulation of glucosinolates. On the other hand, the markers coded BAC-MYB28 (1-7) and BAC-MAM (1-2) had known physical distance from the transcription factor *Myb28* and the GLS-*Elong* methylthioalkylmalate genes (*MAM*). Both of these genes mapped under the previously identified major glucosinolate QTL (Lou et al. 2008) (Pino Del Carpio Chapter 5). Although, *Myb29* does not map to this genetic region, we also profiled the markers BAC-MYB29 (1-2) physically linked to this transcription factor, since this transcription factor has been studied for its effect on glucosinolate regulation. In

short, the markers BAC-MYB28, MYB29 and MAM are markers targeting candidate genes.

Together random and targeted markers made a total of 37 genomic microsatellites located along the linkage group A03 (Table 1). After profiling the microsatellites over the 168 accessions we could identify 374 alleles (Supplementary table 2). All the microsatellite markers had at least one allele that had a frequency higher than 10%. From the 374 total number of alleles only 98 alleles had a frequency higher than 10% and the markers BAC-MYB28-6, BAC-MYB28-7 and BAC-MAM3-1 had a major allele with a frequency of 90% or higher over all the accessions.

For our study we used previously reported subgroup classification obtained in STRUCTURE in this same collection (Pino Del Carpio et al. Chapter 3).

The number of alleles varied across markers on different chromosomal positions and between subpopulations (Figure 2). The average number of alleles per microsatellite marker that was found after the screening of 37 microsatellites over the populations was 10.

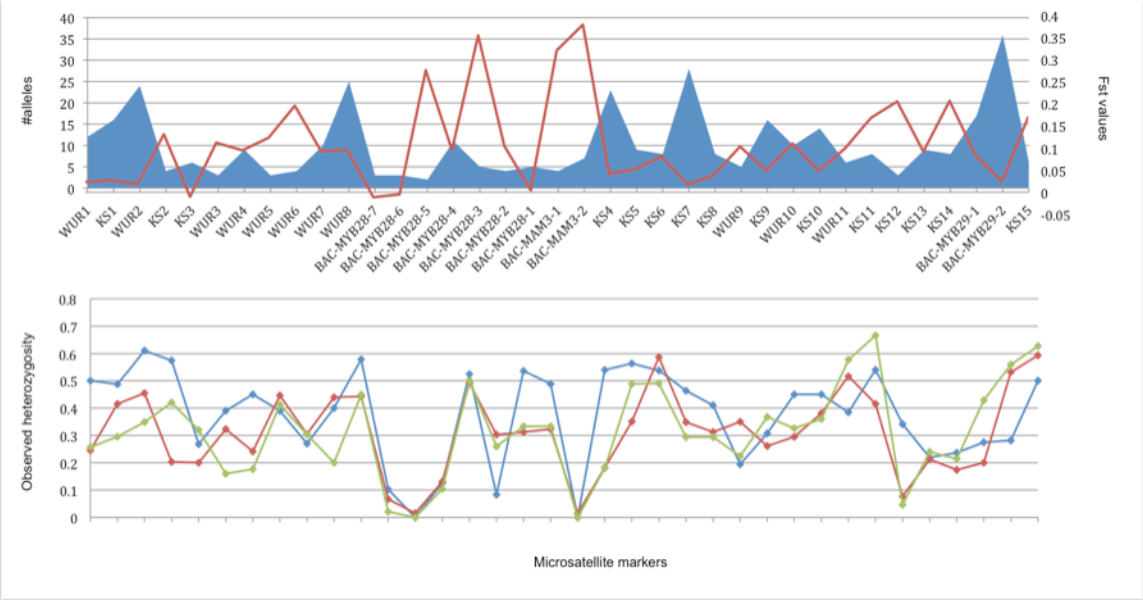


Figure 2. Fst statistics and locus by locus AMOVA. Top image : blue peaks indicate number of alleles, red lines indicate Fst values.Bottom image for observed heterozygosity values the markers are ordered according to their position across linkage group A03 and population: blue (population 1:FT+VT), red (population 2:PC+T) and Green(population 4:CC), population 3(YS+SO+RC) was not included because of small sample size.

Across populations the highest average number of alleles (n=8) was found for populations 1 (PC+T) and 2 (VT+FT) in comparison to the average for the relatively small population 3 (YS+SO+RC) (n=4.1) and population 4(CC) (n=5.7). (Table 2)

Marker	Population 1	Population 2	Population 3	Population 4	Mean	Total number	PIC
WUR1	8	6	5	8	6.75	12	0.6614
KS1	13	12	5	12	10.5	16	0.9185
WUR2	14	17	7	16	13.5	24	0.8771
KS2	4	3	2	3	3	4	0.5422
KS3	5	6	3	4	4.5	6	0.4510
WUR3	3	3	1	2	2.25	3	0.3480
WUR4	8	5	4	2	4.75	9	0.4177
WUR5	3	3	2	3	2.75	3	0.5359
WUR6	4	4	4	4	4	4	0.6920
WUR7	8	8	2	4	5.5	10	0.4445
WUR8	20	17	8	6	12.75	25	0.8243
BAC-MYB28-7	3	3	1	3	2.5	3	0.1044
BAC-MYB28-6	2	2	1	1	1.5	3	0.0181
BAC-MYB28-5	2	2	2	2	2	2	0.3714
BAC-MYB28-4	10	12	4	7	8.25	13	0.7963
BAC-MYB28-3	5	4	1	2	3	5	0.4600
BAC-MYB28-2	4	4	3	2	3.25	4	0.4259
BAC-MYB28-1	4	3	2	2	2.75	5	0.4205
BAC-MAM3-1	3	3	4	1	2.75	4	0.1704
BAC-MAM3-2	7	6	3	3	4.75	7	0.5465
KS4	20	16	6	16	14.5	23	0.9220
KS5	8	7	6	6	6.75	9	0.7555
KS6	8	6	4	5	5.75	8	0.7372
KS7	19	26	10	15	17.5	28	0.9395
KS8	8	7	4	5	6	8	0.6444
WUR9	5	4	3	4	4	5	0.6510
KS9	12	10	8	8	9.5	16	0.7408
WUR10	8	10	5	6	7.25	10	0.8076
KS10	12	13	6	12	10.75	14	0.8374
WUR11	5	6	3	4	4.5	6	0.6314
KS11	7	6	2	4	4.75	8	0.5053
KS12	2	2	3	2	2.25	3	0.3711
KS13	6	9	5	4	6	9	0.5986
KS14	6	5	5	4	5	8	0.5501
BAC-MYB29-1	13	14	5	8	10	17	0.9011
BAC-MYB29-2	27	29	11	19	21.5	36	0.9423
KS15	4	6	3	4	4.25	6	0.6793
Mean	8.108	8.081	4.135	5.757	6.52	10.162	

Table 2.Summary statistics of microsatellites profiled over the core collection. Values are indicated according to subpopulations: Population 1 (FT+VT), population 2 (PC+T), population 3 (YS+SO+RC) and population 4(CC). PIC=polymorphic index content.

The highest number of alleles per locus was found for marker MYB 29-2 (n=36) and the lowest number was found for marker BAC134-MYB28-5 (n=2). Polymorphic information content (PIC) was mostly related to the number of alleles per marker that was found within the core collection; for example, marker BAC-MYB29-2 with 36 alleles had a PIC of 0.9423. Markers developed from BACs in the closest physical

distance to a candidate gene were variable in the PIC values, BAC-MYB28-5 had a PIC value of 0.3714, BAC-MAM3-1 had a PIC value of 0.1704, BAC-MAM3-2 had a PIC value of 0.5465 and BAC-MYB29-1 had a PIC value of 0.9011 .

AMOVA and F statistics

The Analysis of Molecular Variance (AMOVA) is a method of estimating population differentiation directly from molecular data. Testing hypotheses about differentiation within and between subpopulations can indicate different selective pressures among different subpopulations. This selection can have very different effects on different alleles and allele combinations across linkage group A03 and selection pressure will be reflected in lower heterozygosity values for each subpopulation.

Global ANOVA based on the 4 subgroups as identified using STRUCTURE software indicated that the lower molecular variation was found among populations with a percentage value of 11% and the higher molecular variation was found within populations with a percentage value of 89% (Table 3).

Source of Variation	d.f	Sum of squares	variance components	Percentage of variation
Among populations	3	174.85	0.68226 Va	10.99
Within populations	332	1835.022	5.52717 Vb	89.01
total	335	2009.872	6.20944	
Fixation Index	Fst: 0.10987			

Table 3. AMOVA results between and within population.

In accordance to this classification the Fst values obtained based on the analysis of microsatellite data of linkage group A03 clearly showed the differentiation between populations (Table 4). The largest value (Fst=0.23), which in turn signifies an Fst pairwise value for greatly differentiated populations, was found after comparison between population 3 (YS+SO+RC) and 4 (CC). The second highest values, which also correspond to greatly differentiated populations, were found after comparison between population 1 (VT+FT) and 3 (YS+SO+RC) (Fst=0.16), population 1 (VT+FT) and 4 (CC) (Fst=0.15) and population 2 (PC+T) and 3 (YS+SO+RC) (Fst=0.15). Values for moderately differentiated populations were found in the Fst pairwise comparison between populations 1(PC+T) and 2 (VT+FT) (Fst=0.08) and the comparison between populations 2 (VT+FT) and 4 (CC) (Fst=0.07).

Distance method: Nr of different alleles (Fst)

	Population 1	Population 2	Population 3	Population 4
Population 1	0			
Population 2	0.07992	0		
Population 3	0.15729	0.15247	0	
Population 4	0.15154	0.06974	0.22952	0

Table 4. Fst values compared across subpopulations: Population 1 (VT+FT), population 2 (PC+T) , population 3 (YS+SO+RC) and population 4(CC)

Across the linkage group A03 the Fst values varied greatly between genomic regions ranging from -0.01116 for the marker BAC-MYB28-7 at a distance of 79.7 kb from candidate gene Myb28 to 0.37886 for the marker BAC-MAM3-2 at a distance of 1.5 kb from candidate gene MAM (Fig 2 and Supplementary Figure 1).

The observed heterozygosity varied along different positions on the linkage group A03 in correspondence to the Fst values across subpopulations. The largest variation in heterozygosity values per marker and between subpopulations was found at the position around the MAM locus at the BAC-MAM3-2 marker with values of 0.54 for population 1 (VT and FT), 0.18 for population 2 (PC+T), and 0.18 for population 4 (CC).

Around *Myb28*, a reduced number of alleles was observed in particular between the region from BAC-MYB28-7 to BAC-MYB28-5. The observed heterozygosity in this region was also reduced in comparison to other genomic regions along the linkage group A03 and it followed the same pattern in subpopulations 1, 2 and 4 (Figure 2).

Around the MAM locus the observed heterozygosity was also reduced as observed for markers BAC-MAM3-1 in all populations and BAC-MAM3-2 for populations 1 (VT+FT) and 4 (CC). However, population 2 (PC+T) followed no significant reduction at the position of marker BAC-MAM3-2 in comparison to populations 1 and 4. Interestingly, the peaks for Fst values were found at the position of the marker BAC-MYB28-5, the closest marker physically to *Myb28* gene at a distance of 12.9 kb and BAC-MYB-28-3 at a distance of 68.3 kb of the *Myb28* gene and at the position of the marker BAC-MAM3-2 at a distance of 1.5 kb of one of the triplicated copies of MAM (Figure 2 and supplementary Figure 1).

Association study of glucosinolate variation

Although we included glucosinolates (GLS) of different types in the analysis (aliphatic, indolic and one aromatic), significant association results were only found for the aliphatic glucosinolates: 3-butenyl (gluconapin), 8-methylsulfinyloctyl (glucohirsutin), hexyl, heptyl, hydroxypentenyl; the ratios for aliphatic glucosinolate hydroxylation: 2H3B/2H3B+3B (rat1) and H4P/H4P+4P (rat2); the sum of alkenyl glucosinolates: 2Pr+3B+4P (totALK); and the total content of 4C (tot4C) and 6C (tot6C) aliphatic glucosinolates (Table 6).

The distribution of the glucosinolate-marker associations was not limited to a single locus on A03, but it was found over different loci along linkage group A03.

A total of nine out of the 33 microsatellites included in the analysis had at least one allele associated to a single glucosinolate, a glucosinolate ratio or total glucosinolate content.

The maximum number of alleles found to be significantly associated to one or more glucosinolates was two for the BAC-MYB28-5, WUR9 and KS14 markers.

We denominated the identified marker-glucosinolate associations as linkage disequilibrium QTL ($_{LD}$QTL) in further descriptions.

The chromosomal regions with clusters of more than two associated aliphatic GLS, GLS ratios (1-4), GLS of different chain size 3C-8C or total aliphatic content were considered as major $_{LD}$QTL genomic regions. Five major$_{LD}$QTL were identified and within these regions candidate genes for glucosinolate biosynthesis could be assigned (Table 6).

Table 6. Association results as obtained in Tassel. Indicated are QTL regions, candidate genes, markers and adjusted p-values for 5000 permutations (p_adj_Marker) together with explained variation (Rsq_Marker)

Name/description	Scientific name	Abbreviation	QTL	candidate gene	marker	p-perm_Marker	p-adj_Marker	Rsq_Marker
Glucohirsutin	8-methylsulfinyloctyl glucosinolate *	8MSO			WUR1	0.0016	0.0224	0.0728
	Hexyl*	HXL			WUR7	0.0014	0.0262	0.0414
Gluconapin	but-3-enyl glucosinolate	3B	**LDQTL1**		BAC-MYB28-5	2.00E-04	4.00E-04	0.0462
2H3B/2H3B+3B		rat1			BAC-MYB28-5	0.0014	0.0308	0.048
4C		tot4C		*Myb28*	BAC-MYB28-5	0.0016	0.0182	0.0338
	Hexyl*	HXL			BAC-MYB28-5	0.0018	0.0206	0.0399
6C		tot6C			BAC-MYB28-5	0.0022	0.0458	0.035
	Heptyl*	HPL			BAC-MYB28-4	6.00E-04	0.0018	0.0394
	Heptyl*	HPL	**LDQTL2**	*MAM*	BAC-MAM3-1	0.0028	0.0464	0.0305
	hydroxypentenyl	H4P			BAC-MAM3-2	8.00E-04	0.0014	0.0627
Gluconapin	but-3-enyl glucosinolate	3B	**LDQTL3**		KS6	0.0014	0.0112	0.0354
Talk		totALK		*AOP*	KS6	0.0022	0.0396	0.0345
TOTAL 4C		tot4C			KS6	0.0024	0.038	0.031
Gluconapin	but-3-enyl glucosinolate	3B	**LDQTL4**		WUR9	6.00E-04	8.00E-04	0.0439
pro/pro+nap		rat1			WUR9	8.00E-04	0.0012	0.0654
pro/pro+nap		rat1		*GS-OH*	WUR9	6.00E-04	2.00E-04	0.0752
Talk		totALK			WUR9	1.00E-03	0.0082	0.0422
hxp/gbn+hxp		rat2			WUR9	0.002	0.0488	0.0476
Gluconapin	but-3-enyl glucosinolate	3B			KS13	0.0014	0.0146	0.0395
Gluconapin	but-3-enyl glucosinolate	3B	**LDQTL5**		KS14	8.00E-04	0.0058	0.0372
Talk		totALK		*Myb29*	KS14	1.00E-03	0.005	0.0428
Gluconapin	but-3-enyl glucosinolate	3B			KS14	1.00E-03	0.0074	0.0371

The marker BAC-MYB28-5 ($_{LD}$QTL1), which is a microsatellite located at a 12 kb distance from the *Myb2*8 gene, was found to be associated to 3-butenyl glucosinolate (3B), hexyl (HXL), 6C glucosinolates (tot6C), 4C glucosinolates (tot4C) and the hydroxylation ratio (rat1). The marker BAC-MAM (1-2) ($_{LD}$QTL2), which is a microsatellite located within the BAC containing the *MAM* gene was found to be associated to heptyl (HPL) and hydroxypentenyl (H4P). Although the markers BAC-MYB29-1 and BAC-MYB29-2 located at a distance of only 42.9 kb and 33.5 kb respectively of *Myb29*, neither of them showed association to glucosinolates. However, KS14 ($_{LD}$QTL5), which is >50kb from *Myb29* was associated to but-3-enyl glucosinolate (3B) and to the total amount of alkenyl glucosinolates (totALK).

Two additional regions were found to be associated to glucosinolates; these SSR markers did not physically map on BACs with a candidate gene but were genetically linked to BACs with candidate genes. For example, marker KS6 ($_{LD}$QTL3), with a genetic distance of 2.7 cM of the BAC containing the AOP gene, was found to be associated to but-3-enyl glucosinolate (3B), the total content of 4C glucosinolates (tot4C) and of alkenyl glucosinolates (totALK) and marker N6 ($_{LD}$QTL4), a microsatellite marker of a BAC from the same scaffold as the locus with GS-OH genes, was found to be associated to but-3-enyl glucosinolate (3B) the hydroxylation ratios 1(rat1) and 2 (rat2) and the total content of alkenyl glucosinolates (totALK).

The total explained variation of the associated markers showed a range between 3.05% for the marker BAC-MAM ($_{LD}$QTL3) allele 1, associated to an heptyl glucosinolate, to 7.52% for the marker WUR9 allele 3, associated to the hydroxylation ratio 1 (rat1).

Discussion

In *Brassica rapa* several QTLs for aliphatic glucosinolates were previously identified in two doubled haploid populations (Lou et al. 2008). The major QTL within this study was found at the bottom of linkage group A03 in both populations and in two seasons, while a second minor QTL was located in the middle region of the same linkage group in only one population in two seasons. Based on genome synteny and recent genome sequence information (Korea and China) a possible role in regulation of glucosinolates for chain elongation *MAM* genes and the transcription factor *Myb28*,

mapping under the major QTL was predicted. The *AOP* genes were predicted to map within the genomic region at the minor QTL in the middle of linkage group A03.

Because this QTL study was limited to the variation found in two biparental crosses with a yellow sarson, a pak choi and a turnip parent, we decided to extent the study of glucosinolate content and variation to a collection of accessions of different morphotype and origin.

The accessions from the core collection showed differences based on their glucosinolate contents, which demonstrates the existence of enough variation for the detection of genomic regions involved in genetic regulation of glucosinolate biosynthesis. Furthermore, because of the diversity in the glucosinolate biochemical structures, the possibility exists to study the variation at different levels of glucosinolate biosynthesis, like chain elongation, synthesis of the core structure and side chain modification (i.e oxidation and hydroxylation). These quantitative and qualitative differences are expected to be due to the allelic variation in structural and regulatory genes of the glucosinolate pathway among accessions.

From the 22 glucosinolates that were detected, 3-butenyl (gluconapin) had the highest detection signal over all the glucosinolates. Because 4-methylsulfinyl butyl (glucoraphanin) is converted into 3-butenyl (gluconapin) by *GLS-ALK* (*AOP*) most of the accessions with high 3-butenyl (gluconapin) concentration had a reduced concentration of 4-methylsulfinyl butyl (glucoraphanin) and viceversa.

Within the scope of the present study we decided to further study, through an association mapping approach, the genomic regions leading to the glucosinolate variation in linkage group A03 with particular focus on the major QTL locus found in the DH38 population. Available sequence information (Personal communication Mina Jin Korea, Dr. Xiaowu Wang, IVF CAAS) led us to identify the presence of a copy of a *Myb28* gene at a distance of 240kb from a triplicate copy of *MAM* in the region of interest. Furthermore, a microsatellite marker developed from the BAC containing *Myb28* was mapped in the DH38 population within the major QTL region (Pino Del Carpio et al Chapter 5).

The three markers BAC-MYB28-5, BAC-MAM3-1 and BAC-MAM3-1, which had the closest known physical distance from the two candidate genes *Myb28* and *MAM* showed association to aliphatic glucosinolates. Interestingly, the *Myb28* gene appears to be a major regulator of the variation of glucosinolate content in *Brassica rapa,* showing association to five glucosinolate related-traits ($_{LD}$QTL1). This regulation is

reflected at different levels; for example related to the content of 3-butenyl (gluconapin), the glucosinolate with the highest content in this collection, and the total content of 4C (tot4C) and 6C (tot6C) glucosinolates. Additionally, the association to the ratio of hydroxylation (rat1) reflects possible regulation of the *GLS-OH* locus. On the other hand, the *MAM* markers only showed association to heptyl and hydroxypentenyl ($_{LD}$QTL2), being in this context of less abundance in the overall variation of glucosinolates in *B. rapa*.

One of the most important advantages of the association mapping approach is the possibility to find additional genomic locations related to a trait in comparison to biparental crosses, which are limited in the allelic variation of a cross (Abdurakhmonov et al. 2008). In our study genomic regions, which had not been previously reported, were found to be associated to hydroxylation ratios calculated with the glucosinolate data. The marker (WUR9) in the vicinity of the *GLS-OH* locus, which regulates the hydroxylation of alkenyl glucosinolates (Halkier and Du 1997) showed association to the hydroxylation of 3-butenyl (gluconapin) (rat1) to 2-hydroxybut-3-enyl (progoitrin) and pent-4-enyl (glucobrassicanapin) to hydroxypentenyl (rat2) ($_{LD}$QTL4).

The *GLS-ALK* locus (*AOP*) is involved in the production of alkenyl homologues by removal of the methylthio group followed by the insertion of a double bond (Halkier and Du 1997). In our study KS6 ($_{LD}$QTL3), with a genetic distance of 2.7 cM of the BAC containing the AOP gene showed association to the total content of alkenyl glucosinolates (totALK), the content of 4C glucosinolates (tot4C) and 3-butenyl (gluconapin). The association to the total alkenyl glucosinolates is in accordance to the function of this locus and the association to 3-butenyl (gluconapin) is in correspondence to the minor QTL previously found in two DH populations in *Brassica rapa* and to other studies in Arabidopsis and Brassicas (Lou et al. 2008, Kliebenstein et al. 2001a,b, Gao et al. 2004, Li et al. 2003).

An example of the extent of linkage disequilibrium in this core collection is the presence of the association between KS14 ($_{LD}$QTL5) to but-3-enyl glucosinolate (gluconapin) and total alkenyl glucosinolates (totALK). This particular marker is >50kb to the closest candidate gene *Myb29*; apparently the extent of LD in this region is larger compared to that around the *Myb28* locus which shows the most significant marker-trait association up to a distance of only 13 kb (BAC-MYB28-5), which was the closest SSR linked to *Myb28*.

In Arabidopsis an important group of candidate genes for glucosinolate variation and regulation have previously been found through gene expression and metabolite QTL analyses (Wentzell et al. 2007,Kliebenstein et al 2001a,b). These analyses showed at a population level that variation at *MAM* (*GLS-Elong*) and *AOP* (*GLS-ALK*) controlled the accumulation of glucosinolates and transcripts related to these metabolites. More recently, through Omics-based approaches *Myb28* and *Myb29* have been discovered as transcription factor genes involved in the regulation of aliphatic GSL production (Hirai et al. 2007).

The *B. rapa* genome is sequenced (China and Korea) and with available sequence information several candidate genes for glucosinolate regulation have been recently identified (Zang et al. 2009). However, these candidate genes have not been identified as loci underlying QTL and their functions have not been validated in *B. rapa*.

In the present study we screened microsatellites developed from BACs and sequenced contigs containing the candidate genes *MAM* and *Myb28* (Korea (www.brassica-rapa.org) and China (Dr. Wang Xiaowu, IVF CAAS, Beijing) located within the genomic region of the major QTL on A03 (Lou et al. 2008) and demonstrated their association to several aliphatic glucosinolates in a *B. rapa* core collection.

In general, the resolution with which a QTL can be mapped is a function of how quickly linkage disequilibrium (LD) decays over distance. *Brassica rapa* is an outcrosser and conversely in outcrossers LD generally breaks down more rapidly like for example in maize (Remington et al. 2001). In the present study we showed that LD varied among loci in A03 in the vicinities of known candidate genes, which goes from a distance of 0.5kb (BAC-MAM3-1) to 13 kb (BAC-MYB28-5) or even over 50kb near *Myb29*. This variation can be due to the different mutation rate of SSRs if compared to single nucleotide polymorphisms (SNPs) and insertions or deletions (InDels) that may have caused variation in the causal genes. These differences can affect the identification of LD even when a marker is associated to the gene at a very close range; on the other hand population subdivision and the allelic diversity in the different sub populations can increase LD around a gene and affect the rate of false positive results (Pritchard and Przeworski, 2001).

In conclusion, following an association mapping approach we showed the importance of *Myb28* as the major regulator of the aliphatic glucosinolate accumulation of different types and ratios as well as the effect of the *MAM* genes within this *B. rapa* core collection. More importantly, the screening of markers along linkage group A03

helped us to identify new genomic regions that could be sources of variation for glucosinolate regulation. More work is needed with markers closely linked to candidate genes to confirm our assumptions that *GLS-OH* and *AOP* are within this important genomic region.

Acknowledgements

We thank Liu Nini for her laboratory work on designing primers and screening microsatellites. We thank Alejandra Freire Rios for her work on the analysis of the data.This research was funded by the IOP Genomics project “Brassica vegetable nutrigenomics” IGE 05010.

Supplementary Table 1. A.Identified glucosinolates with common name, chemical name, abbreviation and description.B.Additionally values for ratios and total calculations for marker-trait associations are described.* isoforms

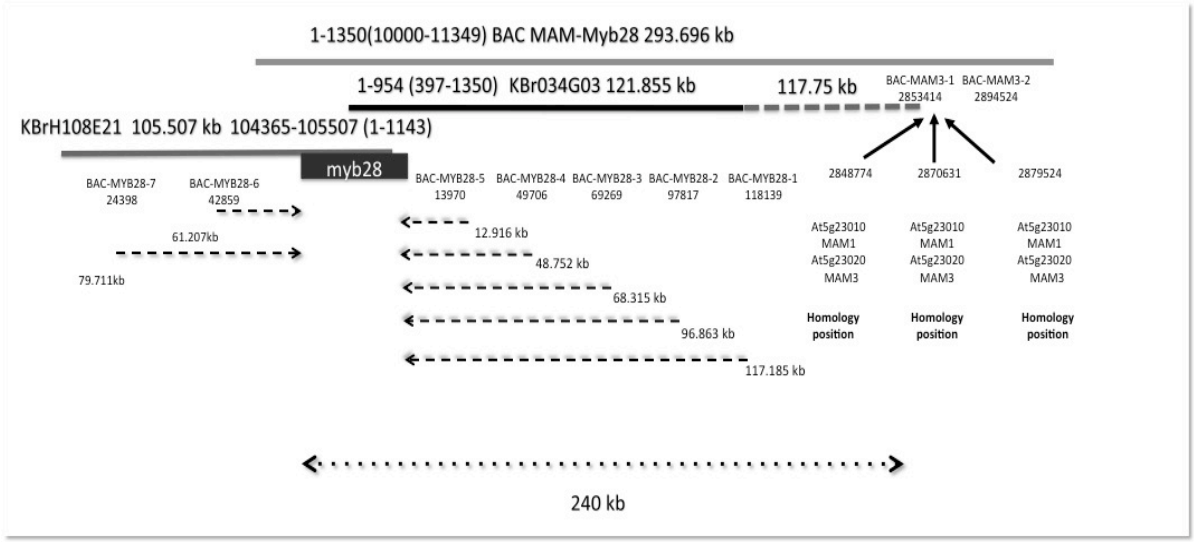
A.

common name	Abbreviation	scientific name	description
Sinigrin	2Pr	prop-2-enyl	double bond straight aliphatic
Gluconapin	3B	but-3-enyl	double bond straight aliphatic
Glucobrassicanapin	4P	pent-4-enyl	double bond straight aliphatic
	HXL	Hexyl *	aliphatic straight chain
	HPL	Heptyl *	aliphatic straight chain
Progoitrin	2H3B	2-hydroxybut-3-enyl	double bond hydroxy straight aliphatic
	H4P	hydroxypentenyl	double bond hydroxy straight aliphatic
Glucoraphanin	4MSB	4-methylsulfinyl butyl	sulphur containing side chain aliphatic
Glucoalyssin	MSP	methylsulfinyl pentyl	sulphur containing side chain aliphatic
	MSOB	methsulphonylbutyl	sulphur containing side chain aliphatic
Glucoerucin	4MTB	4-methylthiobutyl	sulphur containing side chain aliphatic
Glucohirsutin	9MSO	9-methylsulfinyloctyl *	sulphur containing side chain aliphatic
Methglucobrassicin	4MOI3M	4-Methoxyindol-3-ylmethyl	Indole
Neoglucobrassicin	1MOI3M	1-methoxyindol-3-ylmethyl	Indole
Glucobrassicin	I3M	indol-3-ylmethyl	Indole
Gluconasturiin	2PE	2-phenylethyl	Aromatic

B.

Additional traits for association	
tot3C	2Pr (3C)
tot4C	H3B+4MSB+MSOB+4MTB
tot5C	4P+H4P+MSP (5C)
tot6C	HXL(3) (6C)
tot7C	HPL(3) (7C)
tot8C	9MSO(3)(8C)
R3	C4+C5+C6+C7+C8/C3
1/R3	1/C3
R4	C5+C6+C7+C8/C4
R5	C6+C7+C8/C5
R6	C7+C8/C6
R7	C8/C7
rat1	2H3B/2H3B+3B
rat2	H4P/H4P+4P
rat3	3B/3B+4MSB
rat4	4B/4B+MSP
totALK	2Pr+3B+4P

Supplementary Figure 1. Graphical representation of marker position and physical distance to the targeted candidate genes *Myb28* and MAM.The KBr code indicate sequence information from Korea sequencing project, the BAC *MAM-Myb28* is sequence information from China sequencing project



Marker	Allele	Frequency
WUR1	100	0.5000
WUR1	120	0.1641
WUR1	140	0.1641
KS1	240	0.1560
KS1	160	0.1525
KS1	180	0.1418
KS1	140	0.1383
WUR2	315	0.1392
WUR2	319	0.1076
WUR2	323	0.1013
KS2	297	0.5583
KS2	307	0.2117
KS2	303	0.2055
KS3	100	0.5693
KS3	110	0.3675
WUR3	234	0.7349
WUR3	243	0.2289
WUR4	236	0.7298
WUR4	228	0.1180
WUR5	235	0.4940
WUR5	238	0.3413
WUR5	242	0.1647
WUR6	300	0.3133
WUR6	260	0.2733
WUR6	380	0.2333
WUR6	310	0.1800
WUR7	229	0.7048
WUR7	243	0.1596
WUR8	220	0.2975
WUR8	212	0.1772
WUR8	222	0.1329
WUR8	218	0.1139
BAC-MYB28-1	275	0.5370
BAC-MYB28-1	269	0.4290
BAC-MYB28-2	208	0.4939
BAC-MYB28-2	206	0.4697
BAC-MYB28-3	205	0.5484
BAC-MYB28-3	206	0.3806
BAC-MYB28-4	98	0.2987
BAC-MYB28-4	96	0.2201
BAC-MYB28-4	97	0.1289
BAC-MYB28-4	94	0.1195
BAC-MYB28-5	219	0.5597
BAC-MYB28-5	217	0.4403
BAC-MAM3-1	100	0.9048
BAC-MAM3-2	100	0.6049
BAC-MAM3-2	120	0.1759
BAC-MAM3-2	130	0.1265
KS4	116	0.1224
KS4	114	0.1020
KS5	100	0.3106
KS5	110	0.2671
KS5	111	0.1584
KS5	115	0.1149
KS6	104	0.3482
KS6	106	0.2440
KS6	105	0.1696
KS6	101	0.1101
KS7	125	0.1189
KS8	101	0.5162
KS8	103	0.1526
KS8	104	0.1429
KS8	102	0.1039
WUR9	102	0.3704
WUR9	104	0.2870
WUR9	103	0.2593
KS9	100	0.3974
KS9	113	0.2147
KS9	114	0.1442
WUR10	107	0.2699
WUR10	106	0.2086
WUR10	105	0.1687
WUR10	109	0.1135
KS10	100	0.2955
KS10	113	0.1396
KS10	110	0.1201
WUR11	104	0.4049
WUR11	105	0.3620
WUR11	103	0.1043
KS11	100	0.6614
KS11	110	0.1203
KS11	115	0.1076
KS12	100	0.6078
KS12	110	0.3862
KS13-1	100	0.5889
KS13-1	111	0.1259
KS14	100	0.5518
KS14	110	0.2805
BAC-MYB29-1	223	0.1584
BAC-MYB29-1	222	0.1366
BAC-MYB29-1	219	0.1149
KS15	102	0.4672
KS15	104	0.1898
KS15	101	0.1241
KS15	103	0.1058

Supplementary table 2. Microsatellites alleles with frequency > %10 included in the association analysis

Chapter 5

The genetics of the *Brassica rapa* metabolome

Dunia Pino Del Carpio, Ram Kumar Basnet, Danny Arends, Frank Johannes, Ric CH De Vos, Jan Kodde ,Kim Boutilier, Johan Bucher, Ritsert Jansen, Richard Visser, Guusje Bonnema

Abstract

In the present study we followed a genetical genomics approach to identify candidate genes for six biosynthetic pathways: carotenoids, tocopherols, folates, glucosinolates, flavonoids and phenylpropanoids, based on the co-localization of metabolic QTLs and expression QTLs. A Doubled Haploid population was profiled for metabolite content and variation through targeted and LC-MS untargeted approaches. Additionally, the same population was profiled for transcript variation with a newly developed microarray assembled using EST sequences mainly from three species: *B. napus, B. rapa* and *B. oleracea*. Co-localization of mQTLs and eQTLs lead to successful identification of candidate genes for carotenoids, tocopherols and glucosinolates. Using the glucosinolates pathway as model pathway the results revealed the co-localization of eQTLs of a cluster of co-regulated genes and mQTLs for short (3C-5C) chain aliphatic glucosinolates with modified side chains around *AOP* in linkage group A09 and the co-localization of eQTLs for *MAM* genes and mQTL for long chained aliphatic glucosinolates in A03. On the other hand, further work is still needed to identify candidate genes for mQTLs found in A07 for flavonoids. The application of this type of studies in *Brassica rapa* and the future validation approaches for the identification of *cis* and *trans* regulation with the soon available *Brassica rapa* genome sequence are discussed.

Introduction

With the advent of modern techniques for global phenotypic and genotypic profiling, the generation of huge datasets has increased the opportunity to dissect the genetics underlying complex traits.

The dissection of the genetic regulation of a trait initiates with the collection of phenotypic data from a mapping population. Later on with the aid of molecular markers, as tags mapped across the genome, and through statistical analysis (QTL mapping) regulatory genomic regions can be identified. The ultimate goal of the QTL mapping is to determine which genes are responsible for the variation in a group of selected traits (Mackay 2001).

In recent years, breeding for nutritional quality became an important research topic and in this context metabolomics approaches have enabled the parallel assessment of the levels of a broad range of metabolites (Fernie et al. 2009, Verpoorte et al. 2008, Rowe et al. 2008). In Arabidopsis the metabolite variation was found to be abundant and its genetic regulation complex, plausible candidate regulators could be identified after LC-MS mass peaks were assigned to genomic loci (Keurentjes et al. 2006)

The use of transcriptomics, which measures the variation in mRNA transcript abundance, has been recorded across populations in plants and expression profiles can be treated as heritable traits to map expression quantitative trait loci (eQTL); this type of analysis has been denominated as genetical genomics (Jansen and Nap 2001).

Thus, with the integration of metabolomics with other genomic platforms it has been possible to identify candidate genes, which are correlated to the levels of metabolites in plant systems (Goossens et al. 2003, Hirai et al. 2005, Keurentjes et al. 2006). Furthermore, the investigation of selected biochemical pathways of pre-defined metabolites showed that the connections between gene expression and metabolite variation are complex (Wentzell et al. 2007, Kliebenstein et al. 2006).

Brassica rapa is an important source of vegetables. The variation in morphology is huge (oil, turnip,pak choi, Chinese cabbage and several Asian morphotypes) and similarly the variation in metabolite composition is large (Chapter 2). This variation has increased the interest of plant breeders to breed for phytonutrient quality in Brassica. *B.rapa* is a close relative to Arabidopsis, its triplicated genome has a well described genome synteny with Arabidopsis (Parkin et al.2005, Schranz et al. 2006). As a consequence of an evolutionary triplication event many genes have paralogues.

The triplicated nature of the Brassica genome, and the fact that at the moment of this study the genome sequence was not yet known, represented a challenge for the genetical genomics approach to unravel the genetics of metabolic traits

For the present research, we had to face several considerations that could affect the outcome of the QTL study. In principle the choice of parental lines that could harbor enough variation in metabolite composition, and the choice of population, in which the number of recombinations can affect the mapping resolution and statistical power of the study (Doerge 2001). Other considerations like segregation distortion and phenotypic distribution are of relevance for statistical analysis and in particular if metabolite data is collected because different algorithms have to be applied to map the underlying variation (Fu et al. 2007, Broman 2003).

An important step in a QTL mapping study is the selection of parental lines that contrast for the phenotype of interest. This statement also applies if we want to map expression quantitative trait loci (eQTL). However, if parental transcript variation would be used exclusively to select differentially expressed genes for subsequent studies, many informative genes would end up being overlooked because of transgressive segregation in the progeny (West et al. 2007, Keurentjes et al. 2007).

In the present study we performed a metabolic and transcript profiling of leaves of six week old plants from a Doubled Haploid (DH) population developed from a F1 cross between a yellow sarson (R500) and a pak choi (PC175) type. We applied an untargeted metabolomics approach using liquid chromatography-mass spectrometry (LC-MS) and a targeted approach to identify isoprenoids (carotenoids and tocopherols) and folates. Additionally, the whole genome transcript level was performed on all DH lines using a distant pair design with a newly developed 60-mer oligo microarray assembled using EST sequences mainly from three species: *B. napus, B. rapa* and *B. oleracea* (Trick et al. 2009). To prioritize on a number of candidate genes we used data of known biochemical pathways of phytonutrient metabolites. Although we narrowed our search within this group of genes we could find ample correlations between transcript abundance and metabolite level. These data are an important step to gain insight in the genetic factors responsible for the metabolite variation in *B. rapa*. Further work is needed to integrate our results with physical maps to correctly identify the genes that regulate either in cis or trans the identified eQTLs. With the complete *B. rapa* sequence expected to be available in the summer of 2010 this becomes a realistic option.

Materials and Methods

Parental materials to develop a double haploid population

A *Brassica rapa* Doubled Haploid (DH) population was developed from a cross between pak choi PC-175 (cv.Nai Bai Cai: accession number VO2B0226) as the male parent to the accession yellow sarson YS-143(accession number FIL500). The parental accessions were selected based on their differences in phenotypic characteristics and genetic distance (Zhao et al. 2005, Lou et al. 2008). Furthermore, this population is a reciprocal cross of the previously developed population DH 38 as described by Lou et al. (2008).

The Doubled Haploid (DH) population was created using the microspore culture protocol described in Lou et al. (2008), based on Coventry et al. (1988) and Custers et al. (1994, 2001). The progeny of the DH plants from three F1 plants were used for the phenotyping and genotyping. The resulting population was named DH68 and consisted of 92 DH lines and for each line the corresponding F1 parent was known.

Plant growth conditions

The seeds of the DH lines were sown in Jiffy pots in the greenhouse under the following conditions: 16 hrs light and temperature between 18 C° and 21 C°. After a week germinated seedlings were transplanted and randomly distributed over three different blocks. Five weeks after transplanting, the 3rd and 4th leaves of each replicate were collected and placed in liquid nitrogen to be further grinded and stored at -70 C°. Each replicate was grinded individually and the mix with equally weighted amount of the three replicates was used for metabolic and transcriptomic profiling and DNA marker profiling for the construction of a linkage map and QTL analysis.

LC-MS metabolic profiling

Brassica leaf samples were analyzed for variation in semi-polar metabolite composition using LC-QTOF MS, essentially as described in De Vos et al. (2007). In short, 0.5 g FW of frozen leaf powder, from one individual plant per accession, was weighed in a 10 ml glass tube and extracted with 1.5 ml of methanol containing 0.1% formic acid. Samples were sonicated and then filtered (Captiva 0.45 µm PTFE filter

plate, Ansys Technologies) into 96-well plates with 700µl glass inserts (Waters) using a TECAN Genesis Workstation equipped with a 4-channel pipetting robot and a TeVacS 96-wells filtration unit. Samples were injected (5 µl) using an Alliance 2795 HT instrument (Waters), separated on a Phenomenex Luna C18 (2) column (2.0x 150 mm, 3 mm particle size) using a 5-35% acetonitrile gradient in water (acidified with 0.1% formic acid) and then detected on-line firstly by a Waters photodiode array detector (wavelength 220-600nm (Waters) and secondly by a Water-Micromass QTOF Ultima MS with negative electrospray ionization (m/z 80-1500).

Metalign software (www.metalign.nl) was used to automatically extract and align all relevant mass signals (signal to local noise ratio > 3) from the raw data files. A total of 6,673 mass signals was filtered for signals being present in at least 15 samples and having amplitudes of at least 100 (about 6 times the noise value) in at least one of the samples. Then, mass signals originating from the same metabolites were clustered based on their similar retention times and variation over samples, using the in-house developed Metabolite Mass Spectral Reconstruction script (Tikunov et al. 2006). This mass signal clustering retained 228 so-called centrotypes, in which each centrotype represents a unique metabolite.

RNA isolation

Total RNA was extracted using the TRIZOL reagent (Invitrogen) starting with approximately 300 mg of frozen leaf material. RNA concentration and purity were quantified with Nanodrop measurements and the quality of the total RNA was checked on a 1% RNase free agarose gel.

Total RNA (5ul) was treated with the DNase I Amplification Grade kit (Invitrogen) for digestion of single and double stranded DNA according to manufacturer's intructions.

Total RNA was cleaned using the RNeasy Mini Kit (Qiagen) starting with the 100 µl of DNase I treated RNA.The concentration of the cleaned RNA was measured and the samples were diluted with nuclease free water (Qiagen) to 400 ng/µl in a total volume of 10µl

Microarray design

The distant pair design proposed for two colour microarrays experiments by Fu and Jansen (2006) was followed as implemented in the R package designGG (http://gbic.biol.rug.nl/designGG/).

The design uses genetic marker information to identify pairs of individuals with maximum dissimilarity across the mapping population and improves the efficiency of eQTL studies. In our study we used information obtained from 48 pairs of DH lines and the information on parental lines was additionally hybridized in two microarrays with dye swap of Cy3 and Cy5.

QTL mapping analyses

QTL analysis was performed using the basic single marker regression procedure present in R/qtl. This was done for both the expression ratio values and the metabolite datasets in a similar fashion, leading to results that could be easily combined in the end. A total of 78,688 expression probes together with 228 centrotypes (summarizing 2,157 mass peaks from the LCMS analysis) were mapped back to the genetic map of *B. rapa* using the basic model. The expressions were measured using two-color array technology and for the mapping we used the ratio's between two genotypes

$Y_i = \alpha + \beta G_i + \text{Error}$

(Y_i = Probe intensity, G_i =Genetic effect)

In this model the genetic effect was annotated for the expression ratio's as described in Fu & Jansen (2006); β is the effect of the different allele (1 for A>B 0 for A==B and -1 A<B). This model was evaluated at each marker to get an estimate of the allelic effect on the expression probes. This results in a P-value, which was transformed into a LOD score. These scores where then visualized in three different ways to show underlying genetic architecture: Using QTL profile plots, circleplots and heatmaps.

Further analysis was composed of permutation of the data to get estimates of significance thresholds.

To reconstruct the metabolic network based on the annotated glucosinolate data from the total LC-MS output we followed the MetaNetwork computational protocol as described in Fu et al. (2007). Significant second order correlations were plotted using Cytoscape with the generated network file (network.sif) and the edge-attributable file (network.eda). Additionally, the output QTL profiles (-log10P significance values plotted at marker positions along the genome) were ploted for visualization for all the annotated metabolites (folate and isoprenoids) and 14 centrotypes of the LC-MS output.

Results

Construction of a genetic linkage map

A genetic linkage map was constructed for population DH 68. A total of 247 markers were mapped in the DH population (Figure 1). The total map length was 942.25 cM and consisted of 10 linkage groups, corresponding to the 10 chromosomes of *Brassica rapa*. The largest linkage group was A03 with a size of 186.2 cM and the smallest linkage group was A07 with a size of 52.7 cM. Each of the linkage groups had at least one SSR marker, which allowed the identification of the corresponding chromosome and the comparison with previously published maps (Lou et al. 2008, Kim et al. 2006, Choi et al. 2007). In addition to the SSR markers, gene targeted markers related to the glucosinolate pathway were mapped in this population as well. A total of 19 markers related to the glucosinolate (GLS) biosynthetic pathway were mapped in all the linkage groups except for A04, A05 and A10. The linkage group with most GLS genes mapped was A03 with six. The mapping of this particular group of markers together with the SSRs was of aid for the identification of the map orientation and for further syntenic comparison with *Arabidopsis thaliana* in the search of candidate genes for metabolic pathways.

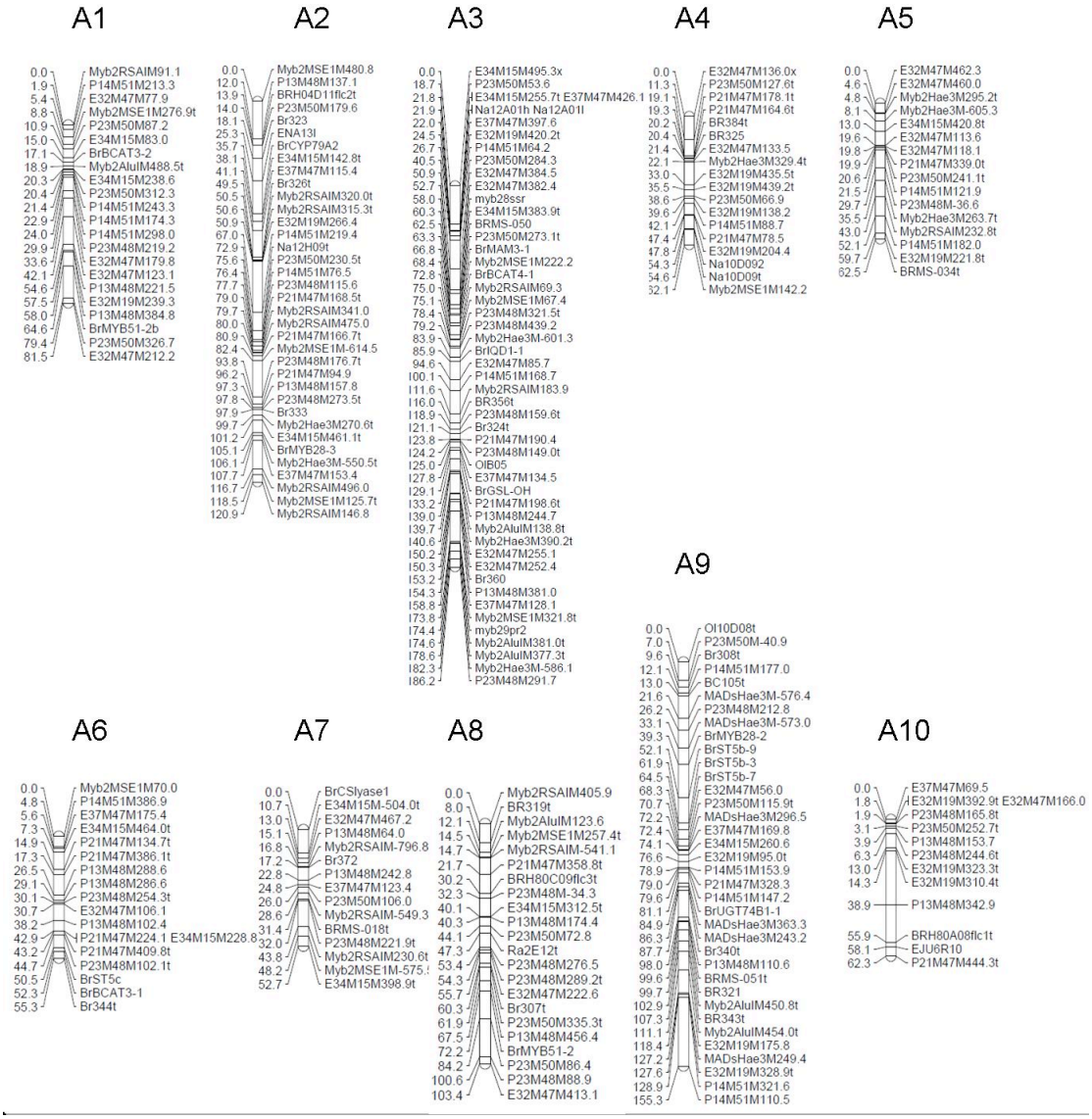


Figure 1. Map Doubled Haploid population DH68

QTL analysis of targeted metabolites

Making use of targeted metabolic extraction and analyses procedures, including quantification using reference compounds, it was possible to measure the variation in absolute levels of health-related phytochemicals in the leaves of *Brassica rapa*. Among the compounds quantified were tocopherols, carotenoids, and folic acid. The levels obtained for the carotenoids, α-, β-, δ- and γ-tocopherol and folate are presented in Table 1. Variation was observed in both the content and the composition within the DH population. Within the maximum values obtained among all the DH lines the highest levels for the tocopherols were observed for α-tocopherol with 22.65 mg/kg FW. The lowest value was observed for δ-tocopherol with a value of 0.21

mg/kg FW. In the case of the carotenoids, the higher value was lutein with a maximum value of 158 mg/kg FW and the lowest maximum value was observed for neoxanthin with a value of 45.41 mg/kg FW. The level of folic acid ranged between 1303.68 to 4115.69 mg/kg FW.

	TOCOPHEROLS				CAROTENOIDS				
	βtocopherol	αtocopherol	γtocopherol	δtocopherol	βcarotene	lutein	neoxanthin	violaxanthin	folate
Mean	0.40	14.71	0.19	0.03	61.68	84.70	28.77	66.11	2304.53
Median	0.41	14.34	0.17	0.03	61.28	78.47	27.71	65.37	2183.27
Minimum	0.12	9.18	0.07	0.00	42.70	44.45	16.11	40.31	1303.68
Maximum	0.82	22.65	0.80	0.22	86.35	158.00	45.41	90.27	4115.69
s.e.m	0.02	0.30	0.01	0.00	1.01	2.74	0.63	1.16	63.25
%CV	36.04	18.92	58.23	104.29	15.41	30.36	20.50	16.44	26.04

Table 1. Summary statistics of the metabolic variation for targeted metabolites. Units for all the metabolites are in mg/kg per FW.

The QTL analysis of β-carotene, lutein, violaxanthin and neoxanthin resulted in 16 significant mQTL (logp > 3.37, FDR=0.05) for these compounds (Table 2). The identified QTL were calculated as significant logp value per marker and were distributed as follows: four for lutein, seven for β-carotene, three for neoxanthin and two for violaxanthin. The highest logp values were found for markers located in linkage group A03 for lutein (logp=11.5) and β-carotene (logp=4.3) and linkage group A09 for violaxanthin (logp=4) and neoxanthin (logp=3.8). Based on the QTL plot profile we also identified overlapping QTL regions located in linkage A05 for lutein and violaxanthin and in linkage group A10 for violaxanthin and neoxanthin.

The QTL analysis of α-, β-, δ- and γ-tocopherol resulted in a total of seven significant mQTL (logp > 3.37, FDR=0.05) for these compounds. The identified QTL with a significant logp value per marker were: two for α-tocopherol, three for β-tocopherol and two for γ tocopherol. No significant results were found for δ-tocopherol.

Table 2. Metanetwork metabolic QTL results of targeted (tocopherols,carotenoids and folates) and untargeted LC-MS annotated results glucosinolates and 14 centrotypes (c-code).LG:linkage group, QTLl:left border,QTLr:right border.

Metabolite	LG	Marker	QTLl(cM)	QTLpeak(cm)	QTLr(cM)(1.5)	Logp value
δ-tocopherol	NS					
γ-tocopherol	2	M35	49.9965	50.913	58.9735	5
γ-tocopherol	5	M131	0	8.057	16.3365	4.1
β-tocopherol	3	M73	61.356	62.454	65.0335	5.3
β-tocopherol	3	M61	20.219	21.785	23.242	7.2
β-tocopherol	3	M68	33.604	40.504	51.781	7.7
α-tocopherol	2	M37	50.5115	72.876	74.218	4.2
α-tocopherol	5	M133	0	19.635	25.614	4.4
Folate	9	M218	75.3365	78.95	80.3405	3.7
Folate	9	M234	127.3955	155.279	155.279	4.3
lutein	3	M88	117.466	118.912	120.019	7.7
lutein	3	M92	124.003	125.032	128.468	9.9
lutein	3	M95	131.1835	133.224	145.3615	11.5
lutein	5	M132	0	13.038	21.0605	4.6
β-carotene	3	M62	9.3265	21.84	33.604	3.4
β-carotene	3	M83	84.8655	85.872	97.351	3.6
β-carotene	3	M104	166.2935	173.753	180.483	4
β-carotene	3	M75	65.0335	66.806	67.627	4.1
β-carotene	3	M109	184.292	186.239	186.239	4.3
β-carotene	5	M143	47.5445	62.52	62.52	3.5
β-carotene	5	M128	0	0	6.4155	3.6
Neoxanthin	3	M88	81.5115	118.912	120.019	3.4
Neoxanthin	9	M230	114.7575	118.417	122.8025	3.8
Neoxanthin	9	M228	92.8325	107.279	109.1885	3.8
Violaxanthin	9	M228	92.8325	107.279	109.1885	3.6
Violaxanthin	9	M230	114.7575	118.417	142.0895	4
glucoraphanin	3	M80	73.8775	78.402	90.249	5.8
glucoraphanin	3	M74	59.121	63.261	65.0335	7.3
sinigrin	1	M9	19.6135	20.293	20.929	4.6
sinigrin	1	M7	12.9325	17.119	18.0265	5
sinigrin	9	M218	63.2375	78.95	80.3405	7.7
progoitrin	3	M90	122.447	123.768	128.468	9.6
progoitrin	9	M214	69.516	72.422	73.2555	4.8
progoitrin	9	M205	23.8815	26.19	45.7	5.2
glucoalyssin	9	M218	66.442	78.95	80.3405	4.5
hydroxypentenyl	3	M74	59.121	63.261	65.0335	6.8
hydroxypentenyl	3	M91	124.003	124.238	128.468	9.7
hydroxypentenyl	3	M80	76.729	78.402	84.8655	11.1
hydroxypentenyl	3	M84	90.249	94.626	113.7975	12.9
gluconapin	9	M205	23.8815	26.19	29.6625	5.3
gluconapin	9	M214	66.442	72.422	73.2555	9.7
methsulphonylbutyl	6	M151	11.1085	29.124	47.5845	5.2
methsulphonylbutyl	8	M187	40.2155	44.124	50.37	4.4
methsulphonylbutyl	9	M217	75.3365	78.915	80.3405	9.4
methsulphonylbutyl	9	M213	71.421	72.157	73.2555	10.1
methsulphonylbutyl	9	M222	83.0145	86.271	86.9845	10.2
glucobrassicapin	3	M80	76.729	78.402	84.8655	5.1
glucobrassicapin	9	M206	23.8815	33.135	36.2225	4.9

glucobrassicapin	9	M222	83.0145	86.271	92.8325	5.3
glucobrassicapin	9	M211	63.2375	68.347	73.2555	5.3
glucobrassicapin	9	M218	75.3365	78.95	80.3405	5.4
glucoerucin	3	M73	61.356	62.454	65.0335	4.4
glucoerucin	3	M69	45.693	50.882	55.332	4.8
glucoerucin	9	M206	23.8815	33.135	36.2225	6.9
glucoerucin	9	M221	83.0145	84.939	86.9845	9.4
glucoerucin	9	M216	75.3365	76.584	80.3405	12.3
glucoerucin	9	M211	66.442	68.347	69.516	12.9
glucoerucin	9	M214	71.421	72.422	73.2555	14
glucobrassicin	NS					
gluconasturtin	3	M74	59.121	63.261	65.0335	4.9
methglucobrassicin	NS					
hexyl GS I	2	M35	49.9965	50.913	58.9735	5.7
hexyl GS II	NS					
hexyl GS III	3	M71	55.332	57.984	65.0335	7.1
neoglucobrassicin	NS					
heptyl GS I	3	M69	45.693	50.882	55.332	7.5
heptyl GS I	3	M79	73.8775	75.056	81.5115	7.8
heptyl GS I	3	M74	59.121	63.261	65.0335	10.2
heptyl GS II	3	M89	120.019	121.126	122.447	4.8
heptyl GS II	3	M84	90.249	94.626	105.8255	8.4
heptyl GS II	3	M79	73.8775	75.056	78.783	10.8
heptyl GS II	3	M74	61.356	63.261	65.0335	13.5
heptyl GS III	3	M69	45.693	50.882	55.332	7.4
heptyl GS III	3	M80	73.8775	78.402	84.8655	8.5
heptyl GS III	3	M74	61.356	63.261	65.0335	9.3
heptyl GS III	9	M219	75.3365	79.591	80.3405	4.7
c1082	3	M59	0	0	9.3265	6.6
c1082	7	M170	23.796	25.967	27.29	21
c1082	7	M173	30.023	31.987	37.8985	15.8
c1096	3	M59	0	0	21.8125	4.5
c1096	7	M169	23.796	24.791	25.379	19.2
c1138	7	M162	0	0	5.3695	4.6
c1138	7	M169	23.796	24.791	27.29	6.6
c1194	7	M162	0	0	5.3695	4.5
c1194	7	M169	23.796	24.791	27.29	6.6
c1194	7	M173	30.023	31.987	37.8985	6.6
c1275	7	M170	25.379	25.967	27.29	16.8
c1275	7	M172	30.023	31.433	37.8985	12.4
c1320	3	M68	33.604	40.504	45.693	4.9
c1320	7	M163	5.3695	10.739	11.8515	7.3
c1320	7	M170	19.9985	25.967	27.29	9.3
c1320	7	M172	30.023	31.433	46.001	12.1
c1355	3	M59	0	0	9.3265	5.6
c1355	3	M71	55.332	57.984	59.121	4.6
c1355	7	M169	23.796	24.791	25.379	14.5
c1435	3	M59	0	0	9.3265	6
c1435	7	M170	23.796	25.967	27.29	9
c1435	7	M173	30.023	31.987	37.8985	6.8
c1435	9	M218	75.3365	78.95	80.3405	5.1
c1435	9	M223	85.605	87.698	92.8325	5.5

c1440	3	M68	25.605	40.504	51.781	6
c1440	7	M165	14.043	15.122	15.984	7.7
c1440	7	M169	23.796	24.791	25.379	12.1
c1440	7	M173	30.023	31.987	37.8985	8.2
c1463	3	M68	33.604	40.504	45.693	4.5
c1463	7	M163	5.3695	10.739	11.8515	8.2
c1463	7	M170	23.796	25.967	27.29	11.3
c1463	7	M173	30.023	31.987	37.8985	14.2
c1500	3	M68	25.605	40.504	45.693	5.7
c1500	7	M167	15.984	17.196	19.9985	9.4
c1500	7	M169	23.796	24.791	27.29	12
c1500	7	M173	30.023	31.987	37.8985	13.9
c1500	7	M176	50.46	52.728	52.728	8.8
c1536	3	M67	25.605	26.704	55.332	5
c1536	3	M98	131.1835	140.568	145.3615	5.2
c1536	7	M162	0	0	5.3695	9.7
c1536	7	M169	23.796	24.791	27.29	12
c1536	7	M173	30.023	31.987	37.8985	10.7
c1683	7	M169	23.796	24.791	25.379	8.9
c1702	3	M68	25.605	40.504	45.693	5.8
c1702	7	M167	15.984	17.196	19.9985	8.3
c1702	7	M169	23.796	24.791	27.29	10.5
c1702	7	M173	30.023	31.987	37.8985	11.8
c1702	7	M176	50.46	52.728	52.728	8.1

The highest logp values were found for markers located in linkage group A03 for β-tocopherol (logp=7.7), linkage group A02 for γ-tocopherol (logp=5) and linkage group A05 for α-tocopherol (logp=4.4). Only one overlapping QTL region was located on linkage group A05 for mQTL results of γ- and α-tocopherol.

Although the folate biosynthesis is regulated by several genes (Sahr et al. 2005) the mQTL analysis of folate content in the double haploid population resulted in the detection of only one significant mQTL (logp > 3.48, FDR=0.05) for this metabolite. The identified mQTL had a peak logp value of 4.3 in A09 although the QTL profile was irregular throughout this linkage group .

QTL analysis of untargeted LC-MS data

The metabolite variation within the double haploid population was high with a total of 2,758 different mass peaks detected. In general, each metabolite is represented by a group of masses with different retention time and mass-scan number. After grouping of the mass peaks because of high correlation between peak signals the whole dataset was reduced to 228 centrotypes, which potentially represent different metabolites (see material and methods). The natural variation in metabolite composition represented by 228 centrotypes was used in the further QTL analyses to unravel the genetic regulation of secondary metabolites. *B. rapa* metabolites detected by this LC-QTOF MS profiling of aqueous-methanol extracts were mostly semi-polar compounds such as glucosinolates, phenylpropanoids and flavonoids (De Vos et al. 2006).

QTLs were detected for 158 out of the 228 centrotypes, the genomic distribution of the QTL profiles showed that these LC-MS detectable metabolites are not evenly distributed over the *B. rapa* linkage groups and cold and hotspots for the genetic regulation of metabolite content could be identified (Figure 2). The most important regulatory region seemed to be located in linkage group A07 where mQTLs were detected for 112 out of the 228 centrotypes. Furthermore, when we randomly selected 14 out of the 112 centrotypes mapping to A07 to be identified they corresponded to metabolites identified as flavonoids (Dr. Ric de Vos, PRI, Wageningen University & Research Center; data not shown). The significant QTL loci detected for these 14 centrotypes with at least one mQTL in A07 showed remarkable overlap in genomic regions of linkage group A03. Interestingly, nine out of the fourteen centrotypes showed the same mQTL pattern in the same regions of the linkage groups A03 and A07, while one centrotype showed an mQTL in the middle region of A03 and A07 plus a region of A09 and four centrotypes only showed an mQTL in the region of A07.

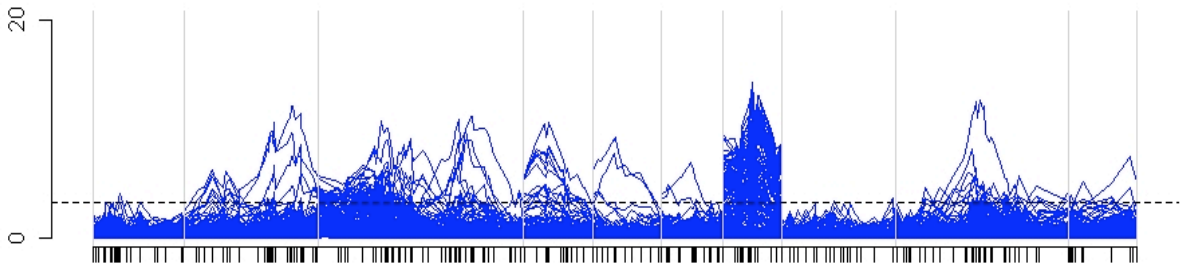


Figure 2. Whole genome QTL analysis of 228 centrotypes from the LC-MS data.Lines indicate division for each linkage group 1-10

QTL analysis of transcriptomics data

To identify the genomic regions responsible for the metabolic regulation in *Brassica rapa* we profiled the transcript abundance of 92 DH lines in the DH68 population. To maximize the differences between pairs that were hybridized to each microarray we followed a distant pair design. A total of 50 pairs were used in the transcriptomics analysis, including dye swap and replicates of the parental lines. However, because it has been reported (West et al. 2007, Keurentjes et al. 2007) that the transcript variation in segregating populations is not limited to genes differentially expressed between parents, we followed a QTL analysis that was not limited to the results obtained comparing the parental lines. Instead, we followed a regression analysis of the transcript abundance represented by the 78,278 informative probes on the microarray against the 247 markers of the map. In total, 24,850 probes were detected as significant against a marker with a logp >3 which in turn corresponded to an average of 8.5 markers found as significant per probe with a total of 210,876 eQTLs. The whole genome profile of the number of eQTL versus chromosome position is displayed in Figure 3. Based on the whole genome profile there is no evidence of eQTL clustered as hotspots. Instead, the eQTL were distributed randomly across the genome with a higher number than average on linkage groups A03, A05 and A09.

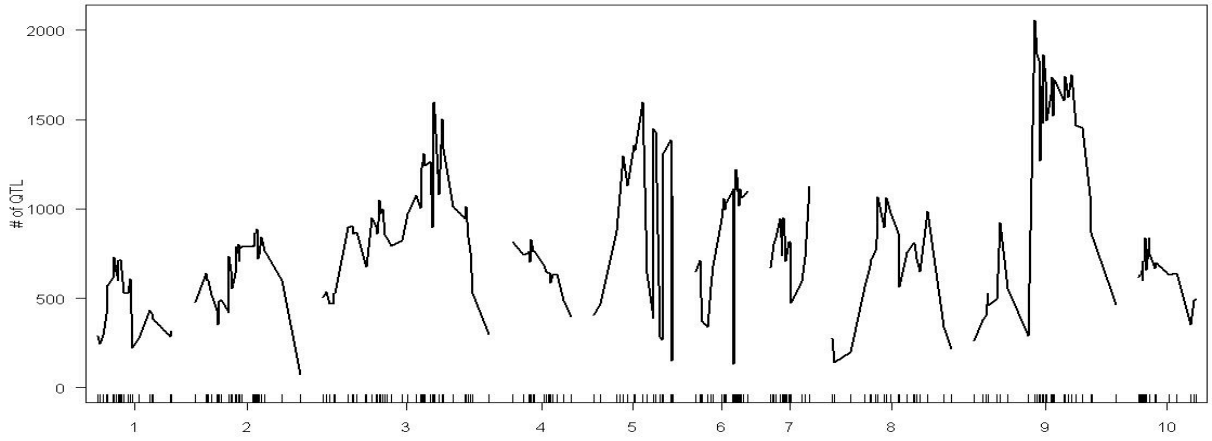


Figure 3. Whole genome QTL analysis of expression data. Numbers in x-axis indicate linkage group and y-axis indicate number of QTL per map position

Unravelling the genomic regulation of metabolic pathways through the combined analysis of eQTL and mQTL loci

To determine whether it was possible to identify regions involved in metabolic regulation we combined our mQTL and eQTL results. In order to combine these results we focused on the mQTL results of the group of targeted metabolites (carotenoids, tocopherols and folate) and the metabolites identified as glucosinolates and flavonoids from the LC-MS data of 228 centrotypes.

In order to identify eQTL involved in the regulation of metabolite production we first screened the probes represented on the microarray against a compiled list of potential candidate genes that are known to be involved in the regulation of six biosynthetic pathways that lead specifically to the production of flavonoids, phenylpropanoids, glucosinolates, carotenoids, tocopherols and folate (Table 3).

Table 3.List of candidate genes for six biosynthetic pathways.The probes of the microarray are identified with a genebank code, origin 1.B.rapa, 2.B.napus,3.B.oleracea and 4 other Brassica species.Abbreviation is included in whole genome eQTL profiling.

CAROTENOIDS			
Genbank code	**origin**	**Abbreviation**	**Description**
CV545014	1	*NCED*	9-cis-epoxycarotenoid dioxygenase
EX074328	1	*CCD8*	Carotenoid cleavage dioxygenase 8, chloroplast precursor
EX104530	1	*NCED6*	9-cis-epoxycarotenoid dioxygenase NCED6, chloroplast precursor
EX106389	1	*eLCY*	Lycopene epsilon cyclase, chloroplast precursor
EX117989	1	*ZDS*	Zeta-carotene desaturase, chloroplast/chromoplast precursor
JCVI_11826	1	*βLCY*	Lycopene beta cyclase, chloroplast precursor
JCVI_14575	1	*εLCY*	Lycopene epsilon cyclase, chloroplast precursor
JCVI_16311	1	*VDE*	Violaxanthin de-epoxidase, chloroplast precursor
JCVI_16900	1	*ZE*	Zeaxanthin epoxidase
JCVI_17398	1	*PDS*	Phytoene dehydrogenase, chloroplast/chromoplast precursor
JCVI_18580	1	*εLCY*	Lycopene epsilon cyclase, chloroplast precursor
JCVI_24842	1	*εLCY*	Lycopene epsilon cyclase, chloroplast precursor
JCVI_2741	1	*ZDS*	Putative zeta-carotene desaturase
JCVI_27625	1	*CCD4*	Probable carotenoid cleavage dioxygenase 4, chloroplast precursor
JCVI_35995	1	*NPQ1*	At2g21860/F7D8.18
JCVI_9110	1	*NCED*	9-cis-epoxycarotenoid dioxygenase
CN727043	2	*NCED2*	9-cis-epoxycarotenoid dioxygenase NCED2, chloroplast precursor
CX187784	2	*MAX1*	Protein At2g26170
CX187784	2	*MAX1*	Protein At2g26170
EE401805	2	*NCED9*	9-cis-epoxycarotenoid dioxygenase NCED9, chloroplast precursor
EV011579	2	*βOHase*	Beta-carotene hydroxylase
EV078086	2	*PDS*	Phytoene desaturase
EV127419	2	*ZDS*	Putative zeta-carotene desaturase
EV160672	2	*CCD4*	Probable carotenoid cleavage dioxygenase 4, chloroplast precursor
EV160758	2	*NCED*	9-cis-epoxycarotenoid dioxygenase
EV193301	2	*PDS*	Phytoene dehydrogenase, chloroplast/chromoplast precursor
EV197471	2	*PSY*	Phytoene synthase, chloroplast precursor
EV200814	2	*ZE*	Zeaxanthin epoxidase
EV216628	2	*εOHase*	Chloroplast carotenoid epsilon-ring hydroxylase
H07750	2	*εLCY*	Lycopene epsilon-cyclase
JCVI_1053	2	*CCD1*	Carotenoid cleavage dioxygenase 1
JCVI_10559	2	*εLCY*	Lycopene epsilon cyclase, chloroplast precursor
JCVI_10896	2	*ZDS*	Zeta-carotene desaturase, chloroplast/chromoplast precursor
JCVI_11039	2	*PSY*	Phytoene synthase, chloroplast precursor
JCVI_12969	2	*PSY*	Phytoene synthase, chloroplast precursor
JCVI_1390	2	*PDS*	Phytoene dehydrogenase, chloroplast/chromoplast precursor
JCVI_14102	2	*ZE*	Zeaxanthin epoxidase
JCVI_14909	2	*εLCY*	Lycopene epsilon cyclase, chloroplast precursor
JCVI_17404	2	*ZE*	Zeaxanthin epoxidase
JCVI_19201	2	*PSY*	Phytoene synthase, chloroplast precursor
JCVI_21332	2	*LUT1*	e-carotene hydroxylase
JCVI_21434	2	*εLCY*	Lycopene epsilon cyclase, chloroplast precursor
JCVI_2356	2	*CCD1*	Carotenoid cleavage dioxygenase 1
JCVI_3236	2	*LUT1*	e-carotene hydroxylase
JCVI_26571	2	*MAX1*	Protein At2g26170
JCVI_26954	2	*ZDS*	Zeta-carotene desaturase, chloroplast/chromoplast precursor
JCVI_2706	2	*PSY*	Phytoene synthase
JCVI_29425	2	*CRISO*	CRTISO (carotenoid isomerase); carotenoid isomerase
JCVI_31271	2	*ZE*	Zeaxanthin epoxidase
JCVI_31309	2	*PDS*	Phytoene dehydrogenase, chloroplast/chromoplast precursor
JCVI_3183	2	*CRTISO*	Putative uncharacterized protein F12K22.18
JCVI_34730	2	*βLCY*	Lycopene beta cyclase, chloroplast precursor
JCVI_3514	2	*NCED9*	9-cis-epoxycarotenoid dioxygenase NCED9, chloroplast precursor
JCVI_36369	2	*VDE*	Violaxanthin de-epoxidase, chloroplast precursor
JCVI_36812	2	*NCED3*	9-cis-epoxycarotenoid dioxygenase NCED3, chloroplast precursor
JCVI_37199	2	*εLCY*	Lycopene epsilon cyclase, chloroplast precursor
JCVI_3911	2	*βOHase*	Putative beta-carotene hydroxylase
JCVI_40121	2	*εLCY*	Lycopene epsilon cyclase, chloroplast precursor
JCVI_4781	2	*CCD4*	Probable carotenoid cleavage dioxygenase 4, chloroplast precursor
JCVI_5261	2	*βOHase*	Putative beta-carotene hydroxylase
JCVI_5697	2	*NCED3*	9-cis-epoxycarotenoid dioxygenase NCED3, chloroplast precursor
JCVI_7482	2	*PDS*	Phytoene dehydrogenase-like
JCVI_7641	2	*NCED*	9-cis-epoxycarotenoid dioxygenase
JCVI_844	2	*ε-OHase*	Chloroplast carotenoid epsilon-ring hydroxylase
JCVI_9750	2	*βLCY*	Lycopene beta cyclase, chloroplast precursor
AM386177	3	*βOHase*	Putative beta-carotene hydroxylase
EH429196	3	*NPQ1*	At2g21860/F7D8.18
JCVI_29716	3	*PSY*	Phytoene synthase, chloroplast precursor

Table 3. continued

GLUCOSINOLATES			
Genbank code	**origin**	**Abbreviation**	**Description**
JCVI_8920	2	*GS-OH*	1-aminocyclopropane-1-carboxylate oxidase homolog 4
JCVI_21288	2	*GS-OH*	1-aminocyclopropane-1-carboxylate oxidase-like protein
JCVI_30455	2	*IPMS*	2-isopropylmalate synthase
JCVI_12366	1	*IPMS1*	2-isopropylmalate synthase 1
JCVI_1943	3	*IPMS1*	2-isopropylmalate synthase 1
JCVI_42142	2	*IPMS1*	2-isopropylmalate synthase 1
JCVI_8389	2	*IPMS1*	2-isopropylmalate synthase 1
ES909518	2	*AOP1*	2-oxoglutarate-dependent dioxygenase
JCVI_19574	1	*AOP2-1*	2-oxoglutarate-dependent dioxygenase
JCVI_31233	1	*AOP*	2-oxoglutarate-dependent dioxygenase
JCVI_33047	2	*AOP*	2-oxoglutarate-dependent dioxygenase
DY011947	2	*Aconitase*	3-isopropylmalate dehydratase, small subunit
JCVI_13240	2	*Aconitase*	3-isopropylmalate dehydratase, small subunit
JCVI_317	3	*Aconitase*	3-isopropylmalate dehydratase, small subunit
JCVI_35164	1	*Aconitase*	3-isopropylmalate dehydratase, small subunit
JCVI_13669	2	*IPM DH2*	3-isopropylmalate dehydrogenase 2, chloroplast precursor
JCVI_19682	1	*IPM DH2*	3-isopropylmalate dehydrogenase 2, chloroplast precursor
JCVI_10889	2	*IPM DH3*	3-isopropylmalate dehydrogenase 3, chloroplast precursor
EE553947	2	*IPM DH*	3-isopropylmalate dehydrogenase, chloroplast precursor
JCVI_1748	2	*IPM DH*	3-isopropylmalate dehydrogenase, chloroplast precursor
JCVI_24090	2	*IPM DH*	3-isopropylmalate dehydrogenase, chloroplast precursor
JCVI_33561	1	*IPM DH*	3-isopropylmalate dehydrogenase, chloroplast precursor
JCVI_3780	2	*IPM DH*	3-isopropylmalate dehydrogenase, chloroplast precursor
ES901517	2	*AOP1*	AOP1.2
JCVI_14016	2	*AOP1*	AOP1.2
JCVI_24139	2	*AOP1*	AOP1.2
JCVI_31713	2	*AOP1*	AOP1.2
JCVI_40366	3	*AOP1*	AOP1.2
JCVI_5015	2	*AOP1*	AOP1.2
JCVI_36433	2	*AOP2*	AOP2
JCVI_24334	2	*CYP79B3*	At2g22330
JCVI_31655	2	*Myb34*	ATR1
EE441525	2	*Myb34*	ATR1 MYB34
JCVI_33924	2	*Myb34*	ATR1 MYB34
JCVI_39476	2	*Myb34*	ATR1 MYB34
EV036453	2	*CYP79A2*	Cytochrome P450 79A2
JCVI_14950	1	*CYP83A1*	Cytochrome P450 83A1
JCVI_5112	2	*CYP83A1*	Cytochrome P450 83A1
JCVI_109	2	*CYP83B1*	Cytochrome P450 83B1
EV126172	2	*DOF1.1-2*	Dof zinc finger protein DOF1.1
ES917650	2	*MAM*	Methylthioalkylmalate synthase a
EX019177	1	*MAM3-1*	Methylthioalkylmalate synthase a
EE401951	2	*MAM1*	Methylthioalkylmalate synthase precursor
EX059080	1	*MAM1-1*	Methylthioalkylmalate synthase precursor
JCVI_12709	2	*BCAT4*	Probable branched-chain-amino-acid aminotransferase 4
JCVI_18213	2	*BCAT4*	Probable branched-chain-amino-acid aminotransferase 4
JCVI_34763	2	*BCAT4*	Probable branched-chain-amino-acid aminotransferase 4
EV215478	2	*DOF1.1-1*	Protein At1g07640
JCVI_32569	2	*DOF1.1-2*	Protein At1g07640
EX112446	1	*SUR1*	Protein At2g20610
JCVI_24441	2	*Myb28*	Protein At5g61420
DY009740	1	*MAM3-2*	Putative 2-isopropylmalate synthase
JCVI_32420	2	*MAM*	Putative 2-isopropylmalate synthase
JCVI_32618	2	*MAM*	Putative 2-isopropylmalate synthase
JCVI_15640	2	*FMOGS-OX1*	Putative flavin-binding monooxygenase protein
JCVI_1081	1	*Aconitase*	Putative uncharacterized protein At4g13430
JCVI_4650	2	*Aconitase*	Putative uncharacterized protein At4g13430
JCVI_486	2	*Aconitase*	Putative uncharacterized protein At4g13430
JCVI_9992	1	*Aconitase*	Putative uncharacterized protein At4g13430
DN961030	1	*SUR1*	ROOTY/SUPERROOT1
JCVI_32185	1	*SUR1*	ROOTY/SUPERROOT1
JCVI_40856	1	*SUR1*	ROOTY/SUPERROOT1
JCVI_531	2	*SUR1*	ROOTY/SUPERROOT1
EE409221	2	*AtSOT17*	Sulfotransferase
JCVI_17335	2	*CYP79F1*	Supershoot 1
JCVI_5227	2	*FMOGS-OX1*	FLAVIN-MONOOXYGENASE GLUCOSINOLATE S-OXYGENASE 2

Table 3. continued

PHENYLPROPANOIDS			
Genbank code	**origin**	**Abbreviation**	**Description**
JCVI_9460	2	*UGT72E1*	AT3g50740/T3A5_120
JCVI_13790	2	*Ugt84a1*	At4g15480
JCVI_16482	2	*Ugt84a1*	At4g15480
AT001707	1	*CCOmt1*	Caffeoyl-CoA 3-O-methyltransferase
JCVI_1016	2	*CCOmt1*	Caffeoyl-CoA 3-O-methyltransferase
EV151983	2	*CCOmt1*	Caffeoyl-CoA O-methyltransferase
AM386090	3	*C4H*	Cinnamate 4-hydroxylase
JCVI_18334	1	*C4H*	Cinnamate 4-hydroxylase
JCVI_566	1	*C4H*	Cinnamate 4-hydroxylase
EX137858	1	*C4H*	Cinnamate 4-hydroxylase isoform 1
JCVI_1047	2	*C4H*	Cinnamate 4-hydroxylase isoform 1
JCVI_121	2	*C4H*	Cinnamate 4-hydroxylase isoform 1
JCVI_1296	2	*C4H*	Cinnamate 4-hydroxylase isoform 1
JCVI_13980	1	*C4H*	Cinnamate 4-hydroxylase isoform 1
JCVI_20750	1	*C4H*	Cinnamate 4-hydroxylase isoform 2
EX132152	1	*CCR2*	Cinnamoyl CoA reductase CCR2
EE559396	2	*CCR1*	Cinnamoyl CoA reductase isoform 1
JCVI_13173	1	*CCR2*	Cinnamoyl CoA reductase, putative
JCVI_4502	2	*CCR2*	Cinnamoyl CoA reductase, putative
JCVI_28683	2	*Ccr6*	Cinnamoyl-CoA reductase-like protein
JCVI_5513	2	*CYP98A3*	Cytochrome P450 98A3
JCVI_9357	2	*CYP98A3*	Cytochrome P450 98A3
JCVI_9360	2	*CYP98A3*	Cytochrome P450 98A3
EV199049	2	*F5H*	Ferulate-5-hydroxylase
JCVI_26392	1	*F5H*	Ferulate-5-hydroxylase precursor
EX099365	1	*CCOmt1*	O-methyltransferase
EE418726	2	*PAL*	Phenylalanine ammonia-lyase
JCVI_10710	2	*PAL*	Phenylalanine ammonia-lyase
JCVI_10975	2	*PAL*	Phenylalanine ammonia-lyase
JCVI_13216	2	*PAL*	Phenylalanine ammonia-lyase
JCVI_15212	2	*PAL*	Phenylalanine ammonia-lyase
JCVI_17602	2	*PAL*	Phenylalanine ammonia-lyase
JCVI_19663	2	*PAL*	Phenylalanine ammonia-lyase
JCVI_31462	2	*PAL*	Phenylalanine ammonia-lyase
JCVI_32380	2	*PAL*	Phenylalanine ammonia-lyase
EV152862	2	*PAL3*	Phenylalanine ammonia-lyase 3
JCVI_22515	2	*PAL4*	Phenylalanine ammonia-lyase 4
JCVI_7599	2	*PAL4*	Phenylalanine ammonia-lyase 4
DT317691	5	*PAL*	Phenylalanine ammonia-lyase class 1
DN961091	1	*4Cl9*	Protein At1g20510
JCVI_39210	2	*4Cl9*	Protein At1g20510
EV080676	2	*4Cl3*	Protein At1g65060
JCVI_36583	2	*CCOmt1*	Putative caffeoyl-CoA O-methyltransferase protein
EV180978	2	*Ccr6*	Putative cinnamoyl CoA reductase
EX040361	1	*Ccr6*	Putative cinnamoyl CoA reductase
JCVI_13824	2	*Ccr6*	Putative cinnamoyl CoA reductase
JCVI_31369	2	*Ccr6*	Putative cinnamoyl CoA reductase
JCVI_19558	1	*4Cl9*	Putative uncharacterized protein At1g20510
EE503308	2	*SCPL10*	Serine carboxypeptidase-like 10 precursor
JCVI_32692	2	*SCPL10*	Serine carboxypeptidase-like 10 precursor
EE458112	2	*SCPL13*	Serine carboxypeptidase-like 13 precursor
JCVI_26719	3	*SCPL13*	Serine carboxypeptidase-like 13 precursor
EX109440	1	*SCPL9*	Serine carboxypeptidase-like 9 precursor
JCVI_28075	1	*SCPL9*	Serine carboxypeptidase-like 9 precursor
EV214117	2	*Myb12*	Transcription factor MYB12
JCVI_29626	1	*Myb12*	Transcription factor MYB12
JCVI_4697	2	*Myb12*	Transcription factor MYB12
JCVI_14659	2	*Myb75*	Transcription factor MYB75
ES903366	2	*UGT72E3*	F9D12.4 protein
ES907905	2	*Ccr1*	T24D18.5 protein
JCVI_13385	2	*Ccr1*	T24D18.5 protein
JCVI_13416	2	*Ccr1*	T24D18.5 protein
JCVI_16611	2	*Ccr1*	T24D18.5 protein
JCVI_18675	1	*Ccr1*	T24D18.5 protein
JCVI_28458	2	*UGT72E3*	F9D12.4 protein
JCVI_35897	1	*UGT72E3*	F9D12.4 protein
JCVI_814	2	*Ccr1*	T24D18.5 protein
JCVI_31825	2	*CCOmt7*	Probable caffeoyl-CoA O-methyltransferase At4g26220
JCVI_11293	3	*CCOmt7*	Probable caffeoyl-CoA O-methyltransferase At4g26220

Table 3. continued

TOCOPHEROLS			
Genbank code	**origin**	**Abbreviation**	**Description**
EV109105	2	*VTE4*	Gamma-tocopherol methyl transferase
JCVI_20479	2	*VTE4*	Gamma-tocopherol methyltransferase
EE423907	2	*GGPS*	Geranyl geranyl pyrophosphate synthase
JCVI_23702	2	*GGPS*	Geranyl geranyl pyrophosphate synthase-like protein
JCVI_2864	2	*GGPS*	Geranylgeranyl pyrophosphate synthase-related protein, chloroplast precursor
ES926228	2	*GGPS2*	Geranylgeranyl pyrophosphate synthetase 2 precursor (GGPP synthetase 2) (GGPS2)
JCVI_14579	1	*GGPS2*	Geranylgeranyl pyrophosphate synthetase 2 precursor (GGPP synthetase 2) (GGPS2)
JCVI_39717	1	*GGPS2*	Geranylgeranyl pyrophosphate synthetase 2 precursor (GGPP synthetase 2) (GGPS2)
ES935671	1	*GGPS3*	Geranylgeranyl pyrophosphate synthetase 3, chloroplast precursor (GGPP synthetase 3)
CD827634	2	*GGPS4*	Geranylgeranyl pyrophosphate synthetase 4 precursor (GGPP synthetase 4) (GGPS4)
ES960774	2	*GGPS*	Geranylgeranyl pyrophosphate synthetase, chloroplast/chromoplast precursor (GGPP synthetase) (GGPS)
JCVI_4351	2	*GGPS*	Geranylgeranyl pyrophosphate synthetase, chloroplast/chromoplast precursor (GGPP synthetase) (GGPS)
EV101195	2	*GGR*	Geranylgeranyl reductase
EV189544	2	*GGR*	Geranylgeranyl reductase
JCVI_123	2	*GGR*	Geranylgeranyl reductase
JCVI_2005	2	*GGR*	Geranylgeranyl reductase
JCVI_961	2	*GGR*	Geranylgeranyl reductase
DY028331	3	*PDS*	Phytoene dehydrogenase-like
JCVI_7482	2	*PDS*	Phytoene dehydrogenase-like
EV193301	2	*PDS*	Phytoene dehydrogenase, chloroplast/chromoplast precursor
JCVI_1390	2	*PDS*	Phytoene dehydrogenase, chloroplast/chromoplast precursor
JCVI_17398	1	*PDS*	Phytoene dehydrogenase, chloroplast/chromoplast precursor
JCVI_31309	2	*PDS*	Phytoene dehydrogenase, chloroplast/chromoplast precursor
EV078086	2	*PDS*	Phytoene desaturase
JCVI_2706	2	*PSY*	Phytoene synthase
EV197471	2	*PSY*	Phytoene synthase, chloroplast precursor
JCVI_11039	2	*PSY*	Phytoene synthase, chloroplast precursor
JCVI_12969	2	*PSY*	Phytoene synthase, chloroplast precursor
JCVI_19201	2	*PSY*	Phytoene synthase, chloroplast precursor
JCVI_29716	3	*PSY*	Phytoene synthase, chloroplast precursor
JCVI_39765	1	*VTE2*	Putative uncharacterized protein At2g18950
EX056312	1	*VTE1*	Tocopherol cyclase, chloroplast precursor
JCVI_15620	2	*VTE1*	Tocopherol cyclase, chloroplast precursor
JCVI_7811	2	*VTE1*	Tocopherol cyclase, chloroplast precursor
EV218406	2	*VTE2*	Putative uncharacterized protein At2g18950
FOLATE			
Genbank code	**origin**	**Abbreviation**	**Description**
EV210859	2	*ATDFB*	ATDFB (A. THALIANA DHFS-FPGS HOMOLOG B); tetrahydrofolylpolyglutamate synthase
JCVI_34479	1	*ATDFB*	ATDFB (A. THALIANA DHFS-FPGS HOMOLOG B); tetrahydrofolylpolyglutamate synthase
JCVI_40645	1	*ATDFC*	ATDFC (A. THALIANA DHFS-FPGS HOMOLOG C); dihydrofolate synthase
EE568745	2	*DHFR-TS1*	Bifunctional dihydrofolate reductase-thymidylate synthase 1 (DHFR-TS1)
JCVI_14706	2	*DHFR-TS1*	Bifunctional dihydrofolate reductase-thymidylate synthase 1 (DHFR-TS 1)
JCVI_30585	3	*DHFR-TS1*	Bifunctional dihydrofolate reductase-thymidylate synthase 1 (DHFR-TS 1)
JCVI_30590	3	*DHFR-TS1*	Bifunctional dihydrofolate reductase-thymidylate synthase 1 (DHFR-TS 1)
CX268767	1	*DHFR-TS2*	Bifunctional dihydrofolate reductase-thymidylate synthase 2 (DHFR-TS 2)
JCVI_10760	2	*DHFR-TS2*	Bifunctional dihydrofolate reductase-thymidylate synthase 2 (DHFR-TS 2)
JCVI_16406	3	*DHFR-TS2*	Bifunctional dihydrofolate reductase-thymidylate synthase 2 (DHFR-TS 2)
JCVI_32112	2	*DHFR-TS2*	Bifunctional dihydrofolate reductase-thymidylate synthase 2 (DHFR-TS 2)
JCVI_34008	2	*DHFR-TS2*	Bifunctional dihydrofolate reductase-thymidylate synthase 2 (DHFR-TS 2)
JCVI_6575	2	*DHFR-TS2*	Bifunctional dihydrofolate reductase-thymidylate synthase 2 (DHFR-TS 2)
CX187539	2	*DHFS*	Dihydrofolate synthetase
EV087170	2	*EMB1997*	EMB1997
JCVI_22500	2	*EMB1997*	EMB1997
JCVI_25196	2	*EMB1997*	EMB1997
JCVI_37402	2	*EMB1997*	EMB1997
EX060511	1	*DFC*	Encodes a protein with tetrahydrofolylpolyglutamate synthase activity that is located in the mitochondrial matrix.
JCVI_18087	2	*FPGS*	Folylpolyglutamate synthase-like protein
JCVI_14052	2	*FPGS*	Folylpolyglutamate synthetase precursor
JCVI_20901	2	*FPGS*	Folylpolyglutamate synthetase, chloroplastic isoform
JCVI_41473	1	*FPGS*	Folylpolyglutamate synthetase, chloroplastic isoform
JCVI_40595	1	*FPGS*	Folylpolyglutamate synthetase, cytosolic isoform
BG543197	1	*THFS*	Formate--tetrahydrofolate ligase
JCVI_10210	2	*THFS*	Formate--tetrahydrofolate ligase
JCVI_28501	2	*THFS*	Formate--tetrahydrofolate ligase
JCVI_35985	1	*THFS*	Formate--tetrahydrofolate ligase
JCVI_832	3	*THFS*	Formate--tetrahydrofolate ligase
JCVI_25564	2	*GLA1*	GLA1 (GLOBULAR ARREST1); tetrahydrofolylpolyglutamate synthase
JCVI_41383	2	*GLA1*	GLA1 (GLOBULAR ARREST1); tetrahydrofolylpolyglutamate synthase
JCVI_23396	2	*MTHFR*	Methylenetetrahydrofolate reductase 1
JCVI_2938	2	*MTHFR*	Methylenetetrahydrofolate reductase 1
JCVI_6438	4	*MTHFR*	Methylenetetrahydrofolate reductase 1
EX043765	1	*MTHFR*	Methylenetetrahydrofolate reductase 2
JCVI_12138	2	*MTHFR*	Methylenetetrahydrofolate reductase 2
JCVI_16284	2	*MTHFR*	Methylenetetrahydrofolate reductase 2
JCVI_20682	2	*MTHFR*	Methylenetetrahydrofolate reductase 2
JCVI_15105	2	*DHFR-TS*	Putative dihydrofolate reductase-thymidylate synthase

Table 3. continued

FLAVONOIDS			
Genbank code	**origin**	**Abbreviation**	**Description**
DN961176	1	*4CL*	4 coumarate CoA ligase
DN964565	1	*4CL1*	4-coumarate--CoA ligase 1
JCVI_1086	2	*4CL1*	4-coumarate--CoA ligase 1
JCVI_28371	1	*4CL1*	4-coumarate--CoA ligase 1
JCVI_7527	2	*4CL1*	4-coumarate--CoA ligase 1
JCVI_32600	2	*4CL2*	4-coumarate--CoA ligase 2
JCVI_21718	2	*4CL3*	4-coumarate--CoA ligase 3
JCVI_26641	2	*4CL3*	4-coumarate--CoA ligase 3
CD813580	2	*4CL4*	4-coumarate--CoA ligase 4
JCVI_22820	1	*4CL4*	4-coumarate--CoA ligase 4
JCVI_38583	2	*4CL4*	4-coumarate--CoA ligase 4
JCVI_39778	1	*4CL4*	4-coumarate--CoA ligase 4
EV123005	2	*4CL2*	4-coumarate--CoA ligase 4CL2
EL590255	2	*4CL*	4-coumarate--CoA ligase family protein / 4-coumaroyl-CoA synthase family protein
JCVI_29296	2	*4CL*	4-coumarate--CoA ligase family protein / 4-coumaroyl-CoA synthase family protein
ES905313	2	*4CL*	4-coumarate--CoA ligase-like protein
JCVI_12961	2	*4CL*	4-coumarate--CoA ligase-like protein
JCVI_28304	2	*4CL*	4-coumarate--CoA ligase-like protein
CX189414	2	*4CL*	4-coumarate-CoA ligase-like protein
EV134945	2	*4CL*	4-coumarate-CoA ligase-like protein
JCVI_1014	2	*4CL*	4-coumarate-CoA ligase-like protein
JCVI_15414	3	*4CL*	4-coumarate-CoA ligase-like protein
JCVI_15580	2	*4CL*	4-coumarate-CoA ligase-like protein
JCVI_16133	2	*4CL*	4-coumarate-CoA ligase-like protein
JCVI_18695	2	*4CL*	4-coumarate-CoA ligase-like protein
JCVI_19940	2	*4CL*	4-coumarate-CoA ligase-like protein
JCVI_40711	2	*4CL*	4-coumarate-CoA ligase-like protein
JCVI_16747	2	*ANS*	Anthocyanidin synthase
JCVI_18158	3	*ANS*	Anthocyanidin synthase
JCVI_7650	2	*ANS*	Anthocyanidin synthase
JCVI_20382	1	*CHI*	Chalcone isomerase, putative
JCVI_512	1	*CHI*	Chalcone isomerase, putative
ES954613	2	*CHI*	Chalcone isomerase, putative; 94270-95700
JCVI_15733	2	*CHS*	Chalcone synthase
JCVI_17935	2	*CHS*	Chalcone synthase
JCVI_2058	2	*CHS*	Chalcone synthase
JCVI_3974	1	*CHS*	Chalcone synthase
JCVI_26515	3	*CHS1*	Chalcone synthase 1
JCVI_1334	2	*CHS3*	Chalcone synthase 3
JCVI_2411	2	*CHS3*	Chalcone synthase 3
JCVI_6210	2	*CHSA1*	Chalcone synthase A1
JCVI_2414	2	*CHSB1*	Chalcone synthase B1
EV141692	2	*CHS*	Chalcone synthase family protein
JCVI_10836	1	*CHS*	Chalcone synthase family protein
JCVI_21743	1	*CHS*	Chalcone synthase family protein
JCVI_2460	2	*CHS*	Chalcone synthase family protein
JCVI_27081	2	*CHS*	Chalcone synthase family protein
JCVI_5	2	*CHS*	Chalcone synthase family protein
EV160950	2	*CHSh*	Chalcone synthase homolog
JCVI_29194	1	*CHSh*	Chalcone synthase homolog
CD834583	2	*CHI*	Chalcone--flavonone isomerase
JCVI_18538	2	*CHI*	Chalcone--flavonone isomerase
JCVI_20189	2	*CHI*	Chalcone--flavonone isomerase
JCVI_31142	1	*CHI*	Chalcone--flavonone isomerase
EV005315	2	*CHS*	CHS-like protein
ES947547	3	*DFR*	Dihydroflavonol 4-reductase
JCVI_11889	2	*DFR*	Dihydroflavonol 4-reductase
JCVI_5913	1	*DFR*	Dihydroflavonol 4-reductase-like
EE513550	2	*DFR*	Dihydroflavonol 4-reductase, putative
JCVI_26257	3	*DFR*	Dihydroflavonol 4-reductase, putative
JCVI_17031	2	*F3'H*	Flavanone 3-hydroxylase 1
JCVI_24830	2	*F3'H*	Flavanone 3-hydroxylase-like protein
JCVI_26368	2	*F3'H*	Flavanone 3-hydroxylase-like protein
JCVI_8351	3	*F3'H*	Flavanone 3-hydroxylase-like protein
JCVI_8838	1	*F3'H*	Flavanone 3-hydroxylase-like protein
JCVI_15136	2	*F3'H*	Flavonoid 3'-hydroxylase
EH419222	3	*F3'5'H*	Flavonoid 3', 5'-hydroxylase-like protein
EV102803	2	*F3'5'H*	Flavonoid 3', 5'-hydroxylase-like protein
JCVI_14153	2	*F3'5'H*	Flavonoid 3', 5'-hydroxylase-like protein
JCVI_32868	3	*F3'5'H*	Flavonoid 3', 5'-hydroxylase-like protein
CN728267	2	*F3'5'H*	Flavonoid 3',5'-hydroxylase-like; cytochrome P450
JCVI_33397	2	*F3'5'H*	Flavonoid 3',5'-hydroxylase-like; cytochrome P450
DY025170	2	*UGT78D2*	Flavonol 3-O-glucosyltransferase-like
JCVI_12895	2	*UGT78D2*	Flavonol 3-O-glucosyltransferase-like protein

Table 3. continued

EE502236	2	*FLS*	Flavonol synthase
EV046864	2	*FLS*	Flavonol synthase
JCVI_21931	2	*FLS*	Flavonol synthase
JCVI_2239	2	*FLS*	Flavonol synthase
JCVI_39687	2	*FLS*	Flavonol synthase
JCVI_41972	2	*FLS*	Flavonol synthase
DY029766	3	*FLS*	Flavonol synthase-like protein
JCVI_19465	2	*F3'H*	Flavonol synthase/flavanone 3-hydroxylase
JCVI_2934	2	*F3'H*	Flavonol synthase/flavanone 3-hydroxylase
ES938339	3	*Myb4*	Myb-related protein Myb4
JCVI_11777	2	*TT12*	Protein TRANSPARENT TESTA 12
JCVI_15282	2	*TT12*	Protein TRANSPARENT TESTA 12
ES992355	2	*TT16*	Protein TRANSPARENT TESTA 16
JCVI_1338	2	*TT16*	Protein TRANSPARENT TESTA 16
JCVI_14265	2	*TT16*	Protein TRANSPARENT TESTA 16
JCVI_27138	2	*TT16*	Protein TRANSPARENT TESTA 16
EV027089	2	*TT8*	Protein TRANSPARENT TESTA 8
EE540565	2	*ANS*	Putative anthocyanidin synthase
EX135089	1	*ANS*	Putative anthocyanidin synthase
JCVI_34470	2	*ANS*	Putative anthocyanidin synthase
JCVI_888	1	*CHS*	Putative chalcone synthase
L46421	1	*CHS*	Putative chalcone synthase
EE443026	2	*UGT78D2*	Putative flavonol 3-O-glucosyltransferase
EV200311	2	*UGT78D2*	Putative flavonol 3-O-glucosyltransferase
JCVI_24301	3	*UGT78D2*	Putative flavonol 3-O-glucosyltransferase
JCVI_28934	2	*UGT78D2*	Putative flavonol 3-O-glucosyltransferase
JCVI_3312	2	*UGT78D2*	Putative flavonol 3-O-glucosyltransferase
JCVI_33709	2	*UGT78D2*	Putative flavonol 3-O-glucosyltransferase
JCVI_34261	3	*UGT78D2*	Putative flavonol 3-O-glucosyltransferase
JCVI_38505	1	*UGT78D2*	Putative flavonol 3-O-glucosyltransferase
JCVI_40360	2	*UGT78D2*	Putative flavonol 3-O-glucosyltransferase
JCVI_41487	2	*UGT78D2*	Putative flavonol 3-O-glucosyltransferase
BQ791509	1	*FLS*	Putative flavonol synthase
JCVI_16142	3	*FLS*	Putative flavonol synthase
EV120446	2	*COmtl*	Quercetin 3-O-methyltransferase 1
JCVI_14670	2	*COmtl*	Quercetin 3-O-methyltransferase 1
JCVI_35351	2	*COmtl*	Quercetin 3-O-methyltransferase 1
JCVI_4160	2	*COmtl*	Quercetin 3-O-methyltransferase 1
JCVI_494	2	*COmtl*	Quercetin 3-O-methyltransferase 1
EE558066	2	*Myb4*	R2R3 MYB protein MYB4
CD820923	2	*Myb4*	Transcription factor MYB4
JCVI_19557	3	*Myb4*	Transcription factor MYB4
JCVI_4870	2	*TT2*	Transparent testa 2 family isoform 1
EV080676	2	*4CL3*	Protein At1g65060
JCVI_15603	2	*4CL*	4-coumarate-CoA ligase-like protein
JCVI_19558	1	*4CL9*	Putative uncharacterized protein At1g20510

Table 3.continued

In general, we searched for genomic regions in which metabolic QTL (mQTL) seem to co-localize with expression QTL (eQTL) of annotated probes which correspond to candidate genes of the selected pathways. Additionally, the probes were classified in correspondence to the EST sequence origin: (1) *B.rapa*, (2) *B.napus*, (3) *B.oleracea* and (4) other Brassica species; this code was added after the annotated gene name. The eQTL results corresponding to each pathway are summarized in Table 4.

	Features on the array					Features with QTL				
Pathway	total	*B.rapa*	*B.napus*	*B.oleracea*	other	total	*B.rapa*	*B.napus*	*B.oleracea*	other
Carotenoids	68	16	49	3	-	29	5	22	2	-
Tocopherols	35	6	27	2	-	12	2	9	1	-
Folate	39	9	25	4	1	20	5	13	2	-
Glucosinolates	66	19	44	3	-	30	11	17	2	-
Phenylpropanoids	69	18	47	3	-	32	3	26	2	1
Flavonoids	120	20	86	14	-	34	6	23	5	-

Table 4. Expression QTL detected in DH68.Results are indicated per metabolic pathway and according to the origin of the probe identified as candidate gene within each pathway.

Carotenoids

A total of 68 probes were selected from the microarray data as being representative of candidate genes of the carotenoids pathway and were further tested for their significance by regression analysis. These probes correspond to ESTs, which have different Brassicas origin, with the largest number in this case from *B. napus* (n=49). A total of 29 probes had at least one significant QTL and within these 22 had a *B. napus* origin.

When the results of both analyses are combined we can identify regions which harbor potential colocalization of QTL as observed in Supplementary figure 1. For example, in linkage group A03 the markers with an eQTL for gene eLCY:2, mapped in an interval which spans a similar region to the QTL found for lutein. In linkage group A05 the markers with an eQTL for BLCY:1 mapped in a similar genomic region found for the QTL of lutein and β-carotene. Finally, in linkage group A09 the probes corresponding to ZE:2 and CRTISO showed a significant QTL in a region where a QTL for violaxanthin and a QTL for neoxanthin were identified.

Tocopherols

From the complete eQTL output we selected a set of thirty-five probes, which are known to be involved in the tocopherol pathway. From this set of probes the largest number had a *B. napus* (n=27) origin. After regression analysis a total of 12 probes had at least one significant QTL and within these nine had a *B. napus* origin.

The combined results allowed us to identify regions in which co-localization of mQTL and eQTL for tocopherols was observed (Supplementary figure 2). In linkage group A03 the markers with an eQTL for gene VTE4:2 and GGPPS:2 mapped in an interval which span a similar region to the QTL found for β-tocopherol. In linkage group A05 eQTL loci for PDS:2, PDS:3 and GGPPS:2 mapped in a similar QTL region found for γ- and α-tocopherol.

Although we found mQTL for γ- and α-tocopherol in linkage group A02 none of the selected candidate genes showed an eQTL neither in the same region nor on that linkage group.

Folates

A total of 39 probes were selected from the microarray data as candidate genes that could potentially lead to the production of folate in *B. rapa*. The majority of these probes again have a *B. napus* (n=25) origin.

Twenty out of 39 probes had at least one significant QTL. The highest logp value being the one found for DHFR-TS1:2 in linkage group A03. However, when the results of the eQTL and mQTL are combined we could not identify regions in which both types of QTL co-localize (Supplementary figure 3). Either the identification of candidate genes did not include all the potential regulators or the relevant transcription factors were not included in the analysis and therefore we missed information on potential regulators in the A09 genomic region.

Flavonoids and Phenylpropanoids

The general phenylpropanoid metabolism leads to the synthesis of the major subgroups of flavonoids. To assess which genomic regions are responsible for the regulation of the flavonoids variation we performed the eQTL analysis with combined expression information obtained from probes that were annotated as candidate genes of the phenylpropanoids and flavonoids pathway.

Out of the total number of probes, 69 were identified as representative for genes from the general phenylpropanoids pathway and 120 were identified as genes leading to the production of flavonoids. For both groups of probes the EST information had mostly a *B. napus* origin: 47 out of 69 for the phenylpropanoids pathway and 86 out of 120 for the flavonoids pathway.

To uncover which loci are controlling the variation in flavonoids within this DH population we combined the eQTL and mQTL data. Based on the whole genome analysis of the LC-MS centrotype data we previously selected a set of 14 centrotypes to be further annotated. The set was identified as a group of flavonoids and mQTL analysis was further performed with this subset of centrotypes (Table 2). The QTL analysis resulted in 48 significant mQTL (logp > 4.37, FDR=0.05) for these metabolites. Strikingly, in comparison with the results obtained for other targeted pathways we obtained more single marker signals in the eQTL analysis (Supplementary figure 4). This will be considered when the eQTL and mQTL results are compared because it suggests that in the case of the transcriptomics analysis the

probe signals could be the result of mishybridizations and/or that significance threshold values for detecting an eQTL need still be adjusted to avoid false positive results.

Even with these considerations, and although the region in linkage group A07 was detected as a hotspot for the mQTL, we were not able to find any suitable candidate gene in that location. Like for the folates, possibly the identification of candidate genes did not include all the potential regulators or the relevant transcription factors, which also may have caused the failure to detect regulators in the A07 genomic region.

Regulatory network of the glucosinolate biosynthesis

To further analyze the complete regulatory network of a pathway we focused on the very well characterized pathway in the Brassicaceae that leads to the glucosinolate biosynthesis. The variation and regulation of the glucosinolate content has been widely studied (Gigolashvili et al. 2007,2008, Kliebestein et al. 2001a,b,2006, Halkier and Du 1997, Mithen et al.2000) at different developmental stages and organs and with different approaches (Hirai et al. 2007). In Arabidopsis in a Ler vs Cvi RIL population two major loci were found through QTL analyses to explain the observed variation for most of the aliphatic glucosinolates (Kliebenstein et al. 2001). The MAM locus was responsible for the variation in length chain (Kroymann et al. 2001) and the AOP locus was responsible for the variation in side chain modification (Kliebenstein et al. 2001b). By using an association mapping approach in *Brassica rapa* the regions containing these loci have also been found to be significant in the regulation of glucosinolate content (this thesis, Chapter 4). In this same study (Chapter 4) in addition to *MAM* and *AOP* the transcription factor *Myb28* was found to play an important role in the variation of the glucosinolate profile in the *B. rapa* accessions.

In our study in a double haploid mapping population of 92 lines a total of forty-seven markers were detected as mQTLs for 15 glucosinolates out of the 19 identified from the LC-MS profiling.

The mass signals for the 15 glucosinolates showed mQTLs that co-localized in genomic regions on linkage groups A03 and A09. Previously in an effort to enrich the genetic map with informative markers related to the glucosinolate pathway we

mapped in linkage group A03 two microsatellite markers that were linked to the *Myb28* and *MAM* genes, located at position 57.98cM and 66.81cM respectively. Within this region in linkage group A03 spanning 22cM between 50.88cM (marker M69) and 72.77cM (marker M77) the heptyl (I,II,III) and glucoraphanin showed a QTL peak with the highest logp value at position 63.26cM (M74). Furthermore, hexyl III glucosinolate also showed an mQTL within the same region but with a peak specifically located at the position of the Myb28 marker.

In linkage group A09 within a genomic region between 78.95cM and 86.27cM a different group of glucosinolates showed overlapping mQTLs. For example, glucobrassicanapin and sinigrin showed the highest logp value at 78.95cM (M218), glucoerucin and gluconapin showed a QTL peak value at position 72.42 (M214) and methsulphonylbutyl showed a QTL peak 86.27(M222).

The network analysis of the mQTL result indicates a high genetic correlation between glucosinolates that have an mQTL in the same region. In the case of linkage group A03 mQTL were detected for the long chained glucosinolates and in A09 the mQTL were mostly detected for C3, C4 and C5 glucosinolates and their modified forms (Supplementary figure 5)

For eQTL analysis a total of sixty-six probes were selected as representative of candidate genes for this pathway with the highest number (n=44) of sequences of *B. napus* origin. From this list of candidate genes 30 showed at least one eQTL (Figure 4). The genes *MAM:2, MAM:1* and Aconitase:3 showed an eQTL in the region on linkage group A03 which colocalized with the identified mQTLs for long chained aliphatic glucosinolates in that same linkage group. Additionally, the genes A*OP:2, AOP:1(2), BCAT4:2(2), CYP83:A1:2 ,CYP83:A1:1, CYP79B3:1, FMO2:2, FMO1:2* and *SUR1:1* showed eQTLs in A09 which co-localizes with the identified mQTLs for modified glucosinolates in that linkage group. Cluster analyses of the microarray expression data of those genes (Figure 5) indicate that multiple genes follow a similar expression pattern. The genes that follow this pattern include all the genes, which showed an eQTL in A09.

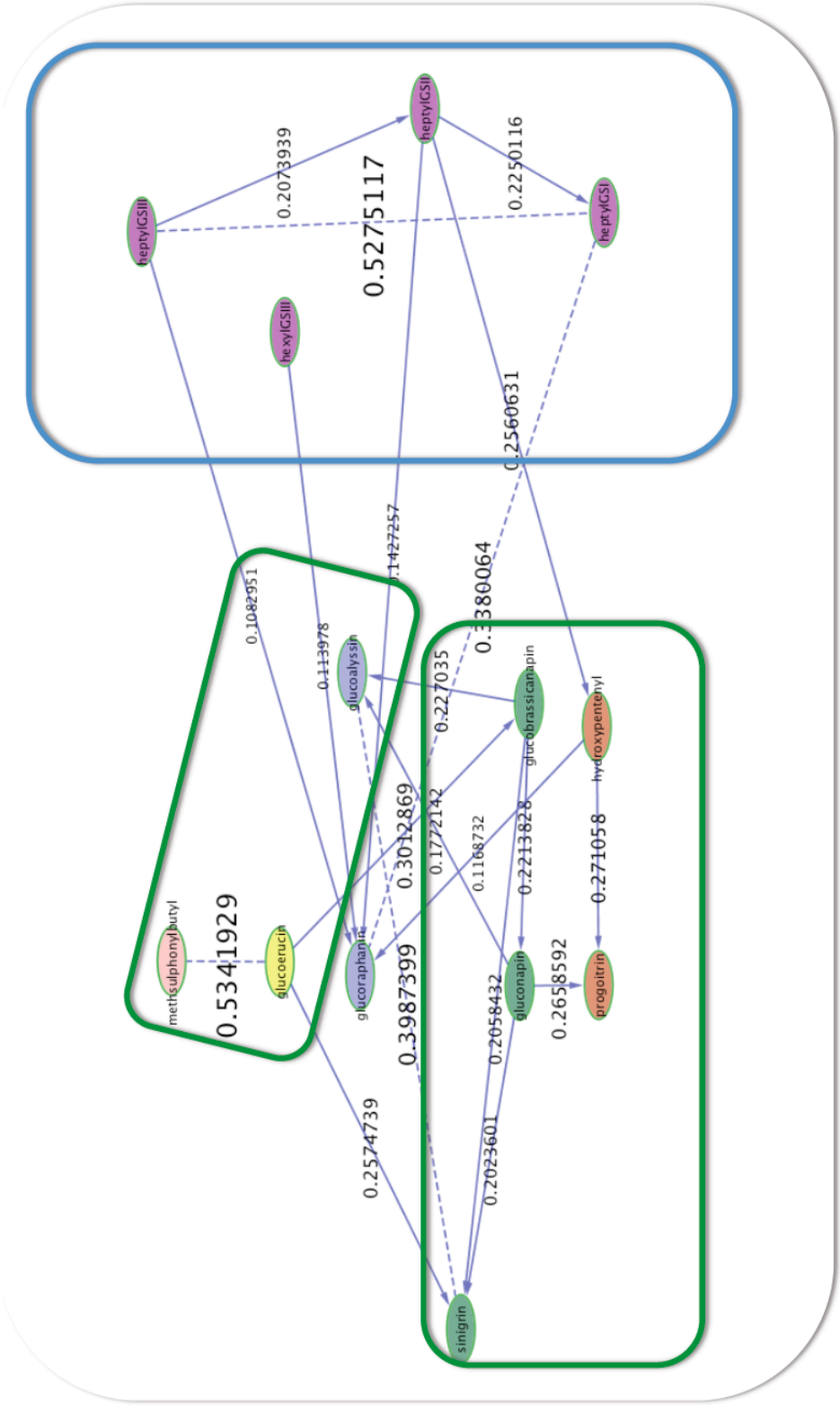


Figure 4. Metanetwork second order correlations results between glucosinolates detected in the DH68 population. Colors indicate glucosinolates regulated at the same step in the pathway. Enclosed in green are glucosinolates with an mQTL in A9 and in blue are enclosed the glucosinolates with an mQTL in A03.

To determine if the expression regulation of these genes, which showed an eQTL is located at a distant (trans-regulation) or at a local (cis-regulation) level, it is necessary to compare the map position of the genes with the eQTL region by anchoring the genetic map to the physical map. However, at the time of this thesis there was no full *B.rapa* sequence infromation. Further studies with the aid of new sequence information (Xiaowu Wang, IVF-CAAS,China) will soon make such elucidations possible.

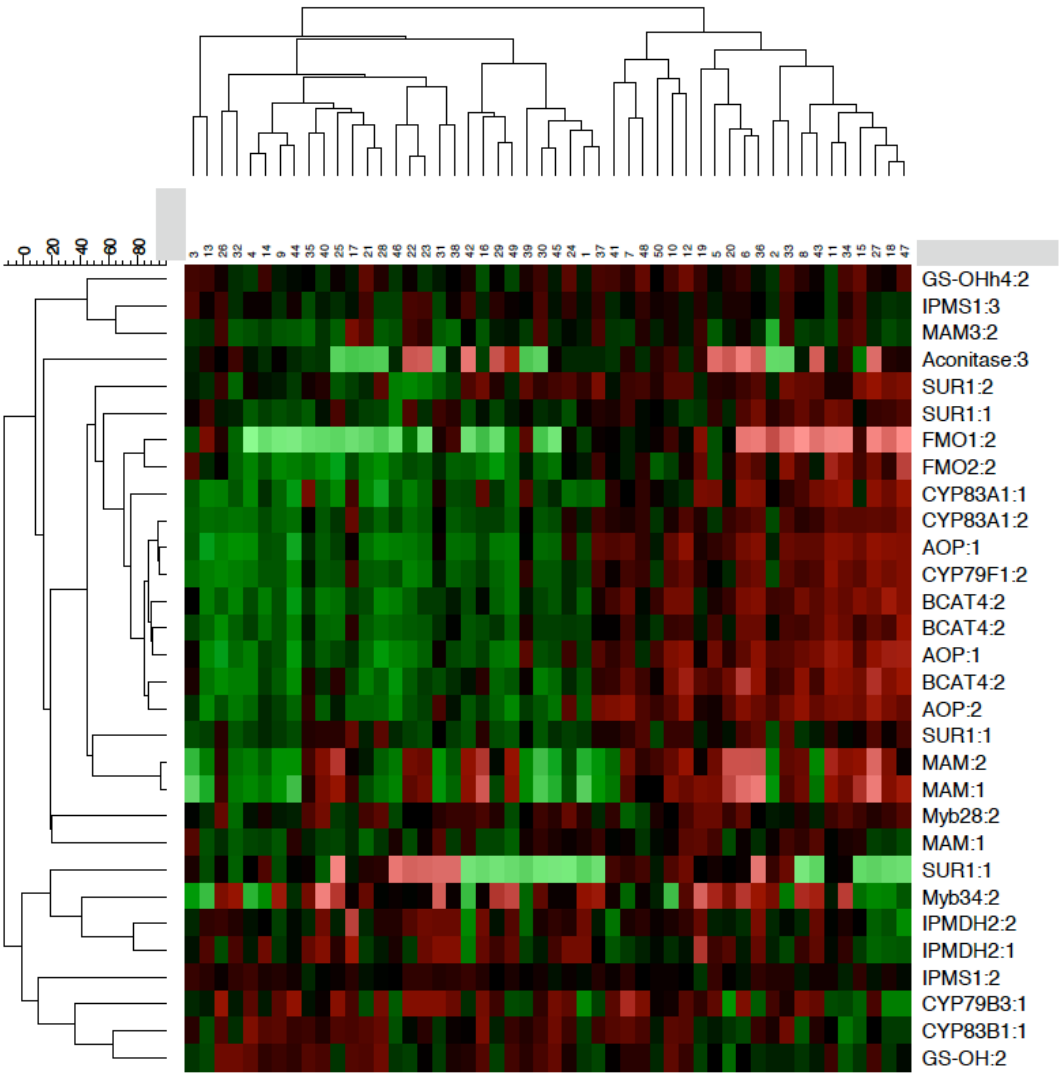


Figure 5.Cluster analysis of microarray intensity ratio result of the DH lines pairs for the probes that showed a significant eQTL.Names of annotated probes corresponding to candidate genes for the glucosinolate pathway are depicted on the right.

Discussion

A genetical genomics approach was chosen to gain insight in the genetics of the *B. rapa* metabolome. To illustrate the strength of this approach we selected a group of six biosynthetic pathways: carotenoids, tocopherols, folates, flavonoids and phenylpropanoids, in order to combine metabolic and expression QTL data obtained from a doubled haploid population (DH68). These pathways have been very well studied in *Arabidopsis thaliana* and most of the genes have been characterized (Gachon et al.2005).

Traditionally, the synteny between Brassica and Arabidopsis has assisted in the prediction of candidate genes in genomic regions where a phenotypic QTL has been detected (Lou et al. 2008, Schranz et al. 2006). In the present study, the metabolites were detected in leaves of a *Brassica rapa* doubled haploid population through targeted methods to quantify isoprenoids and folate and also following an LC-MS untargeted profiling approach with 14 centrotypes representing flavonoids and 22 glucosinolates annotated.

The use of a metabolic targeted approach to identify isoprenoids and folate allowed us to further analyze the QTL results and predict a group of candidate genes based on synteny.

On the other hand QTL analysis of the untargeted metabolites was performed with a group of 228 centrotypes, which potentially represent different compounds. The QTLs detected for the centrotypes clustered in a genomic region in linkage group A7. The co-localization of mQTLs in a cluster was an indication that these centrotypes possibly shared a common genetic regulator. Further identification of 14 centrotypes, which mapped in this QTL cluster, demonstrated that these are biochemically related to the flavonoid pathway. Thus, in our study a preliminary mQTL analysis was of aid to reduce our dataset of untargeted metabolites and to identify a region with an important genetic function for the regulation of the flavonoid pathway.

To identify influential genes and gene products the genetical genomics approach has emerged as a tool to combine expression profiling with molecular marker analysis through the use of quantitative trait loci (QTL) analysis in a segregating population (Jansen and Nap 2001). For our study we profiled the transcript abundance of the 92 DH lines with a newly developed microarray using unigenes assembled from EST sequences from mainly *B. napus*, *B. rapa* and *B. oleracea* (Trick et al. 2009).

The direct comparison of metabolite QTL and eQTL maps has shown the predictive capacity of eQTL to detect candidate genes for phenotypic differences in Arabidopsis (Wentzell et al. 2007).

The comparison between metabolite QTL and eQTL in our study revealed co-localization of both in many cases. The predictive value of the QTL comparison through co-localization was very successful in the case of the isoprenoids and the glucosinolate pathway but not complete either in the case of the folate or for the cluster of mQTLs found for the flavonoid centrotypes.

Our inability to identify eQTLs related to mQTLs could be caused by the lack of inclusion of genetic regulators like transcription factors in our ratcheted analysis. The metabolic specific QTL can also be related to loci, which control the flux of substrates, post translational regulation of enzymes or newly identified regulatory genes.

Nonetheless, for the aliphatic glucosinolates the eQTLs were consistent with the results obtained for the metabolic QTLs. The metabolic network results clearly suggest a difference in the correlation between glucosinolates. Further comparison with the expression differences results showed that co-localization existed with eQTLs found for *AOP* and *MAM*.

The analysis of expression differences of a selected group of candidate genes for these selected pathways helped us to focus on well known and characterized genes. Although these candidate genes are potential regulators of important pathways leading to metabolites of interest, further considerations have to be included in the analysis of this type of data. The triplication of the genomes in Brassica has to be taken into account, with the presence of paralogues in both the A and C genomes. The fact that this microarray was assembled from EST sequences from different Brassicas will very likely influence the hybridization results because of the mRNA sequence diversity in probe regions (Alberts et al. 2007).

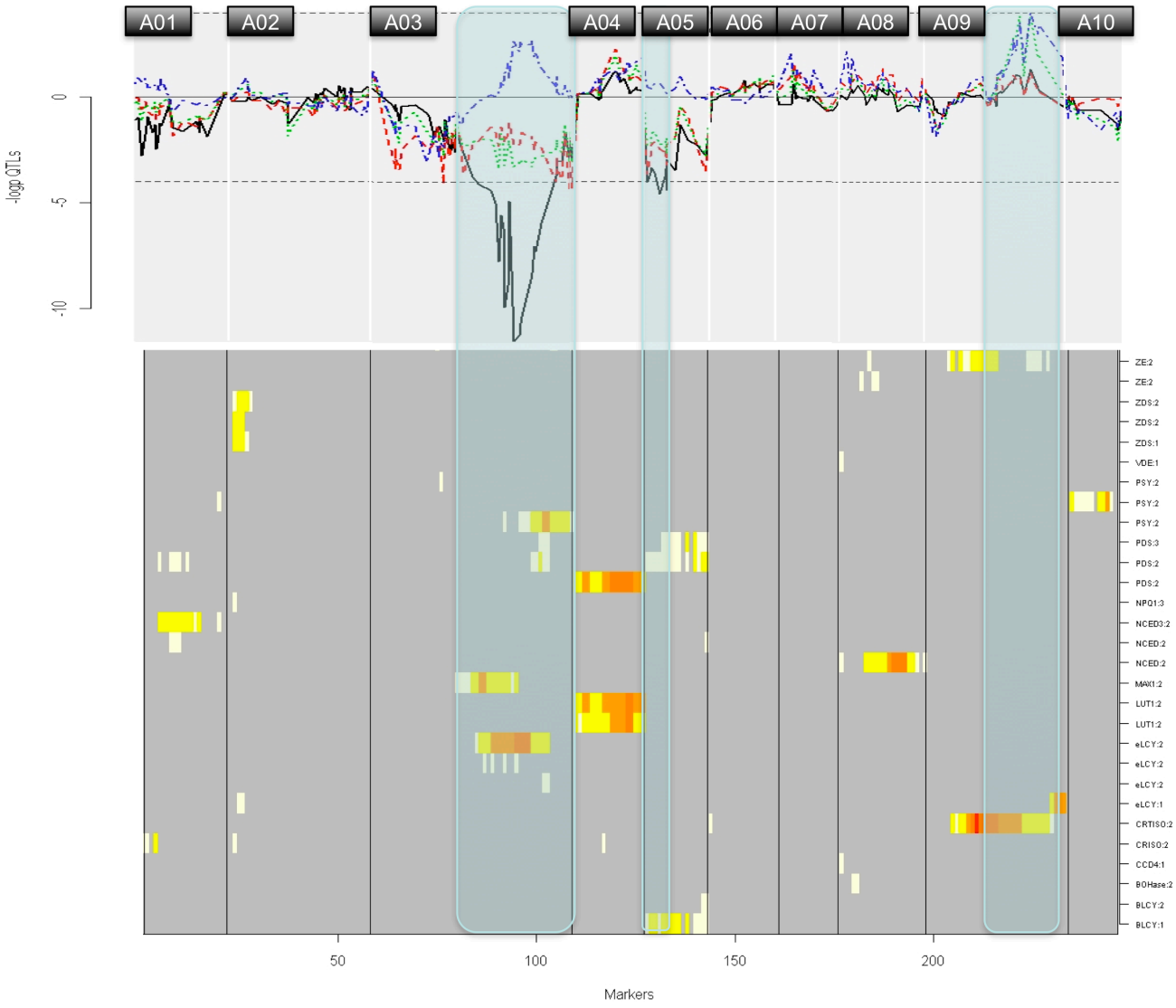
Additionally in our study we selected probes representing structural genes, which probably act as cis-acting factors. In Arabidopsis and barley it has been reported that trans-QTL are more abundant than cis-QTL (West et al. 2007, Chen et al. 2010). Detailed analysis and annotation of all 24,850 probes with eQTL co-localizing with mQTL can lead to detection of these regulating genes. Furthermore, to identify genes and the types of regulation underlying each QTL , it will be necessary to anchor the genetic map to a physical map (Keurentjes et al. 2007). Although we believe the data

generated in our study is valuable for the elucidation of the genetic regulation of the metabolome, additional work is still needed. Currently the amount of *B. rapa* genome sequence data is growing (http://www.brassica.info); this together with bioinformatic tools capable of handling such large datasets will make it possible to maximize the information obtain from this type of approaches.

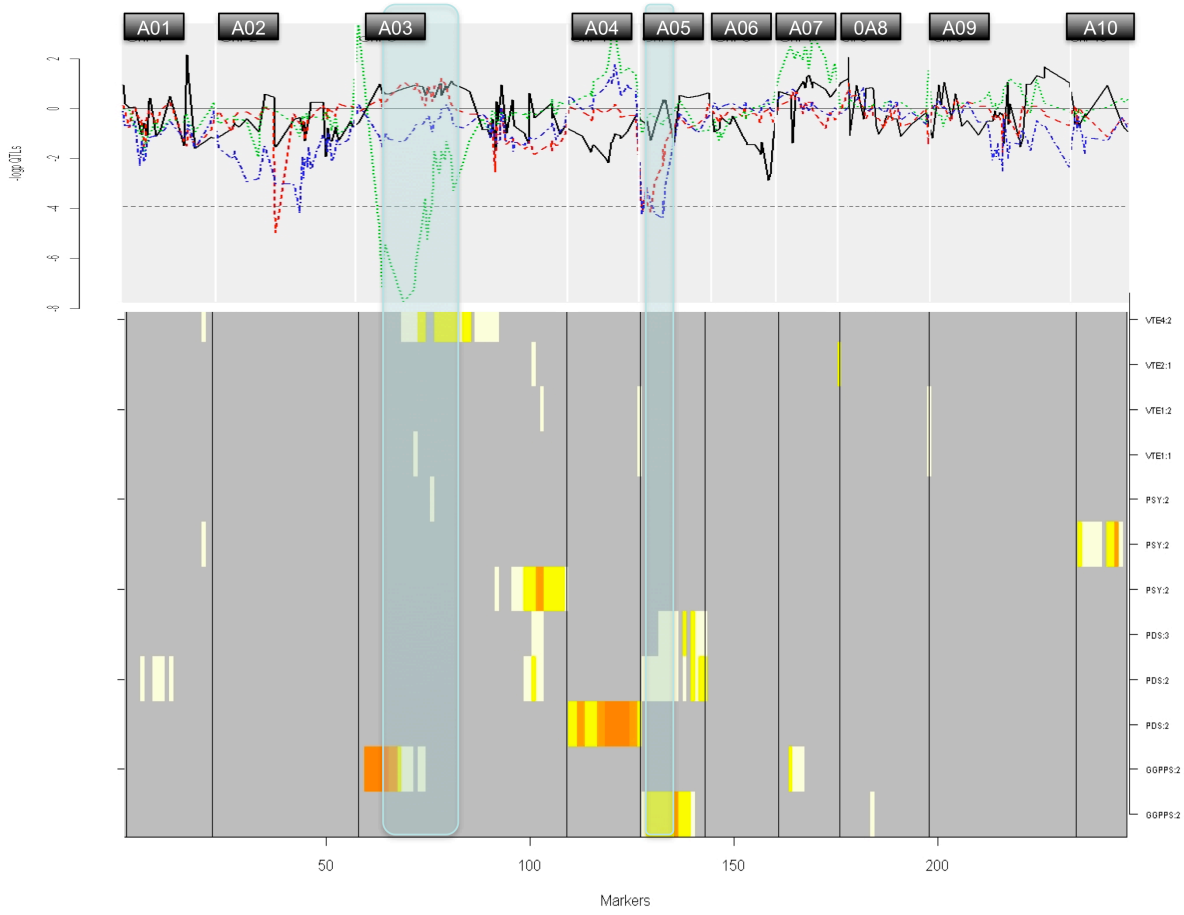
Acknowledgements

We would like to thank Harry Jonker and Yvonne Birnbaum for their help in the isoprenoids and folates analyses.

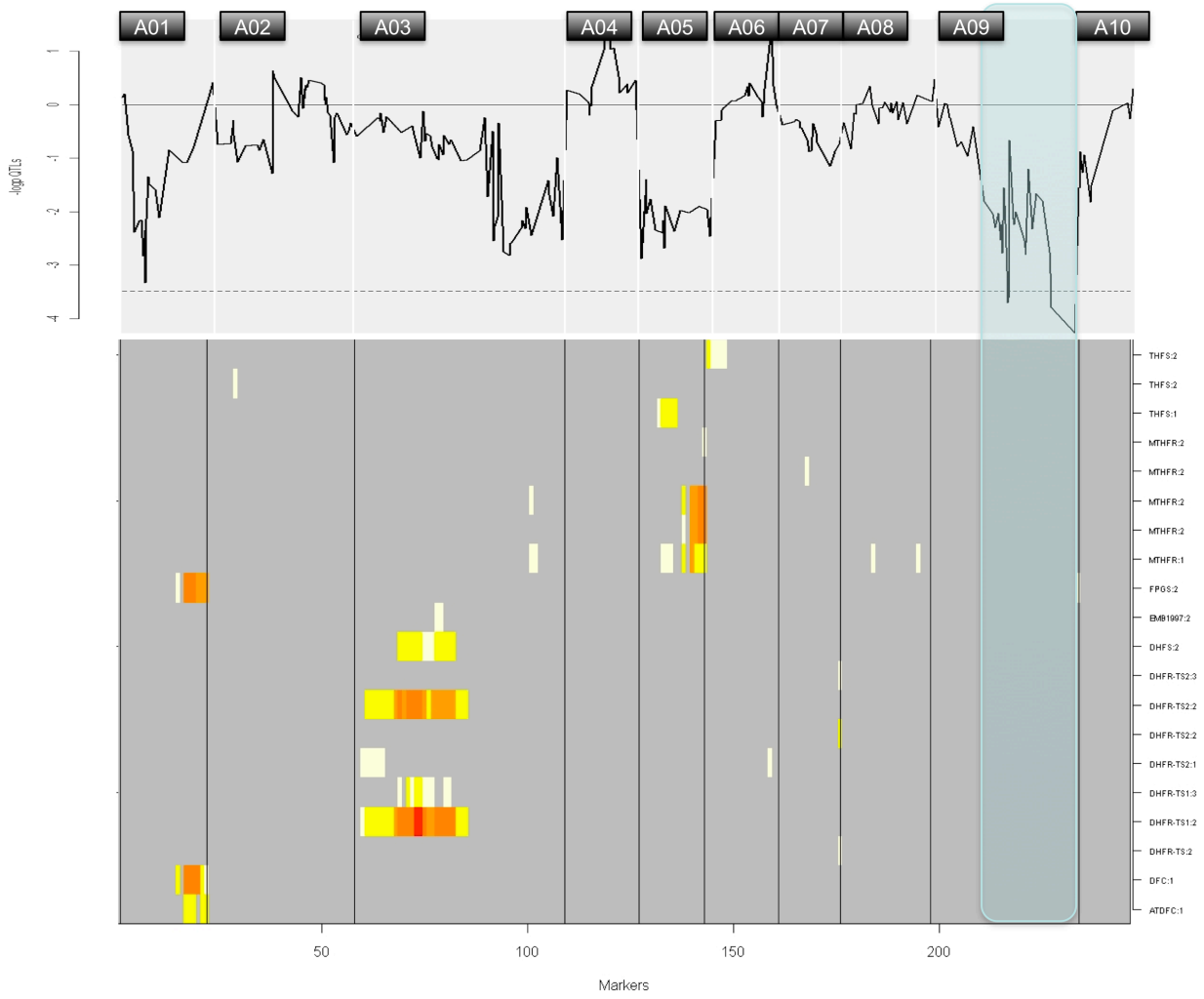
This research was funded by the IOP Genomics project "Brassica vegetable nutrigenomics" IGE 05010.



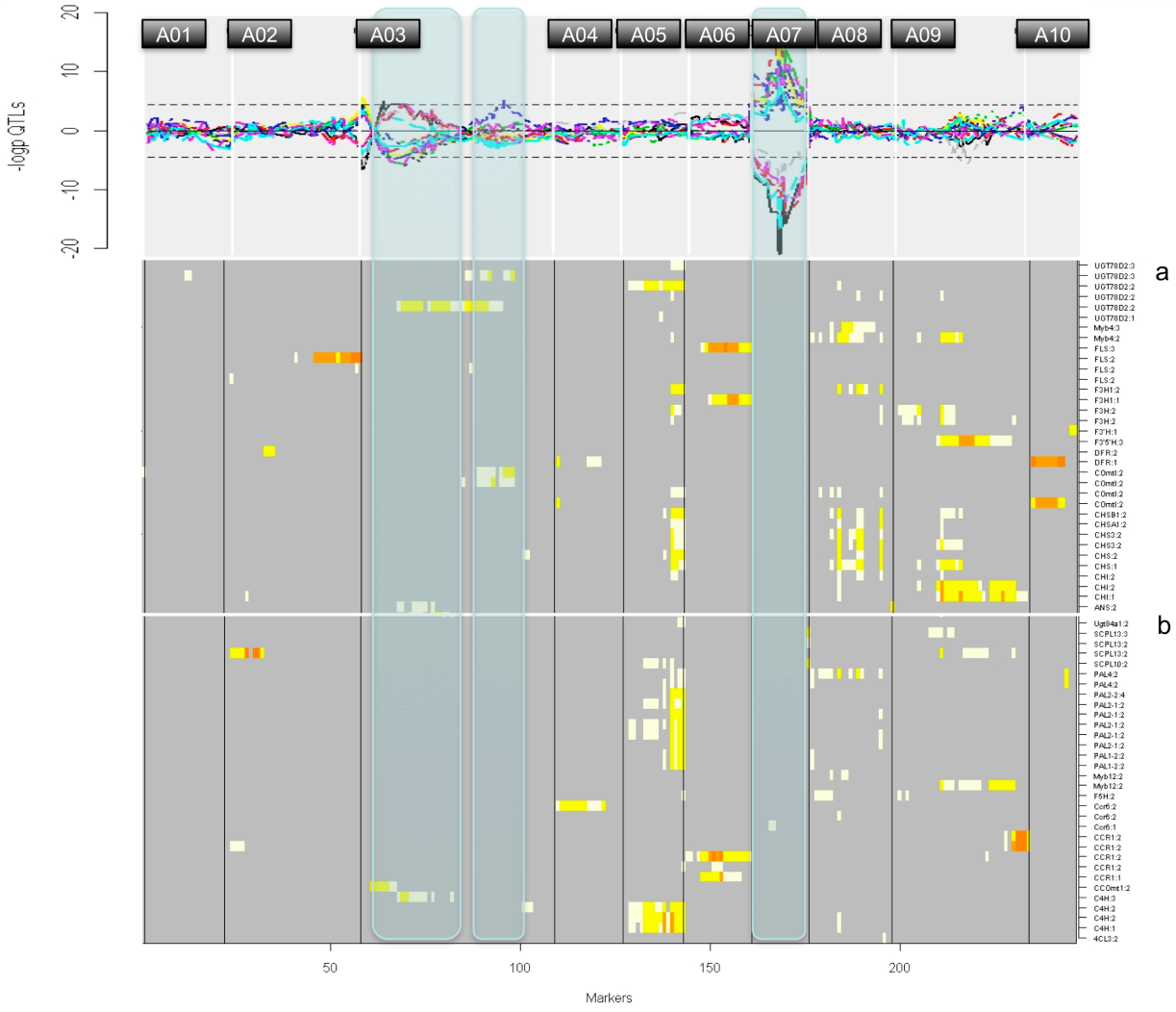
Supplementary Fig 1.. QTL analysis results of the carotenoids pathway data. Top indicates QTL metabolic profiling and the bottom shows QTL expression results of probes representing candidate genes, names are listed on the right. Light Yellow (logp=3),yellow(logp=3-5),orange (logp=5-7),darkorange (logp=7-10),red(logp=>10)



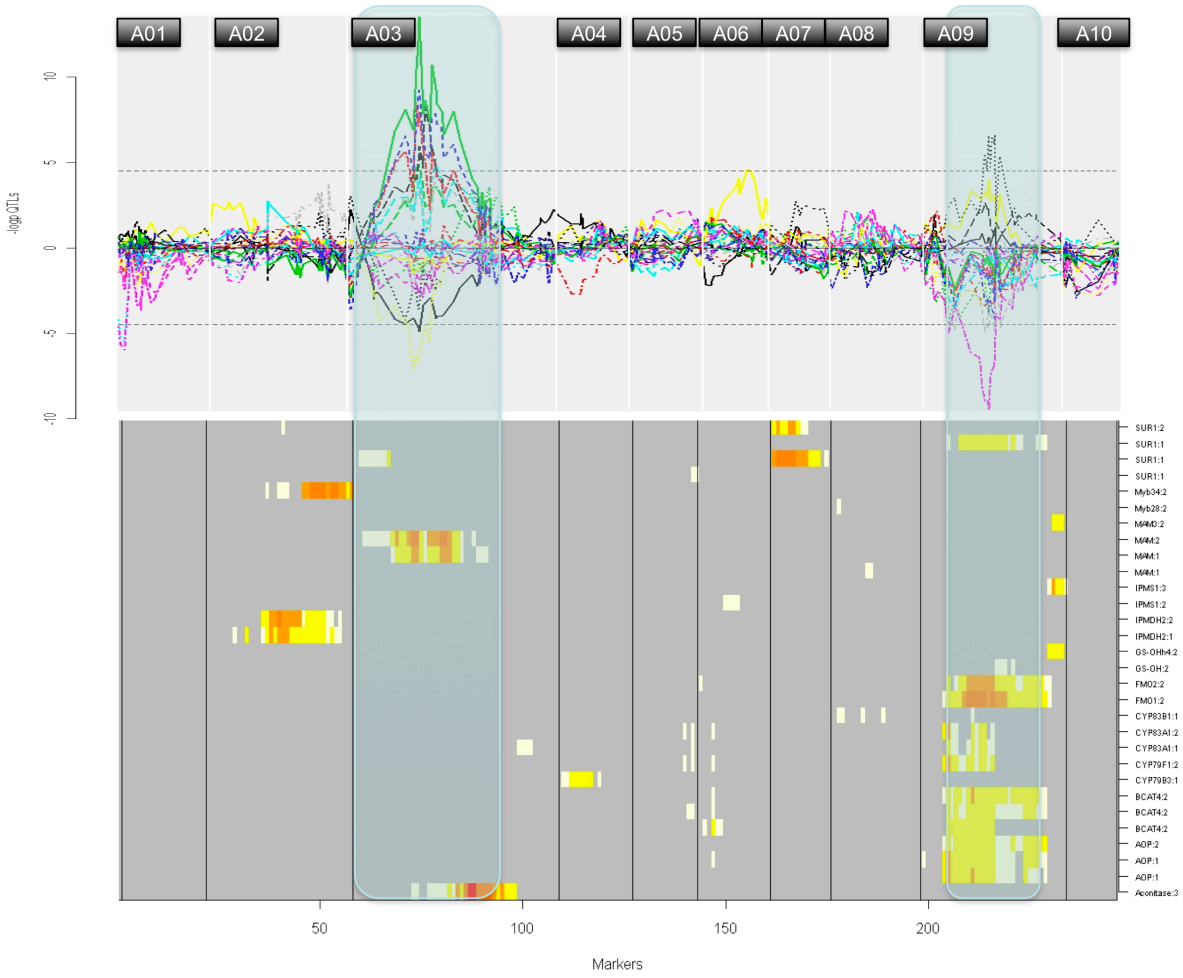
Supplementary Fig 2. QTL analysis results of the tocopherols pathway data. Top indicates QTL metabolic profiling and the bottom shows QTL expression results of probes representing candidate genes, names are listed on the right. Light Yellow (logp=3),yellow(logp=3-5),orange (logp=5-7),darkorange (logp=7-10),red(logp=>10)



Supplementary Fig 3. QTL analysis results of the folates pathway data. Top indicates QTL metabolic profiling and the bottom shows QTL expression results of probes representing candidate genes, names are listed on the right. Light Yellow (logp=3),yellow(logp=3-5),orange (logp=5-7),darkorange (logp=7-10),red(logp=>10)



Supplementary Fig 4. QTL analysis results of the flavonoids pathway data. Top indicates QTL metabolic profiling and the bottom shows QTL expression results of probes representing candidate genes, names are listed on the right. A.flavonoids candidate genes and B.phenylpropanoids candidate genes Light Yellow (logp=3),yellow(logp=3-5),orange (logp=5-7),darkorange (logp=7-10),red(logp=>10)



Supplementary Fig 5. QTL analysis results of the glucosinolates pathway data. Top indicates QTL metabolic profiling and the bottom shows QTL expression results of probes representing candidate genes, names are listed on the right. Light Yellow (logp=3), yellow (logp=3-5),orange (logp=5-7), darkorange (logp=7-10), red(logp=>10)

Chapter 6

General discussion

Within this thesis different approaches are followed to unravel the genetics of the *Brassica rapa* metabolome in a group of accessions comprising a core collection, and in a Doubled Haploid mapping population. To gain insight into the genetic regulation of metabolites we took advantage of the genetic and phenotypic diversity that can be found within and between the different *B. rapa* morphotypes. In this chapter the established relationships between the genetic and metabolic content in these different populations are discussed.

Genetic and phenotypic multivariate analysis for the design of a core collection

The genus *Brassica* has a long history of worldwide cultivation and comprises a large and diverse group of important vegetable, oil, fodder and condiment crops. *B. rapa* is the most diverse species with the longest cultivation history, and encompasses leafy vegetables, turnips and oils. The leafy vegetables include heading Chinese cabbage, pak choi, mizuna and mibuna, komatsuna and neep greens. The turnips include vegetable and fodder types and the oil types include both the annual oil types and biannual oil types (Zhao et al. 2005 and Chapter 1).

In order to characterize the phenotypic and genotypic diversity of a group of accessions we followed a multivariate analysis. Hierarchical cluster analysis based on molecular markers (AFLP and microsatellites) identified three groups, very similar to the groups found in previous studies in *B. rapa* (Zhao et al. 2005 and 2007). A group with Chinese cabbage and pak choi and other Asian leafy types, but also some Japanese turnips, from mainly Asian origin, a group with European turnips, broccoletto and some oil types from European origin, and a small group of annual oils from Indian origin. The group number and composition obtained with molecular markers was highly correlated to the groups based on morphological traits of vernalized accessions. However, the value of allele exchange (admixture) and relatedness among individuals could only be obtained when molecular data was analyzed with STRUCTURE. In further studies we included the STRUCTURE Q matrix as a correction term for confounding in association studies (Chapters 3 and 4). For the breeding of nutritional-related metabolites it is necessary to differentiate subgroups of morphotypes and accessions within these subgroups to identify lines

that carry interesting metabolites or metabolite concentration for crop improvement and/or for the selection of parental lines to create populations for metabolic QTL studies.

The analysis of metabolite data proved to be very valuable in the classification of morphotypes and comparable to the genetic profiling (Fig. 1).

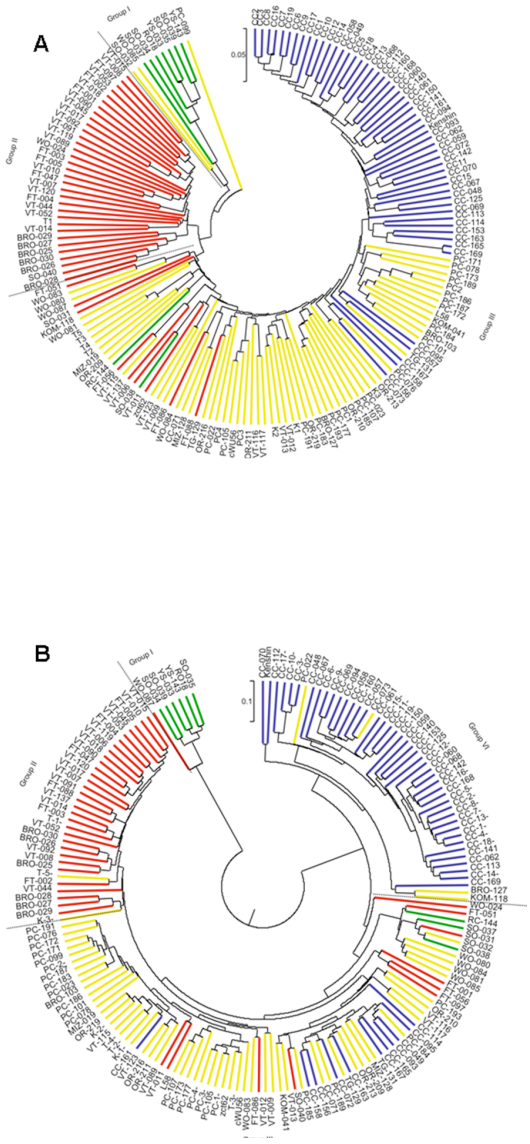


Figure 1.Hierarchical cluster UPGMA obtained with (A) 412 molecular markers and (B) 5546 (LC-MS mass-scan signals) metabolites; the colors indicate the four population subgroups as defined by STRUCTURE.

In terms of the analytical tools Random Forests proved to be valuable to select a small group of variables from the unidentified LC-MS data that represent or define sub populations. Among the metabolites that could be identified after selection of

variables that can define or differentiate subpopulations are: isopropyl glucosinolate, methylpropyl glucosinolate, hexyl glucosinolate 2, caffeoylquinic acid, chlorogenic acid, coumaroylquinic acid, quercetin3- (2-feruloylsophoroside) 7-diglucoside and kaempferol coffeoyl tetraglucoside

Currently, several genebanks have large well-evaluated *Brassica* collections. The ECP/GR (European cooperative programme for plant genetic resources) *Brassica* database (http://documents.plant.wur.nl/cgn/pgr/brasedb/) contains passport data of most of the European *Brassica* collections (Hintum and Boukema 1993; Boukema and Hintum 1998). The CGN (Center for Genetic Resources, the Netherlands, http://www.cgn.wur.nl/UK/) harbors a collection of cruciferous crops. Other resources for germplasm include the national plant germplasm system of the USA (Genetic Resource Information Network (GRIN)) the plant gene resources of Canada (GRIN-Canada) and China has collected more than 7000 oilseed and vegetable *Brassica* accessions in its genebank (Wu et al. 2008).

Presently an effort to define a *Brassica rapa* Diversity Fixed Foundation Set (BrDFFS) representing the genetic diversity within *B. rapa* is ongoing through a collaboration between Wageningen UR Plant Breeding (Wageningen University & Research Center, Wageningen, The Netherlands), Vavilov Institute (St Petersburg, Russia) and IVF-CAAS (Beijing, China) (Zhao et al, accepted in Genome). A general approach for constructing such collections is the selection of a collection after clustering of groups based on genetic distances (Hu et al. 2000). After selection of the accessions a core collection can be constructed with equally sized subgroups with enough within variation to conduct association studies. In future studies our results can serve as a valuable reference for the selection of *Brassica rapa* accessions from groups defined by multivariate analyses. Furthermore we conclude that the high level of admixture between groups expressed at phenotypic and genetic level ensures that it is possible to design a set of accessions that can harbor enough variation for association studies using all subpopulation data combined.

Natural population mapping: a complementary approach and new insights into candidate genes

Brassica rapa crop types are the result of different breeding and domestication histories around the world. Because adaptation, genetic drift, domestication or

selection influence the level of relatedness between types, confounding can be a significant problem in association studies. In order to reduce the rate of false-positive associations it is then necessary to correct for population structure. Previously, we identified with STRUCTURE the presence of four subpopulations, which showed correlation with the origin and morphotypes of *B. rapa* (Chapter 1). Before further following any association method we first evaluated the Principal Coordinates Analysis (PCA) in order to introduce an additional confounding correction term. This approach together with the insertion of only the kinship matrix in the model did not prove to be relevant for our further studies because the result did not show significant differences if compared with a model without correction (Chapter 3).

In general there are two categories of approaches in association studies: candidate gene association mapping and genome wide association mapping (Zhu et al. 2008). In Chapter 3 we considered whole genome association mapping by investigating the genetic association between markers with random positions over the *B. rapa* genome and tocopherols, carotenoids, chlorophylls and folate in a core collection of 168 *B. rapa* accessions. Many of these metabolites showed significant differences in amounts between the different sub populations. 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 AFLP, *Myb*-motif targeted and microsatellite (SSR) markers were included in the analysis. Additionally, for comparison and as a complementary method, we introduced the use of Random Forests for association studies in this type of plant collections. We selected 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 healthy compounds. These markers met different criteria and eight markers were kept as significant after multiple testing correction (q-value) of the results obtained with the model with kinship and Q matrix (STRUCTURE) correction. Furthermore, sixteen markers selected across methods, including Random Forests, are considered as the most promising candidates for marker assisted selection and validation. Within this study we showed the feasibility of the `whole genome association mapping in a structured population of *B. rapa* with limited number of markers. However, because linkage disequilibrium differs between genomic regions and the presence of subpopulations it was not possible to calculate LD decay.

To study the variation in glucosinolate content in the same group of accessions we followed a different approach profiling SSRs from BACs containing candidate genes for the glucosinolates pathway. The profiling of microsatellites allowed us to capture the allelic variation across subpopulations and along the linkage group A03.

The association method was similar to the one applied in Chapter 3, which included the kinship and Q matrix of STRUCTURE in the statistical model. We focused on the genetic dissection of a major QTL in A03 for glucosinolate variation previously identified in a doubled haploid population from a cross between a pak choi and a yellow sarson, which explained a large proportion of the variation in several aliphatic glucosinolates and was consistent over two seasons. Within our study we profiled microsatellite markers linked to the transcription factor *Myb28* and the gene involved in aliphatic GLS chain elongation (*MAM*), which were identified through sequence information as the major candidates for the QTL region. The microsatellite linked to *Myb28* was considered as the major regulator of glucosinolate variation because of its association to 3-butenyl (gluconapin), the total content of 4C (tot4C) and 6C (tot6C) glucosinolates and to the ratio of hydroxylation (rat1). The association of the marker linked to *Myb28* to the ratio of hydroxylation reflects possible regulation at the 2-oxo acid dependent dioxygenase (*GLS-OH*) locus, which is responsible for the biosynthesis of hydroxylated alkenyl glucosinolates. Besides the markers linked to *Myb28* and *MAM*, designed within BACs and contigs with these candidate genes, we profiled markers across the linkage group A03.

In our study the BAC containing the marker KS6, which had a genetic distance of 2.7, cM in the physical map of a Korean Doubled Haploid population used for the reference map (Mina Jin personal communication), to the BAC containing the AOP gene showed association to the total content of alkenyl glucosinolates. These results are of relevance for breeding if we consider that the regulation at the GLS-OH locus controls the conversion of 3-butenyl (gluconapin) to 2-hydroxy-3-butenyl (progoitrin), which is known to cause goiter among animals fed with rapeseed meal (Mithen et al. 2000).

Cruciferous vegetables are known to have potential health effects; the cancer preventing properties of these vegetables is due among others to the activity of 4-MSB isothiocyanates, which are derived from 4-MSB (glucoraphanin) glucosinolates by the action of endogenous myrosinase (Sarikamis 2009). Glucoraphanin can be converted into 3-butenyl (gluconapin) by AOP-2; in our study we found that

gluconapin is the most abundant glucosinolate and is regulated by several loci in A03.

The candidate gene approach was followed to unravel the genetic dissection of an apparently coordinated co-expression and interaction of the genes of the glucosinolate biosynthetic pathway in *B. rapa.* This approach was very successful in separating the roles of the closely linked *Myb28* and *MAM* genes contained within the QTL region previously identified, and in identification of three additional candidate genes *AOP* and *GS-OH* involved in side chain modification and *Myb29* in transcriptional regulation.

Metabolomics and transcriptomics: the genetical genomics approach in *B. rapa*

To identify genetic loci that explain the variation in secondary metabolites in leaves of five-week-old plants we analyzed 92 lines of a doubled haploid population developed form an F1 cross between a yellow sarson and a pak choi. We followed an untargeted metabolic profiling though LC-MS analysis and a targeted approach to identify the isoprenoids and folate composition of the mapping population. QTL analysis of the untargeted metabolites was performed with a group of 228 centrotypes, which potentially represent different compounds and a group of annotated glucosinolates from the LC-MS data. The analysis of the genomic distribution of these QTL showed that they were not evenly distributed over the *B. rapa* genome because 112 of the 228 centrotypes mapped on A07. The identification of some of these centrotypes demonstrated that these are biochemically related to the flavonoid pathway and therefore have a similar or common genetic regulation. However, it is still necessary to investigate whether this high correlation and QTL location of centrotypes in a hotspot on A07 can be due to technical factors. Traditionally the synteny between Brassica and Arabidopsis assists the identification of candidate genes in *B. rapa* as described in Chapter 1. When QTL results of the metabolic targeted approach of carotenoids, tocopherols and folate were analyzed it was possible to predict a group of candidate genes based on genome synteny and colocalization of QTL and candidate gene. However, the genetical genomics approach, which combines expression profiling with molecular marker analysis on a segregating population has made it possible to use quantitative trait loci (QTL) analysis for identification of influential genes and gene products (Jansen and Nap

2001). For our study we profiled the transcript abundance of the 92 DH lines with a newly developed microarray (Trick et al.2009) using unigenes assembled from EST sequences from mainly *B. napus*, *B. rapa* and *B. oleracea*. Using information from 78,278 probes we found that in total 24,850 probes were detected as significantly associated with a marker with a LOD score >3. Furthermore, the whole genome profile of transcript abundance in the doubled haploid population of *B. rapa* showed no evidence for hotspots of eQTLs : QTL were distributed over the whole genome with a few genomic regions with more clustered eQTL (Chapter 5). To gain some insight in the genetic regulation of the metabolites detected by LC-MS profiling (centrotypes) and the targeted metabolites we selected six different biochemical pathways: carotenoids, tocopherols, folate, glucosinolates, flavonoids and phenylpropanoids. From a total of 397 candidate genes related to these pathways, 157 showed at least one significant QTL (LOD>3, Table 1) Although these candidate genes are potential regulators of important pathways leading to metabolites of interest, several considerations have to be included in further analysis of this type of data. The presence of orthologues of the A and C genomes and paralogues within the A and C genomes resulting from the triplication of the genome in Brassica has to be taken into account. For example, this microarray was assembled from EST sequences from different Brassicas, which very likely will influence the hybridization results because of the mRNA sequence diversity in probe regions (Alberts et al. 2007).

Additionally in Arabidopsis and barley it has been reported that trans- QTL are more abundant than cis-QTL (West et al. 2007, Chen et al. 2010). In our study we selected probes representing structural genes, which probably act as cis-acting factors. Currently the *B. rapa* genome sequence data available is growing (http://www.brassica.info), this together with bioinformatics tools capable of handling such large datasets will make it possible to identify genes and the types of regulation underlying each QTL.

	Original					Interest				
Pathway	total	*B.rapa*	*B.napus*	*B.oleracea*	other	total	*B.rapa*	*B.napus*	*B.oleracea*	other
Carotenoids	68	16	49	3	-	29	5	22	2	-
Tocopherols	35	6	27	2	-	12	2	9	1	-
Folate	39	9	25	4	1	20	5	13	2	-
Glucosinolates	66	19	44	3	-	30	11	17	2	-
Phenylpropanoids	69	18	47	3	-	32	3	26	2	1
Flavonoids	120	20	86	14	-	34	6	23	5	-

Table 1. eQTL results for selected biochemical pathways. Candidate genes represented on the array are shown in the Original columns and number of candidate genes with significant QTL (LOD>3) are shown in the interest columns.

In this thesis we followed different approaches to dissect the genetics of complex traits focusing mostly on metabolite variation. We first demonstrated through multivariate methods that *Brassica rapa* carries a wide metabolic variation, which is found across the different morphotypes. Through different association mapping methods and following a genome wide and a candidate gene approach we were able to link this metabolic variation to genomic regions and to dissect QTL results found in DH populations (Chapter 4). Our assumptions about the candidate genes for the isoprenoids and the glucosinolate pathway located within these QTL regions can soon be confirmed with the availability of the genome sequence. However, further work will be needed to validate the function of these genes with for example knock out, over expression or mutational analyses. One major constraint at the moment is the lack of such validation methods in *B. rapa*, even though several groups work on optimization of transformation. The markers identified can still be useful for marker assisted selection of lines, for crossings or parental lines for breeding purposes.

The construction of an expanded core collection based on the data generated within this thesis and following similar multivariate approaches can increase the chance of making better choices for allelic and phenotypic variation. Based on the impact that population structure had on our results we suggest in *Brassica rapa* to follow an association mapping approach in which the confounding effects or population structure are not present. A core collection can be "designed" to contain enough within variation to follow association studies in subpopulations but should also increase the frequency of rare alleles. In *B. rapa* this could mean to rely on experimental crosses to create nested association mapping populations, as developed for maize (Yu et al. 2008) . If accessions of the small distinct group of spring oil types, which coincidentally are the ones that respond better in microspore culture and in vitro systems, are used as standard to cross with 20-30 selected lines of diverse morphotypes, followed by selfing or fixing through DH culture, large RIL populations can be developed.

Although the work done in Arabidopsis has served as a reference for *B. rapa* research for many years, sequence information has already shown that gene map prediction

based on macrosynteny is not very informative. Additionally, one important aspect in the complexity of the *B. rapa* genome is the presence of paralogues, Our preliminary results with the genetical genomics approach, co-localization with metabolic QTLs and the use of a multispecies microarray indicate a possible subfunctionalization of the paralogues in *B. rapa*. In this case the genetic dissection of traits must separate the main effect of the paralog and the interaction with other genomic regions. One research goal in this direction could be the development of introgression line populations, which should include QTL results in the selection of genomic regions for fine mapping and gene expression profiling.

The *Brassica rapa* community will soon benefit from the availability of sequence information from several crop types. It is then necessary to develop new methods for genetic dissection of traits and/or to follow approaches developed for much more complex plant systems with similar domestication and crossing patterns i.e maize than to rely solely on work done in model species like Arabidopsis.

References

Abdel-Farid, I., Kim, H.K., Choi, Y.H. & Verpoorte, R. 2007, "Metabolic Characterization of Brassica rapa Leaves by NMR Spectroscopy", *Journal of Agricultural and Food Chemistry,* vol. 55, no. 19, pp. 7936-7943.

Abdurakhmonov I. & Abdukarimov A. 2008, "Application of Association Mapping to Understanding the Genetic Diversity of Plant Germplasm Resources," *International Journal of Plant Genomics*, vol. 2008, 18 pages.

Agrama, H., Eizenga, G. & Yan, W. 2007, "Association mapping of yield and its components in rice cultivars", *Molecular Breeding,* vol. 19, no. 4, pp. 341-356.

Alberts, R., Terpstra, P., Li, Y., Breitling, R., Nap, J. & Jansen, R.C. 2007, "Sequence Polymorphisms Cause Many False *cis* eQTLs", *PLoS ONE,* vol. 2, no. 7, pp. e622.

Al-Shehbaz, I., Beilstein, M. & Kellogg, E. 2006, "Systematics and phylogeny of the Brassicaceae (Cruciferae): an overview", *Plant Systematics and Evolution,* vol. 259, no. 2, pp. 89-120.

Aranzana, M.J., 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 Genet,* vol. 1, no. 5, pp. e60.

Atwell, S., Huang, Y.S., Vilhjalmsson, B.J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A.M., Hu, T.T., Jiang, R., Muliyati, N.W., Zhang, X., Amer, M.A., Baxter, I., Brachi, B., Chory, J., Dean, C., Debieu, M., de Meaux, J., Ecker, J.R., Faure, N., Kniskern, J.M., Jones, J.D.G., Michael, T., Nemri, A., Roux, F., Salt, D.E., Tang, C., Todesco, M., Traw, M.B., Weigel, D., Marjoram, P., Borevitz, J.O., Bergelson, J. & Nordborg, M. 2010, "Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines", *Nature,* vol. 465, no. 7298, pp. 627-631.

Halkier B.A. & Du, L. November 1997, "The biosynthesis of glucosinolates", *Trends in plant science,* vol. 2, pp. 425-431(7).

Baik, H., Juvik, J.A., Jeffery, E.H., Wallig, M.A., Kushad, M. & Klein, B.P. 2003, "Relating Glucosinolate Content and Flavor of Broccoli Cultivars", *Journal of Food Science,* vol. 68, no. 3, pp. 1043-1050.

Bailey, C.D., Koch, M.A., Mayer, M., Mummenhoff, K., Jr, S.L.O., Warwick, S.I., Windham, M.D. & Al-Shehbaz, I.A. 2006, "Toward a Global Phylogeny of the Brassicaceae", *Molecular biology and evolution,* vol. 23, no. 11, pp. 2142-2160.

Batagelj, V. and Mrvar, A. (2003): Pajek - analysis and visualization of large networks. In: M. Juenger and P. Mutzel (Eds.): *Graph Drawing Software.* Springer (series Mathematics and Visualization), Berlin, 2003. 77-103

Beekwilder, J., van Leeuwen, W., van Dam, N.M., Bertossi, M., Grandi, V., Mizzi, L., Soloviev, M., Szabados, L., Molthoff, J.W., Schipper, B., Verbocht, H., de Vos, Ric C. H., Morandini, P., Aarts, M.G.M. & Bovy, A. 2008, "The Impact of the Absence of Aliphatic Glucosinolates on Insect Herbivory in Arabidopsis", *PLoS ONE,* vol. 3, no. 4, pp. e2068.

Beilstein, M.A., Al-Shehbaz, I.A. & Kellogg, E.A. 2006, "Brassicaceae phylogeny and trichome evolution", *American Journal of Botany,* vol. 93, no. 4, pp. 607-619.

Bino, R.J., Vos, C.H.R.d., Lieberman, M., Hall, R.D., Bovy, A., Jonker, H.H., Tikunov, Y., Lommen, A., Moco, S. & Levin, I. 2005, "The light-hyperresponsive *high pigment-2*^{dg} mutation of tomato: alterations in the fruit metabolome", *New Phytologist,* vol. 166, no. 2, pp. 427-438.

Botstein, D., White, R.L., Skolnick, M. & Davis, R.W. 1980. "Construction of a genetic linkage map in man using restriction fragment length polymorphisms."American Journal of human genetics, vol 32, no.3, pp.:314-31

Boukema, I.W., & Th.J.L. van Hintum ,1998. The European Brassica Database. Proceedings of an International Symposium on Brassicas. Acta Horticulturae 459. ISHS 1998. pp 249-254.

Breiman, L. 2001, "Random Forests", *Machine Learning,* vol. 45, no. 1, pp. 5-32.

Breseghello, F. & Sorrells, M.E. 2006, "Association Analysis as a Strategy for Improvement of Quantitative Traits in Plants", *Crop Science,* vol. 46, no. 3, pp. 1323-1330.

Broman, K.W., Wu, H., Sen, S. & Churchill, G.A. 2003, "R/qtl: QTL mapping in experimental crosses", *Bioinformatics,* vol. 19, no. 7, pp. 889-890.

Chen, X., Liu, C., Zhang, M. & Zhang, H. 2007, "A forest-based approach to identifying gene and gene–gene interactions", *Proceedings of the National Academy of Sciences,* vol. 104, no. 49, pp. 19199-19203.

Chen, X., Hackett, C.A., Niks, R.E., Hedley, P.E., Booth, C., Druka, A., Marcel, T.C., Vels, A., Bayer, M., Milne, I., Morris, J., Ramsay, L., Marshall, D., Cardle, L. & Waugh, R. 2010, "An eQTL Analysis of Partial Resistance to *Puccinia hordei* in Barley", *PLoS ONE,* vol. 5, no. 1, pp. e8598.

Chen, X., Zhu, J., Gerendas, J. & Zimmermann, N. 2008. Glucosinolates in Chinese Brassica campestris vegetables: Chinese cabbage, purple cai-tai, choysum, pakchoi, and turnip. HortScience 43: 571-574.

Chevenet, F., Brun, C., Banuls, A., Jacq, B. & Christen, R. 2006, "TreeDyn: towards dynamic graphics and annotations for analyses of trees", *BMC Bioinformatics,* vol. 7, no. 1, pp. 439.

Choi, S., Teakle, G., Plaha, P., Kim, J., Allender, C., Beynon, E., Piao, Z., Soengas, P., Han, T., King, G., Barker, G., Hand, P., Lydiate, D., Batley, J., Edwards, D., Koo,

D., Bang, J., Park, B. & Lim, Y. 2007, "The reference genetic linkage map for the multinational Brassica rapa genome sequencing project", *TAG Theoretical and Applied Genetics,* vol. 115, no. 6, pp. 777-792.

Churchill, G.A. & Doerge, R.W. 1994, "Empirical Threshold Values for Quantitative Trait Mapping", *Genetics,* vol. 138, no. 3, pp. 963-971.

Cockerham, C.C. 1969. Variance of gene frequencies. Evolution 23:72-84.

Cockerham, C.C. 1973. Analysis of gene frequencies. Genetics 74:679-700.

Cohen, J.H., Kristal, A.R. & Stanford, J.L. 2000, "Fruit and Vegetable Intakes and Prostate Cancer Risk", *JNCI Journal of the National Cancer Institute,* vol. 92, no. 1, pp. 61-68.

Coventry, J., Kott, L. & Beversdorf, W.D. 1988. Manual for Microspore Culture Technique for *Brassica napus*. Department of Crop Science Technical Bulletin OAC Publication 0489. University of Guelph

Custers, J.B.M., Cordewener, J.H.G., Fiers, M.A., Maassen, B.T.H., Lookeren Campagne, M.M.,van & Liu, C.M. 2001, "Androgenesis in Brassica: A model system to study the initiation of plant embryogenesis", .

Custers, J., Cordewener, J., Allen, Y., Dons, H. & Van, L.C. 1994, "Temperature controls both gametophytic and sporophytic development in microspore cultures of Brassica napus", *Plant Cell Reports,* vol. 13, no. 5, pp. 267-271.

D´hoop, B., Paulo, M., Mank, R., Eck, H.v. & Eeuwijk, F.v. 2008, "Association mapping of quality traits in potato (Solanum tuberosum L.)", *Euphytica,* vol. 161, no. 1, pp. 47-60.

De Vos, R.,C.H., Moco, S., Lommen, A., Keurentjes, J.J.B., Bino, R.J. & Hall, R.D. 2007, "Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry", *Nat.Protocols,* vol. 2, no. 4, pp. 778-791.

Denford, K.E. & Vaughan, J.G. 1977, "A Comparative Study of Certain Seed Isoenzymes in the Ten Chromosome Complex of Brassica Campestris and its Allies", *Annals of Botany,* vol. 41, no. 2, pp. 411-418.

Diaz-Uriarte, R. 2007, "GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest", *BMC Bioinformatics,* vol. 8, no. 1, pp. 328.

Diaz-Uriarte, R. & Alvarez, d.A. 2006, "Gene selection and classification of microarray data using random forest", *BMC Bioinformatics,* vol. 7, no. 1, pp. 3.

Dixon, G.R. 2006. Vegetable Brassicas and Related Crucifers. CABI.

Doerge, R.W. 2002, "Mapping and analysis of quantitative trait loci in experimental populations", *Nature reviews. Genetics,* vol. 3, no. 1, pp. 43-52.

Excoffier, L., Laval, G. & Schneider, S. 2005, "Arlequin (version 3.0): An integrated software package for population genetics data analysis", .

Fahey, J.W., Zalcmann, A.T. & Talalay, P. 2001, "The chemical diversity and distribution of glucosinolates and isothiocyanates among plants", *Phytochemistry,* vol. 56, no. 1, pp. 5-51.

Falush, D., Stephens, M. & Pritchard, J.K. 2007, "Inference of population structure using multilocus genotype data: dominant markers and null alleles", *Molecular Ecology Notes,* vol. 7, pp. 574-578(5).

Falush, D., Stephens, M. & Pritchard, J.K. 2003, "Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies", *Genetics,* vol. 164, no. 4, pp. 1567-1587.

Fenwick G.R, Heaney R.K. & Mullin W.J. 1983. Glucosinolates and their breakdown products in food and food plants. Critical Reviews in Food Science and Nutrition 18, 123-201.

Fernie, A.R. & Schauer, N. 2009, "Metabolomics-assisted breeding: a viable option for crop improvement?", *Trends in Genetics,* vol. 25, no. 1, pp. 39-48.

Fiehn, O. 2001, "Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks", *Comparative and Functional Genomics,* vol. 2, no. 3, pp. 155-168.

Field, B., Cardon, G., Traka, M., Botterman, J., Vancanneyt, G. & Mithen, R. 2004, "Glucosinolate and Amino Acid Biosynthesis in Arabidopsis", *Plant Physiology,* vol. 135, no. 2, pp. 828-839.

Flint-Garcia, S.A., Thuillet, A., Yu, J., Pressoir, G., Romero, S.M., Mitchell, S.E., Doebley, J., Kresovich, S., Goodman, M.M. & Buckler, E.S. 2005, "Maize association population: a high-resolution platform for quantitative trait locus dissection", *The Plant Journal,* vol. 44, no. 6, pp. 1054-1064.

Fu, J. & Jansen, R.C. 2006, "Optimal Design and Analysis of Genetic Studies on Gene Expression", *Genetics,* vol. 172, no. 3, pp. 1993-1999.

Fu, J., Swertz, M.A., Keurentjes, J.J.B. & Jansen, R.C. 2007, "MetaNetwork: a computational protocol for the genetic study of metabolic networks", *Nat.Protocols,* vol. 2, no. 3, pp. 685-694.

Gachon, C., Langlois-Meurinne, M., Henry, Y. & Saindrenan, P. 2005, "Transcriptional co-regulation of secondary metabolism enzymes in Arabidopsis : functional and evolutionary implications", *Plant Molecular Biology,* vol. 58, no. 2, pp. 229-245.

Gao, M., Li, G., Yang, B., McCombie, W.R. & Quiros, C.F. 2004, "Comparative analysis of a Brassica BAC clone containing several major aliphatic glucosinolate

genes with its corresponding Arabidopsis sequence", *Genome,* vol. 47, pp. 666-679(14).

Gigolashvili, T., Engqvist, M., Yatusevich, R., Müller, C. & Flügge, U. 2008, "HAG2/MYB76 and HAG3/MYB29 exert a specific and coordinated control on the regulation of aliphatic glucosinolate biosynthesis in *Arabidopsis thaliana*", *New Phytologist,* vol. 177, no. 3, pp. 627-642.

Gigolashvili, T., Yatusevich, R., Berger, B., Müller, C. & Flügge, U. 2007, "The R2R3-MYB transcription factor HAG1/MYB28 is a regulator of methionine-derived glucosinolate biosynthesis in *Arabidopsis thaliana*", *The Plant Journal,* vol. 51, no. 2, pp. 247-261.

Gislason, P.O., Benediktsson, J.A. & Sveinsson, J.R. 2006, "Random Forests for land cover classification", *Pattern Recognition Letters,* vol. 27, no. 4, pp. 294-300.

Gomez-Campo, C. 1999. Developments in plant genetics and breeding, vol 4. Biology of Brassica coenospecies. Elsevier

Goossens, A., Häkkinen, S.T., Laakso, I., Seppänen-Laakso, T., Biondi, S., De Sutter, V., Lammertyn, F., Nuutila, A.M., Söderlund, H., Zabeau, M., Inzé, D. & Oksman-Caldentey, K. 2003, "A functional genomics approach toward the understanding of secondary metabolism in plant cells", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 100, no. 14, pp. 8595-8600.

Goslee, S.C. & Urban, D.L. 2007, "The ecodist Package for Dissimilarity-based Analysis of Ecological Data", *Journal of Statistical Software,* vol. 22, no. 7, pp. 1-19.

Hardy, O.J. & Vekemans, X. 2002, "spagedi: a versatile computer program to analyse spatial genetic structure at the individual or population levels", *Molecular Ecology Notes,* vol. 2, no. 4, pp. 618-620.

Hartings, H., Berardo, N., Mazzinelli, G., Valoti, P., Verderio, A. & Motto, M. 2008, "Assessment of genetic diversity and relationships among maize (Zea mays L.) Italian landraces by morphological traits and AFLP profiling", *TAG Theoretical and Applied Genetics,* vol. 117, no. 6, pp. 831-842.

Hartl, D.L.C. & Andrew, G. 1997. Principles of population genetics. Sinauer Associate.

Hasan, M., Friedt, W., Pons-KÃ¼hnemann, J., Freitag, N., Link, K. & Snowdon, R. 2008, "Association of gene-linked SSR markers to seed glucosinolate content in oilseed rape (Brassica napus ssp. napus)", *TAG Theoretical and Applied Genetics,* vol. 116, no. 8, pp. 1035-1049.

He, Y., Michaels, S.D. & Amasino, R.M. 2003, "Regulation of Flowering Time by Histone Acetylation in Arabidopsis", *Science,* vol. 302, no. 5651, pp. 1751-1754.

Hintum, Th.J.L. van & I.W. Boukema. 1993. The establishment of the European Database for Brassica. FAO/IBPGR Plant Genetic Resources Newsletter no 94/95: 11-13.

Hirai, M.Y., Klein, M., Fujikawa, Y., Yano, M., Goodenowe, D.B., Yamazaki, Y., Kanaya, S., Nakamura, Y., Kitayama, M., Suzuki, H., Sakurai, N., Shibata, D., Tokuhisa, J., Reichelt, M., Gershenzon, J., Papenbrock, J. & Saito, K. 2005, "Elucidation of Gene-to-Gene and Metabolite-to-Gene Networks in Arabidopsis by Integration of Metabolomics and Transcriptomics", *Journal of Biological Chemistry,* vol. 280, no. 27, pp. 25590-25595.

Hirai, M.Y., Sugiyama, K., Sawada, Y., Tohge, T., Obayashi, T., Suzuki, A., Araki, R., Sakurai, N., Suzuki, H., Aoki, K., Goda, H., Nishizawa, O.I., Shibata, D. & Saito, K. 2007, "Omics-based identification of Arabidopsis Myb transcription factors regulating aliphatic glucosinolate biosynthesis", *Proceedings of the National Academy of Sciences,* vol. 104, no. 15, pp. 6478-6483.

Hu, J., Zhu, J. & Xu, H.M. 2000, "Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic values of crops", *TAG Theoretical and Applied Genetics,* vol. 101, no. 1, pp. 264-268.

Husson, F., Josse, J., Le, S. & Mazet, J. 2008. Factor Analysis and Data Mining with R. R package version 1.10.

Jansen R.C., Nap JP .2001. “Genetical genomics: The added value from segregation”. *Trends Genet* 17: 388–391

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,* vol. 10, pp. S65.

Keurentjes, J.J.B., Fu, J., de Vos, ,C.H.Ric, Lommen, A., Hall, R.D., Bino, R.J., van der Plas, L.,H.W., Jansen, R.C., Vreugdenhil, D. & Koornneef, M. 2006, "The genetics of plant metabolism", *Nature genetics,* vol. 38, no. 7, pp. 842-849.

Keurentjes, J.J.B., Fu, J., Terpstra, I.R., Garcia, J.M., van den Ackerveken, G., Snoek, L.B., Peeters, A.J.M., Vreugdenhil, D., Koornneef, M. & Jansen, R.C. 2007, "Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci", *Proceedings of the National Academy of Sciences,* vol. 104, no. 5, pp. 1708-1713.

Kim, J.S., Chung, T.Y., King, G.J., Jin, M., Yang, T., Jin, Y., Kim, H. & Park, B. 2006, "A Sequence-Tagged Linkage Map of Brassica rapa", *Genetics,* vol. 174, no. 1, pp. 29-39.

Kliebenstein, D.J., Kroymann, J., Brown, P., Figuth, A., Pedersen, D., Gershenzon, J. & Mitchell-Olds, T. 2001a, "Genetic Control of Natural Variation in Arabidopsis Glucosinolate Accumulation", *Plant Physiology,* vol. 126, no. 2, pp. 811-825.

Kliebenstein, D.J., Lambrix, V.M., Reichelt, M., Gershenzon, J. & Mitchell-Olds, T. 2001b, "Gene Duplication in the Diversification of Secondary Metabolism: Tandem 2-Oxoglutarate-Dependent Dioxygenases Control Glucosinolate Biosynthesis in Arabidopsis", *The Plant Cell,* vol. 13, no. 3, pp. 681-693.

Kliebenstein D, West M, van Leeuwen H, Loudet O, Doerge R, et al. 2006."Identification of QTLs controlling gene expression networks defined a priori". *BMC Bioinformatics* 7: 308

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,* vol. 17, no. 1, pp. 41-58.

Kroymann, J., Donnerhacke, S., Schnabelrauch, D. & Mitchell-Olds, T. 2003, "Evolutionary dynamics of an Arabidopsis insect resistance quantitative trait locus", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 100, no. Suppl 2, pp. 14587-14592.

Kroymann, J., Textor, S., Tokuhisa, J.G., Falk, K.L., Bartram, S., Gershenzon, J. & Mitchell-Olds, T. 2001, "A Gene Controlling Variation in Arabidopsis Glucosinolate Composition Is Part of the Methionine Chain Elongation Pathway", *Plant Physiology,* vol. 127, no. 3, pp. 1077-1088.

Kumar S, Dudley J, Nei M,Tamura K.2008. "MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences". *Briefings in Bioinformatics* 9: 299-306.

Lagercrantz, U. 1998, "Comparative Mapping Between Arabidopsis thaliana and Brassica nigra Indicates That Brassica Genomes Have Evolved Through Extensive Genome Replication Accompanied by Chromosome Fusions and Frequent Rearrangements", *Genetics,* vol. 150, no. 3, pp. 1217-1228.

Leoni, O., Iori, R., Palmieri, S., Esposito, E., Menegatti, E., Cortesi, R. & Nastruzzi, C. 1997, "Myrosinase-generated isothiocyanate from glucosinolates: Isolation, characterization and in vitro antiproliferative studies", *Bioorganic & medicinal chemistry,* vol. 5, no. 9, pp. 1799-1806.

Li CW. 1981, "Origin, Evolution, Taxonomy, and Hybridization of Chinese Cabbage". pp. 3-11.

Li, F., Kitashiba, H., Inaba, K. & Nishio, T. 2009, "A Brassica rapa Linkage Map of EST-based SNP Markers for Identification of Candidate Genes Controlling Flowering Time and Leaf Morphological Traits", *DNA Research,* vol. 16, no. 6, pp. 311-323.

Liang, Y, Kim, H.K., Lefeber, A.W.M., Erkelens, C., Choi, Y.H. & Verpoorte, R. 2006, "Identification of phenylpropanoids in methyl jasmonate treated Brassica rapa leaves using two-dimensional nuclear magnetic resonance spectroscopy", *Journal of Chromatography A,* vol. 1112, no. 1-2, pp. 148-155.

Liu, J., Liu, L., Hou, N., Zhang, A. & Liu, C. 2007, "Genetic diversity of wheat gene pool of recurrent selection assessed by microsatellite markers and morphological traits", *Euphytica,* vol. 155, no. 1, pp. 249-258.

Lou, P., Zhao, J., He, H., Hanhart, C.J., Pino del Carpio, D., Verkerk, R., Custers, J.B.M., Koornneef, M. & Bonnema, A.B. 2008, "Quantitative trait loci for glucosinolate accumulation in Brassica rapa leaves", .

Lunetta, K., Hayward, L.B., Segal, J. & Van Eerdewegh, P. 2004, "Screening large-scale association study data: exploiting interactions using random forests", *BMC Genetics,* vol. 5, no. 1, pp. 32.

Mackay, T.F.C. 2001, "The genetic architecture of quantitative traits", *Annual Review of Genetics,* vol. 35, no. 1, pp. 303-339.

Malosetti, M., van der Linden, C.G., Vosman, B. & van Eeuwijk, F.A. 2007, "A Mixed-Model Approach to Association Mapping Using Pedigree Information With an Illustration of Resistance to Phytophthora infestans in Potato", *Genetics,* vol. 175, no. 2, pp. 879-889.

Mantel, N. 1967, "The Detection of Disease Clustering and a Generalized Regression Approach", *Cancer research,* vol. 27, no. 2 Part 1, pp. 209-220.

Mithen, R.F., Dekker, M., Verkerk, R., Rabot, S. & Johnson, I.T. 2000, "The nutritional significance, biosynthesis and bioavailability of glucosinolates in human foods", *Journal of the science of food and agriculture,* vol. 80, no. 7, pp. 967-984.

Moco, S., Vervoort, J., Moco, S., Bino, R.J., Vos, R.C.H.D. & Bino, R. 2007, "Metabolomics technologies and metabolite identification", *TrAC Trends in Analytical Chemistry,* vol. 26, no. 9, pp. 855-866.

Mohammadi, S.A. & Prasanna, B.M. 2003, "Analysis of Genetic Diversity in Crop Plants--Salient Statistical Tools and Considerations", *Crop Science,* vol. 43, no. 4, pp. 1235-1248.

Moore, L.E., Brennan, P., Karami, S., Hung, R.J., Hsu, C., Boffetta, P., Toro, J., Zaridze, D., Janout, V., Bencko, V., Navratilova, M., Szeszenia-Dabrowska, N., Mates, D., Mukeria, A., Holcatova, I., Welch, R., Chanock, S., Rothman, N. & Chow, W. 2007, "Glutathione S-transferase polymorphisms, cruciferous vegetable intake and cancer risk in the Central and Eastern European Kidney Cancer Study", *Carcinogenesis,* vol. 28, no. 9, pp. 1960-1964.

Nei, M. 1978, “Estimation of average heterozygosity and genetic distance from a small number of individuals”, *Genetics,* vol. 89, no. 3, pp. 583-590.

de Nooy, W., A. Mrvar, V. Batagelj (*2005*). “Exploratory Social Network. Analysis with Pajek*”*. Cambridge : Cambridge University Press

Onyilagha, J., Bala, A., Hallett, R., Gruber, M., Soroka, J. & Westcott, N. 2003, "Leaf flavonoids of the cruciferous species, Camelina sativa, Crambe spp., Thlaspi arvense

and several other genera of the family Brassicaceae", *Biochemical systematics and ecology,* vol. 31, no. 11, pp. 1309-1322.

Padilla, G., Cartea, M.E., Velasco, P., de Haro, A. & Ordás, A. 2007, "Variation of glucosinolates in vegetable crops of Brassica rapa", *Phytochemistry,* vol. 68, no. 4, pp. 536-545.

Pang, H., Lin, A., Holford, M., Enerson, B.E., Lu, B., Lawton, M.P., Floyd, E. & Zhao, H. 2006, "Pathway analysis using random forests classification and regression", *Bioinformatics,* vol. 22, no. 16, pp. 2028-2036.

Parkin, I.A.P., Gulden, S.M., Sharpe, A.G., Lukens, L., Trick, M., Osborn, T.C. & Lydiate, D.J. 2005, "Segmental structure of the Brassica napus genome based on comparative analysis with Arabidopsis thaliana", *Genetics,* .

Patterson, N., Price, A.L. & Reich, D. 2006, "Population Structure and Eigenanalysis", *PLoS Genet,* vol. 2, no. 12, pp. e190.

Perrier, X. & Jacquemoud-Collet, J.P. 2006. DARwin software http://darwin.cirad.fr/darwin

Pflieger, S., Lefebvre, V. & Causse, M. 2001, "The candidate gene approach in plant genetics: a review", *Molecular Breeding,* vol. 7, no. 4, pp. 275-291.

Podsedek, A. 2007, "Natural antioxidants and antioxidant capacity of Brassica vegetables: A review", *LWT - Food Science and Technology,* vol. 40, no. 1, pp. 1-11.

Poelman, E.H., Dam, N.M., Loon, J.J.A., Vet, L.E.M. & Dicke, M. 2009, "Chemical diversity in Brassica oleracea affects biodiversity of insect herbivores", *Ecology,* vol. 90, no. 7, pp. 1863-1877.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. & Reich, D. 2006, "Principal components analysis corrects for stratification in genome-wide association studies", *Nature genetics,* vol. 38, no. 8, pp. 904-909.

Pritchard, J.K. & Przeworski, M. "Linkage Disequilibrium in Humans: Models and Data".

Pritchard, J.K., Stephens, M. & Donnelly, P. 2000, "Inference of Population Structure Using Multilocus Genotype Data", *Genetics,* vol. 155, no. 2, pp. 945-959.

Reeves, P.A. & Richards, C.M. 2009, "Accurate Inference of Subtle Population Structure (and Other Genetic Discontinuities) Using Principal Coordinates", *PLoS ONE,* vol. 4, no. 1, pp. e4269.

Remington, D.L., Thornsberry, J.M., Matsuoka, Y., Wilson, L.M., Whitt, S.R., Doebley, J., Kresovich, S., Goodman, M.M. & Buckler, E.S. 2001, "Structure of linkage disequilibrium and phenotypic associations in the maize genome", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 98, no. 20, pp. 11479-11484.

Ritland, K. 1996, "Estimators for pairwise relatedness and individual inbreeding coefficients", *Genetics Research,* vol. 67, no. 02, pp. 175-185.

Rochfort, S.J., Imsic, M., Jones, R., Trenerry, V.C. & Tomkins, B. 2006, "Characterization of Flavonol Conjugates in Immature Leaves of Pak Choi [Brassica rapa L. Ssp. chinensis L. (Hanelt.)] by HPLC-DAD and LC-MS/MS", *Journal of Agricultural and Food Chemistry,* vol. 54, no. 13, pp. 4855-4860.

Romani, A., Vignolini, P., Isolani, L., Ieri, F. & Heimler, D. 2006, "HPLC-DAD/MS Characterization of Flavonoids and Hydroxycinnamic Derivatives in Turnip Tops (Brassica rapa L. Subsp. sylvestris L.)", *Journal of Agricultural and Food Chemistry,* vol. 54, no. 4, pp. 1342-1346.

Rowe, H.C., Hansen, B.G., Halkier, B.A. & Kliebenstein, D.J. 2008, "Biochemical Networks and Epistasis Shape the Arabidopsis thaliana Metabolome", *The Plant Cell,* vol. 20, no. 5, pp. 1199-1216.

Sahr, T., Ravanel, S. & Rébeillé, ,F. *Tetrahydrofolate biosynthesis and distribution in higher plants*.

Saito, M., Kubo, N., Matsumoto, S., Suwabe, K., Tsukada, M. & Hirai, M. 2006, "Fine mapping of the clubroot resistance gene, Crr3, in Brassica rapa", *TAG Theoretical and Applied Genetics,* vol. 114, no. 1, pp. 81-91.

Schranz, M.E., Lysak, M.A. & Mitchell-Olds, T. 2006, "The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes", *Trends in plant science,* vol. 11, no. 11, pp. 535-542.

Sharpe, A.G., Parkin, I.A.P., Keith, D.J., & Lydiate, D.J. (1995). Frequent nonreciprocal translocations in the amphidiploid genome of oilseed rape (Brassica napus). Genome 38: 1112–1121

Simko, I. 2004, "One potato, two potato: haplotype association mapping in autotetraploids", *Trends in plant science,* vol. 9, no. 9, pp. 441-448.

Smýkal, P., Hýbl, M., Corander, J., Jarkovský, J., Flavell, A. & Griga, M. 2008, "Genetic diversity and population structure of pea (Pisum sativum L.) varieties derived from combined retrotransposon, microsatellite and morphological marker analysis", *TAG Theoretical and Applied Genetics,* vol. 117, no. 3, pp. 413-424.

Song, K., Osborn, T.C. & Williams, P.H. 1990, "Brassica taxonomy based on nuclear restriction fragment length polymorphisms (RFLPs)", *TAG Theoretical and Applied Genetics,* vol. 79, no. 4, pp. 497-506.

Song, K.M., Osborn, T.C. & Williams, P.H. 1988, "Brassica taxonomy based on nuclear restriction fragment length polymorphisms (RFLPs)", *TAG Theoretical and Applied Genetics,* vol. 75, no. 5, pp. 784-794.

Specht, C.E. & Diederichsen, A. 2001. Cruciferae In: Hanelt, P. (ed.) Mansfeld's Encyclopedia of Agricultural and Horticultural Crops. Springer-Verlag, Berlin, Germany Vol. 3: 1413-1481

Stich, B. 2009, "Comparison of Mating Designs for Establishing Nested Association Mapping Populations in Maize and Arabidopsis thaliana", *Genetics,* vol. 183, no. 4, pp. 1525-1534.

Storey, J.D. & Tibshirani, R. 2003, "Statistical significance for genomewide studies", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 100, no. 16, pp. 9440-9445.

Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P. & Feuston, B.P. 2003, "Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling", *Journal of chemical information and computer sciences,* vol. 43, no. 6, pp. 1947-1958.

Sybesma, W., Starrenburg, M., Tijsseling, L., Hoefnagel, M.H.N. & Hugenholtz, J. "Effects of Cultivation Conditions on Folate Production by Lactic Acid Bacteria", .

Takuno, S., Kawahara, T. & Ohnishi, O. 2007, "Phylogenetic relationships among cultivated types of Brassica rapa L. em. Metzg. as revealed by AFLP analysis", *Genetic Resources and Crop Evolution,* vol. 54, no. 2, pp. 279-285.

Talalay, P. & Fahey, J.W. 2001, "Phytochemicals from Cruciferous Plants Protect against Cancer by Modulating Carcinogen Metabolism", *Journal of Nutrition,* vol. 131, no. 11, pp. 3027S-3033.

Thornsberry, J.M., Goodman, M.M., Doebley, J., Kresovich, S., Nielsen, D. & Buckler, E.S., 2001, "Dwarf8 polymorphisms associate with variation in flowering time", *Nature genetics,* vol. 28, no. 3, pp. 286-289.

Tikunov, Y., Lommen, A., de Vos, C.H.R., Verhoeven, H.A., Bino, R.J., Hall, R.D. & Bovy, A.G. 2005, "A Novel Approach for Nontargeted Data Analysis for Metabolomics. Large-Scale Profiling of Tomato Fruit Volatiles", *Plant Physiology,* vol. 139, no. 3, pp. 1125-1137.

Traka, M. & Mithen, R. 2009, "Glucosinolates, isothiocyanates and human health", *Phytochemistry Reviews,* vol. 8, no. 1, pp. 269-282.

Trick, M., Cheung, F., Drou, N., Fraser, F., Lobenhofer, E., Hurban, P., Magusin, A., Town, C. & Bancroft, I. 2009, "A newly-developed community microarray resource for transcriptome profiling in Brassica species enables the confirmation of Brassica-specific expressed sequences", *BMC Plant Biology,* vol. 9, no. 1, pp. 50.

Vallejo, F., TomÂ·s-BarberÂ·n, F.A. & Ferreres, F. 2004, "Characterisation of flavonols in broccoli (Brassica oleracea L. var. italica) by liquid chromatography-UV diode-array detection-electrospray ionisation mass spectrometry", *Journal of Chromatography A,* vol. 1054, no. 1-2, pp. 181-193.

Verpoorte, R., Choi, Y., Mustafa, N. & Kim, H. 2008, "Metabolomics: back to basics", *Phytochemistry Reviews,* vol. 7, no. 3, pp. 525-537.

Vos P, Hogers R,Bleeker M, Rijan M, van der Lee T et al .1995. "AFLP: a new technique for DNA fingerprinting". *Nucleic Acid Res* 23: 4407-4414

Ward, J.L., Harris, C., Lewis, J. & Beale, M.H. 2003, "Assessment of 1H NMR spectroscopy and multivariate analysis as a technique for metabolite fingerprinting of Arabidopsis thaliana", *Phytochemistry,* vol. 62, no. 6, pp. 949-957.

Warwick, S.I., James, T. & Falk, K.C. 2008, "AFLP-based molecular characterization of Brassica rapa and diversity in Canadian spring turnip rape cultivars", *Plant Genetic Resources,* vol. 6, no. 01, pp. 11-21.

Weckwerth, W. & Morgenthal, K. 2005, "Metabolomics: from pattern recognition to biological interpretation", *Drug discovery today,* vol. 10, no. 22, pp. 1551-1558.

Weir, B.S. & Cockerham, C.C. 1984. Estimating F-statistics for the analysis of population structure. Evolutionary Bioinformatics Online 38: 1358-1370.

Wentzell, A.M., Rowe, H.C., Hansen, B.G., Ticconi, C., Halkier, B.A. & Kliebenstein, D.J. 2007, "Linking Metabolic QTLs with Network and *cis*-eQTLs Controlling Biosynthetic Pathways", *PLoS Genet,* vol. 3, no. 9, pp. e162.

West, M.A.L., Kim, K., Kliebenstein, D.J., van Leeuwen, H., Michelmore, R.W., Doerge, R.W. & St. Clair, D.A. 2007, "Global eQTL Mapping Reveals the Complex Genetic Architecture of Transcript-Level Variation in Arabidopsis", *Genetics,* vol. 175, no. 3, pp. 1441-1450.

Widarto, H., Van, D.M., Lefeber, A., Erkelens, C., Kim, H., Choi, Y. & Verpoorte, R. 2006, "Metabolomic Differentiation of Brassica rapa Following Herbivory by Different Insect Instars using Two-Dimensional Nuclear Magnetic Resonance Spectroscopy", *Journal of chemical ecology,* vol. 32, no. 11, pp. 2417-2428.

Wittstock, U. & Halkier, B.A. 2002, "Glucosinolate research in the Arabidopsis era", *Trends in plant science,* vol. 7, no. 6, pp. 263-270.

Wright, S.I. & Gaut, B.S. 2005, "Molecular Population Genetics and the Search for Adaptive Evolution in Plants", *Molecular biology and evolution,* vol. 22, no. 3, pp. 506-519.

Wright, S. 1951. The genetical structure of populations. Ann Eugen 15:323-354.

Wu, G., Shi, Q., Niu, Y., Xing, M. & Xue, H. 2008, "Shanghai RAPESEED Database: a resource for functional genomics studies of seed development and fatty acid metabolism of Brassica", *Nucleic acids research,* vol. 36, no. suppl_1, pp. D1044-1047.

Yang, T., Kim, J.S., Kwon, S., Lim, K., Choi, B., Kim, J., Jin, M., Park, J.Y., Lim, M., Kim, H., Lim, Y.P., Kang, J.J., Hong, J., Kim, C., Bhak, J., Bancroft, I. & Park,

B. 2006, "Sequence-Level Analysis of the Diploidization Process in the Triplicated FLOWERING LOCUS C Region of Brassica rapa", *The Plant Cell,* vol. 18, no. 6, pp. 1339-1347.

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(Suppl 1):S135. doi: 10.1186/1471-2156-6-S1-S135

Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., Kresovich, S. & Buckler, E.S. 2006, "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness", *Nature genetics,* vol. 38, no. 2, pp. 203-208.

Zang, Y., Kim, H.U., Kim, J.A., Lim, M., Jin, M., Lee, S.C., Kwon, S., Lee, S., Hong, J.K., Park, T., Mun, J., Seol, Y., Hong, S. & Park, B. 2009, "Genome-wide identification of glucosinolate synthesis genes in Brassicarapa", *FEBS Journal,* vol. 276, pp. 3559-3574(16).

Zhang, X., Blair, M. & Wang, S. 2008, "Genetic diversity of Chinese common bean (Phaseolus vulgaris L.) landraces assessed with simple sequence repeat markers", *TAG Theoretical and Applied Genetics,* vol. 117, no. 4, pp. 629-640.

Zhao, J., Paulo, M., Jamar, D., Lou, P., van Eeuwijk, F., Bonnema, G., Vreugdenhil, D. & Koornneef, M. 1 October 2007, "Association mapping of leaf traits, flowering time, and phytate content in Brassica rapa", *Genome,* vol. 50, pp. 963-973(11).

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", *TAG Theoretical and Applied Genetics,* vol. 110, no. 7, pp. 1301-1314.

Zhu, C., Gore, M., Buckler, E.S. & Yu, J. 2008, "Status and Prospects of Association Mapping in Plants", *The Plant Genome,* vol. 1, no. 1, pp. 5-20.

Zickute J., Strumylaite L., Dregyal L., Petrauskiene J., Dudzevicius J. & Strartilovas E. 2005. Vegetables and fruits and risk of stomach cancer. Medicina (Kaunas, Lithiuania) 41: 733-740

Summary

In this thesis the metabolic variation in *Brassica rapa* is described based on results of metabolic profiling of a core collection of 168 accessions representing the different crop types and geographical origin and a Doubled Haploid population. In Chapter 2 we describe the genetic and phenotypic variation of this core collection to explore the possibility of following association mapping methods to identify genes involved in metabolic regulation. We explored through a genome wide and candidate gene approach different association mapping methods in a core collection in Chapters 3 and 4 respectively and in Chapter 5 we combined the QTL analysis of targeted and untargeted metabolites profiled through LC-MS with expression QTLs following a genetical genomics approach aiming to detect genes underlying the metabolite QTL.

The genetic diversity evaluated through the screening of AFLP and SSR markers was correlated with classification of accessions using morphological and metabolic trait values. The relationship between accessions in groups was compared using hierarchical clustering and the STRUCTURE program. Using Random Forests classification a set of metabolites was selected that differentiated the different sub groups as determined by STRUCTURE (Chapter 2). Based on the classification into subpopulations using the STRUCTURE program we included the subpopulations as a correction term in our statistical model for association studies (Chapter 3). Additionally, because of the increasing amount of data that will be soon available through sequencing technology we tested the use of Random Forests in the search for marker-trait association for the isoprenoids pathway. Using the results obtained with the linear models as implemented in TASSEL and the results obtained in Random Forests we found a set of 16 significant markers with potential use for marker assisted selection in breeding for several isoprenoids

The determination of map positions through synteny prediction and genetic mapping of a group of genes from the glucosinolate pathway lead us to identify *Myb28* and *MAM* as candidate genes mapping under a previously detected major QTL for glucosinolates We followed an association mapping approach to investigate their role in the variation in glucosinolates in the core collection by profiling 37 SSR markers, which included markers linked to these candidate genes and markers distributed along different positions in linkage group A03 (Chapter 4). Interestingly, not only *MAM* and *Myb28,* but the *AOP* and *GS-OH* genes involved in side chain modification and

Myb29 in transcriptional regulation were also associated with glucosinolate levels. A genetical genomics approach was followed to identify candidate genes for variation in metabolites of six biosynthetic pathways: carotenoids, tocopherols, folates, glucosinolates, flavonoids and phenylpropanoids, based on the co-localization analysis and comparison between metabolic (m)QTLs and expression (e)QTLs (Chapter 5). A Doubled Haploid (DH) population was profiled for metabolite content and variation through targeted and LC-MS untargeted approaches. Additionally, the same population was profiled for transcript variation with a newly developed 105K Cogenics array assembled using mainly EST sequences from three species: *B. napus, B. rapa* and *B. oleracea*. Co-localization of eQTLs and mQTLs for several isoprenoids (tocopherols and carotenoids) and glucosinolates lead us to the identification of candidate genes for these pathways. However, further work is needed to identify the gene or genes underlying a major cluster of QTLs for 112 centrotypes derived from the LC-MS untargeted data. The results obtained through this combined approach and considerations that need to be taken into account when performing these types of studies with regard to identification of paralogues and the use of a multi Brassica species microarray for transcript profiling in *Brassica rapa* are discussed.

In the final Chapter, the combined use of core collections encompassing the genetic diversity within *B. rapa* and biparental DH populations to unravel the genetic regulation of the metabolome are discussed.

Samenvatting

In dit proefschrift is de variatie en genetische regulatie van metabolietsamenstelling in bladeren van *Brassica rapa* bestudeerd en zijn een aantal genetische loci op het genoom en kandidaatgenen voor die regulatie geïdentificeerd. Voor dit onderzoek is zowel een *B. rapa* core-collectie samengesteld uit materiaal van genenbanken en veredelingsbedrijven, die de verschillende gewastypes en de verschillende geografische herkomsten representeert, als een populatie van verdubbelde haploïden gecreëerd. In hoofdstuk 1 wordt het genus *Brassica* en de soort *B. rapa* beschreven, met nadruk op de enorme morfologische variatie die tot uiting komt in de verschillende gewassen (bladgroentes als Chinese kool en Paksoi, knolgewassen als meiraapjes en voederraap, oliegewassen en groentes waarvan de bloeiwijze geconsumeerd worden zoals broccoletto en taicai). In hoofdstuk 2 is de genetische variatie, gebaseerd op moleculaire merkers, en de fenotypische variatie (zowel morfologische eigenschappen als metabolietsamenstelling) van de 168 accessies beschreven, met als doel te kijken in hoeverre deze verschillende datasets leiden tot vergelijkbare classificaties van de core-collectie en in hoeverre genetische kartering via associatie-genetica een optie is voor *B. rapa*. In de hoofdstukken 3 en 4 zijn verschillende statistische modellen getoetst om de associaties tussen genetische merkers en individuele metabolieten op te sporen, door genoombreed allelische variatie in merkers over de accessies te bepalen, of door juist in te zoomen op merkers die fysisch gekoppeld zijn aan genen die betrokken zijn bij de biosynthese van specifieke metabolieten. In hoofdstuk 5 is de genetische regulatie van metabolietsamenstelling onderzocht door de kwantitatieve variatie in metabolieten en genexpressie te meten in de populatie verdubbelde haploïden: deze aanpak word `genetical genomics`genoemd, en heeft als doel genen die de variatie onder een QTL (genetisch locus dat een deel van de genetische variatie voor een bepaalde kwantitatieve eigenschap verklaart) veroorzaken, te identificeren.

De genetische diversiteit in de core-collectie, geëvalueerd door de accessies te screenen met AFLP- (Amplified Fragment Length Polymorphism) en microsatelliet- (SSR) merkers, is gecorreleerd aan de classificatie gebaseerd op morfologische en metaboliet variatie. De genetische diversiteit wordt berekend via hiërarchische clustering en via het programma STRUCTURE, en de resulterende subgroepen worden met elkaar vergeleken. Met de statistische classificatie-methode RANDOM FORESTS is een aantal metabolieten geselecteerd die op basis van kwantitatieve variatie de verschillende groepen, zoals bepaald in STRUCTURE, kunnen onderscheiden (hoofdstuk 2).

In hoofdstuk 3 wordt de allelische variatie in AFLP- en SSR-merkers gerelateerd aan variatie in glucosinolaten, tocopherolen, carotenoïden en folaat met behulp van verschillende statistische methoden, waarbij de groepen zoals gedefiniëerd in STRUCTURE worden ingevoerd als correctiefactoren. De geschiktheid van een geheel nieuwe statistische methode werd getoetst, namelijk RANDOM FORESTS, die vooral geschikt is voor grote datasets, zoals gegenereerd in metabolomics- en transcriptomicsonderzoek, om associaties van genetische merkers met variatie in metabolieten op te sporen. De gecombineerde resultaten zoals verkregen met de lineaire methodes, gecorrigeerd voor subpopulaties, en RANDOM FORESTS, resulteerde in een set van 16 merkers die significant geassociëerd zijn met gehaltes van specifieke isoprenoïden, en die gebruikt kunnen worden in de merkergestuurde veredeling voor optimale metabolietsamenstelling.

De genetische positie van een groot aantal genen uit de glucosinolaat-biosyntheseroute is bepaald via genetische kartering en voorspeld op basis van Brassica-Arabidopsis syntenie. Samenvallen van de genetische posities van deze genen en QTLs voor glucosinolaat-samenstelling in het blad van *B. rapa*, leidde tot de hypothese dat twee genen, *Myb28* en *MAM*, mogelijk de variatie in glucosinolaat-samenstelling verklaren. Dit werd nader onderzocht in een associatie-studie waarbij de associatie tussen glucosinolaat-samenstelling in het blad van de 168 verschillende accessies en de allelische variatie in 37 SSR merkers gelegen op koppelingsgroep A03, waaronder merkers gekoppeld aan glucosinolaat-biosynthese en transcriptie regulerende genen, werd berekend (hoofdstuk 4). Dit toonde aan dat niet alleen allelische variatie voor *MAM* en *Myb-28*, maar ook voor genen betrokken bij modificatie van de zijketens, *AOP* en *GSL-OH*, en de transcriptiefactor *Myb-29*, geassociëerd waren met variatie in glucosinolaat-gehaltes in de core-collectie.

In hoofdstuk 5 werd de zogenaamde genetical genomics-aanpak gevolgd met als doel kandidaatgenen te identificeren die de variatie in metabolieten uit zes biochemische routes bepalen. Het betrof de carotenoïden, folaat, tocopherolen, glucosinolaten, flavonoïden en phenylpropenoiden. Hiervoor werd de metaboliet-samenstelling bepaald in blad van zes weken oude planten van de verdubbelde haploïden-populatie door middel van LC-QTOF-MS en gerichte analyse van specifieke metabolieten. Van hetzelfde materiaal werd RNA geïsoleerd om variatie in genexpressie te meten met de Cogenics array, die105.000 probes telt, gebaseerd op EST (Expressed Sequence Tags) sequenties van *B. rapa, B. oleracea* en vooral *B. napus*. Co-lokalisatie van QTLs voor metabolieten (mQTL) en een subset van de transcripten (namelijk de genen waarvan de rol in de biosynthese van de geselecteerde metabolieten bekend is) (eQTL) leidde tot een aantal kandidaatgenen voor de regulatie van

een aantal glucosinolaten, tocopherolen en carotenoïden. Voor een cluster van QTLs voor 112 centrotypes geïdentificeerd met LC-QTLOF-MS op koppelingsgroep A07 werden nog geen kandidaat-genen geïdentificeerd. Verdere analyse is nodig als er ersteen annotatie van alle probes op de array beschikbaar is. In dit hoofdstuk worden de resultaten besproken, met focus op de rol van paralogen en de (on)mogelijkheden om met de Cogenics array expressie verschillen van de verschillende paralogen te bepalen.

In het afsluitende hoofdstuk worden de mogelijkheden besproken die een gecombineerde aanpak van QTL-kartering en genetical genomics in segregerende populaties (verkregen via het kruisen van twee ouders) en van associatiestudies in grote core collecties biedt om de genetica van het metaboloom te ontrafelen. Hierbij wordt ingegaan op de genetische resolutie van beide methodes en de beschikbare allelische variatie.

Curriculum Vitae

Dunia Pino Del Carpio was born on the 2nd September 1977 in Lima, Peru. She studied at Universidad Mayor de San Marcos (Peru) and obtained her BSc degree in Biology with a major in Microbiology and Parasitology in 2001. After her studies she worked as a visiting scientist at Ohio State University (USA). She started her Master program at Wageningen University in 2004 and got her degree in Plant Sciences with specialization in Plant Breeding and Genetic resources. In 2006 she started her doctoral studies at Wageningen University funded by an IOP genomics project. She will defend her PhD thesis on October 4th 2010 to obtain her PhD degree. After her doctoral studies she will begin her work as a postdoc at Heinrich Heine University in Duesseldorf, Germany.

Publications

Association mapping in *Brassica rapa*: a case study on metabolite variation. **Dunia Pino Del Carpio**, Ram Kumar Basnet, Ric De Vos, Chris Maliepaard, Joao Paulo, Guusje Bonnema. Paper in preparation.

The Patterns of population differentiation in a *Brassica rapa* Core Collection. **Dunia Pino Del Carpio**, Ram Kumar Basnet, Ric De Vos, Chris Maliepaard, Guusje Bonnema. Paper under review 2010

BrFLC2 (FLOWERING LOCUS C) as a candidate gene for a vernalization response QTL in *Brassica rapa* . Vani Kulkarni, Jianjun Zhao, Nini Liu, **Dunia Pino Del Carpio**, Johan Bucher and Guusje Bonnema. Journal of experimental Botany, 2010.

Identification of seed related QTLs in a new *Brassica rapa* population: Plants with more siliques have more seeds and higher seed oil content but smaller seeds. Hedayat Bagheri, **Dunia Pino Del Carpio**, Guusje Bonnema, Maarten Koornneef and Mark G.M. Aarts. Paper in preparation.

Brassica vegetable book.Chapter 3.Diversity analysis and molecular taxonomy in Brassicas. Book in editing process, editor Jan Sadowsky. **Dunia Pino Del Carpio**, Jianjun Zhao and Guusje Bonnema.

Quantitative trait loci for glucosinolate accumulation in *Brassica rapa* leaves. Ping Lou, Jianjun Zhao, Hongju He, Corrie Hanhart, **Dunia Pino Del Carpio**, Ruud Verkerk, Jan Custers, Maarten Koornneef and Guusje Bonnema. New Phytologist.2008

Quantitative trait loci for flowering time and morphological traits in multiple populations of *Brassica rapa*. Ping Lou, Jianjun Zhao, Jung Sun Kim, Shuxing Shen, **Dunia Pino Del Carpio**, Xiaofei Song, Mina Jin, Dick Vreugdenhil, Xiaowu Wang, Maarten Koornneef, Guusje Bonnema . J Exp Bot. 2007

Acknowledgements

To finish my doctoral studies has been a great achievement in my life and it has only been possible because of the collaboration with multiple people.

I would like to express my gratitude to all who have in one way or another encouraged me or supervised me through this PhD project.

First of all I would like to acknowledge the support of Prof. Richard Visser, thank you for your excellent guidance specially during the manuscript preparation, I am very grateful for the advice and suggestions to improve the thesis.

When I first arrived to Wageningen back in 2004 I began to do research in a vegetable crop which was unknown to me at that moment: *Brassica rapa*. Dr. Guusje Bonnema allowed me to learn through her wide experience about this crop during my master studies and later on she trusted me with a PhD project. I truly believe that she was the best supervisor I could have had to follow my doctoral studies. Guusje, I have learned a lot from you, from our always very stimulating scientific conversations and also from your personal kindness.

I am also very thankful to Prof. Evert Jacobsen, I always enjoyed our conversations not always about science but also about my personal ambitions. Evert, you always understood this side of me and encouraged me to continue my Phd studies with a positive view.

Without the scientific support of many people it would not have been possible to complete this thesis. For the data analysis I would like to express my gratitude for the collaboration with Dr.Chris Maliepaard which was very important to gain more insight in the statistical methods. For the statistical analysis I would also like to thank Frank Johannes and Danny Arends from Groningen University.

From the Brassica group I would like to thank several people. Johan Bucher, Ram Kumar Basnet and Jianjun Zhao. Johan, you have been more than the technician in this project, I am very glad I can consider you my friend, you always made more enjoyable those times when we had to do some tedious lab work. I will always remember our lunch breaks.

Ram, through you I learned so much, your curiosity was always stimulating and I will miss very much our conversations about statistics. Jianjun, I have known you the longest from the Brassica group, your positive attitude and your wide knowledge about Brassica helped me a lot during my Phd study.

During this time in Wageningen I was lucky to meet many people with whom I enjoyed my free time. I would like to thank them for showing me there was also some life outside the Lab. Many thanks to: Alejandra, Miluska, Golda, Andres and Delphine for the good times that will always stay in my memory.

Finally, I would like to thank my family. No tengo palabras para agradecerles el apoyo que he tenido de uds todo este tiempo, se que muchas veces me he distanciado por el trabajo, pero nunca me he sentido sola y siempre han estado en mi corazon, espero que siempre se sientan orgullosas de mi. Now I will begin a new life in Germany and I just want to say gracias por todo tu apoyo y por toda la felicidad. Ich liebe dich.

Dunia Pino Del Carpio
Wageningen, The Netherlands
October, 2010

Education Statement of the Graduate School

Experimental Plant Sciences

The Graduate School EXPERIMENTAL PLANT SCIENCES

Issued to:	**Dunia Pino Del Carpio**
Date:	**4 October 2010**
Group:	**Laboratory of Plant Breeding, Wageningen University**

1) Start-up phase		*date*
►	**First presentation of your project**	
	Brassica Vegetable nutrigenomics	Aug 28, 2006
►	**Writing or rewriting a project proposal**	
►	**Writing a review or book chapter**	
	Vegetable Brassicas, Chapter 3 diversity analysis and molecular taxonomy	2009
►	**MSc courses**	
►	**Laboratory use of isotopes**	
	Subtotal Start-up Phase	*5.5 credits**

2) Scientific Exposure		*date*
►	**EPS PhD Student Days**	
	EPS PhD Student day, Wageningen University	Sep 13, 2007
►	**EPS Theme Symposia**	
	EPS Theme 3 symposium 'Metabolism and Adaptation', Wageningen University	Nov 06, 2007
	EPS Theme 3 symposium 'Metabolism and Adaptation', Leiden University	Feb 16, 2010
►	**NWO Lunteren days and other National Platforms**	
	ALW meeting 'Experimental Plant Sciences', Lunteren	Apr 02-03, 2007
	ALW meeting 'Experimental Plant Sciences', Lunteren	Apr 07-08, 2008
	ALW meeting 'Experimental Plant Sciences', Lunteren	Apr 06-07, 2009
►	**Seminars (series), workshops and symposia**	
	Symposium on quality from soil to healthy people,wageningen	Nov 16, 2006
►	**Seminar plus**	
►	**International symposia and congresses**	
	Crucifer genetics workshop, Wageningen The Netherlands	Sep 30-Oct 04, 2006
	Molecular Mapping and Marker assisted selection in plants, Vienna, Austria	Feb 03-06, 2008
	Crucifer genetics workshop, Lillehammer,Norway	Sep 08-12, 2008
	Genomics assisted conference, Korea	Dec 16, 2008
	Glucosinolates conference, Denmark	May 24-27, 2009
	OECD Association mapping conference, Perth Australia	Nov 09-12, 2009
►	**Presentations**	
	oral: EPS theme 3 symposium 'Metabolism and Adaptation'	Nov 06, 2007
	poster: Molecular Mapping and Marker assisted selection in plants, Vienna, Austria	Feb 03-06, 2008
	oral: Research day 2008, Wageningen, The Netherlands	Jun 17, 2008
	oral: Crucifer genetics workshop, Lillehammer,Norway	Sep 08-12, 2008
	oral: Genomics assisted conference, Korea	Dec 16, 2008
	oral: Glucosinolates conference, Denmark	May 24-27, 2009
	poster: OECD Association mapping conference, Perth Australia	Nov 09-12, 2009
►	**IAB interview**	Dec 05, 2008
►	**Excursions**	
	Subtotal Scientific Exposure	*15.5 credits**

3) In-Depth Studies		*date*
►	**EPS courses or other PhD courses**	
	ETNA summerschool 'Metabolite profiling and data analysis", Potsdam (Germany)	Sep 20-29, 2006
	Statistics for Omics data, Wageningen,The Netherlands	Dec 11-14, 2006
	Basic statistics	Jun-Jul 2009
	Introduction to R	Oct 08-09, 2009
►	**Journal club**	
	Member of the literature discussion group at the Plant Breeding Group	2006-2010
►	**Individual research training**	
	Subtotal In-Depth Studies	*8.7 credits**

4) Personal development		*date*
►	**Skill training courses**	
	Scientific writing	Sep-Nov 2008
►	**Organisation of PhD students day, course or conference**	
	Crucifer genetics workshop 2006	Sep 30-Oct 04, 2006
►	**Membership of Board, Committee or PhD council**	
	Subtotal Personal Development	*3.3 credits**

TOTAL NUMBER OF CREDIT POINTS*	**33**

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

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