Multivariate statistical analyses, association study and network analyses on genetic and phenotypic data from a *Brassica rapa* core collection

By:

Ram Kumar Basnet

Major thesis for Master of Science in Plant Sciences
(Specialization Plant Breeding and Genetic Resources)

Supervisors:
Dr. Chris Maliepaard
Dr. Guusje Bonnema
M. Sc. Dunia Pino Del Carpio

Examiners:
Dr. Chris Maliepaard
Dr. Guusje Bonnema

Department of Plant Breeding
Wageningen University
April, 2009

Multivariate statistical analyses, association study and network analyses on genetic and phenotypic data from a *Brassica rapa* core collection

By:

Ram Kumar Basnet
(Regd. No. 801009037010)

Major thesis for the partial fulfillment of ***Master of Sciences in Plant Sciences***
Specialization ***Plant Breeding and Genetic Resources***

Supervisors:
Dr. Chris Maliepaard
Dr. Guusje Bonnema
M. Sc. Dunia Pino Del Carpio

Examiners:
Dr. Chris Maliepaard
Dr. Guusje Bonnema

Department of Plant Breeding
Wageningen University
April, 2009

Acknowledgement

It is my pleasure to acknowledge for all people who co-operate me for the completion this thesis work. Firstly, I express my gratitude to God for giving me motivation, strength and courage to fulfill my aims. I would like to thank **Dr. Chris Maliepaard** for providing opportunities to work under his supervision and grateful to his sincere care and guidance. Dear Sir, I am able to gain lots of statistical knowledge that will be crucial for my scientific career development. I like your abilities to understand my problems and to make clarity for my confusion during this thesis as well as theory courses. I never felt helpless through out my work; even when I had confusion. I was more encouraged and got exposure by getting opportunity to take part in a course "System Biology: statistical analysis for ~omics data". I thank you so much.

I would also thankful to **Dr. Guusje Bonnema** for her co-operation for the completion of my thesis. During this thesis, her encouragement and providing biological insight on this study had always boosted up for the completion of my thesis even it was late in final submission. I salute the opportunity that you provide me for further specialization, and believe that it will be milestone in developing my scientific career. Besides, advices during course composition as study advisor as well as take care in thesis, I always feel you as a guardian.

I also like to thank **Dunia Pino Del Carpio** for her close supervision and inspiration towards works. Dunia, you really gave me feelings on how to co-relate my results with biological relevance. I impressed with your ability to integrate statistical analysis with biological interpretation. You were so determine and willful in helping me that I was able to manage my huge datasets.

I can not complete my thanks-giving words without acknowledging **Animesh Acharjee**. I express my heartfelt thanks for his very good support for statistical understandings. Dear brother, I can imagine it will be very difficult to learn R-software without your support. The frequent discussions help me a lot to get insight in statistical techniques.

I also acknowledge for NUFFIC for offering fellowship for my Master degree.

Also, I express my thanks to my office; Agriculture Botany Division, Nepal Agricultural Research Council (NARC) for facilitating all the necessary steps in making my Master study abroad.

Finally, my deepest thanks go to all individuals for their direct and indirect support for the completion of my thesis.

Table of contents

Abstract

The *Brassica rapa* is an important vegetable crop grown through out the world. This study aimed to explore the morphological, molecular and metabolic relations among the diverse morphotypes in a core collection of 168 accessions. These accessions were genotyped with AFLP and MYB markers, and morphological observations were recorded on 26 traits under vernalized treatment. ^{1}-H NMR was conducted for untargeted metabolite profiling while LC-MS-PDA-QTOF was used for both untargeted metabolite profiling and targeted analysis of carotenoid and tocopherol pathway. The unsupervised classification methods, such as hierarchical clustering and PCA/PCoA on morphology, marker and LC-MS datasets grouped the accessions according to geographic origins and lesser extent to morphotypes. Two sub-groups of oil types were discovered based on genetic and LC-MS datasets. The highly significant Mantel correlations were found between dissimilarity matrices in all combinations of datasets except with NMR. RF classification, a supervised approach, was also used to identify the distinguishing markers, NMR bins and LC-MS peaks in pair-wise classifications of four STRUCTURE classes as well as morphotypes. In addition, association studies were conducted to identify molecular markers correlated SPAD value of leaf color under both vernalized and non-vernalized conditions by using RF regression and unified mixed model approach, and also LC-MS peaks using RF regression approach. The correction was done only for population structure in RF regression, and for both population structure and kinship relations in unified mixed model approach, but obtained similar results. The associated LC-MS peaks and markers were varied according to vernalization treatment confounded with their growing conditions. Finally, the correlation network analyses were conducted, separately, for targeted metabolites of caretenoid and tocopherol pathway, and RF regression selected LC-MS peaks for SPAD trait under vernalized conditions. The simple and partial correlation analyses were used to explore the direct and indirect correlation of metabolites. The associated markers were integrated in partial correlation network to reveal the genetic regulation of metabolites under the pathway. The annotation of metabolic peaks and identification of map position of markers is essential to know their biological relevance. Thus, these kinds of study can be used for to explore the relations of morphotypes in a population and their distinguishing variables.

Key words: Core collection, multivariate statistical tools, random forest, variable selection, population structure, association study, networking

Chapter 1: Introduction

The genus *Brassica* comprises a large numbers of genetically diverse and economically important species. *Brassica rapa*, the first domesticated *Brassica* species, has a long and independent history of cultivation in Europe and Asia (Zhao *et al.,* 2005) and encompasses some key morphotypes: vegetable or fodder turnip (ssp. *rapa,* formerly subsp. *rapifera*, used for edible swollen roots), oil-seed turnip rape or Chinese turnip rape (*B. rapa* ssp. *oleifera,* used for seed-extracted oil), Chinese cabbage (ssp. *pekinensis,* an Asian heading vegetable with tightly overlapping pale green leaves), Pak choi (ssp. *chinensis,* a Chinese non-heading dark green and thick leaves with broad, thick, white petioles), Komatsuna including mizuna and mibuna (ssp. *nipposinica,* many thin narrow serrated or non-serrated leaves (Takuno *et al.*, 2007), broccoletto (ssp. *ruvo,* a European vegetable with an enlarged and compact inflorescence), Wutacai (sp. *narinosa*), sarson, and Caixin or Caitai (Paterson *et al.,* 2001; Warwick *et al.,* 2008; Zhao *et al.*, 2005). The oil seed type ssp. *oleifera* is further divided into sub-groups based on their growth habit: winter and spring types (Zhao *et al.*, 2005), and sarson into three types: brown sarson (ssp. *dichotoma*), yellow sarson (ssp. *trilocularis*) and toria as a result of history-long breeding efforts in India (Duhoon and Koppar, 1998).

A core collection of *B. rapa,* used in this study, consists of 168 accessions representing the different morphotypes. Based on allele frequencies of AFLP markers, Zhao *et al.,* (2007a) identified four sub-groups: a turnip group, a Pak choi group, a spring oil group and a Chinese cabbage group this core collection.

B. rapa is the most commonly consumed vegetable and oil seed crop through out the world because of its high nutritional importance. *Brassica* vegetables are a rich source of secondary metabolites and phytonutrients, including different types of glucosinolates, carotenoids, flavonoids, vitamin C and folic acid, which have health promoting roles (IOP: *Brassica* vegetable nutrigenomics, 2006-2010). Isoprenoids such as tocopherol and carotenoids are well-known antioxidants; carotenoids are the main precursors of vitamin A synthesis (Zhou *et al.*, 2008). Despite some researches on the nutritional importance and availability of isoprenoid compounds, the study on genetic regulation of the carotenoids and tocopherol pathway in *B. rapa* is still insufficient. The unraveling of the biosynthetic regulation of carotenoids and tocopherol, and finding genetic markers in *Brassica* associated with the presence and abundance of health-promoting compounds is essential to improve the nutritional quality of commercial cultivars.

Breakthroughs in high-throughput technologies and recent advances on statistical tools especially in ~omics have exponentially increased the study of metabolomics. Metabolic characterization is essential for quality improvement of plants, for instances, nutrient content, colour, flavour, defense mechanisms. In an untargeted approach (also called metabolic fingerprinting), a global screening of as many metabolites as possible is performed to generate quantitative measurements of compounds with largely unknown chemical structures, whereas a set of metabolites in a selected biochemical pathway or a specific class of compounds is profiled in a targeted approach, called metabolic profiling (Allwood *et al.,* 2008; Dettmer *et al.,* 2007).

Metabolite levels fluctuate according to developmental, environmental, physiological and pathological conditions. The identifications of species-specific secondary metabolites and differences in their levels between species are essential to know the species-specific metabolites and to understand the interaction of the cell with its environment (Verpoorte *et al.,* 2007).

B. rapa has a wide natural variation in phytonutrient composition and a large range of concentrations, and the study of the biochemical diversity can play a complementary role in metabolomics research. 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 (MS) provides useful information on the mass and is required to identify the molecular formula of the detected metabolites (Moco, 2007). It is known that no single analytical technique is sufficient to extract and detect all the metabolites (De Vos *et al.,* 2007), mostly due to ionization techniques, chromatography and detector capabilities (Weckwerth and Morgenthal, 2005). Therefore, it is preferable to use a wide spectrum of chemical analysis techniques for rapid, reproducible and stable analyses which cover more different types of metabolites present in the biological sample. In addition, 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). Unlike GC-MS, where derivatization is essential, and LC-MS, which is biased against less polar compounds (Hendrawati *et al.,* 2006; Ward *et al.,* 2007), ^{1}H NMR can detect all the proton-bearing (^{1}H) compounds including most of the non-polar "organic" compounds such as carbohydrates, amino acids, organic and fatty acids, amines, esters, ethers and lipids present in a sample (Ward *et al.,* 2003). A combination of MS and NMR techniques is used here to have a large coverage of the *Brassica* metabolome, and is reported as the most powerful and informative to detect and identify metabolites from a complex mixture (Smith *et al.,* 2006; Verpoorte *et al.,* 2007).

Phenotypic and genotypic characterization plays an important role in plant breeding and genetic studies. The use of DNA markers has emerged as a powerful tool for the assessment of genetic relationships and exploring the genetic make-up of collections of accessions. AFLP (Amplified Fragment Length Polymorphism, (Warwick *et al.,* 2008; Zhao *et al.,* 2005) and MYB (myeloblastosis, (Diez, 2008) markers are very useful for genotyping and genetic analysis of *Brassica* species. AFLP markers are highly reproducible and cost efficient, and can generate a large amount of genetic information sampled across the entire genome rather than from a specific location without prior sequence information in a rapid way (van Berloo *et al.,* 2008; Vos *et al.,* 1995; Zhao, 2007). The MYB protein family, the largest transcription family, is involved in metabolic pathway regulation (secondary metabolism), pigmentation, developmental control, regulation of plant responses to environmental factors and hormones, and Myb genes are present throughout the genome of *Arabidopsis* (Riechmann and Ratcliffe, 2000). Diez (2008) reported in a Master thesis that MYB markers such as AtMYB 28, AtMYB29 and MYB 34/ATR1 are distributed throughout the genome and involved in the regulation of biosynthesis of secondary metabolites, for instance glucosinolates, in *Brassica* species.

The application of multivariate statistical analysis tools is useful for the distinguishing different groups of samples based on phenotypic and genotypic information, as well as to filter out the markers or metabolites correlated with specific traits (Weckwerth and Morgenthal, 2005). In this study, we applied both unsupervised multivariate tools (cluster analysis and principal component analysis (PCA)) and supervised tools, such as random forest classification and regression which uses a *priori* information.

Hierarchical clustering is an unsupervised classification technique which discovers and visualizes group structure across accessions at different levels (Dopazo, 2007), while PCA is a dimension reduction technique that allows to visualize and help the interpretation of groupings of accessions based on linear combinations (known as principal components) of the original variables preserving most of the information (Weckwerth and Morgenthal, 2005). These statistical tools enable the visualization of morphological, genetic and metabolic coherence based on their inherent correlative behavior.

Random forest (RF) is a machine learning statistical tool for classification as well as non-parametric regression with variable selection features which can identify the relevant predictor variables even in complex interactions (Diaz-Uriarte and Alvarez de Andres, 2006; Pang *et al.,* 2006; Strobl *et al.,* 2008). The popularity of this tool has been increasing within the scientific community because of its ability to handle wide datasets with higher number of variables than samples, high correlation among the variables (called multi-collinearity), large numbers of noise variables, mixtures of categorical and continuous predictors, multi-class problems with high predictive accuracy (Diaz-Uriarte and Alvarez de Andres, 2006; Strobl *et al.,* 2008). It has an internal cross-validation and few parameters need to be fine-tuned (Gislason *et al.,* 2006). It returns the most important variables which account for either classification of different groups or which explain large parts of the variation present in a continuous trait of interest in a non-parametric regression approach (Pang *et al.,* 2006; Strobl *et al.,* 2008).. Here, we apply RF classification to select metabolic peaks and genetic markers important in classifying the different groups of the core collection. Additionally, the RF regression approach was used to identify the metabolic peaks and genetic markers which are associated with leaf color, an important morphological trait for the characterization of morphotypes.

Metabolic variation is the evident even in the samples from identical genotypes under well-controlled conditions. The variations of the metabolites depend on the levels of other metabolites under the pathway. Metabolic networks can give a glance of the physiological pathways of the plant at a particular development stage by visualizing the observed correlation of metabolic variation (Steuer *et al.,* 2003). Networks can either be based on simple Pearson correlations or on partial correlations. Simple Pearson correlation shows the both direct and indirect correlation between the metabolites, whereas partial correlation used to filter out the indirect correlation. A partial correlation measures the correlation between the two variables after the influence of one or several other variables is removed or controlled (Khanin and Wit, 2007). Such correlations can reflect the underlying biochemical network of the pathways (Morgenthal *et al.,* 2006). Hence, a correlation network is a useful tool to get a “fingerprint” of the underlying metabolic pathways. Metabolites or peaks with higher numbers of connections are regarded as the essential compounds of a particular pathway. The integration of associated markers in the network can benefit from additional genetic information based on marker-metabolite connection (Khanin and Wit, 2007) as well as relatedness of the compounds, and can help identify the unknown compounds.

In this study, a core collection of *B. rapa* species is used to explore the groups and sub-groups of accessions based on the morphology, metabolic composition (NMR and LC-MS) and molecular markers using unsupervised hierarchical clustering and PCA. Univariate as well as multivariate statistical tools are applied to identify phenotypic traits as well as genetic markers which can distinguish the (group?) structure and morphotypes of the core collection. Association studies are conducted with the objective of identifying LC-MS peaks and molecular markers correlated with leaf colour. The associated metabolic peaks are visualized in correlation networks to make the aid in the interpretation of relationships, and are integrated with their associated markers to discover the importance of markers as well as and roles of the metabolic peaks.

In addition, results of the targeted metabolic profiling of the carotenoid pathway of *B. rapa* species is visualized with associated markerss in a correlation network, and compared with the underlying pathway in order to get to know the relations between metabolites and makers. This study will help to understand the variation at morphological, genetic and metabolic levels in *B. rapa* species. Association studies and networking is useful to know the underlying pathways in relation to genetic information.

Chapter 2: Background of the data

2.1 Plant material

A core collection of 169 accessions of *Brassica rapa* with different morphotypes originating from various parts of the world (Table 1; appendix) was studied to explore their morphological, genetic and metabolic variability, and study the inter-relationships between accessions as well as between variables over accessions. The morphotypes included were Broccoletto (BRO), Chinese cabbage (CC), Chinese turnip rape (OR), Fodder turnip (FT), Komatsuma (KOM), Mizuna (MIZ), Pak Choi (PC), Spring oil (SO), Vegetable turnip (VT), Winter oil (WO), Yellow Sarson (YS) and Turnip green (TG) (Fig. 1). Within this core collection, a set of 31 lines consisting of 19 lines of Chinese cabbage (CC), 3 lines of Komatsuna, 4 lines of Pak choi (PC) and 5 lines of turnip (T-1) were obtained from 5 plant breeding companies of the Netherlands (Takii Europe BV, Bejo Zaden BV, Nick , Seminis and Syngenta).

Chinese Cabbage (CC), turnip types (FT + VT) and oil types (WO+SO+YS+OR) were the dominating numbers of accessions in the core collection. These accessions were collected from various parts of the world by three gene banks namely the Dutch Crop Genetic Resources Center (CGN) of Wageningen, and Chinese Academy of Agricultural Sciences (Institute of Vegetables and Flowers (CAAS-IVF) and Oil Crop Research Institute (CAAS-OCRI)) of China (Requena, 2007; Zhao, 2007). Two accessions were kindly obtained from Dr. T. Osborn (University of Wisconsin, USA).

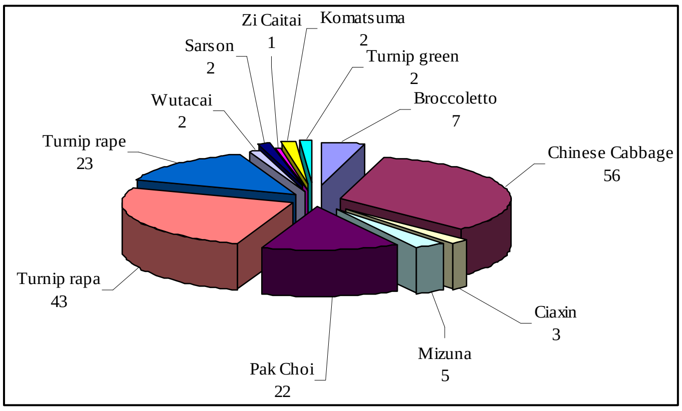


Fig. 1: Composition of a *B. rapa* core collection with number of accessions per morphotype

2.2 Morphological data

Morphological data were recorded from plants in two growing conditions: vernalized and non-vernalized conditions. For the vernalization, germinated seeds were treated with cold temperature (5^0C) in a dark room for 31 days and then transplanted to the greenhouse (16 hours light, 18-21^0C temperature). As a result the plants were vernalized to induce flowering and seed setting. The dataset consists of 26 morphological traits on 164 accessions. The data were acquired in experiments conducted in a randomized complete block design (RCBD) with 4 blocks in the March of year 2007. For the non-vernalized condition, seeds were directly sown in pots without prior cold treatment. Altogether only 11 flower morphological traits in 2006 and SPAD (a quantitative measurement of leaf colour) measurements in August of both 2006 and 2007 have been recorded. Thus, data from vernalized condition that have observations on all traits (Table 2) were only analyzed using multivariate statistical tools to explore the morphological relations of morphotypes and their characterization. The traits were grouped into four categories; flowering traits, leaf traits, flower morphological traits and plant architectural traits. Among a total of 26 morphological traits, leaf color (LC), leaf edge shape (LES), petal shape (pS) and petal color (PC) are qualitative traits, and presence of petiole (PP) is a binary trait, which were visually scored on ordinal and/or binary scale. The traits had been recorded in different measurement units; however, some observations, especially in floral morphological traits of turnip rape, were missing because these accessions flower extremely late and seedling vernalization does not accelerate flowering. The details of traits measured and their descriptions are shown in Table 2. Only SPAD traits from both treatments (vernalized and non-vernalized conditions) were used in a trait-specific association study. SPAD observations correspond to the amount of chlorophyll present in leaves or leaf colour, where the transmittance of red light (650 nm) and infrared (940 nm) radiation through the leaf was measured with the help of a SPAD meter (SPAD-502 (Minolta SPAD Chlorophyll meter)) (Requena, 2007). These measurements were recorded per block in 2 days (between 10:00 and 13:00 hr) on the adaxial side of the premature as well as fully mature leaves, and then averaged.

2.3 Marker data

The marker genotyping was done by using 218 AFLP and 141 MYB motif-directed markers on 168 accessions with few missing values. MYB markers target the largest family of transcription factors of *Arabidopsis*, which are generally present throughout the genome of the *Arabidopsis* but, sometimes, also appear in clusters (Diez, 2008). AFLP markers are generally more scattered across the genome, however, this depends on the restriction enzymes used. Three primer combinations for AFLP markers and 4 enzymes for MYB markers were used for marker genotyping. Among these markers, 90 markers have known map positions in a reference Double Haploid (DH) mapping population of a *Brassica rapa* cross (Yellow Sarson 143 x Pak Choi 175). Both types of markers are dominant markers.

2.4 NMR data

^{1}H-NMR is a powerful tool to identify the structural group of complex mixture of compounds because of presence of hydrogen in almost all molecules of biological samples (Liang *et al.*, 2006, Ward *et al.*, 2003). ^{1}H-NMR, one-dimensional approach, measures the nuclear spin of the H atom of molecules by using radio frequency pulses. The subsequent emission of radiation is detected as signal of compounds.

NMR data with 236 bins in the range of 0.32-10 ppm, bucketed every 0.04 ppm, were made available for analysis. The measurements were carried out on 50 mg sample (dry weight basis) from each 166 accessions (5 weeks old) using 500 MHz Brucker and NMR solvent [MeOD-KH_2PO_4 buffer in D_2O (1:1, v/v), pH 6]. The data were already scaled to total intensity as a pre-processing step so that each bin had relative area of the total intensity. Water bucket effects and technical errors on 4.96-4.76 and 3.32 spectrums were removed before proceeding analyses. In comparison to marker and LC-MS datasets, two accessions; cWU56 and RC-144 were absent in the NMR dataset.

2.5 LC-MS data

LC-MS (Liquid Chromatography-Mass Spectrometry) is widely used method to detect the heat-labile compounds, such as phosphates compounds, co-enzyme (CoA), isoprenes, alkaloids, phenylpropanoids, glucosinolates, and flavonaoids. Unlike gas chromatography, a solvent is used in the mobile phase to isolate the compounds from the samples and photodiode-array (PDA) was used to break the compounds into varying sizes of fragments. Ionization can be done based on positive or negative charge (Hall, 2006) but negative mode ionization was used in this study to detect compounds such as, isoprenoid, flavonoids, glucosinolates. The fragmented masses were detected by mass detection devices quadruple-time-of-flight (QTOF). Finally, LC-MS result a 3-dimensional graph of chromatogram and mass spectrum consisting of signal intensities (abundance), retention time and mass-to-charge (m/z) ratio value. Thus, chromatogram separates the compounds present in original mixture of sample and MS provides their fragmentation patterns which are unique to each compounds.

The dataset of untargeted metabolic profiling was provided for the identification of grouping patterns of accessions, and discriminating metabolic peaks among different morphotypes. In metabolic profiling, LC-PDA-QTOF-MS technique was applied to single-plant-samples of 168 accessions (same as marker dataset) at 5 weeks old (except few early flowering accessions were at flowering stage) with 12 technical replicates from the accession CC-068 and biological controls of RO18 and L58 accessions. Pre-processed signals for accessions where none of the 168 accessions had peak intensities greater than 200 value were deleted to prevent the effect of noise signals, and hence, the dataset was reduced from 46779 peaks to 5546 peaks. Most of the LC-MS peaks were represented by the combination of centrotype, mass and scan number in the form of “centrotype_mass_scan”. The centrotypes had been assigned based on the multivariate correlation of peak intensities of masses in combination with the scan number (retention time), so that a centrotype consists of peaks that are correlated to each other and fall within the same retention time window. Some peaks that were not allocated to any centrotype were coded with the letter “A” followed by a number, for example A1, A2, A3 and so on. In metabolomic studies, a transformation using the logarithm is essential to make a normal distribution with intensity-independent variance (Pietiläinen K.H *et al.*, 2007). Here, a $\log_2$ transformation was done to decrease the influence of very high values as well as stretch the very low values (Steinfath *et al.*, 2008).

In addition, a targeted metabolic profiling of 16 metabolites; folate, chlorophyll-a and -b, β-carotene, lutein and its derivatives, neoxanthin, violaxanthin, and α-, β-, γ-and δ-tocopherol of carotenoids and tocopherol pathway were measured via. LC-PDA-QTOF-MS technique. Those metabolites were used for network analyses.

Table 2: Phenotypic traits used for the measurements of a core collection of *B. rapa* after vernalization

Trait type	Abbreviation	Description
A. Flowering trait		
Flowering in time	DTF	Number of days from transplant till the appearance of the first open flower (days)
B. Leaf traits		
Leaf length	LL	from base of petiole to tip of lamina (cm)
Lamina blade length	Lbl	Distance from the tip lamina to the fist 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 (cm^2)
Leaf perimeter	LP	The edge of full leaf (cm)
Petiole length	PL	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	PP	Score (0= absent, 1= present)
SPAD	SPAD	Chlorophyll content
C. Flower Morphological 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 (mm^2)
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)
D. 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)

Chapter 3: Study of morphological, molecular and metabolic relationships of different *B. rapa* morphotypes

3.1 Materials and Methods

3.1.1 Cluster analysis

Clustering, also called class discovery, is an exploratory data analysis tool which sorts homogeneous groups of samples across the variables into respective categories by maximizing the degree of similarity within a group and dissimilarity between the groups (Wit and McClure, 2004)**.** An agglomerative hierarchical clustering algorithm was used to explore the relationships of accessions at different levels (Dopazo, 2007). This algorithm first considers all the accessions as separate ones and then successively groups accessions into larger and larger clusters until only a single cluster is obtained (Podani, 2001). The Pearson correlation coefficient was used to calculate a dissimilarity matrix. This dissimilarity measure makes groups of the accessions based on their patterns of observations rather than on the size of the raw values (Wit and McClure, 2004) on morphological traits or metabolic profiles. However, Jaccard's distance was used for the molecular marker data because of its better performances in analyzing asymmetric binary variables (Duarte *et al.,* 1999; Everitt, 1980) such as dominant markers. The most popular linkage method, UPGMA (Unweighted Pair-Group Method using Arithmetic average), a type of average linkage method, was used to calculate the distance between two clusters. Unlike the single and complete linkage methods, UPGMA grasp the information about all pairs of distances (Quackenbush, 2001), and joins two clusters having the lowest average distance to form a new cluster.

Cluster analyses of morphological and NMR data were done using the "pvclust" package, an add-on package for R-statistical software, which assesses the uncertainty of hierarchical clustering caused by sampling error of data through two types of measures: Bootstrap probability (BP) and Approximately Unbiased (AU) p-value (Suzuki and Shimodaira, 2006a). The BP of a cluster means the frequency of the cluster that appears in the bootstrap replicates (Shimodaira, 2002). The BP has been widely used in many scientific studies; however, it is biased due to use of constant sample size throughout the bootstrap replicates (López-López *et al.,* 2008), which may not be the case at population level. The AU test is newly devised multi-scale bootstrap technique which reduces the bias in hypothesis testing (López-López *et al.,* 2008; Shimodaira, 2002; Suzuki and Shimodaira, 2006b)**,** and controls the type-I-error (Shimodaira, 2002). Multi-scale bootstrapping procedure, in this study, bootstrap the samples with 10 different sample sizes (ratio of new sample size; N' and original sample size; N ranged from 0.5 to 1.4 by 0.1 as default setting), and generated 10,000 bootstrap replicates of each sample size. Hierarchical clustering was performed to each bootstrap sample to get the sets of bootstrap replications of dendrogram, and computed BP for each sample size. A theoretical curve (z(N') = v * sqrt (N'/N) + c * sqrt (N/N')) is fitted to the observed BP values along the different sample sizes to estimate coefficients v and c, and AU p-value is calculated by using the equation AU = Φ(-v + c) where, Φ is the standard normal distribution function, and v and c are coefficient of each cluster (López-López *et al.,* 2008; Shimodaira, 2002; Suzuki and Shimodaira, 2004; Suzuki and Shimodaira, 2006b). The AU p-value gives the probability of each cluster at population level (Suzuki and Shimodaira, 2004). The greater the p-value, the greater the probability a cluster is the true cluster.

High memory requirement in bootstrap procedure, especially for large dataset (LC-MS), and lack of distance function for binary variables for marker data in “pvclust” package, GeneMaths statistical software (Applied Maths BVBA, Belgium) was used instead of bootstrapping procedure for cluster analyses of LC-MS and marker dataset.

The software STRCTURE was used to identify the groups/classes in the populations based on Bayesian clustering approach, and assign the individual samples with their membership probabilities of being in those classes in unsupervised approach. This approach classifies the accessions into classes on the basis of their marker genotype under the following assumptions: the marker loci linkage disequilibrium between subpopulations (classes) and in Hardy-Weinberg equilibrium state within a subpopulation (Pritchard *et al.*, 2000)..

A supervised approach such as ‘classification’ is used to classify accessions into given input classes; however, unsupervised approach such as ‘cluster analysis’ discovers the classes without prior information. RF predicts the membership probability of each accession in a supervised classification approach in which the class information is used as *prior* information. The comparisons on the alignment of the accessions were made, in this study, among unsupervised hierarchical classification based on different (morphology, marker, NMR and LC-MS) datasets, STRUCTURE software given membership probability (using only the marker dataset) and RF membership probability on different datasets (morphology, marker, NMR and LC-MS) with prior information of STRUCTURE classification. Hence, these comparisons give a visual evaluation of hierarchical clustering of accessions. The details of this random forest analyses are described in Chapter 4.

3.1.2 Principal Component Analysis

PCA is the most commonly used visualization technique in multivariate statistics. It finds the variability of accessions with minimum loss of information available in the dataset. Pearson correlations between all variables were, first, observed to get the overview of suitability of all the datasets for principal component analysis (PCA). PCA was done for the morphological, NMR and LC-MS dataset using their correlation matrices to discover groupings of accessions based on the patterns of metabolic expression level. All these analyses were conducted by using the “FactoMineR” package in R-software (Husson *et al.*, 2008). For PCA, LC-MS data were log (base 2) transformed but for NMR, data were autoscaled with and without log (base 2) transformation. For autoscaling, the mean of each bin was subtracted from individual observations and divided by their respective standard deviation. Despite having various thumb rules namely; 80 % rule, broken stick model (Zuur *et al.*, 2007), Kaiser’s rule (eigenvalue-greater-than-one-rule), Horn’s procedure(Lattin *et al.*, 2003), the most practical and commonly used approach “elbow-effect” in scree plot was used to determine the number of PCs for score plots because of its simplicity, and high limitation on sample size on aforementioned thumb rules (Lattin *et al.,* 2003).

Score plots, a tool of PCA, were used to visualize the highdimensional data in lower dimensions (first few PCs). Similarly, correlation-variable plots were made to visualize the variables (morphological traits) space in relations to accession groupings, and to examine the cosine angles between variables. The cosine angle between the traits in correlation-variable plot approximates the correlation of two traits, while its length is proportional to the variance. Traits which are plotted in the same direction have a high and positive correlation, at angle 90^0 they have a small correlation, and in opposite directions indicate high but negative correlation (Lattin *et al.,* 2003). For each PC, the loadings (or weights) reflect the influence of the original variables on the PCs, whereas the scores (coefficient of the PC) reflect the contribution of each PC in every sample (Colquhoun, 2007). Because of very large numbers of variables and too messy figures without any variable groupings, variable-plots in case of NMR and LC-MS are not shown.

3.1.3 Principal Co-ordinate Analysis

Principal Coordinate Analysis (PCoA), also called metric multidimensional scaling (metric MDS), was applied for the marker dataset. It can be used to visualize any kinds (dis)similarity matrices (Zuur *et al.,* 2007); however, PCA utilizes only either correlation or covariance matrix. The AFLP and myb markers, in this study, are dominant in nature and had properties of asymmetric binary variable. Hence, Jaccard's binary distance function was used to calculate dissimilarity matrix instead of using correlation or covariance function. Similar to PCA, PCoA reduces the dimensionality based on eigenvalue equation, and produces latent variables. An add-on package "ecodist" was employed to visualize the location of accessions in low-dimensional spaces in the R-software (Goslee and Urban, 2007a).

3.1.4 Mantel test

The Mantel test was used to evaluate the pairwise correlations between dissimilarity matrices (Goslee and Urban, 2007b; Luo and Fox, 1996; Pissard *et al.,* 2008). Since Mantel test uses a permutation procedure to test the statistical significance of matrix correspondence with the null hypothesis of random correlations, it is robust to correct type I error (Legendre, 2000) and widely used in population genetics and ecological study (Telles and Diniz-Filho, 2005). Mantel test yields Mantel r statistic based on the normalized Mantel statistic (equation 1). This statistic is normalized via a standard normal transformation where the mean of the matrix is subtracted from each element and then each element is divided by the standard deviation.

$$\frac{\sum\sum(d_{X,ii'}-\bar{d}_X)(d_{Y,ii'}-\bar{d}_Y)}{\sqrt{\left[\sum\sum(d_{X,ii'}-\bar{d}_X)^2\right]\left[\sum\sum(d_{Y,ii'}-\bar{d}_Y)^2\right]}}, \qquad \text{..........(1) (Dutilleul } et\ al.,\text{ 2000)}$$

where, $d_{X,ii'}$ and $d_{Y,ii'}$ for the distances between observational units *i* and *i'*, and $\bar{d}_X$ and $\bar{d}_Y$ are the means of the distances derived from the observations on variables X and Y respectively.

Mantel test resolved the violated assumption of independence by making random permutations between rows and columns of either test matrix (Goslee and Urban, 2007b; Luo and Fox, 1996; Smouse *et al.*, 1986). In this study, only 163 accessions common to all datasets were taken to avoid the influence of differing numbers of samples. Distances matrices were calculated from autoscaled data (subtraction of the mean of each variable from individual observations and divided by their respective standard deviation) in order to create distances based on the correlation matrices (Lattin *et al.,* 2003) because the Pearson correlations were also used in clustering and PCA for morphology, NMR and LC-MS. Jaccard's binary distance measure was used for the marker dataset. An "ecodist" package in the R statistical software was used to calculate all the distance matrices and simple mantel tests (Goslee and Urban, 2007a), whereas Mantel test was conducted with 10,000 permutations .

3.2 Results

3.2.1 Cluster analysis and Principal Component Analysis (PCA)

3.2.1.1 Morphological traits:

Applying cluster analysis on all 26 morphological traits from all 164 accessions, five groups were identified at 95 % confidence level, marked by red box in Fig. 2; groups were mainly composed of Mizuna, Pak Choi (PC), Oil (mainly SO form Bangaldesh and India), European Turnip (ET) and Chinese Cabbage (CC) were identified. However, a mixture of PC and Oil (WO form Pakistan and OR from China), ET and CC could be observed at higher level of dissimilarity. Mizuna consists of only two genotypes. European Turnips were separated far apart from rest of the accessions whereas, PC and Oils were in a close group, and broccoletto from Italy were attached in the group of CC from Asia (Fig. 2). Numbers in red color in each edge meant probability of each cluster (in percentage) to be a true cluster. The clustering of the accessions based on morphological traits was generally in agreement with their geographical origin and had good correspondence with the membership probability calculated through random forest statistical analysis on morphological traits in vernalized conditions (barplot ^{b}), and STRUCTURE class membership probability based on AFLP and MYB markers (barplot ^{c}). Most of the company based lines were within the cluster of older accessions of respective morphotypes indicating the possibility of similar morphological characteristics.

High correlation of the morphological traits (Fig. 3; Appendix) suggested PCA for good visualization of genotypes in lower dimensional spaces, and dimension reduction by forming orthogonal latent variables for easy interpretation. More than 50 % the total variance, (31.22 % by PC1 and 23.87 % by PC2), was explained by first two PCs (Table 3).

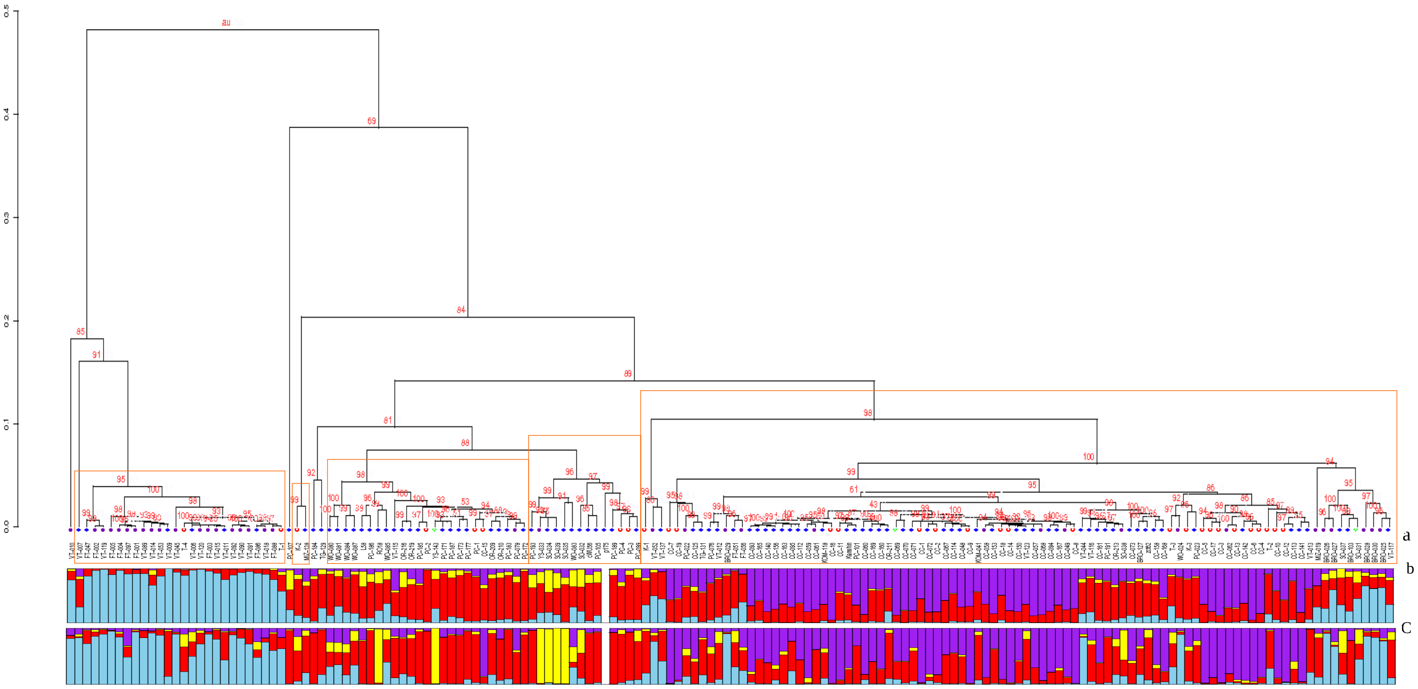


Fig. 2: Dendrogram of *Brassica* core collection based on morphological traits in vernalized condition with Pearson correlation distance function and UPGMA linkage method, ^{a}-indicates geographical origin (purple: Europe, blue: Asia, red: Company lines, green: America), ^{b}-Supervised membership probability of being in four STRUCTURE classes obtained from random forest analysis on morphological traits, ^{c}-Unsupervised membership probability of being in four classes calculated by STRUCTURE software using markers. Colors in barplots ^{b} and ^{c} indicate skyblue: class 1, red: class 2, yellow: class 3, purple: class 4 and white: no class information.

PC1 resulted in good separation of European Turnips (ET) and Chinese Cabbage (CC) (Fig. 4), where leaf traits (PL, LP, LL, LB, LA and LES) and DTF were accountable for ET and plant architecture traits (LN, PB, PH and PfH) and LC characterized CC. CC had yellowish green leaf color (LC) and were taller in plant height at both pre-mature stage (PH = 70.02 cm) and at mature stage (117.79 cm). PC and CC were clearly distinguished by PC2 (Fig. 4), and higher petal area (PA), petal width (PW), corolla length (CL), corolla width (CW), lamina width (LW), leaf area (LA) and petiole presence (PP) were observed in CC in contrast to PC, however, petiole were present in Pak Choi (PC) (Fig. 6). A small group of Mizuna was identified in PC3 which had high petal index (high petal length but low width),

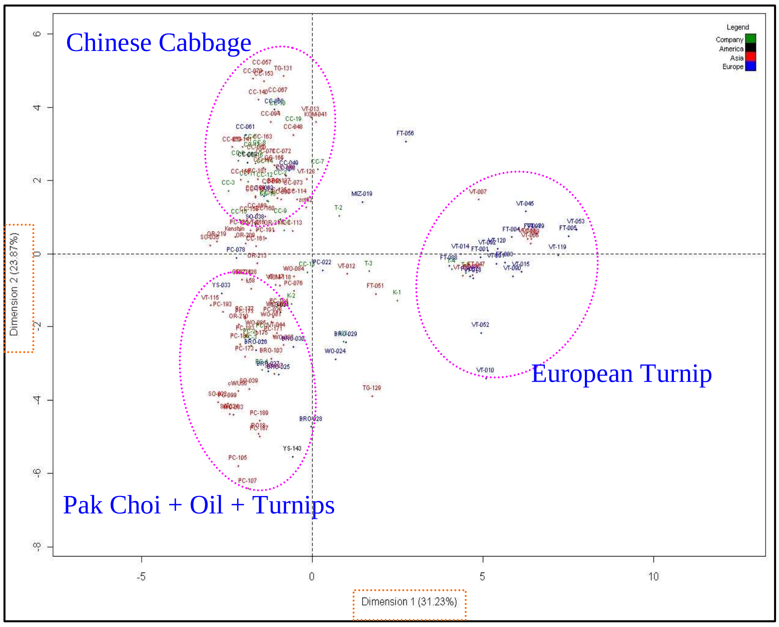


Fig. 4: PCA-Score plot of 163 accessions of B. *rapa* core collection based on 26 morphological traits with PC1 (dimension=1) and PC 2 (Dimension=2)

elongated petal shape, leaf edge shape (LES), and slightly higher number of leaves (12) when the first flowers opens (Fig. 6). Some traits; LA, LW, LbI, PfH and LC were seems more important in distinguishing genotypes because of higher loadings in first two PCs (Table 4). Similarly, different levels of correlation in terms of cosine angle among the traits can be observed in variable plot (Fig. 6) Different colors in score plot represent the origin of the accessions: colors were not well separated in PC2 and PC3. However, all the Asian (red color) and European originated (blue color) *Brassica* accessions, with few exceptions, were distinct in the first PC. Like clustering, almost all company lines and three American accessions were, interestingly, in the closed with older accessions of their morphotypes (Fig. 4 and Fig. 5).

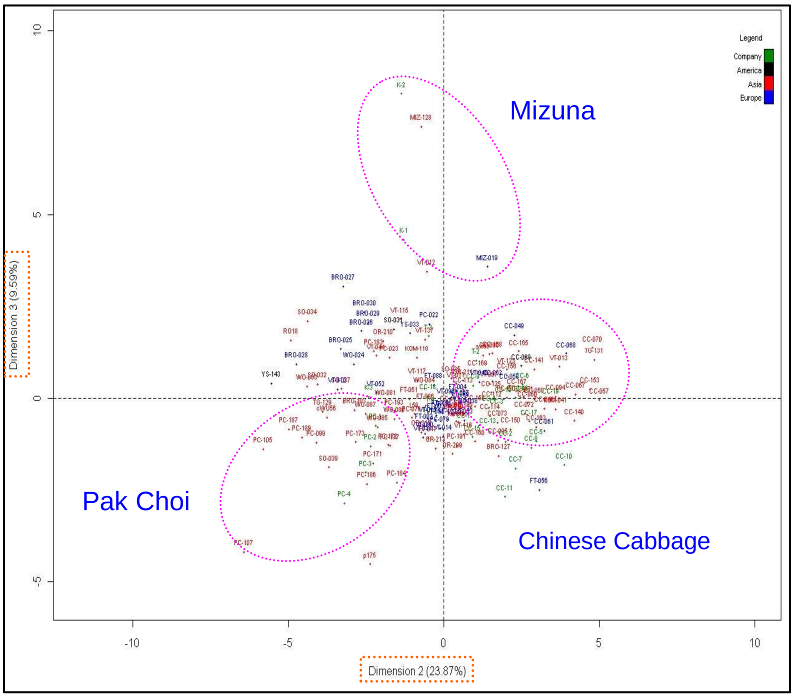


Fig. 5: PCA-Score plot of 163 accessions of B. *rapa* core collection based on 26 morphological traits with PC1 (dimension=1) and PC 2 (Dimension=2)

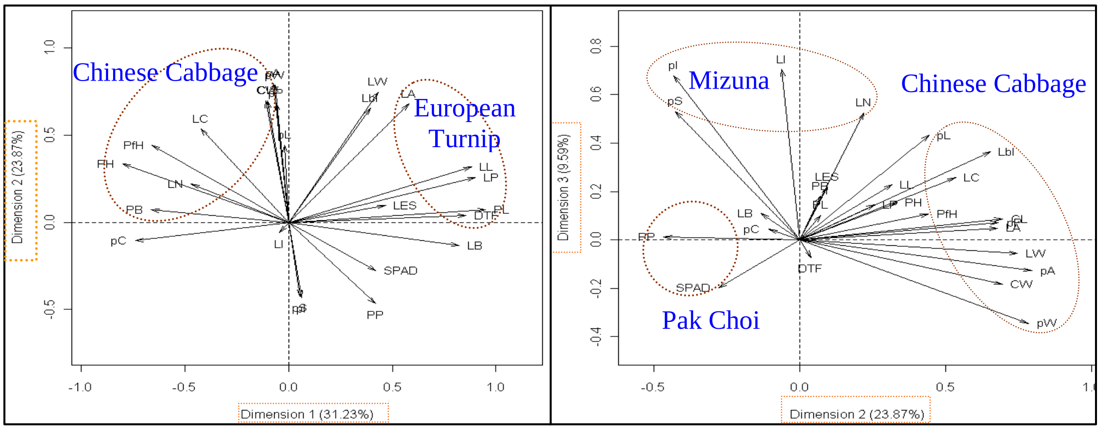


Fig. 6: PCA-variable plot of 26 morphological traits in PC1 (Dim=1), PC2 (Dim=2) and PC3 (Dim=3) in *B. rapa* core collection.

Table 3: Variance and Cumulative variances explained by first few PCs on morphological traits scored over 164 vernalized accessions

PCs	Percentage of variance	Cumulative percentage of variance
PC 1	31.22	31.22
PC 2	23.87	55.10
PC 3	9.58	64.69
PC 4	7.60	72.29
PC 5	5.36	77.65
PC 6	4.24	81.89

3.2.1.2 Molecular markers

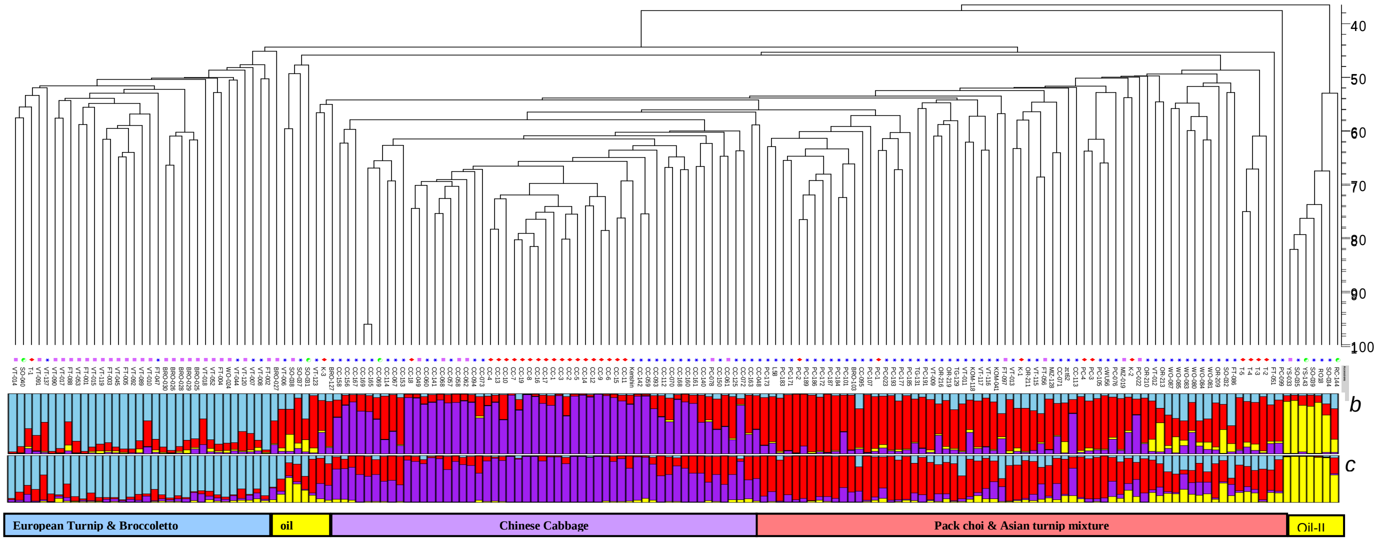


Fig. 7: Dendrogram showing the groups of *Brassica* core collection based on DNA markers with jaccard's distance function and UPGMA linkage method. [a]-indicates geographical origin of accessions (pink: Europe, blue: Asia, green: America, red: Company), [b]-Supervised membership probability of being in four STRUCTURE classes obtained from random forest analysis on markers, [c]-Unsupervised membership probability of being in four classes calculated by STRUCTURE software using markers. Colors in barplots [b] and [c] indicate skyblue: class 1, red: class 2, yellow: class 3, and purple: class 4.

Four distinct groups; European Turnip (ET) and Broccoletto (BRO), Chinese Cabbage (CC), a mixture of Pak Choi and Asian turnips, and oils were well separated in cluster analysis and were in very good agreement with their geographic origin (Fig. 7). Oils were appeared in two sub-groups, however, only three accessions (SO-031, SO-037 and SO-038) in Oil-II sub-group with low membership probability and seven accessions (RC-144, SO-034, SO-035, SO-039, RO-18, YS-033, YS-143) in sub-group Oil-I. The membership probability of each accession calculated by RF was almost similar to STRUCTURE membership probability calculated based on molecular markers. Like morphological traits based clustering, most of the genotypes of breeding company fell within the groups of older accessions indicating the possibility of similar genetic backgroups in their pedigree.

Principal Co-ordinates Analysis (PCoA) also showed three distinct groups of ET, mix group (oil, Asian turnip, PC and CC) and oil in two dimensional spaces (dimension 1 versus 2) were well separated. Similar to clustering, oils groups were also found in two small sub-groups, where Oil-II sub-group (SO-032, SO-034, SO-037, SO-038 and RC-144) was in close distance with Asian turnip, PC and CC; however, Oil-I comprised of YS-033, YS-143, SO-035, SO-039 and RO-18 was distinctly isolated in PCo2 (Fig. 8). The coloring of the genotypes names indicates their respective origin (blue: Europe, red: Asia, black: America, green: Company) which was also well distinguished by PCoA. Only the first three dimensions had eigen value more than 1 (dim 1=3.07, dim 2=2.079 and dim 3=1.098) indicating those dimensions account for more variance than that of one of the original variables.

3.2.1.3 NMR

In NMR, cluster analysis was not effective in identifying different groups of *Brassica* accessions through different distance functions; correlation, Euclidean, Manhattan and data transformation techniques. Pearson correlation distance function was better than Euclidean distance function because of groupings of at least CC groups in PC1. In hierarchical clustering via multi-scale bootstrap, neither geographical origin nor STRUCTURE membership probability matched with the clustering, however, some patches of Chinese cabbage, oils and European turnip were observed (Fig. 9).

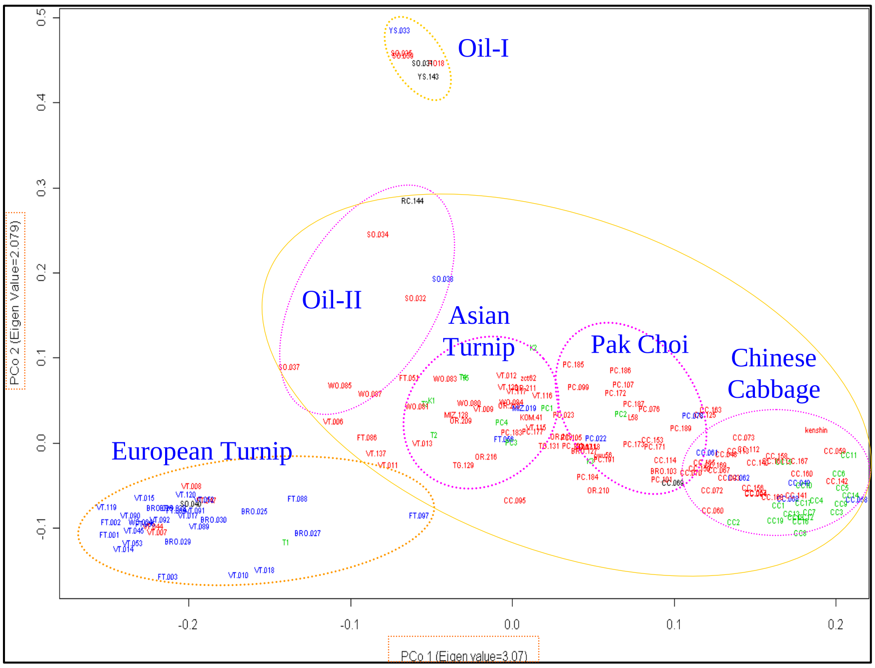


Fig. 8: PCoA-score plot of *B. rapa* core collection based on AFLP and MYB markers with PCo1 and PCo2.

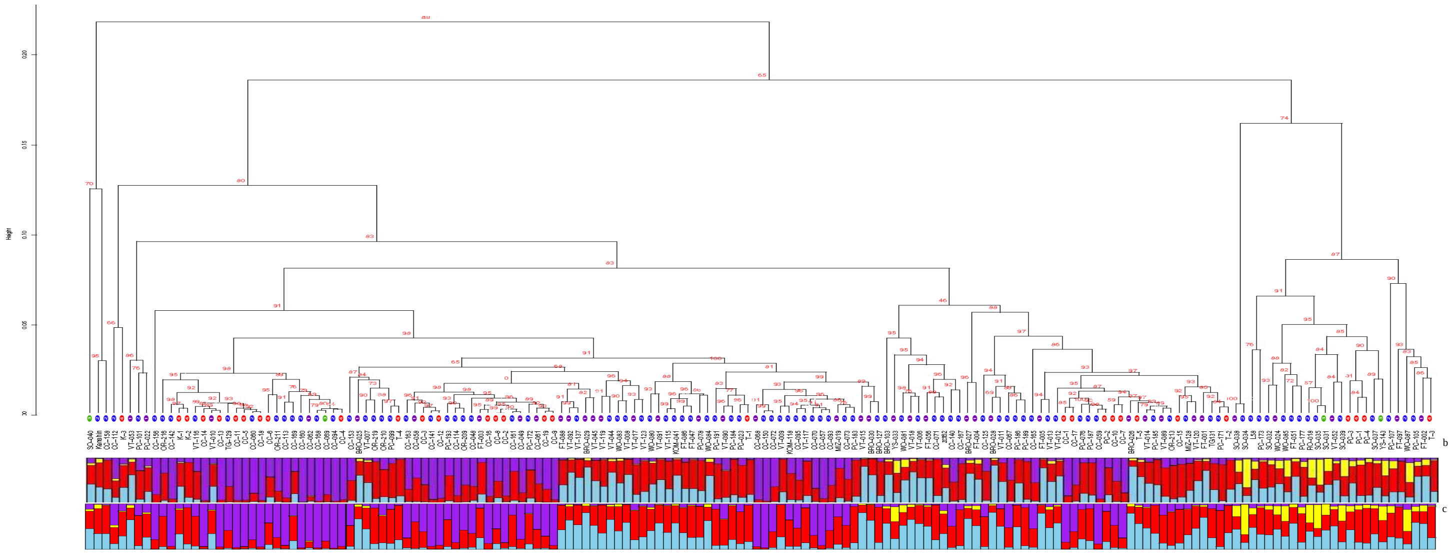


Fig. 9: Dendrogram showing the groups of *Brassica* core collection based on NMR bins with Pearson correlation distance and UPGMA linkage, [a]-indicates geographical origin of accessions (pink: Europe, blue: Asia, green: America, red: Company lines), [b]-Supervised membership probability of being in four STRUCTURE classes obtained via. random forest analysis on NMR dataset, [c]-Unsupervised membership probability of being in four classes calculated by STRUCTURE software on molecular markers. Colors in barplot[b] and [c] -indicate skyblue: class 1, red: class 2, yellow: class 3, and purple: class 4.

NMR bins were highly correlated with each other (Fig. 10; Appendix) because of wide coverage of metabolites. In PCA, albeit PC1 (45.42 %) and PC2 (10.98 %) explained more than 55 % of total variance (Table 5), only Chinese Cabbages (CC) were close together (Fig.11). Origin of the accessions (colored labels) did not comply with PCA score plot. Most of the genotypes obtained from breeding companies, especially CC, were also in the groups of Asian-originated CC (Fig.11). Large number of NMR bins had equal loadings on all PCs, however, almost all NMR bins of 8 and 9 ppm, and some bins of 1 and 6 ppm were important for positive impact on PC1. NMR bins 3.8-60, 3.68-60, 3.84-60, and 3.36-61 had relatively higher loadings and negatively correlated with PC1. Although first two PCs explained > 50 % of total variance; NMR bins had small differences in their loadings values across all PCs (data not shown).

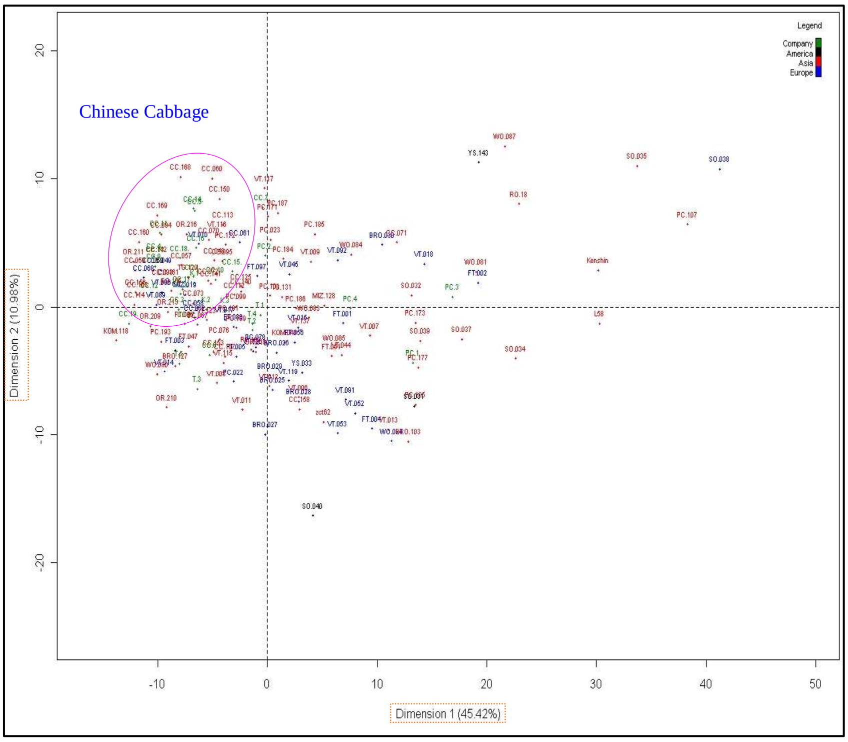


Fig.11: PCA-Score plot of B. *rapa* core collection based on NMR bins with PC1 (dim=1) and PC 2 (Dim=2)

Table 5: Variance and cumulative variances (%) explained by top 10 PCs in PCA on NMR dataset

PCs	percentage of variance	cumulative percentage of variance
PC 1	45.42	45.42
PC 2	10.98	56.40
PC 3	6.34	62.74
PC 4	5.03	67.77
PC 5	4.32	72.09
PC 6	3.37	75.46
PC 7	3.02	78.48
PC 8	2.42	80.90
PC 9	2.08	82.98
PC 10	1.66	84.64

3.2.1.4 LC-MS:

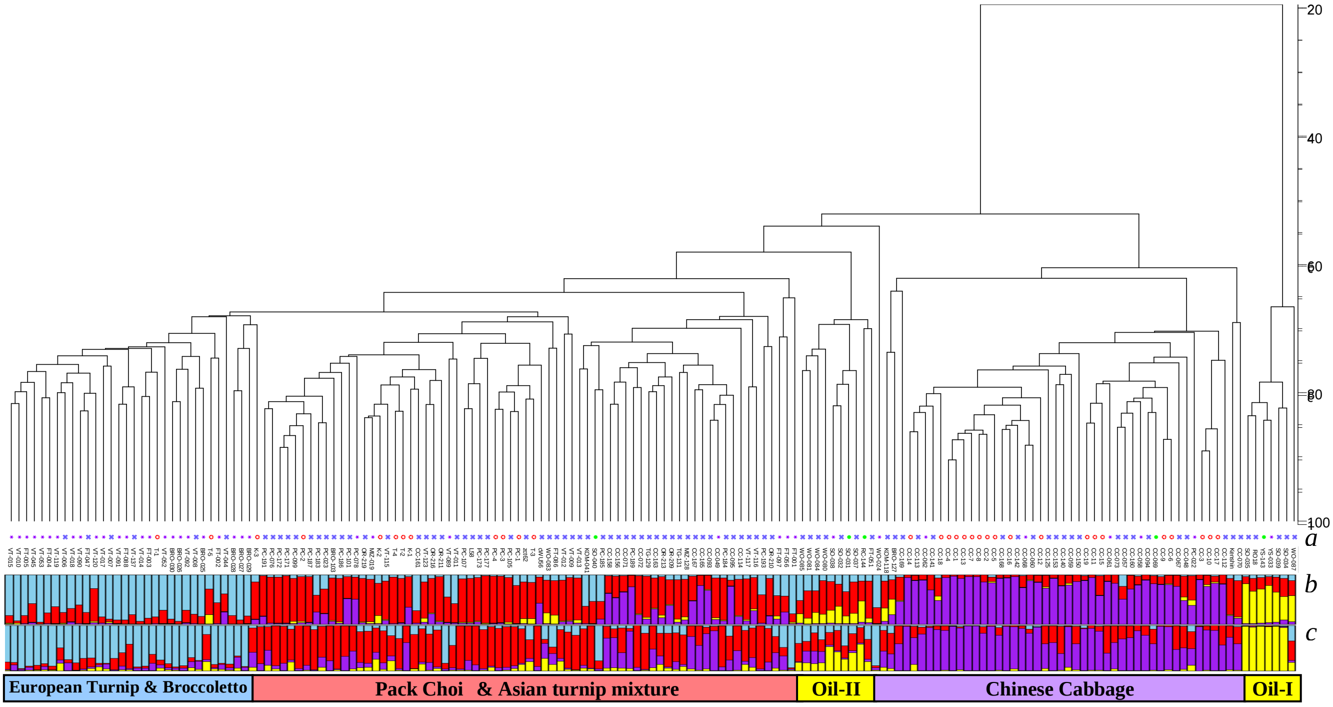


Fig.12: Dendrogram showing the groups of *Brassica* core collection based on LC-MS peaks with correlation distance function and UPGMA linkage method, [a]-indicates geographical origin of accessions (purple: Europe, blue: Asia, green: America, red: Company), [b]-Supervised membership probability of being in four structure classes calculated in random forest on LC-MS peaks, c-Unsupervised membership probability of being in four different classes calculated via STRUCTURE software using molecular markers. Colors in barplot[b] and [c] indicate skyblue: class 1, red: class 2, yellow: class 3, and purple: class 4

Cluster analysis on LCMS data showed the presence of four distinct groups of *Brassica*. Like molecular markers, two small sub-groups of oils; Oil-I (SO-034, SO-035, SO-039, RO-018, YS-033, YS-143) and Oil-II (RC-144, SO-031, SO-032, SO-037, SO-038) were discovered. The Oil-II sub-group had lower membership probability of accessions of being placed in oil groups. The membership probability of being in four classes calculated through RF analysis on this dataset was in good consent with their origin (purple: Europe, blue: Asia, green: America, red: Company) of accessions, and with the membership probability calculated using molecular markers via. STRUCTURE software (Fig.12). Similar to morphology and marker, company lines of Chineses cabbage were in close groups of Asian *Brassica*.

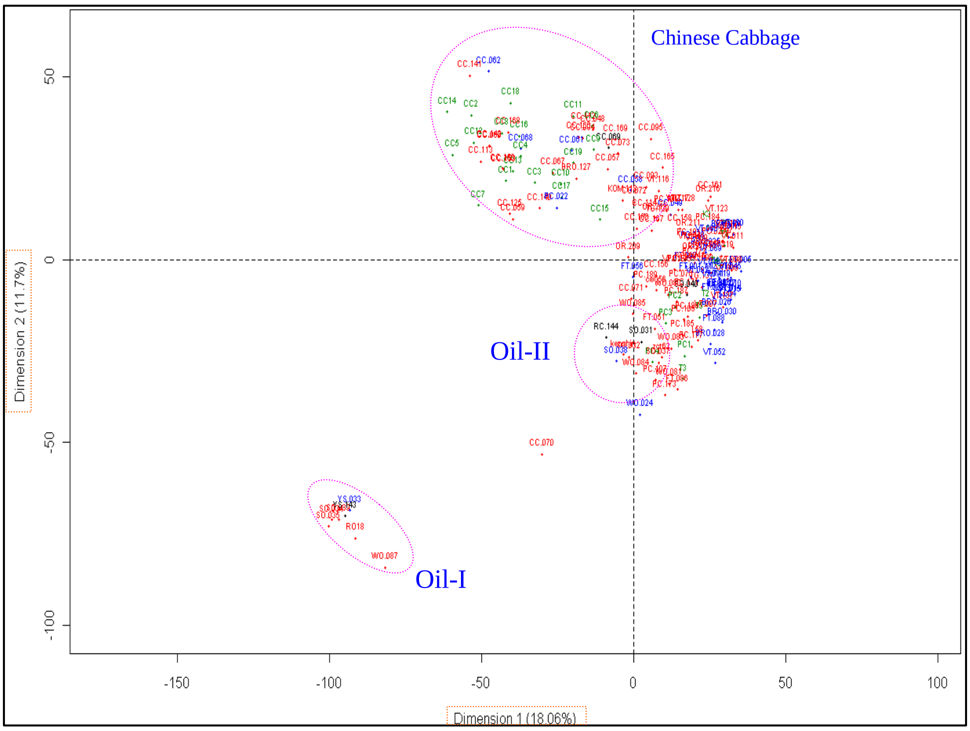


Fig.13: PCA-score plot of *B. rapa* core collection based on LC-MS peaks with PC1 and PC2

Table 6: Variance and cumulative variances (in %) explained by top 10 PCs in PCA on LC-MS dataset

PCs	percentage of variance	Cumulative percentage of variance
PC 1	18.06	18.06
PC 2	11.70	29.76
PC 3	4.74	34.51
PC 4	3.78	38.28
PC 5	2.81	41.09
PC 6	2.42	43.52
PC 7	1.92	45.44
PC 8	1.90	47.34
PC 9	1.75	49.09
PC 10	1.51	50.60

PCA showed two well-separated small groups of oil and a group of CC in correlation based score plot of PC1 (18.06 %) versus PC2 (11.70 %). The accession origin (blue: Europe, red: Asia, black: America, green: Company) did not correspond with the spatial distribution of accessions (Fig.13). The variances explained by PCs were quite low where top 10 PCs had carried out only 50 % of total variances (Table 6). Score plot in combination of either PC 2 versus PC 3 or PC1 versus PC3 did not give any new groupings of collections (figure not shown). Variable plots on some combinations of first few PCs did not provide any information to get idea of contributing variables of different accession groups (figure not shown).

3.2.2 Mantel test:

Simple Mantel test indicates highly significant correlation (α <0.01) between distance matrices of morphological traits (vernalized), molecular markers and LCMS indicating high correlation among them. However, non-significant correlations were observed in all combinations with NMR (Table 7) which was also observed in cluster analysis and PCA score plot.

Table 7: Mantel correlation between dissimilarity matrices

Dissimilarity matrix	Dissimilarity matrix	Mantel r	p-value (2-sided)	C.I. (95%)
Morphology	Marker	0.456	0.0001**	0.43, 0.49
Morphology	LCMS	0.17	0.008**	0.15, 0.21
Morphology	NMR	-0.018	0.70^{ns}	-0.048, 0.01
Marker	LCMS	0.48	0.0001**	0.423, 0.52
Marker	NMR	0.00224	0.967^{ns}	0.045, 0.0435
NMR	LCMS	-0.0222	0.701^{ns}	-0.057, 0.007

**- p-value < 0.01 level of significance, ^{ns}- non-significant at α=0.05

3.3 Discussions

Brassica rapa, composed of several morphotypes, is a widely grown species for vegetable and oil purposes in the world. Although *B. rapa* accessions are diverse in their geographic origins, this study was conducted to analyze their morphological, molecular and metabolic relations.

The application of unsupervised grouping techniques, namely hierarchical clustering, Principal Component Analysis (PCA) and Principal Coordinate Analysis (PCoA) on morphological observations, molecular markers and LC-MS peaks revealed the clusters of 168 accessions with good correspondence with geographic origins, and to lesser extent with morphotypes. A study of genetic relations of the same core collection based on AFLPs also showed that the accessions from the same geographic origins cluster together even if they were from different morphotypes (Zhao *et al.*, 2005). In hierarchical cluster analyses, the distinct groups of morphotypes; European turnip and broccoletto, Pak Choi and Asian turnip, two oil sub-groups, and Chinese cabbage (CC)were discovered based on 359 markers (218 AFLP and 141 myb) and 5548 LC-MS peaks rather than morphological traits and NMR spectroscopy (Fig. 2, 7, 12). Clustering based on both molecular markers and LC-MS peaks, accessions FT-047, VT-007 and VT-044 originated from Soviet Union, and accessions VT-006 and VT-008 from India fell within the group of European turnip. Since the Soviet Union was a very large country, which spanned Eastern Europe and Asia, we cannot be sure whether these accessions should be considered European or Asian. On the other hand, European originated turnips FT-056 (France) and FT-097 (Germany) were close to Asian turnips. This could suggest that these accessions were developed from germplasm from different geographic origins than their present growing area. Among those accessions, FT-056 from France was also reported as outlier accession of European turnip groups (Zhao *et al.*, 2005). Similarly, PC-022 from Netherlands (only one non-Asian Pak Choi in this study), was a strong outlier of Pak Choi group and close to CC in LC-MS profile. The difference in growing area might affect the quantitative levels of metabolites or be developed by crossing with CC. However, McGrath and Quiros (1992) reviewed hybrids between CC and PC is less vigorous than that of CC and turnips. Oil types (SO and YS and WO) accessions were clearly separated into two sub-groups based on LCMS and, to a lesser extent also the case based on AFLP/myb markers, although the numbers of oil accessions were few in comparison to other morphotypes. Based on molecular markers, one oil sub-group (Oil-I) consists of three SO types of Bangladesh (SO-034, SO-035, SO-039), annual oils (YS-033, YS-143) and Rapid cycling (RC-144) at higher genetic distance with high STRUCTURE class membership probabilities. The other oil sub-group (Oil-II) was consists of other three different SO types (SO-031, SO-037, SO-038) with low membership probabilities (Fig. 7). Similarly, in LC-MS, the oil sub-group (Oil-I) was composed with the same accessions as that of marker dataset; three SO (SO-034, SO-035 and SO-039) and annual oils (YS-033, YS-143, while other SO types (SO-031, SO-032, SO-037 and SO-038) were with WO types (WO-080, WO-081, WO-084 and WO-085) and Rapid cycling (RC-144) in the other sub-group (Oil-II) (Fig. 12). But these WO types of Pakistan were grouped together close to the Asian turnips and Pak Choi in dendrogram of marker. Zhao *et al.,* (2005) profiled similar accessions using AFLPs and also found two small groups of oil types, the WO from Pakistan in one sub-group, and the annual oils- YS (originally developed in India) and SO from Bangladesh in the other sub-group. Some SO types aligned in both sub-groups in this study were together with YS in the one oil sub-group of Zhao *et al.,* (2005). This shows that YS and SO formed a strong sub-group, which might be due to the same geographic origins of these morphotypes in the history. The SO and YS might be from the north-east part of Indian sub-continent, and developed independently after the split of India and Bangladesh. But other SO types mainly from India, USA and Germany form weak sub-group with WO types in only LC-MS dataset. The LC-MS dataset contained large number of variables which cover wider range of metabolites, but the available markers may not cover the whole genome. However, we can consider WO types as a separate group from YS because it might be developed separately from the western part (Pakistan) of Indian sub-continent, Although YS accessions in this study were from USA and Germany, McGrath and Quiros (1992) reported that they were likely developed independently in India. In this core collection, 31 company lines of different morphotypes were also included, and they were grouped with older accessions of their respective morphotypes indicating similar genetic backgrounds.

The allocation of the accessions in each dendrogram was compared with the membership probabilities obtained by the software ‘STRUCTURE’ (based on allele frequency of myb and AFLP markers) and RF classification (based on the respective datasets). Although RF classification is a supervised technique which uses, in this case, prior information of the STRUCTURE classification, those comparisons help to make a visual representation of the classification of accessions by cluster analysis carried out on different datasets of this study. Marker dataset provide the chance to compare hierarchical clustering and STRUCTURE software groupings. But, in other datasets, STRUCTURE membership plot from molecular markers can be compared with hierarchical clustering and RF classification based on morphological, ^{1}H-NMR and LC-MS datasets. Those comparisons show that the clustering of accessions based on molecular markers are in good agreement with the LC-MS dataset and to lesser extent with morphological traits.

The hierarchical clustering on morphological traits was more powerful to distinguish broccoletto from European turnips (Fig. 2) although the turnips traits were not taken into account. From visual observation, broccoletto are, indeed, different for several morphological traits as compared to the turnip groups. Broccoletto has relatively shorter leaf perimeter (LP), leaf length (LL), lamina blade length (Lbl), lamina width (LW) and petiole length (PL) than turnips. In addition, less number of leaf lobes but higher leaf number and very early flowering (32 days) were observed in Broccoletto. Zhao *et al.*, (2005) also reported strong stem with short internode length and small flower heads as edible part in Broccoletto but turnips have a swollen hypocotyls and tap root with varied shape and color. But these morphotypes were clustered together in dendrograms based on the genetic and metabolic information. A previous study also showed close genetic relation of broccoletto and European turnips based on AFLP markers (Warwick *et al.,* 2008; Zhao *et al.,* 2005).

PCA and PCoA were conducted on all different datasets to visualize the locations of different morphotypes in two-dimensional plots and identify the distinguishing variables of those morphotypes. PCoA, a variant of PCA, on molecular markers discinctly separated European turnips and oil sub-group (Oil-I) from other accessions indicating their diverse genetic backgrounds. However, CC, PC, Asian turnip and oil sub-group (Oil-II) were clearly separated into sub-groups that were placed together in a big group (Fig. 8) indicating the more similar genetic background among them compared to the European turnip group and Oil-I. In the case of LC-MS, PCA distinguishes only CC and Oil-I in the first two PCs, which explain 29.76 % of the total variance present in the LC-MS dataset (Fig.13). The results of cluster analysis and PCA/PCoA imply that molecular markers and LC-MS analysis were more powerful to discover the natural groupings of *B. rapa* than morphological and NMR datasets. In case of NMR dataset, PCA was able to group only CC although they were not distinctly isolated from other morphotypes (Fig.11), but not any grouping in cluster analysis (Fig. 9), which indicates ^{1}H-NMR did not distinguish groups of accessions according to morphotyes or geographic origins in both cluster analysis and PCA.

In addition, the correlation of the dissimilarity matrices of different datasets that used for clustering and PCA/PCoA were compared using Mantel test, where LC-MS and marker had highly significant correlation (p-value = 0.0001), followed by their combinations with morphology. But NMR had non-significant correlation with other datasets (Table 6). The results of Mantel test also supported a good correspondence of marker and LC-MS datasets on the groupings of accessions, followed by morphological observations.

Molecular markers provide the more reliable information at the gene level for the study of genetic relations. Similarly, LC-MS has high sensitivity and specificity in the detection of the metabolites even if present in low concentration, such as secondary metabolites on the basis of molecular masses of chemicals. Therefore, more distinct groups were discovered according to morphotypes or the geographic locations. But ^{1}H-NMR is less sensitive technique than LC-MS which detect only abundant and proton bearing compounds, such as organic acids. Besides, one dimensional

^{1}H-NMR generates the overlapping NMR bins due to large numbers of contributing compounds and multiple signals (Colquhoun, 2007; Widarto *et al.,* 2006), that might hinder the separation of accessions in this study. PCA on morphological traits distinguished CC, European turnip, PC and mizuna. Interestingly, a group of mizuna morphotypes was discovered only through morphological traits, such as high petal index (high petal length and low width), elongated petal shape, leaf edge shape and relatively higher number of leaves when the first flower opens (Fig. 6). Among two older mizuna accessions (MIZ-019 and MIZ-128) and three company lines (K-1, K-2 and K-3), only MIZ-019 and K-2 formed one sub-group in both clustering and PCA, and MIZ-128 and K-1 were close together only in the PCA (Fig. 5). Zhao *et al.*, (2005) also observed the differences between two old mizuna accessions, which was also supported by their growing regions, where MIZ-019 was from the Netherlands, and MIZ-128 from Japan. These two small sub-groups of mizuna suggest that company lines K-1 and K-2 might be closely related with MIZ-128 and MIZ-019 respectively or developed with the breeding objectives of replacing the older accessions. European turnips were different because of very late flowering, leaf morphology (longer perimeter, higher area, longer length and very long petiole length) and plant architecture (less number leaf and plant branch**)** while Pak Choi, CC and mizuna had distinct flower morphology (Fig. 6). The morphological traits, such as leaf color (LC), leaf area (LA), and presence of petiole (PP) were accountable in distinguishing CC from Pak Choi. CC had yellowish leaf color, higher leaf area and absent petiole in comparison with Pak Choi.

PCA/PCoA form new latent variables (called PCs/PCo) in the linear combination of original variables, such as molecular markers, NMR bins and LC-MS peaks. In this study, PCA on metabolic peaks (NMR and LCMS) and PCoA on molecular markers could not identify the variables accountable for the different morphotypes. In stead, smear of variables were observed in variable plots of PCA/PCoA because large numbers of variables had uniform loadings with small difference over the variables. This analysis suggests that the individual molecular marker/LC-MS peak/NMR bin had non-linear relations with small and interaction effects. Thus, this study was followed by a more powerful variables selection technique, Random Forest, was used to identify the distinguishing variables.

3.4 Conclusions

B. rapa core collection has diverse morphological, metabolic and genetic relations, which corresponds with the geographic origins and also with morphotypes. LC-MS peaks were capable to distinguish two small sub-groups of oil, which also extends to some extent in case of molecular markers. Moreover, morphological traits are also found important in isolating some of the morphotypes, such as mizuna, broccoletto. Different morphological traits that distinguish morphotypes were identified. Among all datasets, LC-MS peaks and molecular marker information were in more congruence in grouping the accessions followed by morphological observations. But ^{1}H-NMR spectroscopy was not effective to distinguish the accessions in accordance with morphotypes as well as geographic origins.

Chapter 4: Identification of unique molecular markers and metabolic peaks in distinguishing different morphotypes of *B. rapa* core collection

4.1 Materials and Methods

4.1.1 Random Forest

Random forest (RF) is an improved version of the Classification and Regression Trees (CART) method. However, CART builds only one classification tree but random forests build a collection of multiple classification (and/or regression trees), and are, therefore, named “forests”. RF uses both a boosting and a bagging strategy, where boosting reduces the both the variance and the bias of the classification and bagging reduces the variance (Gislason *et al.,* 2006). Important features of RF are: good prediction accuracy, relatively robust to outliers and noise, it returns useful internal estimates of prediction error and variable importance (Breiman, 2001), it can handle the “small n large p” problem of high dimensional data and also complex interactions of variables and situations with highly correlated variables (Strobl *et al.,* 2008). RF can have overfitting problem (Segal, 2004) and can estimate the variables importance for the both classification as well as regression (Breiman, 2001; Gislason *et al.,* 2006; Truong *et al.,* 2004).

RF does not allow any kinds of missing values; data imputation was done by considering 20 neighbouring observations using Euclidean metric in a K nearest neighbour algorithm in morphology and marker datasets. An “impute” package in R-software (Hastie *et al.,* 2008) was used by setting the parameters at maximum 20 % missing in row and 50 % column.

4.1.1.1 RF Classification

RF makes an ensemble of trees and may have many forests to produce an unbiased estimate of the classification error (Pang *et al.,* 2006). All trees are more or less independent of each other because each tree is built by taking a bootstrap sample (random sampling of two-third of the samples with replacement) of the original data (called the training set), and a random subset of the variables (at each split as a candidate set of variables). Thus, the computational load is reduced, and it can handle the high dimensional data (n <<< p) (Gislason *et al.,* 2006). Each tree is fully grown (unpruned) to obtain low-bias, high variance and low correlated tree. At end, RF averages over all trees resulting low-bias, low variance and low correlated trees giving good prediction (Svetnik *et al.,* 2003). On an average, about one-third of the original data are not sampled in training set, called out-of-bag (OOB) samples which are used as the test set. The test set data (OOB samples) are then run down the tree to be classified by the random forest, and the classification for the i^{th} tree is predicted by comparing the estimate of the OOB samples to their real class. Based on the majority of votes over the trees of the forest, the OOB samples are assigned to a particular class. The classification error can then simply be estimated by comparing the estimated class with their true class label. Thus, there is no need for an extra cross-validation (Gislason *et al.,* 2006; Pang *et al.,* 2006; Svetnik *et al.,* 2003). For the variable selection, RF starts with a forest using all variables in a random subset of variables at each split node and then builds new forests at each step while discarding the variables with the smallest importance (also called backward elimination). Thus, the OOB error rate is considered as a biased estimator to assess the overall prediction error rate of the algorithm (Diaz-Uriarte and Alvarez de Andres, 2006). The 0.632 + bootstrap method, which use a weighted average of the re-substitution error (the error when a classifier is applied to the training data) with OOB error gives the better prediction error rate of the classification (Diaz-Uriarte and Alvarez de Andres, 2006). The smaller the OOB error rate, the better is classification of the samples(Pang *et al.,* 2006).

RF randomly permutes the values of predictor variables of the OOB samples and then run down the tree to get new classification. The prediction of classification of OOB samples in terms of correct classifications of OOB samples is compared before and after permuting the predictor variables and averaged over all trees. If the prediction accuracy substantially decreases after the permutation indicate the association of those variables in the classification of samples (Gislason *et al.,* 2006; Pang *et al.,* 2006; Svetnik *et al.,* 2003). RF measures the importance of each variable by mean decrease in the prediction accuracy and mean decrease in Gini index. Gini index measure the impurity of split selection criteria in machine learning, such as RF, CART, however, it is biased in favor of variables having higher numbers of categories and continuous variables, and offering more splits in classification (Strobl *et al.,* 2008; Strobl *et al.,* 2007).

A web-based package of random forest classification, “GeneSrF” was used for classification purposes (Diaz-Uriarte, 2007). GeneSrF can handle a large dataset and also deal with the problem of multiplicity (lack of stable selection of variables upon the repetitive analyses). Diaz-Uriarte and Alvarez de Andres (2006) suggested for considering the biological relevance of the selected variables to address the stability problem. In this study, box plots on selected variables were drawn to compare the relevance of selected variables in classification of the accessions. GeneSrf uses the 0.632+ bootstrap method using 200 bootstrap samples to produce an unbiased estimate of the prediction error. For this analysis, all the default settings were retained because of no significant changes on its performances over wide range of settings except in extreme cases (Svetnik *et al.,* 2003). The default setting on parameter “mtry” (the number of random variables selected at each node) for classification is the square root of the total number of variables and number of samples in each end node (nodesize) is at one. The minimum node size determines the minimum size of nodes below which no further split will be attempted.

4.2 Results

4.2.1 Marker

In random forest classification of four STRUCTURE classes, 38 markers were selected as distinguishing markers for those four classes. The markers were repeatedly observed on different pair-wise comparisons of different groups present in this core collection. Among 359 AFLP and MYB markers, 10 markers were found important in distinguishing class 1 versus class 2, 8 markers for class 1 versus 3, 24 markers for class 1 versus 4, 3 markers for class 2 versus 3, 24 markers for class 2 versus 4, 2 markers for class 3 versus 4, 4 markers for Chinese cabbage (CC) versus Pak choi (PC) and 4 markers for European (EU) versus Asian turnips (Table 8). The importance spectrum plot also showed the clear differences between the markers importance before permutation (original data) and after permutation indicating the higher importance of marker in separating the four classes (Fig. 14; Appendix). Form classes comparisons of class 1 versus 2, class 1 versus 4, and class 2 versus 4, 9 markers (out of 10), 23 markers (out of 24) and 21 markers (out of 24), respectively, were found common with the 38 markers selected on four STURCURE classes comparisons. This indicates that class 1, 2 and 4 were the most diverse groups among all groups (classes). However, 2 markers (out of 2), 4 markers (out of 4) and 4 markers (out of 4) of comparisons of class 3 versus 4, CC versus PC, and EU versus Asian turnip, respectively, were also in congruence with markers found important in all 4 classes classification, even though they had very few numbers of distinguishing markers selected. In contrast, all three and 7 (out of 8) markers importantly selected in class comparisons of 2 versus 3, and 1 versus 3, respectively, were not observed in the comparisons of 4 classes. Almost all different combinations of classes (groups) comparisons had low classification prediction error, however, EU versus Asian turnip comparison had highest error (0.148) followed by CC versus PC (0.085) and all 4 classes comparisons (0.083) (Table), and Fig. 15 also showed the high-misclassifications accordingly. All possible comparisons among the different classes (groups), altogether 59 markers were uniquely listed, in which 22 markers have known map position based on DH mapping population of Yellow Sarson 143 and Pak choi 175 bi-parental cross (Table 9).

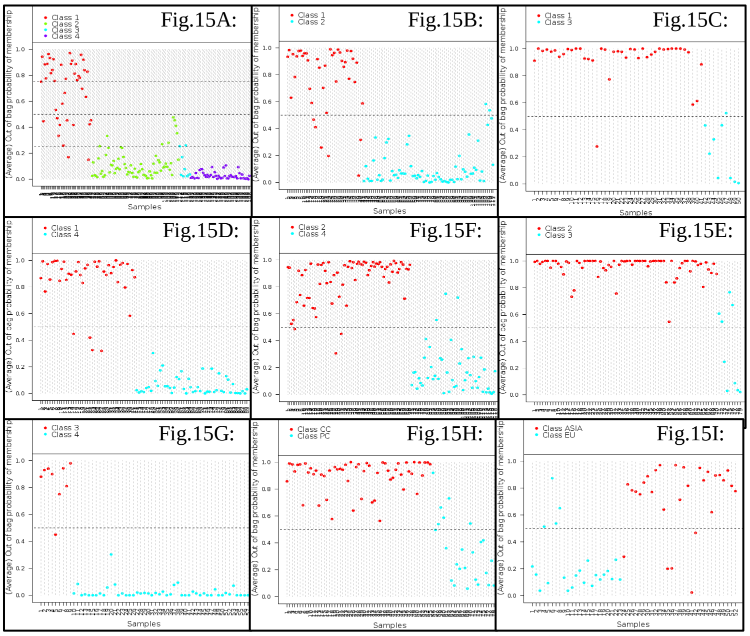


Fig. 15: Class membership probability of accessions (RF classification based on markers) of being in: **A-** Class 1 in four STRUCTURE classes comparison; **B-** Class 1 with respect to class 2; **C-** Class 1 with respect to class 3; **D-** Class 1 with respect class 4; **E-** class 2 with respect to class 3; **F-** class 2 with respect to class 4; **G-**class 3 with respect to class 4; **H-** CC with respect to PC; **I-** Asian turnip with respect to

Table 8: RF classification errors obtained in pair-wise comparisons in Marker dataset

Comparisons	Class prediction error*	Leave-one-out bootstrap error
All four STRUCTURE classes	0.083	0.122
Class 1 vs. 2	0.066	0.093
Class 1 vs. 3	0.055	0.079
Class 1 vs. 4	0.028	0.043
Class 2 vs. 3	0.037	0.053
Class 2 vs. 4	0.043	0.065
Class 3 vs. 4	0.025	0.038
CC vs. PC	0.085	0.107
EU vs. Asian turnip	0.148	0.179

*- Bootstrap (0.632+) estimate of prediction error.

4.2.2 NMR

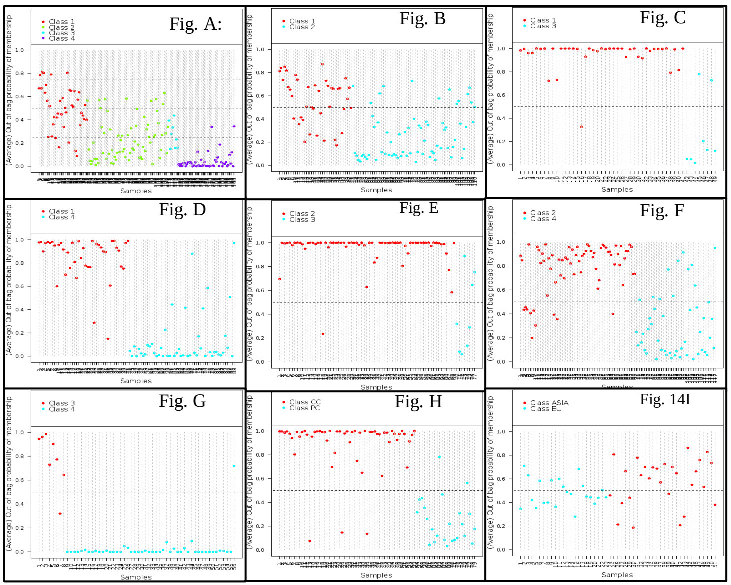


Fig. 17: Class membership probability of accessions (RF classification based on NMR) of being in: **A**-Class 1 in four STRUCTURE classes comparison; **B**-Class 1 with respect to class 2; **C**-Class 1 with respect to class 3; **D**- Class 1 with respect class 4; **E**- class 2 with respect to class 3; **F**-class 2 with respect to class 4; **G**-class 3 with respect to class 4; **H**- CC with respect to PC; **I**-Asian turnip with respect to European Turnip.

RF classification conducted in NMR dataset selected 32 NMR bins accountable for the distinction of four STRUCTURE classes with prediction error 0.27. In classification of accessions into pair-wise classes, 14 bins were identified as important bins to distinguish class 1 versus 2, 2 bins for class 1 versus 3, 32 bins for class 1 versus 4, 3 bins for class 2 versus 3, 4 bins for class 2 versus 4, 2 bins for class 3 versus 4, 2 bins for CC versus PC and 9 bins for and EU turnip versus Asian turnip with different classification errors for class prediction of accessions and variable importance (Table 10 and 11). The importance spectrum plot also showed the clear differences between the importance of NMR bins before permutation (original data) and after permutation indicating the higher importance of bins in separating the four classes (Fig. 16; Appendix). In comparisons of selected bins, 12 bins (out of 14) of class 1 and 2, 24 bins (out of 32) of class 1 versus 4, 4 bins (out of 4) of class 2 versus 4, 2 bins (out of 2) of CC versus PC, and only 3 bins (out of 9) of EU versus Asian turnips were in common with the bins selected on all 4 classes classification (Table 11; Appendix). Based on selected bins, class 1, class 2 and class 4 were appeared as most distinguishing classes among all groups of this core collection, where high numbers of distinguishing bins selected in separate pair-wise classifications were correspondence with bins selected in all 4 classes comparison. Among all these comparisons, the errors for class prediction and variable importance (leave-one-out bootstrap error) were higher in the group classifications of Europe versus Asian turnips, four STRUCTURE classes, class 1 versus group 2, and class 2 versus 4 accordingly. Moreover, Fig. 17 showed the membership probability of each accession of being in different classes during the RF classifications in different groups' combinations.

Table 10: RF classification errors obtained in pair-wise class comparisons in NMR dataset

Comparisons	Class prediction error*	Leave-one-out bootstrap error
All four STRUCTURE classes	0.27	0.35
Class 1 vs. 2	0.22	0.275
Class 1 vs. 3	0.055	0.077
Class 1 vs. 4	0.066	0.097
Class 2 vs. 3	0.054	0.073
Class 2 vs. 4	0.147	0.198
Class 3 vs. 4	0.027	0.041
CC vs. PC	0.0791	0.113
EU vs. Asian turnip	0.393	0.425

*-Bootstrap (0.632 +) estimate of prediction error.

4.2.3 LC-MS

RF classification used for variable selection to distinguish 4 STRUCTURE classes of classification selected 64 LC-MS peaks (noted by centrotype_mass_scan) among 5546 peaks (variables) with class prediction error (0.166) and variable importance error (0.23) (Table 12). Those 64 selected peaks represented 31 different centrotypes, which indicates the possibly that these signals represent the same compounds or compounds from the same chemical groups (isotopes of a chemical derivative). Out of 64 selected peaks, only 21 peaks could not annotate the chemical name, where 16 peaks were from 14 centrotypes, and 5 peaks didn't have any centrotype (Table 13).

Importance spectrum plot of top 200 peaks showed the clear differences between the importance of peaks for the original data and that of random permutation (Fig. 18; Appendix). This plot indicates that top 200 peaks did good job in classification of accessions into four classes, where higher the differences, the higher the importance of peaks in differentiating four STRUCTURE classes. The clustering based on these selected 64 LC-MS peaks gave similar distribution of accessions with that of dendrograms based on whole 5546 peaks (Figure not shown).

In pair-wise comparisons, 3 peaks for class 1 v. 2, 5 for class 1 vs. 3, 2 for class 1 vs. 4, 3 for class 2 vs. 3, 41 for class 2 vs. 4, 2 for class 3 vs. 4, 14 for CC vs. PC and 26 for EU vs. Asian turnip were found important in assigning the accessions into different classes in the RF classification (Table 13) with different classification and variable selection error (Table 12). All the peaks (3 peaks out of 3) of RF classification of class 1 vs. 2, 2 peaks (out of 2) of class 1 vs. 4, 27 peaks (out of 41) of class 2 vs. 4, 3 peaks (out of 26) and 5 peaks (out of 14) were also observed in RF selected peaks in the classification of all the accessions into 4 STRUCTURE classes, however, none of the peaks of classifications of class 1 vs. 3 (5 peaks), and class 2 vs. 3 (3 peaks) were common.

Table 12: RF classification errors obtained in pair-wise class comparisons in LC-MS dataset

Comparisons	Class prediction error*	Leave-one-out bootstrap error
All four STRUCTURE classes	0.166	0.23
Class 1 vs. 2	0.117	0.152
Class 1 vs. 3	0.055	0.078
Class 1 vs. 4	0.053	0.074
Class 2 vs. 3	0.051	0.070
Class 2 vs. 4	0.117	0.162
Class 3 vs. 4	0.028	0.042
CC vs. PC	0.058	0.085
EU vs. Asian turnip	0.221	0.277

*- Bootstrap (0.632 +) estimate of prediction error.

Classification error for class prediction and variable importance for classification were highest in the classification of accessions into EU vs. Asian turnip, followed by the classification into all four STRUCTURE classes, class 2 vs. 4, and class 1 vs. 2 successively which was also visualized in figure (Fig. 19).

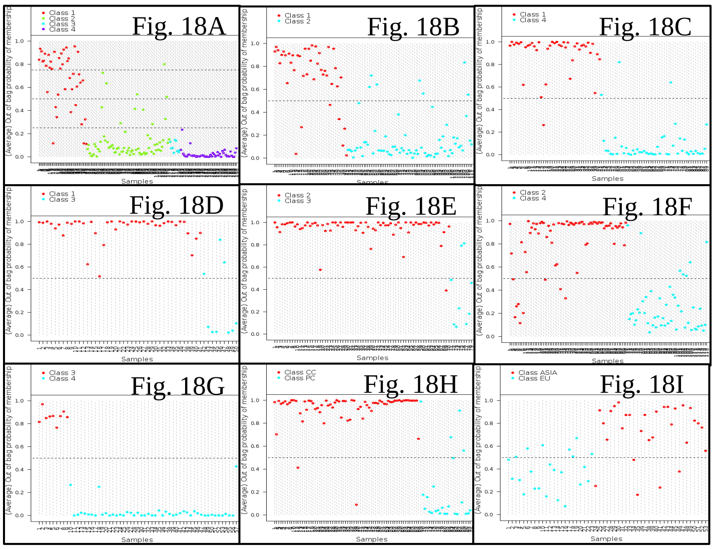


Fig. 19: Class membership probability of accessions (RF classification based on LCMS) of being in: **A**-Class 1 in four STRUCTURE classes comparison; **B**-Class 1 with respect to class 2; **C**-Class 1 with respect to class 3; **D**- Class 1 with respect class 4; **E**- class 2 with respect to class 3; **F**-class 2 with respect to class 4; **G**-class 3 with respect to class 4; **H**- CC with respect to PC; **I**-Asian turnip with respect to European Turnip.

The selected peaks in four classes comparison were visualized in box plot to see the level of metabolites present in different STRUCTURE classes (Fig. 20; Appendix). Box plots showed that Centrotypes: 4108 (Isopropyl glucosinolate), 4815 (Methylpropyl glucosinolate) , 5400 and 4867 (glucopyranoside derivatives), 5441, 8845 and 9223, and peaks: A194_501_706, A190_388_1036 and A86_484_2174 were present only in class 1 (European Turnip) whereas, centrotypes: 4932, 4990, 5248 (Kaempferol caffeoyltetra glucoside), some peaks; 5236_708_784, 5236_707_785, 5236_1191_788 and 5236_1192_787 belonging to a centrotype 5236 (Chlorogenic acid), and centrotype 9722 were only in class 4 (Chinese Cabbage). Peak A62_449_632, and centrotype 5028 (Quercetin diglucoside), 7673 and 5600 were dominant in class 3 (Oils) and 4 (Chinese cabbage) but some of them were also in classes 1 and 2 (Pak Choi and Asian Turnip).

Two small groups of oils type observed in both cluster analysis and PCA, where SO-034, SO-035, SO-039, RP-18, YS-033 and YS-143 were in one group with high level of 6882_666_1076 and 5600_357_849 peaks, and low level of 8515_649_1410 and 8722_565_1442 peaks than another oil groups consisting of SO-037, SO-038 and RC-144 (Fig. 21; Appendix). RF classification prediction error was 0.03915 in comparison of two small groups of oil type.

4.3 Discussion

Accessions were classified by using RF classification approach, which can handle larger numbers of variables (marker loci, NMR bins and LC-MS peaks) than that of accessions. The previous studies reported its ability to explore complex interactions and non-linear relations of predictor variables, its internal cross validation properties; few parameters need to be adjusted and better performances on multi-class situations (Cutler *et al.,* 2007; Barrett and Cairns, 2008). In addition, RF classification approach was used for the estimation of variable importance and error rates in this study.

Molecular marker:

In RF classification based on molecular markers, CC (falls in STRUCTURE class 4) was the most distinct morphotypes from European turnip, Asian turnip and Pak Choi (PC) because higher numbers of markers were needed for the distinguishing those morphotypes in pair wise comparisons (Table 9). European- and Asian- turnips, and CC and PC were close together because they showed differences in fewer (4) markers. In a pair-wise comparison of Class 2 (composed of Asian turnip and PC) with class 4 (CC), high numbers of markers (24) were differ in their presence or absence between the classes, however, few markers (4) had distinguished CC and PC morphotypes. This suggests that CC is close to PC in their genetic background, but distinct from Asian turnips.

European and Asian turnips were differing only in four markers suggesting a very similar genetic background despite being diverse in their geographic origins. Hence, those two geographically different morphotypes might have been differentiated in a later stage of the domestication process. Based on this study, oil morphotypes (represented by class 3) were more close to class 2 followed by classes 4 and 1, which indicated that oils are close to Asian turnips and PC in their genetic backgrounds. However, the identification of map positions of the molecular markers is essential to know whether these markers are in the same position or not. Several genetic studies show that *Brassica* accessions of different morphotypes are more related based on the geographic origin despite being in different morphotypes (McGrath and Quiros, 1992; Warwick *et al.,* 2008; Zhao *et al.,* 2005). This study is also in agreement with those studies in showing that accessions originating from the same geographic origins are closer than morphotypes from more distant geographic origins.

Nuclear Magnetic Resonance (NMR)

Based on NMR metabolic profiles, classification of accessions into European- and Asian- turnip had higher classification error for the prediction of classes as well as variable importance (Table 10) followed by classification of four classes, class 1 versus class2, and class 2 versus class 4 in a decreasing order. Class 4 (dominated by CC) was most distinct from class 1 (dominated by European turnip), where 32 NMR bins had higher importance in differentiating those classes (Fig. 17). Fourteen (14) NMR bins were able to distinguish class 1 versus 2 but only very few NMR bins were found important in distinguishing other pair-wise classifications. Although 9 NMR bins were able to distinguish European- and Asian- originated turnips, the high numbers of miss-classification of accessions were observed in class membership plot (Fig. 17-I) showing no prominent differences in NMR-based metabolome for these two turnips. The results signified that European turnips differ largely in metabolic composition from CC and Pak Choi. However, class 4 (CC) does not show a distinct metabolic composition than that of Asian turnips and Pak Choi (class 2) which might be because of close relation of CC with PC. And CC and PC differ only in two NMR bins (possibly only two metabolites). Class membership probability plots (Fig. 17) also showed a large numbers of outlier accessions in pre-defined classes suggesting either that NMR bins are not able to distinguish the accessions.

A 2-dimensional NMR analysis might be more helpful to identify possibly distinguishing metabolites, since this method improves the resolution by reducing the overlapping signals of NMR bins (Nilsson *et al.*, 2004). In another study of metabolic differentiation of *B. rapa* where response upon the attack of herbivores was investigated, Widarto *et al.,* (2006) suggested 1-dimensional NMR was not effective to identify discriminating metabolites because of the overlapping of NMR signals and spectral complexity. The relatively high classification error might be because of overlapping of chemical compounds over the NMR bins. In contrast with the molecular marker and LC-MS datasets, cluster analysis and PCA based on the NMR dataset in this study were also unable to discover the distinct groups of morphotypes.

LC-MS

Similar to NMR metabolite profiling, the classification error was higher in the classification of accessions into European- versus Asian- originated turnips, followed by classification of the four STRUCTURE classes, class 1 versus 2, and class 2 versus class 4 (Table 12). But in contrast to NMR and marker-based classifications, high numbers of LC-MS peaks were involved in distinguishing class 2 versus class 4 followed by EU- versus Asian- turnips and CC versus PC (Table 13). Class 2 is one of the most heterogeneous groups composed of Asian turnips and PC, Asian turnip was more heterogeneous than others in cluster analysis. This lead to select higher number of peaks in class comparison including differential peaks intensities within Asian turnips morphotype. RF classification resulted the slightly different numbers of variables as important variables during the repeated analyses. Diaz-Uriarte and Alvarez de Andres (2006) also noted the stability problem for variable selection and suggested to consider the biological relevance of the selected variables. Thus, we suggest to take relatively larger numbers of selected variables, and to observe the biological relevances. In this study, biological differences were visualized via. box plots graph (Fig. 20). The cluster analysis based on those selected 64 peaks also showed more similar distributions of accessions to clustering based on whole 5546 peaks (figure not shown). This indicates RF classification had good performance in selecting variables.

The quantitative and qualitative differences of metabolic peaks, that were obtained RF classification, were observed among the four classes. The qualitative differences were observed in two annotated compunds: a glucopyranoside-derivative (4,7-Megastigmadiene-3,9-diol, 3-Ketone, 9-O-[α-L-arabinopyranosyl-(1->6)-β-D-glucopyranoside]) was present only in class 1 (European turnip) and chlorogenic acid and some other metabolites (not yet annotated) were specific to CC morphotypes (class 4). Also other morphotypes, such as Asian turnip, PC, oils have unique non-annotated metabolites (many mass peaks do not have an annotation). Besides, other annotated and non-annotated peaks had quantitative differences among the four classes (Fig. 20). RF classification was reported as an effective method for the classification of samples and detection of unique peaks in metabolomic and proteomic high-dimensional datasets since it can cope with multicollinearity situations, allows the assessment of complex interactions of peaks, it does supposedly not suffer from overfitting problems and allows estimation of the importance score of peaks (Beckmann *et al.,* 2007; Barrett and Cairns, 2008; Enot *et al.,* 2006). However, Segal (2004) reported the problem of overfitting in a study based on data simulation because of unpruned tree in random forest.

In cluster analysis and PCA, oils were separated into two possible sub-groups, each with a small number of accessions. The separation indicates the quantitative variations oil-metabolites between these subgroups. RF classification was used to identify the distinguishing LC-MS peaks for these sub-groups. Among six centrotypes, three centrotypes were present at relatively higher level only in Oil-I sub-group (SO, YS) but the other three centrotypes were at higher level in Oil-II sub-group (SO, WO, RC) (Fig. 21). This result signifies the presence of two oil sub-groups in *B. rapa* species. Thus, RF was able to classify the different morphotypes with the identification of discrimination LC-MS peaks**.**

4.4 Conclusions

RF classification had good performances in classifying the accessions and also identifying the distinguishing molecular markers, NMR bins and LC-MS peaks in all class comparisons as well as pair-wise morphotypes classifications. European- and Asian turnips were less distinguishable as compared to other pair-wise classifications in all datasets. Similarly, CC and PC were closely related on their genetic backgrounds as well as metabolic contents but CC and European turnips were the most distinct apart. However, the annotation of LC-MS peaks and identification of map positions of marker are important to confirm the results.

In addition, RF classification made selections of molecular markers, NMR bins and LC-MS peaks that were found important in classifying the accessions into different morphotypes. The qualitative and quantitative differences of metabolites were observed among the morphotypes. The performance of RF classification was better in molecular markers and LC-MS than that of NMR dataset.

Chapter 5: Identification of LC-MS peaks and molecular markers associated with leaf color of *B. rapa*

5.1 Materials and Methods

5.1.1 Correction of SPAD traits for the population structure and correlation analysis

Plant genetic resources, even a single species, at a population level are genetically and phenotypically diverse due to varied geography, natural and/or artificial selection, mutation, migration, natural recombination and several spontaneous or induced factors. The *B. rapa* core collection used in this study has different sub-populations due to their independent origin in Europe and Asia, and long-history of domestication and breeding. Zhao (2007) also found the presence of four groups (classes) of sub-population in this core collection in his Ph. D. studies. The influence of population structure on correlations is very large and therefore it is necessary to control for false positive associations in association mapping (Balding, 2006; van Berloo *et al.,* 2008; Yu *et al.,* 2006). In this core collection, four distinct classes, identified using the STRUCTURE software (Pritchard *et al.,* 2000), were considered for mean comparisons to see the influence of population structure on SPAD via. Analysis of Variance (ANOVA). The residuals obtained from the ANOVA, which was assumed to be free from the influence of population structure, were stored for the association studies. The model used for ANOVA was $y_{ij} = \mu + G_i + e_{ij}$, where, y_{ij} is the SPAD observation in the ith STRUCTURE group and jth observation, μ for common mean, G_i for the effect of the ith STRUCTURE group (i= 1,..,4) and e_{ij} for random error in ith group and jth observation.

Pearson correlations among SPAD traits of all three conditions (vernalized condition 2007, and non-vernalized conditions 2006 and 2007) were tested to have an idea of how the same traits in different conditions compare to each other. The significance of Pearson correlations was examined using a simple student’s t-test procedure. All the ANOVA and correlation tests were done using R-statistical software.

5.1.2 Random Forest Regression

LC-MS measurement and marker genotyping datasets have a very high number of variables (5546 peaks in LC-MS, 359 markers) with respect to the number of samples (168 accessions). An ordinary multiple regression approach can not handle the problem of “small n large p” as well as multi-collinearity situation. Therefore, in this study, an RF regression was used for SPAD traits (a quantitative measurement of leaf color) of all three conditions to find the associated metabolic peaks and molecular markers, which are robust to handle those problems.

The main goal of a RF regression approach is to find a set of variables that best predicts the variation present in a continuous trait of interest. The percentage of variance explained by 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). This measure shows the performance of set of variables by explaining the variation present in trait of interest. RF regression yields two important measures; mean decrease in accuracy and mean decrease in MSE for all the variables. The mean decrease in accuracy is measure by comparing the true class label of the sample with the plurality of OOB class votes after the permutation of random subset of variables (Breiman, 2001). The higher the amount of decrease in importance measures is, the higher is the importance of those variables.

RF regressions on residuals of SPAD traits were conducted by using the “randomForest” package of the R-software (Breiman *et al.,* 2008). RF regression approach is suitable for association studies especially in case of large numbers of predictor variables, where interactions of predictors are present and no need to specify model (Lunetta *et al.,* 2004). The number of trees (ntree) was adjusted at 5000 because of computer memory limit; however, all other parameters were adjusted in the default settings because of no significant changes on its performances over wide range of settings except in extreme cases (Svetnik *et al.,* 2003). The default settings on parameter “mtry” (the number of random variables selected at each node) is the one-third of the total variables (p/3) and the number of samples in each end node (nodesize) is at 5 samples. The node size determines the minimum size of nodes below which no further split will be attempted.

5.1.3 Unified mixed model

An association mapping study was also conducted for different SPAD traits separately in a unified mixed model approach using the software TASSEL (Zhang *et al.,* 2006). This mixed-model approach takes into account multiple levels of relatedness and has good control of type I and type II error rates over other methods (Yu *et al.,* 2006). A population matrix (Q-matrix) and a kinship matrix (K-matrix) were used in the mixed model (Q + K method), hence, named a unified mixed model. Marker data was used to calculate the Q-matrix in a Bayesian approach via. STRUCTURE software (Pritchard *et al.,* 2000) and the K-matrix in the TASSEL software. Q-matrix provides the genetic relation between the different groups of a population while K-matrix measures the relatedness of accessions with a group. The statistical model of the Q + K method employed in TASSEL is:

$y = X\beta + Zu + e$

where **y** is the trait of interest (here: SPAD); β is an unknown vector containing fixed effects, including genetic markers and population structure (Q); u is an unknown vector of random additive genetic effects from multiple background QTL for individuals/lines; X and Z are the known design matrices containing the marker information; and **e** is the unobserved vector of random residual. The u and e vectors are assumed to be normally distributed with null mean and variance of $\mathrm{Var}\begin{pmatrix} \mathbf{u} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{pmatrix}$ where $G = \sigma^2_a K$ with σ^2_a as the additive genetic variance and K is the kinship matrix. Homogeneous variance was assumed for residual errors, which means $R=I\sigma^2_e$, where σ^2_e is the residual variance. And $\sigma^2_{a\ \mathrm{and}}$ σ^2_e are calculated by Restricted Maximum likelihood (REML) approach (Bradbury *et al.,* 2007).

5.2 Result

Results of association studies of LC-MS peaks and molecular markers with SPAD traits of vernalized conditions 2007, non-vernalized condition 2006 as well as 2007 were presented in the subsequent sections.

5.2.1 Summary statistics of SPAD traits:

5.2.1.1 Mean comparison of four STRUCTURE classes:

The means of SPAD values from accessions from the four STRUCTURE classes were compared for vernalized SPAD of 2007, non-vernalized SPAD of both 2006 and 2007 in one-way ANOVA test. SPAD of vernalized (p-value=0.0003), SPAD non-vernalized 2006 (p-value=0.0007) and SPAD non-vernalized 2007 (p-value=0) were highly significant, however class 1 in for data vernalized SPAD 2007, class 2 for data non-vernalized SPAD 2006, and class 1 and 2 of non-vernalized 2007 were only significantly different from the other groups in that experiment at LSD (at 5 % level) in multiple comparisons (Table 14). In relation to SPAD trait, two clear groups of accessions were found where class 1 had higher mean SPAD trait value indicating dark leaf color in the year 2007 but class 2 was darker in year 2006. Albeit this analysis might be biased because of varied number of accessions among the classes, the general ideas can be drawn regarding the effects of different classes of this core collection on association studies.

Table 14: Summary statistics of SPAD traits across the STRUCTURE classes

Growing conditions	Statistical summary	STRUCTURE classes				$LSD_{at\ 0.05}$
		Class 1	Class2	Class3	Class4	
Vernalized 2007	Mean	36.10^a	33.62^b	31.06^b	30.67^b	3.346
	# of samples	38	69	8	48	-
	Std. error	0.941	0.699	2.052	0.838	-
Non-vernalized 2006	Mean	33.93^b	35.74^a	30.78^b	31.66^b	3.195
	# of samples	38	69	8	48	-
	Std. error	0.899	0.667	1.960	0.800	-
Non-vernalized 2007	Mean	34.461^a	37.56^a	29.867^b	28.186^b	3.795
	# of samples	37	68	8	48	-
	Std. error	1.079	0.796	2.320	0.947	-

^{a} indicates the significant mean differences from ^{b}, and vice versa in ^{b}

5.2.1.2 Correlation of SPAD traits grown in different conditions and years:

Pearson correlations of SPAD vernalized 2007 with SPAD 2006 and SPAD 2007 of non-vernalized conditions was 0.45 and 0.49 respectively. In spite of being different years, correlation between SPAD 2006 and SPAD 2007 within the non-vernalized condition was higher (r = 0.66) than with vernalized condition. This might be the reasons for getting different associated markers as well as LC-MS peaks in relation to SPAD traits measured even in the same years.

5.2.2 Association study of LC-MS peaks with SPAD traits

In RF regression of LC-MS peaks on the residuals of SPAD traits of three different growing conditions, the top 30 peaks were selected from 5546 peaks based on the contribution of each peaks in decreasing error mean sum of square (%IncMSE). Among the top 30 peaks selected for SPAD of vernalized condition of 2007, 24 peaks were within 22 centrotypes while 6 peaks had unknown centrotype (alphabetic code; for example A769, A61 and so on) (Table. 15 and Fig. 22). The variance explained was very low (R^2=2.28 %).

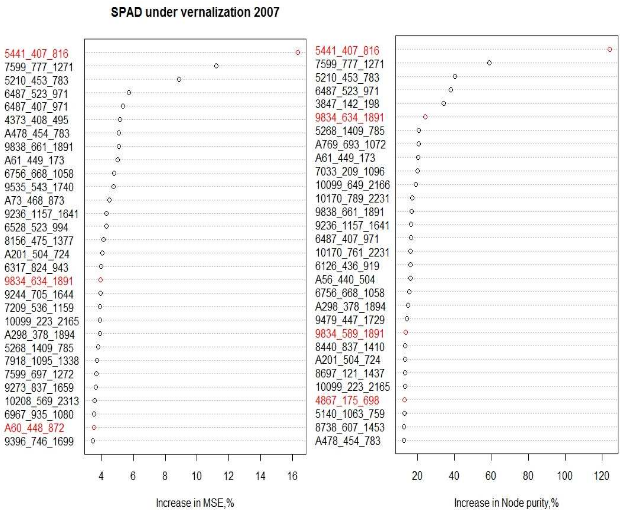


Fig. 22: RF regression selected the top 30 LC-MS peaks for SPAD residuals on vernalized condition in 2007 year. Red color indicates common peaks with non-vernalized condition.

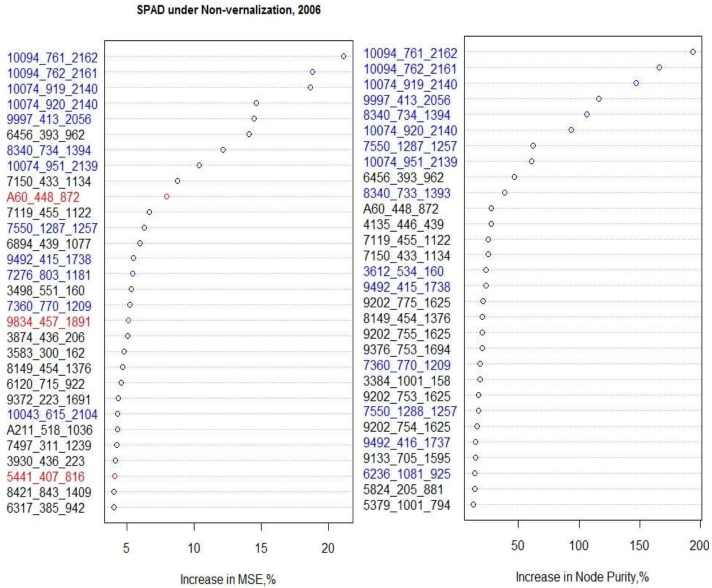


Fig. 23: RF regression selected the top 30 LC-MS peaks for SPAD residuals on non-vernalized condition in 2006. Red color indicates common peaks common with vernalized condition and blue color indicate common within non-vernalized condition.

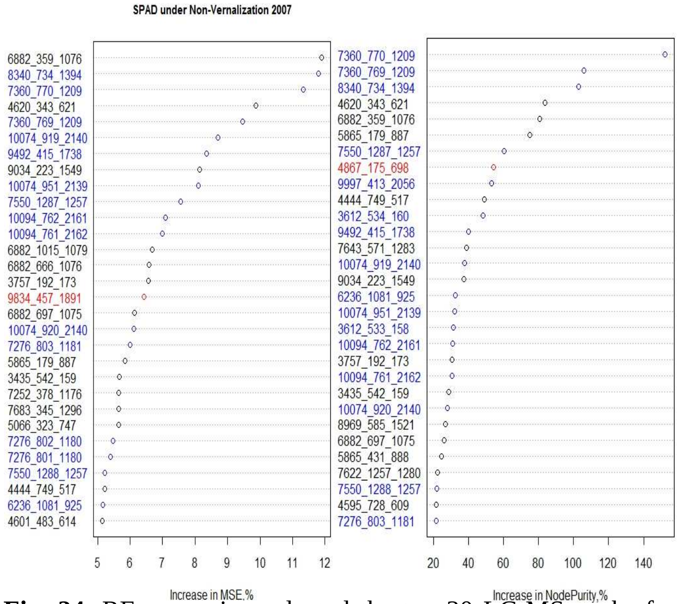


Fig. 24: RF regression selected the top 30 LC-MS peaks for SPAD residuals on non-vernalized condition in 2007. Red color indicates common peaks common with vernalized condition and blue color indicate common within non-vernalized condition.

Similarly, 25 known centrotypes and 2 peaks with unknown centrotype in the top 30 selected peaks found important in explaining the variances of SPAD trait under the non-vernalized condition (R^2=22.39 %) of 2006 (Fig. 23). In case of non-vernalized condition of 2007, the top 30 peaks consisting of 20 centrotypes were importantly associated with SPAD trait (R^2=21.44 %) **(**Fig. 24**)**. Within non-vernalized condition, 8 centrotypes consisting of 12 (based on MSE) and 15 peaks (based on node purity) from SPAD 2006, and 15 peaks (based on MSE) and 16 peaks (Node purity) from SPAD 2007 were found common (Table 15). Hence, these common peaks could signify the important compounds in relation to SPAD traits of *Brassica* under non-vernalized condition.

However, only two centrotypes, namely; 5441, 9834 and one peak: A60_448_872 associated with SPAD of vernalized condition of 2007 were matched with peaks selected for non-vernalized SPAD 2006, and with only two peaks; 9834_457_1891 (based on MSE) and 4867_175_698 (based on node purity) of non-vernalized SPAD 2007 (Table 15). The selection of different peaks for SPAD traits of vernalized and non-vernalized conditions is in agreement with the low correlation between SPAD of these two conditions.

5.2.3 Association study of molecular marker with SPAD traits

In RF regression of markers with residuals of all SPAD traits, four markers, namely; pTAmCAC.90.2, Mse.M476.5, Rsa.M385.1 and Alu.M394.5 were found highly associated with SPAD trait of vernalized condition, 2007 (Fig. 25), and RF regression explained 8.11 % variances of SPAD on this condition. In association mapping studies via TASSEL software, pTAmCAC.90.2 and Alu.M394.5 were also significantly (p-value = 0.05) associated with SPAD vernalized condition.

In non-vernalized condition, SPAD was measured in 2006 and 2007, and association mapping studies were done separately in both conditions. RF regression showed markers; Hae.M461.3, pGGmCAA154.10. Markers pTAmCAC-181.4 and pTAmCAT336.9 were the top most variable accordingly with high impact on explaining the variances of non-vernalized SPAD 2006 (Fig. 27). Same markers were also found the most importantly associated with non-vernalized SPAD under 2007, moreover, one more marker; Hae288.8 was also highly associated (Fig. 29). Also in association mapping studies from TASSEL software, the same markers as found in RF regression analyses were the most significant markers. pTAmCAT.336.9 (p-value=0.0013), Hae.M461.3 (p-value=0.0014), pTAmCAC.438 (p-value=0.0024), pGGmCAA.154.10 (p-value=0.0024), Mse.M271.2 (p-value=0.0027) and pTAmCAC.181.4 (p-value=0.0062) were the top most important markers. Out of top five markers obtained from RF regression, top three markers selected from RF regression were same as TASSEL's result. In these RF regressions, 359 markers explained 10.14 % and 8.98 % variance of non-vernalized SPAD 2006 and 2007 respectively. Box plots in Fig. 26, 28, 30 (see appendix) showed only few markers seem associated with SPAD trait under different conditions.

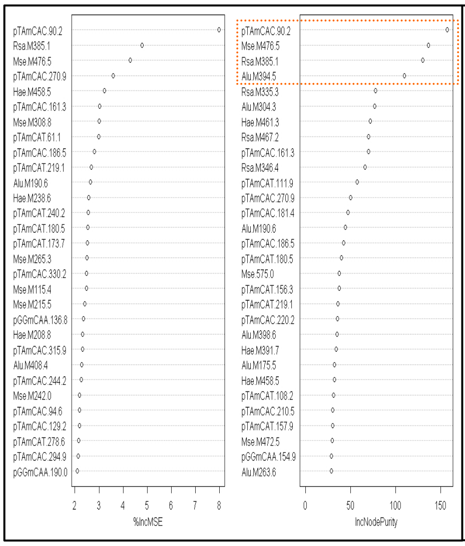


Fig. 25: List of the selected markers on RF regression on the residuals of SPAD on vernalized 2007

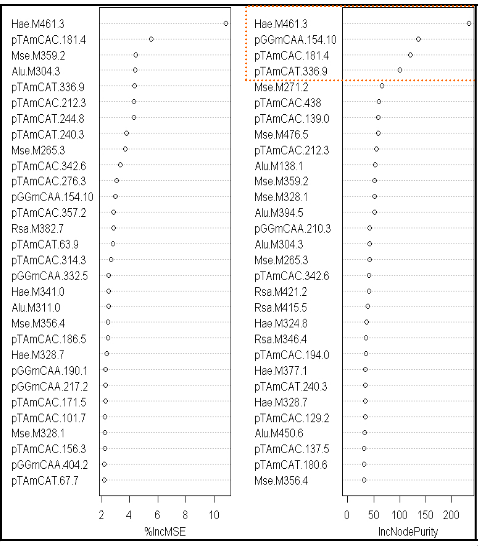


Fig. 27: List of the selected markers on RF regression on the residual of SPAD on Non-vernalized 2006

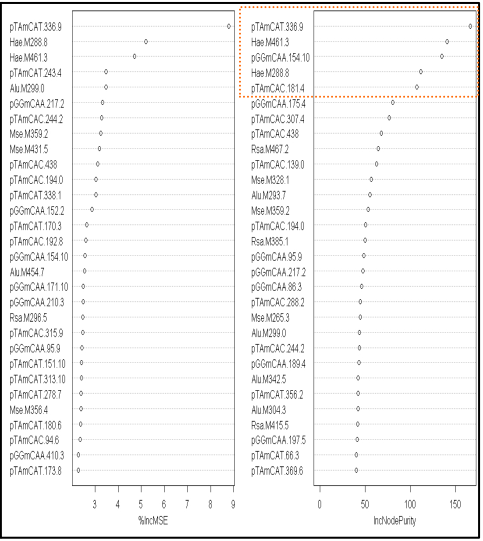


Fig. 29: List of the selected markers on RF regression on the residual of Non-vernalized SPAD 2007

5.3 Discussion

Association study of LC-MS peaks with SPAD traits:

LC-MS peaks associated with SPAD was selected in order to identify metabolites involved in leaf color (measured by SPAD) of *B. rapa*. The differences in leaf color as well as associated LC-MS peaks were observed between with- and without- vernalization treatment. Among the top 30 LC-MS peaks association with SPAD under vernalized condition, 24 peaks were from 22 centrotypes, suggesting the possibility of 22 different compounds; however, the remaining 6 peaks did not have a centrotype assigned. Those peaks with unknown centrotype might be different compounds.

Under the non-vernalized condition, eight centrotypes consisting of 11 peaks associated with SPAD traits of plants grown in 2006 and 15 peaks from plants grown in 2007 were in agreement despite the different growing years. Those 8 centrotypes most probably are the important compounds that are involved in leaf color development under non-vernalized condition. The remaining peaks were not in agreement with the peaks found under the non-vernalized treatment indicating a high influence of environmental conditions, even though plants were grown in a similar growth season (August). In comparison with peaks selected for SPAD of vernalized plants, only three centrotypes and one unidentified peak were in common. Those common centrotypes and peaks suggest a set of basic compounds which are essential for development of different colors.

The SPAD correlation was higher within the non-vernalized conditions despite grown in the same seasons of different years than that of vernalized plants grown in the same year but in different seasons. This showed that leaf color was influenced under the different treatments and growing seasons. Under these two treatments, only the light condition was different because of growing the plants in different months. Vernalized plants were grown in March 2006 and non-vernalized plants were grown and transplanted in August (in both 2006 and 2007) when the day length is longer and the intensity of light is higher. Besides, vernalized plants also develop faster and vigorously that may allocate metabolites to, for example flowers and other reproductive parts earlier. Although the mean SPAD values varied only in class 1 and 2 under with- and without- vernalization treatment (Table 15), the variation SPAD values were observed on morphotypes (Fig. 31). Morphotypes, such as VT, KOM, TG had higher SPAD value under vernalized condition but other morphotypes had high variation. The differences in SPAD values might be due to either light condition during the plant growth or vernalization effect. A separate study is needed to partition the effect of vernalization and / or light in leaf

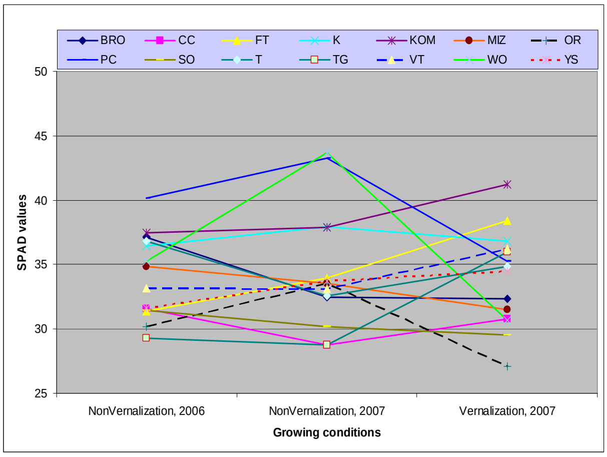


Fig. 31: Changing patterns of SPAD values in different morphotypes under with and without vernalization treatments.

color development. The annotation of selected centrotype or peaks is essential to know the biological relevance of the compounds related to leaf color in *B. rapa*. Some of the earlier studies also reported the influence of vernalization on the growth and development of morphological traits. Guo *et al* (2004) found an increased number of inflorescence branches, and reduction of stem diameter, final leaves number and final plant height under vernalization treatment in *B. oleracea*. Burton *et al* (2008) observed a significant reduction of leaf number in vernalized *B. napus*. During vernalization, the exposure to low temperature slows down the plant growth and development, and the plant resumes normal growth only after being transferred to the higher temperatures. However, for many accessions flowering time is seriously reduced, so they show quicker development with less vegetative growth. Because of the late initiation of plant growth and development, the characteristics of morphological traits are affected by vernalization. A significant variation on leaf angle, length and shape but no effect on leaf ratio (leaf blade length divided by total leaf length) was also recorded due to the effect of vernalization in *Arabidopsis thaliana* (Hopkins *et al.,* 2008). However, there was no previous study regarding the effects of vernalization with respect to leaf color of *B. rapa* accessions, and further study is necessary to confirm the changes in leaf color under vernalized and non-vernalized conditions.

Association study of markers with SPAD traits:

RF regression and a mixed model are two different approaches that were used to identify markers associated with SPAD under vernalization and non-vernalization. In the RF regression approach, SPAD under three different conditions (vernalized 2007, non-vernalized 2006 and 2007) were first corrected for the possible influence of population structure, and then used for marker-trait association studies. Zhao *et al.,* (2007a) found that the correction for population structure significantly reduces the number of false positive results in marker-trait associations in almost similar *B. rapa* core collection to the one used in this study. The top most markers (two in vernalized and six in non-vernalized conditions) were found in common in both approaches. However, association mapping, which takes into account the population structure does not sufficiently address the complex pattern of relatedness of all the accessions (Zhao *et al.*, 2007b). Yu *et al.,* (2006) suggested to include both population structure (Q-matrix) and kinship information (K-matrix) to have better control of spurious marker-trait association than correcting for only one matrix. This mixed model approach includes both population structure (in the Q-matrix) and individual relatedness within and between populations (in the kinship or K-matrix) to correct for multiple levels of relatedness rather than only population structure. This approach had a better performance on controlling both false positives (type I error) and false negatives (type II error) with higher statistical power than a simple regression approach, and also higher than association analysis considering only kinship relatedness (Yu and Buckler, 2006; Yu *et al.*, 2006; Zhao *et al.*, 2007b). RF regression approach can also be used for association studies and it gave the similar markers in this study, with that of the mixed model approach. Although, RF regression approach is not commonly used for association studies, Lunnetta *et al.,* (2004) used it for genome-wide association study of SNP markers with complex human diseases and reported a better performance than standard univariate methods especially in high dimensional data. RF regression can handle large numbers of predictors with small effects using their interaction (Lunnetta *et al.,* 2004); however, other regression methods such as lasso regression, elastic net cannot handle the interaction effects of predictor variables (Personal communication C. Maliepaard).

This study showed that RF regression had a comparable performance with mixed model without necessity of calculating K-matrix separately like in the mixed model approach. Zhu *et al.,* (2008) suggested to use sufficient numbers of co-dominant molecular markers (for example; SSR markers) rather than dominant markers (for example; RAPD, AFLP, myb) distributed across the genome in calculating population structure and familial relatedness of individual accessions. However, only 359 AFLP markers were used for this study, where only 90 markers have their map position. Hence, we are not sure whether these markers are sufficient in number and well-distributed across the whole genome. Therefore, the assignment of map position to markers and maintenance of sufficient markers through out the genome is essential to confirm these results and to draw the conclusion on the necessity of the calculation of kinship relatedness of accessions for the association study.

Similar to LC-MS peaks-SPAD association, the similarity in marker-trait association was found for SPAD traints under non-vernalized treatment (2006 and 2007) than that of vernalized (2007). Among the top 30 RF selected markers, 9 markers were in common in association with SPAD under both vernalilzed and non-vernalized conditions, and 12 markers were common with non-vernalized conditions. However the common markers found within non-vernalized condition were in the top rank with high importance (marked by red box in Fig. 25, 27, 29). Zhao *et al.,* (2007a) also found no correlation in days to flowering (DF) trait under with- and without-vernalization. The previous literatures reported the effect of vernalization treatment on morphological traits (discussed earlier). Thus, this suggests that SPAD might have different QTLs under these two conditions. In a QTL mapping study in a doubled haploid population (namely DH68) of a cross YS 143 x PC 175 under non-vernalized treatment, the same QTLs had recorded for SPAD of two different growing seasons; April and August in 2007 (Personal communication: Dunia Pino Del Carpio), which supports the similarity in the results of association studies within the non-vernalized treatment. But the markers for those QTLs were different from the results of this study, which might be because of different mapping population as well as allelic frequency. The oil type accessions, such as YS-143 (one of the parents of DH68) were very few (only 8 accessions) in core collection, which cause the variation in allelic frequency. Therefore, it is important to take into account for the effect of vernalization treatment and growing conditions when carrying out metabolic and genetic studies in *B. rapa* accessions.

Box plots in Fig. 26, 28 and 30 showed the smaller differences in SPAD trait with respect to marker alleles indicating no strong association of marker-SPAD trait. Reasonably, the SPAD trait could have relatively less variation in this population or the associated markers may not be representing the genomic regions contributing leaf color development. The SPAD value of different morphotypes in Fig. 31 also explained relatively small range of variation within each growing condition. Thus, the use of population with wider variability in leaf color or gene-targeted markers genotyping could be helpful for further research.

5.4 Conclusions

The effects of vernalization together with light conditions during plant growth were discovered in identifying LC-MS peaks and makers associated with SPAD traits in *B. rapa* accessions, which indicate the influences of environmental factors on metabolic profile as well as genetic factors of leaf color development. RF regression only with structure correction had comparable performances with unified mixed model approach correcting including both population structure and kinship relations of the accessions.

Chapter 6: Network Analysis of metabolic peaks associated with carotenoid pathway and leaf color of *B. rapa*

6.1 Materials and Methods

A Network is an extended form of graph with additional information on the vertices and the edges of the graph (de Nooy *et al.*, 2005). In marker integrated metabolites networks, the vertices are metabolites (or LC-MS peaks) and markers, and edges correspond to their correlations, where an edge is given between vertices if the metabolite-metabolite correlations or marker-metabolite associations are higher than some pre-defined thresholds. Simple Pearson correlations and partial correlations were calculated to construct metabolic correlation networks. The stronger the correlations, the thicker the edges while the higher the degree of connections, the bigger the size of vertices. Hence, correlation networks give a good visualization of how metabolites are related to each other, and also give the functional and regulation relations of metabolites by comparing them with known biochemical pathway (Ursem *et al.*, 2008). For the construction of the metabolic networks, Pearson correlations were estimated for 16 targeted metabolites, and also for 30 unknown metabolites selected by random forest (regression approach) statistical analysis on the SPAD trait (vernalized condition 2007). Correlations were calculated after correction for population structure to remove the spurious correlations. This was done by storing residuals of metabolites in analysis of variance (ANOVA) with four structure groups, and then the correlations were calculated on those residuals.

In the simple Pearson correlations (zero-order correlations), there may be overestimation of correlations because of both direct and indirect relations between metabolites. This also estimates false correlations between two metabolites, which might be because of other metabolites of the pathway. Thus, those networks may not have straightforward correspondence with the underlying metabolic reaction network (Khanin and Wit, 2007; Morgenthal *et al.*, 2006; Steuer *et al.*, 2003). Partial correlation coefficients were used to filter the causal relationship of metabolites due to indirect effect of other metabolites from large numbers of potential links in simple correlation which measures the direct correlation after removing the effects of intermediate metabolites or ancestry metabolites of the pathway (Khanin and Wit, 2007). The exact undirected dependency graphs (UDG) were constructed for n metabolites (16 metabolites in the carotenoid pathway and 30 in vernalized SPAD network) by estimating full-order partial correlations. In full order partial correlation, n-2 variables (14 metabolites in the carotenoid pathway and 28 peaks in SPAD network) were controlled while calculating the correlation between the remaining two because it is not possible to control one or a set of specific metabolites in undirected network (Opgen-Rhein and Strimmer, 2007). Partial correlations give the exact correlation between two metabolites, even though two or more than two indirect pathways are present (de la Fuente *et al.*, 2004). Thus, comparison of simple and partial correlation networks indicate the direct or indirect metabolites relations in the underlying pathway.

The associations of markers with metabolites were analyzed by one-sample t-tests of each marker with the residuals of the SPAD traits, and these t-test results were then integrated into the partial correlation network by using the t-test statistic to visualize the association of the markers with the metabolites. All the t-test statistics were divided by the highest one to convert them into 0-1 scale so that it will be in the comparable with metabolite-metabolite correlations for effective visualization. Metabolite-marker integrated networks provide the information of the linked markers of metabolites in the pathway. Markers which are associated with more than one metabolite might linked to genes involved in transcriptional regulation, and, in contrast, more markers associated with one metabolite might signify the presence of multiple QTL effects on that metabolite. All the correlations, partial correlations and t-test results for marker associations were calculated in the R-statistical software. Full-order partial correlations were estimated by using the “corpcor” package (Schaefer *et al.*, 2008) while their significance tests were calculated via the function “pcor.test” (http://www.yilab.gatech.edu/pcor.html). All the networks were constructed using the Pajek graph drawing software (Batagelj and Mrvar, 2003) for visualization of the correlation matrix of metabolites and associated markers. In all networks, simple and partial correlation and marker-metabolites association having q-value < 0.05 were used as a threshold to retain an edge in metabolites networks and metabolites-marker association. An FDR correction algorithm suitable for dependency condition developed by (Benjamini and Yekutieli, 2001) was used to calculate q-values in the R-software. This algorithm is less conservative than Bonferroni (Salvador *et al.*, 2005) and local FDR (fdr) (Aubert *et al.*, 2004), and is superior in controlling false discovery rate in highly correlated variables (Pounds, 2006). For easy interpretation and visualization, correlation of metabolites have > |0.5| were marked with green color, and the remaining ones in grey scale, while all the marker-metabolites association were presented in a red colour.

6.2 Results

Networks constructed based on simple and partial Pearson correlations as well as markers integration for (1) targeted metabolites of caretonoid pathway, analyzed by LC-MS, and (2) selected LC-MS peaks of unidentified compounds related to SPAD traits (leaf color) of vernalized *Brassica* core collection are shown below:

6.2.1 Networking of targeted metabolites of Carotenoids and tocopherol pathway

Simple correlation based networking shows that most of the metabolites are related to eachother; however, two distinct groups of networks (tocopherol and carotenoids) appeared at a high level of correlation (pearson $r \geq 0.5$). Only two negative correlations were found between δ-tocopherol and an unidentified compound, and between lutein-1 and lutein-2. The metabolites γ-tocopherol, δ-tocopherol, lutein-1 and an unidentified compound (named as UNKNOWN in the network figures) were less involved in this pathway; these had less connectivity with other compounds (the smaller the size of vertices, the lower the degree of connectivity) (Fig. 32).

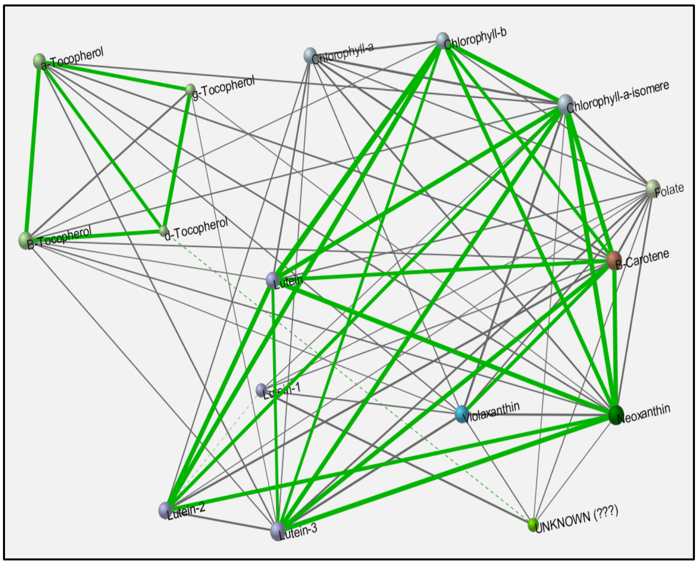


Fig. 32: Network visualization based on simple Pearson correlations of targeted metabolites of the carotenoid pathway. Green lines indicate for r ≥ |0.5| and grey lines for r < |0.5| correlation values, and the thickness of the edges represents the strength of correlations. The dashed lines show negative correlation.

In partial correlation network, many relations observed in zero-order correlation disappeared; however, some new relations, especially negative correlations, were also appeared. Like in the simple correlation network, two branches of the pathway were observed. The tocopherol pathway splits from the carotenoid pathway, and linkage between the γ-tocopherol and chlorophyll-a act as a bridge between the two pathways. Three lutein derivatives, and in between chlorophyll-a, Lutein-2, and Lutein-3 had a negative clique, while chlorophyll-a-isomer, neoxanthin and violaxanthin, and chlorophyll-a, chlorophyll-a-isomer and violaxanthin had a positive clique (Fig. 33).

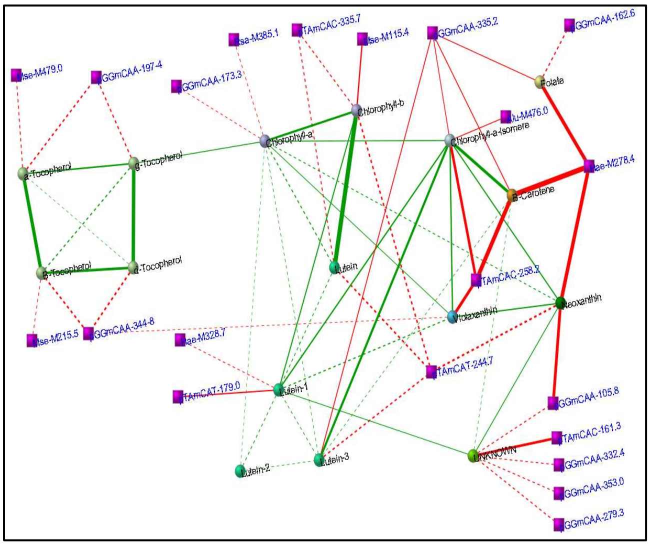


Fig. 33: Partial correlation networks of targeted metabolites of carotenoid pathway and integration of associated markers. Green lines indicate a partial correlation network, where solid lines for positive and dotted lines for negative correlation of the metabolites. Red lines describe the markers' associations. The thickness of the lines represents the strength of relations. Round vertices with black labels symbolize metabolites while square boxes (pink color) with blue label are indicating markers.

Markers' associations with the metabolites were integrated in the partial correlation network to depict the genetic information of the pathway regulation (Fig. 33). Dotted lines between the marker and metabolites described the presence of marker (gene) that had negative association with the metabolites indicating down-regulation in the pathway (Fig. 33). Markers; pTAmCAT-244.7, Hae-M278.4, pTAmCAC-258.2, pGGmCAA-355.2, pGGmCAA-344.8, pGGmCAA-197.4, pTAmCAC-335.7 and pGGmCAA-105.8 have association with at least 2 metabolites. These kinds of associations could delineate genetic loci harboring important genes that regulate the metabolites together, although pGGmCAA-335.2 marker is only weakly association with the respective metabolites. On the other hand, two or more than two markers were affiliated with metabolites; chlorophyll-a, lutein-1 and an unidentified compound. However, box plot analysis gave impression of having false positive association of markers pGGmCAA-105.8, pGGmCAA-332.4, pGGmCAA-279.3 and pGGmCAA-353.0 with an unidentified compound, and pGGmCAA-162.6 marker with folate compound at 0.05 level of q-value, where the level of metabolites did not seem different with alleles of the markers (Fig. 34; appendix).

6.2.2 LC-MS peaks for vernalized SPAD

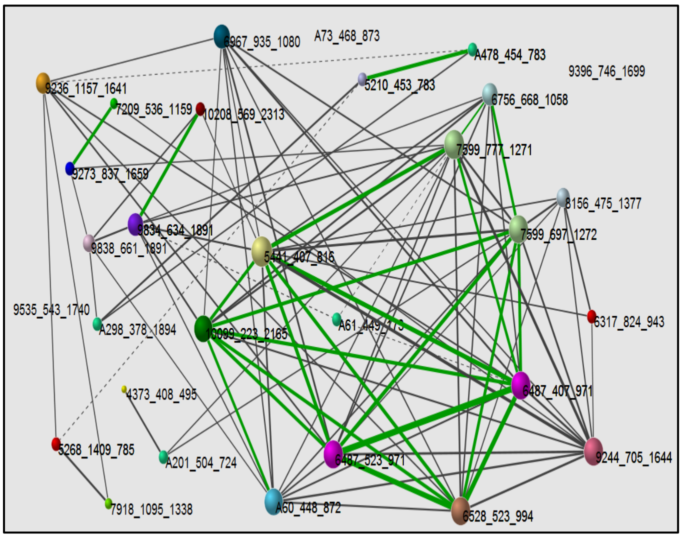


Fig. 35: Simple Pearson correlation network of selected LC-MS peaks related to SPAD in the vernalized *Brassica* core collection. The thickness of lines indicates the strength of correlation, solid lines for positive correlation and dot lines for negative correlations. The green lines show for correlation ≥ |0.5| and grey lines for < |0.5|. The bigger the size of the vertices indicates the higher numbers of connection in the network.

Visualization of Pearson correlation network of RF selected top 30 LC-MS peaks for SPAD trait in vernalized condition shows that most of the peaks have connections to at least other peaks except 9535_543_1740, 9396_746_1699 and A73_468_873. In this network, peaks with the bigger size of vertex symbolize a higher degree of connections; whereas, the smaller sized vertices mean lower connections. Most commonly, peaks with higher connectivity have a higher strength of relations. Negative correlations (dot lines) were also observed between some peaks. Peaks with centrotype 5441, 7599, 10099, 6487, 6528, 9244 and 6967 had high numbers of connections with other metabolites. Centrotypes coded by a letter “A” followed by numbers, for example, A60, A61, A201, A298 and A478 were for all the peaks that did not have centrotype (Fig. 35).

In partial correlation based networking, only few connections remained while some new relations were also established. The LC-MS peaks 9535_543_1740 and A73_468_873 have new relations of positive correlation with A298_378_1894 and 5441_407_816 respectively although they were free in the simple correlation network. Relations between 10208_569_2313 and 9834_634_1891, 9273_837_1659 and 7209_536_1159, 6487_407_971 with 6487_523_971 and 7599_777_1271, and A478_454_783 and 5210_453_783 peaks remained in both simple and partial correlation network with correlation > 0.5. Similarly, A61_449_173 and 7599_777_1271 had negative correlation in both correlation networks. All the green lines described correlations between the metabolic peaks, whereas red lines show associations of markers with metabolites. Marker information was integrated into the partial correlation network, where more than one marker was related to the most of the metabolites. In metabolic correlation network indicated by green lines, the thicker the lines, the higher the strength of correlation. Similarly, markers integrated with metabolic peaks shown in red lines, the thicker the line, the higher the explanation of variance of the metabolic peaks. Multiple peaks which have connections with the same marker(s) may be either the same compounds or the co-regulated compounds. Some markers have negative association (red dotted lines) with metabolites, where the mean of peak intensity was higher with the recessive marker allele. However, some peaks were associated with more than one marker, for example; peak 5210_453_783 has connections with 16 markers (Fig. 36).

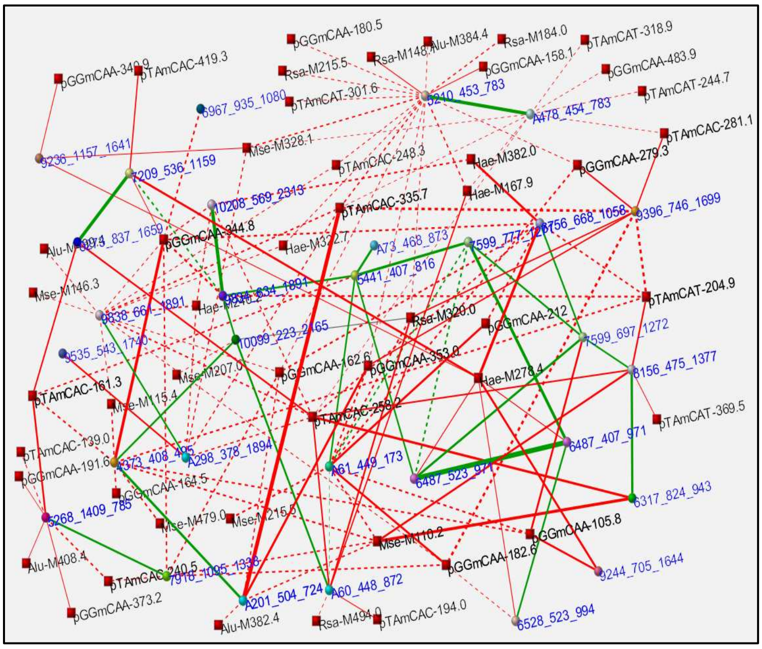


Fig. 36: Partial correlation network of metabolic peaks and integration of markers. Green lines denote correlations, where dotted line for negative and solid lines for positive correlations of LC-MS peaks. Red lines indicate the associations of markers with peaks. Red square vertices represent markers and round coloured vertices represent LC-MS peaks.

6.3 Discussion

Network analyses were conducted on the data from targeted metabolites profile of the carotenoid and tocopherol pathway, and, separately, on the LCMS peaks associated with SPAD under vernalization. The simple Pearson correlation was used to understand the interaction between the metabolites. However, simple correlation based networks are of limited use because of confounding effects of direct and indirect associations of metabolites on each other, which can overestimate the strength and number of links between metabolites (Khanin and Wit, 2007). Hence, the simple correlation can not distinguish between indirect relations of metabolites of the underlying pathway. Therefore, I also investigated the partial correlations. A partial correlation assesses the strength of the relation between two metabolites after controlling the effects of other metabolites (Khanin and Wit, 2007). Hence, partial correlation networks depict only the direct linear associations of metabolites (Khanin and Wit, 2007; Opgen-Rhein and Strimmer, 2007). Full-order partial correlation calculates the relation of two metabolites controlling the influence of all the metabolites of network, which depicts only the independent correlation between two metabolites. However, de la Fuente *et al.*, (2004) suggested the second-order correlation to construct independent network. In this study, the full-order partial correlation network was constructed where lower-order partial correlation is not possible because of unknown direction of topological order of the metabolites in undirected network (Opgen-Rhein and Strimmer, 2007).

Network analysis of targeted metabolites of carotenoid and tocopherol pathway:

In the carotenoid and tocopherol pathway, negative correlations were observed between δ-tocopherol and an unidentified compound, and between δ-tocopherol and lutein-1 and lutein–2 indicating the possibly antagonistic effects between them.

The simple Pearson correlation (simple correlation) showed the correlation of almost all metabolites with each other with different strength at false discovery control (FDR) of 0.05 (Fig. 32). Even though almost all metabolites were connected to each other, two distinct groups of metabolites were observed based on the strength of the correlation. One group consisted of different tocopherol compounds while another group had chlorophyll and its derivatives, folate, lutein and its derivatives, beta-carotene, violaxanthin and neoxanthin. Those two groups indicate the possibility of two branches of a pathway. Partial correlation based networking was incorporated with associated markers (Fig. 33), where marker-metabolites association was calculated at 0.05 level of FDR using simple two-sample t-test after correcting for population structure. The partial correlation also showed the presence of two branches of a carotenoid and tocopherol pathway. Tocopherol and carotenoid were splitted into two branches from geranylgeranyl diphosphate (GGDP) in pathway shown by DellaPenna and Pogson (2006) (Fig 37), and have competition for the substrate GGDP between these two branches (Lindgren, 2003). Within the tocopherol sub-branch, γ- and α-tocopherols, and δ- and β- tocopherols were in separate chains (DellaPenna and Pogson, 2006; Maeda and DellaPenna, 2007). This study discovered negative correlations between γ- and β- tocopherols on the one hand, and α- and δ-tocopherols on the other hand suggesting the possibility of competition between two chains for substrate but not detail previous studies on their relations were found. Similarly, α- and γ-tocopherols had associated with the one marker pGGmCAA-197-4, and δ- and β-tocopherols with another pGGmCAA344-8. But DellaPenna and Pogson (2006) reported the same locus VTE1 involved in the synthesis of both δ- and γ-tocopherol, and VTE4 locus in the conversion of δ- to β-tocopherols, and γ- to α-tocopherol. This suggests that these markers might be linked to the same gene.

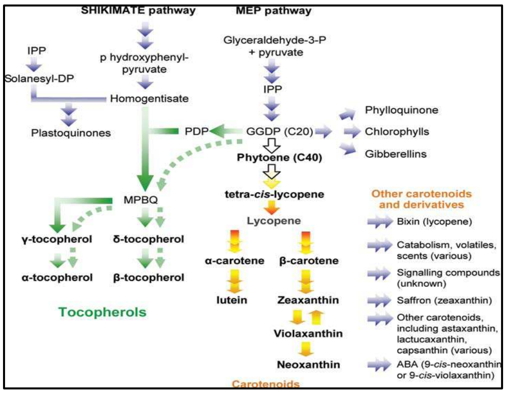


Fig 37: Overview of carotenoid and tocopherol pathways in plants.
(Source: DellaPenna and Pogson, 2006)

The negative correlation of lutein and its derivatives with β-carotene, violaxanthin and neoxanthin indicate for the presence of separate chain and might have competition in substrate for lutein biosynthesis. DellaPenna and Pogson (2006) and Kopsell and Kopsell (2006) demonstrated the pathway with the split of lycopene into α- and β- carotene, where α-carotene involved in the synthesis of lutein, and β-carotene for violaxanthin and neoxanthin. Since both lutein and violaxanthin are the component of carotenoid, the relation of biosynthesis of lutein and violaxanthin was not emphasized in the literatures. However, it has been reported that α-carotene is formed if one β-ring and ε- ring attached to lycopene, and β-carotene formed if two β-rings attached (Lu and Li, 2008). In this study, marker Hae-M278.4 was strongly associated with β-carotene, neoxanthin and folate, and marker pTAmCAC-258.2 with β-carotene, violaxthin and chlorophyll-a-isomer (Fig. 33) indicating these markers could be linked to a common regulatory gene of this side-chain because DellaPenna and Pogson (2006) enlisted the defferent genes were involved in the biosynthesis of each of these metabolites (DellaPenna and Pogson, 2006). In green plant tissue, lutein is found abundant than violaxanthin (Lu and Li, 2008). Hence, there could be competition for the substrates for these two side-chains of carotenoid pathways.

In addition, markers associated with more than two metabolites may indicate that those metabolites might have a regulatory gene in common or be linked to same genes. In contrast, metabolites with more than one marker could indicate either the possibility of markers all linked to the same gene or the presence of multiple QTL effects, if those markers correspond to different genomic regions. The identification of genetic map position of these markers is essential to confirm the possibility of the presence of one or more associated genes. Similarly, negative association of marker-metabolites may indicate the down-regulation of gene associated with corresponding markers will stimulate the metabolites synthesis (Fig. 34; appendix). Thus, this type of network analyses was largely used in the study of metabolic pathway to explore the relations of metabolites and their genetic control in a pathway.

Network of LCMS peaks associated with SPAD traits under vernalized condition:

LC-MS peaks associated with SPAD under vernalization condition were selected in an RF regression approach and used for network analysis. Simple and partial correlations of those selected metabolites were calculated for network analysis to know the relation of metabolites involved in leaf color, and here again associated markers were integrated to allow also a genetic interpretation. The top 30 selected peaks were studied in a correlation network. Most of the peaks were from different centrotypes indicating possibly different compounds. Like in the case of the carotenoid and tocopherol pathway above, the peaks with higher connectivity possibly indicate that important compounds might play a role in the different leaf colors (Fig. 35).

In the partial correlation network, only a few peaks remained in connections than that of simple correlation network, and also some of new connections were discovered (Fig. 36). The peaks connected with very high strength of correlation in both simple and partial correlation networks could indicate the possibility of having the metabolites with similar chemical structures or strongly co-regulated compounds of the pathway. Similarly, some of the peaks with negative correlation in both correlation networks could indicate that compounds are present in different sub-branches of the same pathway and might have competition for the common substrates. Associated markers were also added in the partial correlation network that will helpful to annotate the peaks by prioritizing them. If one of the peaks is annotated, network analysis will facilitate to predict compounds, its function and genetic interpretation of other peaks of network based on the strength of peaks correlations and marker-peaks association as well as to make genetic interpretation. Here again associated markers were identified by simple two-sample t-test procedure after correction for population structure. However, the markers associated with LC-MS peaks selected for SPAD in this network analysis were different from the markers identified via RF regression and the mixed model in the association study (Chapter V). Statistically, those differences might be due to two-step procedures adopted in this t-test procedure, where LC-MS peaks first selected for SPAD trait of vernalized condition via. RF regression and then, marker-LC-MS peaks association after correcting for population structure were done. However, in marker-trait association studies with vernalized SPAD trait in previous chapter, marker-SPAD trait association was directly studied by correcting the population structure. Hence, these were two-step procedures. However, biologically, the markers associated with SPAD trait indicated the correlation of markers with the leaf color phenotypes, whereas in case of marker-LC-MS peaks (selected for vernalized SPAD trait) gave the marker correlated with metabolites involved in leaf color development. Besides, SPAD measurement were at pre-mature and fully mature stages of plants but LC-MS analysis was conducted on 5 weeks old plant samples. This difference in growth stages could affect on the qualitative and quantitative composition of metabolites, and make differences in marker-trait association studies. Therefore, annotation of LCMS peaks and further studies on the roles and function of annotated peaks, and mapping the markers are necessary to have detailed understanding and exploration of pathway related to leaf color of *Brassica rapa* accessions.

6.4 Conclusion

The network analysis based on simple and partial correlation of the metabolites was found useful to know the direct and indirect relation of metabolites under a certain pathway. The network analyses of the targeted metabolites of carotenoid and tocopherol pathways, in this study were in good correspondence with the underlying carotenoid and tocopherol pathway of the plants. The integration of associated markers in the correlation network provides the genetic relation of the metabolites in a pathway. The networking of the selected LC-MS peaks related to leaf color of *B. rapa* showed the possibility of exploring the relations of LC-MS peaks making ease to annotate those peaks and their interpretation under a particular pathway.

Chapter 7: Further direction

Random forest (RF) is an example of machine learning statistical technique. The comparative study with other machine learning techniques, such support vector machine will be helpful to know the performance of RF classification.

RF classification gave the relative importance of the variables in classifying the accessions into different classes and morphotypes. Further study by using appropriate supervised classification techniques, such as different kinds of Discriminant Analyses will be useful to develop discriminating equation of RF selected variables for classifying the accessions into different groups.

Mapping of molecular markers and annotation of LC-MS peaks associated with SPAD trait is important to draw the conclusions on the nature of genes whether major or minor genes are related to leaf color, and understanding the regulatory pathway of leaf color development.

In this study, the combined effect of vernalization and light condition have influence on association studies. Hence the further study to separate effects of these two factors will be useful to confirm the effect of vernalization on association studies.

Network analysis based on lower-order partial correlation might improve the networks of metabolites since the full-order partial correlation, used in this study, consider conservative in explaining the relations of metabolites.

Reference:

Allwood, J.W., Ellis, D.I. and Goodacre, R. (2008) Metabolomic technologies and their application to the study of plants and plant&#x2013;host interactions. *Physiologia Plantarum* **132**, 117-135.

Aubert, J., Bar-Hen, A., Daudin, J.J. and Robin, S. (2004) Determination of the differentially expressed genes in microarray experiments using local FDR. *BMC Bioinformatics* **5**, 125.

Balding, D.J. (2006) A tutorial on statistical methods for population association studies. *Nat Rev Genet* **7**, 781-791.

Batagelj, V. and Mrvar, A. (2003) Pajek - Analysis and Visualization of Large Networks. In: Juenger, M. and Mutzel, P. (eds.): Graph Drawing Software. Springer (series Mathematics and Visualization), Berlin. 77-103. ISBN 3-540-00881-0.

Benjamini, Y. and Yekutieli, d. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 1165-1188.

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ram-doss, Y. and Buckler, E.S. (2007) TASSEL: Software for Association Mapping of Complex Traits in Diverse Samples. *Bioinformatics*, btm308.

Breiman, L. (2001) Random Forests. *Machine Learning* **45**, 5-32.

Breiman, L., Cutler, A., Liaw, A. and Wiener, M. (2008) Breiman and Cutler's Random Forests for Classification and Regression. R package version 4.5-16, . *URL http://CRAN.R-project.org/package=randomForest*.

Burton, W.A., Flood, R.F., Norton, R.M., Field, B., Potts, D.A., Robertson, M.J. and Salisbury, P.A. (2008) Identification of variability in phenological responses in canola-quality *Brassica juncea* for utilisation in Australian breeding programs. *Australian Journal of Agricultural Research* **59**, 874-881.

Colquhoun, I.J. (2007) Use of NMR for metabolic profiling in plant systems. *Journal of Pesticide Science* **32**, 200-212.

de la Fuente, A., Bing, N., Hoeschele, I. and Mendes, P. (2004) Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* **20**, 3565-3574.

de Nooy, W., Mrvar, A. and Batagelj, V. (2005) Exploratory social network analysis with Pajek, First Edn: Cambridge University press, New York, USA.

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

DellaPenna, D. and Pogson, B.J. (2006) Vitamin synthesis in plants: tocopherols and carotenoids. *Annual Review of Plant Biology* **57**, 711-738.

Dettmer, K., Aronov, P.A. and Hammock, B.D. (2007) Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews* **26**, 51-78.

Diaz-Uriarte, R. (2007) GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. *BMC Bioinformatics* **8**, 328.

Diaz-Uriarte, R. and Alvarez de Andres, S. (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7**, 3.

Diez, A.G. (2008) Application of MYB markers in *Brassica rapa.* Master thesis. Wageningen: Wageningen University and Research (WUR), The Netherlands.

Dopazo, J. (2007) Clustering — Class Discovery in the Post-Genomic Era. In *Fundamentals of Data Mining in Genomics and Proteomics*, pp. 123-148.

Duarte, J.M., Santos, J.B.d. and Melo, L.C. (1999) Comparison of similarity coefficients based on RAPD markers in the common bean. *Genetics and Molecular Biology* **22**, 427-432.

Duhoon, S.S. and Koppar, M.N. (1998) Distribution, collection and conservation of bio-diversity in cruciferous oilseeds in India. *Genetic Resources and Crop Evolution* **45**, 317-323.

Dutilleul, P., Stockwell, J.D., Frigon, D. and Legendre, P. (2000) The Mantel test versus Pearson's correlation analysis: Assessment of the differences for biological and environmental studies. *Journal of Agricultural, Biological and Environmental Statistics* **5**, 131-150.

Everitt, B. (1980) Cluster Analysis: Halsted press, Division of John Wiley and sons, New York, USA.

Gislason, P.O., Benediktsson, J.A. and Sveinsson, J.R. (2006) Random Forests for land cover classification. *Pattern Recognition Letters* **27**, 294-300.

Goslee, S. and Urban, D. (2007a) Dissimilarity-based functions for ecological analysis. R package version 1.1.4,. *URL http://cran.r-project.org/package=ecodist*.

Goslee, S.C. and Urban, D.L. (2007b) The ecodist package for dissimilarity-based analysis of ecological Data. *Journal of Statistical Software* **22**, 1-19.

Guo, D.P., Ali Shah, G., Zeng, G.W. and Zheng, S.J. (2004) The Interaction of plant growth regulators and vernalization on the growth and flowering of cauliflower (*Brassica oleracea* var. botrytis). *Plant Growth Regulation* **43**, 163-171.

Hall, R.D. (2006) Plant metabolomics: from holistic hope, to hype, to hot topic. *New Phytologist* 169: 453-468.

Hastie, T., Tibshirani, R., Narasimhan, B. and Chu, G. (2008) impute: Imputation for microarray data. *The impute Package version 1.0-5., URL http://cran.r-project.org/web/package=impute*.

Hendrawati, O., Yao, Q., Kim, H.K., Linthorst, H.J.M., Erkelens, C., Lefeber, A.W.M., Choi, Y.H. and Verpoorte, R. (2006) Metabolic differentiation of Arabidopsis treated with methyl jasmonate using nuclear magnetic resonance spectroscopy. *Plant Science* **170**, 1118-1124.

Hopkins, R., Schmitt, J. and Stinchcombe, J.R. (2008) A latitudinal cline and response to vernalization in leaf angle and morphology in *Arabidopsis thaliana* (*Brassicaceae*). *New Phytologist* **179**, 155-164.

Husson, F., Josse, J., Le, S. and Mazet, J. (2008) Factor Analysis and Data Mining with R. R package version 1.10,. *URL http://cran.r-project.org/package=FactoMineR*.

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

Khanin, R. and Wit, E. (2007) Construction of Malaria Gene Expression Network Using Partial Correlations. In *Methods of Microarray Data Analysis V*, pp. 75-88.

Kopsell, D.A. and Kopsell, D.E. (2006) Accumulation and bioavailability of dietary carotenoids in vegetable crops. *Trends in Plant Science* **11**, 499-507.

Lattin, J.M., Carroll, J.D. and Green, P.E. (2003) Analyzing multivariate data: Pacific Grove, CA (etc): Thomson/Brooks/Cole.

Legendre, P. (2000) Comparison of permutation methods for the partial correlation and partial Mantel tests. *J. Statist. Comput. Simul.* **67**, 37-73.

Liang, Y.S., Kim, H.K., Lefeber, A.W.M., Erkelens, C., Choi, Y.H. and Verpoorte, R. (2006) Identification of phenylpropanoids in methyl jasmonate treated *Brassica rapa* leaves using two-dimensional nuclear magnetic resonance spectroscopy. *Journal of Chromatography A* **1112**, 148-155.

Lindgren, O. (2003). Carotenoid Biosynthesis in seed of Arabidopsis thaliana. Ph. D. thesis. Department of plant biology and forest genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden. pp: 1-33.

López-López, P., Benavent-Corai, J. and García-Ripollés, C. (2008) Geographical assemblages of European raptors and owls. *Acta Oecologica* **34**, 252-257.

Lu S. and Li, L. (2008). Carotenoid metabolism: biosynthesis, regulation, and beyond. *Journal of integrative plant biology* 50 (7): 778-785.

Lunetta, K.L., Hayward, L.B., Segal, J. and Van Eerdewegh, P. (2004) Screening large-scale association study data: exploiting interactions using random forests. *BMC Genetics* **5**, 32.

Luo, J. and Fox, B.J. (1996) A Review of the Mantel Test in Dietary Studies: Effect of Sample Size and Inequality of Sample Sizes. *Wildlife Research* **23**, 267-288.

Maeda, H. and DellaPenna, D. (2007) Tocopherol functions in photosynthetic organisms. *Current Opinion in Plant Biology* **10**, 260-265.

Moco, S. (2007) Metabolomics technologies applied to the identification of compounds in plants. PhD thesis. Wageningen: Wageningen University and Research (WUR).

Morgenthal, K., Weckwerth, W. and Steuer, R. (2006) Metabolomic networks in plants: Transitions from pattern recognition to biological interpretation. *Biosystems* **83**, 108-117.

Opgen-Rhein, R. and Strimmer, K. (2007) From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Systems Biology* **1**, 37.

Pang, H., Lin, A., Holford, M., Enerson, B.E., Lu, B., Lawton, M.P., Floyd, E. and Zhao, H. (2006) Pathway analysis using random forests classification and regression. *Bioinformatics* **22**, 2028-2036.

Paterson, A., Lan, T.-h., Amasino, R., Osborn, T. and Quiros, C. (2001) *Brassica* genomics: a complement to, and early beneficiary of, the Arabidopsis sequence. *Genome Biology* **2**, reviews1011.1011 - reviews1011.1014.

Pietiläinen K.H, Sysi-Aho, M., Rissanen, A., Seppänen-Laakso, T., Yki-Järvinen, H., Kaprio, J. and Oresïc, M. (2007) Acquired Obesity Is Associated with Changes in the Serum Lipidomic Profile Independent of Genetic Effects – A Monozygotic Twin Study. *PLoS ONE* **2**, e218.

McGrath, J.M. and Quiros, C.F. (1992) Genetic diversity at isozyme and RFLP loci in Brassica campestris as related to crop type and geographical origin. *TAG Theoretical and Applied Genetics* **83**, 783-790.

Pissard, A., Arbizu, C., Ghislain, M., Faux, A.-M., Paulet, S. and Bertin, P. (2008) Congruence between morphological and molecular markers inferred from the analysis of the intra-morphotype genetic diversity and the spatial structure of Oxalis tuberosa Mol. *Genetica* **132**, 71-85.

Podani, J. (2001) Numerical Ecology, by P. Legendre and L. Legendre. *Journal of Classification* **18**, 285-288.

Pounds, S.B. (2006) Estimation and control of multiple testing error rates for microarray studies. *Brief Bioinform* **7**, 25-36.

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **155**, 945-959.

Quackenbush, J. (2001) Computational analysis of microarray data. *Nat Rev Genet* **2**, 418-427.

Requena, C. (2007) Population structure of a *Brassica rapa* core collection:Comparison between molecular markers and morphological traits. Master thesis. Wageningen: Wageningen University and Research (WUR), The Netherlands.

Riechmann, J.L. and Ratcliffe, O.J. (2000) A genomic perspective on plant transcription factors. *Current Opinion in Plant Biology* **3**, 423-434.

Salvador, R., Peña, A., Menon, D.K., Carpenter, T.A., Pickard, J.D. and Bullmore, E.T. (2005) Formal characterization and extension of the linearized diffusion tensor model. *Human Brain Mapping* **24**, 144-155.

Schaefer, J., Opgen-Rhein, R. and Strimmer, K. (2008) Efficient Estimation of Covariance and (Partial) Correlation. R package version 1.5-1. *URL http://cran.r-project.org/packages=corpcor*.

Segal, M. R. (2004). Machine learning benchmarks and random forest regressions. Center for Bioinformatics and Molecular Biostatistics, University of California, San Franciso, USA. pp:1-14 http://repositories.cdlib.org/cbmb/bench_rf_regn

Shimodaira, H. (2002) An Approximately Unbiased Test of Phylogenetic Tree Selection. *Systematic Biology* **51**, 492 - 508.

Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R. and Siuzdak, G. (2006) XCMS: processing mass spectrometry data for metabolite porfiling using nonlinear peak alignment, matching and identification. *Analytical Chemistry* **78**, 779-787.

Smouse, P.E., Long, J.C. and Sokal, R.R. (1986) Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Systematic zoology* **35**, 627-632

Steinfath, M., Groth, D., Lisec, J. and Selbig, J. (2008) Metabolite profile analysis: from raw data to regression and classification. *Physiologia Plantarum* **132**, 150-161.

Steuer, R., Kurths, J., Fiehn, O. and Weckwerth, W. (2003) Observing and interpreting correlations in metabolomic networks. *Bioinformatics* **19**, 1019-1026.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. and Zeileis, A. (2008) Conditional variable importance for random forests. *BMC Bioinformatics* **9**, 307.

Strobl, C., Boulesteix, A.-L., Zeileis, A. and Hothorn, T. (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* **8**, 25.

Suzuki, R. and Shimodaira, H. (2004) An application of multiscale bootstrap resampling to hierarchical clustering of microarray data: How accurate are these clusters? *In: Proceedings by the Fifteen Internatinal Conference on Genome Informatics (GIW 2004)*, P034.

Suzuki, R. and Shimodaira, H. (2006a) Hierarchichal clustering with P-Values via Multiscale Bootstrap Resampling. R package version 1.2-0. *URL http://cran.r-project.org/package=pvclust*.

Suzuki, R. and Shimodaira, H. (2006b) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22**, 1540-1542.

Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P. and Feuston, B.P. (2003)

Random Forest:&nbsp; A Classification and Regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences* **43**, 1947-1958.

Takuno, S., Kawahara, T. and Ohnishi, O. (2007) Phylogenetic relationships among cultivated types of *Brassica rapa* L. em. Metzg. as revealed by AFLP analysis. *Genetic Resources and Crop Evolution* **54**, 279-285.

Telles, M.P.C. and Diniz-Filho, J.A.F. (2005) Multiple Mantel tests and isolation-by-distance, taking into account long-term historical divergence. *Genet and Mol Res* **4**, 742-748.

Truong, Y., Lin, X. and Beecher, C. (2004) Learning a complex metabolomic dataset using random forests and support vector machines. In: *Proceedings of 10th ACM SIGKDD international conference on knowledge discovery and data mining.* pp: 835-840

Ursem, R., Tikunov, Y., Bovy, A., van Berloo, R. and van Eeuwijk, F. (2008) A correlation network approach to metabolic data analysis for tomato fruits. *Euphytica* **161**, 181-193.

van Berloo, R., Zhu, A., Ursem, R., Verbakel, H., Gort, G. and van Eeuwijk, F. (2008) Diversity and linkage disequilibrium analysis within a selected set of cultivated tomatoes. *TAG Theoretical and Applied Genetics* **117**, 89-101.

Verpoorte, R., Choi, Y. and Kim, H. (2007) NMR-based metabolomics at work in phytochemistry. *Phytochemistry Reviews* **6**, 3-14.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., Lee, T.v.d., Hornes, M., Friters, A., Pot, J., Paleman, J., Kuiper, M. and Zabeau, M. (1995) AFLP: a new technique for DNA fingerprinting. *Nucl. Acids Res.* **23**, 4407-4414.

Ward, J.L., Baker, J.M. and Beale, M.H. (2007) Recent applications of NMR spectroscopy in plant metabolomics. *FEBS Journal* **274**, 1126-1131.

Ward, J.L., Harris, C., Lewis, J. and Beale, M.H. (2003) Assessment of 1H NMR spectroscopy and multivariate analysis as a technique for metabolite fingerprinting of Arabidopsis thaliana. *Phytochemistry* **62**, 949-957.

Warwick, S.I., James, T. and Falk, K.C. (2008) AFLP-based molecular characterization of *Brassica rapa* and diversity in Canadian spring turnip rape cultivars. *Plant Genetic Resouces* **6**, 11-21.

Weckwerth, W. and Morgenthal, K. (2005) Metabolomics: from pattern recognition to biological interpretation. *Drug Discovery Today* **10**, 1551-1558.

Widarto, H., Van Der Meijden, E., Lefeber, A., Erkelens, C., Kim, H., Choi, Y. and 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* **32**, 2417-2428.

Wit, J. and McClure, J. (2004) Statistics for microarray: design, analysis and inference. Chichester: Jobn Wiley and Sons.

Yu, J. and Buckler, E.S. (2006) Genetic association mapping and genome organization of maize. *Current Opinion in Biotechnology* **17**, 155-160.

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. and Buckler, E.S. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* **38**, 203-208.

Zhang Z., Btadbury P. J., Kroon, D.E., Casstevens, T.M. and Buckler, E.S. (2006) TASSEL 2.0: A Software Package For Association And Diversity Analyses In Plants And Animals. *In: Plant & Animal Genomes Conference, 14th-18th January 2006*.

Zhao, J. (2007) The genetics of phytate content and morphological traits in *Brassica rapa.* PhD thesis. Wageningen: Wageningen University and Research (WUR), The Netherlands.

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

Zhao, J., Wang, X., Deng, B., Lou, P., Wu, J., Sun, R., Xu, Z., Vromans, J., Koornneef, M. and Bonnema, G. (2005) Genetic relationships within *Brassica rapa* as inferred from AFLP fingerprints. *TAG Theoretical and Applied Genetics* **110**, 1301-1314.

Zhao, K., Aranzana, M.J., Kim, S., Lister, C., Shindo, C., Tang, C., Toomajian, C., Zheng, H., Dean, C., Marjoram, P. and Nordborg, M. (2007b) An *Arabidopsis* example of association mapping in structured samples. *PLoS Genet* **3**, e4.

Zhou, X., Van Eck, J., Li, L. and El-Gewely, M.R. (2008) Use of the cauliflower Or gene for improving crop nutritional quality. In *Biotechnology Annual Review*, Volume Volume 14: Elsevier, pp. 171-190.

Zhu, C., Gore, M., Buckler, E.S. and Yu, J. (2008) Status and prospects of association mapping in plants. *The Plant Genome* **1**, 5-20.

Zuur, A.F., Ieno, E.N. and Smith, G.M. (2007) Principal coordinate analysis and non-metric multidimensional scaling. In *Analysing Ecological Data*, pp. 259-264.

Appendices

Table 1: List of accessions used in this study with geographic origins and their STRUCTURE classes membership probabilities

				Structure class membership probabilities			
S.N.	Accession	Geographic origin	Structure class	C1	C2	C3	C4
Broccoletto (ssp. *broccoletto*)							
1	BRO-025	Italy	1	0.77	0.019	0.179	0.032
2	BRO-026	Italy	1	0.91	0.032	0.04	0.018
3	BRO-027	Italy	1	0.672	0.024	0.245	0.059
4	BRO-028	Italy	1	0.853	0.073	0.057	0.018
5	BRO-029	Italy	1	0.9	0.023	0.07	0.007
6	BRO-030	Italy	1	0.838	0.092	0.043	0.028
7	BRO-127	Japan	2	0.17	0.461	0.321	0.049
Caixin (ssp. *parachinensis*)							
8	BRO-103	Indonesia	2	0.009	0.771	0.213	0.006
9	PC-078	Netherlands	2	0.011	0.557	0.393	0.039
Chinese Cabbage (ssp. *pekinensis*)							
10	CC-048	Soviet Union	2	0.017	0.477	0.476	0.03
11	CC-049	Netherlands	4	0.016	0.106	0.869	0.009
12	CC-057	China	4	0.012	0.282	0.701	0.005
13	CC-058	Czech Republic	4	0.006	0.112	0.878	0.004
14	CC-059	Korea	4	0.007	0.024	0.888	0.082
15	CC-060	China	4	0.08	0.093	0.821	0.006
16	CC-061	Yugoslavia	2	0.008	0.605	0.378	0.01
17	CC-062	Germany	4	0.035	0.15	0.809	0.007
18	CC-067	Japan	2	0.009	0.517	0.469	0.005
19	CC-068	Bulgaria	4	0.008	0.057	0.931	0.004
20	CC-069	USA	4	0.086	0.316	0.584	0.014
21	CC-070	Korea	4	0.025	0.375	0.584	0.016
22	CC-071	Japan	2	0.057	0.539	0.346	0.058
23	CC-072	China	4	0.06	0.064	0.822	0.054
24	CC-073	China	4	0.009	0.366	0.583	0.043
25	CC-093	China	4	0.034	0.146	0.773	0.047
26	CC-094	Japan	4	0.012	0.201	0.78	0.007
27	CC-095	China	2	0.236	0.526	0.232	0.006
28	CC-112	China	4	0.005	0.426	0.562	0.007
29	CC-113	China	4	0.022	0.249	0.591	0.138
30	CC-114	China	2	0.013	0.539	0.435	0.012
31	CC-125	Korea	4	0.011	0.391	0.543	0.055
32	CC-140	Japan	4	0.007	0.371	0.616	0.006
33	CC-141	Japan	4	0.015	0.027	0.947	0.01
34	CC-142	Japan	4	0.006	0.018	0.931	0.045
35	CC-150	China	4	0.018	0.389	0.572	0.021
36	CC-153	China	2	0.017	0.604	0.366	0.012
37	CC-156	China	4	0.077	0.195	0.66	0.067
38	CC-158	China	4	0.02	0.262	0.653	0.065
39	CC-160	China	4	0.009	0.257	0.724	0.01
40	CC-161	China	4	0.008	0.367	0.617	0.009

Table 1 (continued)

41	CC-163	China	2	0.013	0.508	0.409	0.07
42	CC-165	China	4	0.012	0.425	0.551	0.012
43	CC-167	China	4	0.01	0.234	0.713	0.043
44	CC-168	China	4	0.045	0.174	0.766	0.015
45	CC-169	China	4	0.009	0.391	0.594	0.007
46	Kenshin	Korea	4	0.007	0.256	0.711	0.026
47	Sumiko F1 (CC1)	Company	4	0.01	0.01	0.973	0.007
48	Manoko F1 (CC2)	Company	4	0.032	0.136	0.827	0.004
49	niZ12-42 (CC3)	Company	4	0.007	0.018	0.973	0.003
50	Optiko F1 (CC4)	Company	4	0.014	0.098	0.87	0.018
51	Bilko F1 (CC5)	Company	4	0.005	0.013	0.975	0.007
52	Morilloxstorido F1 (CC6)	Company	4	0.005	0.019	0.949	0.026
53	Vitimo (CC7)	Company	4	0.005	0.008	0.984	0.004
54	Nikko F1 (CC8)	Company	4	0.007	0.007	0.982	0.003
55	Winter pride (CC9)	Company	4	0.004	0.011	0.981	0.004
56	Sambok rapids (CC10)	Company	4	0.011	0.093	0.884	0.012
57	Tropic emperor (CC11)	Company	4	0.005	0.061	0.918	0.016
58	Bulam plus (CC12)	Company	4	0.027	0.01	0.959	0.004
59	Sun green (CC13)	Company	4	0.024	0.028	0.921	0.027
60	Gold leaf (CC14)	Company	4	0.004	0.013	0.973	0.01
61	m1 (CC15)	Company	4	0.019	0.106	0.862	0.013
62	m2 (CC16)	Company	4	0.005	0.01	0.981	0.003
63	m3 (CC17)	Company	4	0.006	0.107	0.883	0.005
64	m4 (CC18)	Company	4	0.01	0.078	0.909	0.003
65	m5 (CC19)	Company	4	0.052	0.02	0.923	0.005
Turnip (ssp. *rapa*)							
66	FT-001	Netherlands	1	0.93	0.025	0.031	0.014
67	FT-002	UK	1	0.857	0.074	0.029	0.04
68	FT-003	Netherlands	1	0.893	0.043	0.059	0.005
69	FT-004	Denmark	1	0.839	0.063	0.069	0.029
70	FT-005	Germany	1	0.944	0.018	0.033	0.006
71	FT-047	Soviet Union	1	0.723	0.238	0.027	0.012
72	FT-051	Soviet Union	2	0.308	0.437	0.048	0.207
73	FT-056	France	2	0.247	0.355	0.338	0.06
74	FT-086	Pakistan	1	0.477	0.14	0.243	0.141
75	FT-088	Netherlands	1	0.602	0.297	0.076	0.026
76	FT-097	Germany	1	0.486	0.182	0.317	0.015
77	VT-006	India	1	0.495	0.264	0.069	0.173
78	VT-007	Soviet Union	1	0.832	0.017	0.109	0.042
79	VT-008	India	1	0.709	0.089	0.132	0.07
80	VT-009	Japan	2	0.142	0.801	0.026	0.032
81	VT-010	Hungary	1	0.818	0.052	0.127	0.003
82	VT-011	Soviet Union	1	0.435	0.492	0.056	0.017
83	VT-012	Japan	2	0.089	0.64	0.061	0.21
84	VT-013	Japan	2	0.212	0.764	0.012	0.012
85	VT-014	Italy	1	0.912	0.031	0.048	0.01
86	VT-015	Italy	1	0.803	0.164	0.016	0.017
87	VT-017	Netherlands	1	0.834	0.034	0.047	0.085
88	VT-018	Netherlands	1	0.752	0.071	0.173	0.005
89	VT-044	Soviet Union	1	0.835	0.021	0.107	0.037

Table 1 (continued)

90	VT-045	Italy	1	0.892	0.031	0.065	0.013
91	VT-052	Netherlands	1	0.765	0.101	0.076	0.057
92	VT-053	Germany	1	0.875	0.097	0.016	0.012
93	VT-089	France	1	0.828	0.126	0.039	0.006
94	VT-090	France	1	0.822	0.161	0.009	0.007
95	VT-091	United Kingdom	1	0.721	0.253	0.016	0.011
96	VT-092	Netherlands	1	0.884	0.051	0.055	0.009
97	VT-115	Japan	2	0.117	0.708	0.143	0.032
98	VT-116	Japan	2	0.052	0.655	0.216	0.076
99	VT-117	Japan	2	0.112	0.675	0.069	0.144
100	VT-119	Netherlands	1	0.897	0.078	0.012	0.013
101	VT-120	Netherlands	1	0.717	0.195	0.04	0.047
102	VT-123	Japan	2	0.076	0.727	0.045	0.152
103	VT-137	Uzbekistan	2	0.414	0.532	0.025	0.028
104	Turnip primera (T1)	Company	1	0.636	0.194	0.163	0.007
105	Turnip oasis (T2)	Company	2	0.285	0.663	0.032	0.019
106	Turnip natsu komachi (T3)	Company	2	0.312	0.412	0.071	0.204
107	natu-haturei (T4)	Company	2	0.225	0.442	0.108	0.225
108	kt-189 (T5)	Company	2	0.197	0.554	0.043	0.206
Pak Choi (ssp. *chinensis*)							
109	PC-022	Netherlands	2	0.139	0.37	0.365	0.126
110	PC-023	China	2	0.022	0.673	0.223	0.082
111	PC-076	China	2	0.031	0.515	0.335	0.119
112	PC-099	China	2	0.08	0.683	0.04	0.197
113	PC-101	China	2	0.01	0.674	0.313	0.003
114	PC-107	Hong Kong	2	0.009	0.948	0.023	0.021
115	PC-171	China	2	0.019	0.815	0.159	0.007
116	PC-172	China	2	0.01	0.872	0.074	0.044
117	PC-173	China	2	0.01	0.86	0.122	0.008
118	PC-177	China	2	0.088	0.858	0.041	0.012
119	PC-183	China	2	0.092	0.882	0.018	0.008
120	PC-184	China	2	0.011	0.904	0.08	0.004
121	PC-185	China	2	0.014	0.926	0.019	0.041
122	PC-186	China	2	0.006	0.955	0.011	0.028
123	PC-187	China	2	0.008	0.866	0.108	0.018
124	PC-189	China	2	0.005	0.915	0.075	0.005
125	PC-191	China	2	0.032	0.806	0.154	0.008
126	PC-193	China	2	0.014	0.945	0.036	0.005
127	L58	Unknown	2	0.009	0.761	0.212	0.019
128	Green fortune (PC1)	Company	2	0.029	0.858	0.063	0.051
129	White (PC2)	Company	2	0.006	0.938	0.047	0.008
130	Misome (PC3)	Company	2	0.067	0.861	0.065	0.007
131	Tatsoi (PC4)	Company	2	0.028	0.866	0.028	0.078
Neep greens (ssp. *perviridis*)							
132	KOM-041	Japan	2	0.075	0.802	0.06	0.064
133	KOM-118	Japan	2	0.11	0.45	0.362	0.079
134	TG-129	Japan	2	0.173	0.792	0.022	0.013
135	TG-131	Japan	2	0.163	0.636	0.168	0.033
Mizuna (ssp. *nipposinica*)							
136	MIZ-019	Netherlands	2	0.129	0.699	0.09	0.082

Table 1 (continued)

137	MIZ-128	Japan	2	0.204	0.659	0.043	0.094
138	Mizuna (K1)	Company	2	0.197	0.747	0.017	0.039
139	Mibuna (K2)	Company	2	0.053	0.539	0.152	0.256
140	Green boy (K3)	Company	2	0.057	0.536	0.329	0.079
Turnip rape (ssp. *oleifera*)							
141	OR-209	China	2	0.173	0.736	0.031	0.06
142	OR-210	China	2	0.066	0.737	0.19	0.007
143	OR-211	China	2	0.103	0.813	0.028	0.056
144	OR-213	China	2	0.172	0.5	0.189	0.139
145	OR-216	China	2	0.188	0.63	0.167	0.015
146	OR-219	China	2	0.14	0.676	0.112	0.072
147	SO-031	USA	1	0.445	0.274	0.024	0.257
148	SO-032	India	2	0.107	0.447	0.029	0.416
149	SO-034	Bangladesh	3	0.021	0.01	0.015	0.954
150	SO-035	Bangladesh	3	0.004	0.006	0.006	0.984
151	SO-037	Bangladesh	3	0.181	0.375	0.047	0.397
152	SO-038	Germany	3	0.149	0.285	0.032	0.534
153	SO-039	Bangladesh	3	0.004	0.009	0.005	0.982
154	SO-040	Canada	1	0.783	0.176	0.032	0.009
155	WO-024	Sweden	1	0.886	0.047	0.05	0.017
156	WO-080	Pakistan	1	0.315	0.262	0.247	0.176
157	WO-081	Pakistan	1	0.348	0.219	0.255	0.177
158	WO-083	Pakistan	1	0.301	0.092	0.334	0.273
159	WO-084	Pakistan	2	0.164	0.401	0.28	0.155
160	WO-085	Pakistan	2	0.394	0.4	0.022	0.184
161	WO-087	Pakistan	2	0.343	0.427	0.044	0.185
162	RC-144	USA	3	0.048	0.334	0.03	0.589
163	RO18	United Kingdom	3	0.006	0.014	0.017	0.964
Yello sarson (ssp. *tricolaris*)							
164	YS-033	Germany	3	0.006	0.009	0.009	0.976
165	YS-143	USA	3	0.003	0.007	0.005	0.985
Wutacai (ssp. *narinosa*)							
166	PC-105	China	2	0.03	0.895	0.063	0.013
167	cwu56		2	0.009	0.899	0.081	0.01
Zai Caitai (ssp. *chinensis* var. *purpurea* Bailey)							
168	zct62		2	0.073	0.757	0.031	0.139

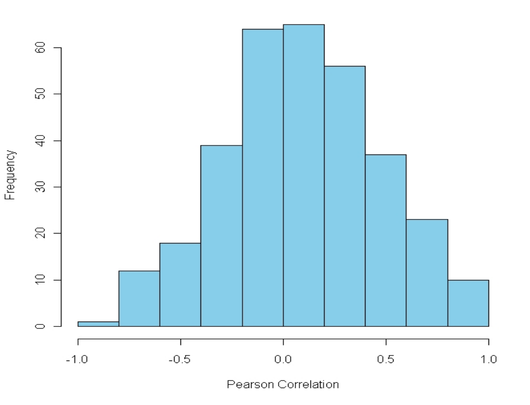


Fig. 3: Frequency distribution of correlation among morphological traits.

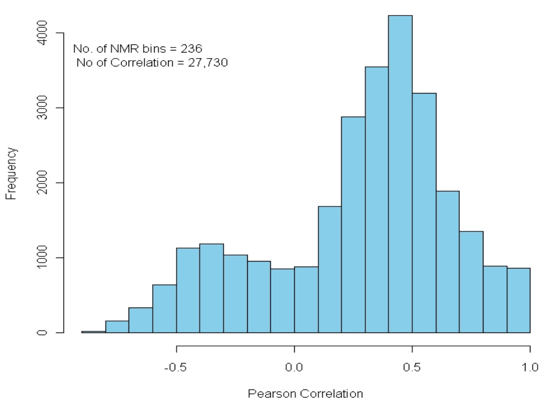


Fig. 10: Frequency distribution of correlation among NMR bins

Table 4: Loading values of morphological traits (Vernalized) on first three PCs

Variables	PC 1	PC 2	PC 3
PL	0.95	0.07	0.10
LP	0.90	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
PP	0.41	-0.46	0.013
Lbl	0.39	0.65	0.36
pI	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	-0.20
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	-0.10	0.05
PH	-0.80	0.33	0.15

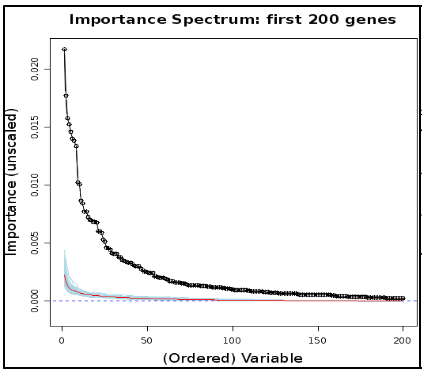


Fig. 14: Variable importance spectrum plot of the top 200 markers in RF classification of accessions into STRUCTURE four classes

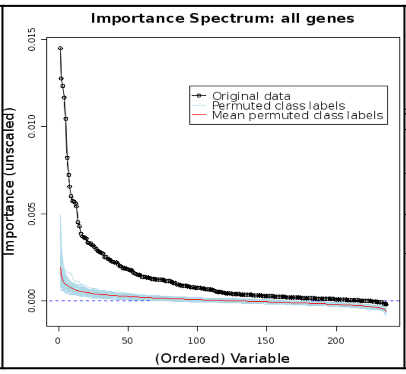


Fig. 16: Variable importance spectrum plot of all NMR bins (236) in RF classification of accessions into STRUCTURE four classes

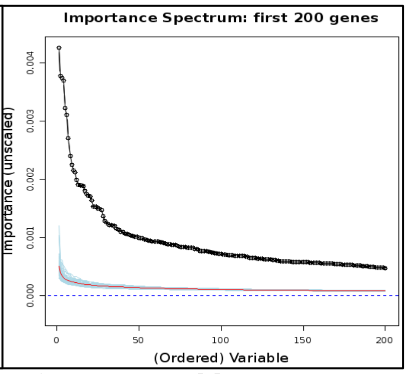


Fig. 18: Variable importance spectrum plot of the top 200 LC-MS peaks in RF classification of accessions into four STRUCTURE classes

Table 9: List of RF (classification) selected markers from comparisons made among all four STRUCTURE groups, CC versus PC, and European versus Asian turnips of a *B. rapa* core collection genotyping with 359 AFLP and MYB markers

S.N.	Markers	Map position	S.N.	Markers	Map positon
1	Alu-M175.5[1,c,e]	-	31	pTAmCAC-165.6[c]	Chr.10: 78 cM
2	Alu-M258.1[1,c,e]	-	32	pTAmCAC-171.5[1,c]	-
3	Alu-M357.1[e]	Chro.1: 60.3 cM	33	pTAmCAC-192.8[1,c,g]	-
4	Alu-M380.0[b,d,f]	Chro.3: 6.1 cM	34	pTAmCAC-244.2[1]	Chr.10: 76.1 cM
5	Alu-M401.8[1,a,c]	Chr.7: 83 cM	35	pTAmCAC-273.0[1,e]	Chr.2: 84.5 cM
6	Alu-M487.5[1,e]	Chr.1: 61.6 cM	36	pTAmCAC-290.9[b]	Chr.3: 1.2 cM
7	Hae-M199.0[1,c]	-	37	pTAmCAC-293.0[1,e]	Chr.3: 0.6 cM
8	Hae-M202.2[1,a,c]	Chr.7: 83.5 cM	38	pTAmCAC-315.9[1,c,e]	-
9	Hae-M341.0[1,c,e]	-	39	pTAmCAC-326.7[g]	-
10	Hae-M356.5[e]	-	40	pTAmCAC-63.6[b]	Chr.5: 77 cM
11	Hae-M373.7[d]	Chr.7: 15.4 cM	41	pTAmCAC-94.6[1,a,h]	-
12	Hae-M458.5[1]	-	42	pTAmCAT-170.3[1,a]	-
13	Mse-575.0[1,a,h]	Chr.7: 40.2 cM	43	pTAmCAT-173.8[1,c,e,g]	-
14	Mse-M197.3[1,c]	-	44	pTAmCAT-175.7[1,e]	-
15	Mse-M232.6[1,c]	-	45	pTAmCAT-199.6[1,c,e]	-
16	Mse-M242.0[a]	-	46	pTAmCAT-209.9[1,c]	-
17	Mse-M308.8[1,c,e]	-	47	pTAmCAT-240.3[e]	Chr.5: 54.2 cM
18	Mse-M354.6[b]	-	48	pTAmCAT-243.4[1,e]	-
19	Mse-M356.4[1,c,e]	-	49	pTAmCAT-252.2[e]	Chr.10: 77.8 cM
20	Mse-M431.5[b]	-	50	pTAmCAT-278.7[1]	-
21	Mse-M455.7[1,c,e]	-	51	pTAmCAT-282.5[1,a,h]	-
22	pGGmCAA-127.7[c]	-	52	pTAmCAT-313.10[1,c,e]	Chr.1: 63.7 cM
23	pGGmCAA-152.2[1,a,h]	-	53	pTAmCAT-334.2[d]	Chr.8: 78 cM
24	pGGmCAA-165.8[1]	Chr.2: 67.8 cM	54	pTAmCAT-336.9[b]	-
25	pGGmCAA-181.5[1,c]	-	55	Rsa-M124.4[e]	-
26	pGGmCAA-217.2[1,e]	-	56	Rsa-M241.6[1,c,e,g]	-
27	pGGmCAA-224.3[c]	Chr.6: 46.9 cM	57	Rsa-M268.8[b]	Chr.10: 78.7 cM
28	pGGmCAA-359.4[b]	Chr.8: 34.5 cM	58	Rsa-M346.4[e]	-
29	pTAmCAC-112.3[1,c,e]	-	59	Rsa-M489.6[1,f]	Chr.7: 84 cM
30	pTAmCAC-157.8[1,c,a]	-	-	-	-

[1]-all 4 STR groups, [a]-1 vs. 2, [b]-1 vs. 3, [c]- 1 vs. 4, [d]-2 vs. 3, [e]-2 vs. 4, [f]-3 vs. 4, [g]-CC vs. PC, [h]-EU vs. Asian turnip

Table 11: List of RF (classification) selected NMR bins from comparisons made on all four STRUCTURE groups, CC versus PC, and European versus Asian turnips of a *B. rapa* core collection

S.N.	NMR bins	S.N.	NMR bins	S.N.	NMR bins
1	nmr-0.88-66[1c]	19	nmr-2.28-63[c]	37	nmr-5.2-57[f]
2	nmr-0.92-66[c]	20	nmr-2.36-63[1c]	38	nmr-5.24-56[1ce]
3	nmr-0.96-66[1c]	21	nmr-2.4-63[a]	39	nmr-5.84-55[1a]
4	nmr-1-66[c]	22	nmr-2.48-63[h]	40	nmr-5.88-55[1c]
5	nmr-1.08-66[1c]	23	nmr-2.64-62[f]	41	nmr-5.92-55[1h]
6	nmr-1.24-65[1cg]	24	nmr-2.68-62[1ac]	42	nmr-5.96-55[1]
7	nmr-1.28-65[1ac]	25	nmr-2.8-62[1]	43	nmr-6.16-54[b]
8	nmr-1.32-65[1ce]	26	nmr-2.84-62[1ac]	44	nmr-6.28-54[bd]
9	nmr-1.36-65[1]	27	nmr-2.92-62[c]	45	nmr-6.52-54[1]
10	nmr-1.48-65[1ceg]	28	nmr-3.36-61[c]	46	nmr-7-53[1ac]
11	nmr-1.52-65[1c]	29	nmr-3.4-61[c]	47	nmr-7.4-52[h]
12	nmr-1.56-65[1ac]	30	nmr-3.84-56[d]	48	nmr-7.64-51[1a]
13	nmr-1.6-65[1c]	31	nmr-4.04-59[d]	49	nmr-7.68-51[1a]
14	nmr-1.76-64[1c]	32	nmr-4.56-58[c]	50	nmr-7.76-51[1]
15	nmr-1.8-64[1ach]	33	nmr-4.6-58[c]	51	nmr-7.84-51[1a]
16	nmr-2.08-64[1c]	34	nmr-5.04-57[1ce]	52	nmr-8.08-50[h]
17	nmr-2.12-63[ac]	35	nmr-5.08-57[h]	53	nmr-8.2-50[1ach]
18	nmr-2.16-63[1ac]	36	nmr-5.12-57[ch]	54	nmr-9.84-46[h]

[1]-all 4 STR groups, [a]-1 vs. 2, [b]-1 vs. 3, [c]- 1 vs. 4, [d]-2 vs. 3, [e]-2 vs. 4, [f]-3 vs. 4, [g]-CC vs. PC, [h]-EU vs. Asian turnip

Table 13: List of RF (classification) selected LC-MS peaks from comparisons made on all four STRUCTURE classes, CC versus PC, and European versus Asian turnip of a *B. rapa* core collection

S. N.	Centrotype_mass _scan No.	Chemical formula	Chemical compounds
1	3435_479_161[d]		
2	3384_259_159[d]		
3	4108_360_404[1]	$C_{10}H_{19}NO_9S_2$	Isopropyl glucosinolate
4	4108_361_402[1]	$C_{10}H_{19}NO_9S_2$	Isopropyl glucosinolate
5	4128_402_417[h]		
6	4128_403_416[h]		
7	4128_404_415[h]		
8	4128_470_416[h]		
9	4294_454_489[h]		
10	4294_667_489[h]		
11	4373_210_501[h]		
12	4373_792_499[h]		
13	4492_351_520[e]		
14	4492_352_520[e]		
15	4492_368_520[1,e]	$C_{16}H_{18}O_9$	caffeoylquinic acid (isotope 353)
16	4492_369_519[e]		
17	4492_385_522[1,e]		
18	4492_453_519[h]		
19	4492_705_520[e]		
20	4492_706_518[e]		
21	4492_707_519[e]		
22	4492_708_520[e]		
23	4492_709_519[e]		
24	4492_721_521[1,e]		
25	4492_730_520[e]		
26	4492_731_520[e]		
27	4492_761_521[e]		
28	4504_440_518[h]		
29	4577_517_581[1]		
30	4815_750_689[1]	$C_{11}H_{21}NO_9S_2$	Methylpropyl glucosinolate (isotope)
31	4838_405_693[1,f]		
32	4867_502_700[1,a,h]	$C_{24}H_{38}O_{11}$	4,7-Megastigmadiene-3,9-diol, 3-Ketone, 9-O-[?-D-apiofuranosyl-(1?2)-?-D-glucopyranoside]
33	4932_691_714[e]		
34	4932_692_718[1]		
35	4990_1479_738[1]		
36	5028_1124_746[g]		
37	5028_1125_746[g]		
38	5028_1126_746[1]	$C_{49}H_{58}O_{30}$	Quercetin 3-(2-feruloylsophoroside) 7-diglucoside
39	5028_1128_746[1,g]		
40	5028_1139_746[1,g]		
41	5028_1140_746[1,g]		
42	5028_1142_746[1]		
43	5028_1193_746[g]		
44	5028_1215_746[g]		
45	5028_1216_746[1,g]		
46	5028_1217_745[1]		
47	5028_562_746[g]		
48	5028_563_746[g]		
49	5236_1191_788[1,e]	$C_{16}H_{18}O_9$	chlorogenic acid
50	5236_1192_787[1]	$C_{16}H_{18}O_9$	chlorogenic acid
51	5236_353_785[1,e]	$C_{16}H_{18}O_9$	chlorogenic acid
52	5236_354_785[1,e]	$C_{16}H_{18}O_9$	chlorogenic acid
53	5236_355_784[1,e]	$C_{16}H_{18}O_9$	chlorogenic acid
54	5236_443_785[1,e]	$C_{16}H_{18}O_9$	chlorogenic acid
55	5236_707_785[1,e]	$C_{16}H_{18}O_9$	chlorogenic acid
56	5236_708_784[1,e]	$C_{16}H_{18}O_9$	chlorogenic acid
57	5400_501_797[1,a]	$C_{24}H_{38}O_{11}$	a glucopyranoside-derivative (4,7-Megastigmadiene-3,9-diol, 3-Ketone, 9-O-[α-L-arabinopyranosyl-(1->6)-β-D-glucopyranoside])
58	5400_502_796[1,a,h]		
59	5248_1193_786[1]	$C_{48}H_{56}O_{29}$	unknown
60	5248_1272_785[1,e]	$C_{48}H_{56}O_{29}$	kaempferolcaffeoyltetraglucoside
61	5248_1273_785[1,e]	$C_{48}H_{56}O_{29}$	unknown
62	5248_1274_785[1]	$C_{48}H_{56}O_{29}$	unknown
63	5441_433_819[1]		
64	5441_434_820[1]		
65	5600_357_849[1]		
66	6124_173_920[1,e]		fragment
67	6124_337_920[1,e]	$C_{16}H_{18}O_8$	Coumaroylquinic acid I
68	6124_338_920[1,e]		
69	6124_339_919[1,e]		
70	6124_359_920[1,e,f]		
71	6124_675_920[1,e]		
72	6481_173_971[1,e]		fragment
73	6481_337_971[e]		
74	6481_338_971[1]	$C_{16}H_{18}O_8$	Coumaroylquinic acid II
75	6481_339_972[1,e]		
76	6481_359_972[1,e]		
77	6481_405_971[1,e]		
78	6481_675_970[e]		
79	6481_697_969[1,e]		
80	7622_291_1278[1]		
81	7673_431_1290[1]		
82	7746_115_1312[1,e]		
83	6573_431_1014[h]		
84	6736_214_1054[h]		
85	7033_925_1092[h]		
86	7448_658_1229[g]		
87	7550_1287_1257[g]		
88	7550_643_1257[g]		
89	7606_445_1274[h]		
90	7746_279_1312[1]		Malic acid, O-(4-Hydroxycinnamoyl)?
91	7781_422_1321[h]		
92	7781_425_1318[h]		
93	7784_643_1319[1,c]		
94	8845_387_1482[1]		
95	8149_308_1378[b]		
96	8253_989_1388[b]		
97	8515_649_1410[b]		
98	8515_650_1410[b]		
99	9153_429_1611[1,h]		
100	9223_415_1631[1]		
101	9244_436_1640[h]		
102	9399_402_1705[1,e]	$C_{13}H_{25}NO_9S_2$	Hexyl glucosinolate 2
103	9640_479_1787[1]		
104	9722_462_1829[1,e]		
105	9722_463_1829[1,e]		
106	10224_748_2357[h]		
107	10038_418_2103[h]		
108	A190_388_1036[1]		
109	A194_501_706[1,h]		
110	A42_431_1949[1]		
111	A201_504_724[h]		
112	A220_221_943[d]		
113	A45_433_1247[h]		
114	A682_609_2068[b]		
115	A62_449_632[1]		
116	A86_484_2174[1,h]		

[1]- 4 STR groups, [a]-1 vs. 2, [b]-1 vs. 3, [c]- 1 vs. 4, [d]-2 vs. 3, [e]-2 vs. 4, [f]-3 vs. 4, [g]-CC vs. PC, [h]-EU vs. Asian turnip.

Note: Centrotypes coded by an alphabet "A" followed by numbers, such as A42, A62 and so on meant peaks with no centrotype.

Table 15: List of selected LC-MS peaks for SPAD traits by random forest regression based on percentage decrease in MSE

SPAD vernalized condition 2007		SPAD non-vernalized condition 2006		SPAD non-vernalized condition 2007	
Peaks	%IncMSE	Peaks	%IncMSE	Peaks	%IncMSE
10099_223_2165	3.891396	10043_615_2104	4.30E+00	10074_919_2140	8.713213
10208_569_2313	3.543923	10074_919_2140	1.87E+01	10074_920_2140	6.124533
4373_408_495	5.174491	10074_920_2140	1.47E+01	10074_951_2139	8.094092
5210_453_783	8.889851	10074_951_2139	1.04E+01	10094_761_2162	6.98667
5268_1409_785	3.77244	10094_761_2162	2.12E+01	10094_762_2161	7.097879
5441_407_816	16.35216	10094_762_2161	1.89E+01	3435_542_159	5.665481
6317_824_943	3.963386	3498_551_160	5.33E+00	3757_192_173	6.556786
6487_407_971	5.361152	3583_300_162	4.79E+00	4444_749_517	5.221399
6487_523_971	5.715385	3874_436_206	5.04E+00	4601_483_614	5.150911
6528_523_994	4.297159	3930_436_223	4.15E+00	4620_343_621	9.874111
6756_668_1058	4.770999	**5441_407_816**	4.06E+00	5066_323_747	5.647579
6967_935_1080	3.513804	6120_715_922	4.59E+00	5865_179_887	5.851048
7209_536_1159	3.897753	***6317_385_942***	4.04E+00	6236_1081_925	5.16328
7599_697_1272	3.664648	6456_393_962	1.41E+01	6882_1015_1079	6.683104
7599_777_1271	11.22918	6894_439_1077	5.96E+00	6882_359_1076	11.90077
7918_1095_1338	3.725454	7119_455_1122	6.67E+00	6882_666_1076	6.57919
8156_475_1377	4.127084	7150_433_1134	8.78E+00	6882_697_1075	6.133402
9236_1157_1641	4.2973	7276_803_1181	5.41E+00	7252_378_1176	5.655705
9244_705_1644	3.920757	7360_770_1209	5.20E+00	7276_801_1180	5.391558
9273_837_1659	3.640021	7497_311_1239	4.24E+00	7276_802_1180	5.477931
9396_746_1699	3.459316	7550_1287_1257	6.32E+00	7276_803_1181	6.010609
9535_543_1740	4.743934	8149_454_1376	4.66E+00	7360_769_1209	9.454385
9834_634_1891	3.939378	8340_734_1394	1.22E+01	7360_770_1209	11.32978
9838_661_1891	5.092471	8421_843_1409	4.05E+00	7550_1287_1257	7.55464
A201_504_724	4.032738	9372_223_1691	4.37E+00	7550_1288_1257	5.229669
A298_378_1894	3.879301	9492_415_1738	5.48E+00	7683_345_1296	5.649296
A478_454_783	5.095435	***9834_457_1891***	5.10E+00	8340_734_1394	11.79727
A60_448_872	3.504676	9997_413_2056	1.45E+01	9034_223_1549	8.137505
A61_449_173	5.008768	A211_518_1036	4.29E+00	9492_415_1738	8.34907
A73_468_873	4.482524	***A60_448_872***	7.93E+00	***9834_457_1891***	6.425288

Note: Color represents the matching of LC-MS peaks

Fig. 20: Box plots of RF selected 64 LC-MS peaks on classification of all accessions into four STRUCTURE classes. In box plots, 1 indicate for class 1, 2 for class 2, 3 for class 3 and 4 for class 4, and log (base 2) of peaks intensities were in y-axis.

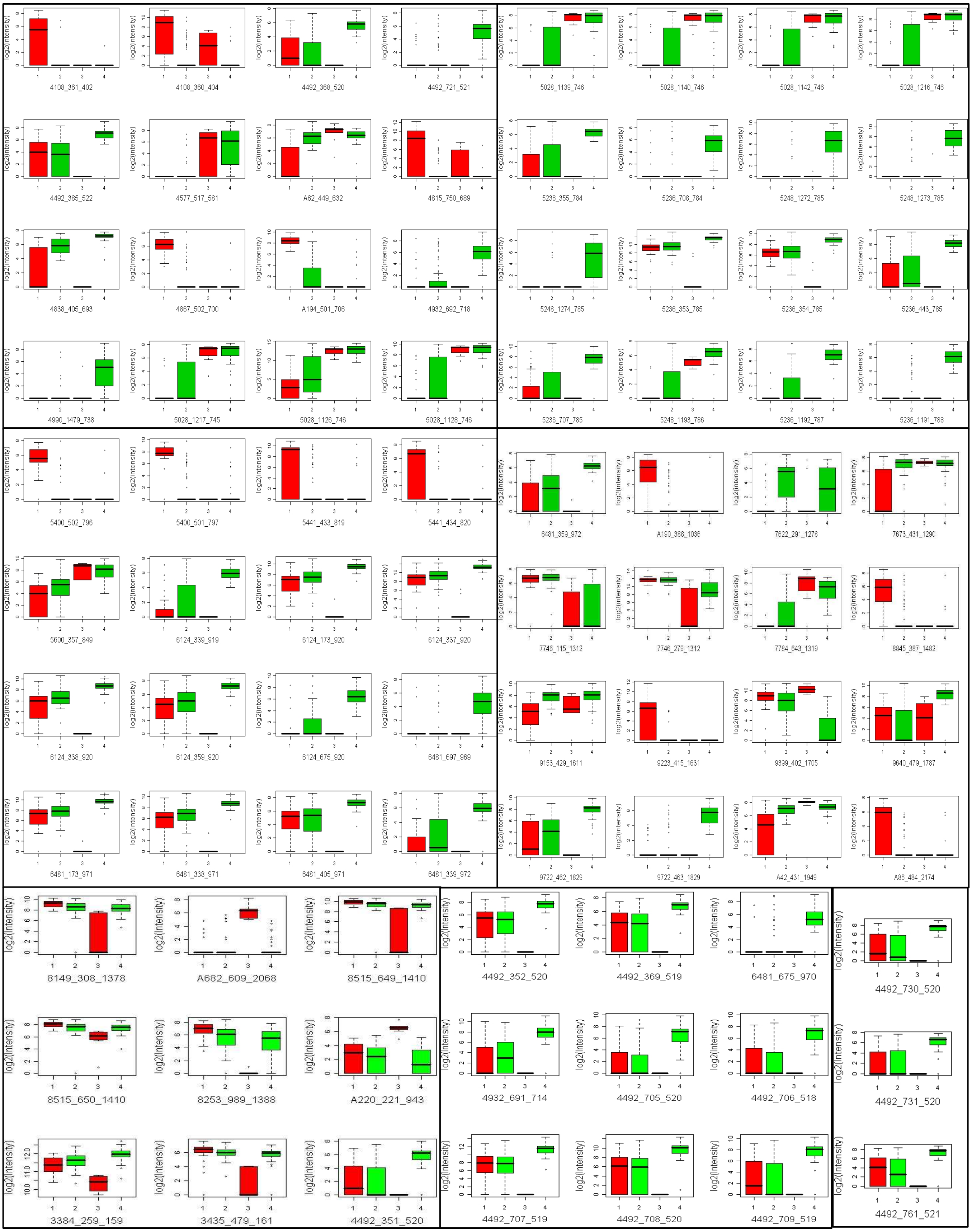


Fig. 21: Comparisons of two small oil groups on the level of the distinguishing LC-MS peaks identified in RF (Classification)

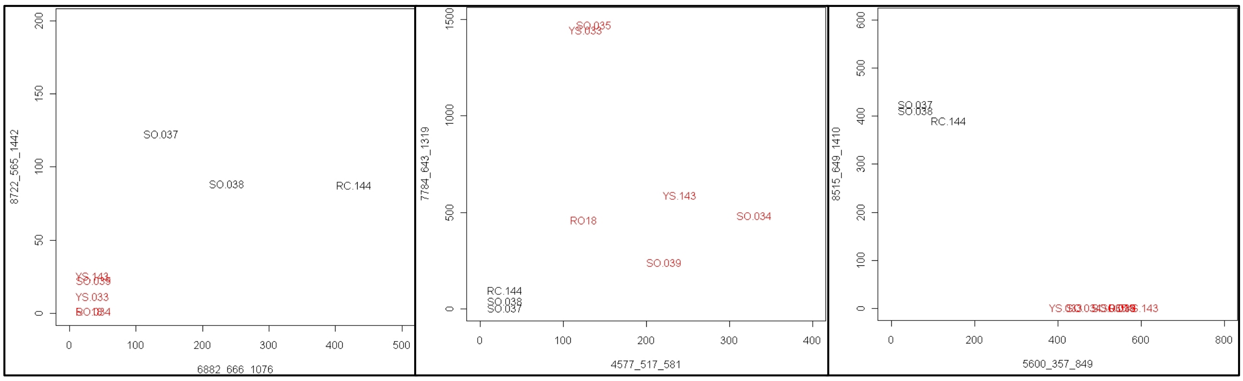
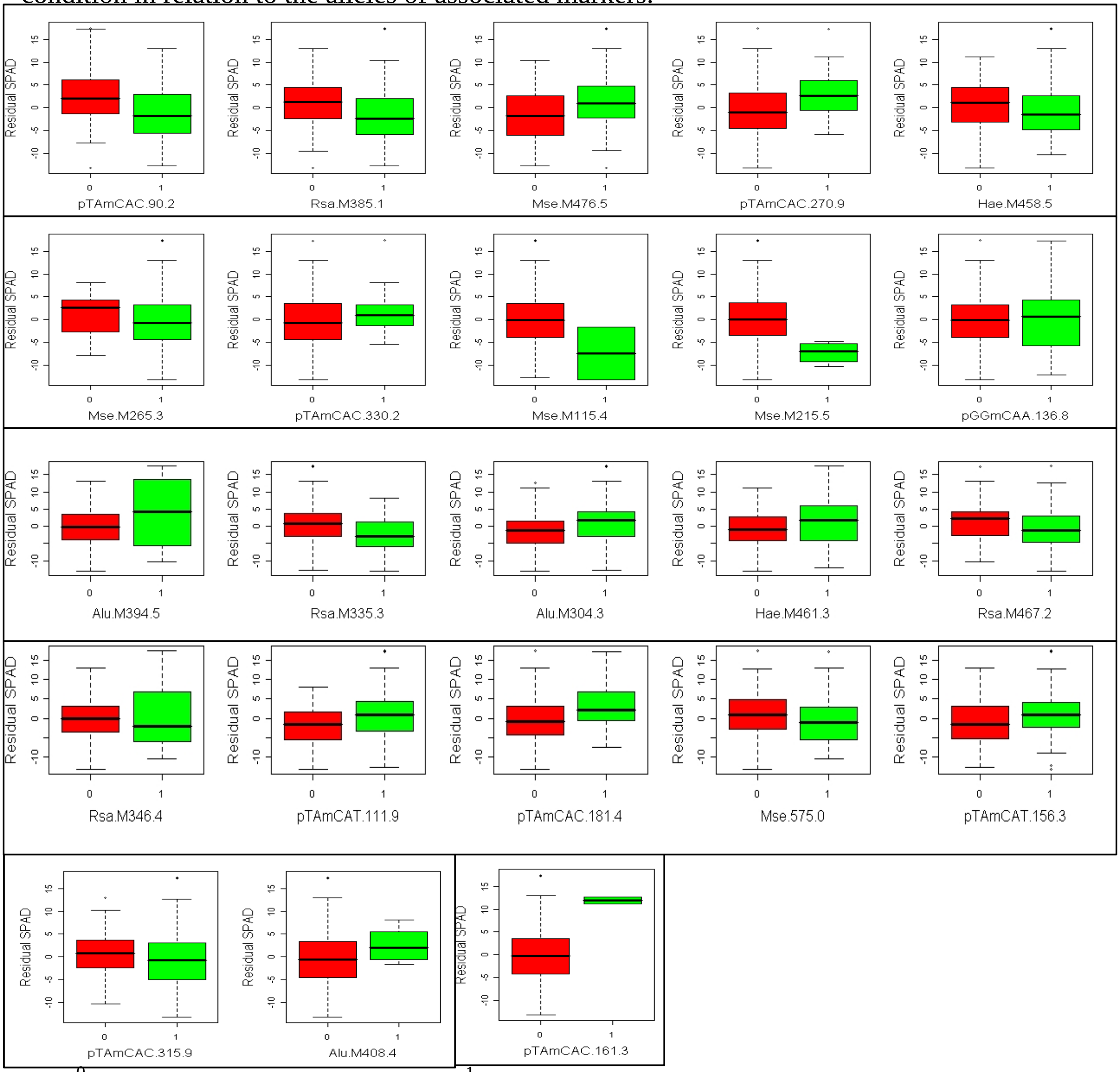
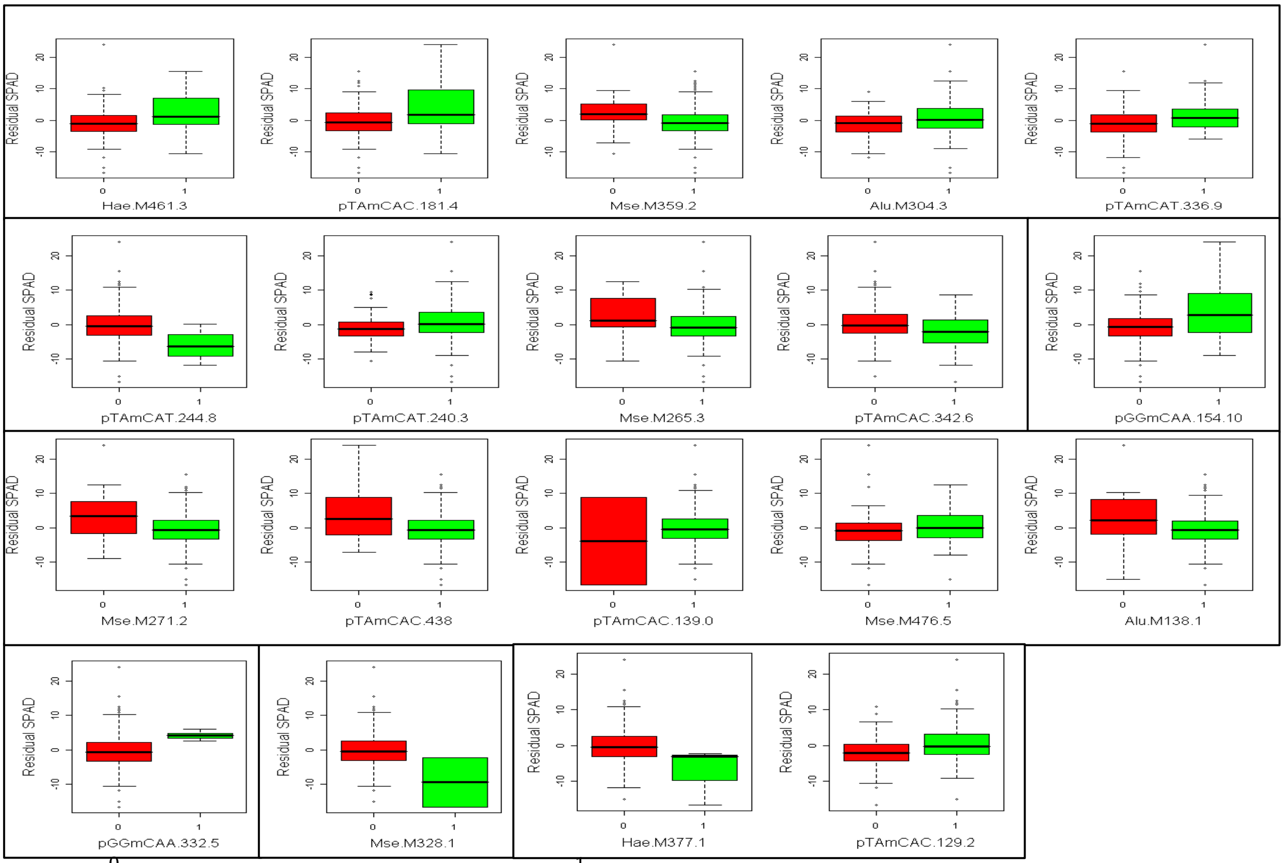


Fig. 26: Box plots showing the level of SPAD (corrected for population structure) of vernalized condition in relation to the alleles of associated markers.



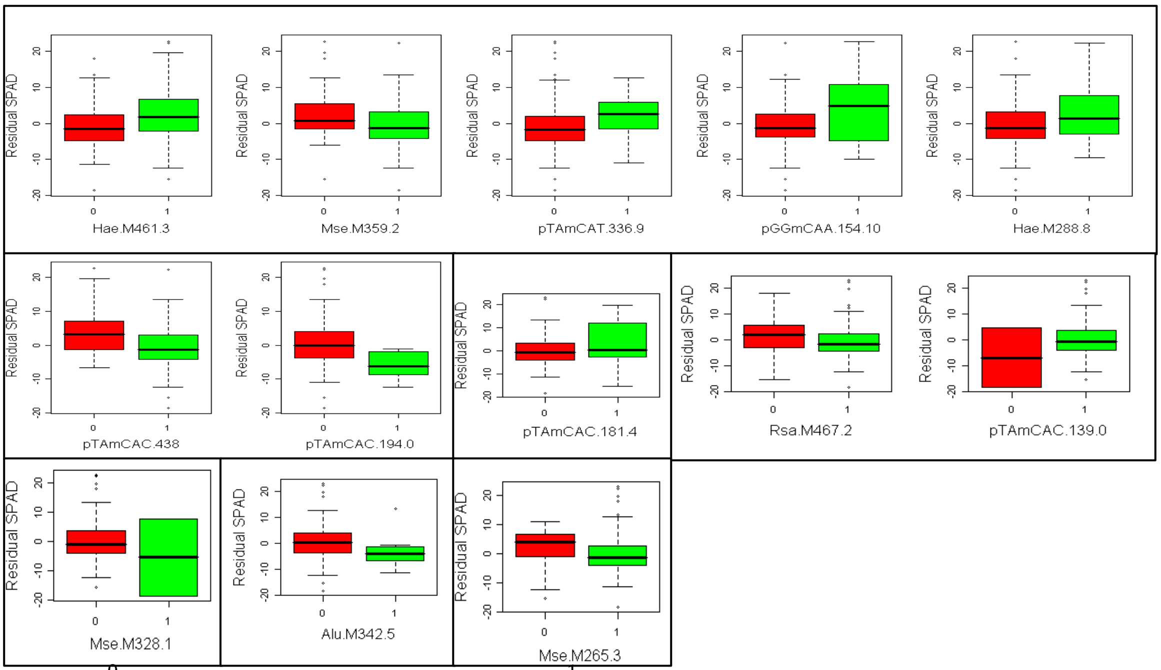
Note: ^{0}-absence of band (red color) and ^{1}-presence of band (green color) of dominant marker. Markers showing allelic differences with SPAD values were shown here.

Fig. 28: Box plots showing the level of SPAD (corrected for population structure) of non-vernalized condition, 2006 in relation to the alleles of associated markers.



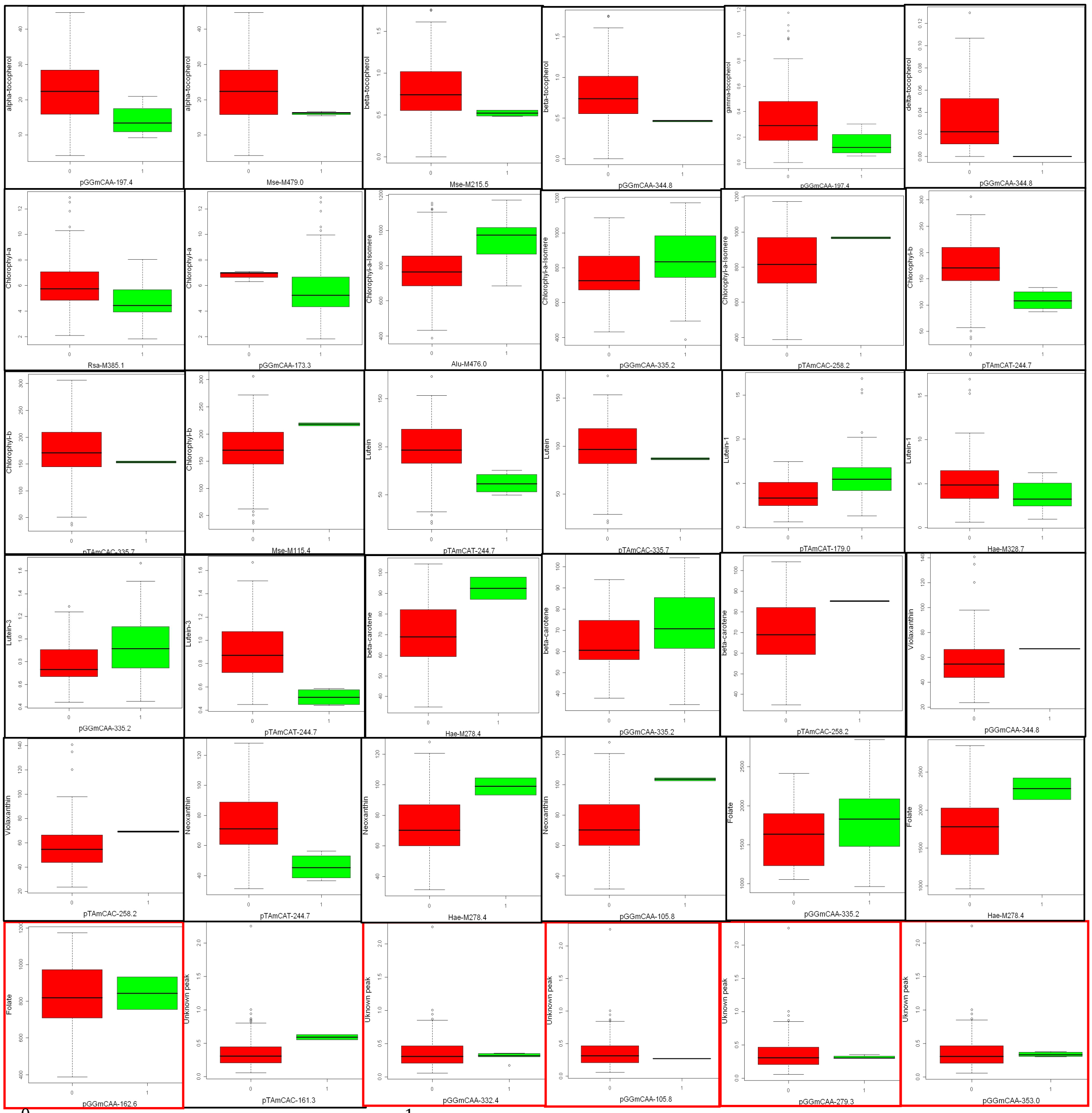
Note: ^{0}-absence of band (red color) and ^{1}-presence of band (green color) of dominant marker. Markers showing allelic differences with SPAD values were shown here.

Fig. 30: Box plots showing the level of SPAD (corrected for population structure) of non-vernalized condition, 2007 in relation to the alleles of associated markers.



Note: ^{0}-absence of band (red color) and ^{1}-presence of band (green color) of dominant marker. Markers showing allelic differences with SPAD values were shown here.

Fig. 34: Box plots showing the level of targeted metabolites in relation to the alleles of the markers



^{0}-absence of band (red colour) and ^{1}-presence of band (green colour) of dominant marker in gel electrophoresis. Box plots circled by red color indicate false positive association.