Genetic Architecture of Glucosinolate Variation in *Brassica napus*

Varanya Kittipol

DOCTOR OF PHILOSOPHY

University of York

Biology

September 2019

This work is dedicated to God of my salvation. Only by the grace of God that I could accomplish anything. Thank you LORD, praise your holy name!

'The LORD is my strength and my shield;
My heart trusted in Him, and I am helped;
Therefore my heart greatly rejoices,
And with my song I will praise Him.'[1]

This degree is dedicated with love and honour TO MY MOTHER, whose sacrificial care and love for her children, provided us with invaluable educational opportunities for a brighter future.

[1] Psalm 28:7 (NKJV)

Abstract

Glucosinolates (GSLs) are a group of secondary metabolites prevalent in the important oilseed rape crop (*Brassica napus L.*). The GSL hydrolysis products show diverse bioactivities and thus play significant biological and economical roles in the defence system and nutritional qualities of rapeseed protein meal. Hence, there is an increasing desire to harness the defensive properties of GSLs to improve pest resistance properties in the vegetative tissues while maintaining low GSLs in the seeds for animal feed.

This thesis aims to identify the genetic controls underlying natural GSL variations in the leaves and roots of *B. napus*, and also develop understanding of their connections with seed GSLs. To address these aims, Associative Transcriptomics (AT), was performed on a panel of 288 *B. napus* accessions. AT correlates GSL trait variations to the variations in either gene sequences or gene expression across these accessions to identify highly associated quantitative trait loci for GSL contents.

This thesis provides five key findings. Firstly, the GSL profiles differ extensively between the leaves and roots in both type and amount. Secondly, both the single nucleotide polymorphism and gene expression marker associations identify the *MYB28/HAG1* orthologues on chromosomes A9 and C2 as key regulators for aliphatic GSLs in leaves. Thirdly, the reduced GSL levels in seeds reflect the reduced level of GSLs in leaves, and is due to the genetic variations caused by homoeologous exchanges in the genomic regions containing *Bna.HAG1.A9* and *Bna.HAG1.C2*. Fourthly, AT and differential expression analyses of roots implicate *Bna.HAG3.A3*, an orthologue of *MYB29/HAG3*, as the main controlling factor for root aromatic GSL variations. Lastly, significant relationships exist between different classes of GSLs, suggesting some metabolic cross-talks between pathways. This work improves our understanding of the genetic regulatory of GSL natural variations in *B. napus* that could lead to crop improvement.

Contents

CHAPTER 2 - General Material and Methods 57

List of Tables

List of Figures

List of Accompanying Material

Datasets with large spreadsheets are uploaded with this thesis as zip files. It contains the following datasets:

Spreadsheet 1. Raw and processed glucosinolate data from leaf and root tissues of 288 *B. napus* accessions. This data has been published as Appendix 1 in *Kittipol et al.* (2019b).

Spreadsheet 2. Markers and genomic regions showing single nucleotide polymorphism association with variation for aliphatic GSL content in the leaf tissues. This data has been published as Appendix 9 in *Kittipol et al.* (2019b).

Spreadsheet 3. Top gene expression markers for leaf aliphatic glucosinolates. This data has been published as Appendix 11 in *Kittipol et al.* (2019b).

Spreadsheet 4. Markers and genomic regions showing single nucleotide polymorphism association with variation for aliphatic GSL content in the root tissues. This data has been published as Appendix 14 in *Kittipol et al.* (2019b).

Spreadsheet 5. Root differential expression analysis of top BLAST hit to annotated *A. thaliana* genes. Genes with $\log_2$ fold-change ≥ 4 and stringent significance value $p \leq 1\times10^{-10}$ are shown. This data has been published as Appendix 16 in *Kittipol et al.* (2019b).

Acknowledgements

I am in deep gratitude to God and in awe of the wonderful things He has done in bringing me safely to this point. He has provided all that I need at the time I needed them. He has surrounded me with incredible people who have supported me throughout this journey. The last four years of my PhD have been one of the best moments of my life, thanks to all of you.

Firstly, I would like to thank my supervisor, Prof. Ian Bancroft, for all your advice and guidance with the work. I am really grateful for the support from you when I felt overwhelmed with the amount of work, especially during the manuscript planning and writing. My thanks also go to my co-supervisor, Prof. Simon McQueen-Mason, for all your support that started from my undergraduate years and carried through to my entire PhD years. I also want to thank my TAP members, Dr. Michael Schultze and Prof. Neil Bruce, for their constructive advice and feedback throughout the years. I am also very thankful to my examiner, Prof. Stanislav Kopriva, for the delightful discussion during the *viva* and insightful feedback on this work.

As an oversea student, it was not easy to find a funding body that would support my studies in the UK. I am therefore grateful for the Scholarships for Overseas Students from the University of York for their support on the tuition fee and the Radhika V Sreedhar scholarship fund from the department of Biology which helps support some of the living costs in my final year. More importantly, I am grateful for the job opportunity provided by Ian to work in the Bancroft lab as a part-time lab technician while writing up this thesis. This job has helped support a major part of my living costs in the final year.

The past and present members of the Bancroft lab have played an incredible part in shaping the fun PhD experiences I have had. There have been so much support, so much laughter and so much food in the group. Lihong, thank you for teaching me many things in the lab, for all your care and all the good times. Lenka, thank you for your keenness to listen and for your answers my endless scientific queries, and for all the fun things we did together. Zhesi, thanks for your help with many of the computational sides of the research, and your input really helped me to progress in my analysis. I love all the badminton we have played together over the years. A huge thank you also to Helen Riordan, Roxana Teodor, Natalia Stawniak, Harjeevan Kaur and Aoife Sweeney for all the support and time we have spent together. I have never felt lonely in this PhD journey because of all of you, thank you.

I am blessed to have Yi Li in my life. Li, you are like my UK dad and your family has become my UK family. Thank you for all the support, encouragements and invaluable advice on various topics throughout the years. I am grateful for all the time we have spent together playing badminton, sharing meals and strolling around the beautiful campus – all of these have helped me to de-stress and in turn allow me to push forward with the work. Also, thank you for believing in me more than I believe in myself and for boosting my confidence. I truly appreciate the time you have spent proof-reading this thesis for me. Your feedback and input has helped me a lot!

My special thanks also go to my church family at Calvary Chapel York. I am incredibly blessed to be part of the healthy fellowship that has encouraged me in my walk with the Lord. I would like to express my sincere gratitude to the church, in particular the following brothers and sisters in Christ for all their support, their love and constant prayers: Mike and Helen Salmon, Karina Sáenz, Chiew Kin Yung, Sammy Voong, Erin Gutierrez Baragula, Janet Chu, Su Jane Beh, Stephen Parkins, the ladies Bible group, Jonathan and Naomi Anderson.

Lastly, without my family I would not be here today. I am truly grateful for my mum: for all the hard work she has done for her children, for giving me and my brother a chance to go to good schools and for putting us on this path to study abroad. Thank you, mum, for taking care of us, for your love and for all the support. Thank you kor King for being the most caring big brother I could have ever asked for and for believing in me since the beginning. Kor, thank you for always trying your best to look out for me and protect me in every situation, I'm lucky to be your sister. Many thanks to my stepdad and stepbrothers for caring after my mum and my family through the good and bad times. Thank you, dad, for the financial support in my high school and university years.

I acknowledge that all that happens is in the hands of God, may His name be glorified in all things.

Declaration

I declare that this thesis is a presentation of original work and I am the sole author, except for where otherwise stated. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References. The following works, presented in this thesis, have been published:

- Kittipol V, He Z, Wang L, Doheny-Adams T, Langer S, Bancroft I (2019a) Genetic architecture of glucosinolate variation in *Brassica napus*. J Plant Physiol. doi: 10.1016/j.jplph.2019.06.001

- Kittipol V, He Z, Wang L, Doheny-Adams T, Langer S, Bancroft I (2019b) Data in support of genetic architecture of glucosinolate variations in *Brassica napus*. Data Br. doi: 10.1016/j.dib.2019.104402

Please find the reprints of these publications at the end of the thesis.

The following published material is not directly arisen from the works in this thesis, but had been developed at the beginning of the project and provides the methods for glucosinolate quantification used in this thesis:

- Doheny-Adams T, Redeker K, Kittipol V, Bancroft I, Hartley SE (2017) Development of an efficient glucosinolate extraction method. Plant Methods. doi: 10.1186/s13007-017-0164-8

Signature: Varanyat

CHAPTER 1

Introduction

This chapter begins by introducing the research problems and the motivation for the work described in this thesis (Section 1.1). To fully grasp the significance of the research problems, Section 1.2 to Section 1.8 provide comprehensive literature reviews of the research context. Then, the objectives of the work, which address the research problems, are described and the approach taken in the work is discussed (Section 1.9). Finally, the structure of the thesis is outlined.

1.1 Introduction

As sessile organisms, plants depend on a vast diversity of secondary metabolites to cope with the fluctuating abiotic and biotic environmental challenges. Therefore, secondary metabolites are essential to the survival and reproductive fitness of plant in its natural environment. Glucosinolates (GSLs) are a group of amino-acid derived thioglycosidic secondary metabolites (β-glucosides). Their occurrence is limited to the members of the Brassicales order, which include *Brassica* oil crops of economic and nutritional importance such as oilseed rape (*Brassica napus* L.), as well as the related model plant species *Arabidopsis thaliana* (Fahey *et al*., 2001; Wittstock and Halkier, 2002; Halkier and Gershenzon, 2006). GSL degradative products possess a wide range of bioactivities such as chemoprevention (Becker and Juvik, 2016), contribution to flavour (Bell *et al*., 2018), and have long been known for their defensive properties against herbivores and non-adapted pathogens (Glen *et al*., 1990; Verhoeven *et al*., 1996; Potter *et al*., 2000; Hopkins *et al*., 2009;

Bell *et al*., 2018). As defensive secondary metabolites, GSLs tend to accumulate to highest concentrations in the organs which contribute most to plant fitness such as in seeds (Brown *et al*., 2003). From the agricultural perspective, however, accumulation of certain GSLs in *B. napus* seeds are undesirable as it produces goitrogenic products that reduce the nutritional values of the protein-rich seed meal used as livestock feed (Griffiths *et al*., 1998; Tayo *et al*., 2012). The major focus in the past had been on reducing GSLs in the seeds of *B. napus* to allow the use of seed meal as animal feed (Rosa *et al*., 1997), but little attention has been paid to how this may have affect GSL compositions in the leaves and roots.

Despite the successful establishment of the low seed GSL canola cultivars through successive breeding practice, the molecular mechanism underlying the low GSL trait in *B. napus* was unclear. To get better understanding of the modular genetic system that regulates GSL natural variations in *B. napus* as a whole, more work is needed to investigate the regulations of GSL in the vegetative tissues and how these variations relate to the GSL profiles in the seed. This knowledge is essential for targeted marker express breeding of crop traits and exploitation of GSL product potentials, e.g. to improve pest resistance and biofumigation properties in *Brassica* crops while maintaining the requirement of low GSLs in the seeds. So far, significant progress has been made in understanding the biochemistry and the regulatory controls of the two classes of GSLs found in *A. thaliana*, the methionine-derived aliphatic and tryptophan-derived indole GSLs (Grubb and Abel, 2006; Halkier and Gershenzon, 2006; Sønderby *et al*., 2010). However, less information is available for the chain-elongated homophenylalanine-derived aromatic GSLs, which is abundant in *Brassica* crops (Bhandari *et al*., 2015). The inability to use the model plant *A. thaliana* to study aromatic GSLs and the challenges of working with *B. napus* complex polyploidy has limited the advancement in the understanding of the aromatic biosynthetic pathway.

This thesis aims to uncover the underlying genetic bases controlling quantitative variation of the major GSL classes in *B. napus* vegetative tissues using a transcriptome-based GWAS approach, Associative Transcriptomics (AT), to overcome the challenges of working with polyploidy species.

1.2 Glucosinolate structure and diversity

GSLs are a large group of sulphur-rich anionic secondary metabolites, and at least 120 different GSLs have been identified in higher plants (Fahey *et al*., 2001). All GSLs share a chemical structure consisting of β-D-glucopyranose moiety linked via a sulphur atom to a *cis*-*N*-hydroximinosulphate ester with a variable side chain (R) (Figure 1-1) derived from one of eight amino acids (Halkier and Gershenzon, 2006). GSLs are classified into three structural groups depending on the classes of the amino acid precursor of the side chain. Aliphatic GSLs are derived from aliphatic amino acids, mainly from methionine (Met), but can also originated from alanine, valine, leucine or isoleucine. Aromatic GSLs are derived from aromatic amino acids, mainly from phenylalanine (Phe) and less so from tyrosine. Indole GSLs are derived from tryptophan (Trp) (Table 1-1).

The structural diversity of GSLs can be attributed to two features of the biosynthetic pathway. The first feature is the elongation of amino acid precursors to generate variations in side chain length (Tokuhisa *et al*., 2004; Benderoth *et al*., 2006) and the second feature is the extensive secondary modifications of the side chain to generate different functional groups (e.g. hydroxylation, thiol oxidation, desaturation, esterification). This diversity in the structure and accumulation of GSLs are believed to be driven by the reciprocal process of adaptation and counter-adaptation between plants and their biotic attackers (Rask *et al*., 2000; Edger *et al*., 2015).

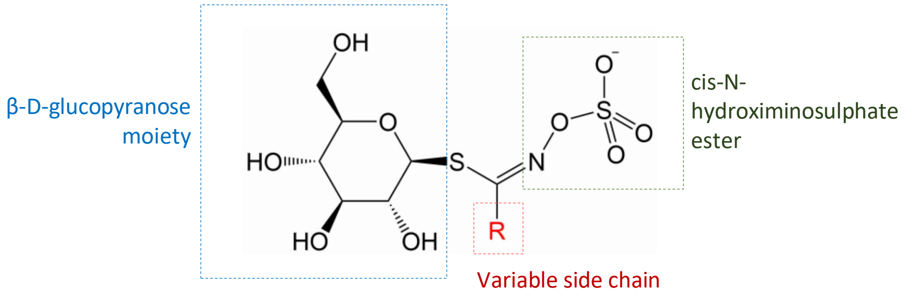
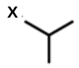
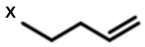
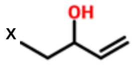
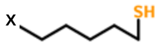
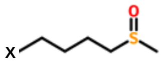
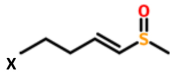
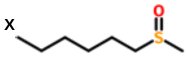
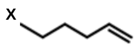
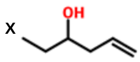
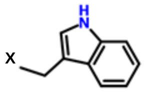
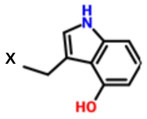
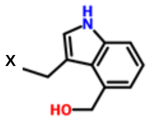
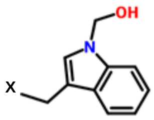
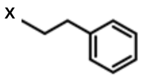


Figure 1-1. Glucosinolates chemical structure, consisting of β-D-glucopyranose moiety linked via a sulphur atom to a *cis*-*N*-hydroximinosulphate ester with a variable R group. This whole structure is represented by 'x' in Table 1-1.

Table 1-1. Representative glucosinolates identified in *Brassica napus* from the three major structural classes. Abbreviation: C_3 to C_5 – methionine-derived aliphatic GSLs with different chain lengths; Ind – tryptophan-derived indole GSLs; Aro – phenylalanine-derived aromatic GSLs. Under R structure heading, the symbol ‘x’ represents the GSL structure in Figure 1-1.

Type	Trivial name	Acronym	R Side chain	R Structure
C_3	Glucoputranjivin	GJV	1-Methylethyl	
C_4	Gluconapin	GNA	3-Butenyl	
	Progoitrin	PRO	(2R)-2-Hydroxy-3-butenyl	
	Glucoerucin	GER	4-Methylthiobutyl	
	Glucoraphanin	GRA	4-Methylsulfinylbutyl	
	Glucoraphenin	GRE	4-Methylsulfinyl-3-butenyl	
C_5	Glucoalyssin	GAL	5-Methylsulfinylpentryl	
	Glucobrassicanapin	GBN	Pent-4-enyl	
	Gluconapoleiferin	GNL	2-Hydroxy-pent-4-enyl	
Ind	Glucobrassicin	GBS	3-Indolylmethyl	
	4-Hydroxyglucobrassicin	4-OHGBS	4-Hydroxy-3-indolylmethyl	
	4-Methoxyglucobrassicin	4-OMeGBS	4-Methoxy-3-indolylmethyl	
	Neoglucobrassicin	neo-GBS	1-Methoxy-3-indolylmethyl	
Aro	Gluconasturtiin	GST	2-Phenethyl	

1.3 Biosynthesis of glucosinolates

The biosynthetic pathway of GSLs proceeds in three stages via (i) amino acid side chain elongation; (ii) the amino acid precursor undergoing metabolic configurations to form the core GSL structure; and (iii) secondary modifications of the side chain to generate a wide spectrum of GSL compounds (Figure 1-2). Many of the genes responsible for biosynthetic steps have been identified in *Arabidopsis thaliana* (reviewed in Wittstock & Halkier 2002; Grubb & Abel 2006; Halkier & Gershenzon 2006; Sønderby *et al*. 2010), which has also helped clarify the core biosynthesis steps and identify orthologous genes in the closely related *Brassica* species. Since methionine-derived aliphatic and tryptophan-derived indole GSLs are the two main classes of GSLs found in *A. thaliana* (Brown *et al*., 2003), significant progress has been made in understanding the genetic of biosynthesis and regulation of these two classes of GSLs. However, less information is available for the chain-elongated homophenylalanine-derived aromatic GSL, which is abundant in *Brassica* species (Bhandari *et al*., 2015) but produced only in minor amounts by a few ecotypes of *A. thaliana* (Brown *et al*., 2003).

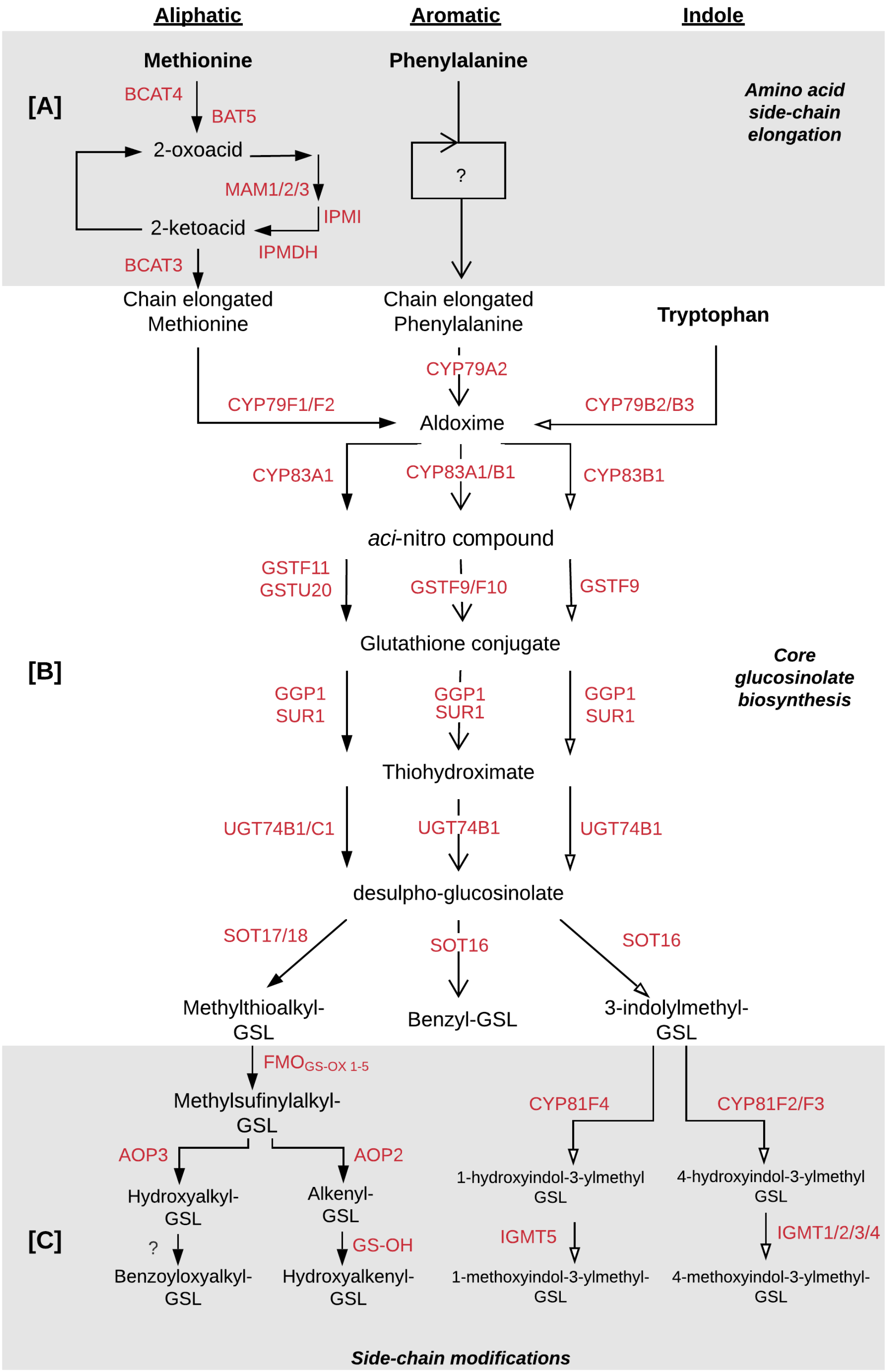


Figure 1-2. The aliphatic, aromatic and indole glucosinolate biosynthetic pathways. A) Amino acid side-chain elongation. B) Biosynthesis of the core glucosinolate structure. C) Secondary modifications of the side-chain. The aliphatic, aromatic and indole pathways are represented by the different types of arrows. Information collated from Halkier & Gershenzon, 2006; Sønderby *et al.*, 2010; Pfalz *et al.*, 2016.

1.3.1 Side chain elongation

The major GSLs found in *B. napus,* such as (2R)-2-hydroxy-3-butenyl GSL (progoitrin) and 2-phenylethyl GSL (gluconasturtiin) (Porter *et al*., 1991; Velasco *et al*., 2008), are derivatives of chain-elongated methionine and chain-elongated phenylalanine respectively. Before entering the core structure biosynthesis pathway, Met and Phe undergo chain elongation (Dörnemann *et al*., 1974; Graser *et al*., 2000). To date, the genes involved in the chain elongation of Phe are unclear, whereas the genes involved in the chain elongation of Met has been described in great detail.

The process of Met elongation is analogous to the conversion of branched-chain amino acid valine to chain-elongated leucine homolog, in which the amino acids go through five reactions: an initial transamination, acetyl-CoA condensation, isomerisation, oxidative decarboxylation and a final transamination (Sønderby *et al*., 2010). The process starts with a deamination of Met by branched-chain amino acid aminotransferase (BCAT) to produce a 2-oxo acid. This first enzyme is localised in the cytosol (Knill *et al*., 2008). However, the remaining enzymes involved in chain elongation are localised in the chloroplast. This indicates the requirement to transport 2-oxo acid into chloroplast. Feeding studies *in planta* has suggested that bile acid transporter (BAT5) imports 2-oxo acid into the chloroplast (Gigolashvili *et al*., 2009). Once imported into the chloroplast, the 2-oxo acid enters a three-step cycle of transformation: condensation with acetyl-CoA by a methylthioalkylmalate (MAM) synthase, isomerisation by an isopropylmalate isomerase (IPMI), and oxidative decarboxylation by an isopropylmalate dehydeogenase (IPM-DH) (Figure 1-2A), yielding a 2-oxo acid that has been elongated by a single methylene group (CH_2). At this stage, the extended 2-oxo acid can be transaminated by BCAT to yield homo-Met and enter the core structure pathway or proceed through another round of elongation. In plants, up to nine cycles are known to happen (Fahey *et al*., 2001).

1.3.1.1 MAM synthase is the key enzyme controlling chain-length variation

The *GSL-Elong* locus has been identified as one of the three major loci genetically controlling quantitative variation in *Arabidopsis* GSLs (Kliebenstein *et al*., 2001a). Fine-mapping of the *GSL-Elong* quantitative trait locus (QTL) in *Arabidopsis* ecotype Columbia had led to the identification of three genes encoding the MAM synthases (*MAM1, MAM2* and *MAM3*) (Kroymann *et al*., 2001).

The three MAM enzymes have similar properties but differ in their substrate specificity. *In vitro* and *in vivo* studies have showed *MAM2* to be involved in the production of aliphatic GSL derived from one elongation cycle (C_3), while *MAM1* is involved in the production of aliphatic GSL derived from two elongation cycle converting C_3 to C_4 aliphatic GSLs (Kroymann *et al*., 2001; Textor *et al*., 2007). MAM3 (formerly known as MAM-L) is able to catalyse all six elongation reactions of Met chain elongation that occur in *A. thaliana* and is involved in all aliphatic GSL production (C_3 to C_8) but mainly contributed to GSLs derived from one, five and six elongation cycles (Textor *et al*., 2007). Though less preferential to Met-derived 2-oxo acid substrates, MAM3 is also able to metabolise a range of non-Met derived 2-oxo acids such as phenylpyruvate (Textor *et al*., 2007), which represents a condensation reaction leading to 2-phenylethyl GSL, an aromatic GSL that is abundant in Brassica. This condensation process catalysed by MAM synthases is the key enzymatic step in determining the length of the amino acid derived side chain and therefore plays a major role in generating GSL structural diversity.

1.3.2 Glucosinolate core structure biosynthesis

Most of the steps for the biosynthesis of glucosinolate core structure and the genes responsible have been identified in *A. thaliana* (Figure 1-2B). This has helped clarify the core biosynthesis steps and identify orthologous genes in the closely related *Brassica* species. The first step in the core structure formation is the conversion of either a primary amino acids or chain elongated amino acids to aldoximes. This conversion is catalysed by the cytochrome P450s belonging to the CYP79 family, each of which has substrate specificity for different amino acid precursors (Halkier and Gershenzon, 2006). Five functional *CYP79* homologues found in *A. thaliana* genome have been characterised. CYP79A2 specifically metabolises L-phenylalanine (Wittstock and Halkier, 2000), CYP79B2 and CYP79B3 convert Trp to their aldoximes (Hull *et al*., 2000), and CYP79F1 and CYP79F2 catalyse the conversion of chain-elongated Met derivatives (Reintanz *et al*., 2001; Chen *et al*., 2003). CYP79F1 and CYP79F2 have different substrate specificity. CYP79F1 is able to metabolise both short- and long-chain Met derivatives, whereas CYP79F2 is restricted to pentahomo- and hexa-homomethionine (Chen *et al*., 2003). The characterisation of these *CYP79* has resolved the first step of core structure formation of most GSLs found in Arabidopsis. Although CYP79A2 is able to catalyse phenylalanine substrates into phenylacetaldoxime, it has a narrow substrate specificity and is unable to metabolise DL-homophenylalanine (Wittstock and Halkier, 2000), the amino acid precursor of 2-phenylethyl GSL. To date, the enzyme that controls the flux into the biosynthetic pathway of homophenylalanine-derived aromatic GSLs is yet to be identified.

The second step is the conversion of aldoximes to nitrile oxides or *aci*-nitro compounds by cytochromes P450 of the CYP83 family (Figure 1-2B). Both CYP83A1 and CYP83B1 can catalyse all of the tested aldoximes, but CYP83A1 has higher affinity for aliphatic aldoximes whereas CYP83B1 prefers Trp-derived and Phe-derived acetaldoximes as substrates (Naur *et al*., 2003). The product, an activated, oxidised form of the aldoxime, of this CYP83 enzymes catalysed step is then conjugated to a sulphur donor. Although cysteine has been considered a probable sulphur donor in GSL biosynthesis (Halkier and Gershenzon, 2006), study by Schlaeppi *et al* (2008) has indicated glutathione (GSH) as a more plausible sulphur donor. Upon herbivore challenging and fungal attack, mutants impaired in GSH biosynthesis displayed a reduction in the level of GSL production, which was restored upon feeding with GSH (Schlaeppi *et al*., 2008). The hypothesis that GSH is the sulphur donor would require a specific Gluthathione-S-transferase (GST) to catalyse the sulphur incorporation step. Candidate GSTs have been identified based on the co-expression of their corresponding genes with genes encoding other GSL biosynthetic enzymes (Hirai *et al*., 2005; Hirai, 2009), which supports this hypothesis. Furthermore, it has been reported that GSTF11 and GSTU20 are involved in the aliphatic GSL biosynthesis, GSTF9 and GSTF10 are predicted to be involved in indole GSL pathway (Hirai *et al*., 2005; Sønderby *et al*., 2010).

The glutathione conjugate are subsequently metabolised by γ-glutamyl peptidase 1 (GGP1) (Geu-Flores *et al*., 2009) and C-S lysase SUR1 to form thiohydroximate. The *sur1* mutant in *A. thaliana* does not produce detectable levels of GSLs when fed with a labelled aldoximes, which indicates that only one C-S lysase is involved in the GSL biosynthesis across the three structural classes (Mikkelsen *et al*., 2004). Thiohydroximates are then glucosylated by glucosyltransferases of the UGT74 family to produce desulfoglucosinolates (ds-GSLs). Based on the co-expression with genes involved in the biosynthesis of aliphatic GSLs, UGT74C1 has been proposed to metabolise Met-derived thiohydroximates (Gachon *et al*., 2005), while UGT74B1 had been shown to glucosylate the Phe-derived thiohydroximates (Grubb *et al*., 2004). Finally, the ds-GSLs are sulphated by the sulphotransferases SOT16, 17 and 18 to form GSLs. These three enzymes accept a broad range of ds-GSL as substrates. However, SOT16 has higher affinity for Trp-derived and Phe-derived ds-GSLs, whereas SOT17 and SOT18 prefers long-chained Met-derived substrates (Piotrowski *et al*., 2004).

1.3.3 Secondary side chain modifications

In addition to the side chain length variation in GSL structures, secondary modifications of the GSL side chain are also of great biological interest because the nature of these modifications contribute largely to the biological activity of GSLs (Hopkins *et al*., 2009). Secondary modifications are most extensive in aliphatic GSLs and include processes such as oxygenations, hydroxylations, alkenylations and benzoylations (Figure 1-2C). In Arabidopsis*,* these modifications are controlled by four polymorphic genetic loci called *GS-OX*, *GS-OH*, *GS-ALK* and *GS-OHP* (Kliebenstein *et al*., 2001b). While hydroxylations and methoxylations of indole GSLs have also been reported (Pfalz *et al*., 2009; Pfalz *et al*., 2011), modifications of aromatic GSLs have not yet been described.

1.3.3.1 Secondary modifications of aliphatic GSLs

Secondary modifications of aliphatic GSLs starts with the oxidation of sulphur in the methylthioalkyl-GSL to form methylsulfinyl moieties. The flavin monooxygenase FMO$_{GS\text{-}OX1}$, localised within the limits of *GS-OX* QTL, was identified as a candidate for the S-oxygenation enzyme activity based on the co-expression with aliphatic GSL genes and their biochemistry activity in catalysing heteroatom oxygenations (Hansen *et al*., 2007). Along with the FMO$_{GS\text{-}OX1}$, a crucifer-specific subgroup of FMO genes called FMO$_{GS\text{-}OX\ 2\text{-}5}$, was identified through phylogenetic analysis (Li *et al*., 2008). Knockout mutants and overexpression studies of FMO$_{GS\text{-}OX\ 1\text{-}5}$ showed the decrease in the ratio of S-oxygenated aliphatic GSLs to total aliphatic GSLs (Li *et al*., 2008), indicated that FMO$_{GS\text{-}OX\ 1\text{-}5}$ is involved in the S-oxygenation of methylthioalkyl-GSL to methylsulfinylalkyl-GSL.

The second round of side chain modification is where methylsulfinylalkyl-GSL get transformed to hydroxyl- and alkenyl- aliphatic GSLs. These reactions are the branching point in aliphatic GSL biosynthesis and are controlled by the tightly linked *GS-ALK* and *GS-OHP* loci (Kliebenstein *et al*., 2001c). Collectively the *GS-ALK* and *GS-OHP* QTLs are refer to as *GS-AOP*. Within these QTLs, two 2-oxoglutarate-dependent dioxygenases has been identified: *AOP2* and *AOP3*. *AOP2*, identified within the *GS-ALK*, catalyses the conversion of S-oxygenated GSLs to alkenyl- GSLs; whereas *AOP3*, localised to the *GS-OHP* locus, controls the reaction toward hydroxylalkyl GSLs (Kliebenstein *et al*., 2001c).

The third round of modification is controlled by the *GS-OH* locus, which is responsible for the production of hydroxylated alkenyl GSL, including 2-hydroxy-3-butenyl GSL (progoitrin). Accumulation of progoitrin poses a problem for using *B. napus* crops as animal feed because its specific hydrolysis derivative, oxazolidine-2-thione, causes goiter in animals (Tayo *et al*., 2012). The production of progoitrin is controlled by *GS-OH*, which is dependent on the presence of both *MAM1* (to convert C_3 to C_4 aliphatic GSLs) and *AOP2* (for the alkenylation step) (Figure 1-2C). Fine-scale mapping of *GS-OH* identified a 2-oxoglutarate-dependent dioxygenase, encoded by At2g25450, as essential for the hydroxylation of 3-butenyl to 2-hydroxy-3-butenyl (Hansen *et al*., 2008). *In planta*, null mutations and T-insertion mutations in the At2g25450 displayed a complete lack of 2-hydroxy-3-butenyl GSL when fed with the precursor GSL compared to wild-type (Hansen *et al*., 2008).

1.3.3.2 Secondary modifications of indole GSLs

Both *B. napus* and *A. thaliana* can produce four different indole GSLs (Kliebenstein *et al*., 2001b; Velasco *et al*., 2008): the unmodified 3-Indolylmethyl (GBS) and its downstream products 4-hydroxy-3-indolylmethyl (4-OHGBS), 4-methoxy-3-indolylmethyl (4-OMeGBS) and 1-methoxy-3-indolylmethyl (neo-GBS) (Table 1-1).

The hydroxylation of GBS to 4-OHGBS is catalysed by the cytochrome P450s belonging to the CYP81 family, CYP81F2 (Pfalz *et al*., 2009) and CYP81F3 (Pfalz *et al*., 2011). Combination of QTL fine-mapping and transcript profiling first identified *CYP81F2* as a gene underlying the *Indole Glucosinolate Modifier 1* (*IGM1*), the QTL for the accumulation of 4-OHGBS and 4-OMeGBS GSLs (Pfalz *et al*., 2009). Knockouts in *CYP81F2* significantly reduced 4-OHGBS and 4-OMeGBS production and increase in the precursor GBS pool (Pfalz *et al*., 2009). However, the knockouts still produce detectable amount of 4-OHGBS and 4-OMeGBS in plant tissue which suggest at least one other gene was able to catalyse the same reaction. Through biochemical assays and mutant analysis, *CYP81F3* was identified as responsible for the same reaction as *CYP81F2* (Pfalz *et al*., 2011).

Despite analyses of GSL profiles from many different plant species , 1-hydroxy-3-indolylmethyl GSL (1-OHGBS), an intermediate upstream of neo-GBS, has never been reported (Fahey *et al*., 2001; Velasco *et al*., 2008; Clarke, 2010). This is possibly due to the instability of its structure where the hydroxyl group is attracted to the nitrogen of the indole ring. The use of *N. benthamiana* transient expression system enabled coupling of the expression of the genes encoding indole GSLs biosynthetic enzymes with O-methyl-transferase genes. This has led to the characterisation of *CYP81F4*, and identification of *IGMT1* and *IGMT2* (Pfalz *et al*., 2011). Disruption of *CYP81F4* virtually abolished production of 1-OHGBS in *cyp81f4* mutants which suggests that its encoded protein is mainly responsible for the production of 1-OHGBS from GBS. *IGMT* tandem genes are a small gene cluster of O-methyl-transferase genes that are fairly specific to secondary modification of indole GSLs. *IGMT1* and *IGMT2* were shown to be able to catalyse the transfer of methyl groups to hydroxy indole GSLs (Pfalz *et al*., 2011). While *IGMT3* and *IGMT4* predicted to have the same function as *IGMT1* and *IGMT2* due to the high sequence similarities (95 to 98% identical at the amino acid level), *IGMT5* shared the least identity with other IGMTs (~70%) but its expression was highly correlated to *CYP81F4* expression. Works from the same group showed that knockouts of *IGMT5* in *Arabidopsis* roots significantly reduced the amount of neo-GBS abundance (Pfalz *et al*., 2016).

Indole GSL modifications occur at two positions of the indole ring, either position 1 or position 4, which form two parallel pathways. At position 1, hydroxylation and methoxylation was carried out by enzymes encoded by *CYP81F4* and *IGMT5* to form neo-GBS as the final product, whereas the modifications at position 4 of the indole ring were catalysed by proteins encodes by *CYP81F1-F3* and *IGMT1-4* which form 4-OMeGBS as the final product (Pfalz *et al*., 2016) (Figure 1-2C).

1.4 Regulation of glucosinolate biosynthesis

As secondary metabolites, GSLs are biochemically expensive to synthesise because they are derived from intermediates or end products of primary metabolic pathways. In Arabidopsis, GSL biosynthesis has been estimated to increase photosynthetic requirements by at least 15% (Bekaert *et al.*, 2012). It is, therefore, important for plants to regulate the production of GSL for specific needs and balance the metabolic expenditure for growth-defence trade-offs. Many biological processes in plants are regulated at the transcription level. In Arabidopsis, a large number of activator- and repressor-type transcription factors that regulate GSL biosynthetic genes have been characterised (reviewed in Frerigmann, 2016). In this review, special attention is on the R2R3-type myeloblastosis (MYB) and the basic-Helix-loop-Helix (bHLH or MYC) transcription factor families (Table 1-2). The physical interaction between these MYB and MYC factors plays a crucial role in the combinatorial control of GSL biosynthesis. On one hand, the MYB factors control the specificity of the aliphatic and indole GSL pathway, and they also regulate the activity of genes in the core structure pathway. On the other hand, the MYC factors control the basal level of all GSL types by interacting with both types of MYBs, and fine-tune the production of this defence compound by integrating phytohormone signals, particularly jasmonate. In essence, the specificity and intensity of GSL regulation relies on the MYB transcription factors while MYCs provide baseline level control but not rate limiting in the GSL pathway.

Table 1-2. List of R2R3-type MYB transcription factors reviewed in this chapter.

Names	AGI code	References
MYB28 / HAG1	AT5G61420	(Gigolashvili *et al.*, 2007b; Hirai *et al.*, 2007; Sonderby *et al.*, 2010)
MYB29 / HAG3	AT5G07690	(Gigolashvili *et al.*, 2008; Sonderby *et al.*, 2010)
MYB76 / HAG2	AT5G07700	(Gigolashvili *et al.*, 2008; Sonderby *et al.*, 2010)
MYB34 / ATR1	AT5G60890	(Celenza, 2005; Frerigmann and Gigolashvili, 2014)
MYB51 / HIG1	AT1G18570	(Gigolashvili *et al.*, 2007a; Frerigmann and Gigolashvili, 2014)

MYB122 / HIG2	AT1G74080	(Gigolashvili *et al*., 2007a; Frerigmann and Gigolashvili, 2014)
bHLH04 / MYC4	AT4G17880	(Schweizer *et al*., 2013; Frerigmann *et al*., 2014)
bHLH05 / MYC3	AT5G46760	(Schweizer *et al*., 2013; Frerigmann *et al*., 2014)
bHLH06 / MYC2	AT1G32640	(Schweizer *et al*., 2013; Frerigmann *et al*., 2014)
bHLH28 / MYC5	AT5G46830	(Frerigmann *et al*., 2014)

1.4.1 R2R3-MYB transcription factors

The direct transcriptional regulators of aliphatic and indole GSL biosynthesis are a group of six homologous R2R3-MYB transcription factors. MYB34/ATR1, MYB51/HIG1, and MYB122/HIG2 are thought to regulate the tryptophan-derived indole GSL pathway (Celenza, 2005; Gigolashvili *et al*., 2007a; Frerigmann and Gigolashvili, 2014), while MYB28/HAG1*,* MYB29/HAG3 and MYB76/HAG2 regulate expression levels of the methionine-derived aliphatic GSL biosynthetic genes (Gigolashvili *et al*., 2007b; Hirai *et al*., 2007; Gigolashvili *et al*., 2008; Sonderby *et al*., 2010).

1.4.1.1 MYB28, MYB29 and MYB76 are regulators of aliphatic GSL biosynthesis

Through co-expression analysis of publicly available Arabidopsis transcriptomic datasets, *MYB29/HAG3* and in particular *MYB28/HAG1* have been shown to be highly co-expressed with the biosynthetic genes of aliphatic GSLs (Hirai *et al*., 2007). These two MYBs, together with MYB76/HAG2, form a regulatory network to shape the profile of aliphatic GSLs (Gigolashvili *et al*., 2007b; Gigolashvili *et al*., 2008; Sonderby *et al*., 2010). Even though *trans*-activation assay shows that all three Arabidopsis MYB-HAGs can induce the promoters of aliphatic GSL genes (Gigolashvili *et al*., 2008), MYB28/HAG1 showed the highest transactivation potential towards aliphatic GSL biosynthetic genes. For instance, *MAML* and *CYP79F2* were strongly activated by MYB28/HAG1 but less so by MYB29/HAG3 (Sønderby *et al*., 2007; Gigolashvili *et al*., 2008). Characterisation of T-DNA mutants in *myb28* has shown the repression of gene expression of most aliphatic GSL biosynthetic genes as well as a reduction in both the short- and long-chained aliphatic GSLs (Gigolashvili *et al*., 2007b; Hirai *et al*., 2007; Sønderby *et al*., 2007). However, single knock-outs in *myb29* or *myb76* only reduce the production of short-chain aliphatic GSLs (chain length of C_3 to C_6: Sønderby *et al*., 2007; Gigolashvili *et al*., 2008). Based on these data, MYB28/HAG1 is regarded as the main regulator of the biosynthesis of aliphatic GSLs, including both short- and long-chain aliphatic GSLs.

In comparison, MYB29/HAG3 is considered only to play a minor role under non-stress conditions. However, the expression of *MYB29/HAG3* is induced more than 50-fold upon wounding, suggesting a role for MYB29/HAG3 in GSL production under mechanical stress (Gigolashvili *et al*., 2008). In addition, the expression of *MYB29/HAG3* is also induced in response to exogenous methyl jasmonate (MeJA) (Gigolashvili *et al*., 2008). This induced expression is specific to *MYB29/HAG3* and not the other MYB genes, leading to the hypothesis that MYB29/HAG3 can integrate stress signals from MeJA for the production of aliphatic GSLs.

Single knockout of *myb76* has little effects on accumulation and gene expression of aliphatic GSL in the mutant plants, leading to the conclusion that MYB76/HAG2 proteins only play an accessory role in the regulation (Gigolashvili *et al*., 2008). Upon wounding *MYB76/HAG2* expression is induced more than 50-fold and could accelerate the production of aliphatic GSL in concert with the other MYBs, indicating a particular role in response to mechanical stimuli. In addition, systematic analysis of metabolite distribution and knock-out mutants suggest a potential role in which MYB76/HAG2 may be involved in regulating the distribution of short-chained aliphatic GSL from the vein to the edge of leaves (Sonderby *et al*., 2010).

The expression patterns of *MYB28/HAG1*, *MYB29/HAG3* and *MYB76/HAG2* in Arabidopsis plants differ from one another but overlap with the known site of aliphatic GSL accumulation. *MYB28/HAG1* is expressed mainly in mature leaves and reproductive organs such as inflorescences (Gigolashvili *et al*., 2007b). *MYB76/HAG2* is the only MYB genes found to be expressed in the transition zone between root and foliar parts, flowers and secondary veins in leaves, and *MYB29/HAG3* expression has been observed in young siliques, trichomes and roots (Gigolashvili *et al*., 2008).

1.4.1.2 *MYB34, MYB51 and MYB122 roles in regulating indole GSLs and auxin homeostasis*

MYB34/ATR1, *MYB51/HIG1*, and *MYB122/HIG2* are close homologues in Arabidopsis and they have been shown to play important but different roles in the regulation of indole GSL and auxin (IAA) biosynthetic pathways. All three MYBs can potentially upregulate the expression of genes involved in the first steps of the indole GSL biosynthetic pathway, e.g. *CYP79B2* and *CYP79B3*, and therefore positively control the main Arabidopsis indole GSL accumulation, i.e. 1-methoxy-3-indolylmethyl (neo-GBS) (Celenza, 2005; Gigolashvili *et al*., 2007a). Nevertheless, *MYB51/HIG1* acts as the key player in the regulatory network controlling indole GSL biosynthesis, while *MYB34/ATR1*, and *MYB122/HIG2* are more involved in regulating the link between indole GSL and auxin biosynthetic pathways.

Plants with overexpressed *MYB51/HIG1* transcripts displayed a significant increase in several indole GSLs, including 4-Methoxy-3-indolylmethyl (4-OMeGBS) and 1-methoxy-3-indolylmethyl (neo-GBS), compared to wild-type (Gigolashvili *et al*., 2007a). Overexpression of either *MYB34/ATR1* or *MYB122/HIG2* leads to high indole GSL chemotype, but with increased production of only neo-GBS but not other indole GSLs (Celenza, 2005; Gigolashvili *et al*., 2007a). Plants with overexpressed *MYB34/ATR1*, *MYB51/HIG1*, and *MYB122/HIG2* have shown increased expression of key pathway genes of indole GSL biosynthesis, such as *CYP79B2*, *CYP79B3* and *CYP83B1* (Celenza, 2005; Gigolashvili *et al*., 2007a). However only MYB51/HIG1 not the other two MYBs can upregulate expression of the genes further downstream of the indole GSL pathways, i.e. *UGT74B1* and *AtST5a* (Gigolashvili *et al*., 2007a). These findings support MYB51/HIG1 as the key player in the regulation of indole GSLs.

On the other hand, the overexpression of *MYB122/HIG2* and *MYB34/ATR1* leads to the high auxin accumulation and causes aberrant growth phenotypes in some *MYB34/ATR1* overexpression lines. Such phenotype has not been observed in *MYB51/HIG1* overexpression lines. In fact, *MYB34/ATR1* preferentially regulate the expression levels of genes that are common to both indole GSL and auxin biosynthesis (*CYP79B2/B3* and *CYP83B1*), indicating its role in regulating the homeostasis between these two pathways (Celenza, 2005).

1.4.2 Interactions between the MYBs

The transcriptional data of MYB genes reveals a highly interconnected network among them. Co-regulation and interdependent between *MYB51/HIG1* and *MYB122/HIG2* has been observed in response to pathogen treatment (Gigolashvili *et al*., 2007a). Similarly, *MYB28/HAG1*, *MYB29/HAG3* and *MYB76/HAG2* is shown to have positive reciprocal activation (Figure 1-3) (Gigolashvili *et al*., 2007b; Hirai *et al*., 2007; Gigolashvili *et al*., 2008). For instance, overexpression of *MYB28/HAG1* causes an increase in the other two MYB transcripts (Gigolashvili *et al*., 2008) while *MYB29/HAG3* is thought to be involved forming a positive feed-forward loop in aliphatic GSL regulation by integrating signals from *MYB28/HAG1* and *MYB76/HAG2* (Sonderby *et al*., 2010).

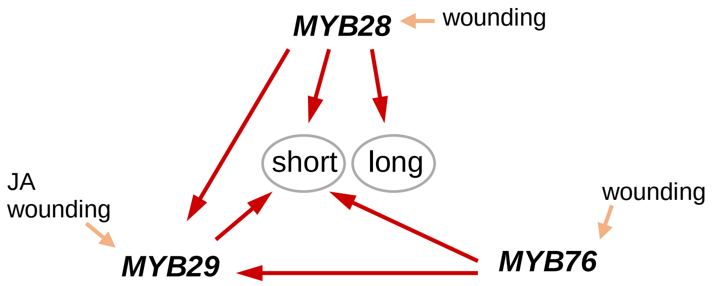


Figure 1-3. Models for the role of MYB28, MYB29 and MYB76 in the regulation of aliphatic GSLs in *Arabidopsis* (accession Col-0), from Sonderby *et al*. (2010). Circles represent short- and long-chain aliphatic GSLs. Short-chained aliphatic GSLs are one to three cycles of chain elongation. Long-chained aliphatic GSLs are four to six cycles of chain elongation. Arrows represent induction of the genes from actual regulation in *planta*.

1.4.2.1 Reciprocal negative crosstalk between aliphatic and indole GSLs

Several studies have demonstrated the reciprocal negative controls between aliphatic and indole GSL biosynthetic pathways through knockout mutants and overexpression lines (Gigolashvili *et al*., 2008; Malitsky *et al*., 2008; Sonderby *et al*., 2010). However, between these studies, no consensus in the results for this negative control between the pathways have been found. For instance, transactivation assay of *HAG*[2] in *Nicotiana benthamiana* had shown to repress indole GSL regulators gene expression (Gigolashvili *et al*., 2008), yet overexpression of *HAG* genes in *Arabidopsis thaliana* had not reduced the expression of indole GSL regulators (Sønderby *et al*., 2007; Malitsky *et al*., 2008). When *MYB29/HAG3* and *MYB76/HAG2* had been overexpressed, no changes in the expression of *HIG*[3] regulator had been detected, though there is a significant reduction in the expression of indole GSL biosynthetic genes and the amounts of GBS and neo-GBS indole GSLs (Malitsky *et al*., 2008). Similarly, the *myb34 myb51 myb122* triple mutant have shown to produce higher amount of aliphatic GSLs in the roots even though this was not caused by the induction of *HAG* gene expression (Frerigmann and Gigolashvili, 2014). These results suggest that *HAG*- and *HIG-MYBs* may not negatively regulate each other as previously thought. Rather, it can be interpreted as that the changes in the amounts of aliphatic and indole GSLs in the knockout lines are caused by the reduced competition for the common enzymes and intermediates such as GSH, PAPs, C-S lyase and UGT74B1 between the aliphatic and indole biosynthetic pathways (Frerigmann, 2016).

[2] 'HAG' is a collective term for the MYB transcription factors of aliphatic GSLs (i.e. *MYB28/HAG1*, *MYB29/HAG3* and *MYB76/HAG2*).

[3] 'HIG' is a collective term for the MYB transcription factors of indole GSLs (i.e. *MYB34/ATR1*, *MYB51/HIG1*, and *MYB122/HIG2*).

1.4.3 bHLH/MYC is an interacting partner of MYB

MYB transcription factors often form complexes with bHLH/MYC proteins to regulate biosynthetic pathways in plants (Zimmermann *et al*., 2004). Four members of the bHLH subgroup IIIe have been found to interact with the MYB-interaction-region of all six of the GSL-related MYB factors. These bHLH/MYC transcription factors are bHLH04/MYC4, bHLH05/MYC3, bHLH06/MYC2 and bHLH28/MYC5 (Schweizer *et al*., 2013; Frerigmann *et al*., 2014). In addition to forming complexes with the MYBs, Schweizer *et al.* (2013) showed that MYC transcription factors can directly regulate the expression of GSL gene by binding to their promoter region. Single *myc* mutants retain wild-type expression of GSL biosynthetic genes (Schweizer *et al*., 2013), and shows no significant reduction in GSL phenotype (Frerigmann *et al.*, 2014*)*. However, the triple *myc2 myc3 myc4* mutants abolish the production of GSLs; revealing the importance of bHLH/MYC in the GSL regulation and certain levels of redundancy of these MYC copies.

Out of all *myc* double mutants, only the *myc3 myc4* plants show significantly reduced aliphatic and indole GSL production in the absence of jasmonic acid (Frerigmann *et al.,* 2014*)*. This suggests that MYC3 and MYC4 are redundant but they are most crucial for basal GSL control. Nevertheless, upon jasmonic acid treatment, bHLH06/MYC2 has been shown to have the ability to complement the low GSL level in *myc3 myc4* double mutants, implying a specific role of MYC2 in jasmonic acid signalling. The observation could be explained by the higher affinity of MYC2 to Jasmonate ZIM-domain proteins (JAZs) compared to other MYCs (Cheng *et al.*, 2011). In the absence of jasmonate, MYC2 forms an inactive complex with JAZ and its activities are hindered. Upon jasmonate treatment, MYC2 are released from the inactive complexes and therefore able to complement the knocked-out MYC3 and MYC4 in the *myc3 myc4* mutant plants (Frerigmann, 2016).

To date, it is unclear whether the fourth homolog of the bHLH subgroup IIIe, *i.e.* bHLH28/MYC5, involves in the regulation of GSL biosynthesis. The expression level of bHLH28/MYC5 is very low *in planta* (Winter *et al.*, 2007; Frerigmann, 2016) and single knock out *myc5* mutant plants have shown no effects on GSL level (Schweizer *et al*., 2013).

Together the MYBs and MYCs have been shown to assert combinatorial control on the GSL biosynthetic pathway. Overexpression of MYCs alone does not result in higher GSL level when not in conjunction with higher abundance of MYB factors (Frerigmann *et al.,* 2014*)*. These results suggest that the abundance of MYCs does not form the bottleneck in the pathway.

1.4.3.1 A model for the regulation of GSL by MYB-MYC complex in Arabidopsis

In the absence of jasmonate signalling, JAZ competitively bind to MYC factors, which prevents the interactions between MYCs and MYBs (Frerigmann, 2016). The preferential binding of some MYCs, particularly MYC3 and MYC4, to the GSL-regulating MYBs rather to the JAZ proteins allows basal transcription activities of GSL genes in wild-type Arabidopsis plants and therefore maintains moderate levels of GSLs (Frerigmann *et al.,* 2014*)*. In the presence of jasmonic acid, JAZs interact preferentially with jasmonate, thus release MYCs from the transcriptional inactive complex. As a consequent, more MYB-MYC protein complexes are formed in the presence of jasmonate (Fernández-Calvo *et al.*, 2011).

In the regulation model proposed by Frerigmann *et al.* (2014), both MYB and MYC factors are recruited to the promoter of GSL genes. MYB factors bind to the MYB-boxes that are present in the promoter region of the GSL genes while MYCs bind to the G-boxes that are present in the same gene (Schweizer *et al*., 2013). The role of MYCs is probably to tether the mediator complex and chromatin modifying factors to DNA, which unwinds the chromatin and makes the DNA accessible to MYBs and RNA polymerase II (Frerigmann *et al.,* 2014*)*. This mechanism then activates the transcription of the GSL gene.

1.5 Systemic distribution of glucosinolates

The abundance and types of GSLs have a pronounced effect on their biological activities and differ significantly between plant species and between cultivars of the same subspecies. These variations in GSL contents are affected by various plant-specific factors such as developmental stages, cultivar genotypes, and site of GSL accumulation in plants (Petersen *et al*., 2002; Brown *et al*., 2003; Bhandari *et al*., 2015), as well as environmental conditions such as temperature, light, water, nutrient availability, agricultural practices, and post-harvest practices (Ji *et al*., 2006; Schonhof *et al*., 2007; Velasco *et al*., 2007; Park *et al*., 2013). Therefore, GSL profiles may vary under both natural and artificial selection. This review will focus more on the plant-specific factors that influence GSL profiles under controlled environmental conditions. Understanding of GSL profiles is important as it provides a groundwork for the study of biosynthesis, transport and metabolism of GSLs because GSL accumulation is a net result of these physiological processes.

1.5.1 Diversity of GSL composition in *A. thaliana* and *B. napus*

In *A. thaliana* (ecotype Columbia), highest accumulation of GSLs are found in the organs that contribute most to plant fitness, and these include seeds, siliques and flowers (Petersen *et al*., 2002; Brown *et al*., 2003). Different GSLs have been found to accumulate in various organs. Seeds of *A. thaliana* are dominated with six aliphatic GSLs that are specific to seeds, while roots contain two major GSLs, aliphatic 4-methylsulfinylbutyl (glucoraphanin), and indole GSL, 1-methoxy-3-indolylmethyl (neoglucobrassicin) (Brown *et al.*, 2003). In leaves, major class of GSLs changes with developmental stages. Similar to roots, leaves are dominated with glucoraphanin aliphatic GSL in vegetative stage and younger leaves accumulate higher concentrations of GSLs than older ones. As plants age, the total amount of GSLs decreases. At the same time, proportion of aliphatic GSLs decreases, resulting in an increased proportion of indole GSLs, mainly 3-Indolylmethyl (glucobrassicin) and neoglucobrassicin (Brown *et al*., 2003).

Similar pattern to *A. thaliana* has been observed in *Brassica* crops where GSLs accumulate to highest levels in reproductive organs, especially seeds (Velasco *et al*., 2008; Bhandari *et al*., 2015). Similarly, levels of GSL concentration in leaves decline with plant maturity in *B. napus* (Porter *et al*., 1991). However, despite the similar pattern with regards to accumulation, GSL profiles in *B. napus* differ considerably to that of *A. thaliana*. Seeds and leaves of *B. napus* are dominated with aliphatic GSLs. Seeds have high concentration of 2-hydroxy-3-butenyl (progoitrin) and 3-butenyl (gluconapin), whereas in leaves pent-4-enyl (glucobrassicanapin) is the most abundant GSL (57% of all GSLs), which is followed by PRO (14% of all GSLs) (Velasco *et al*., 2008) (Figure 1-4). Though no information of GSL profiles in root tissues has been reported for *B. napus*, it has been found that aromatic GSL 2-phenethyl (gluconasturtiin) is the most abundant in the roots of nine other related *Brassica* crops (Bhandari *et al*., 2015).

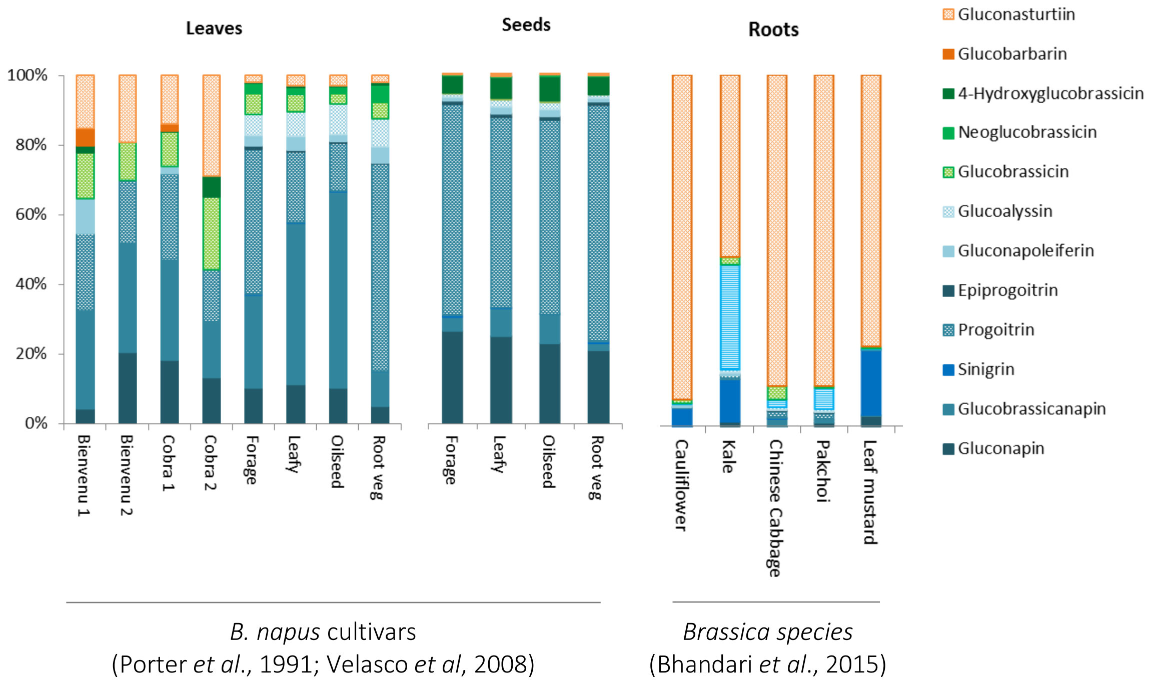


Figure 1-4. Variation of glucosinolate profiles in leaves and seeds between different *B. napus* cultivars and root GSL profiles from various *Brassica* species, summarised from literature. Colour: blue are aliphatic GSLs, green are indole GSLs and orange are aromatic GSLs.

1.5.2 Long-distance transport of glucosinolates

The specific accumulation pattern of GSL compounds in different parts of the plant may be established by *in situ* biosynthesis and/or long-distance transport. Transport of GSLs into sink tissues such as seeds is thought to occur because seeds of both *A. thaliana* and *B. napus* accumulate aliphatic GSLs to high concentrations, but Arabidopsis seeds lack the *in situ* capability for chain elongation and core biosynthesis steps of the aliphatic GSLs (Nour-Eldin and Halkier, 2009). This indicates that aliphatic GSLs must be translocated from other tissues into seeds. In a previous study, *B. napus* F1 hybrid cross between cv. Cobra and a synthetic line showed an identical aliphatic GSL profile in seeds to the profiles in the leaves of maternal parent (Magrath and Mithen, 1993), which suggests that unmodified intact GSLs are transferred from maternal tissue to developing seeds. Investigation into the GSL transport properties confirmed their physiochemical mobility in both xylem and phloem according to the Kleier model (Brudenell *et al*., 1999) thus fulfilling the criteria for long-distance transport. In support of the assimilation translocation of GSLs from source to sink via phloem transport hypothesis, Chen *et al.* (2001) demonstrated the formation of the radiolabelled exogenous *p*-hydroxybenzyl GSL after *A. thaliana* plants were fed with [^{14}C]tyrosine to their rosette leaf. Labelled *p*-hydroxybenzyl GSL was found in other parts of the plants such as other leaves, roots, stem, flower buds and siliques after 24 hours, as well as in the seeds upon bolting. Moreover, intact GSLs has been found in the phloem exudates, indicating that GSLs are transported as intact form, not as desulfoglucosinolates, via phloem in *Arabidopsis* (Chen *et al*., 2001). Similarly, radiolabelled GSLs fed to leaves of *B. napus* has been shown to accumulate in seeds and roots following the phloem pathway (Brudenell *et al*., 1999).

In recent years, specific transporters of GSLs have been identified and characterised. The work by Nour-Eldin *et al.* (2012) had identified two members of the nitrate/peptide transporter family, GTR1 and GTR2, as proton-dependent symporter transporters highly specific to GSLs. While knocking out *gtr1* in *A. thaliana* did not significantly reduce seed GSL content, *gtr2* mutant showed a 48% reduction. In addition, double mutant *gtr1gtr2* in *A. thaliana* plants almost completely abolished the total GSL amount of in seeds. This reduction in seed GSLs was accompanied by the increased in aliphatic GSL accumulation in leaves and silique walls (Nour-Eldin *et al*., 2012), supporting the role of GTRs in the transport of GSLs from source tissues to seeds. In addition to their role in phloem-loading for GSL accumulation in the seeds, GTRs have also been shown to involve in the bidirectional distribution of GSLs between roots and leaves in Arabidopsis. Therefore, GTRs contribute to the establishment of organ-specific GSL compositions (Andersen *et al*., 2013). This grafting experiment has shown to support the proposal that the distinct spatiotemporal distribution of GSL profiles may have been established through a combination of *in situ* biosynthesis and GTR1/GTR2 regulating long-distance transport in Arabidopsis (Andersen *et al.*, 2013).

1.6 The roles of glucosinolates in plants

1.6.1 Glucosinolate-myrosinase system

GSLs are normally sequestered as intact stable form in the vacuoles of ‘S-cells’ that lined the phloem (Koroleva *et al*., 2000), spatially separated from their hydrolytic enzyme myrosinases which is localised in idioblasts called myrosin cells (Andréasson *et al*., 2001). Their potency arises when plant tissue is disrupted, and GSLs come into contact with myrosinase in the presence of moisture. Myrosinases are specific β-glucosidases that remove the β-glucose moiety from the GSLs by hydrolysing the glucosidic bond (Rask *et al*., 2000). This reaction leads to the formation of unstable intermediates. These compounds would then rearrange spontaneously into a wide range of biologically active and sometimes toxic products such as isothiocyanates (ITC), nitriles, epithionitriles, oxozolidine-2-thiones and thiocyanates (Wittstock and Halkier, 2002; Radojcic Redovnikovic *et al*., 2008). The final chemical products and their effect largely depends on side chain structure of the initial GSL, plant species but also the reaction conditions such as pH and presence of interacting proteins such as epithiospecifier proteins (Figure 1-5).

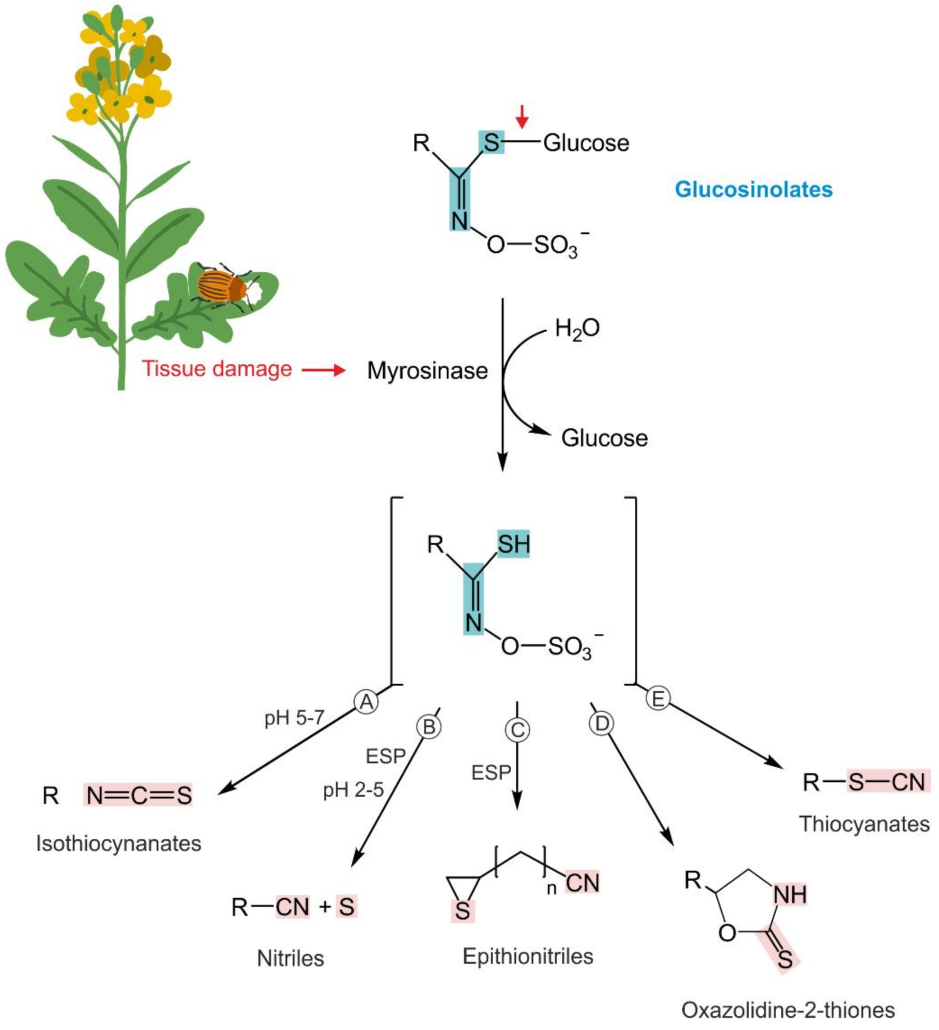


Figure 1-5. The glucosinolate-myrosinase system. Upon tissue damage such as from insect feeding, myrosinase hydrolyses the starting glucosinolate into an unstable intermediate which then rearranges spontaneously and form several breakdown products. At neutral pH, isothiocyanates are typically formed (A). At acidic pH and in the presence of epithiospecifier protein (ESP), nitriles are formed (B). If the starting glucosinolate has a terminal double bond in the side chain, epithionitriles are formed in the presence of ESP after capturing the sulphur released during nitrile formation in the double bond (C). If a hydroxyl group is present at the second carbon position in the GSL side chain, the isothiocyanates formed are unstable and will spontaneously cyclise to oxazolidine-2-thiones (D). Thiocyanates are formed specifically from benzyl-, allyl-, and 4-methylsufinylbutyl-glucosinolates (E). The glycosidic bond where myrosinase acted on glucosinolates is marked with the red arrow. Glucosinolate core structures are highlighted in blue, rearrangements upon hydrolysis are highlighted in pink. Abbreviation: R, variable side chain. Image adapted from Wittstock and Halkier (2002); Redovnikovic *et al.* (2008).

1.6.2 Potential roles of glucosinolates

As GSL hydrolysis products are responsible for the biological functions, the type of GSLs in plants are likely to have been selected for their ability to form specific hydrolysis products (Table 1-3). Changes in environmental conditions such as sulphur status of plant (Falk *et al.*, 2007; Schonhof *et al.*, 2007), and biotic stress such as herbivorous damage and pathogen infection (Hopkins *et al.*, 2009) can significantly alter the GSL compositions in the plant. These alterations in the GSL profile with environmental conditions has brought forward several thoughts about their potential roles. Among them, the most widely accepted role of GSLs is their involvement in the defence against herbivores and pathogens.

Table 1-3. Glucosinolates identified in *Brassica* species and their major hydrolysis products (Mithen, 1992; Rosa *et al.*, 1997). Only the ones with known hydrolysis products are listed.

Chain length	**Trivial name**	**R Side chain**	**Major hydrolysis products**
C_3	Sinigrin	2-Propenyl	Isothiocyanates, Nitrile
C_4	Gluconapin	3-Butenyl	Isothiocyanates, Nitrile
	Progoitrin	(2R)-2-Hydroxy-3-butenyl	Oxazolidine-2-thiones
	Glucoerucin	4-Methylthiobutyl	Isothiocyanates, Nitrile
	Glucoraphanin	4-Methylsulfinylbutyl	Isothiocyanates, Nitrile
C_5	Glucoalyssin	5-Methylsulfinylpentryl	Isothiocyanates, Nitrile
	Glucobrassicanapin	Pent-4-enyl	Isothiocyanates, Nitrile
	Gluconapoleiferin	2-Hydroxy-pent-4-enyl	Oxazolidine-2-thiones
Ind	Glucobrassicin	3-Indolylmethyl	Indolyl-3-carbinol, thiocyanate
	4-Hydroxyglucobrassicin	4-Hydroxy-3-indolylmethyl	Indolyl-3-carbinol, thiocyanate
	4-Methoxyglucobrassicin	4-Methoxy-3-indolylmethyl	Indolyl-3-carbinol, thiocyanate
	Neoglucobrassicin	1-Methoxy-3-indolylmethyl	Phytoalexins
Aro	Gluconasturtiin	2-Phenethyl	Isothiocyanates

1.6.2.1 The role of glucosinolate in plant-pest interactions

The role of GSL-myrosinase system in plant-pest interaction differs depending on the adaptability of the attackers. GSLs can serve as toxic and deterrent to generalist feeders such as slugs and pigeons while at the same time can attract and stimulate feeding and egg laying of specialist insects like adult flea beetles on *Brassica* plants (Giamoustaris and Mithen, 1995).

It has been observed that tissue damage from herbivore feedings and pathogen infections led to changes in GSL accumulation in plants, presumably aiming at increasing plant resistance from further attacks. White mustard plants (*Sinapis alba*) respond to herbivore feedings by accumulating higher levels of GSLs locally, particularly benzyl GSL and 3-indolylmethyl GSL, as well as showing increased myrosinase activities both locally and systematically (Martin and Müller, 2007). It was assumed that indole GSLs are less toxic or less deterrent compared to aliphatic GSLs because they do not produce stable ITCs upon hydrolysis. However, *in vitro* analysis of GSL activity by Mithen *et al.* (1986) has shown that 3-indolylmethyl GSL (neoglucobrassicin) hydrolysis products, especially indolyl-3-carbinol, can be effective inhibitors of fungal pathogen *Leptosphaeria maculans*. In fact, all of the GSLs tested in the study (2-propenyl, 3-butenyl, 3-indolylmethyl, and 1-methoxy-3-indolylmethyl), except 2-hydroxy-3-butenyl (progoitrin), are effective in reducing fungal growth in the presence of myrosinase (Mithen *et al*., 1986). The volatile ITCs, from the hydrolysis of alkenyl GSLs such as sinigrin and glucoraphanin, are highly toxic and strongly inhibitory to fungal pathogens growth (Mithen *et al*., 1986) and downy mildew (Greenhalgh and Mitchell, 1976). In addition to toxicity, volatiles ITCs derived from the hydrolysis of 3-butenyl (gluconapin) and pent-4-enyl GSLs (glucobrassicanapin) can reduce the palatability of cruciferous vegetative tissues, which is considered to deter grazing from herbivores such as pigeons (Mithen, 1992) and slugs (Glen *et al*., 1990). This hypothesis is consistent with the field observation that the rapeseed lines with lower alkenyl GSLs in the leaves are more susceptible to grazing by mammals and birds compared to the ungrazed high alkenyl GSL lines (Mithen, 1992).

To specialist insects, the volatile hydrolysis products may act as an attraction signal from long-distance and the intact GSLs serve as contact cues for feeding and oviposition stimulation (Hopkins *et al*., 2009). Direct comparisons between a specialist and a generalist moth feeding on cotyledon of *B. juncea* with differing GSL profiles and myrosinase activity revealed fundamental differences in the responses of these two species (Li *et al*., 2000). Cotyledons with varying myrosinase activity but high GSL concentrations were fed on less and shorter time by generalist moth, while cotyledons with high myrosinase activity were fed on less by specialist moth. This suggests that levels of intact GSLs have a stronger effect for the defence against generalists, but that of hydrolytic products, as indicated by higher myrosinase activity, might be more essential for defence against specialists (Li *et al*., 2000).

1.6.2.2 Glucosinolates as biofumigants

When GSLs in *Brassica* green manures or rotation crops are hydrolysed, the same hydrolytic products that act as natural protection for plants can be used to suppress soil-borne pests and pathogens in the soil (Kirkegaard and Sarwar, 1998). This concept is called 'biofumigation'. The biofumigation potential of *Brassica* crops is thought to contribute to their attractiveness as break crops because cereal crops grown after *Brassica* crops have higher yield compared to that following non-*Brassica* crops (Kirkegaard *et al*., 1994). These major hydrolysis products, especially ITC from 2-phenylethyl GSL which is abundant in *Brassica* roots, are thought to be responsible for the biofumigation potential. This is because they are most toxic and have the chemical properties suitable for the soil environment. Due to their low volatility and hydrophobic properties, 2-phenylethyl isothiocyanates (2PE-ITC) are less likely to be lost from soil via volatile and leaching losses (Sarwar *et al.*, 1998; Laegdsmand *et al*., 2007). It is also one of the most toxic upon contact to a range of soil-borne organisms, including pathogenic fungi (Sarwar *et al*., 1998), root lesion nematode (Potter *et al*., 2000), and polyphagous soil insects (Borek *et al*., 1998).

Breeding experiments by Potter *et al.* (2000) have shown that it is possible to selectively breed plant containing higher levels of 2-phenylethyl GSL in root without changing GSL levels in shoot or seed. This suggests that GSL profiles may be regulated and selected independently in shoots and roots.

1.6.2.3 Anti-nutritional factor in rapeseed cake

Hydrolysis of *β*-hydroxyalkenyl GSLs (e.g. progoitrin and gluconapoleiferin), produces unstable *β*-hydroxyalkenyl ITCs that spontaneously cyclise to form oxazolidine-2-thiones (Mawson *et al*., 1993). Although thiocyanates and ITC also contribute to the goitrogenic effects in mammals by competing with thyroidal transport thus decreases iodide uptake and thyroid hormone production, oxazolidine-2-thiones have a greater impact because they interfere with the synthesis of thyroid hormones in animals (Langer and Greer, 1977). The effect from oxazolidine-2-thiones cannot be alleviated by iodine supplementation (Langer and Greer, 1977). The anti-nutritional effect of oxazolidine-2-thiones on animals including rats, swine, poultry and ruminants health have been reviewed (Mawson *et al*., 1994).

The high content of goitrogenic GSLs in the seeds of oilseed rape plants have led to major breeding efforts to minimise these substances in seed meal for animal feeds. The development of the low GSL rapeseed or canola will be reviewed in the next section (Section 1.7.2).

1.7 *Brassica napus* – crop of global importance and genetic complexity

1.7.1 Rapeseed global production

To date, rapeseed is the world's third largest oil-producing crops after palm and soybean (Table 1-4) and is the second largest oilseed source of protein meal in the world for the 2018/19 (Table 1-5) (USDA, 2019). The term 'rapeseed' used in the trading and market reports is a collective name for a number of *Brassica* oilseed species, which includes *B. napus* (oilseed rape), *B. rapa* (turnip rape) and *B. juncea* (brown mustard). The exact number for each species varies greatly between countries. The leading producers of rapeseed and products are shown on Table 1-6. In Europe and Canada, the production is mainly from *B. napus* as it is a higher yield crop in cool temperate climate with higher rainfalls. On the other hand, *B. juncea* thrives in hot regions with lower rainfalls and is grown mainly in India and China (Mawson *et al*., 1993).

Table 1-4. World production of vegetable oils (Million Metric Tons) (USDA, 2019). Top four oilseeds are shown

Oilseed	2016/17	2017/18	2018/19
Palm	65.23	70.46	73.48
Soybean	53.72	55.17	56.97
Rapeseed	27.54	28.10	28.03
Sunflowerseed	18.16	18.25	19.45

Table 1-5. World production of protein meals (Million Metric Tons) (USDA, 2019). Top four oilseeds are shown.

Oilseed	2016/17	2017/18	2018/19
Soybean	225.55	232.55	238.10
Rapeseed	38.80	39.54	39.54
Sunflowerseed	19.35	19.59	20.90
Cottonseed	13.44	15.77	15.76

Table 1-6. Top four world supplier of Rapeseed and products in 2018/19 (Thousand Metric Tons) (USDA, 2019).

Production	Meal	Oil	Oilseed
Canada	5,275	4,150	21,100
European Union	13,167	9,656	19,600
China	10,032	6,630	12,850
India	4,050	2,584	8,000

1.7.2 Low glucosinolate rapeseed and canola

To allow the use of seed meal as animal feed, extensive breeding efforts have been made to select for oilseed rape cultivars with low seed GSLs. The identification of Bronowski, a low-GSL Polish spring rape cultivar in the 1970s, has provided the genetic source for the basis of all other commercial low-seed GSL cultivars through selective breeding (Rosa *et al.*, 1997). This reduction is almost entirely due to the decrease of aliphatic GSLs level in Bronowski (Kondra and Stefansson, 1970; Rucker and Rudloff, 1991). 'Double-zero' or 'canola', a double low cultivar with low seed erucic acid (<2%) and seed GSLs (<30 µmol/g), has been subsequently developed.

The introduction of the double-zero cultivars has led to the concern that these cultivars could be more susceptible to pests and diseases, due to reduction of the presumed defensive role of GSL. Nevertheless, levels of GSLs and their interaction with plant pests may be more intricate than previously thought because the same GSL profile can acts as both deterrent to generalist pests and stimulant to specialist pests (as discussed in Section 1.6.2.1) (Mithen, 1992; Giamoustaris and Mithen, 1995; Hopkins *et al.*, 2009). The loci controlling GSL content of seeds have been shown to coincide with *MYB28/HAG1 (Harper et al., 2012; Chalhoub et al*., 2014; Lu *et al*., 2014) indicating that control of biosynthesis plays a part in forming the double-zero cultivars. On the other hand, it is speculative to what extent long-distance transport play in the establishment of the low GSL trait in *B. napus*.

1.7.3 *B. napus* genomes

Within Brassicaceae family, six species from *Brassica* genus have great economic values to humanity as culinary vegetables, condiments and oilseeds. The genetic relationships of these six close relatives can be described by the U's triangle model (Figure 1-6). The model has been developed through extensive experimental crosses between the species and microscopic inspection of meiosis in these crosses (U, 1935). *B. napus* (AC genome, n=19) is a recently formed allotetraploid arisen from a hybridisation between a pair of ancient diploid progenitors, *B. rapa* (A genome, n=10) and *B. oleracea* (C genome, n=9) (U, 1935). This hybridisation of two genetically related genomes can cause genetic instability in the newly formed allopolyploid because of the genome-wide gene redundancy and a high risk of homoeologous chromosome pairing during meiosis which may reduce its fertility (Ma and Gustafson, 2005). To achieve genome stabilisation, newly formed allopolyploid genomes undergo a process termed 'diploidisation', which includes rapid structural changes such as changes in gene copy numbers (Adams and Wendel, 2015). In fact, homoeologous exchanges, where a lost chromosomal region is replaced by a duplicated copy of corresponding homoeologous region of the other genome, are frequent between the A and C subgenomes in *B. napus* (Chalhoub *et al*., 2014; He *et al*., 2016). These homoeologous exchanges cause changes in the gene copy numbers that may affect functional traits.

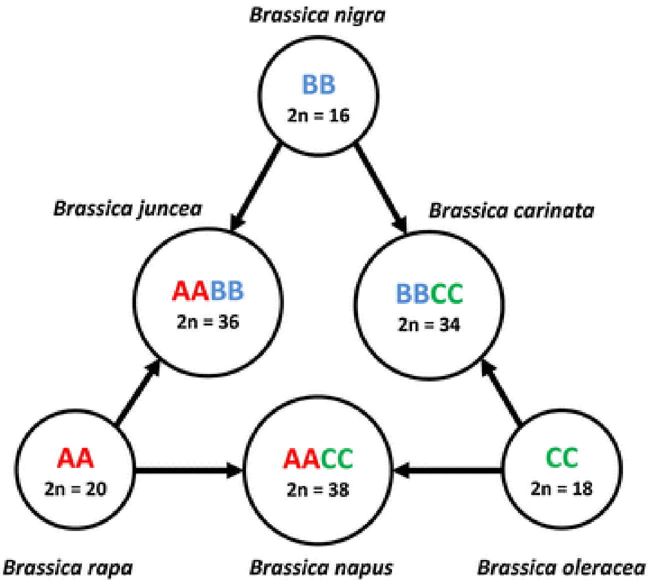


Figure 1-6. Genetic relationship of Brassica species within the U's triangle. The three diploid species (*B. rapa, B. nigra* and *B. oleracea*) represent the AA, BB and CC genomes. Hybridisation of the diploid species give rise to three allotetraploid species (*B. juncea*, *B.napus* and *B. carinata*). Diploid chromosome number (2n) is shown. Image from Koh *et al.* (2017).

1.7.3.1 Relatedness with the model plant Arabidopsis thaliana

B. napus and *A. thaliana*, a model plant for genome analysis, have a close phylogenetic relationship as both are members of the Brassicaceae family. The divergence of the ancestral *Arabidopsis* and *Brassica* has been estimated as approximately 20 million years ago (Figure 1-7). Comparative mapping between *Brassica* species and *A. thaliana* has found that syntenic regions in *A. thaliana* are present in a multiple of three within the diploid *Brassica* genomes (O'Neill and Bancroft, 2000; Parkin *et al*., 2002), supporting the hypothesis that diploid *Brassica* species may have evolved through a hexaploid ancestor. High density comparative mapping has shown the gene content of *A. thaliana* to be very similar to that of the *Brassica* diploids, with more than 87% sequence identity in the coding regions (Parkin *et al*., 2005). Furthermore, work by the same group has found that the 21 conserved segments identified within the Arabidopsis genome could be duplicated and rearranged to cover almost 90% of the mapped length of *B. napus*. The majority of the identified conserved segments revealed an average of 4.4 functional gene copies present in *B. napus* (Parkin *et al*., 2005). Although difference exists between *Brassica* and Arabidopsis orthologues (Rana *et al*., 2004), the highly conserved gene content and gene order between the *A. thaliana* and *B. napus* genomes has enabled the exploitation of the model plant Arabidopsis to infer genetic basis of traits in the *Brassica* crops.

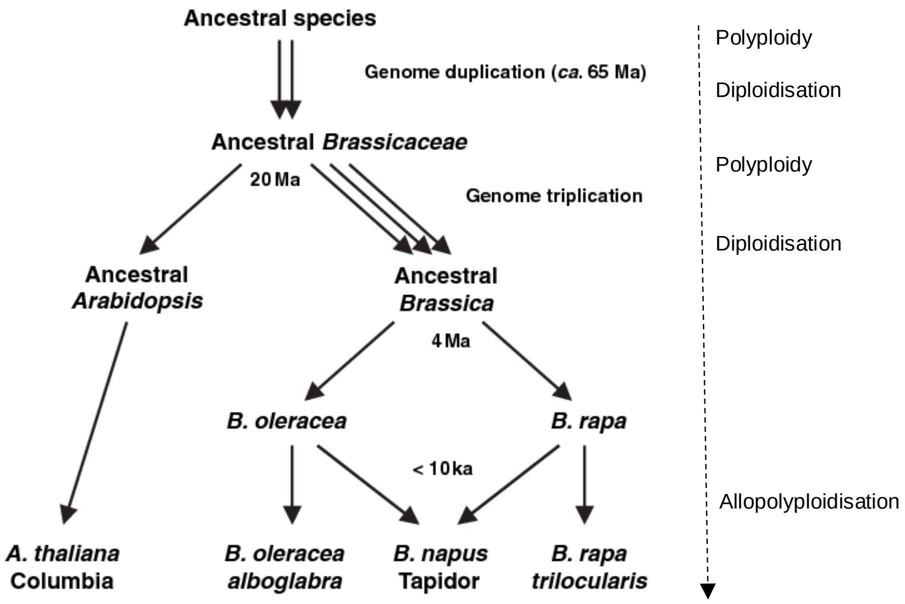


Figure 1-7. Schematic summary of selected *Brassica* species and *Arabidopsis thaliana* genome evolution. Image from Rana *et al.* (2004).

1.8 Transcriptome-based molecular marker system

1.8.1 The need to develop new tools for *B. napus* research

Although the number of chromosomes differs between *B. rapa* and *B. oleracea*, genetic mapping has confirmed the organisation of their genomes to be highly collinear (Lagercrantz and Lydiate, 1996). Most of the ~1200 Mb *B. napus* genome (Arumuganathan and Earle, 1991) comprises of highly related (homoeologous) fragments from A and C genomes progenitors that are difficult to distinguish from one another (Bancroft *et al*., 2015). Although the draft genome sequence has been obtained for *B. napus* (Chalhoub *et al*., 2014), the large genome size and the highly duplicated nature has made it difficult to be assembled to a very high standard.

To address these research challenges, rapid and cost-effective mRNAseq-based technologies have been developed and implemented with great success in *B. napus* in the following areas including: the identification of single nucleotide polymorphism (SNP) among cultivars (Trick *et al*., 2009), linkage mapping and genome organisation studies (Bancroft *et al*., 2011), transcript quantification of homoeologue gene expression (Higgins *et al*., 2012), and association genetics termed 'Associative Transcriptomics' (Harper *et al*., 2012) (see section 1.8.2). In these studies, mRNAseq sequence reads are mapped to an appropriate transcriptome reference sequence for both SNP discovery and transcript quantification. The recent development of a pan-transcriptome resource incorporating the diploid *Brassica* A and C genomes has provided a more reliable reference sequence supporting an existing transcriptome-based technologies as well as a solution to determine genome-of-origin of any given genes in the *B. napus* genome (He *et al*., 2015). This development enables homoeologous genes to be assigned to a particular genome, a task that was difficult to accomplish previously. The reference sequence was achieved by constructing the coding DNA sequence (CDS) gene models to form a comprehensive reference sequence. These CDS gene models are primarily derived from the components of the genomes from the *B. napus* diploid progenitors, i.e. the A genome from *B. rapa* and the C genome from *B. oleracea*, plus other

B. napus-specific CDS models that do not have any orthologues in the genomes of the two progenitor species (Figure 1-8) (He *et al*., 2015).

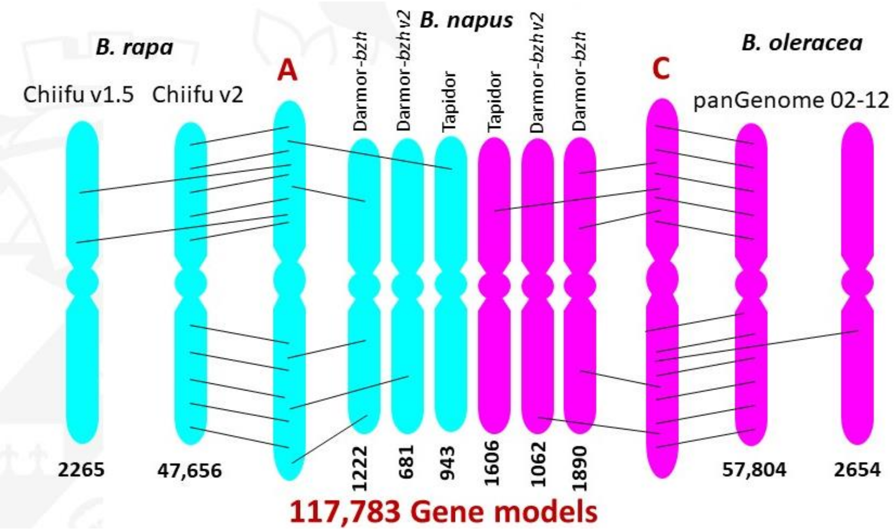


Figure 1-8. The construction of *Brassica* A and C coding DNA sequence (CDS) gene model-based pan-transcriptome reference. Detailed method of the assembly are described in He *et al.* (2015). In brief, the ordering and mapping of *B. rapa* Chiifu genome sequence scaffolds formed majority of the A genome, while *B. oleracea* T1000 genome sequence formed majority of the C genome. The *B. napus*-specific CDS models from *B. napus* Darmor-*bzh* and Tapidor are also interpolated to the A and C genomes. The current number from each sources contributing to the final CDS gene models are shown under the chromosomes (Image from Bancroft lab's presentation July 2018).

1.8.2 Associative Transcriptomics – a tool to study genes associated with complex traits

One of the aims of genetic mapping studies is to identify quantitative trait loci (QTL) that determine variations in phenotypic traits. Genome-wide association study (GWAS) is a powerful method of dissecting genes associated with complex traits and has several advantages over bi-parental QTL analysis (Zhu *et al.*, 2008). The main advantage of association studies is that it uses genetically diverse population to identify QTLs by exploiting all recombination events that have occurred in the evolutionary history of a sample, resulting in a much higher mapping resolution compared to QTL mapping (Yu and Buckler, 2006). The resolution provided by association studies is dependent on the degree of linkage disequilibrium (LD) between the genotyped marker and the functional variant within a genome. LD is defined as the non-random correlation between alleles or polymorphisms (e.g. SNPs) in a population that is caused by their shared history of mutation and recombination (Flint-Garcia *et al.*, 2003). Natural population undergo several rounds of historical recombination which break up the genome into small fragments of highly correlated alleles in high LD. The key to GWAS is to genotype enough markers across the genome, so that when a marker shows an association with a phenotype of interest it is more likely that the genotyped markers will be in LD and physically linked to the causative allele (Myles *et al.*, 2009).

Associative Transcriptomics (AT) is a transcriptome-based GWAS approach that has been developed to overcome the challenges from polyploidy genome complexity that hinder genomic-based study in *B. napus* (Harper *et al.*, 2012). By using transcriptome sequencing instead of genome sequencing, high levels of functional genetic variants can be captured in the coding sequences, regions that may be considered the most functionally important, while reducing the scale and cost of the sequencing project (Trick *et al.*, 2009). Moreover, apart from discovering SNP markers in tight LD with causative genes as in conventional GWAS, the use of transcriptome sequencing in AT allows identifying of gene expression markers (GEMs) from the correlation between expression patterns and trait variation (Harper *et al.*, 2012). The combined power of SNP and GEM variation provides a powerful approach to study the underlying genetic architecture of marker-trait association.

Many studies have demonstrated the potential of AT in identifying genes controlling complex traits in various species, such as erucic acid in *B. napus* (Havlickova *et al.*, 2018), stem strength in bread wheat (Miller *et al.*, 2016) and dieback disease tolerance of European ash (Harper *et al.*, 2016). With regards to GSL studies, AT has also been effectively applied to identify SNPs and GEMs on chromosome A9, C2 and C9 that are highly associated with variations in total seed GSLs (Harper *et al*., 2012). Following this initial AT analysis, two additional association peaks have also been detected on chromosome A2 and C7 relating to variations in total seed GSLs (Lu *et al*., 2014). Within these five associated loci is the orthologues of *HAG1* (Harper *et al*., 2012; Lu *et al*., 2014).

1.9 Thesis objectives

There are implications for understanding the modular genetic system that regulates GSL natural variations in *B. napus* as a whole. This knowledge could lead to crop improvement by exploiting GSL potentials and manipulating GSL profiles for modulation of interactions between important crop plants and its pests. The aim of this research project is to identify the genetic controls underlying natural variations of GSLs in *B. napus* leaves and roots, as well as develop understanding of their connections with seed GSLs. Thus far, previous association studies have been solely focused on identifying genetic markers associated with seed GSL traits (Li *et al*., 2014; Lu *et al*., 2014; Gajardo *et al*., 2015) whereas the study of the genetic controls of GSLs in vegetative tissues have been neglected. The identification of biosynthetic genes and regulators of aliphatic and indole GSLs, as well as characterisation of the GSL transporters in *A. thaliana*, have helped clarify and identify orthologous GSL pathway genes in closely related *Brassica* species. However, several research questions remained to be investigated, these will be addressed in following chapters:

i) What are the genetic controls of GSL variations in *B. napus* leaf and root tissues? (Chapter 4 & Chapter 5)

ii) Which genes are involved in the side chain elongation and its regulation of the homophenylalanine-derived aromatic GSLs, a class that is abundant in *Brassica* roots? (Chapter 5)

iii) What is the relationship between GSLs in the leaves, roots and seeds of *B. napus*? (Chapter 3 & Chapter 6)

iv) Since accumulation pattern of GSLs may be established by biosynthesis and/or transport, what is the underlying basis of low GSL varieties in *B. napus*: regulation of biosynthesis or transport? (Chapter 6)

To address these research questions, the AT platform with 355,536 SNP markers on the transcriptome reference comprising 116,098 ordered CDS gene models (Havlickova *et al.*, 2018) was used on a panel of 288 *B. napus* accessions with leaf and root GSLs as traits. This study provides the first comprehensive analysis of the leaf and root GSL profiles from a large diversity panel, generating a total of 32,256 data points of GSL data. The focus of this thesis was on the analyses of GSL structural classes (as aliphatic, indole and aromatic GSLs) rather than individual GSLs, with the intention to elucidate potential master regulators that control the general GSL accumulation patterns. Prior to the essential large scale GSL profiling required for hundreds of accession lines, a simpler and efficient method was developed for extracting GSLs from *Brassica* leaf and root tissues (Doheny-Adams *et al.*, 2017). This method was then used for all the quantification of GSLs presented in Chapter 2 of this thesis.

Chapter 3 provides the foundation for studying the genetic and biological activities of GSLs by analysing the GSL profiles in the leaves and roots of *Brassica napus*. Variations of GSLs in vegetative tissues will be examined through population analysis and frequency distribution of phenotypic traits will be shown. As the diversity panel is made up of seven crop types based on the clustering of SNPs, GSL contents will be examined to compare the pattern of GSL accumulation in different crop types. Estimated heritability of the trait will be calculated to understand whether and to what extent the observed variation is influenced by genetic components. Correlation analysis of the GSL profile will be conducted to investigate relationships within and between the vegetative tissues.

Chapter 4 focuses on the investigation of the genetic control of leaf GSL variation, with the emphasis on identifying the genetic control of aliphatic GSLs since these are the most abundant type in the leaf tissue and largely determine the variations of the total leaf GSLs. AT will be used to identify the candidate genes. As a polyploidy organism, *B. napus* contains multiple copies of the same gene. The pattern of homoeologous gene expression will be studied with variations in levels of leaf aliphatic GSLs to investigate which copies of the candidate genes are involved. Finally, the impact of structural rearrangement will be investigated to identify genomic regions containing candidate genes which influences variation in aliphatic GSL concentrations. Parallel analysis on indole and aromatic leaf GSLs will also be carried out and discussed.

Chapter 5 aims to identify the genetic control of root GSL variation, with a focus on the genetic control of aromatic GSLs as these largely determine the variation of total root GSLs. Candidate gene will be identified through SNP-based association studies. In parallel with AT analysis, differential expression analysis will also be performed to investigate root-specific gene expressions. Genes with highly correlated expression pattern to candidate genes will be analysed through weighted gene co-expression network analysis. Additional AT analyses using ratios of GSLs as trait will be carried out to get a further insight into the underexplored root GSLs.

Chapter 6 will explore the relationship between GSL content of vegetative tissues and seeds, by performing a Spearman's correlation analysis after adding the seed GSL dataset from Lu *et al.* (2014) to the leaf and root dataset described in Chapter 3. AT will be analysed for the association of known GSL transporters to further investigate whether variation in transport or biosynthesis processes explained the variation in aliphatic GSL contents between leaf and seed. Manhattan plots and patterns of GSL accumulation will be compared and analysed for the underlying genetic basis for the observed variation.

Lastly, chapter 7 will conclude with the discussion of the results. It will address how evolution and breeding may impact the variations in the population structure and how this link to the variations in GSL concentrations of *B. napus* modern varieties. Evaluation and implication of the study will be reviewed. Finally, the directions of future work will be discussed.

CHAPTER 2

General Material and Methods

The following are the general material and methods that are relevant to the works carried out on the diversity panel as a whole. Where relevant, specific method and subset of accessions used for further analyses will be described in each chapter.

2.1 Diversity panel plant material and harvesting

A diversity panel of 288 *B. napus* accessions from the 'RIPR panel' (Renewable Industrial Products from Rapeseed) (Havlickova *et al*., 2018) were grown in long day (16/8 h, 20 °C/14 °C) under controlled glasshouse conditions (University of York, UK). Within this panel, there are 56 Modern Winter OSR, 65 Winter OSR, 6 Winter Fodder, 121 Spring OSR, 26 Swede and 14 Exotic varieties (Appendix 1 & 2). Four biological replicates of each accession were grown in root trainers with Terra-Green for ease of root harvesting, supplemented weekly with a half concentration of Murashige and Skoog growth medium (Murashige and Skoog, 1962) adjusted to pH6.5 with KOH. The experiment was arranged as randomised four-block design with one plant per lines in each block. Four weeks after sowing, the third true leaf and the whole root system were harvested from each plant. At harvest, leaves were cut at the base, wrapped in a labelled aluminium foil and immediately frozen in liquid nitrogen. Plants were removed from the tray, the roots were washed, dried with paper towel and cut at the base. All samples were wrapped in labelled aluminium foils and immediately frozen in liquid nitrogen and stored at -80 °C.

2.2 Determination of glucosinolate by HPLC

The GSL quantification method used in this project had been developed to increase the extraction efficiency while reduce time consumption and liberate the use of hazardous hot methanol, which makes this method suitable for large-scale GSL profiling on vegetative tissues. This method simply uses methanol at room temperature and has comparable or improved GSL extraction efficiency relative to the commonly used ISO method (Doheny-Adams *et al*., 2017). An overview of the GSL analysis is illustrated in Figure 2-1.

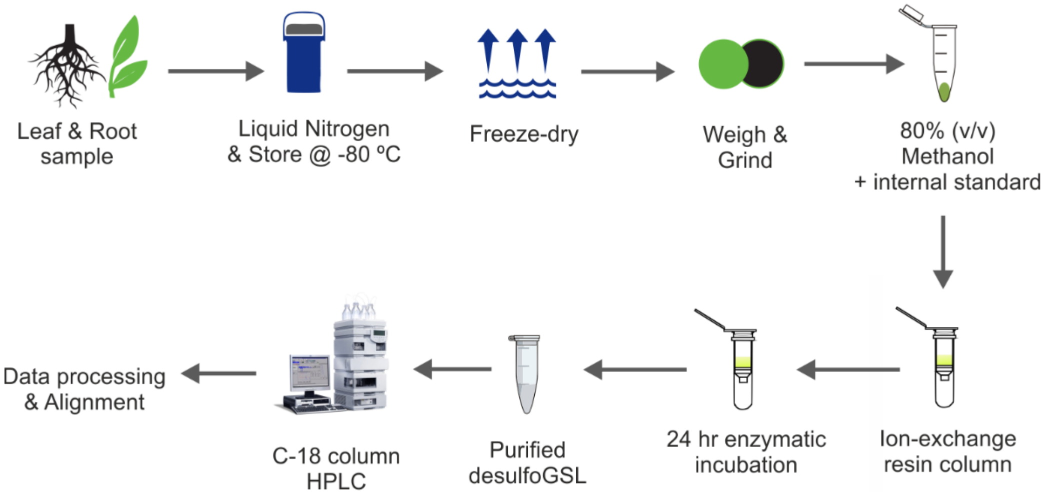

Leaf & Root sample
Liquid Nitrogen & Store @ -80 °C
Freeze-dry
Weigh & Grind
80% (v/v) Methanol + internal standard
Ion-exchange resin column
24 hr enzymatic incubation
Purified desulfoGSL
C-18 column HPLC
Data processing & Alignment

Figure 2-1. Overview of the glucosinolate determination process

2.2.1 Sample preparation and glucosinolate extraction

Myrosinase is an enzyme that breaks down GSL in the presence of water upon plant tissue disruption. To avoid GSL breakdown by myrosinase in plant tissues and get an accurate measurement of GSLs, lyophilisation or freeze-drying was carried out on frozen samples prior tissue grinding to remove water from the tissues. This process also inactivates myrosinase hydrolysis on GSL through thermal inhibition (Doheny-Adams *et al*., 2017). After 10 hrs of lyophilisation or completely dried, samples were homogenised to fine powder with two steel beads for 10 min at a frequency of 30 Hz (TissueLyser II, Qiagen). To 50 mg of homogenate, 1975 µl of 80% (v/v) methanol was used as the extracting solvent at room temperature (23 °C), and 25 µl of 5 mM glucotropaeolin was added as the internal standard. The samples were mixed and left to stand for 30 min at 20 °C, and then mixed with orbital shaker (Vibrax, IKA) at 1200 rpm for 30 min before centrifugation at 8000 rpm for 10 min. Each of the supernatant methanol extract was then transferred to a pre-conditioned Sephadex column in the following purification step.

2.2.2 Purification and desulfication

Purification and desulfation of GSLs was carried out following the ISO 9167-1 (1992) methods. Desulfation with sulfatase was performed on an ion-exchange column consisted of Sephadex beads, which also served for the sample clean-up step. Columns (Sigma-Aldrich, C2728) were prepared with 0.5 ml ion-exchange resin (DEAE Sephadex beads in 1:1 ratio with 2 M acetic acid), conditioned with 2 ml imizadole formate (6 M) and washed twice with 1 ml water. Conditioning and washing of columns were carried out under a fume hood. One millilitre of the sample extract was transferred to a prepared column. Columns were gently washed twice with 1 ml 20 mM sodium acetate (pH 4) (Figure 2-2A) before adding 75 µl of purified sulfatase from Helix pomatia type H-1 (5 U/ml) (Sigma, S9626) (Figure 2-2B). Columns were sealed with parafilm and incubated for 24 h. Desulfoglucosinolates (ds-GSLs) were then eluted from the columns with two 1 ml portions of water (Figure 2-2C) and stored at -20 °C before HPLC analysis. The non-ionic ds-GSLs are a more stable form and are well-suited for reverse phase HPLC.

2.2.2.1 Chemistry of the anion exchange column

Intact GSLs are anionic (negatively charged) due to its *cis*-N-hydroximino sulphate ester. The ion-exchange resin, DEAE Sephadex, is a weak anion exchanger with cationic (positively charged) stationary phase particles. During sample loading, negatively charged sulphate ester of GSLs bind to the Sephadex beads, while the uncharged or cationic contaminants in the sample are washed out with the sodium acetate. Application of sulfatase catalyses the hydrolysis of sulphate ester, which releases the non-ionic ds-GSLs in the final elution with water to form a mixture of pure ds-GSLs for HPLC analysis (Figure 2-2).

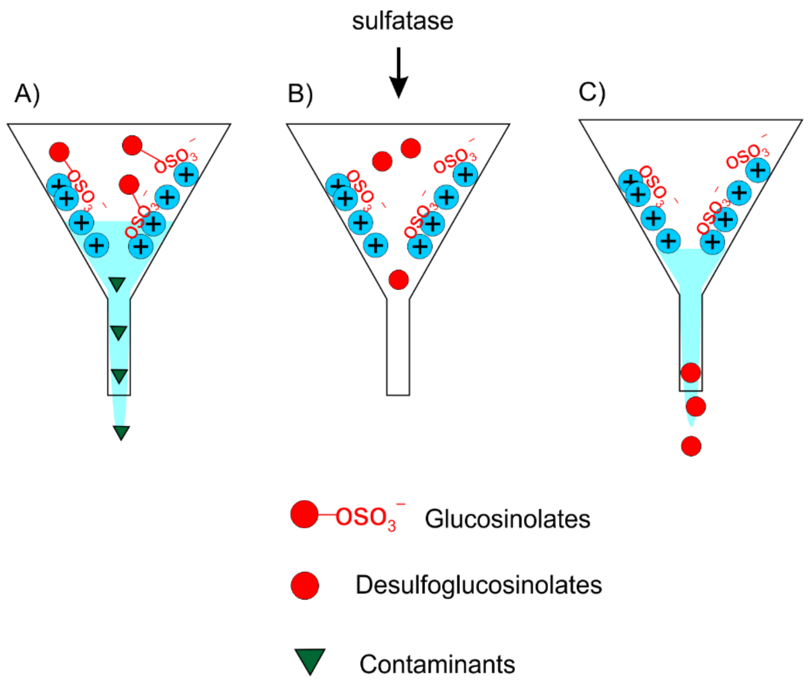


Figure 2-2. Anion exchange chromatography. A) DEAE Sephadex resins bind the anionic sulphate ester moiety of glucosinolates to the column. Washing of the column removes impurities from the sample. B) Incubation with sulfatase catalyses the hydrolysis of ester sulphate, which disulphates glucosinolates and releases them into the solution. C) Pure mixture of desulfoglucosinolates are collected with the final elution.

2.2.3 HPLC analysis of desulfoglucosinolates

Desulfoglucosinolates were separated by HPLC (Millipore 600E system, Waters) on a reverse phase C18 column at 30°C (Phenomonex, SphereClone 5µ ODS(2), 150 mm × 4.6 mm) with mobile phase solutions consisting of 100% diH$_2$O and 30% (v/v) acetonitrile, as detailed in Doheny-Adams *et al*. (Doheny-Adams *et al.*, 2017). Injection was at 10 µl and flow rate was set to 1 ml/min. Samples were separated according to the program shown on Table 2-1. The absorbance of the eluates was monitored at 229 nm wavelength within the UV spectrum.

Table 2-1. Mobile phase conditions used in HPLC for separation of ds-GSL

Time	100% diH$_2$O	30% (v/v) acetonitrile	Transition
0	100	0	
30	0	100	Linear gradient
35	0	100	
40	100	0	Linear gradient
50	100	0	

2.2.3.1 *Glucosinolate identification*

Through standard injections, HPLC-MS identification, retention time and photodiode array (PDA) UV spectra, the identity of all major GSLs in this study were confirmed (see Table 2-2 for the identification methods of each GSL). Ds-GSLs were first identified by comparing the retention time of the peaks and UV spectrum to those of commercial purified ds-GSL standards under the same HPLC apparatus and conditions. Further identification of major GSLs for which no commercial standard is available was carried out by electrospray ionisation mass spectrometry (ESI-MS) (Biology Technology Facility, University of York). Ds-GSL samples going through the mass spectrometry were ionised and separated according to their molecular mass. The molecular weight (MW) of each ds-GSLs in the samples was calculated from this molecular mass of ionised Na$^+$ bound ds-GSLs (M+Na$^+$) and compared to the literature (Hirani *et al.*, 2012). The use of HPLC PDA detector aids the identification of GSLs by revealing UV-absorbance characteristics that are specific to each GSL side-chain groups (Figure 2-3). Associating UV spectra to retention times are particularly useful because retention times and elution order of some ds-GSLs can change depending on the type of columns, solvent flow, column pressure and elution gradient conditions (Wathelet *et al.*, 2004). Once the order and retention time of the GSL has been assigned, subsequent GSL identification uses the UV-absorbance characteristic and alignment of the peak against the reference retention time (Table 2-2). Figure 2-4 shows an example of HPLC chromatograms ds-GSL peaks. Using combination of these analytical methods together provides a reliable method for identifying GSLs.

Table 2-2. Information for the identification of ds-GSLs: retention times, response factors, molecular weight and analytical methods used to confirm identity.

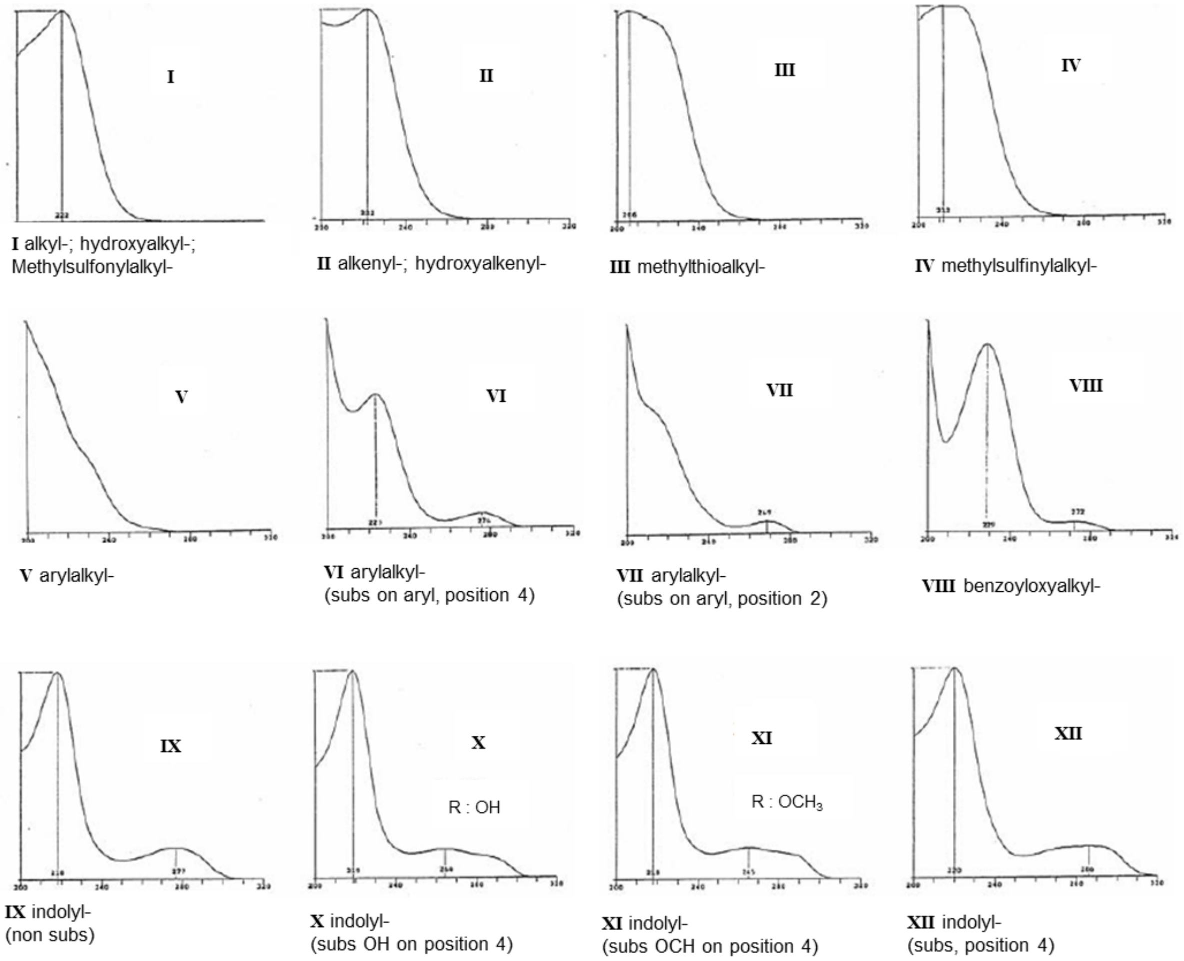
Ds-GSL	RT (min)	RF[1]	RRF	MW ds-GSL[2]	Identification methods		
					Stds	ESI-MS	UV characteristic[3]
GIB	3.9	1.07	1.13	343	✓		Methylsulfinylalkyl (IV)
PRO	4.8	1.09	1.15	309	✓	✓	Hydroxyalkenyl (II)
SIN	5.5	1.00	1.05	279	✓		Alkenyl (I)
GRA	6.9	1.07	1.13	357	✓		Methylsulfinylalkyl (IV)
GRE	7.8	0.90	0.95	355	✓		Methylsulfinylalkyl (IV)
GJV	11.0	1.00	1.05	281		✓	Alkenyl (II)
GAL	12.6	1.07	1.13	371	✓	✓	Methylsulfinylalkyl (IV)
GNA	13.5	1.11	1.17	293	✓		Alkenyl (II)
4-OHGBS	15.2	0.28	0.29	384	✓	✓	Indolyl (X)
GBN	17.6	1.15	1.21	307	✓	✓	Alkenyl (II)
GTL	18.3	0.95	1.00	329	✓	✓	Arylalkyl (V)
GER	18.6	1.04	1.09	341	✓		Methylthioalkyl (III)
GBS	20.6	0.29	0.31	368	✓	✓	Indolyl (IX)
GST	23.3	0.95	1.00	343	✓	✓	Arylalkyl (VII)
4-OMeGBS	24.2	0.25	0.26	398	✓	✓	Indolyl (XI)
Neo-GBS	26.8	0.20	0.21	398		✓	Indolyl (XII)

RT = retention time; RF = response factor; RRF = relative response factor; MW = molecular weight; Stds = standards; ESI-MS = electrospray ionisation mass spectrometry

[1] Collated recommended RF from Clarke (2010)

[2] Hirani *et al.* (2012)

[3] Based on side-chain group then compare to PDA on Figure 2-3 (Wathelet *et al.*, 2004)



I alkyl-; hydroxyalkyl-; Methylsulfonylalkyl-

II alkenyl-; hydroxyalkenyl-

III methylthioalkyl-

IV methylsulfinylalkyl-

V arylalkyl-

VI arylalkyl- (subs on aryl, position 4)

VII arylalkyl- (subs on aryl, position 2)

VIII benzoyloxyalkyl-

IX indolyl- (non subs)

X indolyl- (subs OH on position 4)

XI indolyl- (subs OCH on position 4)

XII indolyl- (subs, position 4)

Figure 2-3. UV spectra of different groups of ds-GSLs, generated from spectrophotometer with spectral scanning 200-350 nm (Wathelet *et al*., 2004).

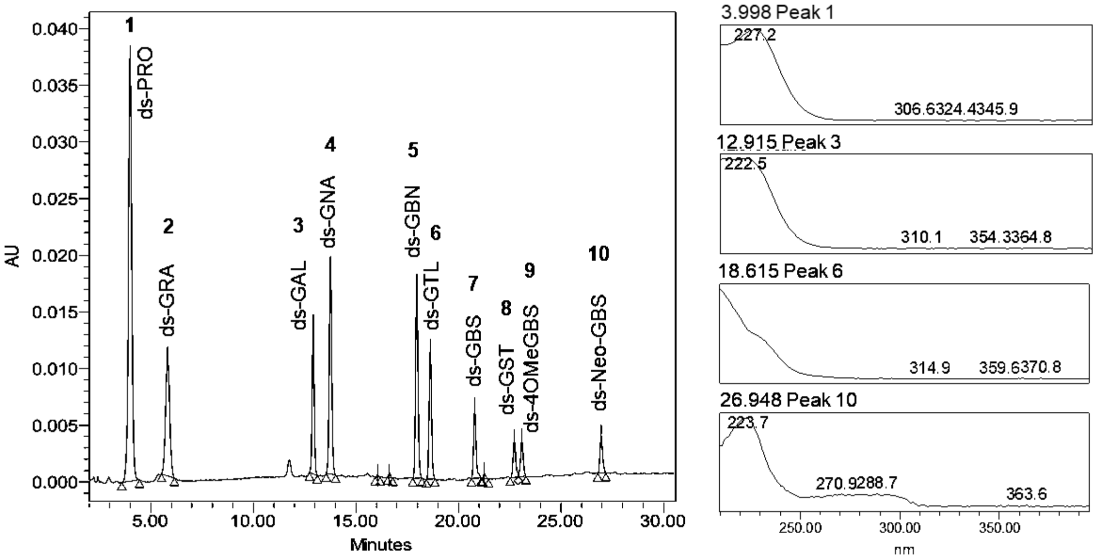


Figure 2-4. HPLC chromatograms example of ds-GSLs from a leaf of *B. napus* showing peak retention times with example of UV-absorbance characteristic.

2.2.3.2 Calculation of glucosinolate content

GSL contents were determined from the peak area of the ds-GSLs. The content of GSLs, expressed in µmol/g, were calculated according to ISO 9167-1 (1992):

$$Glucosinolate\ content = \frac{A_g}{A_s} \times \frac{n}{m} \times K_g \times \frac{100}{100\ - w}$$

where A_g and A_s are the peak areas corresponding to a single type of GSLs in the sample and the internal standard GSL (i.e. glucotropaeolin), respectively; and K_g is the response factor; *m* is mass in grams; *n* is the quantity of glucotropaeolin internal standard in µmol and *w* is the moisture and volatile matter content. The response factor for each analyte was taken from the collated recommended values (Clarke, 2010), and was calculated relative to the internal standard glucotropaeolin (i.e. GTL) for this study (see RRF on Table 2-2). For the unassigned GSLs, class-specific response factor values, i.e. 1.0 for aliphatic, 0.3 for indoles and 0.7 for aromatic, were used. Unlike the default method of assigning the value of 1.0 to any unassigned GSLs, these values were assigned to be close to the average RF for each structural classes as to improve accuracies of the calculation.

2.3 Data cleaning

During the raw data processing, GSLs that occurred less than 5% frequency across all the samples or occurred at low levels of detection (<0.01 µmol/g) were filtered out. This removed all of the uncharacterised GSLs from the dataset and resulted in the final list of fourteen identified GSLs. Outlier detection was performed using standard deviation method to limit erroneous in the final dataset which may be due to variability in the measurement or experimental error. Three standard deviations from the mean is a common cut-off practice for identifying outliers. Outliers were excluded since they can cause problems in statistical analysis. This data quality control removed 5% of the samples from the dataset. The resulting dataset contains at least two biological replicates for each accession, with majority of the accessions having four biological replicates. The replicates were used to calculate the mean GSL contents for each accession in this final dataset.

2.4 Associative Transcriptomics

The same accession RIPR panel was used to generate functional genotypes from leaf RNA for AT analysis. Figure 2-5 illustrates an overview of the AT platform. Plant growth conditions and methods for material sampling, RNA extraction and Illumina sequencing were as described in He *et al.* (2016). AT analysis was performed using custom scripts in R (2013) based on an adaption of the first AT methods (Harper *et al*., 2012). The R scripts were modified to accommodate for larger dataset, as detailed in Havlickova *et al.* (2018). The transcriptome reference sequences used in the current AT platform are ordered *Brassica* A and C pan-transcriptomes, which comprised of 116,098 ordered coding DNA sequence (CDS) gene models derived primarily from progenitor species *Brassica* A and C genomes and interpolated *B. napus*-specific CDS models (He *et al*., 2015). AT analysis included two parallel approaches based on two different types of data matrix derived from the same sequencing dataset: single nucleotide polymorphisms (SNPs) and transcript abundance.

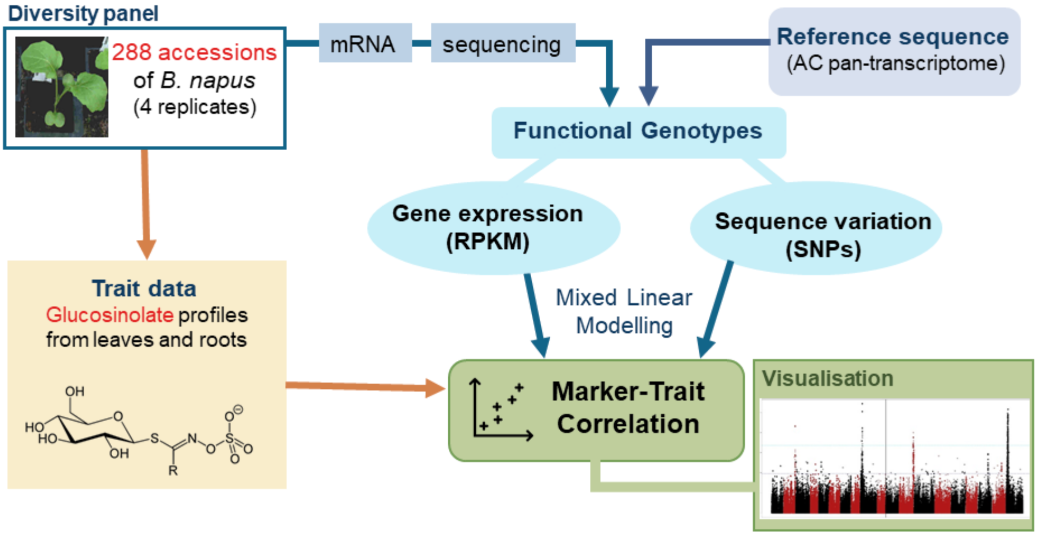


Figure 2-5. Overview of Associative Transcriptomic analysis

2.4.1 Association analysis for SNPs

SNPs were called by the meta-analysis of mRNAseq read alignments from each of the *B. napus* accessions as described previously (Bancroft *et al*., 2011). SNP positions were excluded from further analysis if they have a read depth below 10, or a base call quality below Q20, or missing data below 0.25, or contained three SNP alleles or more. A SNP matrix with a total of 355,536 SNP markers was generated after this rigorous filtering and quality checking with the above parameters to reduce errors in SNP identification and an assessment of linkage disequilibrium as detailed in Havlickova *et al.* (2018). To reduce the risks of false positive associations from undetected population structure that can mimic the signal of association, population structure inference using kernel-PCA and optimisation (PSIKO) (Popescu *et al*., 2014) was used for Q-matrix generation to correct for population stratification. Genome Association and Prediction Integrated Tool (GAPIT), an `R` package with compressed mixed linear model that includes both fixed and random effects (Lipka *et al*., 2012), was used for AT analysis of the 288 *B. napus* accessions in combination with the GSL trait data.

For the Manhattan plots of SNP associations, SNP markers with minor allele frequencies (MAF) below 0.01 were removed from the SNP dataset leaving 256,397 SNPs for the associations (Havlickova *et al*., 2018). SNP markers that can be assigned with confidence to the genomic position of the CDS model were rendered dark points. However, due to the high sequence similarities between the A and C genomes, it is not always possible to assign the genome of the polymorphism with confidence. For such markers, these ambiguous points were rendered pale colouration. SNP markers were positioned on the *x*-axis based on the genomic order of the CDS gene model in which the polymorphism was scored. The significance of the trait association, as $-\log_{10}P$ values, was plotted on the *y*-axis. Presence of large peaks in the Manhattan plot passing both the false discovery rate (FDR) threshold at 5% (lower line, purple) and the threshold for Bonferroni significance of 0.05 (upper line, cyan) were used to suggest that the surrounding genomic region has a strong association with the trait. For instance, Figure 2-6 reveals strong SNP association peaks on chromosome A2, A9, C2 and C9 with a minor association on chromosome C7.

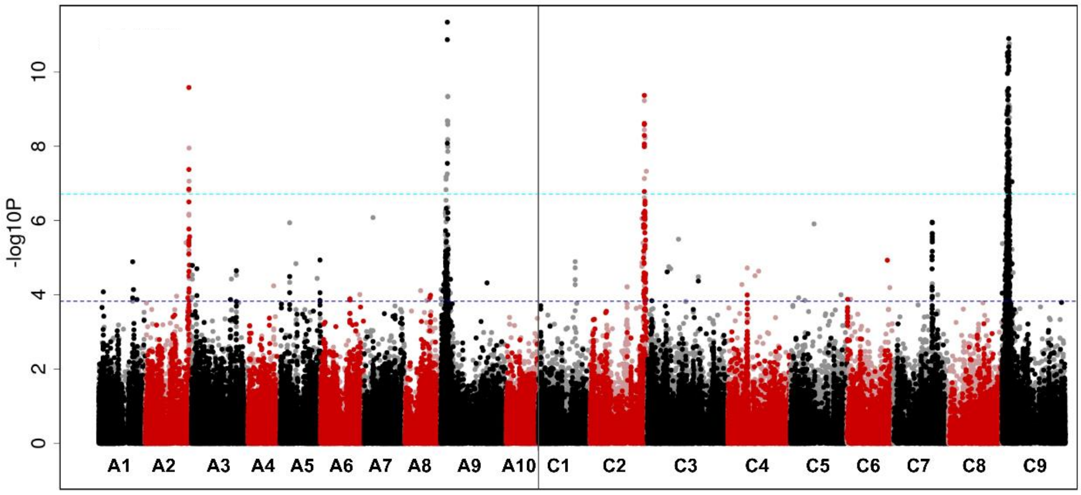


Figure 2-6. Example SNP Manhattan plot. The *x*-axis is the genomic position of the SNPs based on the order of the CDS model and the *y*-axis is the significance value displayed as negative log base 10 of the *P*-value. Chromosomes of *B. napus* are labelled A1 – A10 and C1 – C9, shown in alternating black and red colours to allow boundaries to be clearly distinguished. The lower purple line marks the false discovery rate (FDR) threshold at 5% and the upper cyan line marks the threshold for Bonferroni significance of 0.05. Dark opaque points are simple SNP markers (i.e. polymorphisms between resolved bases) and hemi-SNPs that have been directly linkage-mapped, both of which can be assigned to one genome, whereas pale points are hemi-SNP markers (i.e. polymorphisms involving multiple bases called at the SNP position in one allele of the polymorphism) for which the genome of the polymorphism cannot be assigned.

2.4.1.1 SNP association analysis

The region of associations can be examined in more detail using the GAPIT SNP result table and the ordered pan-transcriptome CDS models. The resulting table was used to provide useful information on the SNP markers including annotated *Arabidopsis thaliana* orthologues, base-pair position, chromosome, *P*-values, second allele frequency (saf) and FDR-adjusted *P*-values corresponding to the results displayed on the Manhattan plot. A marker may be highly associated to the investigated phenotype either because it has direct influence on the trait or because it is in linkage disequilibrium (LD) with a causal polymorphism, although majority of associated SNP markers are more likely to be in LD with a causative gene. ThaleMine, an integrative database of *Arabidopsis thaliana* genomic information (Krishnakumar *et al*., 2017), was used to explore gene functions and relevant pathway of the associated marker. Ordered pan-transcriptome CDS models were used to examine neighbouring genes in the regions with highly associated SNP markers (e.g. markers with $–\log_{10}P$ above Bonferroni threshold). Neighbouring genes in close proximity to top SNP markers with relevant annotated gene functions to the observed trait in *Arabidopsis* orthologues were selected as candidate genes for further investigation.

2.4.2 Association analysis for GEMs

In addition to SNP analysis, AT uses sequence read depths from the same mRNAseq dataset as a measure of gene expression. The method for GEM association analysis in this work was described in Havlickova *et al.* (2018). Transcript abundance was quantified and normalised as reads per kb per million aligned reads (RPKM) using the whole CDS models of pan-transcriptome reference for each sample. Prior to regression, CDS models with a mean expression below 0.4 RPKM across the panel were removed, leaving 53,889 CDS models. For GEM association, fixed-effect linear model was calculated in `R` software, which used RPKM values and the Q matrix inferred by PSIKO as explanatory variables, and trait score as the response variable (Havlickova *et al*., 2018). When genomic inflation factor (λ) was >1, genomic control with *P*-value adjustment (Devlin and Roeder, 1999) was applied to the GEM analysis to correct for false associations.

For GEM Manhattan plots, genomic position of the GEMs based on the order of the CDS model are plotted on the *x*-axis. The significance of the trait association, as $–\log_{10}P$ values, was plotted on the *y*-axis. Individual point on the Manhattan plot that passes both false discovery rate (FDR) threshold at 5% (lower line, purple) and threshold for Bonferroni significance of 0.05 (upper line, cyan) were regarded as GEMs showing strong association with the investigated trait. For instance, Figure 2-7 identifies the GEMs on chromosome C1, A6, A2 and C3.

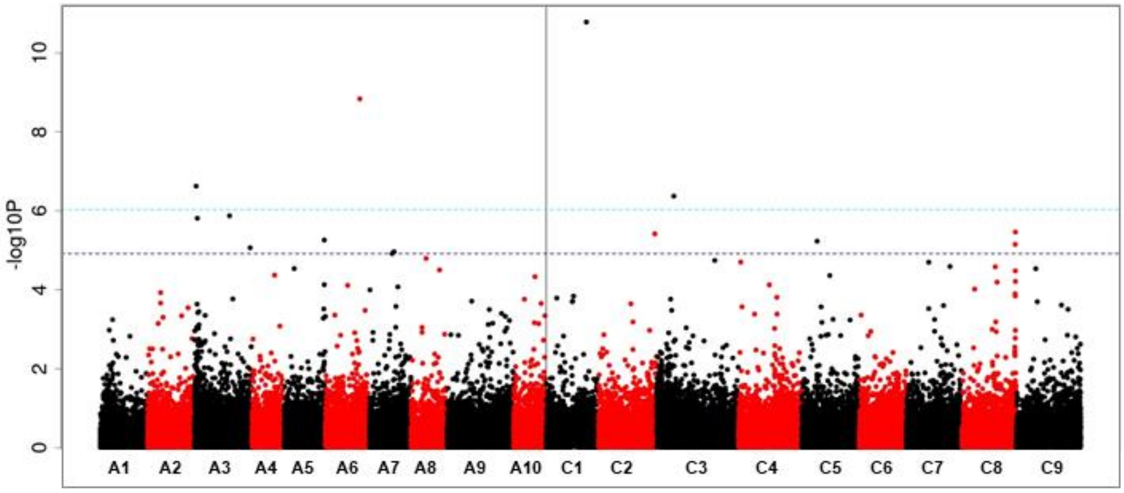


Figure 2-7. Example GEM Manhattan plot. The *x*-axis is the genomic position of the GEMs based on the order of the CDS model and the *y*-axis is the significance value displayed as negative log base 10 of the *P*-value. Chromosomes of *B. napus* are labelled A1 – A10 and C1 – C9 for chromosomes on the A and C genome, respectively. The lower purple line marks the false discovery rate (FDR) threshold at 5% and the upper cyan line marks the threshold for Bonferroni significance of 0.05.

2.4.2.1 GEM association result analysis and candidate selection

In each linear regression analysis of the gene expression matrix, R^2, regression coefficients, constants, F-value and significance *P*-values were produced. CDS models were ranked based on significance *P*-values and top ranked CDS models in this list were considered as top associated GEMs. These were then examined for gene functions that may be connected to the investigated trait in *Arabidopsis* orthologues using ThaleMine (Krishnakumar *et al*., 2017). Other information such as R^2 coefficient of determination and slope of regression line were used to help assess the strength of the relationship between the fitted regression line and the response variable, and examine the direction of the relationship (positive or negative).

CHAPTER 3

Glucosinolate profiles in *B. napus* leaves and roots

This chapter focuses on the analysis of GSL phenotypic traits as a foundation for studying the genetic and biological activities of GSL. The chapter begins with the examination into the variations of GSLs in the leaf and root tissues across a panel of 288 *B. napus* accessions (Section 3.3). This is then further divided into three parts. Section 3.3.1 analyses the frequency distribution of phenotypic traits through histograms and Section 3.3.2 compares the pattern of GSL accumulation in different crop types across the panel. To understand whether and to what extent the observed variation is influenced by genetic components, Section 3.3.3 calculates the contribution of genotype to trait variation from the estimated narrow-sense heritability. Finally, Section 3.4 investigates the relationship of GSLs within and between the vegetative tissues through correlation analysis.

3.1 Introduction

Investigation of the GSL profiles are important as it provides the foundation for studying their biological activities and physiological processes linked to GSLs. To date, the leaf GSL profiles have been reported for 33 *B. napus* accessions (Velasco *et al*., 2008) but there is no information on root GSLs in this important crop species. This chapter will attempt to fill in this knowledge gap by providing a comprehensive analysis of the leaf and root GSL profiles in 288 *B. napus* accessions. The primary aim of the chapter is to investigate the GSL profiles at the population level, which will support the analysis of genetic association studies in subsequent chapters (Chapter 4, 5 and 6). In addition, data generated in this chapter can benefit oilseed rape researchers and agribusinesses in general. For instance, the estimated heritability can be used to assist the breeder, or in the selection of *B. napus* genotype with desirable profiles to modulate plant-pest interactions.

3.2 Methods - Statistical analyses

The general material and methods for GSL quantifications have been presented in Chapter 2. The following methods were used for the statistical analyses specific to this chapter, which were carried out with script packages written in R language (version 3.5.1; 2013).

3.2.1 Population structure analysis

The clustering of single-nucleotide polymorphism (SNP) genotypes and population structure analysis was reproduced based on Havlickova *et al*. (2018), using an R script developed in-house by Dr. Zhesi He, the group's bioinfomatician (shown on Figure 3-2A to 3-2C). The relatedness of the accessions in the panel were based on 355,536 scored genome-wide SNPs as described previously (Havlickova *et al*., 2018). The distance matrix of accessions, visualised as dendrogram, was generated using R package 'PHANGORN'.

3.2.2 Frequency distribution histograms

To visualise the distribution of GSLs, frequency histograms were generated using the 'geom_histogram' function of R package 'ggplot2', which was used to count the number of observed GSL concentration into range of bins. A bin width of 30 was set for all variables. Multiple histograms were generated with the 'facet_wrap' function. The following script was used to produce the histograms:

```
# Plot multiple histograms, change ncol for number of columns
 ggplot(melt(data),aes(x=value)) + geom_histogram(bins = 30) +
        facet_wrap(~variable, ncol = 3, scales = "free") +
        labs(y="No. of Accessions",
              x="Glucosinolate concentration("*mu~"mol/g)") +
        theme(panel.background=element_rect(fill = "white"),
              axis.line = element_line(size = 0.1, colour = "black"),
              text = element_text(size = 11),
              axis.text = element_text(colour="black", size = 11))
```

3.2.3 Comparison between crop types

One-way analysis of variance (ANOVA) and Dunnett’s T3 post hoc test were performed on GSL content between crop types. One-way ANOVA was selected as a statistical test to compare the means of three or more independent groups and determines whether the means between groups are significantly different. Since ANOVA provides a measure of overall difference between groups, a follow-up post hoc test is needed to determine which specific groups differed from one another. Dunnett’s T3 was chosen because it is a suitable test for pairwise comparisons of a dataset with unequal variance and sample size, which is the case for the GSL content dataset. Boxplots were generated with the following script to visualise the difference in the distribution of GSL content between crop types.

```
# Order data from low to high based on mean
L2H = with(<data_file>, reorder(<group e.g.”Crop”>, <variable e.g.GSL
content>, mean, na.rm=T))
# Plot
boxplot(<variable> ~ L2H, ylab = "GSL conc. ("~mu~"mol/g)",
        xlab = "Crop type", cex.axis=0.8, col=c("white","grey"))
```

3.2.4 Correlations of GSLs

Due to the large variabilities and skewed distributions in the GSL dataset, Spearman’s correlation was calculated as the correlation coefficient values because it measures the monotonic relationship between variables and is more robust with extreme values. This Spearman’s correlation table was generated using the `R` package ‘`psych`’ with the following script:

```
pairs.panels(<data_file>, method=”spearman”, stars = TRUE)
```

3.2.5 Quantifying heritability

The compressed mixed linear model (MLM), a population stratification statistical method of GAPIT, was used to quantify the derivative of heritability for GSL traits. MLM uses both fixed and random effect to account for the population structure and unequal relatedness among individuals (Lipka *et al.*, 2012) and calculates estimates of heritability as part of the GAPIT output. Degree of heritability (h^2) is defined as the proportion of total observable variance that is due to genetic variance:

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}$$

where σ_a^2 is the additive genetic variance and σ_e^2 is the residual variance. Derivative of the heritability quantification from the MLM is summarised in Figure 3-1. In this work, narrow-sense heritability in the GSL contents between crop types was derived from the calculations for compressed MLM (Buckler and Zhang, 2018).

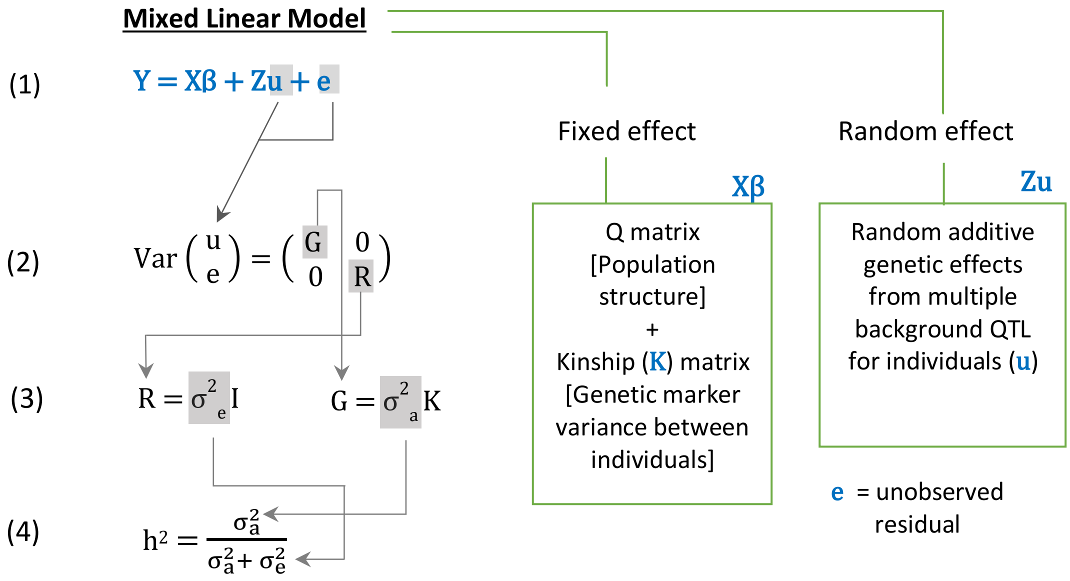

Mixed Linear Model
(1)
Y = Xβ + Zu + e
(2)
Var (u e) = (G 0 0 R)
(3)
R = σ²eI
G = σ²aK
(4)
h² = σ²a / (σ²a + σ²e)
Fixed effect
Random effect
Xβ
Zu
Q matrix [Population structure] + Kinship (K) matrix [Genetic marker variance between individuals]
Random additive genetic effects from multiple background QTL for individuals (u)
e = unobserved residual

Figure 3-1. Derivative of narrow-sense heritability equation from mixed linear model (MLM) used by GAPIT (summarised from Buckler and Zhang, 2018). **(1)** MLM incorporates both random and fixed effects, where **Y** is the vector of observed phenotypes; **β** is a vector containing fixed effects (Q+K); **u** is an unknown vector of random additive genetic effects; **X** and **Z** are known design matrices; and **e** is a vector of unobserved residual. **(2)** **u** and **e** vectors are assumed to be normally distributed with a null mean and a variance as shown in the equation. **(3)** The genetic effect (**G**) with σ^2_a as the additive genetic variance and **K** as kinship matrix. Constant variance is assumed for the residual effect (**R**) where σ^2_e is the residual variance. **(4)** Finally, the proportion the total variance explained by the genetic variance is defined as heritability (h^2).

Results

3.3 Glucosinolate variations in leaves and roots

A diversity panel of 288 *B. napus* accessions was analysed for GSL compositions in the leaves and roots of 3-week old plants using a high throughput quantification method developed for this work as described in Chapter 2. A total of 1,152 plants were grown, which generated 2,304 samples (leaf and root samples). From these, fourteen different known GSLs have been identified. Out of these, nine are classed as aliphatic (of different chain lengths C_3, C_4 and C_5), four indole and one aromatic GSL (Table 3-1). Overall, 32,256 data points have been processed to quantify the GSL concentrations. Absolute value of GSL concentrations are provided in Appendix 1 for leaves and Appendix 2 for roots. This dataset has been published in Kittipol *et al.* (2019b).

Extensive phenotypic variations have been observed in both leaves and roots across the panel. Higher variability is found in leaves but lower in roots. The total GSL content ranges from 0.26 to 21.6 µmol/g in leaves, with aliphatic GSLs as the predominant class (64.0% of all leaf GSLs), particularly PRO (28.61% of all leaf GSLs) and GBN (19.21% of all leaf GSLs). Approximately one third (32.9%) of all leaf GSLs belong to the indole class and a small amount are GST (3.1%), an aromatic GSL (Figure 3-2D). In roots, the total GSL content falls within a narrower range from 2.4 to 17.1 µmol/g and the profile of GSLs is different from that of the leaf tissue. Of all root GSLs, 47.7% are indole GSLs and 45.0% are aromatic GSL. These two types jointly form the major classes of GSLs that have contributed to the root GSL profile. However, GST is the most abundant GSL found in roots (45.0% of all root GSLs) and is the only type of aromatic GSLs present in this 288 accessions *B. napus* population. The second most abundant GSL in the roots is an indole GSL, namely neo-GBS (37.0% of all root GSLs). Unlike leaf, aliphatic GSL was found as a minor class of GSL in the roots (7.3%) (Figure 3-2E). The population structure (Figure 3-2A to 3-2C) in relation to the GSL variations will be discussed in the following section (Section 3.3.2).

Table 3-1. Glucosinolates identified in this study

Type	Trivial name	Acronym	Systematic R Side chain
Aliphatic C_3	Glucoputranjivin	GJV	1-Methylethyl
Aliphatic C_4	Gluconapin	GNA	3-Butenyl
	Progoitrin	PRO	(2R)-2-Hydroxy-3-butenyl
	Glucoerucin	GER	4-Methylthiobutyl
	Glucoraphanin	GRA	4-Methylsulfinylbutyl
	Glucoraphenin	GRE	4-Methylsulfinyl-3-butenyl
Aliphatic C_5	Glucoalyssin	GAL	5-Methylsulfinylpentryl
	Glucobrassicanapin	GBN	Pent-4-enyl
	Gluconapoleiferin	GNL	2-Hydroxy-pent-4-enyl
Indole	Glucobrassicin	GBS	3-Indolylmethyl
	4-Hydroxyglucobrassicin	4-OHGBS	4-Hydroxy-3-indolylmethyl
	4-Methoxyglucobrassicin	4-OMeGBS	4-Methoxy-3-indolylmethyl
	Neoglucobrassicin	neo-GBS	N-Methoxy-3-indolylmethyl
Aromatic	Gluconasturtiin	GST	2-Phenethyl

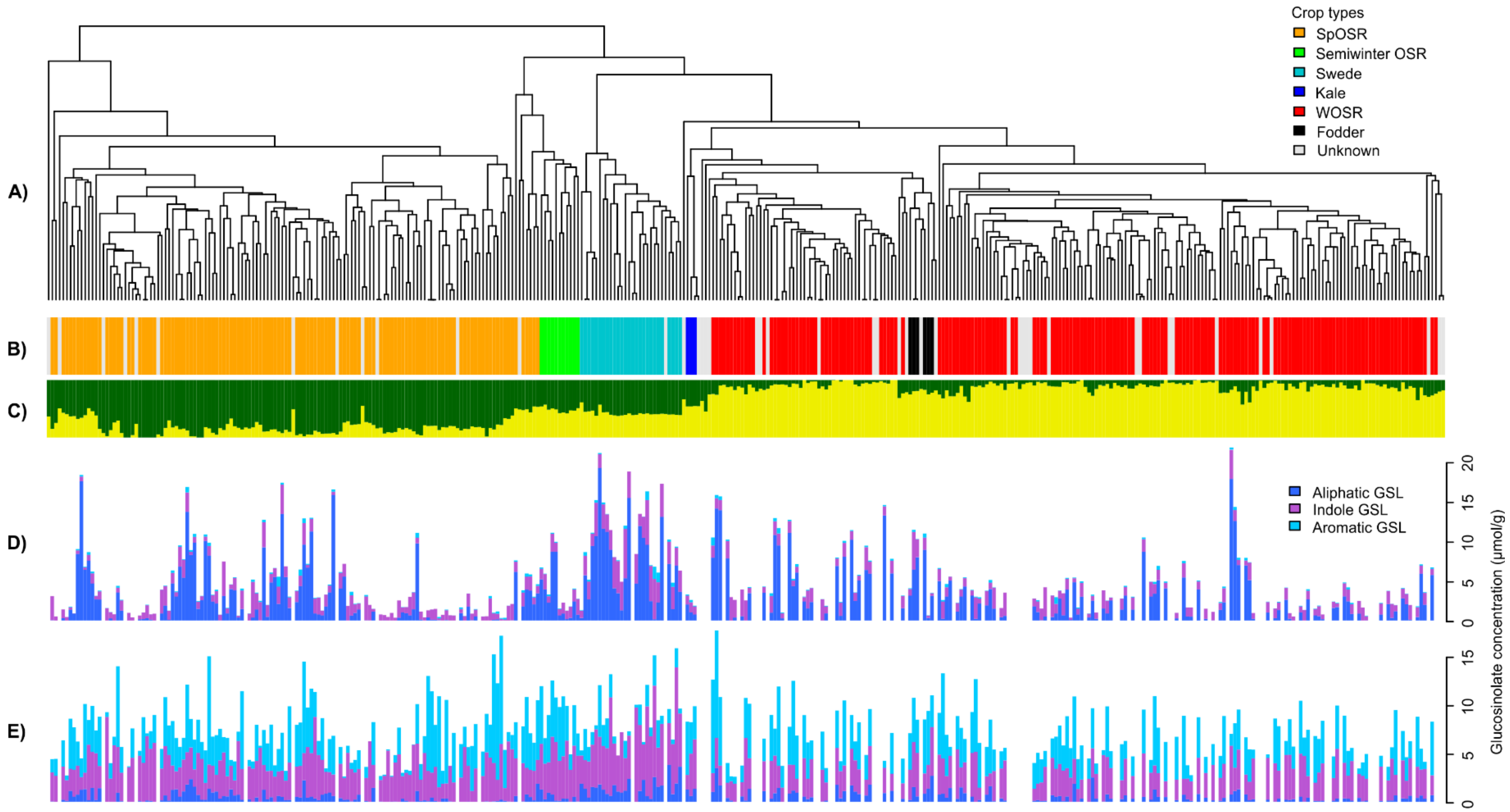


Figure 3-2. Population structure and Glucosinolate variation across the Renewable Industrial Products from Rapeseed (RIPR) Panel. (A) Relatedness of accessions in the panel based on 355 536 scored single-nucleotide polymorphisms (SNPs). **(B)** Main crop types, colour coded: orange for spring oilseed rape (SpOSR); green for semi-winter oilseed rape; light blue for swede; dark blue for kale; red for winter oilseed rape (WOSR); black for winter fodder and grey for crop type not assigned. **(C)** Population structure for highest likelihood *k* = 2. Variation for glucosinolates content in **(D)** leaf and **(E)** root of 288 *B. napus* accessions. Individual glucosinolates were grouped according to their structural class as aliphatic (dark blue), indole (magenta) and aromatic (light blue). Panels A, B and C reproduced from Havlickova *et al*. (2018).

3.3.1 Distribution of phenotypic traits

To examine the distribution of phenotypic traits, frequency histograms were generated for individual leaf GSLs (Figure 3-3 and Figure 3-4), individual root GSLs (Figure 3-5 and Figure 3-6), as well as the distribution of GSL structural classes (Figure 3-7).

All individual GSLs in the leaf, except GBS, display a reverse J-shaped right skewed distribution, with the highest frequency of the population at the extreme low value (Figure 3-3 and Figure 3-4). Traits that are related to aliphatic GSLs including total leaf GSLs, leaf aliphatic GSLs and root aliphatic GSLs have shown the same extreme skewed right distribution (Figure 3-5 and Figure 3-7). Similarly, leaf aromatic GSL traits also exhibit the same extreme skewed right distribution (Figure 3-7). This reverse J-shaped distribution towards extreme low value of aliphatic GSLs is a characteristic of truncation selection over a course of time, a consequence of the intensive breeding of oilseed *B. napus* for low seed GSL traits.

The distribution of the leaf indole GBS, individual root GSLs (Figure 3-6) and root indole GSL (Figure 3-7), though skewed right, does not show the reverse J-shaped distribution similar to the aliphatic GSLs. The skewed right distributions of indole GSLs indicate an effect from the selective breeding of *B. napus*, but it is to a lesser extent compared to aliphatic GSLs.

The frequency distribution of root GST, which is reflected in root aromatic GSL (Figure 3-7), has shown to be skewed to the right and appears to be discontinuous. At a certain threshold (approximately 5.5 µmol/g), the phenotype seems to have a discrete pattern that separates the population into the low and high GSL groups. The discrete distribution of root aromatic GSL implies a simple genetic basis underlying the variation of root aromatic GSL.

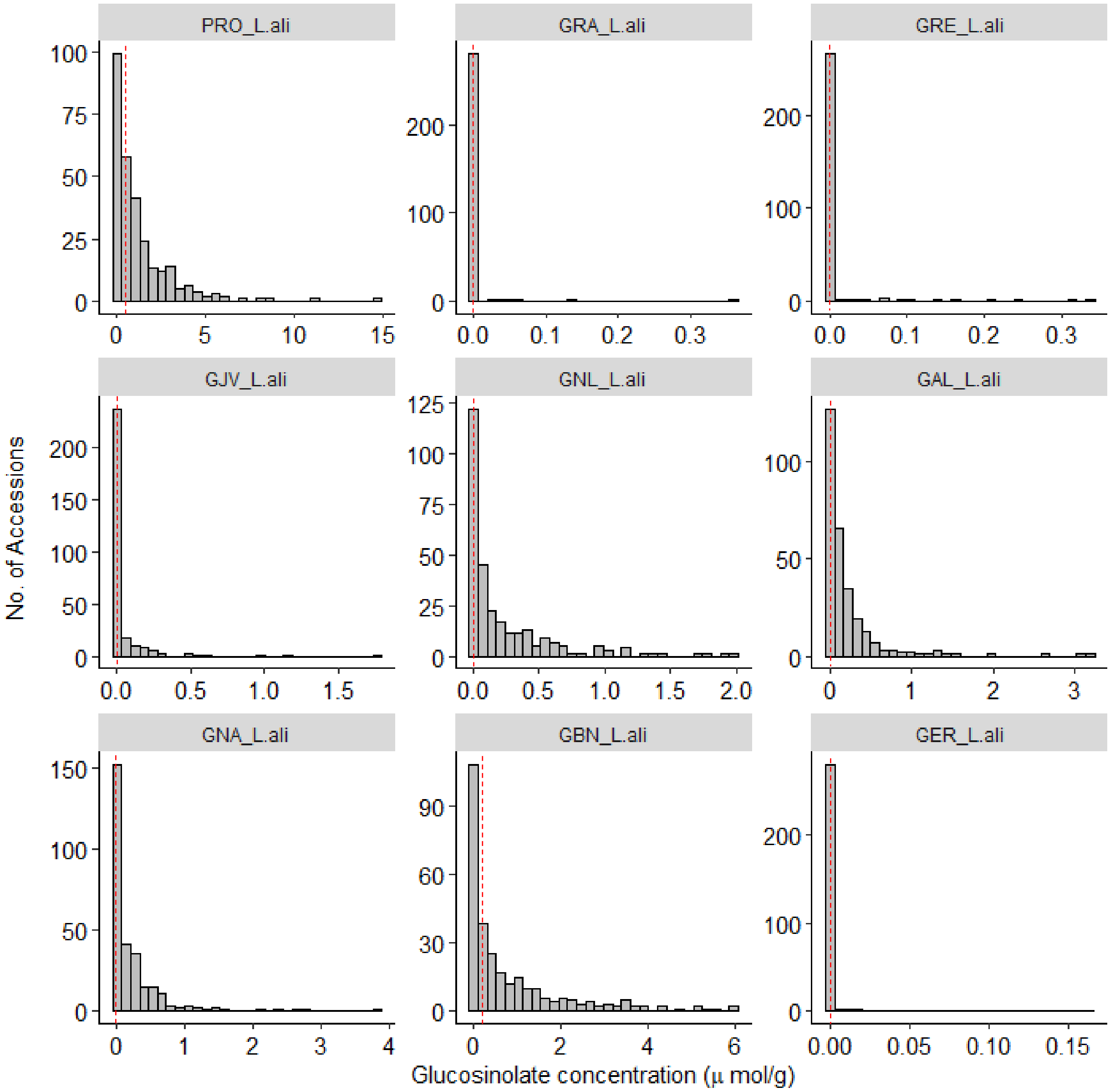


Figure 3-3. Frequency distributions of individual aliphatic glucosinolates in the leaf. The dashed red lines are the medians, which is a better representative of the central tendency of the data with skewed distribution than mean. On top of each plot are the abbreviated glucosinolate name and the type of glucosinolate separated by an underscore. Abbreviation: L.ali, leaf aliphatic; PRO, progoitrin; GRA, glucoraphanin; GRE, glucoraphanin; GJV, glucoputranjivin; GNL, gluconapoleiferin; GAL, glucoalyssin; GNA, gluconapin; GBN, glucobrassicanapin; GER, glucoerucin.

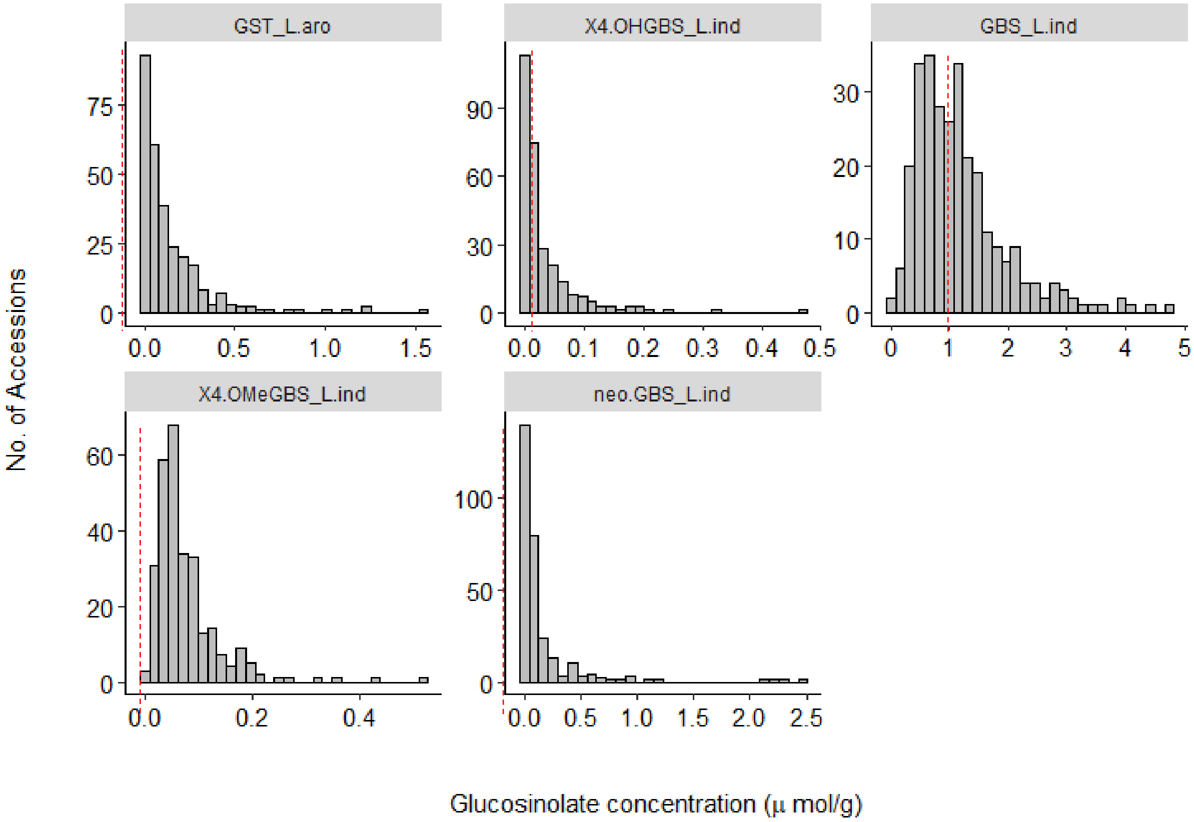


Figure 3-4. Frequency distributions of individual aromatic and indole glucosinolates in the leaf. The dashed red lines are the medians, which is a better representative of the central tendency of the data with skewed distribution than mean. On top of each plot are the abbreviated glucosinolate name and the type of glucosinolate separated by an underscore. Abbreviation: L.aro, leaf aromatic; L.ind, leaf indole; GST, gluconasturtiin; X4.OHGBS, 4-hydroxyglucobrassicin; GBS, glucobrassicin; X4.OMeGBS, 4-methoxyglucobrassicin; neo.GBS, neoglucobrassicin.

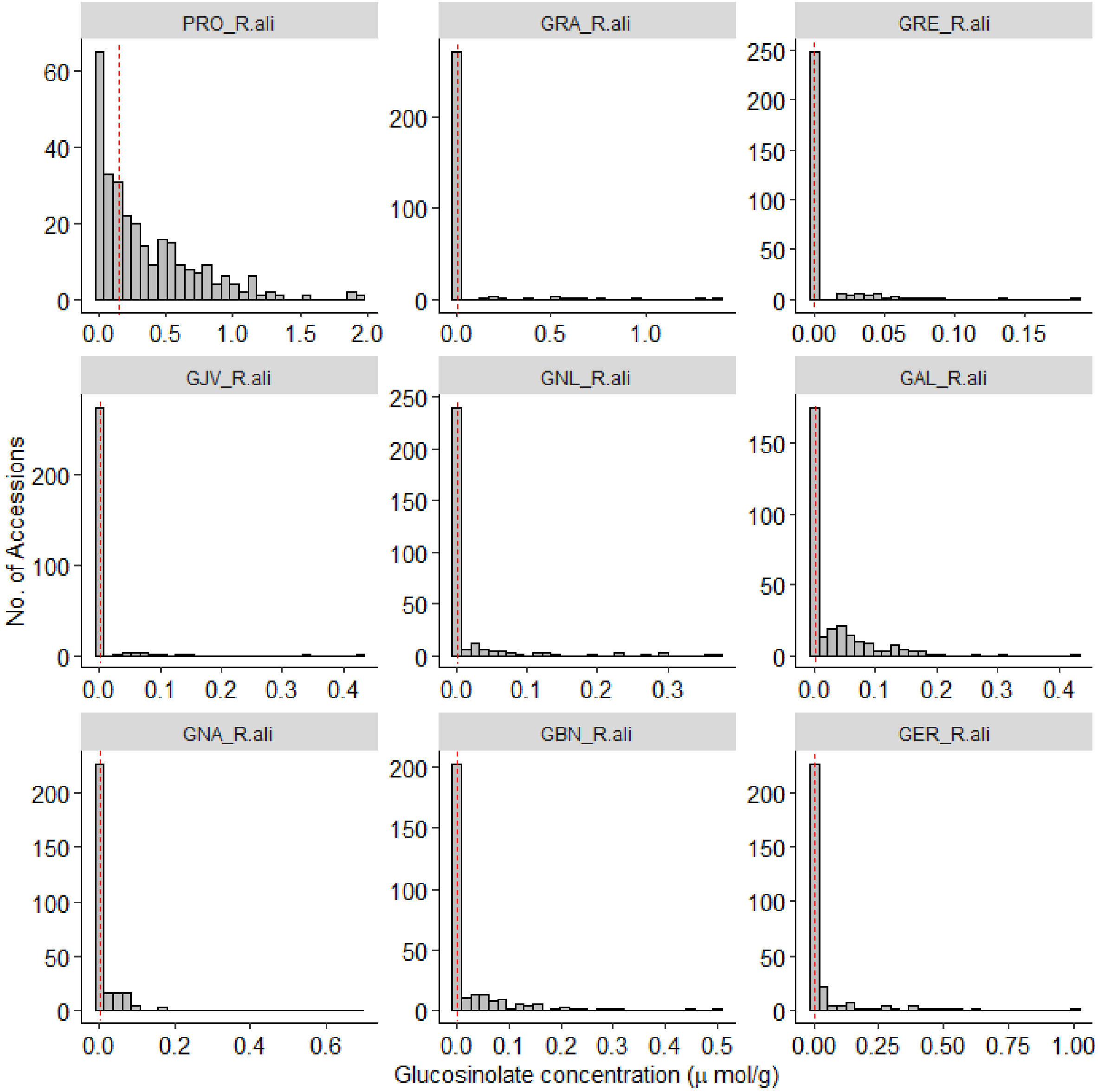


Figure 3-5. Frequency distributions of individual aliphatic glucosinolates in the root. The dashed red lines are the medians. On top of each plot are the abbreviated glucosinolate name and the type of glucosinolate separated by an underscore. Abbreviation: R.ali, root aliphatic; PRO, progoitrin; GRA, glucoraphanin; GRE, glucoraphanin; GJV, glucoputranjivin; GNL, gluconapoleiferin; GAL, glucoalyssin; GNA, gluconapin; GBN, glucobrassicanapin; GER, glucoerucin.

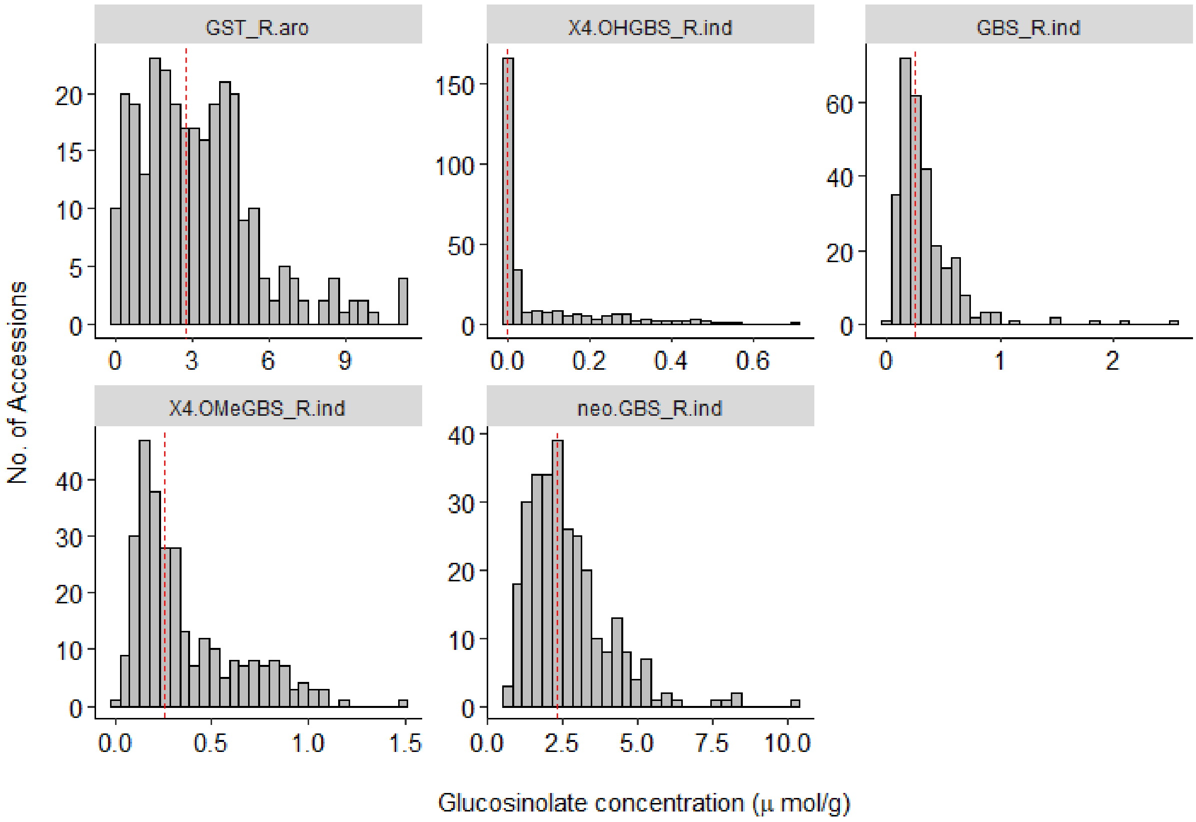


Figure 3-6. Frequency distributions of individual aromatic and indole glucosinolates in the root. The dashed red lines are the medians. On top of each plot are the abbreviated glucosinolate name and the type of glucosinolate separated by an underscore. Abbreviation: R.aro, root aromatic; R.Ind, root indole; GST, gluconasturtiin; X4.OHGBS, 4-hydroxyglucobrassicin; GBS, glucobrassicin; X4.OMeGBS, 4-methoxyglucobrassicin; neo.GBS, neoglucobrassicin.

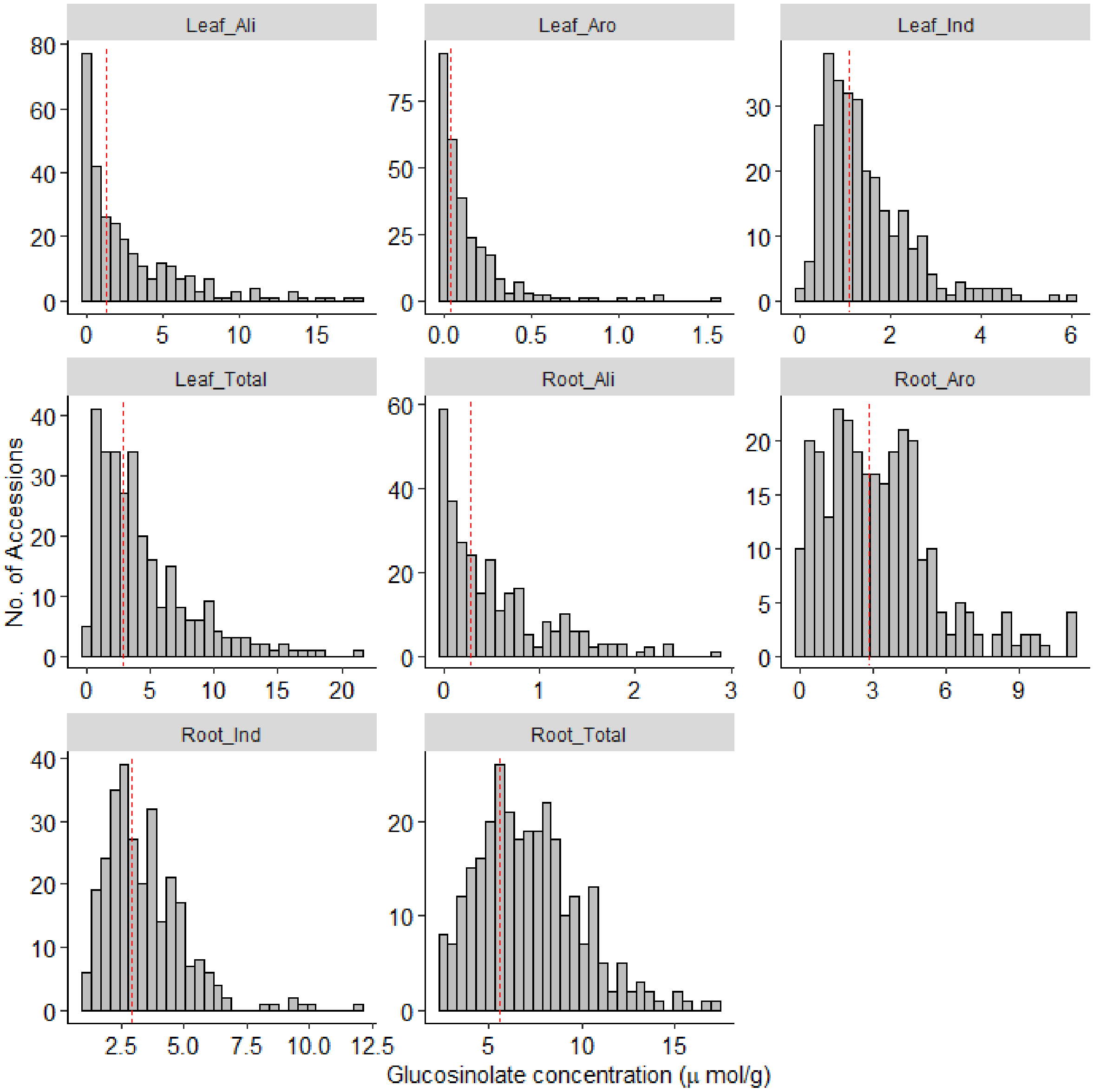


Figure 3-7. Frequency distributions of total glucosinolates and distribution of GSL structural classes in each tissue type. The dashed red lines are the medians, which is a better representative of the central tendency of the data with skewed distribution than mean. Abbreviation: Total, total amount of GSLs; Ali, aliphatic GSLs; Aro, aromatic GSLs; Ind, indole GSLs.

3.3.2 Analysis of glucosinolate by crop types

To get an idea of the population structure, relatedness of the accessions was analysed based on the scored genome-wide SNPs across the panel (Figure 3-2A). The seven assigned crop types show the expected clustering of accessions (Figure 3-2B) with the highest likelihood of two differentiated subpopulations (k = 2), which separated into the spring or winter oilseed rape crop types or a mixture of the two (Figure 3-2C). The genetic architecture of the population in this study is consistent with the previously reported population structure of the full RIPR panel (Havlickova *et al*., 2018).

In addition to differences in GSL composition between leaf and root, the contents also vary considerably between crop types. Analysis of GSL characteristics by crop type revealed three distinct statistical groups in the leaf, with the swede crop type at the higher end of the spectrum while winter and spring oilseed rape crop types are on the lower end (Figure 3-8). The significantly higher total concentration of the leaf GSLs in swede compared to other crop type is due to the increased concentration of both aliphatic and indole classes (Figure 3-9A and Figure 3-9B).

There is less variability observed for the total root GSLs than that for the total leaf GSLs, as shown by the Dunnett's T3 post hoc test (Figure 3-8). Similar to the case of swede leaves, swede roots have also shown increased levels of aliphatic and indole GSLs. It is worth noting that GST, the major root GSL, is significantly lower in swede compared to other crop types (Figure 3-9C), which balances out the concentration of total root GSLs of swede to reside in the mid-range.

Modern winter and spring oilseed rape crop types are among the groups showing lower GSL content in both leaf and root tissues, with a reduction in all three GSL classes being observed (Figure 3-9). This lower GSL content can be partly explained by the extensive breeding efforts for low GSL contents, since many of the commercial varieties commonly used today came from the winter and spring oilseed rape that had been subjected to extensive breeding program for germplasm with low GSL contents during the 1970s (Rosa *et al*., 1997). Based on the population structure, although winter and spring oilseed rape are the most distant subgroups (Figure 3-2A and Figure 3-2C), levels of GSL concentration between these two crop types do not differ significantly either in leaf or root tissues (Figure 3-8).

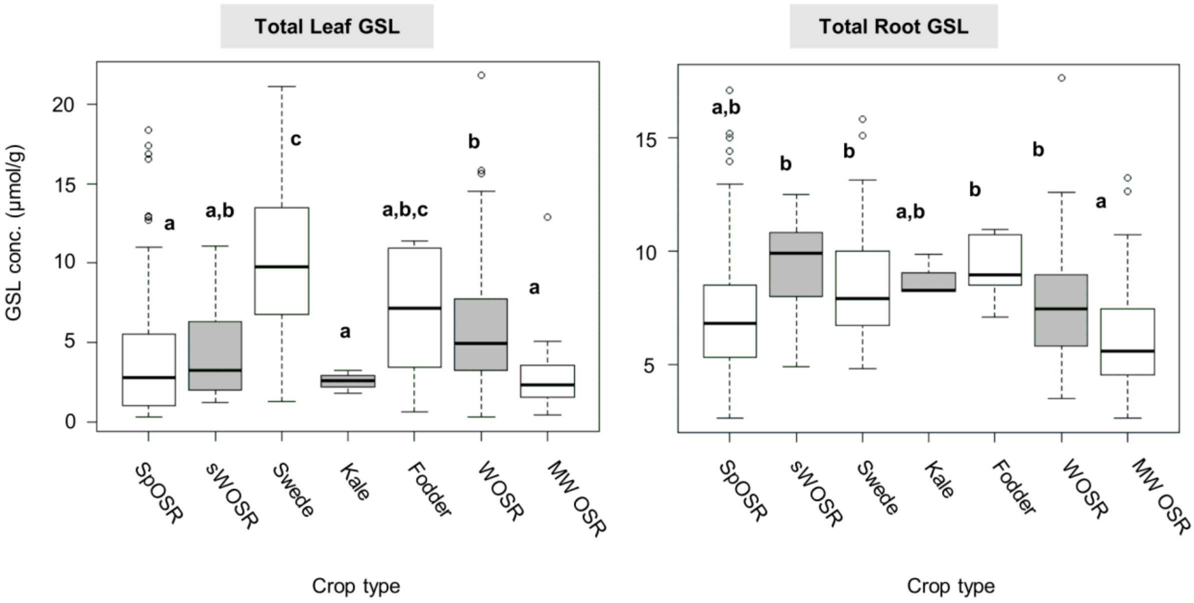


Figure 3-8. Total glucosinolate content differs between crop types of 288 *B. napus* accessions. Box-whisker plots of the total glucosinolates in leaves and roots. The box represents the upper and lower quartiles and the dark line is the median value. The whisker extend to the maximum and minimum values. Open circles indicate outliers. Same letter indicate no significant difference between the groups (ANOVA with Dunnett T3 as post-hoc test for unequal variance and sample size between crop types, p >0.05). Abbreviation: MW OSR, Modern winter oilseed rape; SpOSR, Spring oilseed rape; WOSR, Winter oilseed rape; sWOSR, semiwinter oilseed rape, Fodder, Winter fodder.

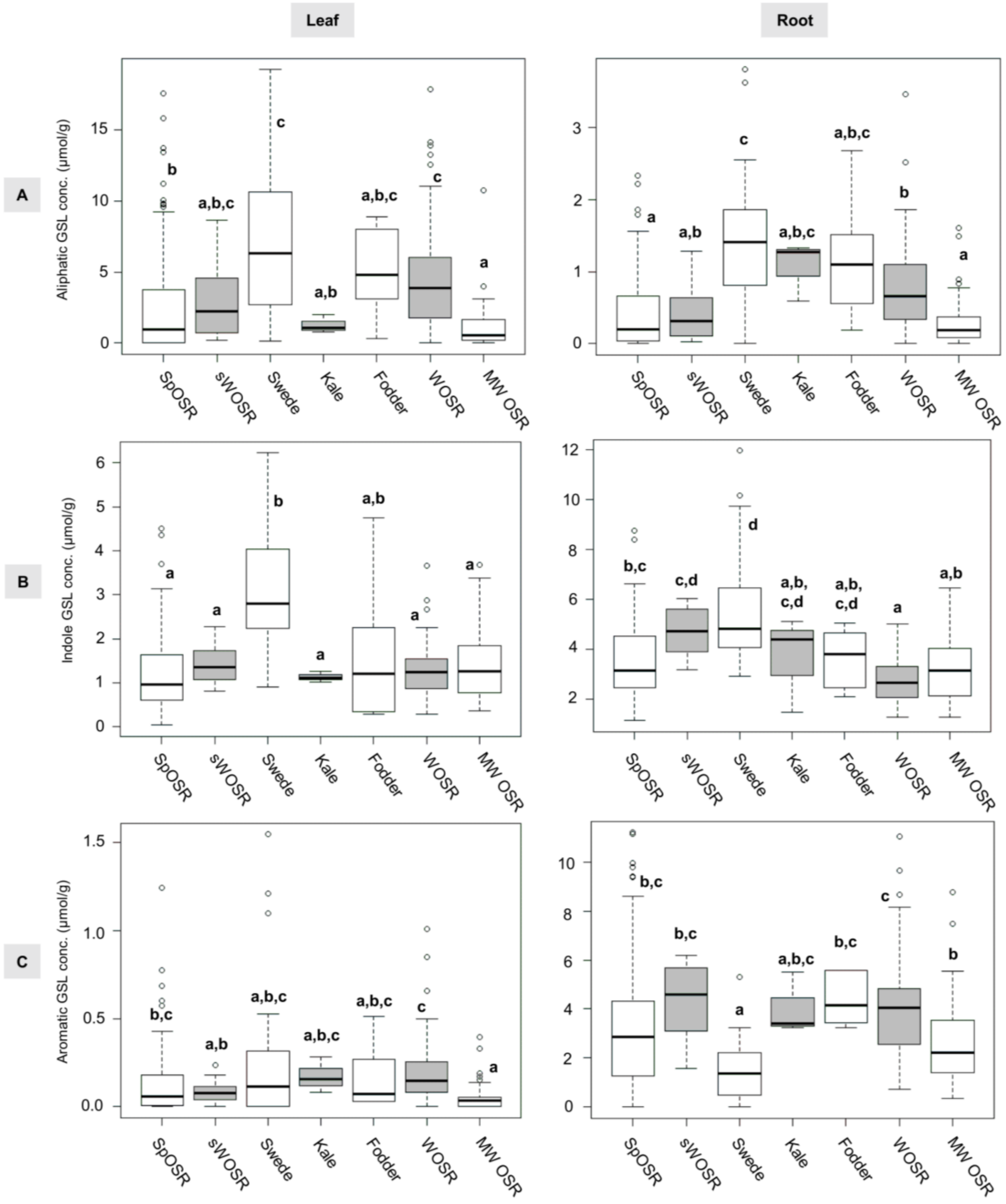


Figure 3-9. Comparison of glucosinolate contents between different crop types of 288 *B. napus* accessions. Box-whisker plots of (A) aliphatic, (B) indole and (C) aromatic GSLs in leaves and roots. The box represents the upper and lower quartiles and the dark line is the median value. The whisker extend to the maximum and minimum values. Open circles indicate outliers. Same letter indicate no significant difference between the groups (ANOVA with Dunnett T3 as post-hoc test for unequal variance and sample size between crop types, p >0.05). Abbreviation: MW OSR, Modern winter oilseed rape; SpOSR, Spring oilseed rape; WOSR, Winter oilseed rape; sWOSR, semiwinter oilseed rape, Fodder, Winter fodder.

3.3.3 Heritability of the traits

To estimate the contribution of genotype to the trait variation, heritability (h^2) has been calculated to show the proportion of trait variation in a population attributable to genetic influences. As shown in Table 3-2, majority of the GSL traits are highly heritable. All leaf GSL traits have greater than 50% in the estimated heritability. Four of the aliphatic GSL traits show that 100% of their phenotypic variations are attributed to genetic variances. Similarly, the root GSL variations can be largely explained by genetic variations. The only exception is the root glucoerucin, where 82.4% of its trait variation is influenced by residual variance. Seven of the root GSL trait variations have 100% estimated heritability, indicating that all observed phenotypic variations in these traits across the population are due to the differences in their genotypes.

Table 3-2 . Estimated heritability from GAPIT output for leaf and root glucosinolate traits.

Leaf GSL	Estimated heritability (%)
Glucoraphenin	100.0
Glucobrassicanapin	100.0
Glucoerucin	100.0
Neoglucobrassicin	100.0
Gluconapoleiferin	96.7
Leaf Aliphatic	92.0
4-Methoxyglucobrassicin	87.7
Total Leaf GSL	85.9
Leaf Aromatic	74.5
Glucobrassicin	74.5
Glucoputranjivin	66.8
Gluconasturtin	65.9
4-Hydroxybrassicin	64.6
Glucoraphanin	61.2
Gluconapin	61.2
Progoitrin	60.0
Glucoalyssin	58.0
Leaf Indole	50.6

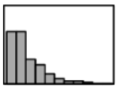
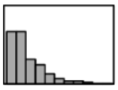
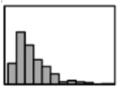
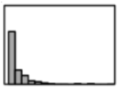
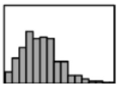
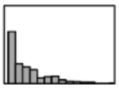
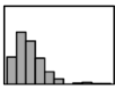
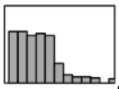
Root GSL	Estimated heritability (%)
Root Aliphatic	100.0
Root Indole	100.0
Root Aromatic	100.0
Gluconasturtin	100.0
4-Hydroxybrassicin	100.0
4-Methoxyglucobrassicin	100.0
Glucoraphanin	100.0
Gluconapoleiferin	89.9
Glucoputranjivin	89.2
Gluconapin	87.0
Progoitrin	79.3
Glucoraphenin	70.6
Glucobrassicin	65.5
Neoglucobrassicin	63.9
Total Root GSL	60.4
Glucobrassicanapin	60.1
Glucoalyssin	59.2
Glucoerucin	17.6

3.4 Relationships among glucosinolates

In order to understand the relationship of GSLs within and between different tissues, Spearman's correlation analysis was performed on the GSL traits (Table 3-3). Within leaves, the total amount of GSLs accumulated is dominated largely by the level of aliphatic GSLs (r = 0.91, *p* ≤0.001). On the other hand, both indole and aromatic GSLs are the major GSL classes found in roots. Between the two classes, aromatic GSLs (i.e. GST) provide a much stronger indication of the total amount of root GSLs (r = 0.75, *p* ≤0.001) and indole GSLs contribute less (r = 0.44, *p* ≤0.001) comparatively.

Aliphatic GSLs have exhibited the strongest correlation between the two tissues (r = 0.68, *p* ≤0.001), indicating that aliphatic GSLs in the leaf and root tissues could be regulated by long-distance transport or a master regulator of aliphatic GSL pathway that controls the biosynthesis in both tissues. Significant positive correlations were observed between aliphatic and aromatic GSLs within the same tissue (Leaf: r = 0.62, *p* ≤0.001; Root: r = 0.30, *p* ≤0.001), as well as between leaf and root (r = 0.50, *p* ≤0.001; r = 0.29, *p* ≤0.001), which suggest the possibility of co-regulation that is shared between these two classes of GSLs. On the contrary, correlations between indole and aromatic GSL within roots (r = –0.18, *p* ≤0.01) and between root and leaf tissues (r = –0.15, *p* ≤0.05; r = –0.22, *p* ≤0.001) are either weak or negative, implying a possible antagonistic relationship between these two GSL classes. Given that different GSL profiles have been found between aliphatic-dominated leaf and indole/aromatic-dominated root tissues (Figure 3-2), the GSL metabolic pathways between above- and below- ground tissues appear to be regulated differentially. Some cross-talk may exist between these pathways, which is supported by the weak but significant correlation between total GSLs in the leaf and root (r = 0.28, *p* ≤0.001).

Table 3-3. Spearman's correlation coefficient analysis of glucosinolate traits with significance value.

	TL	L-ali	L-ind	L-aro	TR	R-ali	R-ind	R-aro
Total Leaf (TL)								
Leaf Aliphatic (L-ali)	0.91 ***							
Leaf Indole (L-ind)	0.45 ***	0.14 *						
Leaf Aromatic (L-aro)	0.62 ***	0.62 ***	0.12 *					
Total Root (TR)	0.28 ***	0.30 ***	0.00	0.37 ***				
Root Aliphatic (R-ali)	0.64 ***	0.68 ***	0.10	0.50 ***	0.43 ***			
Root Indole (R-ind)	0.01	–0.10	0.24 ***	–0.15 *	0.41 ***	–0.04		
Root Aromatic (R-aro)	0.18 **	0.29 ***	–0.22 ***	0.46 ***	0.75 ***	0.30 ***	–0.18 **	

Correlation of mean trait values from 288 accessions of the diversity panel. Significant correlations are indicated; ****p* ≤0.001, ***p* ≤0.01, **p* ≤0.05. The diagonal shows the distribution of data as histograms.

3.5 Chapter discussion

This work is the first report of a large scale GSL profiling from the leaf and root tissues of *Brassica napus*. Consistent with the general leaf profile previously reported (Porter *et al.*, 1991; Velasco *et al.*, 2008), aliphatic GSLs are predominant in the leaf and largely determined the total level of GSLs in the leaves of *B. napus*. As for the root GSL profile, aromatic GSL (GST) is the most abundant in *B. napus*, which is similar to other *Brassica* crops (Bhandari *et al.*, 2015). Unlike other *Brassica* crops, *B. napus* roots are predominated by both aromatic GSL as well as indole GSLs (Figure 3-2), though levels of total GSLs in *B. napus* roots are largely determined by the levels of aromatic GSLs (Table 3-3). The differences in GSL profile and the weak correlation of total GSLs between leaf and root suggests different underlying mechanisms that regulate the GSL variations in these vegetative tissues. Moreover, correlation analyses reveal some insightful significant relationships between different types of GSLs: notably the strong connection of aliphatic GSLs between the leaf and root tissues, the possibility of co-regulation between aliphatic and aromatic GSLs, and an antagonistic relationship between aromatic and indole GSLs.

Across the panel, variations in the levels of GSLs have been observed, leading to the question whether these variations are influenced by the differences in the genes, i.e. whether the traits are heritable. The analysis of estimated narrow-sense heritability (h^2) shown in Table 3-2 reveals that the variations in most of GSL traits are largely attributed to the variance in the genetic components contributing to the genotype. The effect of selection depends on the amount of additive genetic variance (σ_a^2) not on the genetic variance in general, therefore calculation of h^2 is relevant and useful for a prediction of response to breeding selection. This is because additive genetic variance (σ_a^2) of narrow-sense heritability measures the degree in which an individual's genetic makeup contributes to the phenotypic value of the next generation using the sum of the average effects of all alleles the individual carries (Griffiths *et al.*, 1999). Since heritability is high for most GSL traits mean of the population would be expected to respond quickly to the imposed breeding selection because the genetic variation that associates with the desired trait is likely to be inherited in the offspring (Griffiths *et al.*, 1999). Breeding-directed selection for low GSL traits over the years must have caused trait variance in the population to decrease and the mean to be shifted to the extreme end, resulting in the right skewed distributions and truncation of GSL traits as illustrated on Figure 3-3 to Figure 3-7. The discrete distributions imply that the genetic control of GSL variations is likely to be involved with simple genetic loci rather than involving multiple transcriptional regulators across the genome.

In order to identify the underlying genetic basis controlling GSL variations in each of the tissues, GSL structural classes will be used to perform AT analyses for the genetic control of leaf GSL variations in Chapter 4 and for the genetic control of root GSL variations in Chapter 5. As genetic association studies require variations in the phenotypic traits and large sample size, this dataset is suitable to be used in genome-wide association studies.

CHAPTER 4

Genetic control of leaf glucosinolate variation

As the total amount of leaf GSLs is largely determined by variations in leaf aliphatic GSLs (Section 3.4; r = 0.91, $p \leq 0.001$), identifying the genetic controls of aliphatic GSLs will uncover the mechanism controlling the pattern of GSL accumulation in the leaf tissue of *B. napus*. This chapter focuses on the identification of the genetic controls of the levels of aliphatic GSLs (Section 4.2), which is the main GSL class in *B. napus* leaves. This will be further divided into three parts. Section 4.2.1 describes the discovery of candidate genes, *Bna.HAG1.A9 and Bna.HAG1.C2*, through SNP-based and GEM-based associations. *B. napus* is a tetraploid species and its genome contains multiple copies of the same genes. Therefore, section 4.2.2 examines the pattern of *Bna.HAG1* homoeologous gene expression with variations in levels of leaf aliphatic GSL in order to investigate which copies of *Bna.HAG1* are involved. Finally, section 4.2.3 uncovers the impact of structural rearrangement in the genomic regions containing functional copies of *Bna.HAG1* which influences the variations in aliphatic GSL concentrations. Although weak genetic associations have been detected for indole and aromatic GSLs in leaves, the results of analyses are described in section 4.3.

4.1 Methods

4.1.1 Associative Transcriptomics

In order to elucidate genetic loci controlling GSL variations in leaf, AT was performed on a panel of 288 *B. napus* accessions (Section 4.2.1). The current AT platform consists of 355,536 SNP markers and gene expression matrix with a transcriptome reference of 116,098 ordered coding DNA sequence gene models (Havlickova *et al.*, 2018). Detail method of the AT is described in Chapter 2.

4.1.2 Homoeologous gene expressions

As part of a different research project in the lab, mRNAseq data were generated from the leaf tissues of 27 *B. napus* accessions. Each of the accessions had four biological replicates, which improved the robustness and reliability of transcript abundance measurement (He *et al*., 2016). Transcript abundance was quantified for all CDS models after mapping the reads to the same transcriptome reference and subsequently normalised as reads per kilo base per million aligned reads (RPKM) (He *et al*., 2016). These 27 accessions are a subset of the 288 accessions used in the AT analysis. Thus, the availability of these quantified transcript abundance data has provided an excellent opportunity for additional analysis in this PhD project. To compare homoeologous gene expression between accessions, nine accessions were selected from this subset based on the above transcript abundance data as well as the differences in the levels of aliphatic GSLs in leaves, as described in section 4.2.2.

Identification of *Bna.HAG1* homoeologous gene copies was done via two methods: 1) Local BLASTn search analyses with an E-value threshold of 0.00 against the custom 'Pan AC 08 12 2017 genes' database on SequenceServer (https://sequenceserver.york.ac.uk/) (Priyam *et al.*, 2015); and 2) Orthologous ID searches using the orthologue *Arabidopsis* ID to identify the corresponding orthologues within the ordered pan-transcriptome CDS model (V11).

4.1.3 Transcriptome Displayed Tile Plots

Transcriptome Displayed Tile Plots (TDTP), a visualisation method based on mRNAseq data, was used to visualise the relative transcript abundance of homoeologous gene pairs on a genome scale and examine regions of the genome involved in genome structural rearrangement. The method uses in-house R scripts (He *et al.*, 2016). The quantified gene expression of homoeologous gene pairs was derived from the same 27 *B. napus* accessions as Section 4.1.2. Previous study showed that through the analysis of differential gene expression of homoeologous gene pairs, genome dosage changes has been detected in the newly formed allotetraploid *B. napus* genome (He *et al.*, 2016). Through TDTP, certain regions of the genome has shown to be under-expressed with the homoeologous region in other genome over-expressed, which is indicative of structural variation in the genome specifically homoeologous exchange. The works in this thesis utilised the same visualisation method developed by He *et al.* (2016) with slight modification to analyse the impact of structural variations on GSL variations. The tile plot works by assigning quantitative transcript abundance for each member of homoeologous pair a normalised RPKM value from 1 to 0 for the population. Instead of using CMYK colour like He *et al.* (2016), the value in this thesis corresponded to shades of grey where darker shade shows higher transcript abundance and lighter shade of grey shows lower transcript abundance. Inverted colour of the same region on the co-linear chromosome of different genome show homoeologous exchange had occurred. Regions in the genome that correspond to the GEM association peaks are marked with red horizontal lines on the tile plot.

Results

4.2 Genetic control of leaf aliphatic glucosinolates

4.2.1 Uncovering the genetic control through Associative Transcriptomics

AT was performed on all leaf GSLs grouped by type (aliphatic, indole and aromatic) (Appendix 3 and Kittipol *et al.*, 2019b). In the SNP-based association study, the total amount of aliphatic GSLs, has revealed remarkably strong associations with markers in tight regions of chromosome A9, C2 and C9 above the threshold for Bonferroni significance (p = 0.05) as well as weaker associations on chromosome A2 and C7 (Figure 4-1A). All five associated regions detected with leaf aliphatic GSL dataset here had been previously observed in the AT study with a smaller total seed GSL dataset (Lu *et al*., 2014 & Appendix 4). This indicates that leaf aliphatic GSLs and total seed GSLs, which are predominantly aliphatic GSLs, are controlled by the same loci in oilseed rape[4]. Using the pan-transcriptome CDS model V11, regions within the association peaks have been examined for highly associated markers that can be used to locate possible causative genes in close proximity. Orthologues of *HAG1/MYB28* (AT5G61420), a transcription factor that positively regulated aliphatic GSL biosynthesis, have been re-discovered within these associated regions in the leaf. For full details of the markers and associated genomic regions, see Spreadsheet 2 and Kittipol *et al.* (2019b).

In parallel with the analysis using SNP markers, AT was also performed using the gene expression matrix to identify gene expression markers (GEMs) that are correlated with the GSL trait variation (Figure 4-1B). Above the threshold for the false discovery rate (FDR) at 5%, six GEMs have been detected and shown to be involved directly in aliphatic GSL biosynthesis (Table 4-1 for summary and Spreadsheet 3 for more detail). Out of these, two genes are known to be involved in the aliphatic amino acid chain elongation: an orthologue of AT5G23020, a methythioalkymalate synthase (MAM3) on chromosome A3 and an

[4] The relationship between GSL contents of vegetative tissues and seeds will be further explored in Chapter 6

orthologue of AT5G23010 (MAM1) on C7. The other two are involved in the core GSL structure biosynthesis: an orthologue of AT1G16410, a cytochrome P450 CYP79F1, on C5 and an orthologue of AT1G78370, a glutathione S-transferase TAU 20 (GSTU20) on A7, respectively. Two orthologues of *HAG1*, *Bna.HAG1.A9* and *Bna.HAG1.C2*, are also identified as the top GEMs in tight linkage (p = 4.61×10^{-12} and p = 2.58×10^{-6} respectively). In particular, *Bna.HAG1.A9* has passed both the FDR and Bonferroni significance threshold. The identification of *Bna.HAG1.A9* and *Bna.HAG1.C2* as top GEMs along with SNP-based association peaks provides a strong indication that *HAG1* orthologues at these loci act as the causative genes controlling the levels of aliphatic GSLs in the leaf tissue.

Table 4-1. Summary of gene expression marker (GEM) associations for leaf aliphatic glucosinolate. These genes are amongst the top GEMs and are known to be involved in GSL biosynthetic pathway. Full detail of top GEMs is provided in Spreadsheet 3 and Appendix 11 of Kittipol *et al.* (2019b).

Candidate genes	Chromosome	Marker	TAIR	P value
Myb domain protein 28 (**HAG1**)	A09	Cab038298.3	AT5G61420.1	4.61 × 10^{-12}
	C02	Bo2g161590.1	AT5G61420.2	2.58 × 10^{-6}
Glutathione S-transferase TAU 20 (**GSTU20**)	A07	Cab018975.1	AT1G78370.1	5.86 × 10^{-7}
methythioalkymalate synthase 3 (**MAM3**)	A03	Cab001421.1	AT5G23020.1	3.94 × 10^{-5}
methythioalkymalate synthase 1 (**MAM1**)	C07	Bo7g098000.1	AT5G23010.1	4.40 × 10^{-5}
Cytochrome P450 79F1 (**CYP79F1**)	C05	Bo5g021810.1	AT1G16410.1	6.05 × 10^{-5}

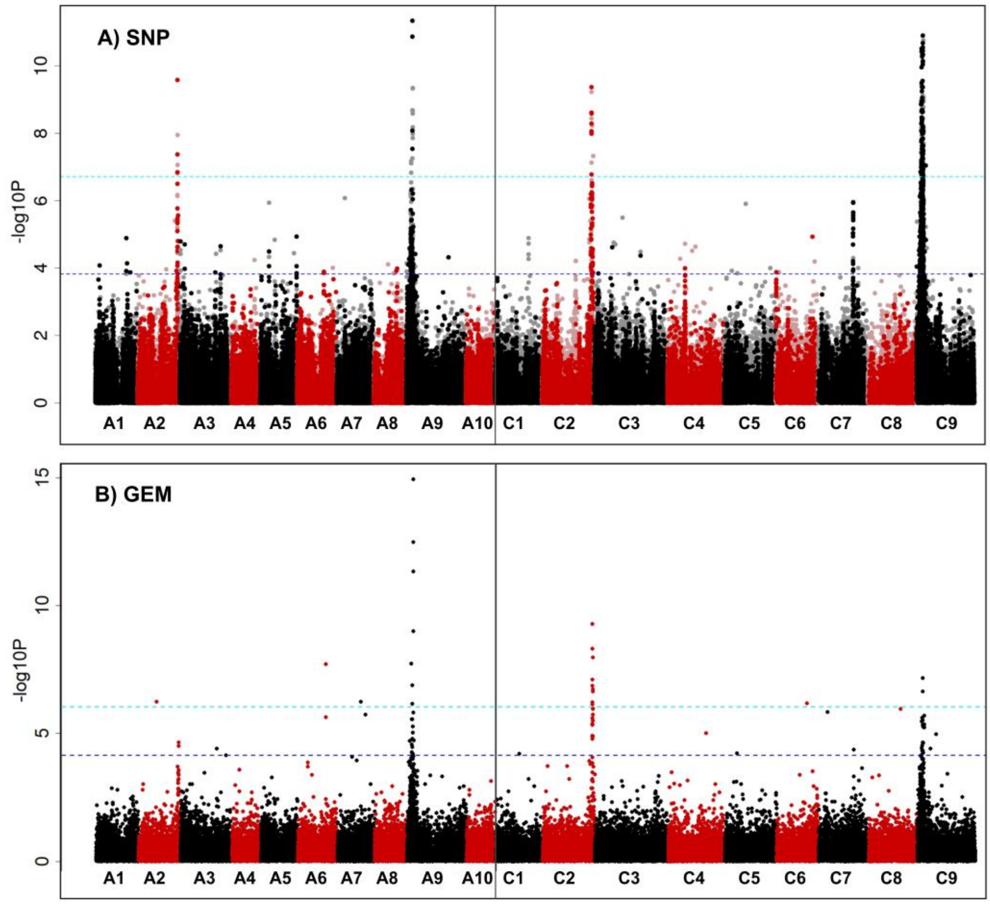


Figure 4-1. Association analysis for leaf aliphatic glucosinolate content. Manhattan plot showing genome-wide associations for the identification of transcriptome **(A)** single-nucleotide polymorphism (SNP) markers of 288 *B. napus* accessions with leaf GSL content. Marker associations was calculated using mixed linear models which incorporated population structure and relatedness. Dark opaque points are simple SNP markers and hemi-SNPs that have been directly linkage-mapped, both of which can be assigned to one genome, whereas light points are hemi-SNP markers (i.e. polymorphisms involving multiple bases called at the SNP position in one allele of the polymorphism) for which the genome of the polymorphism cannot be assigned. **(B)** Association analysis of expression variation-based markers (GEM) with leaf aliphatic glucosinolate. Reads per kb per million aligned reads (RPKM) were regressed against the trait, and R^2 and P values were calculated for each gene. The SNP and GEMs are positioned on the x-axis based on the genomic order of the gene models. The significance of the trait association, as $-\log_{10}$P values, plotted on y-axis. For both plots, the horizontal purple and cyan lines represent false discovery rate (FDR) threshold at 5% and the threshold for Bonferroni significance of 0.05, respectively. Chromosomes of *B. napus* are labelled A1– A10 and C1 – C9, shown in alternating black and red colours to allow boundaries to be clearly distinguished.

4.2.2 *Bna.HAG1* homoeologous gene expressions

Six copies of *HAG1* orthologues are found in the allopolyploid *B. napus* genome. To determine potential impacts of the expression of all of the *HAG1* orthologues on the GSL contents in *B. napus*, leaf transcript abundance[5] data of all six *HAG1* orthologues have been extracted and analysed from nine *B. napus* accessions. These include five accessions showing high levels of aliphatic GSLs in leaves and the other four with low leaf aliphatic GSL contents. The expression levels of both *Bna.HAG1.A9* and *Bna.HAG1.C2* have shown strong positive correlation with level of aliphatic GSL, whereas other homoeologous copies on A3, C7 and C9 are expressed at very low levels (Figure 4-2). The remaining homoeologous copy on A2, though highly expressed, was not correlated with the observed aliphatic GSL variation. This is consistent with the AT results described earlier that *Bna.HAG1.A9* and *Bna.HAG1.C2* play an important role in controlling the levels of leaf aliphatic GSLs. The homoeologous copy on A2, on the other hand, appears to either encode a non-functional protein or has lost its role in the control of leaf GSL biosynthesis by subfunctionalisation.

[5] Transcript abundance are quantified as reads per kilo base per million mapped reads (RPKM), which is a normalised unit of gene expression. Since the first step of gene expression is transcription of the genetic information in DNA into RNA, measuring the levels of mRNA transcripts (transcript abundance) provides a measurement of gene expression.

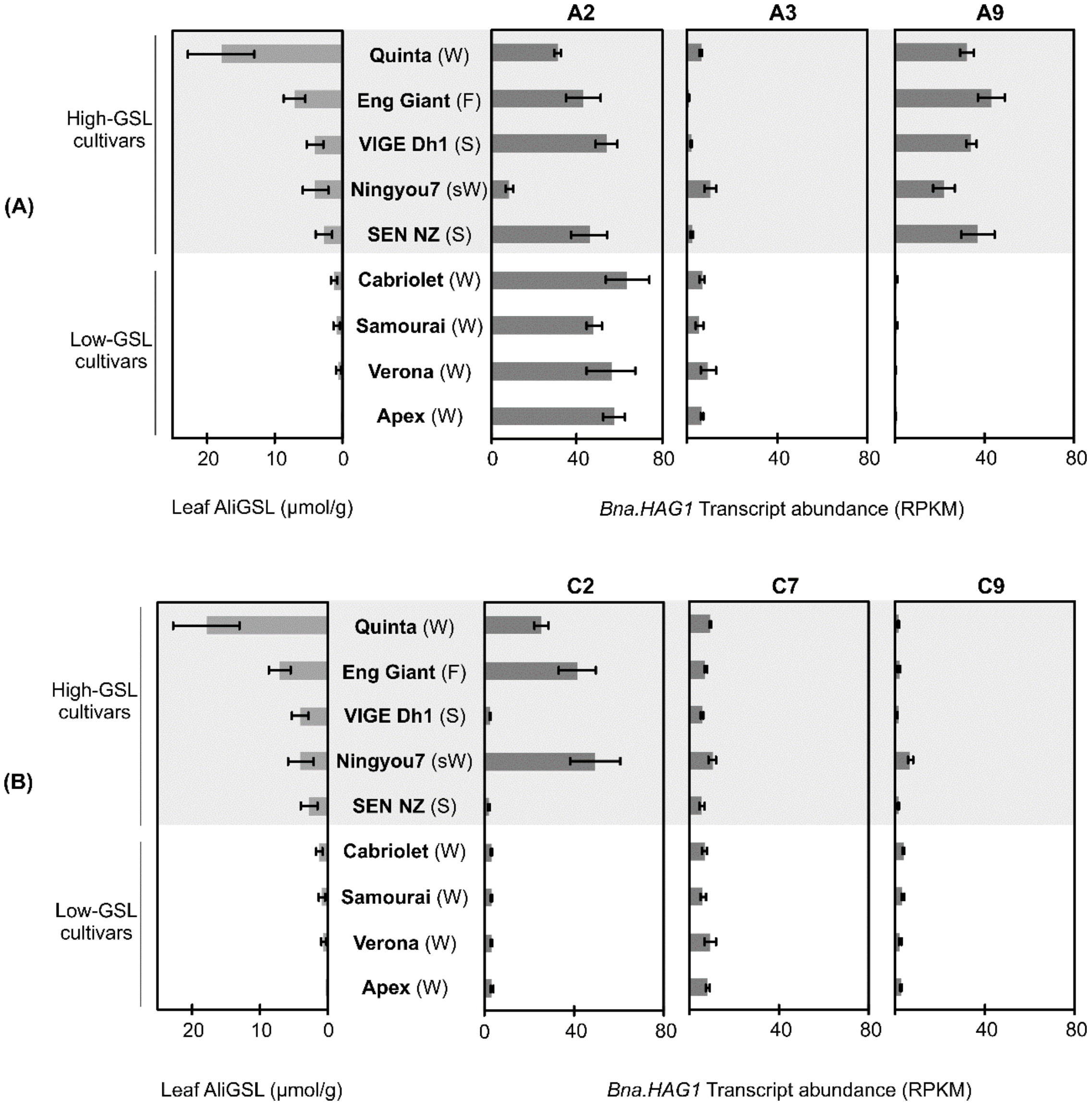
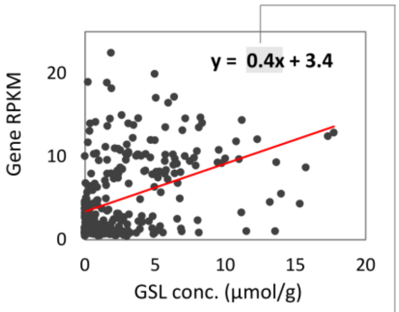


Figure 4-2. Expression of *Bna.HAG1* homoeologues in high- and low- leaf aliphatic GSL *B. napus* accessions. Six orthologues of *HAG1* (AT5G61420) are found in *B. napus*, on **(A)** chromosome A2, A3 and A9 in the A genome and on **(B)** chromosome C2, C7 and C9 in the C genome. Transcript abundance of Bna.HAG1 is expressed as reads per kb per million aligned reads (RPKM), with error bars to indicate standard deviation from four biological replicates of each accessions. Crop type abbreviation: W, winter oilseed rape; F, winter fodder; sW, semiwinter oilseed rape; S, swede.

4.2.3 Impact of structural genome variations on levels of aliphatic GSLs

The presence of GEM association peaks on chromosome A9, C2 and C9 (Figure 4-1B) suggests that many nearby genes have the same directionality of expression, which is indicative of structural changes in the genome causing variation that influence the level of aliphatic GSLs. Investigation into the associated GEM markers has revealed that all associated GEMs on chromosomes A9 and C2 are positively correlated to the aliphatic GSL trait. However, GEMs on chromosome C9 are negatively correlated. To see whether homoeologous[6] gene pairs within these peaks showed opposite effect to each other, directionality of gene expression was compared. This was achieved by plotting the transcript abundance of homoeologous genes against leaf aliphatic GSL content. The slope (m) of the simple linear regression line $(\mathrm{y} = \mathrm{mx} + \mathrm{c})$ was then calculated and used to indicate either a positive or a negative direction of the correlation. The result shows an opposite direction of expression of homoeologous genes between related chromosome pairs A9/C9 and A2/C2 (Figure 4-3). This pattern is suggestive of potential blocks of homoeologous exchange as many neighbouring genes have shown the same directionality: one of the A or C genomes over-expressed and the counterpart genome under-expressed.

[6] The term ‘homoeologues’ refers to pairs of genes or ‘corresponding’ genes derived from different species that were brought back together in the same genome by allopolyploidisation (Glover *et al.*, 2016)



	Chromosome	
Orthologue	**A2**	**C2**
AT5G61960.2	-0.2	0.3
AT5G62130.1	-0.2	0.4
AT5G62090.2	-0.1	0.1
AT5G62200.1	-0.1	0.1
AT4G28360.1	-0.1	0.1

	Chromosome	
Orthologue	**A9**	**C9**
AT2G19400.1	0.7	-0.8
AT5G48385.1	0.1	-0.6
AT5G23080.1	0.7	-0.5
AT5G23370.1	0.4	-0.3
AT5G26610.3	0.1	-0.3
AT3G04970.1	0.5	-0.3
AT5G25150.1	0.3	-0.3
AT5G24710.1	1.5	-0.3
AT3G27000.1	0.3	-0.3
AT5G63110.1	0.3	-0.2
AT5G23890.1	0.1	-0.2
AT5G25060.1	0.2	-0.2
AT5G23090.1	0.2	-0.2
AT5G25265.1	0.3	-0.2
AT5G48160.2	0.1	-0.1
AT3G27240.1	0.1	-0.1
AT3G27230.1	0.1	-0.1
AT5G67590.1	0.1	-0.1
AT5G60920.1	0.0	0.0
AT5G23920.1	0.0	0.0
AT5G23900.1	0.1	0.0
AT5G23740.1	0.0	0.0
AT5G66570.1	0.0	0.0

Figure 4-3. Opposite expression of homoeologous genes on chromosome A2, A9, C2 and C9 in the highly associated GEM clusters. Orthologous *Arabidopsis* TAIR ID were used to identify homoeologous genes on different chromosomes. Transcript quantification, of *B. napus* homoeologous gene, expressed as reads per kb per million aligned reads (RPKM), was plotted against leaf aliphatic GSL content. Regression was calculated as shown by the example of a regression plot. The value of m from the simple linear regression line $(\mathrm{y} = \mathrm{mx} + \mathrm{c})$ are shown in the table. The value of m shows the direction of the regression as positive (green) or negative (red).

Transcriptome display tile plot was used to visualise regions of the genome for presence of homoeologous exchanges. The result shows that homoeologous exchanges has occurred within the associated regions. The segments of C genome are lost and replaced by the duplicated copies of their corresponding homoeologous regions in the A genome (Figure 4-4). This explains the opposite polarity of homoeologous genes and the presence of GEM peaks. The presence of homoeologous exchanges in this particular region of the genome affects the gene copy numbers, which may subsequently affect levels of aliphatic GSLs in *B. napus*. In fact, it has been found that in high-aliphatic GSL cultivars, functional *Bna.HAG1.A9* is duplicated but the homoeologous copy of non-functional *Bna.HAG1.C9* is lost as a consequence of homoeologous exchanges, and vice versa for low-aliphatic GSL cultivars (Figure 4-5). Homoeologous exchanges between chromosome A2 and C2 have been found to occur to a lesser extent in most crop types (Figure 4-6). The only exception is observed in swede where homoeologous exchange polarity displays a different pattern to other crop types. In particular, a large region within the association peak on chromosome C2, containing a functional *HAG1* orthologue, is lost in all swede accessions including the high-GSL cultivars (Figure 4-2B and Figure 4-6). This suggests that *Bna.HAG1.A9* alone is sufficient in controlling variations of leaf aliphatic GSLs in the swede crop type, whereas both *Bna.HAG1.A9* and *Bna.HAG1.C2* are the key regulators of aliphatic GSL variations in the leaf tissue of other crop types.

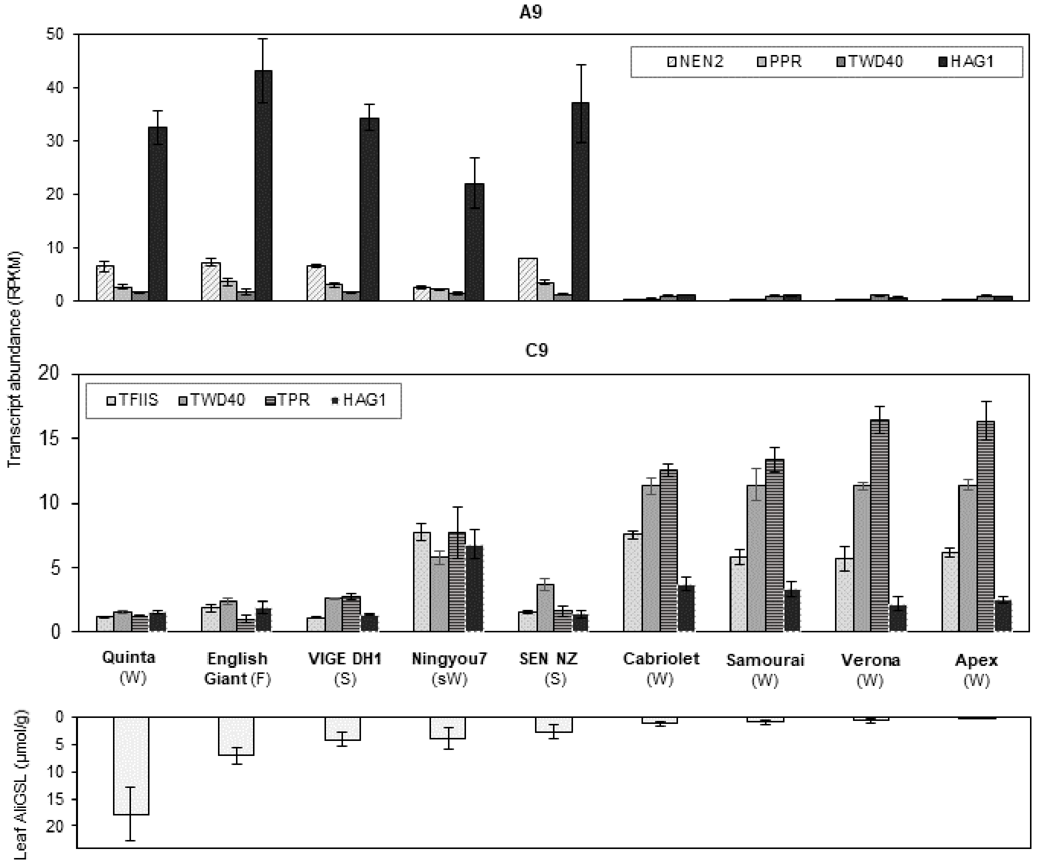


Figure 4-4. Comparison of the top four GEMs transcript abundance on chromosome A9 and C9 between high- and low- leaf aliphatic glucosinolate *B. napus* cultivars. Transcript abundance is expressed as reads per kb per million aligned reads (RPKM), with error bars to indicate standard deviation from four biological replicates of each accessions. Orthologue of *HAG1* on chromosome C9 is not one of the top markers but included for comparison. Crop type abbreviation: W, Winter oilseed rape; F, Winter fodder; S, Swede; sW, Semiwinter oilseed rape.

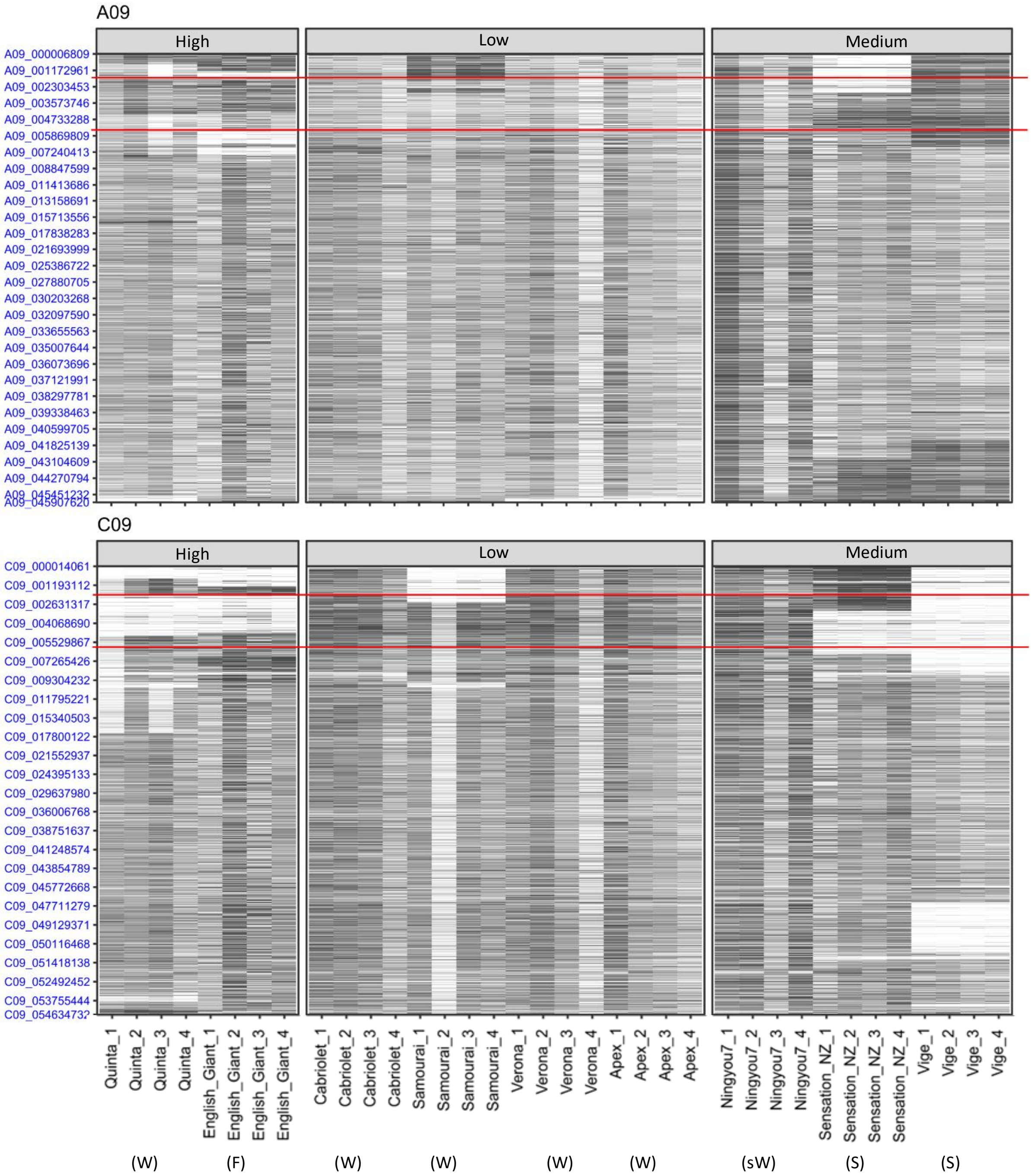


Figure 4-5. Transcriptome Display Tile Plots illustrate homoeologous genome exchanges between A9/C9 chromosome pair in *B. napus* based on leaf gene expression data. The positions of the GEM association peaks on chromosome pair A9 and C9 are marked with red horizontal lines. The top plot shows the relative transcript abundance of A genome on chromosome A9, the bottom plot shows the relative transcript abundance of C genome on chromosome C9, plotted with four biological cultivar replicates. Darker or lighter shade of grey represent increased or decreased gene expression. Inverted colour of the same region on the different chromosome show homoeologous exchange. Cultivars are grouped according to the total aliphatic GSL content in the leaves: high (>7 µmol/g), medium (2 - 7 µmol/g) and low (<2 µmol/g). Crop type abbreviation: W, Winter oilseed rape; F, Winter fodder; S, Swede; sW, Semiwinter oilseed rape.

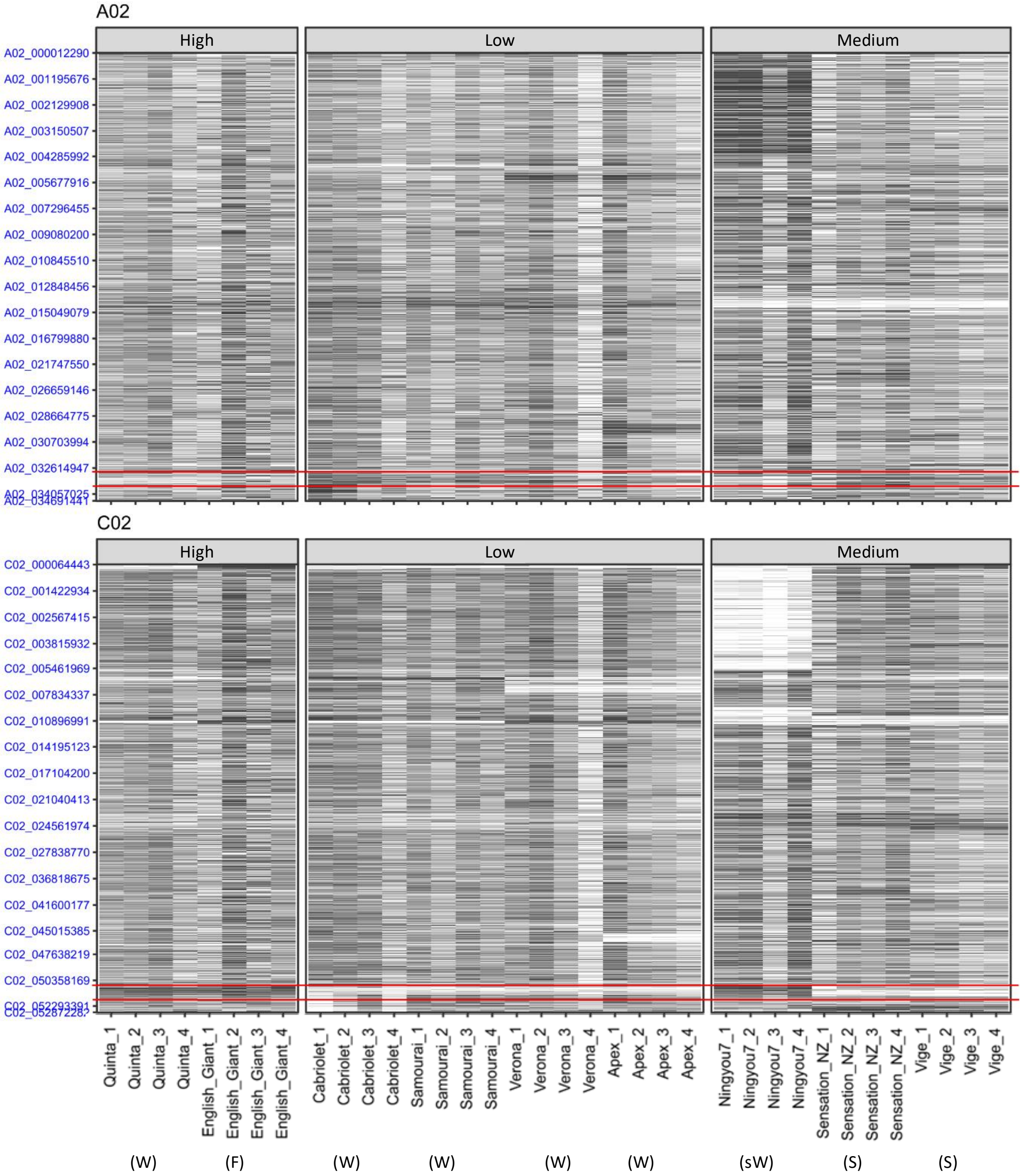


Figure 4-6. Homoeologous genome exchanges between A2/C2 chromosome pair have occurred to a lesser extent within the leaf aliphatic GSL associated regions in *Brassica napus*. The positions of GEM association peaks on chromosome A2 and C2 are marked with horizontal red lines. The top plot shows the relative transcript abundance from leaf mRNA sequence data of A genome on chromosome A2, the bottom plot shows the relative transcript abundance of C genome on chromosome C2, plotted with four biological cultivar replicates. Darker or lighter shade of grey represent increased or decreased gene expression. Inverted colour of the same region on the different chromosome show homoeologous exchange. Cultivars are grouped according to the total aliphatic GSL content in the leaves: high (>7 µmol/g), medium (2 - 7 µmol/g) and low (<2 µmol/g). Crop type abbreviation: W, Winter oilseed rape; F, Winter fodder; S, Swede; sW, Semiwinter oilseed rape.

4.3 Minor leaf glucosinolates

The results for genetic associations are weak for leaf indole and aromatic GSLs. Some observations have been noted and described here as these may help to illuminate the underlying mechanisms of these minor leaf GSLs in future studies.

In the AT analysis of the leaf indole GSLs, none of the GEM-based associations have passed the FDR threshold (Appendix 3). Nonetheless, the top three markers are on chromosome A1, A6 and A9, but none of these genes has been known to be directly involved in indole GSL metabolism. It is possible that they may have an indirect effect on the levels of indole GSLs. These three markers are: two homoeologous copies of FatA thioesterase (AT3G25110) on chromosomes A1 and A6 and the orthologue of PP2A-2 (protein phosphatase 2A-2, AT1G10430) on chromosome A9. FatA is involved in oil content and fatty acid composition in seeds, and PP2A-2 act as a negative regulator of abscisic pathway. Unlike the case of aliphatic GSLs, homoeologous exchanges affecting the levels of indole GSLs seems to occur at the bottom of the chromosomes between functional copies on A9 and non-functional copies on C9. The SNP-based association studies for the leaf indole GSL trait has not identify any association peaks passing the FDR threshold (Appendix 3), therefore it is not possible to detect candidate genes for this trait with the current data.

Similarly, the AT analysis with the current dataset has not yielded any meaningful associations for the leaf aromatic GSL trait. The top associated GEM markers above the Bonferroni significance threshold (p = 0.05) are orthologous genes of AT1G31260, AT1G45688 and AT1G14730 that encode a protein involving in zinc ion transmembrane transport activity, a transmembrane protein and a membrane-associated protein respectively. There is no obvious candidate genes in the region of associations from the SNP association analysis.

4.4 Chapter discussion

Since the same associated regions has been found for both the total GSLs and for aliphatic GSLs in the leaves of *B. napus* (Appendix 3), identifying the genetic controls of aliphatic GSLs can help to elucidate patterns of overall GSL accumulation in this tissue. The *MYB28/HAG1* transcription factor has been identified as the main regulator of aliphatic GSLs in studies on transcriptional regulation of GSL biosynthesis in *A. thaliana* (Gigolashvili *et al*., 2007b; Hirai *et al*., 2007; Sonderby *et al*., 2010). Orthologues of *HAG1* have also been identified as the main regulator of aliphatic GSLs in the leaves of closely related *B. napus*. Unlike the diploid Arabidopsis, *B. napus* has a polyploidy genome which contains multiple copies of *Bna.HAG1*. AT and homoeologous gene expression analyses have been proven invaluable to differentiate the roles of these homoeologous copies. Only *MYB28/HAG1* orthologues on A9 and C2 chromosomes have shown to control the natural variations of aliphatic GSLs and therefore controlling variations of total GSLs in *B. napus* leaves (Figure 4-1, Figure 4-2 & Kittipol *et al*., 2019a).

Homoeologous exchanges occur frequently in recently formed allopolyploids such as *B. napus*, particularly near the telomeres where homoeologues are able to pair most efficiently (He *et al*., 2016). This pattern of structural changes in the genome matches the region of structural rearrangements in Figure 4-5 and Figure 4-6. Thus, genetic variation for the GSL level in leaves is due to the structural changes via homoeologous exchanges in the region of *B. napus* genome containing orthologues of *HAG1*. In low-leaf aliphatic GSL accessions such as 'Cabriolet and 'Apex', the functional *HAG1* orthologue near the top of chromosome A9 is lost and the functional *HAG1* orthologue near bottom of chromosome C2 is replaced by the non-functional A2 orthologue in a homoeologous substitution event. The findings from this experiment are consistent with the draft genome assembly of the low-GSL cultivar 'Darmor-bzh' where orthologues of *HAG1* are absent on A9 and C2 chromosomes (Chalhoub *et al*., 2014).

CHAPTER 5

Genetic control of root glucosinolate variation

Roots of *B. napus* are predominated by both aromatic GSLs and indole GSLs, but the level of total GSLs in roots is largely determined by the level of aromatic class (Section 3.4; r = 0.75, $p \leq 0.001$). This chapter examines the genetic control of aromatic GSL variations, which will elucidate the regulation of majority of the root GSL accumulation pattern in *B. napus* (Section 5.3). For this investigation, three analytical techniques have been used, which divided the section into two parts. First, Section 5.3.1 uses both AT and differential expression analysis to identify potential candidate gene, *Bna.HAG3.A3,* for aromatic GSL control. Second, Section 5.3.2 uses weighted gene co-expression network analysis to examine genes that co-express with the transcription factor *Bna.HAG3.A3.* Then, Section 5.4 describes the results for the genetic associations of aliphatic and indole GSLs. Finally, Section 5.5 uses GSL ratios as traits for additional AT analyses to explore loci controlling proportion of a class of GSL in relation to other structural classes.

5.1 Introduction

Little information is available on the biosynthetic pathway of the chain-elongated homophenylalanine-derived aromatic GSLs (i.e. GST). This is because few ecotypes of *A. thaliana*, a model organism in plant molecular biology, produce aromatic GSLs and in minor amounts when they do (Brown *et al.*, 2003). The inability to use *A. thaliana* and the challenges of working with the complex polyploidy of *B. napus* have limited the advancement in the understanding of the biosynthetic pathway of these aromatic GSLs. In this chapter, three powerful experimental techniques, namely Associative Transcriptomics (AT), differential expression (DE) analysis and weighted gene co-expression network analysis (WGCNA), are described to demonstrate the power of the combinational analyses which results in deciphering the underexplored root GSL trait.

Many studies have demonstrated the capacity of AT in identifying genes controlling various traits in *B. napus* such as erucic acid (Havlickova *et al*., 2018), seed GSL (Lu *et al*., 2014) and leaf aliphatic GSL as described in Chapter 4. This chapter reveals the power of SNP-based association studies of AT for root traits and has led to the identification of a potential candidate gene of a root trait through strong linkage disequilibrium. Transcriptome from juvenile leaves provides a wide dynamic range of transcript abundance levels that allows SNP calling from very highly expressed genes to those with barely detectable expression (Trick *et al*., 2009). Since these SNPs represent variations within a genome which are uniform in every cells with a nucleus throughout the plant, it is possible for SNP-based association of AT to identify marker associated with root traits. On the other hand, the power of GEM association studies with the same leaf RNA-seq is likely to be limited for deciphering root traits, as some may have root-specific gene expression and not represented in the leaf derived GEM dataset. To investigate whether expression of the candidate genes correlate to the variations in root traits in a time- and cost-efficient approach, root RNA-seq data are used to perform root DE analysis from a small subset of accessions either with high- or low- root aromatic GSL levels. In conjunction with root DE analysis, WGCNA will be carried out to explore the genes with similar pattern of expression to the candidate genes identified from AT analysis. In WGCNA, highly correlated genes are grouped into modules based on the similarities of their expression profiles. Each modules are often enriched for genes that share similar biological functions. Therefore, genes that co-expressed with the candidate genes are likely to be part of the same biosynthetic pathway and potentially regulated by the same regulator. Hub genes are highly connected to other genes in the network. Identification of hub genes in the network can provide candidates that may play important roles in a biological system. Through the combination of three analytical techniques, this chapter aims to address the research question on which genes are involved in the genetic control and the molecular basis of the side-chain elongation of homophenylalanine-derived aromatic GSLs, a class that is abundant in *Brassica* roots but with a gap in the knowledge of their pathways and regulation.

5.2 Methods

5.2.1 Associative Transcriptomics

To elucidate genetic loci controlling GSL variations in root, AT was performed on a panel of 288 *B. napus* accessions. The genotypes used in the AT was derived from re-sequencing of leaf transcriptome, consisting of 355,536 SNP markers and gene expression matrix with a transcriptome reference of 116,098 ordered coding DNA sequence gene models (Havlickova *et al*., 2018). Estimate of allelic effects were produced from the Compressed Mixed Linear Model as part of GAPIT output. Detail method of the AT is described in Chapter 2 (Section 2.4).

5.2.1.1 Using ratio as traits

In section 5.5, ratios of GSLs were used as traits to investigate loci controlling proportion of a class of GSL present in relation to other classes of GSLs. The ratio was calculated by dividing each GSL classes with total root GSL, for instance

$$aliphatic\ ratio = \frac{root\ aliphatic\ GSL}{total\ root\ GSL}.$$

5.2.2 Root Differential Expression analysis

5.2.2.1 Subset of plant materials for root DE experiment

Eight accessions were selected for the root DE analysis based on their differed root aromatic GSL content. The GSL profiles of this subset of high- or low-root aromatic GSL groups are shown in Table 5-1. Plants were grown following the same conditions as the diversity panel described in Chapter 2 (Section 2.1). Four biological replicates of each accessions were grown in root trainers with Terra-Green for ease of root harvesting, supplemented weekly with half concentration of Murashige and Skoog growth medium adjusted to pH6.5 with KOH. Four weeks after sowing, plants were removed from the tray, the roots were washed, dried with paper towel and cut at the base. Each root was separated into two samples: one for RNA extraction and the other for GSL quantification. At harvest, the sample for RNA extraction was collected into an Eppendorf tube with metal beads on dry ice, while the sample for GSL quantification was wrapped in labelled aluminium foils and immediately frozen in liquid nitrogen. All samples were stored at -80 °C for further processing. GSL quantification was carried out following the details in Chapter 2 (Section 2.2). For RNA extraction, samples were homogenised to fine powder with a steel bead for 10 min at a frequency of 30 Hz (TissueLyser II, Qiagen) before extracting the RNA using the E.Z.N.A® Plant RNA Kit (Omega Bio-Tek) following the manufacturer's manual and sent for sequencing.

Table 5-1. Eight accessions with different levels of aromatic glucosinolates were selected for root differential analysis. The two accessions, N01D-1330 and KARAT, were selected for further stringent analysis. These accessions, highlighted in yellow, were different only in the levels of root aromatic glucosinolate but similar in other glucosinolates. Mean GSL from four biological replicates are shown. Abbreviation: Aro, aromatic GSL; Ali, aliphatic GSL; Ind, indole GSL.

Group	Accession	Cultivar	Root GSL (µmol/g)			Leaf GSL (µmol/g)		
			Aro	Ali	Ind	Aro	Ali	Ind
High	a-0000255	N01D-1330	11.19	0.05	3.95	0.07	0.04	0.81
High	a-0000348	Olivia	11.15	1.35	2.50	0.41	8.13	1.19
High	a-0000162	Moldavia	11.04	3.46	3.14	0.41	13.98	1.18
High	a-0000107	JANETZKIS SCHLESISCHER	9.64	1.02	1.94	1.01	7.91	1.28
Low	a-0000248	KARAT	0	0.05	3.66	0	0	0.63
Low	a-0000270	Bronowski	0	0	6.23	0	0.03	1.91
Low	a-0000431	Brandhaug	0	1.07	3.71	0.11	2.12	3.99
Low	a-0000442	Troendersk Kvithamar	0	0.96	6.21	0	0.91	3.45

5.2.2.2 *Root transcript quantification*

Following the same method as leaf RNA-seq, as described in Chapter 2 (Section 2.4.2), root transcript abundance was quantified and normalised as reads per kb per million aligned reads (RPKM) for all coding DNA sequence (CDS) models of pan-transcriptome reference for each sample. CDS models with a mean expression across the panel below 0.4 RPKM were removed.

5.2.2.3 *Differential expression analysis*

The root transcript abundance dataset, represented as RPKM values derived from four biological replicates (i.e. using root RNA-seq from 4 separate plants of each accession), was used to perform DE analysis. The methods in Bioconductor package EdgeR (Robinson *et al*., 2009) were used to identify the differential expressed genes between the 'High' and 'Low' GSL groups. In multiple comparisons, both fold change (FC) > 2 and false discovery rate (FDR) < 0.05 were used to flag a gene being differentially expressed. The processing of DE data from raw RNA-seq was carried out by Dr. Zhesi He at the University of York. Interpretation and analysis of the processed results was carried out by the author of this thesis.

To limit potential confounding effect between GSL pathways in the analysis, an even more stringent root differential expression experiment ($\log_2$fold-change ≥ 4; $p \leq 1 \times 10^{-10}$) was performed between accessions, N01D-1330 and KARAT, which differ in root aromatic GSLs but both of which are low in aliphatic GSLs (Table 5-1).

5.2.3 Weighted Gene Co-expression Network analysis

Weighted correlation network analysis was performed using the 'WGCNA' `R` software package (Langfelder and Horvath, 2008).

5.2.3.1 Data input and cleaning

WGCNA between N01D-1330 and KARAT used a subset of the root transcript abundance dataset from the stringent root differential expression experiment to limit potential confounding between aliphatic and aromatic GSL pathways. In order to reduce noise in the expression data, records of genes with fold change (FC) <2 and significance level (*p*) >0.05 were removed from further analyses, leaving a total of 603 gene records in the list. This subset has eight data points in total (two accessions, each with four biological replicates) for each of the 603 genes. Concentrations of root aromatic GSLs for corresponding accession samples were used as trait data. Data input was carried out following the methods in Langfelder and Horvath (2014a).

5.2.3.2 Network construction and module detection

Automatic network construction using soft thresholding power was carried out following the methods described in Langfelder and Horvath (2014b). The soft thresholding power of 16 was selected based on the criterion of approximate scale-free topology. The gene network was constructed with a medium module size of 20 and a medium sensitivity to cluster splitting (`deepSplit=2`). Six modules were identified, with the size ranging from 214 to 41 genes.

5.2.3.3 Quantifying module-trait associations

Summary profiles (eigengenes) of each module were used to correlate with trait values in order to identify modules that are significantly associated with the traits. The `R` codes in Langfelder and Horvath (2014c) were used for this step. Module correlation and correlations (weight) of individual genes with aromatic GSL trait was quantified.

5.2.3.4 Network visualisation with Cytoscape

Cytoscape software (Shannon *et al.*, 2003) was used to visualise the 'red' module network data. Gene correlations (weight) were used in the edge file and the gene names were used in the node file. Genes with weight threshold more than 0.3 were exported and visualised on Cytoscape.

5.2.3.5 Analysis of biological network

'Network Analyzer', a plugin for Cytoscape, was used to perform analysis of the 'red' module network. In a gene network, nodes represent genes and edges represent the interactions between genes as determined by the pairwise correlations between gene expressions. Several parameters can be used to provide information about the network and help identify hub genes (Vlab.amrita.edu, 2012). Degree (or connectivity) of a node is the number of edges linked to it. Neighbourhood connectivity and topological coefficient gives an average connectivity of all neighbouring nodes and a measure for the extent to which a node shares neighbours with other nodes, respectively (Mpi-inf, 2018). All three of these parameters essentially provide information on how central and well connected a node is, which helps in the identification of the hub gene.

5.2.4 Gene ontology (GO) analysis

AgriGO v2.0, a gene ontology (GO) database for agricultural species, was used for GO analysis (Tian *et al*., 2017). The *B. napus* genes in the 'red' module were used to identify their matching Arabidopsis orthologues. The gene IDs of these Arbidopsis orthologues were used as an input data for the analysis. Singular Enrichment Analysis of *A. thaliana* orthologue genes was compared against 'TAIR10_2017' reference database. The Fisher's exact test and the Yekutieli (FDR) for multi-test adjustment method were selected to map the query list to reference list and provide *p*-value of significant GO terms. Queries not reaching the significance level threshold at 0.01 and a minimum number of five mapping entries were filtered out.

Results

5.3 Genetic control of root aromatic glucosinolates

5.3.1 The combined power of AT and root DE analysis to uncover the genetic control of root glucosinolates

To identify potential loci controlling the level and composition of GSLs in roots, AT has been conducted for all GSLs grouped by structural class (aliphatic, indole or aromatic) (Appendix 5 & Kittipol *et al*., 2019b). Only one type of aromatic GSL, 2-phenethyl (GST) glucosinolate, has been found in *B. napus*. SNP association analysis of this aromatic GSL (i.e. GST) has revealed a strong association peak at the top of chromosome A3 (Figure 5-1A). This A3 SNP association peak lies in a region of the genome that has not been seen previously to be associated with GSL contents. This novel locus includes seven genome-assigned SNP markers above the Bonferroni corrected significance threshold (*p* = 0.05), encompassing approximately 95 genes (Spreadsheet 4). These genes are involved in a wide range of biological processes such as transcriptional regulation, vesicle-mediated transport and defence responses to pathogen, but none has been previously identified as involving in the aromatic GSL pathway. Interestingly, an orthologue of *HAG3/MYB29* (AT5G07690), a transcription factor that has been shown to control aliphatic GSL biosynthesis in *A. thaliana*, is in close proximity to the top associated SNP markers in this region (Spreadsheet 4 & Kittipol *et al.*, 2019b).

While the SNP Manhattan plot has shown the above significant associated locus on A3, none of the gene expression markers (GEM) showed significant association above the false-discovery rate (FDR) threshold value in the parallel GEM association analysis (Figure 5-1B). This may be due to the fact that young leaves were used as the source materials for RNA-seq and the reads of these were used to generate both the SNP and GEM datasets in the AT analyses. Therefore, GEMs would not be identifiable for genes with root-specific expression patterns from this GEM dataset.

To address this, root RNA was extracted from four high root GSL and four low root aromatic GSL accessions and the root transcriptome re-sequencing was used to perform DE analyses. Root DE analysis used normalised RPKM read count data in statistical analysis to examine the quantitative changes of the expression levels between the groups. Within the SNP associated region on chromosome A3 described above, *Bna.HAG3.A3* has shown the biggest significant differences in expression between the high- and low-root aromatic GSL groups ($\log_2$FC = 14.76; $p = 5.47\times10^{-11}$) and the strongest correlation of transcript abundance with aromatic GSL content compared to other genes (Table 5-2). The transcript abundance of *Bna.HAG3.A3* is high in the high-root aromatic GSL group and very low in the low-root aromatic group (Figure 5-2). Although orthologue of *HAG3* has not been implicated previously in the control of root aromatic GSL, results from both the SNP association and root DE analysis indicate *Bna.HAG3.A3* as an excellent candidate for controlling this trait.

Since the correlation result from Chapter 3 (Section 3.4) has shown a significant positive relationship between the aliphatic and aromatic GSLs, it is possible that the findings on *Bna.HAG3.A3* from the root DE analysis is due to the connection between the aliphatic and aromatic GSL pathways. In order to limit potential confounding effect between GSL pathways in the analysis, a more stringent root differential expression analysis ($\log_2$fold-change ≥ 4; $p \leq 1 \times 10^{-10}$) was performed between N01D-1330 and KARAT accessions. These two accessions differ in root aromatic GSLs but both are low in aliphatic GSLs (Table 5-1). The result is consistent with the previous finding, supporting *Bna.HAG3.A3* as an excellent candidate gene located within the associated region.

In addition, this stringent analysis has revealed 107 genes with top BLAST hits to annotated *A. thaliana* genes. These include: an orthologue of *MAM3* (AT5G23020) on chromosome A3, an orthologue of *IPMI2* (AT2G43100) on chromosome C4, an orthologue of *UGT74C1* (AT2G31790) on chromosome A5, and orthologues of *CYP83A1* (AT4G13770) on each of chromosome A4 and C4 (Table 5-3) (For more detail see Spreadsheet 5 & Kittipol *et al*., 2019b). All of these have been reported in Arabidopsis studies to be involved in the aliphatic GSL pathways. In *B. napus*, the high root aromatic GSL accessions have shown higher expression levels of these genes in roots (Figure 5-3), indicating that these genes may also be involved in the biosynthesis of the aromatic GSLs.

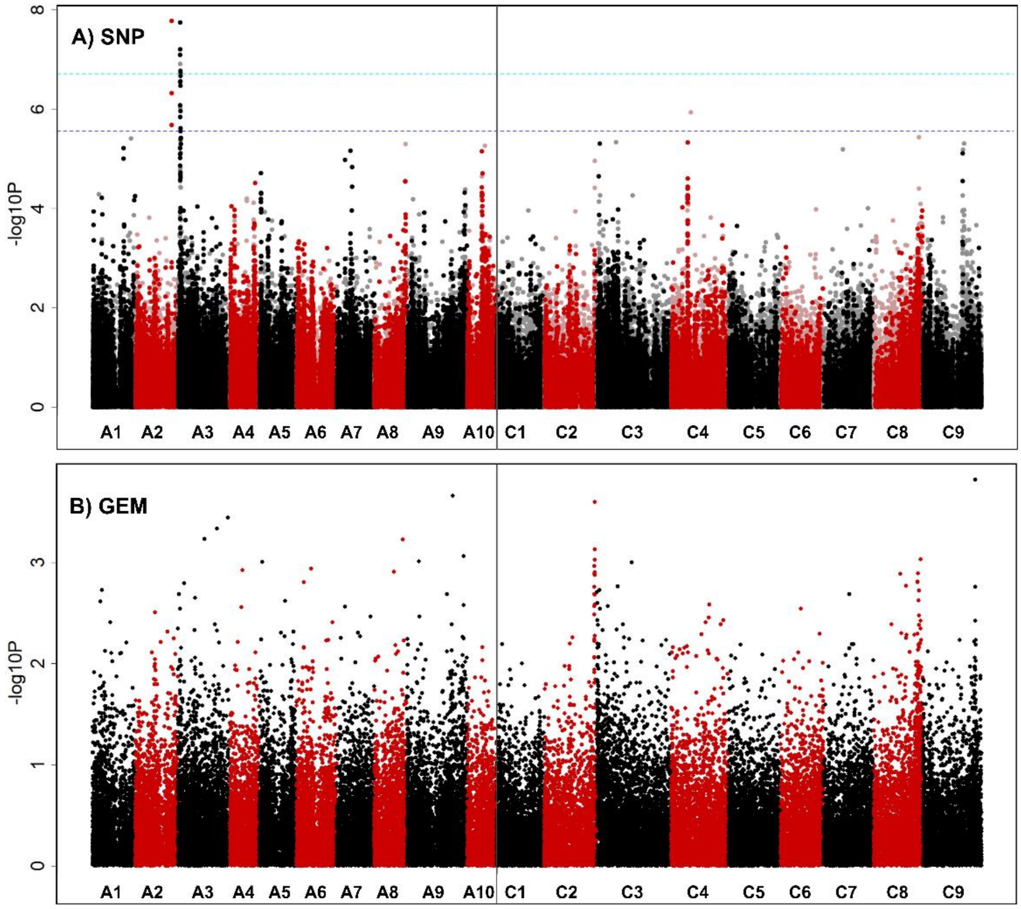


Figure 5-1. Association analysis for root aromatic glucosinolate content. Manhattan plot showing genome-wide associations for the identification of potential candidate markers **(A)** single-nucleotide polymorphism (SNP) markers of 288 *B. napus* accessions with root aromatic GSL content. Marker associations was calculated using mixed linear models which incorporated population structure and relatedness. Dark opaque points are simple SNP markers and hemi-SNPs that have been directly linkage-mapped, both of which can be assigned to one genome, whereas light points are hemi-SNP markers (i.e. polymorphisms involving multiple bases called at the SNP position in one allele of the polymorphism) for which the genome of the polymorphism cannot be assigned. **(B)** Association analysis of expression variation-based markers (GEM) with root aromatic GSL. Reads per kb per million aligned reads (RPKM) were regressed against the trait, and R^2 and P values were calculated for each gene. The SNP and GEMs are positioned on the x-axis based on the genomic order of the gene models. The significance of the trait association, as –log10P values, plotted on y-axis. For both plots, the horizontal purple and cyan lines represent false discovery rate (FDR) threshold at 5% and the threshold for Bonferroni significance of 0.05, respectively. Chromosomes of *B. napus* are labelled A1– A10 and C1 – C9, shown in alternating black and red colours to allow boundaries to be clearly distinguished.

Table 5-2. Root differential expression analysis of the associated region on chromosome A3. Genes with significance value $p < 0.001$ are shown and arranged in genomic position. Log_2FC (fold-change) is the log-ratio of a gene's expression value between the high- and low- aromatic GSL groups. Root's RPKM were regressed against root aromatic GSL level and R^2 were calculated. Orthologue of *HAG3/MYB29* is highlighted. Colour gradient was applied to help distinguish the high and low range of values in the columns.

Gene	TAIR	Annotation	Log_2FC	*P* value	R^2
Cab015489.1			1.66	8.26E−06	0.27
Cab015514.1	AT5G05110.1	Cystatin family protein	-6.32	9.13E−15	0.05
Cab015525.1	AT5G05200.1	Kinase superfamily protein	2.06	2.10E−06	0.54
Cab015527.1	AT5G05340.1	Peroxidase superfamily protein	-9.18	1.26E−07	0.35
Cab015528.1			-4.73	1.12E−06	0.05
Cab015532.1	AT5G05480.1	Peptide−N4−amidase A protein	2.10	9.50E−06	0.32
Cab015543.2	AT5G05780.2	RP non−ATPase subunit 8A	4.28	5.26E−12	0.36
Cab015544.1	AT5G05820.1	Nucleotide−sugar transporter	3.17	8.34E−16	0.43
Cab015547.1	AT5G05980.2	DHFS−FPGS homolog B	4.26	9.44E−04	0.31
Cab015548.1	AT5G05987.1	Prenylated RAB acceptor 1.A2	1.51	6.35E−05	0.22
Cab015551.1	AT5G06110.2	DnaJ and Myb−like DNA−binding domain	2.32	6.06E−08	0.45
Cab015552.1	AT5G43690.1	P−loop hydrolases superfamily protein	-11.35	1.56E−13	0.40
Cab015553.1	AT5G06110.2	DnaJ and Myb−like DNA−binding domain	-1.70	1.40E−05	0.25
Cab015556.1			-4.85	1.52E−04	0.22
Cab015558.2	AT5G06160.1	splicing factor−related	2.48	1.69E−06	0.28
Cab015561.1	AT5G06220.2	LETM1−like protein	-2.14	1.07E−04	0.31
Cab015568.1			14.26	5.46E−11	0.17
Cab015575.1			4.06	5.26E−04	0.74
Cab015589.2			-12.90	5.96E−18	0.11
Cab015616.1	AT5G07300.1	BONZAI 2	1.38	1.02E−04	0.14
Cab015618.1	AT5G07350.2	TUDOR−SN protein 1	1.47	5.90E−04	0.13
Cab015626.1			3.66	2.46E−05	0.25
Cab015634.1	AT5G07690.1	HAG3/MYB29	14.76	5.47E−11	0.80
Bra005951	AT5G07730.1		5.65	1.47E−09	0.78

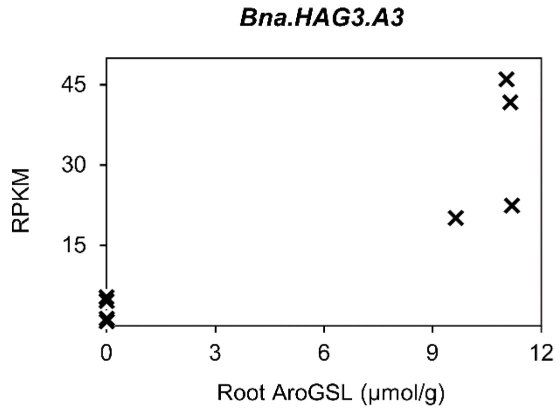


Figure 5-2. ***Bna.HAG3.A3*** **root expression against root aromatic GSL levels.** Root expression shown as reads per kilo base of transcript per million mapped reads (RPKM). Means of four biological replicates of eight *B. napus* accessions from the root differential expression experiment are shown.

Table 5-3. Stringent root differential expression analysis ($p \leq 1\times10^{-10}$) between N01D-133 and KARAT has revealed genes that are known to be involved in the glucosinolate biosynthetic pathway. For the all the top differentially expressed genes see Spreadsheet 5 or Appendix 16 of Kittipol *et al.* (2019b).

Annotation	**Position (bp)**	**Gene**	**TAIR**	**Log$_2$FC**	***P* value**
MAM3	A03_022833784	Cab001420.1	AT5G23020.1	7.83	1.68E−14
CYP83A1	A04_006011661	Cab016602.1	AT4G13770.1	6.99	4.61E−15
UGT74C1	A05_007631928	Cab034792.1	AT2G31790.1	6.77	2.31E−17
IPMI2	C04_002750004	Bo4g018590.1	AT2G43100.1	4.93	3.08E−11
CYP83A1	C04_035540042	Bo4g130780.1	AT4G13770.1	6.35	2.43E−11

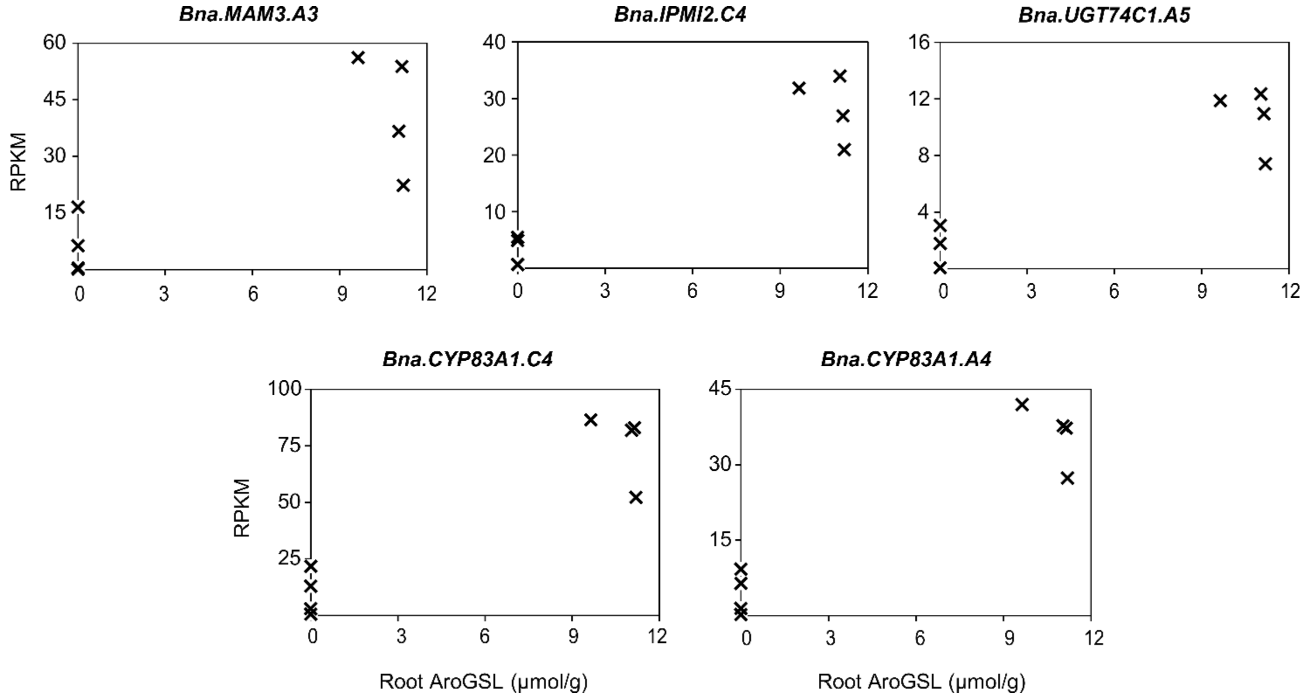


Figure 5-3. Correlation of root transcript abundance with levels of root aromatic glucosinolate. These genes, identified from the stringent root differential expression analysis, are known to be involved in the glucosinolate biosynthetic pathway. Average of four biological replicates are shown.

5.3.2 Weighted Gene Co-expression Network analysis (WGCNA)

To investigate the co-expression patterns between genes, a weighted gene co-expression network analysis (WGCNA) was performed using differentially expressed genes identified from the root DE analysis (Section 5.3.1). Since little information are available on the genes that are involved in root aromatic GSL trait, a hypothesis-driven approach has been taken to examine the genes that are highly connected to the candidate gene *Bna.HAG3.A3,* which is a known transcription factor. Genes that are regulated by *Bna.HAG3.A3* are expected to be co-expressed with the regulator based on the assumption that genes with strongly correlated expression are likely to be functionally associated.

Following the methods described in 5.2.3, a total of six co-expression modules (i.e. clusters of interconnected genes) have been generated. Number of genes per module ranges from 41 (red) to 214 (turquoise) with an average module size of 100 genes. Functional enrichment gene ontology (GO) analysis of the resulting modules indicate that the smallest 'red' module is biologically relevant and meaningful to the trait under investigation (Figure 5-4). The 'red' module contains 41 members in total and 12 genes that are known to be involved in GSL biosynthesis pathway. As a cluster, the representative gene expression profile of the 'red' module shows a significant positive correlation with root aromatic GSL content (r = 0.69; $p < 0.05$). Therefore, this module was chosen for further analysis.

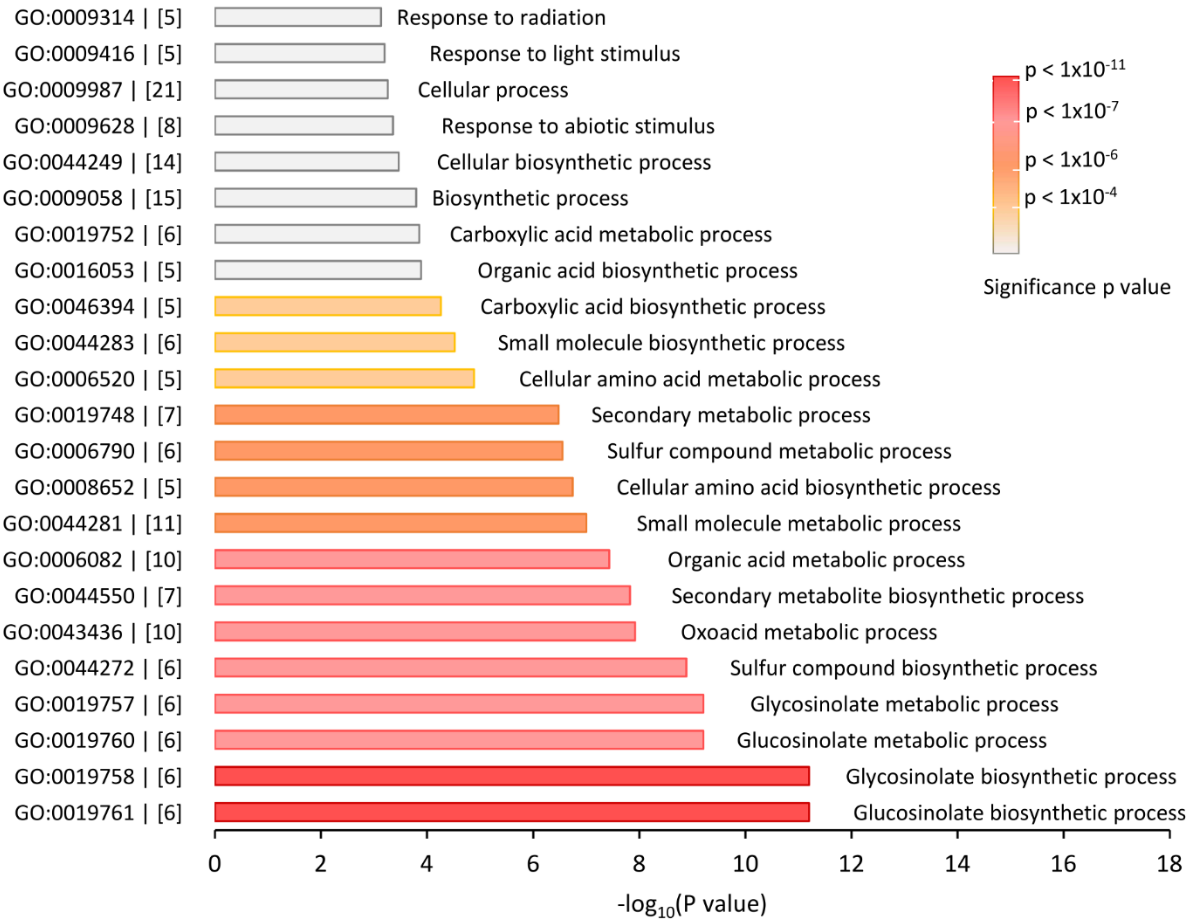


Figure 5-4. Functional enrichment Gene Ontology (GO) analysis of the 'red' module. The *P* values was calculated using FDR-corrected Fisher's exact test, against *Arabidopsis thaliana* genome locus 'TAIR10_2017' database. GO terms, input gene numbers in square brackets are displayed left of the bar chart and description of biological processes are displayed on the right.

The ‘red’ module shows that four genes connect to 17 nodes have the highest degree and nine genes connect to 16 other nodes have the second highest degree (Table 5-4) (Appendix 6 for detail parameters of all nodes). *Bna.HAG3.A3* (Cab015634.1) has shown high neighbourhood connectivity, high topological coefficient as well as being the only transcription factor out of all the nodes with high degree. This indicates *Bna.HAG3.A3* as a hub gene in the ‘red’ module (Table 5-4). As the hub gene with the second highest degree in the module, *Bna.HAG3.A3* (Cab015634.1) shares association with sixteen connected genes, including twelve involved in the aliphatic GSL pathway (Figure 5-5). Using Spearman’s correlation coefficient as the co-expression weight of *Bna.HAG3.A3* and other GSL genes, the expression patterns are highly correlated among the genes, ranging from 0.37 to 0.61 (Table 5-5). WGCNA shows that physically linked genes often appear in the same gene cluster as they share similar expression patterns. This may lead to false positive result. However, the genes connecting to *Bna.HAG3.A3* within this cluster are not linked as they are located on different chromosomes across the genome, suggesting that the findings are from true associations. Consistent with the findings from root DE analysis, the co-expression of aliphatic GSL genes with *Bna.HAG3.A3* indicates that part of the pathways are shared between aliphatic and aromatic GSLs.

As well as the interaction with the known GSL genes, Table 5-5 showed four genes outside of GSL biosynthetic pathway with high expression correlation to *Bna.HAG3.A3.* These are orthologues of aspartate kinase 3 (AT3G02020), myo-inositol-1-phosphate synthase 2 (AT2G22240) and nitrate transporter 2.5 (AT1G12940). Even though so far none of these genes has been reported to have any connection with GSL traits, it is worth noting for future studies.

Table 5-4. Parameters of the highest degree nodes in the red module.

Gene	Annotation	TAIR	Degree	Neighbourhood connectivity	Topological coefficient	GSL gene?
Bo4g018590.1	IPMI2	AT2G43100.1	17	14.7	0.74	✓
Cab034792.1	UGT74C1	AT2G31790.1	17	14.7	0.74	✓
Cab036831.1	IPMI2	AT2G43100.1	17	14.7	0.74	✓
Bo9g159950.1	MATE efflux	AT5G17700.1	17	12.1	0.67	
Cab015634.1	HAG3/MYB29	AT5G07690.1	16	15.3	0.90	✓
Bo4g130780.1	CYP83A1	AT4G13770.1	16	15.3	0.90	✓
Bo4g191120.1	CYP83A1	AT4G13770.1	16	15.3	0.90	✓
Bo8g101260.1	MIPS2	AT2G22240.2	16	15.3	0.90	
Bol004799	MAM3	AT5G23020.1	16	15.3	0.90	✓
Cab001421.1	MAM3	AT5G23020.1	16	15.3	0.90	✓
Cab016602.1	CYP83A1	AT4G13770.1	16	15.3	0.90	✓
Cab021103.1	SOT18	AT1G74090.1	16	15.3	0.90	✓
Bo3g122030.1	FAR5	AT3G44550.1	16	12.2	0.68	

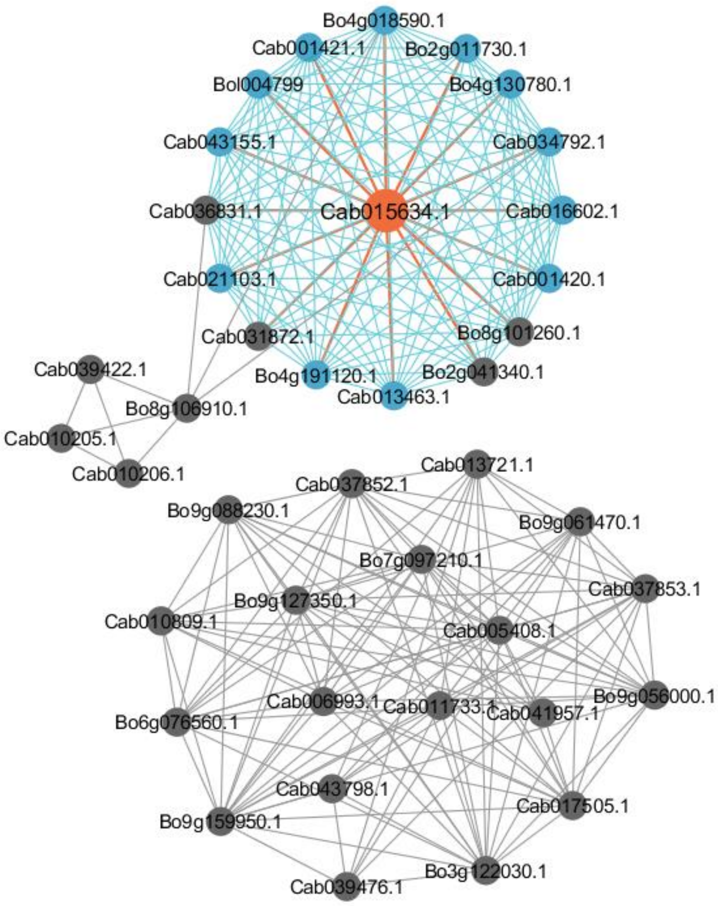


Figure 5-5. The network of genes with *Bna.HAG3.A3* as a hub gene, derived from the red module of the weighted gene co-expression network analysis. Each circular node represent a gene. The orange node represents Cab015634.1 *(Bna.HAG3.A3)*, the blue nodes are the genes that have been identified as GSL biosynthetic genes from *A. thaliana* studies, and grey nodes are other genes. The edges of the network are the lines connecting each nodes, which represents an association or relationship.

Table 5-5. Co-expression of genes connected to the *Bna.HAG3.A3* (Cab015634.1) hub gene. The weight of the co-expression is defined as Spearman's correlation coefficient. Genes involved in the glucosinolate biosynthetic pathway are highlighted in blue.

Node with direct connection to *Bna.HAG3.A3*			
***B. napus* gene**	***Arabidopsis* orthologue**	**Annotation**	**Weight**
Cab034792.1	AT2G31790.1	UGT74C1	0.605
Cab036831.1	AT2G43100.1	IPMI2	0.605
Cab016602.1	AT4G13770.1	CYP83A1	0.566
Bo4g018590.1	AT2G43100.1	IPMI2	0.559
Bo4g191120.1	AT4G13770.1	CYP83A1	0.546
Bo2g041340.1	AT3G02020.1	Aspartate kinase 3	0.517
Bo8g101260.1	AT2G22240.2	Myo-inositol-1-phosphate synthase 2	0.498
Bo4g130780.1	AT4G13770.1	CYP83A1	0.491
Cab001421.1	AT5G23020.1	MAM3	0.48
Bol004799	AT5G23020.1	MAM3	0.476
Cab013463.1	AT2G22240.2	Myo-inositol-1-phosphate synthase 2	0.466
Cab021103.1	AT1G74090.1	SOT18	0.462
Cab043155.1	AT4G13770.1	CYP83A1	0.452
Cab001420.1	AT5G23020.1	MAM3	0.404
Cab031872.1	AT1G12940.1	Nitrate transporter 2.5	0.401
Bo2g011730.1	AT5G14200.3	IMD1	0.372

5.4 Minor root glucosinolates

Indole GSLs are equally abundant in the root as aromatic GSL, however the genetic associations are weak as no clear SNP-based association peaks have been seen for these GSLs (Appendix 5). The GEM-based association analysis shows two markers above the Bonferroni's threshold for the root indole GSLs. These are: orthologues of an F-box domain-containing protein (AT1G13780) and an unknown protein (AT1G06980). The molecular function and their involvement with indole GSL metabolism are unknown. In *Arabidopsis* studies, *MYB34*, *MYB51* and *MYB122* had been reported as regulators of indole GSLs (Chapter 1, Section 1.4.1.2). However, none of the experiments carried out in this project have detected any associations of these MYB transcription factors with indole, aliphatic nor aromatic GSLs. At present, not enough information is available to identify candidate gene for the production of root indole GSLs.

SNP association analysis has revealed identical controlling loci on A2/C2 and A9/C9 for root aliphatic GSLs as in leaves, and *Bna.HAG1.A9* is also one of the top GEMs ($p = 2.10\times10^{-9}$) in the GEM-based association analysis (Appendix 5). Though aliphatic GSLs are only a minor class of GSLs in roots, AT is sensitive enough to detect the SNP associations of *Bna.HAG1.A9* and *Bna.HAG1.C2* with aliphatic GSL trait variations in *B. napus* roots.

5.5 Ratios of glucosinolates as trait in the AT analysis

To get further insight into the underexplored root GSLs, ratios of GSLs from total amount of root GSLs were used as traits for additional AT analyses (Appendix 7). The use of GSL ratios allows the analysis of loci controlling proportion of a class of GSL in relation to other structural classes. SNP associations shows the same peak on chromosome A3 for aromatic GSL ratio and *Bna.HAG1* loci (chromosomes A2/C2 and A9/C9) for aliphatic ratio, consistent with the results using absolute GSL concentration traits. More significantly, SNP association peaks for indole GSL ratio has revealed superimposed controlling loci for both aromatic and aliphatic GSLs (Figure 5-6). Interestingly, these common SNP markers within the shared associated loci between indole and the other two GSL ratios consistently showed opposing allelic effects with the GSL levels (Figure 5-7). For instance, an 'A' allele of a common SNP marker on chromosome A3 that is positively correlated to the levels of aromatic GSL is shown to have a negative correlation to indole GSLs. This opposite effect is also present between the aliphatic ratio and indole ratio common SNP markers, which suggests that these loci are likely to be involved in controlling the flow of GSL biosynthesis from one class to another. Since aromatic and indole GSLs are the major component in roots (Figure 3-2), SNP-based associations of indole and aromatic GSL ratio identify the associated region on chromosome A3 as the key locus in roots for controlling the amount of root aromatic GSLs being made while limiting the flow into indole GSL biosynthesis.

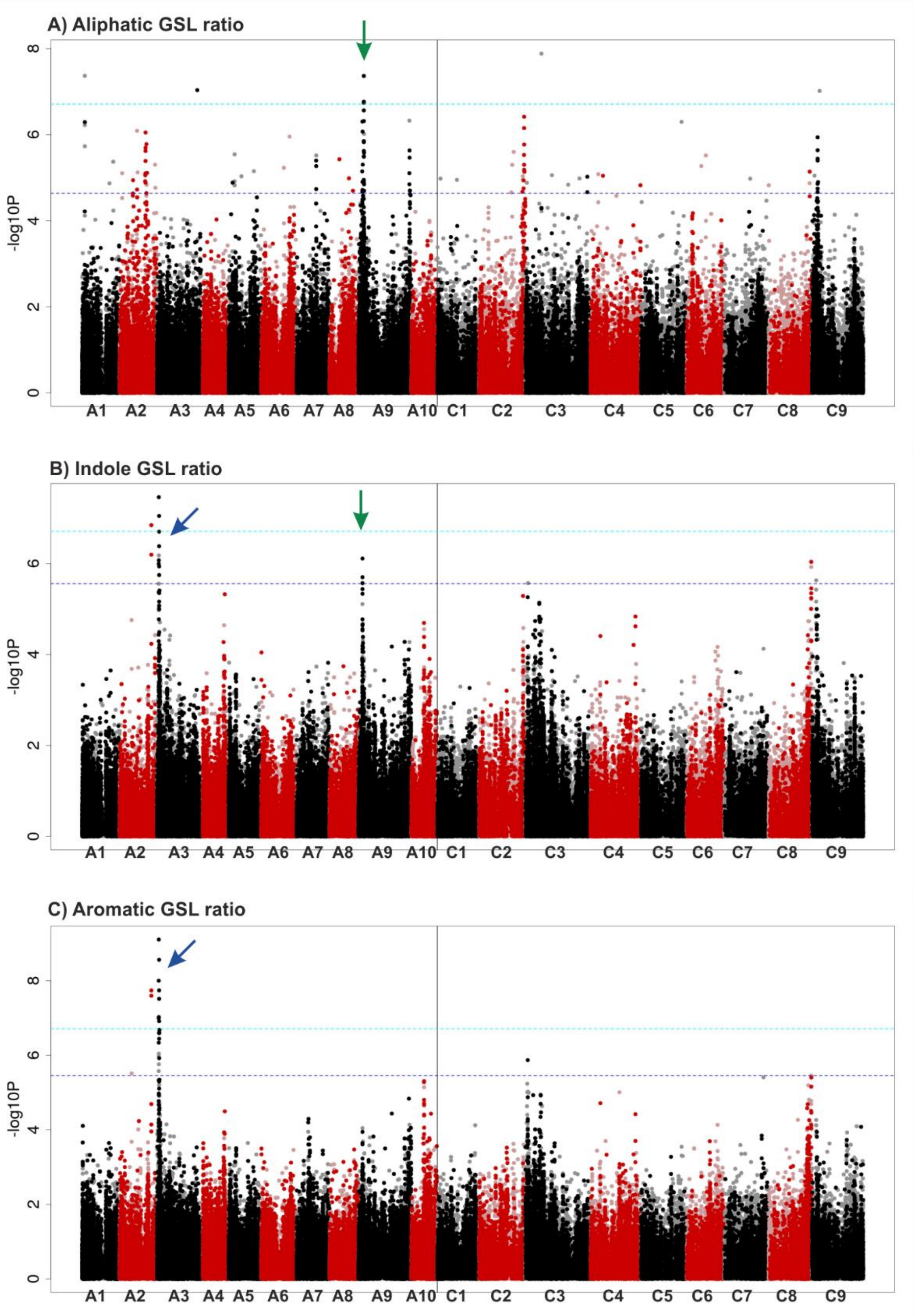


Figure 5-6. SNP association analysis of proportion of root glucosinolates (GSL) as ratio for 288 *B. napus* accessions. Ratio of root GSL were used as traits for **A)** root aliphatic GSL ratio, **B)** root indole GSL ratio and **C)** root aromatic GSL ratio. Root indole GSL revealed a common SNP association peak with aliphatic GSL ratio on chromosome A9 (green arrow) and a common SNP peak with aromatic GSL ratio on chromosome A3 (blue arrow).

A)

	Chrm A9	
Common SNP	Indole Ratio	Aliphatic Ratio
Cab038434.1:42:G	-0.06	0.02
Cab038449.1:1701:T	-0.07	0.02
Cab038449.1:1794:G	-0.07	0.02
Cab038449.1:1341:A	-0.07	0.02
Cab038387.1:540:G	-0.06	0.02
Cab038449.1:1392:A	-0.07	0.02
Cab038387.1:270:T	-0.06	0.02
Cab038387.1:675:A	0.06	-0.01
Cab038449.1:1116:T	0.07	-0.02
Cab038449.1:1320:G	0.08	-0.02
Cab038449.1:1359:G	0.07	-0.02
Cab038449.1:1215:G	0.06	-0.02
Cab038449.1:423:A	0.06	-0.02
Cab038449.1:162:A	0.06	-0.02
Cab038347.2:5100:G	0.07	-0.02
Cab038449.1:1710:A	0.07	-0.02
Cab038449.1:1203:T	0.08	-0.02
Cab038408.1:229:A	0.10	-0.04

B)

	Chrm A3	
Common SNP	Indole Ratio	Aromatic Ratio
Cab015572.1:276:C	-0.09	0.08
Cab015557.1:909:T	-0.08	0.08
Cab015566.1:171:G	-0.08	0.09
Cab015645.2:1213:T	-0.08	0.08
Cab015557.1:918:G	-0.08	0.08
Cab015608.1:807:G	-0.08	0.08
Cab015610.2:1461:T	-0.08	0.08
Cab015722.1:499:G	-0.07	0.07
Cab015658.1:351:A	-0.07	0.08
Cab015616.1:1056:C	-0.07	0.08
Cab015618.1:2691:A	-0.07	0.07
Cab015616.1:429:G	-0.07	0.08
Cab015566.1:165:T	-0.07	0.07
Cab015657.1:1044:C	-0.07	0.08
Cab015618.1:543:G	-0.07	0.08
Cab015614.1:894:A	-0.06	0.06
Cab015679.3:1263:G	0.06	-0.06
BnaA03g01910D:685:G	0.06	-0.07
Cab015572.1:393:C	0.06	-0.06
Cab015614.1:675:T	0.07	-0.07
Cab015616.1:285:C	0.07	-0.08
Cab015616.1:957:T	0.07	-0.07
BnaA03g01910D:1035:C	0.07	-0.07
Cab015723.1:479:A	0.07	-0.06
Cab015723.1:447:C	0.07	-0.06
Cab015614.1:744:T	0.07	-0.07
Cab015618.1:2469:C	0.07	-0.08
Cab015572.1:723:A	0.07	-0.08
Cab015610.2:1436:G	0.08	-0.08
Cab015557.1:825:C	0.08	-0.08
Cab015614.1:699:G	0.08	-0.08
BnaA03g01910D:1498:C	0.09	-0.11
Cab002611.1:2901:G	0.09	-0.08
Cab016333.3:1491:C	0.10	-0.08
Cab015540.1:1266:C	0.14	-0.13
Cab002705.1:630:C	0.17	-0.14

Allelic Effect
-0.20
-0.15
-0.10
-0.05
0
0.05
0.10
0.15
0.20

Figure 5-7. Allelic effect estimates of the common SNP markers within the associated loci between **A)** indole and aliphatic glucosinolate ratios on chromosome A9 and between **B)** indole and aromatic glucosinolate ratios on chromosome A3. The common SNP markers showed opposite allelic effect between the GSLs. Allelic effect estimates were calculated as part of the GAPIT analysis. A positive allelic effect estimate indicates that the allele is favourable over the other allele.

5.6 Chapter discussion

In this chapter, AT SNP-based association analysis has identified a highly associated controlling locus of root-dominant GST on chromosome A3 (Figure 5-1) corresponding to the location of an orthologue of *HAG3* (Spreadsheet 4 & Appendix 14 of Kittipol *et al*. 2019b). Compared to other homoeologous copies, *Bna.HAG3.A3* has the highest frequency of polymorphisms, particularly SNPs, including the ones that are associated with changes in the levels of GST in roots. Though polymorphisms associated with the leaf or root traits can be detected from the AT dataset from either tissues, GEM data from leaf RNA-seq cannot be used to assess root traits because of tissue-specific gene expression. For example, the expression of *HAG3* orthologues is detected strongly only in the roots but not in the leaves of 3-weeks old *B. napus* plants. The expression data from root RNA-seq show that the *Bna.HAG3.A3* gene has increased expression in high-root aromatic GSL lines but reduced expression in low-root aromatic GSL lines. The results presented in this chapter suggest that *Bna.HAG3.A3*, orthologue of a known regulator of aliphatic GSL in *A. thailana*, is a key regulator of root aromatic GSL natural variations in *B. napus*. It is possible that the *HAG3* orthologues do not share the same function across these two genomes. Mutation, gene duplication and polyploidisation events may result in a functional difference after the divergence of *A. thaliana* and *B. napus*. However, it is equally plausible that the *A. thaliana HAG3* has retained the regulatory role on aromatic GSLs but this function has been masked by the lack of aromatic GSL production in this species. To date, this is the first instance of reporting such a role of *HAG3* in regulating root aromatic GSL.

In addition to the discovery of a potential role of a known aliphatic GSL regulator in the regulation of root aromatic GSL, this study has uncovered several genes that are potentially important in the shared biosynthetic pathways between aliphatic and aromatic GSLs. The results from a combination of root DE and WGCN analyses have revealed the expression of four genes that are involved in the aliphatic GSL biosynthetic pathway are highly correlated to the variations in root aromatic GSL (Table 5-3) and co-expressed with the transcription factor *Bna.HAG3.A3* (Table 5-5). These four genes are: the orthologues of *CYP83A1* (AT4G13770), *UGT74C1* (AT2G31790), *IPMI2* (AT2G43100), and *MAM3*

(AT5G23020). Among them, the cytochrome P450 CYP83A1 has been shown to metabolise aliphatic oximes as well as aromatic oximes in Arabidopsis (Naur *et al*., 2003). Outside of this PhD project no data of *UGT74C1* and *IPMI2* activities in the aromatic GSL pathway has been reported. Through DE analysis *Bna.MAM3.A3* has shown the largest changes in their expression between accessions amongst these genes (Table 5-3). Roots of *B. napus* are dominated by GST, a chain-elongated homophenylalanine aromatic GSL. However, genes involved in the chain-elongation of phenylalanine are still unknown. Based on the findings from this work, *Bna.MAM3.A3,* previously known to be part of aliphatic pathway, has been proposed to also involve in the chain-elongation of phenylalanine in *B. napus*. This hypothesis is supported by the observation that MAM3 has a broad substrate specificity in addition to methionine-derived 2-oxoacids where MAM3 is able to form condensation reaction with phenylpyruvate leading to GST production (Textor *et al*., 2007). In a quantitative trait loci mapping study in Arabidopsis for aromatic GSL, the *GS-Elong* locus (comprising of *MAM1*, *MAM2* and *MAM3*) that controls total leaf aliphatic GSLs is also the major QTL for controlling phenylalanine elongation (Kliebenstein *et al.*, 2001a). This observation supports our hypothesis that chain elongation of methionine-derived aliphatic GSLs and phenylalanine-derived aromatic GSLs could be controlled via the same gene.

CHAPTER 6

Relationships between glucosinolate content of vegetative tissues and seeds in *B. napus*

This chapter explores the relationship of GSL composition in the seeds and vegetative tissues (Section 6.3) and elucidates the underlying basis of low GSL cultivars in *B. napus* (6.3.1).

6.1 Introduction

The economic value of *B. napus* is in its seeds. Accumulation of goitrogenic GSLs such as progoitrin reduces the value of *B. napus* seeds, which is undesirable to growers. This is because the hydrolysis of *β*-hydroxyalkenyl GSLs (e.g. progoitrin and gluconapoleiferin) in the seed meal produces toxic products that are detrimental to animals and thus prevents their uses as animal feed (Mawson *et al*., 1993). Consequently, extensive breeding were carried out to select for oilseed rape cultivars with low seed GSLs, leading to the establishment of 'canola' cultivars. CanOLA (Canadian Oil Low Acid) or 'double-zero' are terms used to described rapeseed cultivars with low contents of both seed erucic acid (<2% in oil) and low seed GSLs (<30 µmol/g) (CanolaCouncil, 2017). Despite the commercial success of these double-zero cultivars being grown worldwide, the molecular mechanism underlying the low GSL trait in *B. napus* is unclear.

Transport of GSLs into seeds are thought to occur because *A. thaliana* seeds lack the *in situ* capability for the core biosynthesis steps of the aliphatic GSLs (Nour-Eldin and Halkier, 2009) even though seeds of *A. thaliana* and *B. napus* accumulate aliphatic GSLs to high concentrations (Brown *et al*., 2003; Velasco *et al*., 2008). Past studies have shown no correlation between the low GSL content in the seeds and the GSL content in the leaves of *B. napus* low seed GSL canola cultivars (Porter *et al*., 1991; Fieldsen and Milford, 1994), leading to the assumption that inhibition of GSL transport processes could have given rise to the low-seed GSL trait in these cultivars. This idea is supported by the finding of the nitrate/peptide transporter family in *A. thaliana*, GTR1 and GTR2, being highly specific to transporting GSLs (Nour-Eldin *et al*., 2012). *A. thaliana gtr1 gtr2* double mutant plants almost abolish the total amount of GSL in seeds while increase accumulation in leaves and silique walls (Nour-Eldin *et al*., 2012). Regardless of the difficulty in translating the loss of function into the *Brassica* polyploidy genome, a successful knockout of four out of twelve *B. juncea* GTR orthologues has resulted in a reduction of seed GSL content by 60% (down to approx. 43 µmol/g of seed GSL) but unaltered GSL levels in leaves, stem and roots (Nour-Eldin *et al.*, 2017). Though this reduction of the seed GSL contents is not at the same level as the low seed GSL trait in *B. napus* canola cultivars (<30 µmol/g), the finding supports the idea that the low seed GSL traits of *Brassica* oilseed crops could be derived from mutation in the coding sequences of GTR orthologues.

Since specific accumulation pattern of GSL compounds in different parts of the plant may be established by *in situ* biosynthesis and/or long-distance transport, this chapter aims to gain more understanding of the roles of these two mechanisms by investigating the relationship of GSLs in vegetative tissues and seeds. Furthermore, understanding the genetic control underlying the low GSL traits may provide important targets for modulating precise GSL content in *B. napus* crops.

6.2 Methods

6.2.1 Leaf-seed glucosinolate relationship

To explore the GSL profile relationship between leaf and seed, fifteen accessions representing a subset of the 288 accessions were randomly selected to include at least one of each crop types using random number generator to avoid sample selection bias (Table 6-1). Four biological replicates of each accessions were grown, following the same growth condition described in Section 2.1. GSL compositions were measured from the leaves and seeds. The content of GSLs of leaf samples were quantified as described in Section 2.2. Apart from some minor differences in the sample preparation and extraction procedure described below, purification and HPLC analysis of the seed GSLs followed the same procedure as leaf GSLs.

Table 6-1. Fifteen accessions selected for the analysis of leaf-seed GSL relationship.

Accession	Cultivar	Crop type
a-0000036	Lisek	Modern winter OSR
a-0000056	Musette	Modern winter OSR
a-0000084	NK Nemax	Modern winter OSR
a-0000085	NK Passion	Modern winter OSR
a-0000169	Sarepta	Winter OSR
a-0000172	Slovenska Krajova	Winter OSR
a-0000176	Trebicska	Winter OSR
a-0000179	Wolynski	Winter OSR
a-0000186	MOANA	Winter fodder
a-0000189	FORA	Winter fodder
a-0000234	Zhouyou	Semiwinter OSR
a-0000236	STELLAR DH	Spring OSR
a-0000238	YUDAL	Spring OSR
a-0000421	Scotia	Swede
a-0000424	Tina	Swede

6.2.1.1 Sample preparation and extraction of glucosinolates from seeds

Unlike the leaf and root tissues, seeds do not need to be freeze-dried. Seed batches from single plants were weighed to 50 mg. For each seed batch, the seeds were submerged in 500 µL of 80% (v/v) methanol as extracting solvent, and 25 µl of 5 mM glucotropaeolin were added to the mixture as the internal standard. During homogenisation, methanol was added to prevent myrosinase from breaking down GSL in the sample. The lids of Eppendorf tubes were tightly sealed to prevent liquid spillage and seed samples were homogenised to fine powder with one steel bead for 10 min at a frequency of 25 Hz (TissueLyser II, Qiagen). Once homogenised, a further 1475 µL of 80% (v/v) methanol was added and mixed. The samples were left to stand for 30 min at 20 °C and further mixed with orbital shaker (Vibrax, IKA) at 1200 rpm for 30 min before centrifugation at 8000 rpm for 10 min. Supernatant methanol extracts were then transferred to the pre-conditioned Sephadex columns in the purification step, following the methods in Section 2.2.2.

6.2.1.2 Three-dimensional scatter plot

The pattern of distribution between three variables, i.e. levels of GSL in leaves and seeds as well as the expression level of the functional *HAG1* orthologues, was illustrated with an `R` package called '`plot3D`'. Levels of leaf aliphatic GSL were plotted on the *x*-axis, levels of seed aliphatic GSL were plotted on the *y*-axis and on the *z*-axis were the RPKM values of the *Bna.HAG1.A9* and *Bna.HAG1.C2* for each accessions.

Results

6.3 Relationships of glucosinolate composition in the seeds and vegetative tissues

In order to explore the relationship of GSLs between vegetative tissues and seeds, Spearman's correlation analysis was extended to include the seed GSL data from Lu *et al.* (2014) in addition to the leaf and root data collected from this study (Table 6-2). Of all GSL structural classes, aliphatic GSLs exhibit the strongest correlation between tissues, in particular between leaves and the other two tissues (Leaf-Root: r = 0.69, *p* ≤0.001; Leaf-Seed: r = 0.54, *p* ≤0.001; Seed-Root: r = 0.44, *p* ≤0.001). These significant positive correlations indicate that the natural variations observed in aliphatic GSLs between the tissues are likely to be regulated by long-distance transport or a master regulator of the aliphatic biosynthetic pathway that controls the biosynthesis of aliphatic GSLs in all of these tissues.

Table 6-2. Spearman's correlation coefficient analysis of glucosinolate traits in vegetative tissues and seeds.

	TL	**L-ali**	**L-ind**	**L-aro**	**TR**	**R-ali**	**R-ind**	**R-aro**
Total Leaf (TL)	–							
Leaf Aliphatic (L-ali)	0.91***	–						
Leaf Indole (L-ind)	0.45***	0.14*	–					
Leaf Aromatic (L-aro)	0.63***	0.62***	0.13*	–				
Total Root (TR)	0.28***	0.30***	0.00	0.37***	–			
Root Aliphatic (R-ali)	0.65***	0.69***	0.09	0.50***	0.42***	–		
Root Indole (R-ind)	0.00	–0.10	0.25***	–0.15*	0.41***	–0.04	–	
Root Aromatic (R-aro)	0.18**	0.29***	–0.21***	0.46***	0.75***	0.30***	–0.18**	–
† Total Seed GSL	0.48***	0.54***	0.00	0.40***	0.02	0.44***	–0.20*	0.09

Correlation of mean trait values from 288 accessions of the diversity panel. Significant correlations are indicated; ****p* ≤0.001, ***p* ≤0.01, **p* ≤0.05. † Data for total seed glucosinolates for 151 *B. napus* accessions came from Lu *et al.*(2014).

6.3.1 Basis of aliphatic GSL variations between tissues

To investigate whether it is the variations in transport or in biosynthesis processes that explain the natural variations of aliphatic GSL patterns between the leaves and seeds in *B. napus*, additional seed data have been analysed for associations with the orthologues of Arabidopsis GSL transporters, *GTR1* (AT3G47960) and *GTR2* (AT5G62680). In the *B. napus* genome, four orthologues of *GTR1* (on chromosome C3 and A6) and five orthologues of *GTR2* (on chromosome C3, C9, A6 and A9) have been identified. None of these copies show SNP nor GEM associations above false-discovery rate threshold with seed GSL or leaf or root aliphatic GSL contents (Table 6-3 and Table 6-4). Although *Bna.GTR2.A9* and *Bna.GTR2.C9* is in the region of the genome that are close to the SNP association peaks on chromosome A9 and C9, these loci does not correspond to the *GTR2* control because no significant GEM association has been found and no correlation between gene expression and aliphatic contents has been observed across all three tissues for the two genes (Figure 6-1). On the other hand, comparison of the AT plots for total seed GSLs, leaf and root aliphatic GSLs has shown that they all share the four common association peaks on chromosome A2, A9, C2 and C9 corresponding to the positions of the *HAG1* orthologue-containing control loci (Figure 6-2).

Based on the correlation of seed data from Lu *et al* (2014), total seed GSLs have shown a strong relationship with total leaf GSLs but not with total root GSL (Table 6-2). To explore the strength of this correlation of GSL profiles between leaves and seeds, a subset of *B. napus* accessions were grown and the leaf and seed GSLs were measured (list of accessions on Table 6-1). Both the leaf tissues and seeds are predominated with aliphatic GSLs that makes up the total GSL concentration (83.9% and 98.7% of all GSLs in leaves and seeds are aliphatic respectively), while the other two GSL classes are in very small amounts (Figure 6-3). Comparison between aliphatic GSLs in leaf tissues and seeds has revealed a significant positive relationship (r = 0.63, p ≤0.05). A pattern of two distinct classes is shown: either one with relatively high or the other with relatively low GSL contents in both tissues (Figure 6-3). The absence of any accessions having high GSL contents in leaves and low in seeds indicates that the basis of aliphatic GSL variations between plant tissues is from the amount synthesised, as controlled by orthologues of *HAG1*, and not by variations in the transport processes. In addition, a positive pattern has been observed in the correlation plots with three variables. This pattern shows the levels of *HAG1* expression correspond to the amounts of GSLs both in leaves and seeds (Figure 6-4). Thus, this observation further supports the conclusion that variations in the expression of *Bna.HAG1.A9* and *Bna.HAG1.C2* shape the variations of the aliphatic GSL levels between these two tissues in *B. napus*.

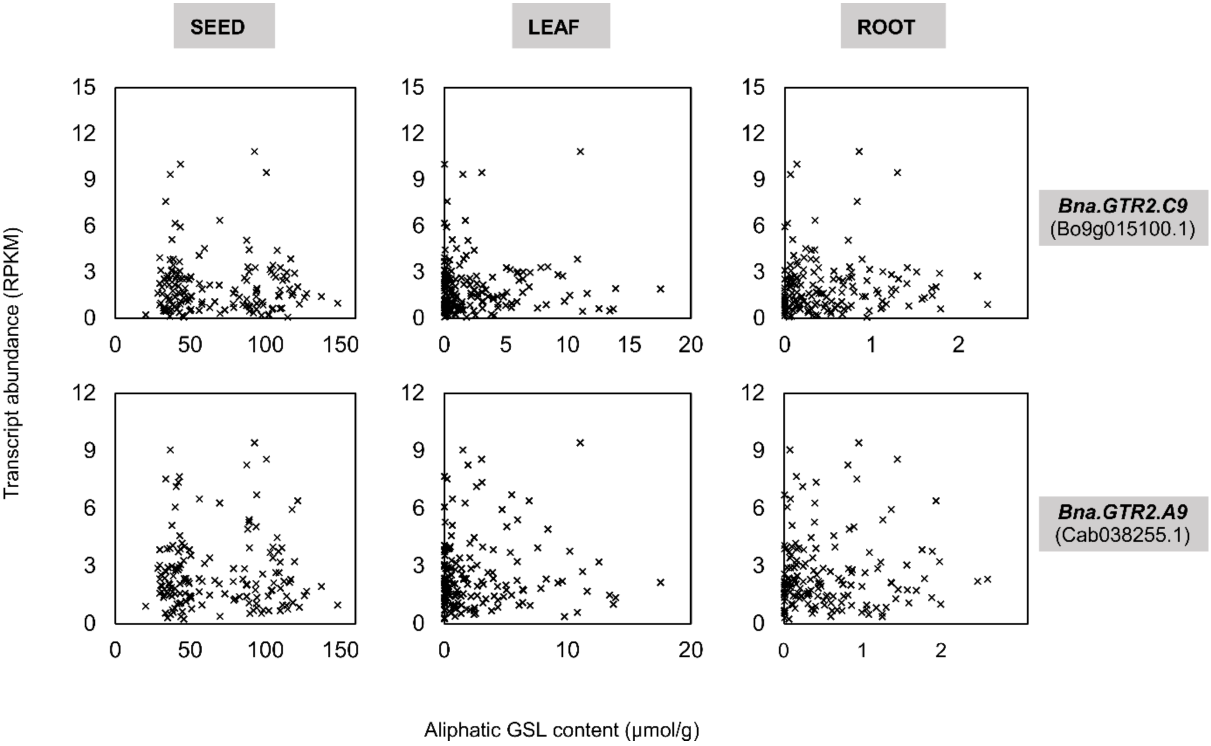


Figure 6-1. Correlation of glucosinolate transporters, *Bna.GTR2.A9* and *Bna.GTR2.C9*, transcript abundance with levels of aliphatic glucosinolates in leaves, roots and seeds. Transcript abundance was quantified and normalised as reads per kb per million aligned reads (RPKM). No correlation between gene expressions and changes in levels of aliphatic GSL was observed across all three tissues.

Table 6-3. GEM association of glucosinolate transporter 1, *GTR1* (AT3G47960), with aliphatic glucosinolate traits. No association was observed.

Trait Source	Gene	Chrm	R^2	$-\log_{10}P$	*P* value
Leaf	Cab005960.1	A06	0.021	0.853	0.028
	Cab028807.1	A06	0.006	0.096	0.708
	Bo3g113800.1	C03	0.007	0.099	0.699
	Bo3g137030.1	C03	0.006	0.019	0.936
Root	Cab005960.1	A06	0.009	0.286	0.286
	Cab028807.1	A06	0.006	0.016	0.942
	Bo3g113800.1	C03	0.009	0.184	0.461
	Bo3g137030.1	C03	0.005	0.032	0.882
Seed	Cab028807.1	A06	0.021	0.677	0.050
	Cab005960.1	A06	0.019	0.579	0.080
	Bo3g113800.1	C03	0.006	0.135	0.591
	Bo3g137030.1	C03	0.004	0.068	0.774

Table 6-4. GEM association of glucosinolate transporter 2, *GTR2* (AT5G62680), with aliphatic glucosinolate traits. No association was observed.

Trait Source	Gene	Chrm	R^2	$-\log_{10}P$	*P* value
Leaf	Cab006180.1	A06	0.019	0.916	0.021
	Cab038255.1	A09	0.016	0.607	0.084
	Bo3g107270.1	C03	0.007	0.315	0.295
	Bo9g015100.1	C09	0.013	0.403	0.204
Root	Cab006180.1	A06	0.019	0.803	0.020
	Cab038255.1	A09	0.015	0.504	0.097
	Bo3g107270.1	C03	0.007	0.028	0.897
	Bo9g015100.1	C09	0.008	0.228	0.377
Seed	Cab006180.1	A06	0.021	0.677	0.050
	Cab038255.1	A09	0.003	0.062	0.792
	Bo3g107270.1	C03	0.003	0.071	0.764
	Bo9g015100.1	C09	0.005	0.104	0.673

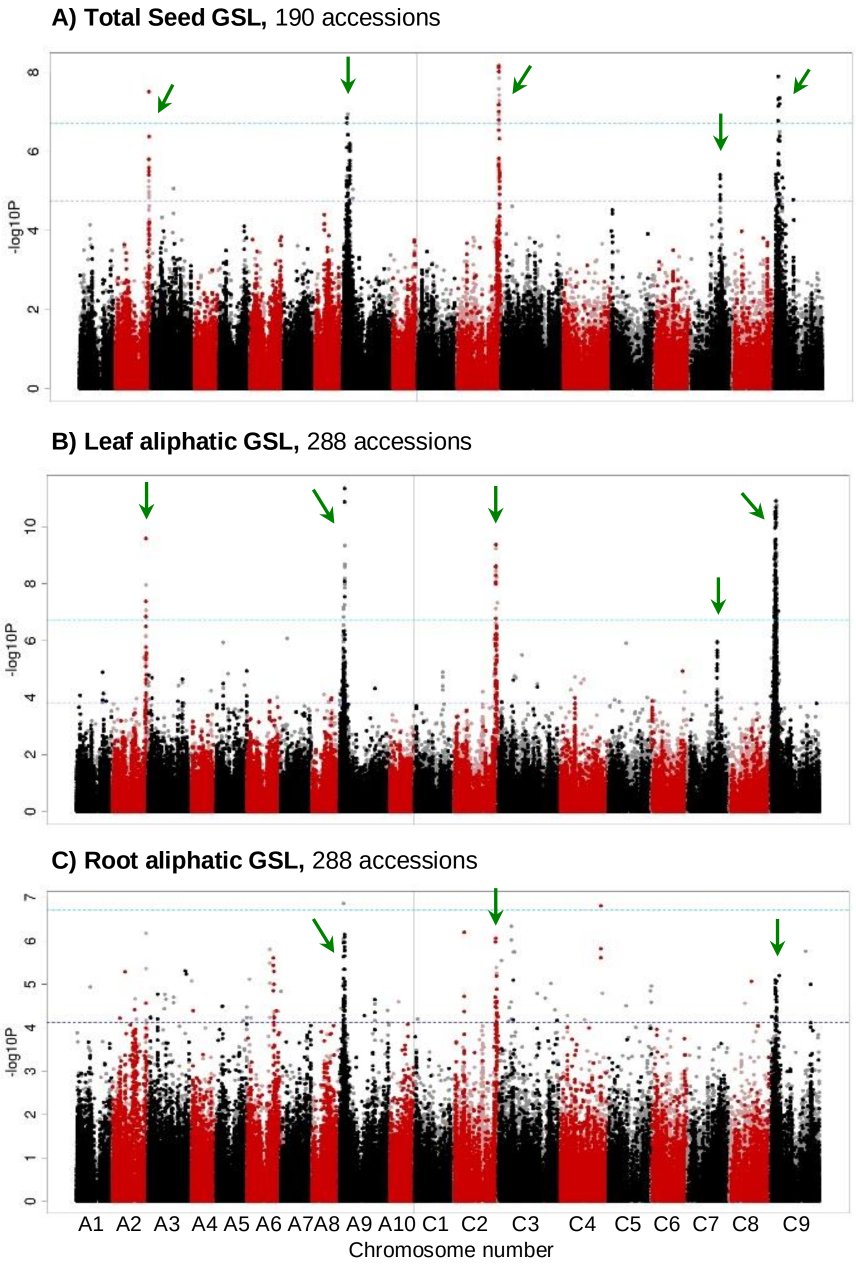


Figure 6-2. Comparison of the aliphatic GSL SNP Manhattan plots between **A)** Total seed GSLs (data from Lu *et al*, 2014), **B)** Leaf aliphatic GSLs and **C)** root aliphatic GSLs. Common SNP association peaks are marked with arrows, these loci are corresponded to the *HAG1* loci.

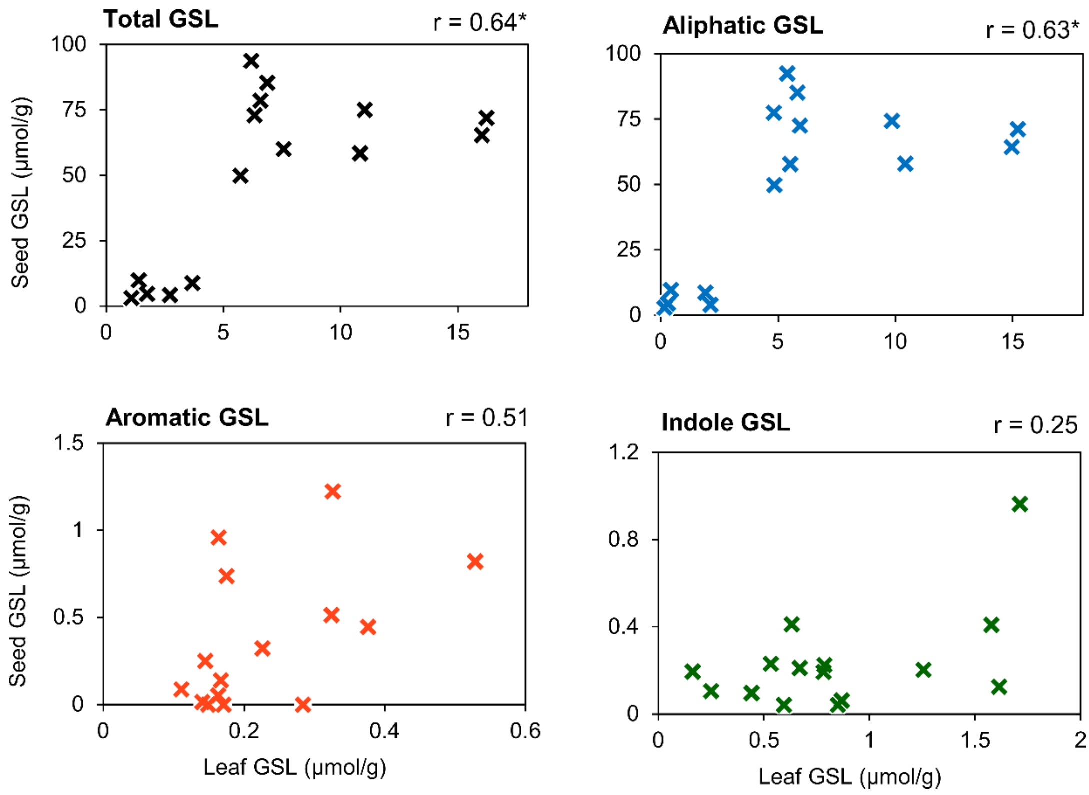


Figure 6-3. Correlation of glucosinolates in the leaf and seed tissues from 15 *B. napus* accessions. Mean concentration of four biological replicates are shown. Spearman's correlation coefficient are shown for each glucosinolate class. Significant correlations are indicated; ****p* ≤0.001, ***p* ≤0.01, **p* ≤0.05.

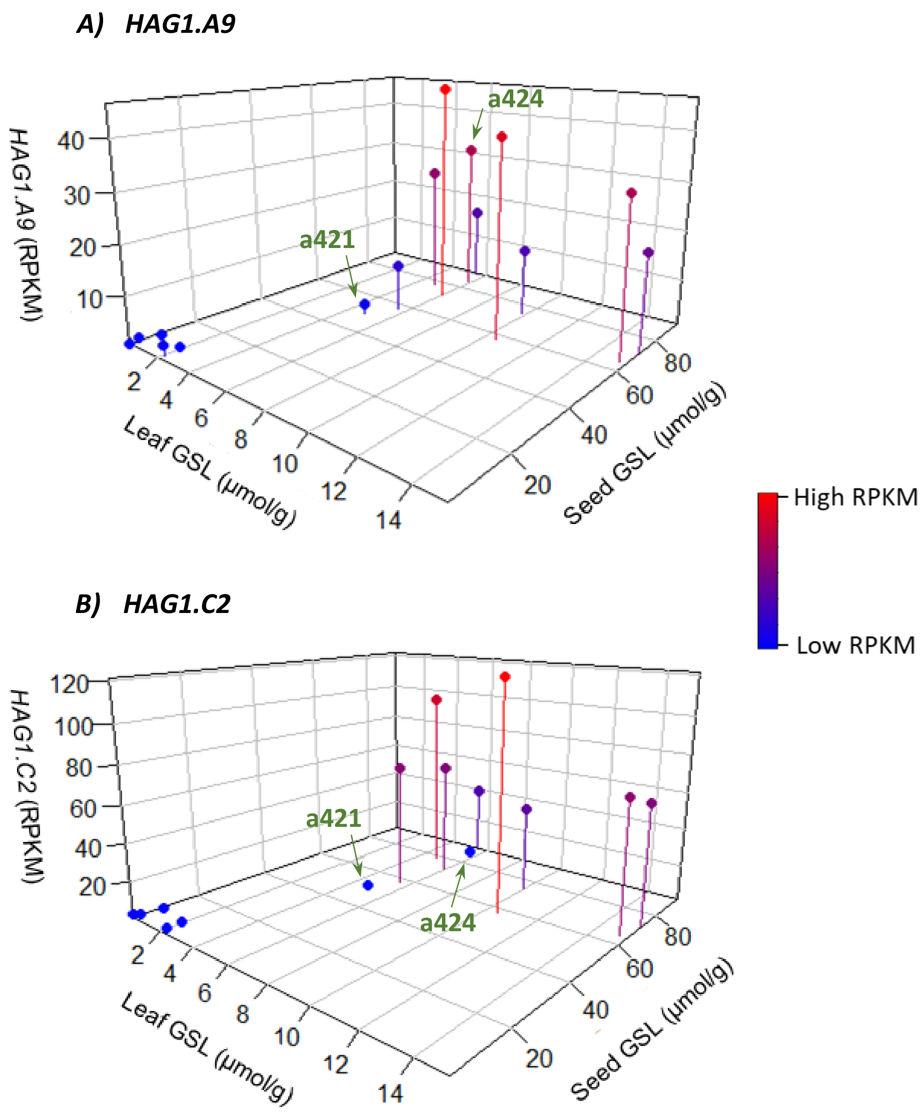


Figure 6-4. Three-dimensional correlation of levels of leaf aliphatic GSL, seed aliphatic and expression of orthologues of *HAG1*. Based on the same data as Figure 6-3, the leaf aliphatic GSL levels are plotted on the *x*-coordinate and the seed aliphatic GSLs levels are plotted on the *y*-coordinate. The expressions of A) *HAG1.A9* and B) *HAG1.C2*, derived from leaf transcriptome RPKM, are plotted on the vertical *z*-coordinate respectively. The vertical lines help to distinguish the coordinate of the points in a 3D space and the colour gradient help to distinguish the level of *HAG1* expression on the *z*-axis. The two labelled accessions, a421 and a424, will be discussed in Chapter 7 (7.2).

6.4 Discussion

The identification of Bronowski, the low-GSL Polish spring rape cultivar of *B. napus*, in the 1970s has provided the genetic source for all commercial *B. napus* cultivars of low seed GSLs through selective breeding (Rosa *et al.*, 1997). This reduction in seed GSLs is due almost entirely to the reduction in aliphatic GSL levels in Bronowski (Kondra and Stefansson, 1970; Rucker and Rudloff, 1991). However, past studies reported no significant correlation of GSL levels between seeds and leaves in canola cultivars, *B. napus* with low seed GSLs (Porter *et al*., 1991; Fieldsen and Milford, 1994), leading to the assumption that inhibition of GSL transport processes could have given rise to the low-seed GSL trait in *B. napus*. This hypothesis was supported by a report on the role of controlling GSL accumulation in *A. thaliana* seeds by GTR1 and GTR2, two members of the nitrate/peptide transporter family (Nour-Eldin *et al*., 2012). Though orthologues of *GTR2* are found in close proximity in the genomic region to the causative loci controlling low-seed GSL trait in *B. napus* (Lu *et al*., 2014), data from this chapter have shown no accessions exhibit GSL profiles that could have arisen from the inhibition of GSL transporters. Such accessions would be expected to have high levels of GSLs in leaves but low in seeds. Furthermore, no SNP or GEM associations of *GTR1* or *GTR2* orthologues with leaf aliphatic or total seed GSL traits have been found. On the other hand, data analysis reveals significant positive correlation between seed and leaf GSLs. As such, the seed GSL profile is a good reflection of GSLs in the leaf tissues (Table 6-2 and Figure 6-3). Previous work in *A. thaliana* has shown a similar positive correlation with the level of aliphatic GSLs in the leaves representing the minimal concentration of aliphatic seed GSL assuming there is no variation in GSL transport from the leaves to the seeds (Kliebenstein *et al.*, 2001b).

Aliphatic GSLs are the most abundant class of GSLs in both the *B. napus* leaf tissues and seeds, therefore it is not surprising that the same associated loci have been detected for total seed GSL (Harper *et al*., 2012) and for total leaf GSL (Appendix 3 & Kittipol *et al.*, 2019b). Data analysis has also shown that the genetic variations for the reduced GSL level in seed, which are also reflected in the reduced GSL level in leaf, are due to structural changes in the region of *B. napus* genome containing *Bna.HAG1.A9* and *Bna.HAG1.C2*, the transcription factors that regulates the aliphatic GSL biosynthetic, as a result of the selective breeding practice. The data from this thesis is consistent with the study by Chalhoub *et al.* (Chalhoub *et al.*, 2014) where they found the loss of *HAG1* orthologues on chromosome A9 and C2 but no sequence changes in *GTR1* and *GTR2* orthologues in the genome of the low GSL *B. napus* cultivar 'Darmor-bzh'.

CHAPTER 7

Final discussion and future directions

This chapter concludes with the discussion of the results (Section 7.1). It addresses how evolution and breeding impact the variations in the population structure of *B. napus* and how this links to the variations in GSL concentrations in the modern varieties. Further to this, Section 7.1.1 examines the interactions between transcriptional regulators that shape the GSL profiles in *B. napus*. Then, evaluation and implications of the study is reviewed in Section 7.2 and Section 7.3, respectively. Afterwards, Section 7.4 examines the directions of future work. Finally, the key findings of this work is summarised in Section 7.5.

7.1 Discussion

As a principal oilseed crop, *Brassica napus* is the third most profitable crop in the UK (£643m) spanning over 600 thousand hectares of land across the nation (DEFRA, 2018). Its commercial viability and environmental footprint has led to an increasing demand to develop beneficial traits in terms of economic and environmental sustainability in this rapeseed crop. With this aim, a potential avenue to explore is to harness the defensive properties of GSLs to improve pest resistance and biofumigation properties in the vegetative tissues while maintaining the requirement of low GSLs in the seeds. To achieve such aim, it is necessary to have more genetic research carried out within *B. napus* rather than to infer the genetic basis of traits from the studies of other diploid plant species such as *A. thaliana*, the model plant for genetic and biochemical research. There are many great advantages of working with *A. thaliana*, for example, in the identification of biosynthetic genes and regulators of aliphatic and indole GSL pathways (Wittstock and Halkier, 2002; Grubb and Abel, 2006; Halkier and Gershenzon, 2006; Sonderby *et al*., 2010). However, differences in traits exist between the species and certain traits, such as aromatic GSL, cannot be investigated in *A. thaliana* Columbia. The works in this thesis aims to increase the understanding of the genetic controls underlying the natural variations of the GSL classes that are dominant in the leaves and roots of *B. napus*, as well as to develop the knowledge of their connection with seed GSLs.

Phenotypic analysis of GSL profiles (Chapter 3) reveals that the total levels of GSL in the leaf of *B. napus* are largely determined by the variation in aliphatic GSL, whereas the total levels of GSL in the root are largely determined by the variation in aromatic GSL (Table 3-3). These variations in the GSL traits are highly heritable and are attributed to the allelic differences of the 288 accessions in this germplasm set (Table 3-2). Analysis of the allelic diversity of this panel has shown two differentiated subpopulations with each representing the winter or the spring crop types respectively (Figure 3-2A to 3-2C). This finding of the large genetic distance between these two crop types has been described previously by Hasan *et al.* (2006) and Bus *et al.* (2011). The distinct allelic clustering of winter and spring crop types could be explained by the adaptation of crop types to their distinct environment and by their breeding history (Bus *et al*., 2011). The winter cultivars have been selected to adapt to the temperate climate of Western Europe and require vernalisation for flowering, while the spring cultivars have low winter hardiness and do not need vernalisation to flower. The spring cultivars are generally grown in regions with short summer growing season like Canada or where winter does not become cold enough to induce flowering like Australia. Despite the genetic distance between these two crop types, the levels of GSL concentrations do not differ significantly in the leaf and root tissues between spring and winter types (Figure 3-8). Findings from Chapter 4 and Chapter 5 show that the levels of aliphatic and aromatic GSLs are controlled by specific and simple genetic loci. This could explain how the large genetically distant spring and winter crop types show similar GSL phenotypes. Selection for low seed GSL in the past has resulted in a preference for specific loci and alleles inherited from the ancestral cultivar (cv Bronowski); these loci are still identifiable in the genome of the modern oilseed rape cultivars (Bancroft *et al.*, 2011). The reduction in total oilseed GSLs is due almost entirely to the reduction in aliphatic GSLs in Bronowski (Kondra and Stefansson, 1970; Rucker and Rudloff, 1991), which reflected the reduced levels of GSLs in the leaves (Chapter 6). Indeed, results from Chapter 4 and Chapter 6 show that the genetic variation for the reduced GSL level in seeds and leaves is due to the homoeologous exchanges of specific blocks in the regions containing *Bna.HAG1.A9* and *Bna.HAG1.C2*, the two key regulators that control the natural variations of aliphatic GSLs. Results also show that seeds and roots have different GSL profiles, with aliphatic class being the dominant type in leaves and aromatic class dominant in roots, and these two GSL classes are regulated by different loci. While aliphatic GSLs are controlled by orthologues of *HAG1*, works from Chapter 5 support the hypothesis that orthologue of *HAG3* on chromosome A3 controls the natural variation of aromatic GSL, particularly GST, which determined total GSL levels in *B. napus* roots. This suggests that selection for low seed GSL phenotypes is not likely to cause a major impact on root GSL levels.

7.1.1 Interactions between transcriptional regulators shape the GSL profiles in *B. napus*

The three GSL classes were previously thought to be independently biosynthesised and regulated by different sets of gene family that are highly substrate-specific (Kliebenstein *et al.*, 2001a). For instance, cytochrome P450 monooxygenases have shown selectivity for specific amino acid substrates (Wittstock and Halkier, 2000). Yet, several studies carried out in *A. thaliana* have shown feedback mechanisms and co-regulations between the pathways, especially between the aliphatic and indole GSL pathways via *HAG-MYBs* and *HIG-MYBs* (Gachon *et al.*, 2005; Frerigmann, 2016). In *B. napus* where all three GSL classes are prevalent, a bigger picture of the intricate relationship between different classes can be observed.

In this thesis, several significant relationships have been detected (Table 3-3) indicating a degree of co-regulation between the levels of different GSL classes. Noticeably, two types of relationships have been observed: the strong positive correlations between the aliphatic and aromatic GSLs, and the negative relationship of indole GSLs with either the aliphatic or aromatic types. Aliphatic is the dominant GSL type in the leaf tissues, thus it is possible that it may mask the extent of the relationships of the other two minor GSL classes in this tissue type. In contrast, analyses of the root GSLs can provide a clearer picture of the relationship between these GSL classes since no single GSL class dominates, with both aromatic and indole being the major GSLs (Figure 3-2).

Based on the AT data and correlation analyses, the significant positive correlation between aliphatic and aromatic GSLs suggests positive co-regulation between these two classes of GSLs. This co-regulation could be explained by the transactivation activity of *HAG1* on *HAG3* (Gigolashvili *et al.*, 2008; Sonderby *et al.*, 2010) and the biosynthetic gene *MAM3* which is shared between these two GSL classes (Kliebenstein *et al.*, 2001a). The distinct aliphatic-dominated leaf and aromatic-dominated root GSL profiles can be interpreted as the coordinated control of *HAG1* and *HAG3* orthologues, regulating the variation of aliphatic and aromatic GSL respectively. The activity of *HAG1* and *HAG3* orthologues are spatially separated as shown by their localised expression in separate tissues. For instance, high expression of *HAG1* orthologues in the leaves of high GSL-lines has been observed but minimal expression in the roots has been detected, whereas *HAG3* orthologues are respectively highly expressed in roots but negligible in leaves. Additionally, the significant negative correlation between aromatic and indole root GSLs could be explained by the negative regulation of *HAG3* orthologues on the indole GSL biosynthetic pathway which is consistent with data from previous work in *A. thaliana* (Gigolashvili *et al.*, 2008).

In *A. thaliana*, a feedback regulatory control is found between aliphatic and indole GSL biosynthesis (Levy *et al.*, 2005), leading to a hypothesis that a metabolic cross-talk mechanism must exist between the aromatic and indole branch of GSL biosynthesis. This cross-talk mechanism could also be present in *B. napus* to maintain GSL homeostasis, possibly via similar feedback regulatory control between aliphatic and indole GSL biosynthesis. AT analysis using indole GSL ratios as the phenotypic trait (Figure 5-6) reveals an opposite allelic effect of the common SNP markers in roots between aromatic and indole GSLs on the associated A3 locus (Figure 5-7), providing support for *Bna.HAG3.A3* as a key regulator for the amount of aromatic GSL while limiting the flow into indole GSL biosynthesis. The exact mechanism of how the pathway switches from one major root GSL type to another is unknown.

No significant associations with indole GSLs has been observed from the AT analysis. As a result, no candidate gene responsible for indole GSL variations has been identified. Works from this thesis did not detect any SNP nor GEM associations above the false-discovery threshold of any orthologues in *B. napus* of the previously known indole GSL regulators from *A. thaliana* studies (i.e. *MYB34/ATR1*, *MYB51/HIG1*, and *MYB122/HIG2*) responsible for the variations of indole GSLs. Data from this thesis show that indole GSLs are the most abundant class in *B. napus* roots (47.7% of all root GSLs are indole vs. 45.0% for aromatic GSL), however levels of indole GSLs are least correlated to the total amount of root GSLs (r = 0.41, p ≤0.001) compared with aromatic and aliphatic GSLs (aromatic: r = 0.75, p ≤0.001; aliphatic r = 0.44, p ≤0.001). This observation indicates that indole GSL variations in *B. napus* may not be controlled by simple loci like aliphatic and aromatic GSLs. The biosynthetic pathway of indole GSLs is known to be metabolically connected to the indole-3-acetic acid (IAA) [7] biosynthesis by their common metabolic intermediate indole-3-acetaldoxime (IAOx) (Hull *et al.*, 2000; Grubb and Abel, 2006). This connection is likely to add to the complexity of indole GSL regulation. Blocking the indole GSL pathway downstream of IAOx has resulted in an overflow of IAOx converting to IAA (Halkier and Gershenzon, 2006). Moreover, overexpression of *MYB34* and *MYB122* in *A. thaliana* elevated the production of both indole GSLs as well as IAA levels (Celenza, 2005; Gigolashvili *et al*., 2007a), indicating that the known indole GSL transcription factors could also regulate auxin homeostasis in *A. thaliana*. It is likely that there is a complex connection between the GSL pathways for plants to maintain a degree of GSL homeostasis, therefore controlling the natural variations of indole GSL may depends on the regulation of other GSL pathways as well as maintaining homeostasis with other metabolic pathway like IAA.

[7] IAA is the most common auxin class of plant hormone

7.2 Evaluation of the study

The quality of association studies depends on high density of molecular markers across the genome, high quality reference gene model, a large panel of accessions, good experimental designs and reliability of the trait data. The AT platform used in this thesis has 355,536 SNP markers, which is equivalent to one SNP in every 0.33 kb across the genome (Havlickova *et al.*, 2018). Compared to the 26,841 or 21,117 SNPs of the commercially available 60K Brassica Infinium® SNP array used in other *B. napus* GWAS studies (Li *et al.*, 2014; Xu *et al.*, 2015), this AT platform provides a much higher SNP density and increased the probability of identifying trait-marker associations in this thesis.

Apart from the higher density of molecular markers, the development of the pan-transcriptome resource has provided a more reliable reference sequence by incorporating the diploid *Brassica* A and C genomes, supporting the transcriptome-based technologies as well as a solution to determine genome-of-origin of any given gene in the *B. napus* genome (He *et al.*, 2015). This enables an insight into the structural changes and analysis into the functional interactions between *B. napus* AC genomes. Such advantage has been demonstrated in previous studies (He *et al.*, 2016; Havlickova *et al.*, 2018; Kittipol *et al.*, 2019a) as well as in this thesis by the detection of homoeologous exchanges within the regions containing functional orthologues of *HAG1*, which impact the variations of aliphatic GSLs (Section 4.2.3).

In addition, the 288 accessions panel used in this thesis is a subset of the 383 RIPR panel (Havlickova *et al.*, 2018). The panel is made up of genetically diverse *B. napus* accessions and large enough to provide sufficient association power. The association power of the RIPR panel and the AT platform has been demonstrated in identifying genes controlling various traits in *B. napus*, such as GSL variations in this thesis (Kittipol *et al*., 2019a), erucic acid (Havlickova *et al.*, 2018), lodging resistance (Miller *et al.*, 2018), and plant nutrients (Alcock *et al.*, 2017; Alcock *et al.*, 2018).

The discovery of GEM markers derived from the juvenile leaf transcriptome dataset can be limited by their transcription in different tissues or developmental stages. For instance, Chapter 5 (Figure 5-1B) has shown the limitation of identifying GEM associations for genes with root-specific expression patterns. Nevertheless, genes associated with traits establishing at different tissues or developmental stages can be identified from leaf transcriptome through SNP-based association, as described in Chapter 5 (Figure 5-1A) and in previous AT studies (Lu *et al*., 2014; Havlickova *et al*., 2018; Kittipol *et al*., 2019a). This is possible because the high density of markers (one SNP per 0.33kb) that provides sufficient coverage to detect variations of SNP markers in LD with the causative gene, resulting in associated region containing the control gene.

With regards to the reliability of the GSL data, good laboratory practices have been deployed throughout the process to ensure high quality trait data. A panel of 288 accessions of *B. napus* were grown in a controlled environment free of pest or pathogen infestation, therefore the levels of GSL extracted from the samples should reflect the natural variation of GSLs found in these tissues. With over 1,152 plants to sample (four biological replicates per accession) and a total of 2,304 samples to process (one leaf and root sample per plant), it is critical for the plants to be harvested at the same growth stage with minimal GSL degradation. Although ideally the harvest should be done in one day, harvesting have been achieved within in a realistically short time window of five consecutive days. Plant age also affects GSL levels as total amount of GSLs decreases with the age. To minimise the effect of potential changes in GSL levels across accessions during harvesting period, one replicate for each accession was harvested at a time in batches, flash-frozen with liquid nitrogen and stored at -80 °C. All analyses have been completed on the same platform for all 288 accessions, i.e. using the same HPLC, column and freeze-dryer. This is to minimise sample variations that could be introduced by these factors during data processing. These quality operating procedures ensure the generation of a robust GSL dataset (Kittipol *et al.*, 2019b) that is reliable and valuable for researcher in this field as well as for the agriculture community.

The scope of this association study is to explore the regulatory control of GSL variations at the population level. While aliphatic GSL variations in the majority of the accessions can be explained by the variations of *Bna.HAG1*, some anomalies has been observed at the individual level. For example, Figure 6-4 shows that the expression level of *Bna.HAG1.A9* and *Bna.HAG1.C2* does not follow the expected trend for two data points ('a421' and 'a424') of the accessions with high concentrations of aliphatic GSL in the leaf and seeds. The GSL traits for these accessions were re-measured and found to be consistent and reliable as the variation of GSL levels between the biological replicates are within the normal standard deviation range. Closer inspection of these accessions has revealed that both are of swede crop type. Swede cultivars have been found to be phenotypically and genetically quite distinctive from other crop types. As a group, swedes display significantly higher concentration of aliphatic and indole GSLs in both leaf and root tissues as well as significantly reduced aromatic GSL in the roots compared to other crop types (Figure 3-9). At the genetic level, Bus *et al.* (2011) reported swede cultivars to be genetically diverse from the spring and winter oilseed rapes based on the principal coordinate analysis of 89 primer combinations for loci distributed evenly across *B. napus* genome. This thesis also reveals a distinctive structural variation in the swede genome. A large region on chromosome C2, which contains the functional *HAG1* orthologue *Bna.HAG1.C2,* is lost in all swede crop type including the high-GSL cultivars (Figure 4-2B) but present in all other crop types with high GSL phenotypes. This indicates that the other *HAG1* orthologue *Bna.HAG1.A9* alone is sufficient in controlling aliphatic GSL variations in swede, and this may thus explain the anomaly behaviour of data point 'a424' on Figure 6-4.

On the other hand, variations in the expression of *HAG1* orthologues cannot explain the accumulation of aliphatic GSLs of data point 'a421' on Figure 6-4. The unusual pattern in this individual may be explained by the low expression of myrosinase (AT5G26000) orthologues. Two copies of myrosinase genes, Cab040081.1 and Cab040082.1, have been found in close proximity to the associated region on chromosome A9. Across the panel, the mean transcript abundance of Cab040081.1 and Cab040082.1 are 1008 and 1005 RPKM, respectively. However, in accession a-000421 (also known as 'a421') these myrosinase gene copies have RPKM values at 184 and 95 respectively, a much lower transcript abundance to the other accessions. Since the levels of GSLs in both leaf and seed tissues may also be determined by the regulation of both biosynthesis and degradation, compromising degradation may also lead to higher accumulation of GSLs. Therefore, lower transcript abundance of the GSL hydrolytic enzymes, myrosinases, provides a plausible explanation of the high GSL traits of data point ‘a421’. Nevertheless, this observation is speculative. Further investigation is required to explore the role of myrosinase and how it may be involved in establishing GSL variations when plants are not under attack.

7.3 Implications

There are implications for understanding the modular genetic system that regulates GSL natural variations in *B. napus* as a whole. This knowledge could be used for crop improvement by exploiting GSL potentials and manipulating GSL profiles for modulation of interactions between important crop plants and its pests. Since *Bna.HAG1.A9* and *Bna.HAG1.C2* are the main transcription factors controlling the production of aliphatic GSLs in both leaves and seeds, loss of function mutation in these genes is likely to reduce the GSL content in both tissues. Such scenario has been shown in a study that RNA interference was used to downregulate *BjMYB28* in *B. juncea*. The highest levels of a knockdown has a significant 89% reduction in seed GSL content, and leaf GSLs also show 80-90% reduction in content (Augustine *et al*., 2013). Low seed GSLs are desirable for oilseed rape quality as it improves nutritional qualities of its protein meal, however low levels of GSLs in leaves would make the plant more vulnerable to pests and could lead to increased pesticide usage. Therefore, alternative approaches may be required to achieve low seed GSL content but high aliphatic GSL content in vegetative tissues. For example, this may be achieved via a blockage of GSL transport into seeds. This transport engineering concept has been demonstrated in a closely related allotetraploid *B. juncea* by Nour-Eldin *et al*. (2017). In their study, they have successful downregulated four out of twelve GTRs, GSL transporter orthologues and resulted in a reduction of seed GSL content by 60% but unaltered GSL levels in leaves, stem and roots. Similar transport engineering approach has not yet been achieved in *B. napus* and this represents an opportunity to be explored.

In roots, there is a simple genetic basis for the variation of root aromatic GSL content that does not influence the variations of seed GSL contents. A single locus on chromosome A3 containing *Bna.HAG3.A3* reported in this thesis has been shown to be associated with root aromatic GSL traits (Kittipol *et al.*, 2019a), which is consistent with the single locus encoding 2-phenylethyl aromatic GSL root trait reported in Potter *et al*. (2000). Increased production of aromatic GSL in roots for pest resistant and biofumigation potential can contribute to the crop's attractiveness. Breeding experiment by Potter *et al*. (2000) has shown that it is possible to selectively breed plant for containing higher levels of 2-phenylethyl GSL in root without changing GSL levels in shoot or seed. The findings from this thesis provide the first underlying genetic basis of regulating root aromatic GSLs in *B. napus* and lay a foundation in which variation on below-ground plant-pest interactions can now be explored.

7.4 Future directions

Through this study, the ability of AT in identifying potential candidate gene for aromatic GSL trait has been demonstrated and known transcription factors for aliphatic GSL control predicted from *A. thaliana* studies have been confirmed in *B. napus*. Association study offers the opportunity to identify genes that contribute to variations in quantitative traits, but it relies entirely on statistical associations. Therefore, functional validation of candidate genes via molecular complementation is still required to make stronger claims about the exact molecular mechanism of these functions. One of the fastest and common methods used to validate potential role of a given gene in *B. napus* is through Arabidopsis T-DNA insertion mutation lines (Alonso *et al*., 2003). Arabidopsis is a closely related species to *B. napus* and provides a much simpler and fully sequenced genome to study the roles of a gene of interest. For example, *MYB28/HAG1* has been functionally characterised from T-DNA *A. thaliana myb28* for their regulation of aliphatic GSLs (Gigolashvili *et al*., 2007b; Hirai *et al*., 2007; Sonderby *et al*., 2010).

However, not all oilseed rape genes can be functionally assessed with this approach especially when Arabidopsis lacks the ability to produce equivalent products. This is evident in the case of the molecular function of *MYB29/HAG3* orthologue for controlling aromatic GSLs. This function cannot be assessed using T-DNA mutants because *A. thaliana* Columbia does not produce this type of GSLs (Brown *et al*., 2003). Therefore, an alternative approach to study these gene functions, such as the role of *Bna.HAG3.A3* in aromatic GSL control, is to generate mutant lines via mutagenesis directly in *B. napus* despite its longer generation time, and assess the resulting phenotypes following the alteration of the target genes. With the recent establishment of radiation mutagenesis lines in *B. napus* within the Bancroft lab, there is a potential for future study to identify *B. napus* mutants that carry deletion and/or duplication in *Bna.HAG3.A3* and assess how this may affect the aromatic GSL phenotype. Unlike mutagenesis by chemical mutagens such as ethyl methanesulfonate (EMS), radiation-mediated mutagenesis induces a wide range of mutations at higher frequency, including small insert-deletions, large-scale deletions and segmental duplications (Bolon *et al*., 2014; Li *et al*., 2016). This radiation panel of mutagenised lines, which is an undergoing development at present, will provide a platform of functional validation for candidate genes, where candidate genes identified from *B. napus* association studies can be validated. These lines can in turn be used to assist mutation breeding for crop improvement.

The scope of this thesis is on the overarching regulations of GSL structural classes, i.e. orthologue of *HAG1* and *HAG3* as the master regulator of aliphatic and aromatic GSLs, respectively. Future studies could be fine-tuned to investigate the mechanism by which individual GSLs are preferentially regulated. For example, there is a desire to reduce the goitrogenic GSL derivative oxazolidine-2-thionine from progoitrin in *B. napus* seed meal, while enriching sulforaphane isothiocynate from glucoraphanin which is a potent anti-carcinogen in humans (Liu *et al*., 2012) and highly toxic to fungal pathogen growth in plants (Mithen *et al*., 1986). Such profile has been attempted by downregulating *GSL-ALK* downstream of glucoraphanin, which prevents the conversion of methylsulfinyl GSL (glucoraphanin) to alkenyl GSL (gluconapin) that is a precursor of progoitrin (Liu *et al*., 2012). This has resulted in a reduction of progoitrin and accumulation of glucoraphanin in *B. napus* seeds. From this work and others, aliphatic GSLs are known as the most diverse class of GSLs in *B. napus* (Velasco *et al.*, 2008; Kittipol *et al.*, 2019b), with nine out of the fourteen different structures being identified as aliphatic in this thesis. As different GSLs form different final hydrolytic products that are responsible for various biological functions, understanding how individual aliphatic GSLs are controlled will enable fine-tuning of GSL profiles and maximising the desirable biological properties of GSLs.

7.5 Key Findings

The followings are the key findings from the present study:

- Leaves and roots of *B. napus* are comprised of different GSL profiles. In the leaves, aliphatic GSLs are predominant and largely determine the total level of GSLs. While both aromatic and indole GSLs are predominant in the roots, levels of aromatic GSLs largely influence the total root GSLs.

- The genetic variations for the GSL levels in leaves are due to the homoeologous exchanges in the region of the *B. napus* genome containing *Bna.HAG1.A9* and *Bna.HAG.C2*. In low-GSL accessions, the functional *HAG1* orthologues on chromosome A9 and C2 are replaced by the non-functional orthologues in a homoeologous substitution.

- Results from both the SNP association and root DE analysis indicate *Bna.HAG3.A3*, an orthologue of a known regulator of aliphatic GSL in *A. thaliana,* as the key regulator for the natural variations of root aromatic GSLs.

- Aliphatic GSL contents in seeds and roots reflect those in the leaves. The positive correlation of aliphatic GSLs between seeds and leaves is due to the amount synthesised, as controlled by *Bna.HAG1.A9 and Bna.HAG1.C2*, rather than by variations in the transport processes.

- Two notable relationships between different classes of GSLs has been observed. First, there is a positive relationship between aliphatic and aromatic classes, which can be explained by the transactivation activity of *HAG1* on *HAG3* and the shared biosynthetic genes such as *MAM3* between these two GSL classes. Second, there is a negative relationship between indole and aromatic GSLs which can be due to a metabolic cross-talk between pathways to maintain GSL homeostasis, regulated by *Bna.HAG3.A3*.

Appendices

Appendix 1. Mean quantity of glucosinolate compositions in the leaves of 288 *B. napus* accessions (µmol/g). For individual GSL data, please see Spreadsheet 1 in Accompanying Material or Appendix 1 of Kittipol *et al*., 2019b.

York accession_id	Cultivar name	Crop type	Leaf_Ali	Leaf_Aro	Leaf_Ind	Leaf_Total	Leaf_Total StdDev
a-0000003	Robust	WOSR	0.18	0.00	1.28	1.45	0.75
a-0000004	Alaska	WOSR	0.48	0.00	3.68	4.16	2.17
a-0000005	Pirola	WOSR	0.00	0.00	0.50	0.50	0.59
a-0000008	Allure	WOSR	1.45	0.04	2.60	4.09	3.98
a-0000009	Agalon	WOSR	0.09	0.00	1.81	1.90	1.17
a-0000014	Rodeo	WOSR	2.99	0.03	1.79	4.82	3.34
a-0000018	Montego	WOSR	0.23	0.00	0.78	1.01	0.21
a-0000020	Pacific	WOSR	2.61	0.06	1.46	4.12	4.57
a-0000022	Missouri	WOSR	0.75	0.09	1.35	2.18	0.83
a-0000023	Manitoba	WOSR	2.76	0.08	1.20	4.05	4.98
a-0000024	Ladoga	WOSR	1.98	0.03	1.64	3.64	3.48
a-0000025	Atlantic	WOSR	1.59	0.04	0.78	2.40	3.82
a-0000026	Cooper	WOSR	0.00	0.00	1.61	1.61	0.89
a-0000027	Licapo	WOSR	1.34	0.12	0.87	2.33	3.03
a-0000028	Capitol	WOSR	2.06	0.04	1.28	3.38	4.01
a-0000029	Idol	WOSR	0.22	0.00	0.72	0.94	0.91
a-0000030	Vivol	WOSR	0.87	0.00	1.60	2.46	1.60
a-0000031	BRISTOL	WOSR	1.00	0.04	1.18	2.22	1.69
a-0000033	Lisabeth	WOSR	0.37	0.00	0.96	1.33	0.24
a-0000034	Lipid	WOSR	0.91	0.06	0.63	1.60	1.99
a-0000036	Lisek	WOSR	0.00	0.00	2.95	2.95	0.48
a-0000037	Contact	WOSR	10.77	0.40	1.59	12.75	7.89
a-0000038	Lion	WOSR	1.52	0.00	2.17	3.70	3.39
a-0000040	Apex	WOSR	0.03	0.00	0.91	0.95	0.91
a-0000042	Magnum	WOSR	1.10	0.03	0.97	2.10	2.37
a-0000044	Laser	WOSR	0.03	0.05	0.40	0.48	0.22
a-0000045	Fortis	WOSR	0.06	0.00	1.33	1.39	1.36
a-0000048	NK Bravour	WOSR	0.00	0.00	2.42	2.42	0.95
a-0000049	NK Fair	WOSR	0.88	0.04	1.51	2.42	1.52
a-0000050	Aviso	WOSR	0.31	0.04	1.07	1.42	1.26
a-0000053	Verona	WOSR	0.58	0.00	2.92	3.50	1.70
a-0000054	Tenor	WOSR	0.21	0.08	2.70	2.99	0.87
a-0000055	Expert	WOSR	3.14	0.06	0.66	3.86	4.73
a-0000056	Musette	WOSR	0.00	0.00	1.33	1.33	0.85
a-0000057	Kvintett	WOSR	1.91	0.04	1.83	3.78	3.63
a-0000058	Falstaff	WOSR	0.61	0.00	2.78	3.39	2.60
a-0000059	SW Sinatra	WOSR	0.22	0.00	0.73	0.95	0.40

Appendix 1 (continued)

York accession_id	Cultivar name	Crop type	Leaf_Ali	Leaf_Aro	Leaf_Ind	Leaf_Total	Leaf_Total StdDev
a-0000060	Viking	WOSR	0.18	0.00	0.59	0.77	0.59
a-0000062	Aurum	WOSR	2.49	0.04	0.36	2.89	4.33
a-0000065	Rasmus	WOSR	1.67	0.04	0.49	2.20	3.09
a-0000066	Gefion	WOSR	3.09	0.15	1.08	4.32	5.17
a-0000067	Nugget	WOSR	0.43	0.02	1.15	1.61	0.57
a-0000068	Zephir	WOSR	0.03	0.01	1.76	1.80	0.72
a-0000069	SLM 0413	WOSR	0.52	0.07	1.25	1.84	1.63
a-0000071	LSF 0519	WOSR	0.65	0.03	0.75	1.43	1.67
a-0000072	Beluga	WOSR	3.94	0.23	0.80	4.97	4.84
a-0000073	Amor	WOSR	0.28	0.03	0.59	0.91	0.30
a-0000075	Campari	WOSR	0.17	0.19	2.17	2.53	1.38
a-0000077	Duell	WOSR	0.38	0.02	2.70	3.10	1.95
a-0000079	Jessica	WOSR	3.03	0.05	0.88	3.96	5.49
a-0000080	Orlando	WOSR	0.39	0.04	0.47	0.90	0.55
a-0000081	Pollen	WOSR	2.12	0.17	0.62	2.90	0.75
a-0000082	Prince	WOSR	0.44	0.00	1.17	1.61	0.91
a-0000083	Wotan	WOSR	0.29	0.00	2.24	2.53	0.70
a-0000084	NK Nemax	WOSR	0.00	0.00	3.31	3.31	1.18
a-0000085	NK Passion	WOSR	0.65	0.05	0.80	1.50	1.29
a-0000090	APEX-93_5 X GINYOU_3 DH LINE	WOSR	5.73	0.19	0.78	6.70	2.66
a-0000093	CANBERRA x COURAGE DH LINE	WOSR	0.08	0.00	0.28	0.35	0.11
a-0000096	HANSEN X GASPARD DH LINE	WOSR	4.06	0.15	0.64	4.86	2.36
a-0000097	MADRIGAL x RECITAL DH LINE	WOSR	2.72	0.22	1.52	4.45	3.05
a-0000099	TAPIDOR DH	WOSR	1.97	0.10	1.33	3.40	2.35
a-0000101	EUROL	WOSR	2.52	0.40	0.84	3.76	2.26
a-0000105	LICROWN X EXPRESS DH LINE	WOSR	0.88	0.05	1.28	2.22	2.60
a-0000106	SHANNON x WINNER DH LINE	WOSR	0.91	0.08	1.45	2.44	1.76
a-0000107	JANETZKIS SCHLESISCHER	WOSR	7.91	1.01	1.28	10.20	1.51
a-0000110	OLIMPIADE	WOSR	1.56	0.06	1.12	2.74	2.51
a-0000113	Samourai	WOSR	0.83	0.15	1.07	2.05	1.23
a-0000114	Sollux	WOSR	7.12	0.66	1.43	9.20	9.71
a-0000115	Akela	WOSR	2.93	0.27	0.94	4.15	3.55
a-0000117	Maplus	WOSR	0.97	0.34	1.89	3.20	2.08
a-0000118	Askari	WOSR	6.30	0.20	2.66	9.16	3.14
a-0000119	Lirabon	WOSR	4.57	0.28	1.58	6.43	2.84
a-0000121	JetNeuf	WOSR	4.95	0.50	1.01	6.47	3.02
a-0000122	Cobra	WOSR	12.29	0.41	1.31	14.01	3.91
a-0000123	Falcon	WOSR	1.03	0.16	1.34	2.53	2.15
a-0000124	Mohican	WOSR	1.76	0.03	0.97	2.77	2.25

Appendix 1 (continued)

York accession_id	Cultivar name	Crop type	Leaf_Ali	Leaf_Aro	Leaf_Ind	Leaf_Total	Leaf_Total StdDev
a-0000125	Flip	WOSR	0.55	0.09	1.66	2.30	1.22
a-0000127	Phil	WOSR	3.04	0.11	1.66	4.81	3.10
a-0000128	Leopard	WOSR	0.26	0.00	0.70	0.96	0.02
a-0000129	RESYN-H048	WOSR	4.97	0.13	1.51	6.62	5.41
a-0000130	Resyn-G_ S4	WOSR	5.06	0.10	0.58	5.74	4.20
a-0000132	Anja	WOSR	2.42	0.26	1.50	4.18	2.95
a-0000133	Baltia	WOSR	1.51	0.10	2.25	3.85	1.53
a-0000135	Brink	WOSR	7.78	0.20	2.18	10.15	3.90
a-0000137	Coriander	WOSR	6.95	0.32	0.62	7.89	5.84
a-0000138	Diamant	WOSR	7.98	0.28	1.74	10.00	6.59
a-0000140	Doral	WOSR	5.94	0.23	1.61	7.78	3.70
a-0000141	Edita	WOSR	5.09	0.05	2.19	7.33	4.33
a-0000144	G™lzower Älquell	WOSR	5.78	0.15	1.52	7.45	1.23
a-0000146	Janpol	WOSR	4.99	0.02	0.49	5.50	0.46
a-0000148	Jupiter	WOSR	13.65	0.45	1.00	15.10	8.21
a-0000149	Krapphauser	WOSR	3.80	0.10	1.05	4.95	5.13
a-0000150	Kromerska	WOSR	0.91	0.07	0.67	1.64	1.09
a-0000151	Librador	WOSR	0.44	0.05	1.10	1.59	0.91
a-0000152	Libritta	WOSR	2.63	0.20	0.93	3.76	5.18
a-0000156	Lirakotta	WOSR	6.57	0.27	0.65	7.49	8.15
a-0000157	Madora	WOSR	5.90	0.15	0.92	6.97	4.20
a-0000162	Moldavia	WOSR	13.98	0.41	1.18	15.57	9.65
a-0000163	Mytnickij	WOSR	3.23	0.07	0.74	4.05	3.39
a-0000164	Nemertschanskij 1	WOSR	9.53	0.12	1.10	10.75	2.16
a-0000166	Panter	WOSR	5.73	0.35	0.71	6.79	5.08
a-0000168	Ramses	WOSR	13.18	0.15	1.08	14.41	10.67
a-0000169	Sarepta	WOSR	2.37	0.06	0.96	3.38	2.47
a-0000170	Skrzeszowicki	WOSR	3.42	0.11	0.47	4.00	3.87
a-0000171	Skziverskij	WOSR	10.99	0.19	1.50	12.67	2.84
a-0000172	Slovenska Krajova	WOSR	4.32	0.18	0.71	5.21	6.29
a-0000175	Start	WOSR	2.25	0.21	1.77	4.23	4.63
a-0000176	Trebicska	WOSR	2.40	0.07	0.55	3.01	3.37
a-0000178	Vinnickij 15/59	WOSR	2.49	0.15	0.62	3.26	2.85
a-0000179	Wolynski	WOSR	6.82	0.13	0.96	7.91	2.90
a-0000185	CANARD	Fodder	3.11	0.03	0.34	3.48	3.28
a-0000188	EMERALD	Fodder	0.33	0.04	0.29	0.66	0.57
a-0000189	FORA	Fodder	6.41	0.27	4.75	11.42	4.17
a-0000190	Aphid Resistant Rape	Fodder	4.23	0.03	2.26	6.52	7.66
a-0000193	Dwarf Essex	Fodder	3.11	0.10	0.83	4.05	3.35
a-0000200	Samo	Fodder	7.16	0.51	1.55	9.22	3.62
a-0000209	RAGGED JACK	Kale	1.00	0.28	1.25	2.53	2.02
a-0000210	RED RUSSIAN	Kale	0.76	0.08	1.01	1.85	0.70
a-0000211	SIBERISCHE BOERENKOOL	Kale	1.79	0.15	1.10	3.05	1.97

Appendix 1 (continued)

York accession_id	Cultivar name	Crop type	Leaf_Ali	Leaf_Aro	Leaf_Ind	Leaf_Total	Leaf_Total StdDev
a-0000226	SWU Chinese 1	Semiwinter OSR	1.59	0.24	0.83	2.65	2.05
a-0000227	SWU Chinese 2	Semiwinter OSR	0.14	0.02	1.03	1.18	0.97
a-0000228	SWU Chinese 3	Semiwinter OSR	0.27	0.08	0.80	1.14	0.73
a-0000232	SWU Chinese 8	Semiwinter OSR	0.28	0.05	1.38	1.71	0.48
a-0000233	SWU Chinese 9	Semiwinter OSR	2.93	0.18	1.32	4.43	2.62
a-0000234	Zhouyou	Semiwinter OSR	5.42	0.11	1.12	6.65	6.03
a-0000235	Drakkar	SpOSR	0.26	0.00	1.10	1.36	1.17
a-0000236	STELLAR DH	SpOSR	0.00	0.00	1.57	1.57	0.80
a-0000237	WESTAR DH	SpOSR	0.00	0.00	0.48	0.48	0.32
a-0000238	YUDAL	SpOSR	3.81	0.41	1.56	5.78	6.79
a-0000239	BRUTOR	SpOSR	0.35	0.02	0.22	0.59	0.25
a-0000241	COMET	SpOSR	0.00	0.01	0.71	0.72	0.96
a-0000242	CRESOR	SpOSR	7.49	0.33	0.70	8.52	6.61
a-0000247	INDUSTRY	SpOSR	0.82	0.08	1.87	2.77	2.39
a-0000248	KARAT	SpOSR	0.00	0.00	0.63	0.63	0.48
a-0000249	MARINKA	SpOSR	0.00	0.02	0.53	0.55	0.21
a-0000250	NIKLAS	SpOSR	1.85	0.07	2.47	4.39	2.53
a-0000251	TARGET	SpOSR	3.46	0.24	0.57	4.27	5.40
a-0000253	KAROO-057DH	SpOSR	0.00	0.28	2.76	3.05	2.02
a-0000254	MONTY-028DH	SpOSR	0.00	0.11	0.15	0.26	0.19
a-0000255	N01D-1330	SpOSR	0.04	0.07	0.81	0.92	0.57
a-0000256	N02D-1952	SpOSR	0.00	0.00	1.82	1.82	1.14
a-0000257	SURPASS400-024DH	SpOSR	0.05	0.00	3.03	3.08	0.74
a-0000258	CUBS ROOT	SpOSR	1.89	0.11	0.84	2.84	1.76
a-0000259	DUX	SpOSR	0.04	0.01	0.81	0.86	0.53
a-0000260	ERGLU	SpOSR	0.31	0.00	2.18	2.49	0.67
a-0000261	HELIOS	SpOSR	1.69	0.01	1.64	3.34	3.82
a-0000262	KROKO	SpOSR	2.67	0.02	0.42	3.11	2.42
a-0000264	LINETTA	SpOSR	8.12	0.57	0.72	9.42	13.37
a-0000265	MAZOWIECKI	SpOSR	6.39	0.34	3.11	9.84	9.23
a-0000267	WEIHENSTEPHANER	SpOSR	3.13	0.13	0.88	4.14	2.04
a-0000269	Alku	SpOSR	2.99	0.00	4.36	7.35	4.16
a-0000270	Bronowski	SpOSR	0.03	0.00	1.91	1.94	1.09
a-0000271	Ceska Krajova	SpOSR	0.63	0.02	0.59	1.24	0.93
a-0000272	Duplo	SpOSR	0.00	0.00	0.96	0.96	0.45
a-0000273	Janetzkis Sommerraps	SpOSR	4.01	0.10	1.96	6.08	4.17
a-0000274	Line	SpOSR	0.00	0.00	0.35	0.35	0.13
a-0000275	Marnoo	SpOSR	1.47	0.09	0.90	2.46	1.72

Appendix 1 (continued)

York accession_id	Cultivar name	Crop type	Leaf_Ali	Leaf_Aro	Leaf_Ind	Leaf_Total	Leaf_Total StdDev
a-0000276	Nugget	SpOSR	0.37	0.06	0.31	0.73	1.29
a-0000277	Olga	SpOSR	2.37	0.05	0.82	3.23	2.34
a-0000278	Spaeths Zollerngold	SpOSR	0.71	0.06	2.09	2.87	1.23
a-0000279	Sval_f⍰s Gulle	SpOSR	0.90	0.04	0.66	1.60	1.37
a-0000281	Tribute	SpOSR	0.00	0.00	0.96	0.96	1.24
a-0000282	Wesway	SpOSR	0.65	0.00	0.04	0.69	0.66
a-0000283	Fido	SpOSR	13.56	0.69	2.44	16.69	5.89
a-0000284	Oro	SpOSR	1.85	0.03	0.51	2.39	3.21
a-0000285	Tower	SpOSR	0.22	0.05	1.63	1.90	0.78
a-0000286	Regina II	SpOSR	1.58	0.09	1.34	3.01	2.32
a-0000288	Campino	SpOSR	2.31	0.06	1.26	3.63	3.72
a-0000289	Clipper	SpOSR	0.95	0.20	0.43	1.58	1.66
a-0000290	Larissa	SpOSR	0.53	0.08	0.77	1.38	0.92
a-0000291	Magma	SpOSR	0.05	0.00	0.75	0.80	0.20
a-0000292	Monty	SpOSR	0.00	0.20	0.10	0.31	0.29
a-0000294	Pauline	SpOSR	0.15	0.30	0.65	1.11	0.52
a-0000295	Sophia	SpOSR	0.07	0.01	1.25	1.34	0.78
a-0000296	Tribune	SpOSR	0.00	0.00	0.44	0.44	0.26
a-0000297	Trigold	SpOSR	0.00	0.02	0.47	0.49	0.12
a-0000298	Rivette	SpOSR	0.00	0.01	0.39	0.40	0.18
a-0000300	Adamo	SpOSR	3.02	0.10	0.56	3.68	1.55
a-0000301	Altex	SpOSR	0.14	0.01	0.50	0.64	0.47
a-0000302	Andor	SpOSR	0.00	0.00	0.42	0.42	0.08
a-0000303	Astor	SpOSR	0.00	0.00	0.42	0.42	0.26
a-0000305	Bingo	SpOSR	0.16	0.00	0.92	1.09	0.53
a-0000306	Callypso	SpOSR	0.00	0.00	0.77	0.77	0.30
a-0000307	Concord	SpOSR	1.33	0.00	1.62	2.95	3.79
a-0000308	Conny	SpOSR	1.22	0.08	1.29	2.59	0.10
a-0000309	Czyzowska	SpOSR	5.32	0.02	1.12	6.46	5.65
a-0000310	Daichousen (fuku)	SpOSR	3.64	0.26	1.80	5.70	3.82
a-0000311	Daichousen (mizuyasu)	SpOSR	0.06	0.18	0.21	0.45	0.30
a-0000312	Daichousen (nakano)	SpOSR	0.07	0.03	1.14	1.23	0.36
a-0000313	Erake	SpOSR	1.04	0.02	0.63	1.69	1.82
a-0000314	Furax	SpOSR	1.43	0.25	2.33	4.01	3.15
a-0000316	Galant	SpOSR	3.95	0.02	0.57	4.54	3.64
a-0000317	Giant Xr707	SpOSR	5.70	0.17	1.28	7.15	1.83
a-0000318	Gisora	SpOSR	0.00	0.00	0.72	0.72	0.44
a-0000321	Granit	SpOSR	2.62	0.08	1.05	3.75	2.63
a-0000323	Hankkija's Lauri	SpOSR	0.00	0.01	0.64	0.65	0.36
a-0000326	Kajsa	SpOSR	0.00	0.00	1.84	1.84	2.37
a-0000327	Korall	SpOSR	6.28	0.25	3.60	10.13	7.94
a-0000328	Korinth	SpOSR	10.00	0.28	0.56	10.83	7.08
a-0000329	Kosa	SpOSR	0.09	0.00	1.31	1.40	0.88

Appendix 1 (continued)

York accession_id	Cultivar name	Crop type	Leaf_Ali	Leaf_Aro	Leaf_Ind	Leaf_Total	Leaf_Total StdDev
a-0000330	Kruglik	SpOSR	0.00	0.00	0.73	0.73	0.67
a-0000332	Lirafox	SpOSR	0.10	0.00	0.94	1.05	0.65
a-0000333	Lirasol	SpOSR	0.15	0.01	0.71	0.88	0.61
a-0000334	Liraspa	SpOSR	0.93	0.01	0.68	1.62	0.94
a-0000335	Lirawell	SpOSR	0.03	0.00	0.67	0.69	0.30
a-0000336	Lisandra	SpOSR	0.66	0.00	2.23	2.89	2.26
a-0000337	Loras	SpOSR	0.00	0.00	0.36	0.36	0.33
a-0000338	Mali	SpOSR	2.51	0.00	1.22	3.73	4.27
a-0000339	Maris Haplona	SpOSR	8.24	0.22	0.32	8.79	7.14
a-0000340	Masora	SpOSR	1.25	0.03	2.04	3.32	1.73
a-0000342	Miyauchi Na	SpOSR	1.79	0.30	1.66	3.75	3.44
a-0000343	Mlochowski	SpOSR	0.45	0.02	1.40	1.87	1.30
a-0000345	Nakate Chousen	SpOSR	3.09	0.22	0.93	4.23	2.47
a-0000346	Nosovskij 9	SpOSR	4.80	0.19	1.07	6.06	3.61
a-0000347	Odin	SpOSR	0.04	0.18	3.00	3.22	1.70
a-0000348	Olivia	SpOSR	8.13	0.41	1.19	9.73	6.30
a-0000349	Omega	SpOSR	0.00	0.04	0.86	0.90	0.57
a-0000350	Optima	SpOSR	2.13	0.17	1.55	3.85	1.50
a-0000351	Orpal	SpOSR	6.77	0.16	2.52	9.45	6.21
a-0000352	Orriba	SpOSR	5.20	0.09	2.24	7.54	2.69
a-0000353	Pera	SpOSR	15.73	0.28	0.41	16.42	6.24
a-0000354	Pivot	SpOSR	6.61	0.12	0.35	7.09	8.26
a-0000355	Pobeda	SpOSR	8.36	0.16	0.45	8.97	7.31
a-0000356	Puma	SpOSR	0.10	0.01	0.69	0.79	0.32
a-0000357	Pura	SpOSR	0.04	0.00	2.05	2.09	1.17
a-0000359	Reston	SpOSR	4.67	0.24	1.27	6.17	3.98
a-0000360	Rsio	SpOSR	1.38	0.06	0.75	2.19	2.12
a-0000361	Rucabo	SpOSR	4.68	0.32	1.32	6.32	5.32
a-0000362	Sabine	SpOSR	1.10	0.01	0.34	1.45	0.98
a-0000365	Sv 705152	SpOSR	3.76	0.09	1.52	5.37	3.35
a-0000366	Sv 706118	SpOSR	0.05	0.07	1.23	1.35	0.84
a-0000367	Sv 716	SpOSR	9.46	0.60	2.69	12.76	2.83
a-0000368	Sv 75716	SpOSR	1.76	0.02	1.49	3.27	1.39
a-0000369	Tanka	SpOSR	0.00	0.00	0.86	0.86	0.64
a-0000370	Toro	SpOSR	0.49	0.43	2.57	3.49	1.42
a-0000371	Triton	SpOSR	0.86	0.04	2.98	3.88	1.90
a-0000372	Uranus	SpOSR	0.95	0.06	1.08	2.09	2.16
a-0000376	Vega	SpOSR	6.37	0.07	0.34	6.77	5.68
a-0000378	Wesbell	SpOSR	17.32	0.20	0.60	18.12	15.47
a-0000380	Wesreo	SpOSR	9.74	0.38	0.71	10.83	8.65
a-0000381	Wesroona	SpOSR	1.37	0.08	2.03	3.49	3.19
a-0000382	Willi	SpOSR	0.22	0.28	4.51	5.01	1.38
a-0000383	Ww 1286	SpOSR	0.01	0.15	2.25	2.41	1.67
a-0000385	Ww1289	SpOSR	11.18	0.15	1.57	12.91	5.46

Appendix 1 (continued)

York accession_id	Cultivar name	Crop type	Leaf_Ali	Leaf_Aro	Leaf_Ind	Leaf_Total	Leaf_Total StdDev
a-0000386	Zachodni	SpOSR	2.07	1.24	2.17	5.49	2.40
a-0000387	Zairai Chousenshu	SpOSR	4.33	0.77	1.30	6.41	1.57
a-0000389	VIGE DH1	Swede	1.83	0.01	3.89	5.73	3.44
a-0000396	VOGESA	Swede	0.19	0.07	0.91	1.16	0.32
a-0000399	JAUNE A COLLET VERT	Swede	3.67	0.47	2.44	6.57	5.17
a-0000401	PIKE	Swede	5.21	0.40	6.04	11.65	9.88
a-0000402	SENSATION NZ	Swede	1.36	0.22	1.92	3.50	3.31
a-0000403	Wilhelmsburger	Swede	11.51	0.25	2.31	14.06	9.99
a-0000406	Altasweet	Swede	15.31	0.00	2.63	17.94	15.59
a-0000409	Bangholm PT	Swede	4.09	0.28	2.09	6.47	5.43
a-0000411	Britannia	Swede	0.17	1.55	4.59	6.30	4.14
a-0000412	Conqueror Bronze Green Top	Swede	2.59	0.14	2.12	4.85	5.26
a-0000414	Drummonds Purple Top	Swede	6.58	0.02	2.26	8.87	8.63
a-0000415	Essex Model	Swede	1.69	0.00	2.69	4.38	0.56
a-0000418	Parkside	Swede	1.00	0.10	4.25	5.36	1.39
a-0000419	Peerless (Acme)	Swede	6.91	0.00	1.46	8.37	7.68
a-0000420	Purple Top	Swede	4.83	1.10	5.54	11.47	9.46
a-0000421	Scotia	Swede	0.84	1.21	2.48	4.53	3.06
a-0000422	Tankard Bronze Top	Swede	11.15	0.16	1.63	12.94	9.63
a-0000423	The Bell	Swede	6.84	0.11	1.50	8.45	7.19
a-0000424	Tina	Swede	4.99	0.04	2.11	7.15	4.47
a-0000426	YORK	Swede	1.52	0.00	3.63	5.15	2.12
a-0000431	Brandhaug	Swede	2.12	0.11	3.99	6.22	1.47
a-0000433	Laugabolsrofa	Swede	4.89	0.32	4.16	9.37	9.61
a-0000435	Rotabaggeue	Swede	0.63	0.53	2.75	3.91	2.94
a-0000439	Kalfafellsrofa	Swede	5.35	0.00	2.38	7.73	2.95
a-0000440	Magres Pajberg	Swede	5.56	0.00	4.10	9.66	8.48
a-0000442	Troendersk Kvithamar	Swede	0.91	0.00	3.45	4.37	1.55
a-0000497	Cabernet	WOSR	2.29	0.03	1.57	3.89	4.12
a-0000498	Cabriolet	WOSR	1.15	0.02	0.58	1.75	1.01
a-0000499	Castille	WOSR	3.92	0.11	0.83	4.86	4.07
a-0000500	Catana	WOSR	8.03	0.22	1.43	9.68	8.42
a-0000501	Chuanyou 2	Semiwinter OSR	5.70	0.12	2.12	7.94	6.95
a-0000505	Huron x Navajo	WOSR	2.73	0.04	0.89	3.66	3.72
a-0000508	Ningyou 7	Semiwinter OSR	1.60	0.06	1.80	3.45	4.34
a-0000510	POH 285, Bolko	WOSR	0.00	0.01	2.56	2.57	0.97
a-0000511	Quinta	WOSR	17.73	0.26	3.56	21.55	11.68
a-0000512	Rocket	WOSR	3.82	0.12	1.26	5.20	3.68
a-0000514	Shengliyoucai	Semiwinter OSR	2.40	0.10	1.42	3.92	2.85

Appendix 1 (continued)

York accession_id	Cultivar name	Crop type	Leaf_Ali	Leaf_Aro	Leaf_Ind	Leaf_Total	Leaf_Total StdDev
a-0000517	Xiangyou 15	Semiwinter OSR	0.68	0.03	1.81	2.51	0.78
a-0000518	Zhongshuang II	Semiwinter OSR	0.87	0.00	1.22	2.09	1.97
a-0000519	Bronze-535	WOSR	4.81	0.08	0.40	5.30	4.46
a-0000521	Cracker-531	WOSR	1.26	0.08	0.95	2.30	1.71

Appendix 2. Mean quantity of glucosinolate compositions in the roots of 288 *B. napus* accessions (µmol/g). For individual GSL data, please see Spreadsheet 1 in Accompanying Material or Appendix 1 of Kittipol *et al*., 2019b.

York accession_id	Cultivar name	Crop type	Root_Ali	Root_Aro	Root_Ind	Root_Total	Root_Total StdDev
a-0000003	Robust	WOSR	0.41	1.66	2.39	4.45	0.38
a-0000004	Alaska	WOSR	0.10	1.64	3.51	5.25	1.29
a-0000005	Pirola	WOSR	0.23	2.28	3.15	5.66	2.19
a-0000008	Allure	WOSR	0.04	2.22	3.10	5.37	1.88
a-0000009	Agalon	WOSR	0.62	1.77	3.84	6.23	3.47
a-0000014	Rodeo	WOSR	0.21	1.30	2.79	4.31	1.96
a-0000018	Montego	WOSR	0.34	1.07	2.54	3.95	0.15
a-0000020	Pacific	WOSR	0.21	1.44	4.63	6.28	1.99
a-0000022	Missouri	WOSR	0.08	1.12	3.56	4.76	1.79
a-0000023	Manitoba	WOSR	0.36	3.03	3.91	7.30	3.40
a-0000024	Ladoga	WOSR	0.13	1.40	2.08	3.61	1.24
a-0000025	Atlantic	WOSR	0.11	2.23	2.34	4.68	0.92
a-0000026	Cooper	WOSR	0.06	1.40	2.61	4.07	1.53
a-0000027	Licapo	WOSR	0.10	2.24	2.37	4.72	2.23
a-0000028	Capitol	WOSR	0.23	2.43	1.85	4.50	2.37
a-0000029	Idol	WOSR	0.14	1.46	4.49	6.08	2.93
a-0000030	Vivol	WOSR	0.34	0.58	2.30	3.23	0.82
a-0000031	BRISTOL	WOSR	0.07	0.37	4.15	4.59	2.08
a-0000033	Lisabeth	WOSR	0.29	1.76	2.97	5.02	2.29
a-0000034	Lipid	WOSR	0.53	5.45	4.44	10.42	5.02
a-0000036	Lisek	WOSR	0.12	0.50	1.95	2.57	1.20
a-0000037	Contact	WOSR	0.82	2.78	3.87	7.47	1.99
a-0000038	Lion	WOSR	0.03	0.50	4.72	5.25	2.46
a-0000040	Apex	WOSR	0.14	2.15	3.46	5.75	2.54
a-0000042	Magnum	WOSR	0.27	3.43	4.90	8.61	5.26
a-0000044	Laser	WOSR	0.12	2.56	2.26	4.93	1.83
a-0000045	Fortis	WOSR	0.08	2.37	1.92	4.37	3.27
a-0000048	NK Bravour	WOSR	0.14	4.12	6.46	10.71	7.35
a-0000049	NK Fair	WOSR	0.74	5.22	3.83	9.80	0.94
a-0000050	Aviso	WOSR	0.00	4.61	3.95	8.55	5.48
a-0000053	Verona	WOSR	0.14	1.30	4.13	5.57	3.15
a-0000054	Tenor	WOSR	0.00	7.48	3.75	11.23	5.69
a-0000055	Expert	WOSR	0.42	4.22	4.50	9.13	3.31
a-0000056	Musette	WOSR	0.00	1.91	4.69	6.60	4.45
a-0000057	Kvintett	WOSR	0.68	3.03	1.90	5.61	4.34
a-0000058	Falstaff	WOSR	0.01	2.59	3.15	5.76	2.78
a-0000059	SW Sinatra	WOSR	0.10	0.90	4.22	5.22	3.07
a-0000060	Viking	WOSR	0.00	4.40	3.84	8.24	4.53
a-0000062	Aurum	WOSR	0.29	1.73	3.17	5.18	3.31
a-0000065	Rasmus	WOSR	0.25	4.45	5.12	9.82	5.48
a-0000066	Gefion	WOSR	0.51	1.83	3.78	6.12	3.18

Appendix 2 (continued)

York accession_id	Cultivar name	Crop type	Root_Ali	Root_Aro	Root_Ind	Root_Total	Root_Total StdDev
a-0000067	Nugget	WOSR	0.18	1.36	2.14	3.69	1.90
a-0000068	Zephir	WOSR	0.01	1.16	4.32	5.49	2.50
a-0000069	SLM 0413	WOSR	0.00	3.45	2.96	6.41	1.60
a-0000071	LSF 0519	WOSR	0.00	3.08	1.32	4.40	1.07
a-0000072	Beluga	WOSR	1.38	5.54	1.32	8.25	2.15
a-0000073	Amor	WOSR	0.00	2.22	1.47	3.69	1.10
a-0000075	Campari	WOSR	0.76	8.76	3.64	13.16	3.25
a-0000077	Duell	WOSR	0.65	1.99	3.86	6.50	1.49
a-0000079	Jessica	WOSR	1.29	4.08	2.06	7.43	3.62
a-0000080	Orlando	WOSR	0.00	2.11	2.36	4.47	1.72
a-0000081	Pollen	WOSR	0.19	4.61	1.24	6.04	2.96
a-0000082	Prince	WOSR	0.08	0.63	2.03	2.74	1.01
a-0000083	Wotan	WOSR	0.00	0.71	1.88	2.59	1.34
a-0000084	NK Nemax	WOSR	0.48	1.72	1.45	3.66	2.16
a-0000085	NK Passion	WOSR	0.09	4.81	1.89	6.78	2.85
a-0000089	AMBER X COMMANCHE DH LINE	WOSR	0.51	3.31	2.90	6.72	3.83
a-0000090	APEX-93_5 X GINYOU_3 DH LINE	WOSR	0.61	5.53	2.05	8.18	3.99
a-0000093	CANBERRA x COURAGE DH LINE	WOSR	0.00	3.03	2.50	5.53	1.10
a-0000096	HANSEN X GASPARD DH LINE	WOSR	0.46	4.04	1.25	5.75	4.17
a-0000097	MADRIGAL x RECITAL DH LINE	WOSR	0.34	3.51	1.51	5.36	3.94
a-0000099	TAPIDOR DH	WOSR	0.30	0.75	2.69	3.74	0.99
a-0000101	EUROL	WOSR	0.42	4.26	1.61	6.28	3.65
a-0000105	LICROWN X EXPRESS DH LINE	WOSR	0.28	3.71	1.50	5.50	1.46
a-0000106	SHANNON x WINNER DH LINE	WOSR	0.29	2.02	1.74	4.05	0.53
a-0000107	JANETZKIS SCHLESISCHER	WOSR	1.02	9.64	1.64	12.30	3.70
a-0000110	OLIMPIADE	WOSR	0.30	2.55	1.31	4.15	0.55
a-0000113	Samourai	WOSR	0.12	1.36	3.68	5.16	1.05
a-0000114	Sollux	WOSR	1.42	8.16	1.94	11.51	4.85
a-0000115	Akela	WOSR	1.28	7.14	2.16	10.57	3.59
a-0000117	Maplus	WOSR	0.22	4.77	3.51	8.51	2.42
a-0000118	Askari	WOSR	0.66	2.54	1.25	4.45	2.56
a-0000119	Lirabon	WOSR	0.73	6.49	1.34	8.56	4.35
a-0000121	JetNeuf	WOSR	0.42	4.69	1.86	6.97	4.56
a-0000122	Cobra	WOSR	1.11	4.57	2.41	8.09	2.61
a-0000123	Falcon	WOSR	0.18	3.17	1.71	5.05	0.50
a-0000124	Mohican	WOSR	0.10	2.63	1.26	4.00	1.97
a-0000125	Flip	WOSR	0.12	1.03	2.78	3.92	2.12

Appendix 2 (continued)

York accession_id	Cultivar name	Crop type	Root_Ali	Root_Aro	Root_Ind	Root_Total	Root_Total StdDev
a-0000127	Phil	WOSR	0.36	4.86	3.66	8.88	7.36
a-0000128	Leopard	WOSR	0.18	4.41	2.12	6.71	4.11
a-0000129	RESYN-H048	WOSR	0.62	4.30	4.11	9.03	7.04
a-0000130	Resyn-G_ S4	WOSR	0.74	4.84	2.55	8.13	4.73
a-0000132	Anja	WOSR	0.30	2.55	2.39	5.25	1.64
a-0000133	Baltia	WOSR	0.15	3.74	2.10	5.99	2.72
a-0000135	Brink	WOSR	0.33	2.02	1.59	3.94	1.36
a-0000137	Coriander	WOSR	0.71	3.83	2.05	6.59	3.07
a-0000138	Diamant	WOSR	0.83	6.69	2.65	10.17	2.96
a-0000140	Doral	WOSR	1.13	4.19	1.77	7.09	3.51
a-0000141	Edita	WOSR	0.16	3.33	1.91	5.41	3.53
a-0000144	G™lzower Älquell	WOSR	1.54	3.81	3.78	9.13	4.12
a-0000146	Janpol	WOSR	1.39	0.92	3.13	5.44	2.38
a-0000148	Jupiter	WOSR	1.49	7.06	1.88	10.44	3.63
a-0000149	Krapphauser	WOSR	0.75	2.40	3.06	6.20	2.04
a-0000150	Kromerska	WOSR	0.09	1.48	1.72	3.29	1.48
a-0000151	Librador	WOSR	0.35	5.15	3.10	8.59	3.78
a-0000152	Libritta	WOSR	0.25	4.78	2.08	7.11	1.97
a-0000156	Lirakotta	WOSR	1.13	4.12	3.96	9.21	4.06
a-0000157	Madora	WOSR	0.58	2.15	2.54	5.26	1.81
a-0000162	Moldavia	WOSR	2.83	11.04	3.01	16.89	3.23
a-0000163	Mytnickij	WOSR	0.95	2.33	2.98	6.27	4.56
a-0000164	Nemertschanskij 1	WOSR	1.47	3.46	3.66	8.59	6.51
a-0000166	Panter	WOSR	0.70	4.45	1.31	6.46	1.44
a-0000168	Ramses	WOSR	0.77	0.73	2.51	4.02	0.84
a-0000169	Sarepta	WOSR	1.06	1.95	3.13	6.13	1.87
a-0000170	Skrzeszowicki	WOSR	1.30	5.61	3.16	10.08	5.12
a-0000171	Skziverskij	WOSR	0.47	3.90	2.99	7.35	0.67
a-0000172	Slovenska Krajova	WOSR	1.17	8.66	2.65	12.47	4.98
a-0000175	Start	WOSR	0.29	4.92	3.02	8.24	0.70
a-0000176	Trebicska	WOSR	0.47	4.42	3.56	8.45	1.73
a-0000178	Vinnickij 15/59	WOSR	2.38	4.12	2.45	8.96	2.11
a-0000179	Wolynski	WOSR	1.08	3.86	3.19	8.13	3.96
a-0000185	CANARD	Fodder	2.05	3.25	4.06	9.37	4.96
a-0000188	EMERALD	Fodder	1.35	5.57	2.46	9.38	2.35
a-0000189	FORA	Fodder	1.39	4.57	4.64	10.60	3.21
a-0000190	Aphid Resistant Rape	Fodder	0.54	3.43	4.48	8.45	0.44
a-0000193	Dwarf Essex	Fodder	0.19	3.75	2.74	6.67	0.79
a-0000200	Samo	Fodder	0.78	5.56	2.09	8.43	1.21
a-0000209	RAGGED JACK	Kale	0.54	3.23	4.36	8.12	1.45
a-0000210	RED RUSSIAN	Kale	1.32	3.40	5.12	9.84	3.92
a-0000211	SIBERISCHE BOERENKOOL	Kale	0.92	5.50	1.48	7.90	2.47

Appendix 2 (continued)

York accession_id	Cultivar name	Crop type	Root_Ali	Root_Aro	Root_Ind	Root_Total	Root_Total StdDev
a-0000226	SWU Chinese 1	Semiwinter OSR	0.52	4.87	3.16	8.55	1.25
a-0000227	SWU Chinese 2	Semiwinter OSR	0.23	3.59	5.99	9.81	3.31
a-0000228	SWU Chinese 3	Semiwinter OSR	0.03	5.56	4.13	9.72	1.93
a-0000232	SWU Chinese 8	Semiwinter OSR	0.03	1.59	3.28	4.89	1.15
a-0000233	SWU Chinese 9	Semiwinter OSR	1.04	5.94	4.63	11.62	5.39
a-0000234	Zhouyou	Semiwinter OSR	0.10	4.60	5.96	10.65	1.89
a-0000235	Drakkar	SpOSR	0.14	2.69	3.32	6.16	3.08
a-0000236	STELLAR DH	SpOSR	0.00	0.48	8.64	9.12	2.95
a-0000237	WESTAR DH	SpOSR	0.00	3.97	4.91	8.89	3.31
a-0000238	YUDAL	SpOSR	0.68	7.31	2.73	10.72	2.63
a-0000239	BRUTOR	SpOSR	0.70	0.96	5.54	7.20	2.68
a-0000241	COMET	SpOSR	0.33	3.26	3.88	7.47	6.48
a-0000242	CRESOR	SpOSR	0.13	2.54	5.64	8.32	2.26
a-0000247	INDUSTRY	SpOSR	0.28	2.58	8.35	11.21	3.24
a-0000248	KARAT	SpOSR	0.00	0.00	3.59	3.59	0.64
a-0000249	MARINKA	SpOSR	0.00	2.10	5.58	7.68	2.46
a-0000250	NIKLAS	SpOSR	0.57	5.05	5.03	10.65	5.52
a-0000251	TARGET	SpOSR	0.87	8.33	4.65	13.85	2.97
a-0000253	KAROO-057DH	SpOSR	0.00	3.39	4.57	7.96	2.45
a-0000254	MONTY-028DH	SpOSR	0.21	11.15	5.74	17.10	4.36
a-0000255	N01D-1330	SpOSR	0.05	11.19	3.95	15.20	1.85
a-0000256	N02D-1952	SpOSR	0.04	0.96	6.15	7.15	2.68
a-0000257	SURPASS400-024DH	SpOSR	0.02	1.36	2.28	3.66	1.93
a-0000258	CUBS ROOT	SpOSR	0.16	2.91	2.56	5.63	1.75
a-0000259	DUX	SpOSR	0.00	4.71	4.01	8.72	3.95
a-0000260	ERGLU	SpOSR	0.00	2.76	4.84	7.60	1.82
a-0000261	HELIOS	SpOSR	0.00	1.93	3.68	5.61	1.67
a-0000262	KROKO	SpOSR	0.19	2.35	4.80	7.34	0.90
a-0000264	LINETTA	SpOSR	1.12	4.07	4.88	10.08	3.70
a-0000265	MAZOWIECKI	SpOSR	1.23	3.60	2.57	7.40	0.63
a-0000267	WEIHENSTEPHANER	SpOSR	0.82	3.40	2.11	6.32	2.55
a-0000269	Alku	SpOSR	0.58	1.56	3.80	5.94	2.62
a-0000270	Bronowski	SpOSR	0.00	0.00	4.68	4.68	2.66
a-0000271	Ceska Krajova	SpOSR	0.66	4.28	3.61	8.56	3.57
a-0000272	Duplo	SpOSR	0.03	0.09	2.30	2.42	1.51
a-0000273	Janetzkis Sommerraps	SpOSR	0.25	1.71	3.55	5.51	2.77
a-0000274	Line	SpOSR	0.00	0.76	4.00	4.76	1.17
a-0000275	Marnoo	SpOSR	0.68	3.43	4.33	8.44	1.96

Appendix 2 (continued)

York accession_id	Cultivar name	Crop type	Root_Ali	Root_Aro	Root_Ind	Root_Total	Root_Total StdDev
a-0000276	Nugget	SpOSR	1.10	5.16	2.24	8.51	1.10
a-0000277	Olga	SpOSR	0.90	2.99	2.76	6.65	1.07
a-0000278	Spaeths Zollerngold	SpOSR	0.17	1.46	4.56	6.19	3.13
a-0000279	Sval_f⍰s Gulle	SpOSR	0.49	3.96	2.08	6.53	2.90
a-0000281	Tribute	SpOSR	0.00	0.09	3.57	3.66	1.83
a-0000282	Wesway	SpOSR	0.53	4.37	2.78	7.67	3.74
a-0000283	Fido	SpOSR	1.71	6.96	3.32	11.99	3.73
a-0000284	Oro	SpOSR	0.00	1.74	4.22	5.96	0.96
a-0000285	Tower	SpOSR	0.12	0.28	6.62	7.01	2.07
a-0000286	Regina II	SpOSR	0.19	5.79	3.86	9.84	3.07
a-0000288	Campino	SpOSR	0.49	4.31	3.15	7.94	6.69
a-0000289	Clipper	SpOSR	0.00	3.44	2.51	5.95	2.23
a-0000290	Larissa	SpOSR	0.00	4.57	3.45	8.01	3.83
a-0000291	Magma	SpOSR	0.00	4.63	1.13	5.75	2.35
a-0000292	Monty	SpOSR	0.10	3.94	1.31	5.35	1.80
a-0000294	Pauline	SpOSR	0.08	9.95	2.19	12.21	1.74
a-0000295	Sophia	SpOSR	0.44	8.37	2.57	11.38	4.70
a-0000296	Tribune	SpOSR	0.00	1.76	2.67	4.43	2.71
a-0000297	Trigold	SpOSR	0.00	2.53	2.66	5.18	2.97
a-0000298	Rivette	SpOSR	0.16	1.73	4.87	6.76	3.01
a-0000300	Adamo	SpOSR	0.28	3.42	1.85	5.55	2.71
a-0000301	Altex	SpOSR	0.07	0.51	4.69	5.27	1.80
a-0000302	Andor	SpOSR	0.00	3.01	4.55	7.56	3.76
a-0000303	Astor	SpOSR	0.06	0.49	2.14	2.69	1.05
a-0000305	Bingo	SpOSR	0.05	9.80	3.11	12.96	0.81
a-0000306	Callypso	SpOSR	0.00	0.42	2.21	2.62	1.11
a-0000307	Concord	SpOSR	0.20	0.15	2.26	2.61	0.43
a-0000308	Conny	SpOSR	0.06	2.64	2.94	5.65	1.73
a-0000309	Czyzowska	SpOSR	0.00	1.15	2.61	3.76	1.37
a-0000310	Daichousen (fuku)	SpOSR	0.38	3.78	2.72	6.88	3.62
a-0000311	Daichousen (mizuyasu)	SpOSR	0.00	9.37	1.50	10.87	3.10
a-0000312	Daichousen (nakano)	SpOSR	0.13	3.04	4.40	7.57	2.02
a-0000313	Erake	SpOSR	0.03	0.35	2.34	2.72	0.62
a-0000314	Furax	SpOSR	0.21	8.00	3.18	11.38	3.65
a-0000316	Galant	SpOSR	1.24	0.59	2.42	4.25	1.55
a-0000317	Giant Xr707	SpOSR	0.47	3.72	3.88	8.07	2.99
a-0000318	Gisora	SpOSR	0.28	1.34	1.50	3.12	1.65
a-0000321	Granit	SpOSR	0.89	5.10	2.27	8.25	3.78
a-0000323	Hankkija's Lauri	SpOSR	0.00	5.11	2.00	7.11	2.35
a-0000326	Kajsa	SpOSR	0.20	0.73	4.82	5.76	2.81
a-0000327	Korall	SpOSR	0.42	3.90	2.63	6.95	1.87
a-0000328	Korinth	SpOSR	0.20	3.35	1.74	5.30	2.47

Appendix 2 (continued)

York accession_id	Cultivar name	Crop type	Root_Ali	Root_Aro	Root_Ind	Root_Total	Root_Total StdDev
a-0000329	Kosa	SpOSR	0.07	4.33	3.42	7.82	3.62
a-0000330	Kruglik	SpOSR	0.00	0.89	5.64	6.53	2.59
a-0000332	Lirafox	SpOSR	0.08	1.60	1.98	3.66	2.41
a-0000333	Lirasol	SpOSR	0.00	1.58	2.25	3.83	1.18
a-0000334	Liraspa	SpOSR	0.04	0.31	2.73	3.08	0.96
a-0000335	Lirawell	SpOSR	0.00	0.15	2.66	2.81	0.72
a-0000336	Lisandra	SpOSR	0.00	0.42	2.64	3.06	0.65
a-0000337	Loras	SpOSR	0.00	0.12	5.01	5.13	2.97
a-0000338	Mali	SpOSR	0.17	1.00	1.67	2.84	2.11
a-0000339	Maris Haplona	SpOSR	2.20	2.94	3.16	8.29	2.13
a-0000340	Masora	SpOSR	0.00	0.04	5.58	5.62	1.02
a-0000342	Miyauchi Na	SpOSR	0.05	3.78	3.74	7.57	3.11
a-0000343	Mlochowski	SpOSR	0.26	2.86	2.68	5.80	2.17
a-0000345	Nakate Chousen	SpOSR	1.86	6.15	2.46	10.46	3.13
a-0000346	Nosovskij 9	SpOSR	0.42	4.76	2.68	7.87	1.71
a-0000347	Odin	SpOSR	0.07	1.12	2.40	3.59	1.74
a-0000348	Olivia	SpOSR	1.32	11.15	2.50	14.97	4.27
a-0000349	Omega	SpOSR	0.22	0.65	3.39	4.26	1.36
a-0000350	Optima	SpOSR	0.83	4.08	1.72	6.62	2.23
a-0000351	Orpal	SpOSR	0.48	2.11	3.03	5.61	0.94
a-0000352	Orriba	SpOSR	0.66	2.25	3.28	6.20	2.71
a-0000353	Pera	SpOSR	0.52	3.98	2.23	6.72	1.00
a-0000354	Pivot	SpOSR	0.41	4.08	5.19	9.68	3.89
a-0000355	Pobeda	SpOSR	0.75	2.48	2.95	6.17	1.27
a-0000356	Puma	SpOSR	0.08	1.80	5.35	7.23	1.09
a-0000357	Pura	SpOSR	0.00	1.46	5.42	6.89	2.47
a-0000359	Reston	SpOSR	1.42	4.76	3.72	9.90	4.46
a-0000360	Rsio	SpOSR	0.74	1.93	2.64	5.31	3.24
a-0000361	Rucabo	SpOSR	1.23	4.58	2.50	8.31	2.15
a-0000362	Sabine	SpOSR	1.23	4.39	2.38	8.01	4.16
a-0000365	Sv 705152	SpOSR	0.57	1.96	1.54	4.07	0.74
a-0000366	Sv 706118	SpOSR	0.27	8.61	1.55	10.43	5.25
a-0000367	Sv 716	SpOSR	2.14	9.42	2.79	14.36	3.74
a-0000368	Sv 75716	SpOSR	0.83	0.36	2.23	3.42	2.35
a-0000369	Tanka	SpOSR	0.00	2.88	6.08	8.96	1.71
a-0000370	Toro	SpOSR	0.05	0.65	3.30	4.01	1.42
a-0000371	Triton	SpOSR	0.03	2.01	3.27	5.31	2.53
a-0000372	Uranus	SpOSR	0.24	1.28	2.78	4.30	1.26
a-0000376	Vega	SpOSR	0.72	7.09	1.69	9.50	0.12
a-0000378	Wesbell	SpOSR	1.07	1.94	1.31	4.32	0.61
a-0000380	Wesreo	SpOSR	1.49	3.06	1.78	6.34	3.24
a-0000381	Wesroona	SpOSR	0.35	3.74	3.74	7.83	1.22
a-0000382	Willi	SpOSR	0.02	0.74	3.40	4.16	1.64
a-0000383	Ww 1286	SpOSR	0.00	0.20	5.18	5.38	2.77

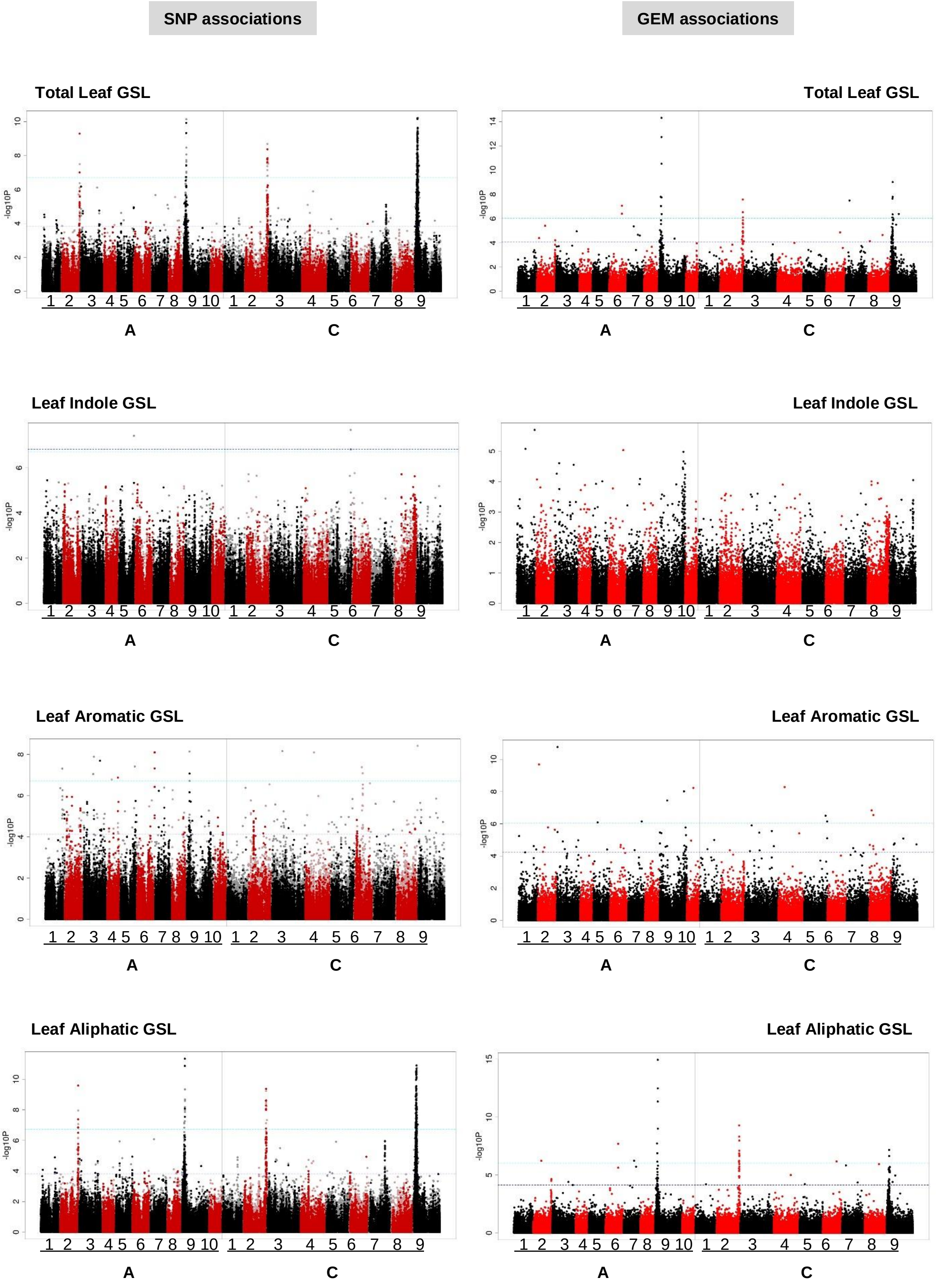
Appendix 2 (continued)

York accession_id	Cultivar name	Crop type	Root_Ali	Root_Aro	Root_Ind	Root_Total	Root_Total StdDev
a-0000385	Ww1289	SpOSR	0.37	6.64	3.99	11.00	3.92
a-0000386	Zachodni	SpOSR	0.49	2.31	2.17	4.96	1.92
a-0000387	Zairai Chousenshu	SpOSR	0.45	4.48	2.93	7.86	2.54
a-0000389	VIGE DH1	Swede	1.70	0.72	3.77	6.19	1.34
a-0000396	VOGESA	Swede	1.77	0.52	2.61	4.90	1.27
a-0000399	JAUNE A COLLET VERT	Swede	1.14	3.22	2.86	7.22	4.09
a-0000401	PIKE	Swede	1.15	1.79	3.24	6.18	2.52
a-0000402	SENSATION NZ	Swede	0.38	1.48	3.63	5.50	2.32
a-0000403	Wilhelmsburger	Swede	1.26	2.15	4.49	7.91	2.79
a-0000406	Altasweet	Swede	2.35	0.44	4.47	7.27	3.93
a-0000409	Bangholm PT	Swede	2.37	1.47	4.30	8.14	7.08
a-0000411	Britannia	Swede	1.38	3.12	9.42	13.91	3.86
a-0000412	Conqueror Bronze Green Top	Swede	1.90	1.98	11.95	15.84	7.16
a-0000414	Drummonds Purple Top	Swede	1.18	0.24	3.54	4.97	3.11
a-0000415	Essex Model	Swede	1.81	0.51	5.48	7.79	4.25
a-0000418	Parkside	Swede	0.30	0.91	6.45	7.66	2.65
a-0000419	Peerless (Acme)	Swede	1.11	2.10	6.80	10.02	2.32
a-0000420	Purple Top	Swede	0.58	2.22	9.26	12.06	2.78
a-0000421	Scotia	Swede	1.54	2.80	4.06	8.40	6.42
a-0000422	Tankard Bronze Top	Swede	1.80	5.32	4.91	12.03	5.89
a-0000423	The Bell	Swede	1.51	1.29	4.37	7.16	0.86
a-0000424	Tina	Swede	0.81	3.07	4.72	8.59	2.32
a-0000426	YORK	Swede	0.00	0.29	9.72	10.01	1.73
a-0000431	Brandhaug	Swede	0.66	0.00	3.71	4.37	0.48
a-0000433	Laugabolsrofa	Swede	1.26	0.41	6.01	7.67	3.43
a-0000435	Rotabaggeue	Swede	0.79	0.47	5.15	6.42	2.83
a-0000439	Kalfafellsrofa	Swede	0.61	2.38	10.15	13.14	5.39
a-0000440	Magres Pajberg	Swede	1.69	0.82	6.36	8.87	3.79
a-0000442	Troendersk Kvithamar	Swede	0.82	0.00	6.21	7.03	4.20
a-0000497	Cabernet	WOSR	0.45	1.94	3.91	6.30	2.03
a-0000498	Cabriolet	WOSR	0.62	6.70	2.10	9.41	3.07
a-0000499	Castille	WOSR	0.67	4.16	3.74	8.58	4.43
a-0000500	Catana	WOSR	0.50	4.51	4.49	9.50	2.60
a-0000501	Chuanyou 2	Semiwinter OSR	0.72	5.78	5.95	12.45	3.91
a-0000505	Huron x Navajo	WOSR	0.52	3.96	3.30	7.79	4.29
a-0000508	Ningyou 7	Semiwinter OSR	0.22	2.62	3.67	6.51	3.33
a-0000510	POH 285, Bolko	WOSR	0.00	3.92	3.29	7.21	1.19
a-0000511	Quinta	WOSR	0.77	1.24	5.01	7.02	1.91
a-0000512	Rocket	WOSR	0.53	5.35	2.91	8.79	2.37

Appendix 2 (continued)

York accession_id	Cultivar name	Crop type	Root_Ali	Root_Aro	Root_Ind	Root_Total	Root_Total StdDev
a-0000514	Shengliyoucai	Semiwinter OSR	0.49	4.16	5.27	9.92	3.22
a-0000517	Xiangyou 15	Semiwinter OSR	0.04	2.28	4.72	7.05	1.54
a-0000518	Zhongshuang II	Semiwinter OSR	0.26	6.17	4.35	10.78	4.87
a-0000519	Bronze-535	WOSR	1.82	6.41	2.59	10.82	5.53
a-0000521	Cracker-531	WOSR	0.31	4.56	2.89	7.76	4.04

Appendix 3. Association analysis of glucosinolate structural classes found in leaves. Absolute amount of GSL (µmol/g) were used as traits for 288 accessions. For individual GSL AT see Appendix 5 in Kittipol *et al.* 2019b.

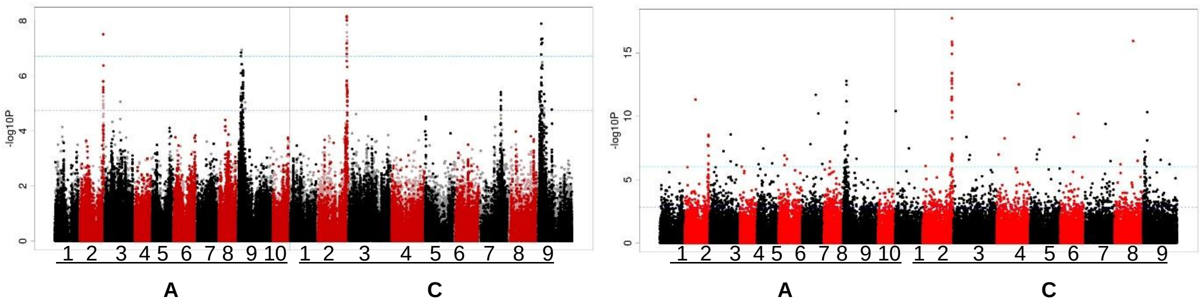


Appendix 4. Association analysis of total seed GSL. Trait data as absolute amount of total seed GSL (µmol/g) measured with near-infrared spectroscopy for 190 *B. napus* accessions came from Lu *et al.* (2014).

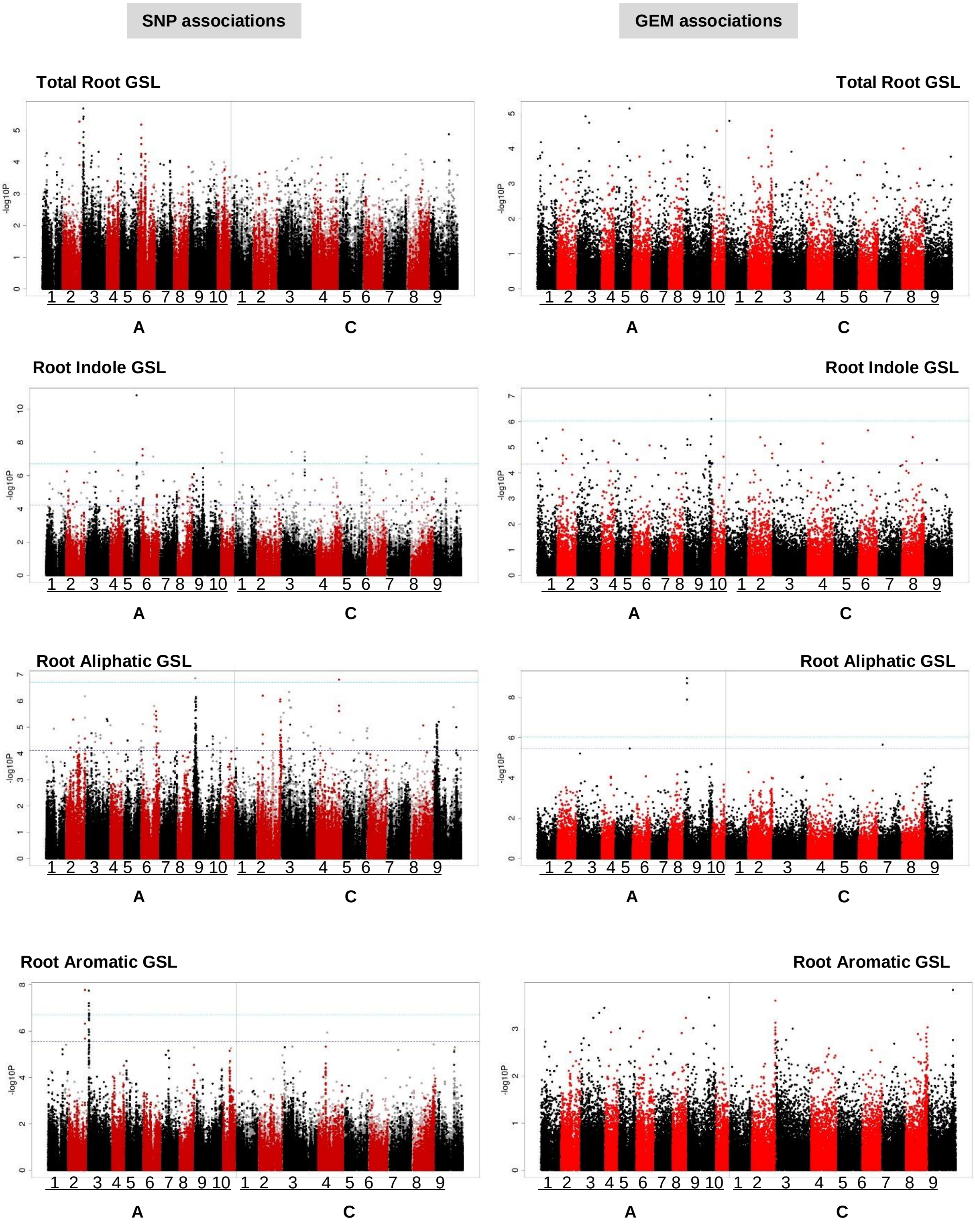
SNP associations

GEM associations

Total Seed GSL – 190 accessions (trait data from Lu *et al*., 2014)



Appendix 5. Association analysis of glucosinolate structural classes found in roots. Absolute amount of GSL (µmol/g) were used as traits for 288 accessions. For individual GSL AT see Appendix 4 in Kittipol *et al.* 2019b.



SNP associations

GEM associations

Total Root GSL

Total Root GSL

Root Indole GSL

Root Indole GSL

Root Aliphatic GSL

Root Aliphatic GSL

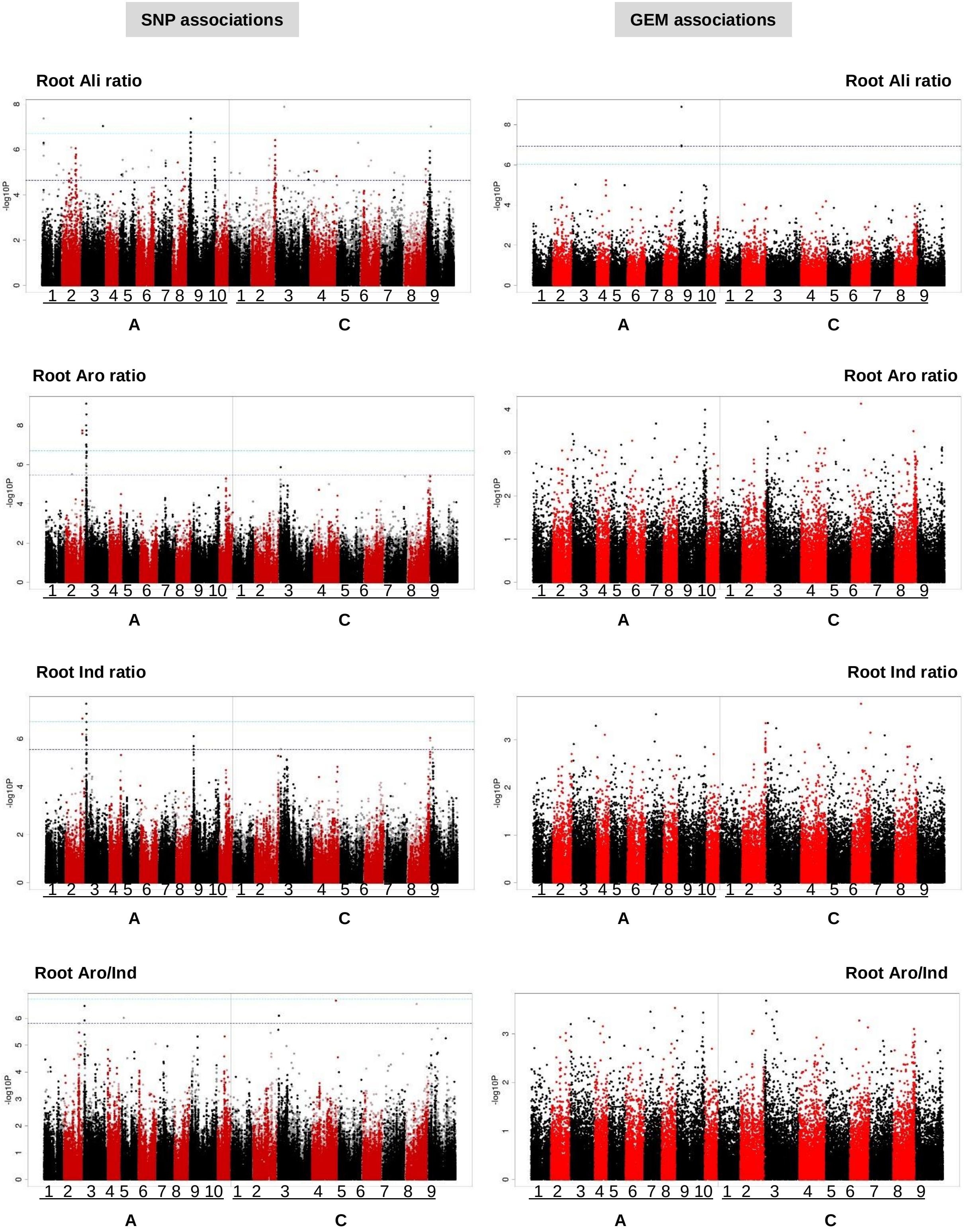
Root Aromatic GSL

Root Aromatic GSL

Appendix 6. WGCNA 'red' module node parameters. Degree is a number of edges linked to it (connectivity). Neighbourhood connectivity is an average connectivity of all neighbours. Topological coefficient is a measure for the extent to which node shares neighbours with other nodes. Clustering coefficient is a ratio of number of edges between the neighbours of nodes divided by the maximum number of edges that could possibly exist between the node.

Gene ID	TAIR	Degree	Neighbourhood connectivity	Topological Coefficient	Clustering Coefficient
Bo4g018590.1	AT2G43100.1	17	14.7	0.74	0.85
Bo9g159950.1	AT5G17700.1	17	12.1	0.67	0.67
Cab034792.1	AT2G31790.1	17	14.7	0.74	0.85
Cab036831.1	AT2G43100.1	17	14.7	0.74	0.85
Bo3g122030.1	AT3G44550.1	16	12.2	0.68	0.68
Bo4g130780.1	AT4G13770.1	16	15.3	0.90	0.94
Bo4g191120.1	AT4G13770.1	16	15.3	0.90	0.94
Bo8g101260.1	AT2G22240.2	16	15.3	0.90	0.94
Bol004799	AT5G23020.1	16	15.3	0.90	0.94
Cab001421.1	AT5G23020.1	16	15.3	0.90	0.94
Cab015634.1	AT5G07690.1	16	15.3	0.90	0.94
Cab016602.1	AT4G13770.1	16	15.3	0.90	0.94
Cab021103.1	AT1G74090.1	16	15.3	0.90	0.94
Cab013463.1	AT2G22240.2	15	15.7	0.92	0.98
Cab043155.1	AT4G13770.1	15	15.7	0.92	0.98
Bo2g011730.1	AT5G14200.3	14	15.9	0.94	1.00
Bo9g056000.1	AT2G03750.1	14	13.4	0.74	0.84
Bo9g127350.1	AT5G56840.1	14	13.0	0.72	0.81
Cab001420.1	AT5G23020.1	14	15.9	0.94	1.00
Cab005408.1	AT4G10310.1	14	12.9	0.72	0.80
Cab017505.1	AT5G56840.1	14	13.0	0.72	0.81
Cab037852.1	AT1G48800.1	14	13.4	0.74	0.84
Cab037853.1	AT1G48800.1	14	12.4	0.69	0.70
Bo2g041340.1	AT3G02020.1	13	16.1	0.95	1.00
Bo6g076560.1	AT3G59580.2	13	13.5	0.75	0.87
Bo7g097210.1	AT5G24090.1	13	13.2	0.73	0.86
Bo9g061470.1	AT1G48800.1	12	12.3	0.68	0.71
Bo9g088230.1	AT4G10310.1	12	13.2	0.73	0.89
Cab006993.1	AT5G49350.2	12	13.7	0.76	0.91
Cab010809.1	AT4G35160.1	11	13.0	0.72	0.91
Cab013721.1	AT1G17710.2	11	12.3	0.68	0.75
Cab031872.1	AT1G12940.1	11	16.3	0.96	1.00
Cab011733.1	AT1G74460.1	7	12.0	0.71	1.00
Cab039476.1	AT5G04120.1	7	12.0	0.71	1.00
Cab043798.1	AT3G02020.1	7	12.0	0.71	1.00
Bo8g106910.1	AT1G14250.1	6	10.0	0.50	0.40
Cab041957.1	AT4G19230.2	6	13.0	0.87	1.00
Cab010205.1	AT5G43350.1	3	4.0	0.67	1.00
Cab010206.1	AT5G43350.1	3	4.0	0.67	1.00
Cab039422.1	AT5G05500.1	3	4.0	0.67	1.00

Appendix 7. Association analysis of proportion of root glucosinolates as ratio for 288 *B. napus* accessions. Ratio of GSL were used as traits. Ali, Ind and Aro ratio were calculated by dividing respective GSL classes with total GSL. Abbreviation: Ali, Aliphatic; Ind, indole; Aro, Aromatic GSL.

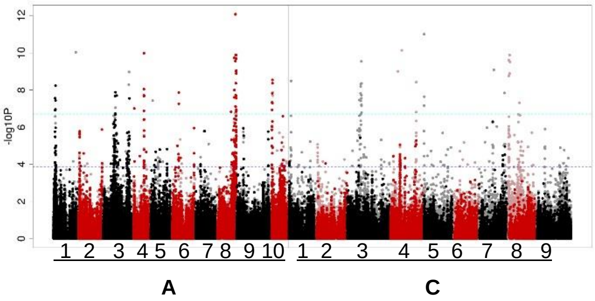


Appendix 7 (continued)

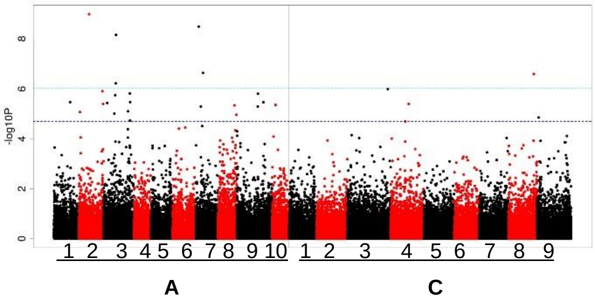
SNP associations

GEM associations

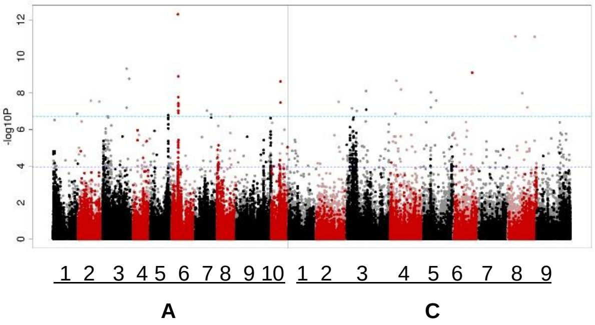
Root Aro/Ali



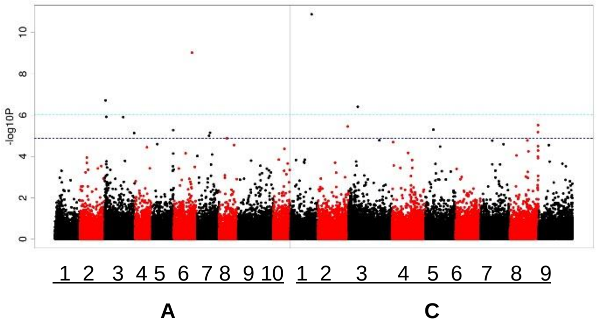
Root Aro/Ali



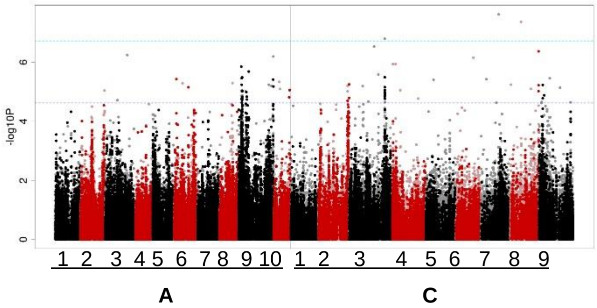
Root Ind/Aro



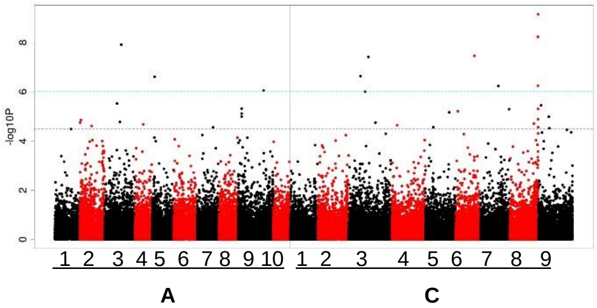
Root Ind/Aro



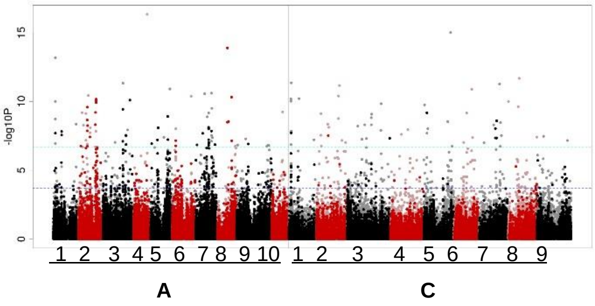
Root Ind/Ali



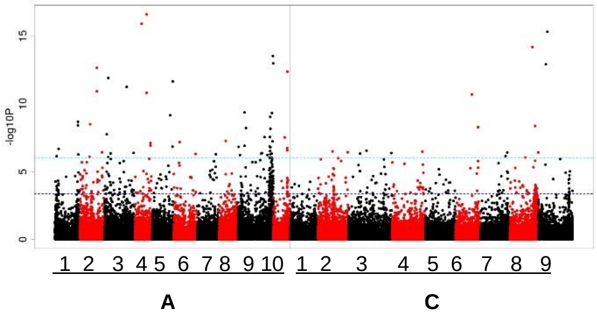
Root Ind/Ali



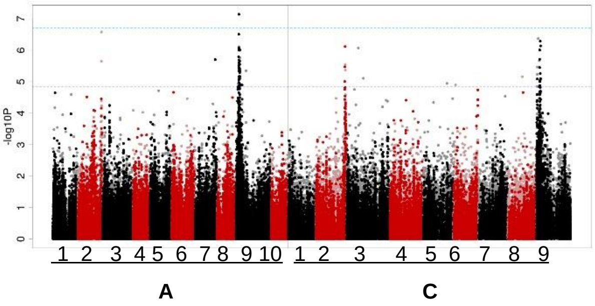
Root Ali/Aro



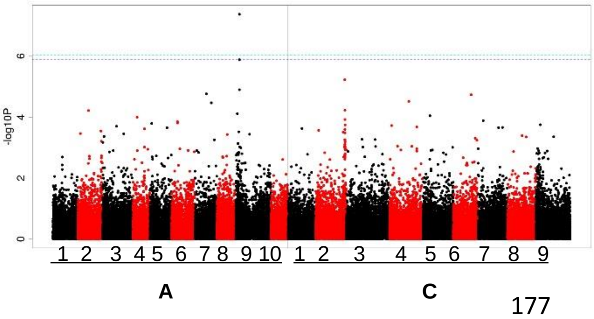
Root Ali/Aro



Root Ali/Ind



Root Ali/Ind



Glossary

Allotetraploid	Hybridisation of two or more diploid genomes from different species, resulting in natural double of chromosomes. For example, the allotetraploid *B. napus* (2n=38, AACC) is derived from the hybridisation between *B. rapa* (2n=20, AA) and *B. oleracea* (2n=18, CC).
Associative Transcriptomics (AT)	Transcriptome-based genome-wide association study using both variations in gene sequences (as SNP markers) and variations in gene expression (as GEMs) for the association with trait variations (Harper *et al.*, 2012)
Biofumigation	Suppressive action of decomposing *Brassica* tissues to control soil-borne pathogens, nematodes, insects and weeds.
Bonferroni significance	The Bonferroni threshold is a family-wise error threshold, designed to control the probability of detecting any positive tests in the family (set) of tests, if the null hypothesis is true (Brett, 2016).
Coding DNA sequence (CDS)	The region of DNA or RNA whose sequence determines the sequence of amino acids in a protein.
Degree	Degree (or connectivity) of a node is the number of edges linked to it.
Desulfoglucosinolate (ds-GSL)	Desulfonated glucosinolates (having sulphate group removed)
Diploid	A cell or organism that has two copies of each chromosome (paired chromosomes), one from each parent.
Edge	Interactions between genes in a gene network.

False discovery rate (FDR)

An alternative approach to the Bonferroni correction. FDR controls for a low proportion of false positive (type I error), instead of guarding against making any false positive conclusions like Bonferroni correction (Glen, 2017).

Gene expression marker (GEM)

A marker that indicates the relationship between variations in gene expression and traits, calculated by fixed-effect linear model. Gene expression or transcript abundance was quantified and normalised as reads per kb per million aligned reads (RPKM) (Havlickova *et al.*, 2018).

Gene ontology (GO)

GO is a major bioinformatics initiative to unify the vocabulary of gene and gene product attributes across all species.

Genome-wide association study (GWAS)

Observational study that scans for markers across the genomes to find genetic variations in different individuals that associated with a particular trait.

Glucosinolate (GSL)

A group of amino-acid derived thioglycosidic secondary metabolites, found exclusively in the members of Brassicales order.

Heritability (h^2)

A statistic that estimates the proportion of variation in a phenotypic trait that is due to genetic variation between individuals in that population.

Homoeologous exchange

Exchanges of large segments of homoeologous chromosomes. This event is unique to allopolyploids.

Homoeologue

Pairs of genes or 'corresponding' genes derived from different species that were brought back together in the same genome by allopolyploidisation (Glover *et al.*, 2016)

Hub gene

A gene that is highly connected to other genes in the network (Horvath and Langfelder, 2011).

Linkage disequilibrium (LD)

The non-random correlation between alleles or polymorphisms (e.g. SNPs) in a population that is caused by their shared history of mutation and recombination (Flint-Garcia *et al*., 2003).

Lyophilisation	Also known as freeze-drying. A technique of dehydration which utilises low pressure and low temperature to induce sublimation of water from a material.
Minor allele frequencies (MAF)	The frequency at which the second most common allele occurs in a population.
Module	Clusters of highly interconnected genes (Horvath and Langfelder, 2011).
Neighbourhood connectivity	An average connectivity of all neighbouring nodes (Mpi-inf, 2018)
Node	A gene in a gene network.
Orthologue	Genes that are found in different species that evolved from a common ancestral gene by speciation.
PSIKO	A nonmodel-based population structure inference using kernel-PCA and optimisation (Popescu *et al*., 2014)
Quantitative trait loci (QTL)	Loci (genes) that correlate with variation of quantitative trait in the phenotype of a population of organisms (Miles and Wayne, 2008).
RPKM	Reads per kilobase of transcript per million mapped reads.
Single nucleotide polymorphism (SNP)	A variation at a single position in a DNA sequence among individuals. A variation can be classified as a SNP when more than 1% of a population carry a different nucleotide at a specific position in the DNA sequence.
Topological coefficient	A measure for the extent to which a node shares neighbours with other nodes (Mpi-inf, 2018)
WGCNA	Weighted correlation network analysis, a technique for studying biological networks based on pairwise correlations between variables.

References

Adams KL, Wendel JF (2015) Polyploidy and genome evolution in plants. Curr Opin Genet Dev **35**: 119–125

Alcock TD, Havlickova L, He Z, Bancroft I, White PJ, Broadley MR, Graham NS (2017) Identification of Candidate Genes for Calcium and Magnesium Accumulation in *Brassica napus* L. by Association Genetics . Front Plant Sci **8**: 1968

Alcock TD, Havlickova L, He Z, Wilson L, Bancroft I, White PJ, Broadley MR, Graham NS (2018) Species-Wide Variation in Shoot Nitrate Concentration, and Genetic Loci Controlling Nitrate, Phosphorus and Potassium Accumulation in *Brassica napus* L. . Front Plant Sci **9**: 1487

Alonso JM, Stepanova AN, Leisse TJ, Kim CJ, Chen H, Shinn P, Stevenson DK, Zimmerman J, Barajas P, Cheuk R, et al (2003) Genome-Wide Insertional Mutagenesis of *Arabidopsis thaliana*. Science (80-) **301**: 653 LP – 657

Andersen TG, Nour-Eldin HH, Fuller VL, Olsen CE, Burow M, Halkier BA (2013) Integration of biosynthesis and long-distance transport establish organ-specific glucosinolate profiles in vegetative *Arabidopsis*. Plant Cell **25**: 3133–45

Andréasson E, Bolt Jørgensen L, Höglund a S, Rask L, Meijer J (2001) Different myrosinase and idioblast distribution in *Arabidopsis* and *Brassica napus*. Plant Physiol **127**: 1750–1763

Arumuganathan K, Earle ED (1991) Nuclear DNA Content of Some Important Plant Species. Plant Mol Biol Report **9**: 208–218

Augustine R, Mukhopadhyay A, Bisht NC (2013) Targeted silencing of BjMYB28 transcription factor gene directs development of low glucosinolate lines in oilseed *Brassica juncea*. Plant Biotechnol J **11**: 855–866

Bancroft I, Fraser F, Morgan C, Trick M (2015) Collinearity analysis of Brassica A and C genomes based on an updated inferred unigene order. Data Br **3**: 51–55

Bancroft I, Morgan C, Fraser F, Higgins J, Wells R, Clissold L, Baker D, Long Y, Meng J, Wang X, et al (2011) Dissecting the genome of the polyploid crop oilseed rape by transcriptome sequencing. Nat Biotechnol **29**: 762–6

Becker T, Juvik J (2016) The Role of Glucosinolate Hydrolysis Products from Brassica Vegetable Consumption in Inducing Antioxidant Activity and Reducing Cancer Incidence. Diseases **4**: 22

Bekaert M, Edger PP, Hudson CM, Pires JC, Conant GC (2012) Metabolic and evolutionary costs of herbivory defense: Systems biology of glucosinolate synthesis. New Phytol **196**: 596–605

Bell L, Oloyede OO, Lignou S, Wagstaff C, Methven L (2018) Taste and Flavor Perceptions of Glucosinolates, Isothiocyanates, and Related Compounds. Mol Nutr Food Res. doi: 10.1002/mnfr.201700990

Benderoth M, Textor S, Windsor AJ, Mitchell-Olds T, Gershenzon J, Kroymann J (2006) Positive selection driving diversification in plant secondary metabolism. Proc Natl Acad Sci **103**: 9118–9123

Bhandari S, Jo J, Lee J (2015) Comparison of Glucosinolate Profiles in Different Tissues of Nine *Brassica* Crops. Molecules **20**: 15827–15841

Bolon Y-T, Stec AO, Michno J-M, Roessler J, Bhaskar PB, Ries L, Dobbels AA, Campbell BW, Young NP, Anderson JE, et al (2014) Genome resilience and prevalence of segmental duplications following fast neutron irradiation of soybean. Genetics **198**: 967–981

Borek V, Elberson LR, McCaffrey JP, Morra MJ (1998) Toxicity of Isothiocyanates Produced by Glucosinolates in Brassicaceae Species to Black Vine Weevil Eggs. J Agric Food Chem **46**: 5318–5323

Brett M (2016) Notes on the Bonferroni threshold. Github, https://matthew-brett.github.io/teaching/bonferroni_correction.html

Brown PD, Tokuhisa JG, Reichelt M, Gershenzon J (2003) Variation of glucosinolate accumulation among different organs and developmental stages of Arabidopsis thaliana. Phytochemistry **62**: 471–481

Brudenell AJP, Griffiths H, Rossiter JT, Baker DA (1999) The phloem mobility of glucosinolates. J Exp Bot **50**: 745–756

Buckler E, Zhang Z (2018) User Manual for Genomic Association and Prediction Integrated Tool (GAPIT).

Bus A, Körber N, Snowdon RJ, Stich B (2011) Patterns of molecular variation in a species-wide germplasm set of *Brassica napus*. Theor Appl Genet **123**: 1413–1423

CanolaCouncil (2017) What is Canola? Canola Counc. Canada, https://www.canolacouncil.org/oil-and-meal/what-is-canola/

Celenza JL (2005) The Arabidopsis ATR1 Myb Transcription Factor Controls Indolic Glucosinolate Homeostasis. PLANT Physiol **137**: 253–262

Chalhoub B, Denoeud F, Liu S, Parkin IAPP, Tang H, Wang XXX, Chiquet J, Belcram H, Tong C, Samans B, et al (2014) Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. Science (80-) **345**: 950–953

Chen S, Glawischnig E, Jørgensen K, Naur P, Jørgensen B, Olsen CE, Hansen CH, Rasmussen H, Pickett JA, Halkier BA (2003) CYP79F1 and CYP79F2 have distinct functions in the biosynthesis of aliphatic glucosinolates in *Arabidopsis*. Plant J **33**: 923–937

Chen S, Petersen BL, Olsen CE, Schulz A, Halkier BA (2001) Long-distance phloem transport of glucosinolates in *Arabidopsis*. Plant Physiol **127**: 194–201

Cheng Z, Sun L, Qi T, Zhang B, Peng W, Liu Y, Xie D (2011) The bHLH transcription factor MYC3 interacts with the jasmonate ZIM-domain proteins to mediate jasmonate response in *Arabidopsis*. Mol Plant **4**: 279–288

Clarke DB (2010) Glucosinolates, structures and analysis in food. Anal Methods **2**: 310–325

DEFRA (2018) Agriculture in the United Kingdom. Dep. Environ. Food Rural Aff.

Devlin B, Roeder K (1999) Genomic Control for Association Studies. Biometrics **55**: 997–1004

Doheny-Adams T, Redeker K, Kittipol V, Bancroft I, Hartley SE (2017) Development of an efficient glucosinolate extraction method. Plant Methods **13**: 17

Dörnemann D, Löffelhardt W, Kindl H (1974) Chain Elongation of Aromatic Amino Acids: the Role of 2-Benzylmalic Acid in the Biosynthesis of a C6C4 Amino Acid and a C6C3 Mustard Oil Glucoside. Can J Biochem **52**: 916–921

Edger PP, Heidel-Fischer HM, Bekaert M, Rota J, Glöckner G, Platts AE, Heckel DG, Der JP, Wafula EK, Tang M, et al (2015) The butterfly plant arms-race escalated by gene and genome duplications. Proc Natl Acad Sci U S A **112**: 8362–6

Fahey JW, Zalcmann AT, Talalay P (2001) The chemical diversity and distribution of glucosinolates and isothiocyanates among plants. Phytochemistry **56**: 5–51

Falk KL, Tokuhisa JG, Gershenzon J (2007) The effect of sulfur nutrition on plant glucosinolate content: Physiology and molecular mechanisms. Plant Biol **9**: 573–581

Fernández-Calvo P, Chini A, Fernández-Barbero G, Chico JM, Gimenez-Ibanez S, Geerinck J, Eeckhout D, Schweizer F, Godoy M, Franco-Zorrilla JM, et al (2011) The *Arabidopsis* bHLH transcription factors MYC3 and MYC4 are targets of JAZ repressors and act additively with MYC2 in the activation of jasmonate responses. Plant Cell **23**: 701–715

Fieldsen J, Milford GFJ (1994) Changes in glucosinolates during crop development in single- and double-low genotypes of winter oilseed rape (*Brassica napus*): I. Production and distribution in vegetative tissues and developing pods during development and potential role in the recycling. Ann Appl Biol **124**: 531–542

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of Linkage Disequilibrium in Plants. Annu Rev Plant Biol **54**: 357–374

Frerigmann H (2016) Glucosinolate Regulation in a Complex Relationship – MYC and MYB – No One Can Act Without Each Other. Adv Bot Res. doi: 10.1016/bs.abr.2016.06.005

Frerigmann H, Berger B, Gigolashvili T (2014) bHLH05 Is an Interaction Partner of MYB51 and a Novel Regulator of Glucosinolate Biosynthesis in Arabidopsis. Plant Physiol **166**: 349–369

Frerigmann H, Gigolashvili T (2014) MYB34, MYB51, and MYB122 distinctly regulate indolic glucosinolate biosynthesis in *Arabidopsis thaliana*. Mol Plant **7**: 814–828

Gachon CMM, Langlois-Meurinne M, Henry Y, Saindrenan P (2005) Transcriptional co-regulation of secondary metabolism enzymes in *Arabidopsis*: functional and evolutionary implications. Plant Mol Biol **58**: 229–245

Gajardo HA, Wittkop B, Soto-Cerda B, Higgins EE, Parkin IAP, Snowdon RJ, Federico ML, Iniguez-Luy FL (2015) Association mapping of seed quality traits in *Brassica napus* L. using GWAS and candidate QTL approaches. Mol Breed **35**: 1–19

Geu-Flores F, Nielsen MT, Nafisi M, Møldrup ME, Olsen CE, Motawia MS, Halkier BA (2009) Glucosinolate engineering identifies a gamma-glutamyl peptidase. Nat Chem Biol **5**: 575–577

Giamoustaris A, Mithen R (1995) The effect of modifying the glucosinolate content of leaves of oilseed rape (*Brassica napus* ssp. oleifera) on its interaction with specialist and generalist pests. Ann Appl Biol **126**: 347–363

Gigolashvili T, Berger B, Mock HP, Müller C, Weisshaar B, Flügge UI (2007a) The transcription factor HIG1/MYB51 regulates indolic glucosinolate biosynthesis in *Arabidopsis thaliana*. Plant J **50**: 886–901

Gigolashvili T, Engqvist M, Yatusevich R, Müller C, Flügge UI (2008) HAG2/MYB76 and HAG3/MYB29 exert a specific and coordinated control on the regulation of aliphatic glucosinolate biosynthesis in *Arabidopsis thaliana*. New Phytol **177**: 627–642

Gigolashvili T, Yatusevich R, Berger B, Müller C, Flügge UI (2007b) The R2R3-MYB transcription factor HAG1/MYB28 is a regulator of methionine-derived glucosinolate biosynthesis in *Arabidopsis thaliana*. Plant J **51**: 247–261

Gigolashvili T, Yatusevich R, Rollwitz I, Humphry M, Gershenzon J, Flugge U-I (2009) The Plastidic Bile Acid Transporter 5 Is Required for the Biosynthesis of Methionine-Derived Glucosinolates in *Arabidopsis thaliana*. Plant Cell Online **21**: 1813–1829

Glen DM, Jones H, Fieldsend JK (1990) Damage to oilseed rape seedlings by the field slug *Deroceras reticulatum* in relation to glucosinolate concentration. Ann Appl Biol **117**: 197–207

Glen S (2017) False Discovery Rate: Simple Definition, Adjusting for FDR. Stat. How To, https://www.statisticshowto.datasciencecentral.com/false-discovery-rate/

Glover NM, Redestig H, Dessimoz C (2016) Homoeologs: What Are They and How Do We Infer Them? Trends Plant Sci **21**: 609–621

Graser G, Schneider B, Oldham NJ, Gershenzon J (2000) The methionine chain elongation pathway in the biosynthesis of glucosinolates in *Eruca sativa* (Brassicaceae). Arch Biochem Biophys **378**: 411–419

Greenhalgh JR, Mitchell ND (1976) The involvement of flavour volatiles in the resistance to downy mildew of wild and cultivated forms of *Brassica oleracea*. New Phytol **77**: 391–398

Griffiths AJ, Gelbart WM, Miller JH, Lewontin RC (1999) Modern Genetic Analysis. W. H. Freeman, New York

Griffiths DW, Birch ANE, Hillman JR (1998) Antinutritional compounds in the Brassicaceae: Analysis, biosynthesis, chemistry and dietary effects. J Hortic Sci Biotechnol **73**: 1–18

Grubb CD, Abel S (2006) Glucosinolate metabolism and its control. Trends Plant Sci **11**: 89–100

Grubb CD, Zipp BJ, Ludwig-Müller J, Masuno MN, Molinski TF, Abel S (2004) *Arabidopsis* glucosyltransferase UGT74B1 functions in glucosinolate biosynthesis and auxin homeostasis. Plant J **40**: 893–908

Halkier BA, Gershenzon J (2006) Biology and Biochemistry of Glucosinolates. Annu Rev Plant Biol **57**: 303–333

Hansen BG, Kerwin RE, Ober JA, Lambrix VM, Mitchell-Olds T, Gershenzon J, Halkier BA, Kliebenstein DJ (2008) A novel 2-oxoacid-dependent dioxygenase involved in the formation of the goiterogenic 2-hydroxybut-3-enyl glucosinolate and generalist insect resistance in Arabidopsis. Plant Physiol **148**: 2096–2108

Hansen BG, Kliebenstein DJ, Halkier BA (2007) Identification of a flavin-monooxygenase as the S-oxygenating enzyme in aliphatic glucosinolate biosynthesis in *Arabidopsis*. Plant J **50**: 902–910

Harper AL, McKinney LV, Nielsen LR, Havlickova L, Li Y, Trick M, Fraser F, Wang L, Fellgett A, Sollars ESA, et al (2016) Molecular markers for tolerance of European ash (*Fraxinus excelsior*) to dieback disease identified using Associative Transcriptomics. Sci Rep **6**: 19335

Harper AL, Trick M, Higgins J, Fraser F, Clissold L, Wells R, Hattori C, Werner P, Bancroft I (2012) Associative transcriptomics of traits in the polyploid crop species *Brassica napus*. Nat Biotechnol **30**: 798–802

Hasan M, Seyis F, Badani AG, Pons-Kühnemann J, Friedt W, Lühs W, Snowdon RJ (2006) Analysis of genetic diversity in the *Brassica napus* L. gene pool using SSR markers. Genet Resour Crop Evol **53**: 793–802

Havlickova L, He Z, Wang L, Langer S, Harper AL, Kaur H, Broadley MR, Gegas V, Bancroft I (2018) Validation of an updated Associative Transcriptomics platform for the polyploid crop species *Brassica napus* by dissection of the genetic architecture of erucic acid and tocopherol isoform variation in seeds. Plant J **93**: 181–192

He Z, Cheng F, Li Y, Wang X, Parkin IAP, Chalhoub B, Liu S, Bancroft I (2015) Construction of Brassica A and C genome-based ordered pan-transcriptomes for use in rapeseed genomic research. Data Br **4**: 357–362

He Z, Wang L, Harper AL, Havlickova L, Pradhan AK, Parkin IAP, Bancroft I (2016) Extensive homoeologous genome exchanges in allopolyploid crops revealed by mRNAseq-based visualization. Plant Biotechnol J 1–11

Higgins J, Magusin A, Trick M, Fraser F, Bancroft I (2012) Use of mRNA-seq to discriminate contributions to the transcriptome from the constituent genomes of the polyploid crop species *Brassica napus*. BMC Genomics. doi: 10.1186/1471-2164-13-247

Hirai MY (2009) A robust omics-based approach for the identification of glucosinolate biosynthetic genes. Phytochem Rev **8**: 15–23

Hirai MY, Klein M, Fujikawa Y, Yano M, Goodenowe DB, Yamazaki Y, Kanaya S, Nakamura Y, Kitayama M, Suzuki H, et al (2005) Elucidation of Gene-to-Gene and Metabolite-to-Gene Networks in *Arabidopsis* by Integration of Metabolomics and Transcriptomics. J Biol Chem **280**: 25590–25595

Hirai MY, Sugiyama K, Sawada Y, Tohge T, Obayashi T, Suzuki A, Araki R, Sakurai N, Suzuki H, Aoki K, et al (2007) Omics-based identification of *Arabidopsis* Myb transcription factors regulating aliphatic glucosinolate biosynthesis. Proc Natl Acad Sci U S A **104**: 6478–6483

Hirani AH, Li G, Zelmer CD, McVetty PBE, Asif M, Goyal A (2012) Molecular Genetics of Glucosinolate Biosynthesis in Brassicas: Genetic. Manipulation and Application Aspects. Crop Plant 189–216

Hopkins RJ, van Dam NM, van Loon JJ a (2009) Role of glucosinolates in insect-plant relationships and multitrophic interactions. Annu Rev Entomol **54**: 57–83

Horvath S, Langfelder P (2011) Tutorials for the WGCNA package for R: WGCNA Background and glossary. 7–9

Hull AK, Vij R, Celenza JL (2000) *Arabidopsis* cytochrome P450s that catalyze the first step of tryptophan-dependent indole-3-acetic acid biosynthesis. Proc Natl Acad Sci **97**: 2379–2384

ISO 9167-1 (1992) Determination of glucosinolates content - Part 1: Method using high-performance liquid chromatography. Int. Stand.

Ji YK, Ibrahim KE, Juvik JA, Doo HK, Wha JK (2006) Genetic and environmental variation of glucosinolate content in Chinese cabbage. HortScience **41**: 1382–1385

Kirkegaard JA, Gardner PA, Angus JF, Koetz E (1994) Effect of *Brassica* break crops on the growth and yield of wheat. Aust J Agric Res **45**: 529–545

Kirkegaard JA, Sarwar M (1998) Biofumigation potential of brassicas: I. Variation in glucosinolate profiles of diverse field-grown brassicas. Plant Soil **201**: 71–89

Kittipol V, He Z, Wang L, Doheny-Adams T, Langer S, Bancroft I (2019a) Genetic architecture of glucosinolate variation in *Brassica napus*. J Plant Physiol **240**: 152988

Kittipol V, He Z, Wang L, Doheny-Adams T, Langer S, Bancroft I (2019b) Data in support of genetic architecture of glucosinolate variations in *Brassica napus*. Data Br **25**: 104402

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

Kliebenstein DJ, Kroymann J, Brown P, Figuth A, Pedersen D, Gershenzon J, Mitchell-Olds T (2001b) Genetic Control of Natural Variation in *Arabidopsis* Glucosinolate Accumulation. Plant Physiol. 126:

Kliebenstein DJ, Lambrix VM, Reichelt M, Gershenzon J, Mitchell-Olds T (2001c) Gene Duplication in the Diversification of Secondary Metabolism: Tandem 2-Oxoglutarate-Dependent Dioxygenases Control Glucosinolate Biosynthesis in *Arabidopsis*.

Knill T, Schuster J, Reichelt M, Gershenzon J, Binder S (2008) *Arabidopsis* Branched-Chain Aminotransferase 3 Functions in Both Amino Acid and Glucosinolate Biosynthesis. Plant Physiol **146**: 1028–1039

Koh JCO, Barbulescu DM, Norton S, Redden B, Salisbury PA, Kaur S, Cogan N, Slater AT (2017) A multiplex PCR for rapid identification of Brassica species in the triangle of U. Plant Methods **13**: 49

Kondra ZP, Stefansson BR (1970) Inheritance of the major glucosinolates of Rapesed (*Brassica Napus*) meal. Can J Plant Sci **50**: 643–647

Koroleva OA, Davies A, Hedrich R, Thorpe MR, Deeken R, Tomos AD (2000) Identification of a New Glucosinolate-Rich Cell Type in *Arabidopsis* Flower Stalk. Plant Physiol **124**: 599–608

Krishnakumar V, Contrino S, Cheng CY, Belyaeva I, Ferlanti ES, Miller JR, Vaughn MW, Micklem G, Town CD, Chan AP (2017) Thalemine: A warehouse for *Arabidopsis* data integration and discovery. Plant Cell Physiol. doi: 10.1093/pcp/pcw200

Kroymann J, Textor S, Tokuhisa JG, Falk KL, Bartram S, Gershenzon J, Mitchell-Olds T (2001) A Gene Controlling Variation in *Arabidopsis* Glucosinolate Composition Is Part of the Methionine Chain Elongation Pathway. PLANT Physiol **127**: 1077–1088

Laegdsmand M, Gimsing AL, Strobel BW, Sørensen JC, Jacobsen OH, Hansen HCB (2007) Leaching of isothiocyanates through intact soil following simulated biofumigation. Plant Soil **291**: 81–92

Lagercrantz U, Lydiate DJ (1996) Comparative genome mapping in *Brassica*. Genetics **144**: 1903–1910

Langer P, Greer MA (1977) Antithyroid Substances and Naturally Occurring Goitrogens.

Langfelder P, Horvath S (2008) WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics. doi: 10.1186/1471-2105-9-559

Langfelder P, Horvath S (2014a) Tutorial for the WGCNA package for R: I. Network analysis of liver expression data in female mice. 1. Data input and cleaning.

Langfelder P, Horvath S (2014b) Tutorial for the WGCNA package for R: I. Network analysis of liver expression data in female mice. 2.a Automatic network construction and module detection.

Langfelder P, Horvath S (2014c) Tutorial for the WGCNA package for R: I. Network analysis of liver expression data in female mice. 3. Relating modules to external information and identifying important genes.

Levy M, Wang Q, Kaspi R, Parrella MP, Abel S (2005) *Arabidopsis* IQD1, a novel calmodulin-binding nuclear protein, stimulates glucosinolate accumulation and plant defense. Plant J **43**: 79–96

Li F, Chen B, Xu K, Wu J, Song W, Bancroft I, Harper AL, Trick M, Liu S, Gao G, et al (2014) Genome-wide association study dissects the genetic architecture of seed weight and seed quality in rapeseed (*Brassica napus* L.). DNA Res **21**: 355–67

Li J, Hansen BG, Ober JA, Kliebenstein DJ, Halkier BA (2008) Subclade of Flavin-Monooxygenases Involved in Aliphatic Glucosinolate Biosynthesis. PLANT Physiol **148**: 1721–1733

Li QUN, Eigenbrode SD, Stringam GR (2000) Feeding and growth of *Plutella xylostella* and *Spodoptera eridania* on *Brassica juncea* with varying glucosinolate concentrations and myrosinase activities. J. Chem. Ecol. 26:

Li S, Zheng Y-C, Cui H-R, Fu H-W, Shu Q-Y, Huang J-Z (2016) Frequency and type of inheritable mutations induced by gamma rays in rice as revealed by whole genome sequencing. J Zhejiang Univ Sci B **17**: 905–915

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z (2012) GAPIT: Genome association and prediction integrated tool. Bioinformatics **28**: 2397–2399

Liu Z, Hirani AH, McVetty PBE, Daayf F, Quiros CF, Li G (2012) Reducing progoitrin and enriching glucoraphanin in *Brassica napus* seeds through silencing of the GSL-ALK gene family. Plant Mol Biol **79**: 179–189

Lu G, Harper AL, Trick M, Morgan C, Fraser F, O'Neill C, Bancroft I (2014) Associative Transcriptomics Study Dissects the Genetic Architecture of Seed Glucosinolate Content in *Brassica napus*. DNA Res **21**: 613–625

Ma XF, Gustafson JP (2005) Genome evolution of allopolyploids: A process of cytological and genetic diploidization. Cytogenet Genome Res **109**: 236–249

Magrath R, Mithen R (1993) Maternal Effects on the Expression of Individual Aliphatic Glucosinolates in Seeds and Seedlings of *Brassica napus*. Plant Breed **111**: 249–252

Malitsky S, Blum E, Less H, Venger I, Elbaz M, Morin S, Eshed Y, Aharoni A (2008) The Transcript and Metabolite Networks Affected by the Two Clades of *Arabidopsis* Glucosinolate Biosynthesis Regulators. PLANT Physiol **148**: 2021–2049

Martin N, Müller C (2007) Induction of plant responses by a sequestering insect: Relationship of glucosinolate concentration and myrosinase activity. Basic Appl Ecol **8**: 13–25

Mawson R, Heaney RK, Piskula M, Kozlowska H (1993) Rapeseed meal-glucosinolates and their antinutritional effects Part 1. Rapeseed production and chemistry of glucosinolates. Dic Nahrung **37**: 131–140

Mawson R, Heaney RK, Zdunczyk Z, Kozlowska H (1994) Rapeseed meal-glucosinolates and their antinutritional effects Part 4. Goitrogenicity and internal organs abnormalities in animals. Food / Nahrung **38**: 178–191

Mikkelsen MD, Naur P, Halkier BA (2004) Arabidopsis mutants in the C-S lyase of glucosinolate biosynthesis establish a critical role for indole-3-acetaldoxime in auxin homeostasis. Plant J **37**: 770–777

Miles CM, Wayne M (2008) Quantitative Trait Locus (QTL) Analysis. Nat. Educ.

Miller CN, Harper AL, Trick M, Wellner N, Werner P, Waldron KW, Bancroft I (2018) Dissecting the complex regulation of lodging resistance in *Brassica napus*. Mol Breed. doi: 10.1007/s11032-018-0781-6

Miller CN, Harper AL, Trick M, Werner P, Waldron K, Bancroft I (2016) Elucidation of the genetic basis of variation for stem strength characteristics in bread wheat by Associative Transcriptomics. BMC Genomics **17**: 500

Mithen R (1992) Leaf glucosinolate profiles and their relationships to pest and disease resistance in oilseed rape. Euphytica **63**: 71–83

Mithen RF, Lewis BG, Fenwick GR (1986) In vitro activity of glucosinolates and their products against Leptosphaeria maculans. Trans Br Mycol Soc **87**: 433–440

Mpi-inf (2018) NetworkAnalyzer Online Help. Max Planck Inst. Informatics, https://med.bioinf.mpi-inf.mpg.de/netanalyzer/help/2.5/

Murashige T, Skoog F (1962) A Revised Medium for Rapid Growth and Bio Assays with Tobacco Tissue Cultures. Physiol Plant **15**: 473–497

Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich DE, Buckler ES (2009) Association Mapping: Critical Considerations Shift from Genotyping to Experimental Design. PLANT CELL ONLINE **21**: 2194–2202

Naur P, Petersen BL, Mikkelsen MD, Bak S, Rasmussen H, Olsen CE, Halkier BA (2003) CYP83A1 and CYP83B1, Two Nonredundant Cytochrome P450 Enzymes Metabolizing Oximes in the Biosynthesis of Glucosinolates in *Arabidopsis*. PLANT Physiol **133**: 63–72

Nour-Eldin HH, Andersen TG, Burow M, Madsen SR, Jørgensen ME, Olsen CE, Dreyer I, Hedrich R, Geiger D, Halkier BA (2012) NRT/PTR transporters are essential for translocation of glucosinolate defence compounds to seeds. Nature **488**: 531–4

Nour-Eldin HH, Halkier BA (2009) Piecing together the transport pathway of aliphatic glucosinolates. Phytochem Rev **8**: 53–67

Nour-Eldin HH, Madsen SR, Engelen S, Jørgensen ME, Olsen CE, Andersen JS, Seynnaeve D, Verhoye T, Fulawka R, Denolf P, et al (2017) Reduction of antinutritional glucosinolates in *Brassica* oilseeds by mutation of genes encoding transporters. Nat Biotechnol **35**: 377–382

O'Neill CM, Bancroft I (2000) Comparative physical mapping of segments of the genome of *Brassica oleracea* var. alboglabra that are homoeologous to sequenced regions of chromosomes 4 and 5 of Arabidopsis thaliana. Plant J **23**: 233–243

Park M-H, Arasu MV, Park N-Y. N-Y, Choi Y-JY-J., Lee S-W. S-W, Al-Dhabi NANA., Kim JBJB., Kim S-J. S-J, Valan Arasu M., Park N-Y. N-Y, et al (2013) Variation of glucoraphanin and glucobrassicin: Anticancer components in *Brassica* during processing. Food Sci Technol **33**: 624–631

Parkin IA, Lydiate DJ, Trick M (2002) Assessing the level of collinearity between *Arabidopsis thaliana* and *Brassica napus* for A. thaliana chromosome 5. Genome **45**: 356–366

Parkin IAP, Gulden SM, Sharpe AG, Lukens L, Trick M, Osborn TC, Lydiate DJ (2005) Segmental structure of the *Brassica napus* genome based on comparative analysis with *Arabidopsis thaliana*. Genetics **171**: 765–781

Petersen BL, Chen S, Hansen CH, Olsen CE, Halkier BA (2002) Composition and content of glucosinolates in developing *Arabidopsis thaliana*. Planta **214**: 562–571

Pfalz M, Mikkelsen MD, Kroymann J, Halkier BA, Bednarek P, Olsen CE (2011) Metabolic Engineering in Nicotiana benthamiana Reveals Key Enzyme Functions in *Arabidopsis* Indole Glucosinolate Modification. Plant Cell **23**: 716–729

Pfalz M, Mukhaimar M, Perreau F, Kirk J, Hansen CIC, Olsen CE, Agerbirk N, Kroymann J (2016) Methyl Transfer in Glucosinolate Biosynthesis Mediated by Indole Glucosinolate O - Methyltransferase 5. Plant Physiol **172**: 2190–2203

Pfalz M, Vogel H, Kroymann J (2009) The Gene Controlling the Indole Glucosinolate Modifier1 Quantitative Trait Locus Alters Indole Glucosinolate Structures and Aphid Resistance in *Arabidopsis*. PLANT CELL ONLINE **21**: 985–999

Piotrowski M, Schemenewitz A, Lopukhina A, Mu A, Janowitz T, Weiler EW, Oecking C (2004) Desulfoglucosinolate Sulfotransferases from *Arabidopsis thaliana* Catalyze the Final Step in the Biosynthesis of the Glucosinolate Core Structure. J Biol Chem **279**: 50717–50725

Popescu AA, Harper AL, Trick M, Bancroft I, Huber KT (2014) A novel and fast approach for population structure inference using Kernel-PCA and optimization. Genetics **198**: 1421–1431

Porter AJR, Morton AM, Kiddle G, Doughty KJ, Wallsgrove RM (1991) Variation in the glucosinolate content of oilseed rape (*Brassica napus* L.) leaves: I. Effect of leaf age and position. Ann Appl Biol **118**: 461–468

Potter MJ, Vanstone VA, Davies KA, Rathjen AJ (2000) Breeding to increase the concentration of 2-phenylethyl glucosinolate in the roots of *Brassica napus*. J Chem Ecol **26**: 1811–1820

Priyam A, Woodcroft BJ, Rai V, Munagala A, Moghul I, Ter F, Gibbins MA, Moon H, Leonard G, Rumpf W, et al (2015) Sequenceserver: a modern graphical user interface for custom BLAST databases. Biorxiv 1–18

R core team (2013) R: a language and environment for statistical computing. doi: 10.1007/978-3-540-74686-7

Radojcic Redovnikovic I, Glivetic T, Delonga K, Vorkapic-Furac J (2008) Glucosinolates and their potential role in plant. Period Biol **110**: 297–309

Rana D, Van Den Boogaart T, O'Neill CM, Hynes L, Bent E, Macpherson L, Jee YP, Yong PL, Bancroft I (2004) Conservation of the microstructure of genome segments in *Brassica napus* and its diploid relatives. Plant J **40**: 725–733

Rask L, Andréasson E, Ekbom B, Eriksson S, Pontoppidan B, Meijer J (2000) Myrosinase: Gene family evolution and herbivore defense in Brassicaceae. Plant Mol Biol **42**: 93–113

Reintanz B, Lehnen M, Reichelt M, Gershenzon J, Kowalczyk M, Sandberg G, Godde M, Uhl R, Palme K (2001) Bus, a bushy *Arabidopsis* CYP79F1 knockout mutant with abolished synthesis of short-chain aliphatic glucosinolates. Plant Cell **13**: 351–367

Robinson MD, McCarthy DJ, Smyth GK (2009) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**: 139–140

Rosa EAS, Heaney RK, Fenwick GR, Portas CAM (1997) Glucosinolates in crop plants. Hortic Rev (Am Soc Hortic Sci). doi: 10.1002/9780470650622.ch3

Rucker B, Rudloff E (1991) Investigations of the inheritance of the glucosinolate content in seeds of winter oilseed rape (*Brassica napus* L.). Proc 8th Int Rapeseed Congr Saskatoon, Canada 191–196

Sarwar M, J.A.Kirkegaard, Wong PTW, Desmarchelier JM (1998) Biofumigation potential of brassicas: III. In vitro toxicity of isothiocyanates to soil-borne fungal pathogens. Plant Soil **201**: 71–89

Schlaeppi K, Bodenhausen N, Buchala A, Mauch F, Reymond P (2008) The glutathione-deficient mutant pad2-1 accumulates lower amounts of glucosinolates and is more susceptible to the insect herbivore Spodoptera littoralis. Plant J **55**: 774–786

Schonhof I, Blankenburg D, Müller S, Krumbein A (2007) Sulfur and nitrogen supply influence growth, product appearance, and glucosinolate concentration of broccoli. J Plant Nutr Soil Sci **170**: 65–72

Schweizer F, Fernández-Calvo P, Zander M, Diez-Diaz M, Fonseca S, Glauser G, Lewsey MG, Ecker JR, Solano R, Reymond P (2013) *Arabidopsis* basic helix-loop-helix transcription factors MYC2,MYC3, andMYC4 regulate glucosinolate biosynthesis, insect performance, and feeding behavior. Plant Cell **25**: 3117–3132

Shannon P, Markiel A, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang DR, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res **13**: 6

Sonderby IE, Burow M, Rowe HC, Kliebenstein DJ, Halkier BA (2010) A Complex Interplay of Three R2R3 MYB Transcription Factors Determines the Profile of Aliphatic Glucosinolates in *Arabidopsis*. PLANT Physiol **153**: 348–363

Sønderby IE, Geu-Flores F, Halkier BA (2010) Biosynthesis of glucosinolates – gene discovery and beyond. Trends Plant Sci **15**: 283–290

Sønderby IE, Hansen BG, Bjarnholt N, Ticconi C, Halkier BA, Kliebenstein DJ (2007) A systems biology approach identifies a R2R3 MYB gene subfamily with distinct and overlapping functions in regulation of aliphatic glucosinolates. PLoS One **2**: 1322

Tayo T, Dutta N, Sharma K (2012) Effect of Feeding Canola Quality Rapeseed Mustard Meal on Animal Production - a Review. Agric Rev **33**: 114–121

Textor S, de Kraker J-W, Hause B, Gershenzon J, Tokuhisa JG (2007) MAM3 Catalyzes the Formation of All Aliphatic Glucosinolate Chain Lengths in *Arabidopsis*. PLANT Physiol **144**: 60–71

Tian T, Liu Y, Yan H, You Q, Yi X, Du Z, Xu W, Su Z (2017) AgriGO v2.0: A GO analysis toolkit for the agricultural community, 2017 update. Nucleic Acids Res **45**: W122–W129

Tokuhisa J, de Kraker JW, Textor S, Gershenzon J (2004) The biochemical and molecular origins of aliphatic glucosinolate diversity in *Arabidopsis thaliana*. Recent Adv. Phytochem. Elsevier, pp 19–38

Trick M, Long Y, Meng J, Bancroft I (2009) Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using Solexa transcriptome sequencing. Plant Biotechnol J **7**: 334–346

U N (1935) Genome analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. Japanese J Bot **7**: 389–452

USDA (2019) Oilseeds: World Markets and Trade.

Velasco P, Äa M, Cartea E, Gonza C, Lez Ä, Vilar M, Orda A (2007) Factors Affecting the Glucosinolate Content of Kale (*Brassica oleracea* acephala Group). doi: 10.1021/jf0624897

Velasco P, Soengas P, Vilar M, Cartea ME, Rio M Del (2008) Comparison of Glucosinolate Profiles in Leaf and Seed Tissues of Different *Brassica napus* Crops. J Am Soc Hortic Sci **133**: 551–558

Verhoeven DTH, Goldbohm RA, Poppel G Van, Verhagen H, Brandt PA van den (1996) Epidemiological Studies on *Brassica* Vegetables and Cancer Risk. Cancer Epidemiol Biomarkers Prev **5**: 733–748

Vlab.amrita.edu (2012) Analysis of biological networks for feature detection. http://vlab.amrita.edu/?sub=3&brch=276&sim=1475&cnt=6

Wathelet J-P, Iori R, Leoni O, Rollin P, Quinsac A, Palmieri S (2004) Guidelines for glucosinolate analysis in green tissues used for biofumigation. Agroindustria 3:

Winter D, Vinegar B, Nahal H, Ammar R, Wilson G V., Provart NJ (2007) An ‘electronic fluorescent pictograph’ Browser for exploring and analyzing large-scale biological data sets. PLoS One **2**: 1–12

Wittstock U, Halkier BA (2002) Glucosinolate research in the *Arabidopsis* era. Trends Plant Sci **7**: 263–270

Wittstock U, Halkier BA (2000) Cytochrome P450 CYP79A2 from *Arabidopsis thaliana* L . catalyzes the conversion of L -phenylalanine to phenylacetaldoxime in the biosynthesis of benzylglucosinolate. J Biol chemstry **275**: 14659–14666

Xu L, Hu K, Zhang Z, Guan C, Chen S, Hua W, Li J, Wen J, Yi B, Shen J, et al (2015) Genome-wide association study reveals the genetic architecture of flowering time in rapeseed (*Brassica napus* L.). DNA Res **23**: 43–52

Yu J, Buckler ES (2006) Genetic association mapping and genome organization of maize. Curr Opin Biotechnol **17**: 155–160

Zhu C, Gore M, Buckler ES, Yu J (2008) Status and Prospects of Association Mapping in Plants. Plant Genome J **1**: 5

Zimmermann IM, Heim MA, Weisshaar B, Uhrig JF (2004) Comprehensive identification of *Arabidopsis thaliana* MYB transcription factors interacting with R/B-like BHLH proteins. Plant J **40**: 22–34

Publications arising from this work

https://doi.org/10.1016/j.jplph.2019.06.001

Journal of Plant Physiology 240 (2019) 152988

Contents lists available at ScienceDirect

Journal of Plant Physiology

journal homepage: www.elsevier.com/locate/jplph

Genetic architecture of glucosinolate variation in *Brassica napus*

Varanya Kittipol, Zhesi He, Lihong Wang, Tim Doheny-Adams, Swen Langer, Ian Bancroft*

Department of Biology, University of York, Heslington, York, YO10 5DD, UK

ARTICLE INFO

Keywords:
Glucosinolates
Brassica napus
HAG1
HAG3

ABSTRACT

The diverse biological activities of glucosinolate (GSL) hydrolysis products play significant biological and economical roles in the defense system and nutritional qualities of *Brassica napus* (oilseed rape). Yet, genomic-based study of the *B. napus* GSL regulatory mechanisms are scarce due to the complexity of working with polyploid species. To address these challenges, we used transcriptome-based GWAS approach, Associative Transcriptomics (AT), across a diversity panel of 288 *B. napus* genotypes to uncover the underlying genetic basis controlling quantitative variation of GSLs in *B. napus* vegetative tissues. Single nucleotide polymorphism (SNP) markers and gene expression markers (GEMs) associations identify orthologues of *MYB28/HAG1* (AT5G61420), specifically the copies on chromosome A9 and C2, to be the key regulators of aliphatic GSL variation in leaves. We show that the positive correlation observed between aliphatic GSLs in seed and leaf is due to the amount synthesized, as controlled by *Bna.HAG1.A9 and Bna.HAG1.C2*, rather than by variation in the transport processes. In addition, AT and differential expression analysis in root tissues implicate an orthologue of *MYB29/HAG3* (AT5G07690), *Bna.HAG3.A3*, as controlling root aromatic GSL variation. Based on the root expression data we also propose *Bna.MAM3.A3* to have a role in controlling phenylalanine chain elongation for aromatic GSL biosynthesis. This work uncovers a regulator of homophenylalanine-derived aromatic GSLs and implicates the shared biosynthetic pathways between aliphatic and aromatic GSLs.

1. Introduction

Glucosinolates (GSLs) are a group of sulfur- and nitrogen-rich secondary metabolites prevalent in Brassicales (Halkier and Gershenzon, 2006). GSLs are economically significant because their bioactive hydrolysates have diverse biological properties that impact agriculturally important *Brassica* crops such as oilseed rape (*Brassica napus* L.) and have been studied extensively in the model plant *Arabidopsis thaliana*. Depending on the reaction conditions and GSL side-chain structure, bioactive hydrolysates such as isothiocyanates, nitriles and oxazolidine-2-thione are produced when myrosinase enzymes came into contact with GSLs after tissue damage (Rask et al., 2000; Wittstock and Halkier, 2002). Some GSLs and their hydrolysis products are thought to defend the plants against non-adapted pathogen and insect pests (Glen et al., 1990; Potter et al., 2000; Hopkins et al., 2009), while other isothiocyanates are suitable as biofumigants to control soil pests and weeds (Gimsing and Kirkegaard, 2009). However, other GSLs have negative impacts. For example, progoitrin can accumulate to high concentrations in seeds. When these are hydrolyzed, it produces goitrogenic products that reduce the nutritional values of the protein-rich seed meal used as livestock feed (Griffiths et al., 1998; Tayo et al., 2012). To allow the use of seed meal as animal feed, extensive breeding efforts have been made to select for oilseed rape cultivars with low seed GSLs (< 30 μmol/g) (Rosa et al., 1997). On the other hand, the introduction of '00' (low seed erucic and GSL) cultivars, has led to the concern that these cultivars could be more susceptible to pests and diseases due to reduction of the presumed defensive role of GSL. Nevertheless, levels of GSLs and their interaction with plant pests may be more intricate than previously thought because the same GSL profile can acts as both deterrent to generalist pests and stimulant to specialist pests (Mithen, 1992; Giamoustaris and Mithen, 1995; Hopkins et al., 2009). Some studies have reported no significant correlation of GSL between seeds and leaves, suggesting that modifying the GSL profiles selectively in different parts of the plant may be feasible (Porter et al., 1991; Fieldsen and Milford, 1994). However, the underlying genetic control of quantitative variations of GSL in vegetative tissues and seeds of *B. napus*, and their interaction, are not well understood.

Based on their amino acid precursor of the side chain, GSLs are divided into three structural groups: aliphatic, indole and aromatic GSLs, which derived from methionine, tryptophan and phenylalanine respectively (Fahey et al., 2001). The biosynthetic pathway of GSLs proceeds in three stages via (i) amino acid side chain elongation; (ii) the

[*] Corresponding author.
E-mail address: ian.bancroft@york.ac.uk (I. Bancroft).

https://doi.org/10.1016/j.jplph.2019.06.001
Received 18 March 2019; Received in revised form 30 May 2019; Accepted 3 June 2019
Available online 11 June 2019

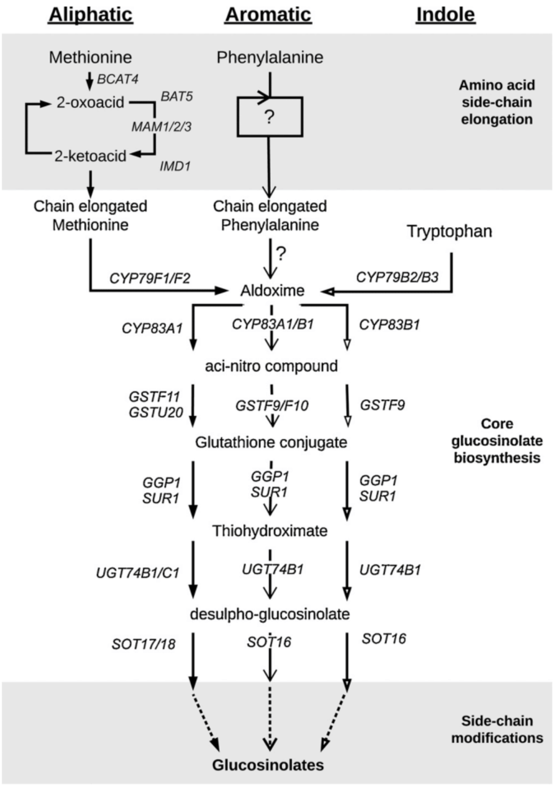


Fig. 1. Simplified aliphatic, aromatic and indole glucosinolate biosynthesis pathways in Brassicaceae, comprising of three stages: amino acid side chain elongation, core moiety biosynthesis and extensive side chain modifications. The enzymes involving in phenylalanine chain elongation and catalyzing the subsequent homophenylalanine are unknown.

amino acid moiety undergoing metabolic configurations to form the core GSL structure; and (iii) secondary modifications of the side chain to generate a wide spectrum of GSL compounds (Fig. 1). Many of the genes responsible for biosynthetic steps have been identified in *Arabidopsis thaliana* (reviewed in Grubb and Abel, 2006; Halkier and Gershenzon, 2006; Sønderby et al., 2010), which has also helped clarify the core biosynthesis steps and identify orthologous genes in the closely related *Brassica* species. A group of R2R3 MYB transcription factors from a single gene family within *Arabidopsis* is known to be involved in the direct transcriptional regulation of GSLs biosynthesis. *MYB34/ATR1*, *MYB51/HIG1*, and *MYB122/HIG2* are thought to regulate the tryptophan-derived indole GSL pathway (Celenza, 2005; Gigolashvili et al., 2007a; Frerigmann and Gigolashvili, 2014), and *MYB28/HAG1, MYB29/HAG3* and *MYB76/HAG2* regulate the methionine-derived aliphatic GSL biosynthetic genes (Gigolashvili et al., 2007b; Hirai et al., 2007; Gigolashvili et al., 2008; Sonderby et al., 2010). Since methionine-derived aliphatic and tryptophan-derived indole GSLs are the two main classes of GSLs found in *A. thaliana* (Brown et al., 2003), significant progress has been made in understanding the biochemistry and the regulatory controls of these two classes of GSLs. However, less information is available for the chain-elongated homophenylalanine-derived aromatic GSL, which is abundant in *Brassica* species (Bhandari et al., 2015). So far, the genes involved in the side chain elongation and the regulatory genes controlling aromatic GSL biosynthesis remain largely uncharacterized. Furthermore, the CYP79A2 that catalyzes phenylalanine substrates has been shown unable to metabolize homophenylalanine into aldoxime (Wittstock and Halkier, 2000), suggesting the enzyme that controls the flux into the biosynthetic pathway of homophenylalanine-derived aromatic GSLs in *B. napus* is yet to be identified.

While some of the natural variation in GSL profiles can be explained by allelic variation of key biosynthetic genes, other differences are likely to be caused by the activity of regulatory loci (Kliebenstein et al., 2001a). Genome-wide association studies (GWAS) provides a powerful method of using genetically diverse population to identify quantitative trait loci (QTLs) at higher resolution by exploiting historical recombination between molecular markers and loci associated with trait variation (Zhu et al., 2008). With the focus on seed quality traits, association studies had been effectively applied to identify clusters of single-nucleotide polymorphisms (SNPs) highly associated with seed GSL content in *B. napus* in recent years (Li et al., 2014; Lu et al., 2014; Gajardo et al., 2015). Nevertheless, to get better understanding of the modular genetic system that regulates GSL natural variations in *B. napus* as a whole, more work is needed to investigate the regulations of GSL in the vegetative tissues and how these variations relate to the GSL profiles in the seed.

In this study we aimed to elucidate the genetic control of GSL biosynthesis in leaves and roots of *B. napus*. We took the approach of firstly undertaking a transcriptome-based GWAS approach. Such a genomics approach was feasible because of the availability of the recently-established full-scale Associative Transcriptomics (AT) platform comprising 355,536 SNP markers and transcriptome reference comprising 116,098 ordered coding DNA sequence gene models (Havlickova et al., 2018). We could deploy this for a large panel of 288 *B. napus* accessions because of the availability of a recently-developed simple and efficient GSL extraction method (Doheny-Adams et al., 2017).

2. Results

2.1. *Glucosinolates identified in B. napus leaves and roots*

A subset of 288 diverse *B. napus* accessions with defined crop types of the RIPR panel (Renewable Industrial Products from Rapeseed) (Havlickova et al., 2018) was analyzed for GSL compositions in the leaves and roots of 4-week old plants. Fourteen different GSLs were identified. Out of these, nine are classed as aliphatic (including C_3, C_4 and C_5 types), four indole and one aromatic GSL (Table 1). Detailed profiles are provided in Appendix 1 of Kittipol et al. (2019). To identify relationships between GSL content of leaves and roots, we performed a Spearman's correlation analysis (Table 2). Within leaf, the total amount of GSL accumulated in the tissue is determined largely by the level of

Table 1
Glucosinolates identified in this study.

Type	Trivial name	Acronym	Systematic R Side chain
Aliphatic C_3	Glucoputranjivin	GJV	1-Methylethyl
Aliphatic C_4	Gluconapin	GNA	3-Butenyl
	Progoitrin	PRO	(2R)-2-Hydroxy-3-butenyl
	Glucoerucin	GER	4-Methylthiobutyl
	Glucoraphanin	GRA	4-Methylsulfinylbutyl
	Glucoraphenin	GRE	4-Methylsulfinyl-3-butenyl
Aliphatic C_5	Glucoalyssin	GAL	5-Methylsulfinylpentryl
	Glucobrassicanapin	GBN	Pent-4-enyl
	Gluconapoleiferin	GNL	2-Hydroxy-pent-4-enyl
Indole	Glucobrassicin	GBS	3-Indolylmethyl
	4-Hydroxyglucobrassicin	4-OHGBS	4-Hydroxy-3-indolylmethyl
	4-Methoxyglucobrassicin	4-OMeGBS	4-Methoxy-3-indolylmethyl
	Neoglucobrassicin	neo-GBS	N-Methoxy-3-indolylmethyl
Aromatic	Gluconasturtiin	GST	2-Phenethyl

Table 2
Spearman's correlation coefficient analysis of glucosinolate traits.

	TL	L-ali	L-ind	L-aro	TR	R-ali	R-ind	R-aro
Total Leaf (TL)	–							
Leaf Aliphatic (L-ali)	0.91***	–						
Leaf Indole (L-ind)	0.45***	0.14*	–					
Leaf Aromatic (L-aro)	0.62***	0.62***	0.12*	–				
Total Root (TR)	0.28***	0.30***	0.00	0.37***	–			
Root Aliphatic (R-ali)	0.64***	0.68***	0.10	0.50***	0.43***	–		
Root Indole (R-ind)	0.01	–0.10	0.24***	−0.15*	0.41***	−0.04	–	
Root Aromatic (R-aro)	0.18**	0.29***	–0.22***	0.46***	0.75***	0.30***	–0.18**	–
†Total Seed GSL	0.48***	0.54***	0.00	0.40***	0.02	0.43***	–0.20*	0.09

Correlation of mean trait values from 288 accessions of the diversity panel. Significant correlations are indicated.

*** P ≤ 0.001.

** P ≤ 0.01.

* P ≤ 0.05.

† Data for total seed glucosinolates for 151 *B. napus* accessions came from Lu et al. (2014).

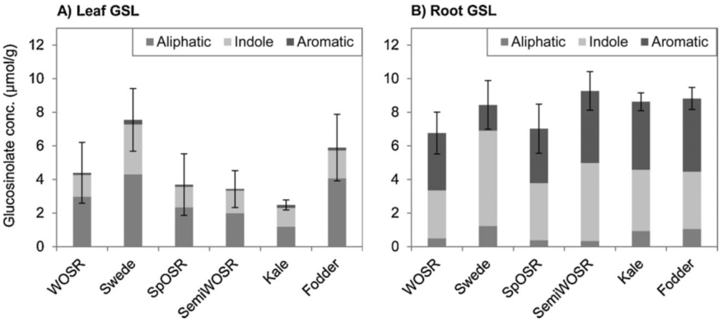


Fig. 2. Glucosinolate variations in *B. napus*. Means of glucosinolate (GSL) content in (A) leaf and (B) root of 288 *B. napus* accessions grouped into six crop types. Individual GSLs were grouped according to their structural classes as aliphatic, indole and aromatic GSLs. Abbreviation: spring oilseed rape (SpOSR), semi-winter oilseed rape (SemiWOSR), winter oilseed rape (WOSR), winter fodder (fodder). Error bars represent standard deviations of total GSL.

leaf aliphatic GSL (r = 0.91 ***). While both indole and aromatic GSLs are the major GSL classes found in roots, aromatic GSL (i.e. GST) provides a much stronger indication of the total amount of root GSLs (r = 0.75 ***) than root indole GSL (r = 0.41 ***). Significant positive correlations were observed between aliphatic and aromatic GSLs within the same tissue (Leaf: r = 0.62 ***, Root: r = 0.30 ***), as well as between leaf and root (r = 0.50 ***, 0.29 ***), suggest the possibility of co-regulation that is shared between these two classes of GSLs. Whereas, the weak and negative correlations between indole and aromatic GSL within root (r = –0.18 **) and between root and leaf tissues (r = –0.15*, –0.22***) indicate antagonistic relationship between this two GSL classes. Given that different GSL profiles were found between aliphatic-dominated leaf and indole/aromatic-dominated root (Fig. 2), the GSL metabolic pathways between above- and below- ground tissues appears to be regulated differentially yet has some cross-talk between the pathways, which is supported by the weak but significant correlation between total GSLs in the leaf and root (r = 0.28 ***).

2.2. Genetic control of leaf glucosinolate variation

Extensive phenotypic variation was observed in leaves for both amount and type of GSLs. The total GSL content ranged from 0.26 to 21.6 µmol/g in leaves, with aliphatic GSLs as the predominant class (64.0% of all leaf GSLs), indole GSLs contributing (32.9%) and a small amount of the aromatic GSL, GST (3.1%). When the *B. napus* diversity panel was assessed by crop type, accessions of the swede crop type were found to contain the greatest amount of GSL (Appendix 3 in Kittipol et al., 2019), with modern winter and spring oilseed rape crop types having the lowest GSL content.

To understand the genetic control of this observed variation we used the established *B. napus* Associative Transcriptomics (AT) platform consisting of 355,536 SNP markers and gene expression matrix with a transcriptome reference of 116,098 ordered coding DNA sequence gene models (Havlickova et al., 2018) to identify molecular marker variation associated with trait variation. As visualized using "Manhattan Plots", clusters of markers with allelic variation correlated with trait variation indicates regions of the genome containing genes controlling the traits. We undertook AT analysis on all individual GSLs, total GSL and GSLs grouped by type (aliphatic, indole or aromatic). The Manhattan plots are shown in Appendix 5 of Kittipol et al. (2019). As illustrated in Fig. 3, associations for aliphatic GSL content exceeding the Bonferroni-corrected 5% significance threshold with SNP markers were observed in regions of chromosomes A2, A9, C2 and C9. A fifth region exceeding the 5% FDR threshold (but not the Bonferroni-corrected 5% significance threshold) was identified on chromosome C7. These five genomic regions had previously been observed in an AT analysis of total seed GSL content (Lu et al., 2014), suggesting that leaf and seed GSL content are both controlled by the same loci. Investigation of the genes underlying the positions of these five association peaks, as shown in Appendix 9 of Kittipol et al. (2019), revealed at every one an orthologue of HAG1 (AT5G61420), a transcription factor that positively regulated aliphatic GSL biosynthesis. In addition, of the associations between Gene Expression Markers (GEM) and leaf aliphatic GSL content that exceeding the 5% FDR threshold six were detected for genes involved directly in aliphatic GSL biosynthesis (Appendix 11 in Kittipol et al., 2019). Two of these genes are known to be involved in the aliphatic amino acid chain elongation, an orthologue of AT5G23020, a methythioalkymalate synthase (MAM3) was found on A3 and an orthologue of AT5G23010, MAM1, on C7. Two genes involved in the core GSL structure biosynthesis, an orthologue of AT1G16410, a cytochrome P450 CYP79F1 and an orthologue of AT1G78370, a glutathione S-transferase TAU 20 (GSTU20) were identified on chromosome C5 and A7 respectively. Two orthologues of HAG1, *Bna.*HAG1.*A9* and *Bna.*HAG1.*C2*, were also identified amongst the top GEMs, implicating the transcript abundance levels of these genes in the control of aliphatic GSL in the leaf. To test this, we analyzed leaf transcript abundance on four biological replicates

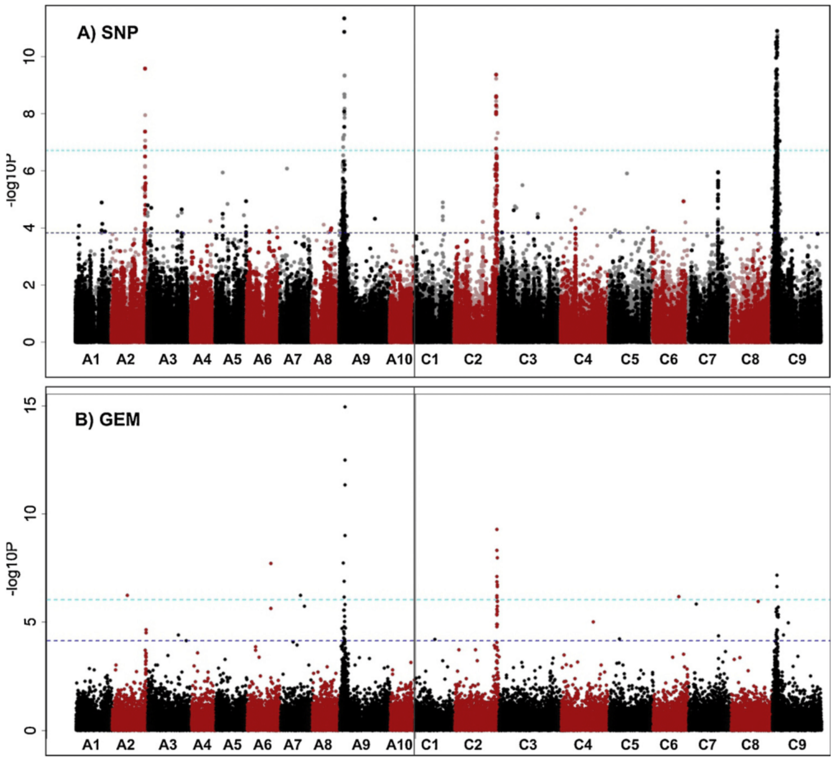


Fig. 3. Association analysis for leaf aliphatic glucosinolate content. (A) Manhattan plot showing genome-wide associations for the identification of transcriptome single-nucleotide polymorphism (SNP) markers of 288 *Brassica napus* accessions with leaf glucosinolate content. Marker associations was calculated using a mixed linear model which incorporated population structure and relatedness. The SNP markers are positioned on the x-axis based on the genomic order of the gene models in which the polymorphism was scored. The significance of the trait association, as -log10 P values, plotted on the y-axis. The horizontal purple and cyan lines represent false discovery rate (FDR) threshold at 5% and the threshold for Bonferroni significance of 0.05, respectively. Chromosomes of *B. napus* are labelled A1– A10 and C1 – C9, shown in alternating black and red colors to allow boundaries to be clearly distinguished. Dark opaque points are simple SNP markers (i.e. polymorphisms between resolved bases) and hemi-SNPs that have been directly linkage-mapped, both of which can be assigned to one genome, whereas light points are hemi-SNP markers (i.e. polymorphisms involving multiple bases called at the SNP position in one allele of the polymorphism) for which the genome of the polymorphism cannot be assigned. (B) Association analysis of expression variation-based markers (GEM) with leaf aliphatic glucosinolate. Reads per kb per million aligned reads (RPKM) were regressed against the trait, and R^2 and P values were calculated for each gene. The gene models are positioned on the x-axis based on their genomic order, with the significance of the associated trait, as -log10 P, plotted on the y-axis. The horizontal purple and cyan lines represent false discovery rate (FDR) threshold at 5% and the threshold for Bonferroni significance of 0.05, respectively (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

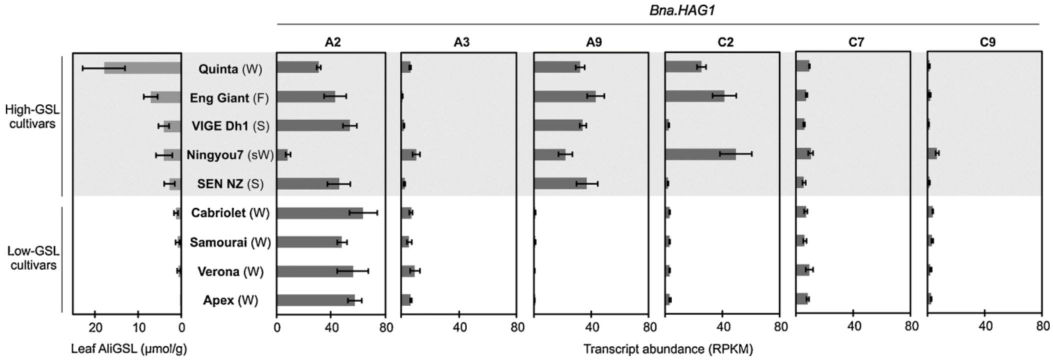


Fig. 4. Expression of *Bna.*HAG1 homoeologues in high- and low- leaf aliphatic GSL *B. napus* cultivars. Six orthologues of HAG1 (AT5G61420) are found in *B. napus*, on chromosome A2, A3, A9, C2, C7 and C9. Transcript abundance of *Bna.*HAG1 is expressed as reads per kb per million aligned reads (RPKM), with error bars to indicate standard deviation from four biological replicates of each accessions. Crop type abbreviation: (W), Winter oilseed rape; (F), Winter fodder; (sW), Semiwinter oilseed rape; (S), Swede.

for all six HAG1 orthologues in 5 high leaf GSL and 4 low leaf GSL *B. napus* accessions. Consistent with the AT results, as shown in Fig. 4, expression of *Bna.*HAG1.*A9* and *Bna.*HAG1.*C2* showed strong positive correlation with level of aliphatic GSL in leaves, whereas the orthologues on A3, C7 and C9 were expressed at relatively low levels. The remaining orthologue, on chromosome A2 was relatively highly expressed in all accessions so this copy appears to be either encode a non-functional protein or has lost its role in the control of leaf glucosinolate biosynthesis by subfunctionalization.

2.3. Genetic control of root glucosinolate variation

Extensive phenotypic variation was observed in roots for both amount and type of GSLs. The total GSL content in roots ranged from 2.4 to 17.1 μmol/g (Appendix 1 in Kittipol et al., 2019). In contrast to leaves, indole GSLs (47.7%) and the aromatic GSL GST (45.0%) formed the major classes, with aliphatic GSLs being a minor component (7.3%).

To identify loci controlling the level and composition of GSLs in roots, we undertook AT analysis on all individual GSLs, total GSL and

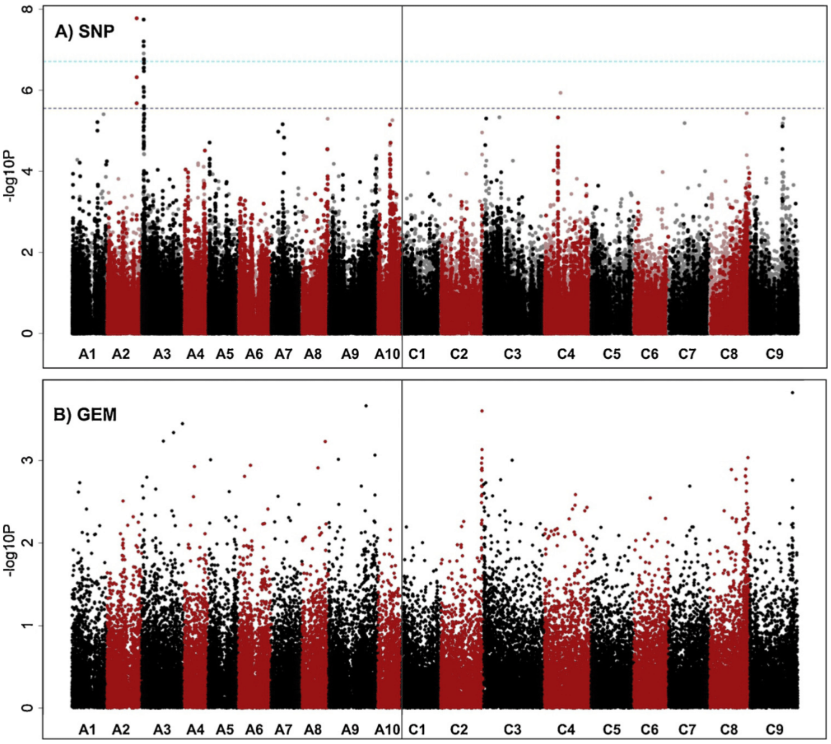


Fig. 5. Association analysis for root aromatic glucosinolate content. (A) Manhattan plot showing genome-wide associations for the identification of transcriptome single-nucleotide polymorphism (SNP) markers of 288 *Brassica napus* accessions with leaf glucosinolate content. Marker associations was calculated using a mixed linear model which incorporated population structure and relatedness. The SNP markers are positioned on the x-axis based on the genomic order of the gene models in which the polymorphism was scored. The significance of the trait association, as -log10 P values, plotted on the y-axis. The horizontal purple and cyan lines represent false discovery rate (FDR) threshold at 5% and the threshold for Bonferroni significance of 0.05, respectively. Chromosomes of *B. napus* are labelled A1– A10 and C1 – C9, shown in alternating black and red colors to allow boundaries to be clearly distinguished. Dark opaque points are simple SNP markers (i.e. polymorphisms between resolved bases) and hemi-SNPs that have been directly linkage-mapped, both of which can be assigned to one genome, whereas light points are hemi-SNP markers (i.e. polymorphisms involving multiple bases called at the SNP position in one allele of the polymorphism) for which the genome of the polymorphism cannot be assigned. (B) Association analysis of expression variation-based markers (GEM) with leaf aliphatic glucosinolate. Reads per kb per million aligned reads (RPKM) were regressed against the trait, and R^2 and P values were calculated for each gene. The gene models are positioned on the x-axis based on their genomic order, with the significance of the associated trait, as -log10 P, plotted on the y-axis (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

GSLs grouped by type (aliphatic, indole or aromatic). The Manhattan plots are shown in Appendix 4 of Kittipol et al. (2019). For the root aliphatic GSLs, SNP associations revealed the same controlling loci on A2/C2 and A9/C9 as in leaves, and furthermore *Bna.HAG1.A9* is also identified as one of the top GEMs ($p = 2.10\times10^{-9}$). For the aromatic GSL (i.e. GST), an exceptionally well-defined association peak, with SNP markers exceeding the Bonferroni-corrected 5% significance threshold, was identified on chromosome A3, as shown in Fig. 5. The genes in this region as listed in Appendix 14 of Kittipol et al. (2019) include an orthologue of HAG3 (AT5G07690), a transcription factor shown (from studies in *A. thaliana*) to regulate aliphatic GSL biosynthesis. The expression level of *Bna.*HAG3.*A3* in the AT platform dataset of Havlickova et al (2018) is low across all accessions. The functional genotypes had been derived from re-sequencing of leaf transcriptome, so GEMs would not be identifiable for genes with root-specific expression patterns. We therefore performed differential expression analyses based on root transcriptome re-sequencing of 4 accessions with high root aromatic GSLs and 4 accessions with low root aromatic GSLs, as listed in Appendix 15 of Kittipol et al. (2019), each with 4 biological replicates. *Bna.*HAG3.*A3* expression was found to be highly correlated with aromatic GSL content ($\log_2$ fold-change = 14.8; $p = 5.47 \times 10^{-11}$) with expression of *Bna.*HAG3.*A3* high in high-root aromatic GSL group and very low in the low-root aromatic group, as shown in Supplementary Figure S1, confirming *Bna.*HAG3.*A3* as an excellent candidate for controlling this trait.

In order to identify differential expression of genes that might be regulated by *Bna.HAG3.A3* in such a way as to limit potential confounding effect between GSL pathways, we performed a stringent root differential expression analysis ($\log_2$ fold-change ≥ 4; $p \leq 1 \times 10^{-10}$) between accessions N01D-1330 and KARAT, which differ in root aromatic GSLs but both of which are low in aliphatic GSLs. This analysis revealed 107 genes with BLAST hits to annotated *A. thaliana* genes, including an orthologue of MAM3 (AT5G23020) on chromosome A3, an orthologue of *IMPI2* (AT2G43100) on chromosome C4 and orthologues of *CYP83A1* (AT4G13770) on each of chromosomes A4 and C4, as shown in Appendix 16 of Kittipol et al. (2019). All of these show higher expression in the high root aromatic GSL accession. In *Arabidopsis*, MAM3 was identified as the key enzyme catalyzing chain elongation of methionine-derived GSLs (Textor et al., 2007) and CYP83A1 can oxidize both aliphatic and aromatic aldoximes (Naur et al., 2003). In *B. napus*, we found that the expression of *Bna.*MAM3.*A3*, *Bna.CYP83A1.A4* and *Bna.CYP83A1.C4* all had significant positive correlations with aromatic GSL in roots (Appendix 17 in Kittipol et al., 2019). GST is a derivative of the chain-elongated homophenylalanine but the genes involved in the chain-elongation of phenylalanine of aromatic GSL pathway are unclear. This result suggests that *Bna.*MAM3.*A3* may play an important role in phenylalanine elongation for aromatic GSL biosynthesis in *B. napus*.

2.4. Relationships between glucosinolate content of vegetative tissues and seeds

In order to understand the relationship of GSLs between vegetative tissues and seeds, we added the seed GSL data from Lu et al. (2014) to the leaf and root data collected from this study and extended the Spearman's correlation analysis shown in Table 2 to include seeds. Aliphatic GSL exhibited the strongest correlations between organs, in particular between leaf and the other two organs (Leaf-Root: r = 0.68 ***, Leaf-Seed: 0.54 ***, Seed-Root: 0.43 ***). These significant positive correlations indicate that natural variation observed in aliphatic GSL between the organs could be regulated by long-distance transport or a master regulator of the aliphatic biosynthetic pathway that controls the biosynthesis of aliphatic GSLs in all of these organs. To investigate whether variation in transport or biosynthesis processes explained the natural variations in aliphatic GSL pattern between leaf and seed *B. napus*, we analyzed additional seed data for associations with the orthologues of *Arabidopsis* GSL transporters, GTR1 (AT3G47960) and GTR2 (AT5G62680). In *B. napus* genome, four orthologues of GTR1 (on C3 and A6) and five orthologues of GTR2 (on C3, C9, A6 and A9) were found but none of the copies showed associations with seed, leaf or root aliphatic GSL. Although *Bna.*GTR2.*A9* and *Bna.*GTR2.*C9* were found in parts of the genome within the SNP and GEM association peaks on chromosome A9 and C9, no correlation between gene expressions and aliphatic contents was observed across the tissues (Appendix 18 in Kittipol et al., 2019). Comparison of the AT plots for total seed GSL, leaf aliphatic and root aliphatic GSLs showed that they all shared four common association peaks on chromosome A2, A9, C2 and C9 which correspond to the HAG1 orthologue-containing control loci. Furthermore, comparison of aliphatic GSLs in leaf and seed, as shown in Supplementary Figure S2 revealed two distinct classes: one with relatively high GSL in both organs and one with relatively low GSL in both organs. The lack of any accession with high GSL in the leaves and low GSL in the seeds indicates the basis of aliphatic GSL variations between plant tissues to be from the amount synthesized, as controlled by orthologues of HAG1, and not by variation in the transport processes.

3. Discussion

3.1. Aliphatic glucosinolates

The Polish spring rape cultivar Bronowski is known to be the genetic source for this trait deployed in all commercial low-seed GSL *B. napus* cultivars through selective breeding (Rosa et al., 1997). This reduction in oilseed GSLs is due to reduction in aliphatic GSLs (Kondra and Stefansson, 1970; Rucker and Rudloff, 1991). However, the molecular mechanism underlying the low seed GSL trait in oilseed rape was unclear. Some studies reported no significant correlation between seed and leaf GSL in *B. napus* canola cultivars (Porter et al., 1991; Fieldsen and Milford, 1994), leading to an assumption that inhibition of the GSL transport processes could have given rise to the low-seed GSL trait in *B. napus*. This hypothesis was supported by the report on the two nitrate/peptide transporter family, GTR1 and GTR2, controlling GSL accumulation in *A. thaliana* seeds (Nour-Eldin et al., 2012). Although orthologues of GTR2 are found in close proximity to causative loci controlling low-seed GSL trait in *B. napus* (Lu et al., 2014), we identified no accession with low seed GSL and the high leaf GSL that would be expected from blocking transport from the leaf, as was observed in *A. thaliana*. Neither did we identify SNP or GEM associations between GTR1 or GTR2 orthologues and GSL traits. Instead, our data reveals significant positive correlation between seed and leaf GSLs where seed GSL profile is a good reflection of the profile found in the leaf (Table 2, Supplementary Fig. S2 and Appendix 19 in Kittipol et al., 2019). Previous work in *A. thaliana* has shown a similar positive correlation with the level of aliphatic GSLs in the leaves representing the minimal concentration of aliphatic seed GSL assuming there were no variation in GSL transport from the leaves to the seeds (Kliebenstein et al., 2001b). Aliphatic GSLs predominate in *B. napus* leaf and seed, so it is not surprising that the same gene associations were detected for total seed GSL (Harper et al., 2012) and total leaf GSL (Appendix 5 & Appendix 10 in Kittipol et al., 2019). Genetic variation for the reduced GSL level in seed, which reflected in the reduced GSL level in leaf, was due to structural changes in the region of *B. napus* genome containing the key regulator of aliphatic GSL biosynthetic genes as a result of breeding-directed selection. Our gene expression analyses confirm the results, i.e. that low-leaf aliphatic GSL lines such as 'Cabriolet' and 'Apex', have non-functional HAG1 orthologues on chromosomes A9 and C2 in place of functional genes in high aliphatic GSL lines (Appendix 12 & Appendix 13 in Kittipol et al., 2019). Our results are consistent with the genome sequence of the low-GSL cultivar Darmor-*bzh*, in which orthologues of HAG1 have been lost on chromosome A9 and C2 but no sequence changes in GTR1 and GTR2 orthologues were identified (Chalhoub et al., 2014).

3.2. Aromatic glucosinolates

Although homophenylalanine-derived GSL is prevalent in *B. napus* roots, few ecotypes of *A. thaliana* produce this class of GSL, and then in very small amounts (Brown et al., 2003). The resulting inability to use the model plant *A. thaliana* to study aromatic GSL and the challenges of working with *B. napus* complex polyploidy has limited the advancement in the understanding of the aromatic biosynthetic pathway. To overcome these challenges, we combined AT with a differential gene expression analysis in root tissues. The region of chromosome A3 showing strong association with variation in root aromatic GSL (Fig. 5) contained an orthologue of HAG3 (Appendix 14 in Kittipol et al., 2019). Compared with other orthologues, *Bna.*HAG3.*A3* contained the highest frequency of polymorphisms, particularly SNPs, which showed strong association with variation in GST content of roots. Using the expression data from root RNA-seq, we have found higher expressions of *Bna.*-HAG3.*A3* gene in high-root aromatic GSL lines and lower expression in low root aromatic GSL lines, supporting our hypothesis. Our interpretation is that *Bna.*HAG3.*A3*, an orthologue of a known regulator of aliphatic GSL in *A. thailana*, is a key regulator of root aromatic GSL biosynthesis in *B. napus*. Furthermore, our results indicate that *Bna.*-HAG3.*A3* regulates a biosynthetic pathways shared between aliphatic and aromatic GSLs. Through differential expression analysis we identified *Bna.*MAM3.*A3* amongst the genes with largest changes in their expression between accessions (Appendix 16 in Kittipol et al., 2019). Roots of *B. napus* are dominated by a chain-elongated homophenylalanine aromatic GSL, GST, but genes involved in the chain-elongation of phenylalanine are unknown. We propose that *Bna.*-MAM3.*A3,* previously known to be part of aliphatic pathway, is also involved in the chain-elongation of phenylalanine in *B. napus*. Consistent with this hypothesis is the observation that MAM3 has a broad substrate specificity in addition to methionine-derived 2-oxoacids where MAM3 is able to form condensation reaction with phenylpyruvate leading to GST production (Textor et al., 2007). Quantitative Trait Locus mapping studies in *A. thaliana* for aromatic GSL reported *GS-Elong* locus (comprising MAM1, MAM2 and MAM3), which controls total leaf aliphatic GSL, to also be the major QTL for controlling phenylalanine elongation (Kliebenstein et al., 2001a). This is also consistent with our hypothesis that chain elongation of methionine-derived aliphatic GSLs and phenylalanine-derived aromatic GSLs share a pathway.

4. Conclusions

Glucosinolate profiles in *B. napus* accessions differ extensively in both type and amount. Aliphatic GSL content in seeds and roots reflect those in leaves and is regulated by *Bna.HAG1*.A9 and *Bna.HAG1*.C2. Aromatic GSLs predominate in the root and we implicate *Bna.HAG3.A3* in their control. There are implications for the manipulation of GSLs for modulation of interactions between the important crops of this species and various pests and diseases. Firstly, blockage of glucosinolate transport into seeds (thus achieving the low seed GSL content needed for oilseed rape quality whilst maintaining high aliphatic GSL content in vegetative tissues) has not yet been achieved in the available germplasm and represents an opportunity to be explored. Secondly, there is a simple genetic basis for the variation observed for root aromatic GSL content and impacts of this variation on below-ground interactions can now be explored.

5. Materials and methods

5.1. Growth of plant material for glucosinolate content

Brassica napus (Oilseed rape) leaves and roots from 288 genotypes of the Renewable Industrial Products from Rapeseed (RIPR) diversity population (Havlickova et al., 2018) were harvested for GSL extraction four weeks after sowing, as described in detail in Kittipol et al. (2019). Four biological replicates of each accessions were grown. At harvest, leaf and root samples were wrapped in labelled foil and immediately frozen in liquid nitrogen. There are 56 Modern Winter oilseed rape (OSR), 65 Winter OSR, 6 Winter Fodder, 121 Spring OSR, 26 Swede and 14 Exotic varieties within this panel (Appendix 1 in Kittipol et al., 2019).

5.2. Glucosinolate quantification

A complete description of the GSL extraction methodology and analysis is presented in Kittipol et al. (2019) and Doheny-Adams et al. (2017). Briefly, GSL mixture from freeze-dried ground leaves or roots were extracted with 80% methanol (v/v), purified and desulfated overnight (Kittipol et al., 2019). Glucotropaeolin was added as an internal standard prior to extraction. Desulfoglucosinolates (dsGSL) were separated by HPLC coupled with photodiode array detector using reverse phase C18 column (5μ ODS(2), 150 mm × 4.6 mm) at 30 °C with mobile phase solutions consisting of 100% diH$_2$O and 30% (v/v) acetonitrile, as described in Doheny-Adams et al (2017).

5.3. Statistical analysis

Statistical analyses were carried out with R statistical software (R core team, 2013). Spearman's correlation analysis was used to analyze the relationship between different groups of GSL in different organs (Table 2). Spearman's correlation was an appropriate type of correlation coefficient because it is more robust to work with the large variabilities and skewed distribution of the levels of GSLs.

5.4. Associative transcriptomics

Functional genotype was constructed (Havlickova et al., 2018) by mapping leaf RNA-sequence data onto the reference sequence of ordered Brassica A and C genome-based pan-transcriptomes (He et al., 2015), using the method described in (Bancroft et al., 2011). To reduce errors in SNP identification and assessment of linkage disequilibrium, filtering and quality checking parameters were applied as described in (Havlickova et al., 2018), producing a set of 355 536 SNP markers, of which 256 397 SNP had a minor allele frequency (MAF) > 0.01. Transcript abundance was quantified and normalized as reads per kb per million aligned reads (RPKM) for each sample and 53 889 CDS models was detected with significant expression (> 0.4 RPKM). Full detail of the methods is described in Kittipol et al. (2019).

The statistical software R was used to perform Associative Transcriptomics was performed using R, as detailed in Kittipol et al. (2019) and Havlickova et al. (2018). SNP-based analyses were performed with Genome Association and Prediction Integrated Tool (GAPIT) R package using mixed linear model that includes both fixed and random effects. SNP markers are positioned on the x-axis based on the genomic order of the CDS gene model in which the polymorphism was scored. The significance of the trait association, as $-\log_{10}P$ values, was plotted on the y-axis. For GEM-based analyses, fixed-effect linear model was calculated in R software, with trait score as the response variable and RPKM values plus the Q matrix inferred by PSIKO as explanatory variables. False discovery rate (FDR) (Benjamini and Hochberg, 1995) and threshold for Bonferroni (Dunn, 1961) corrections were used to set significance threshold at $P < 0.05$.

5.5. Differential expression analysis of root RNA-seq data

Differential gene expression was analyzed using root transcriptome sequences from four biological replicates. The methods in Bioconductor package EdgeR (Robinson et al., 2009) were used to identify differentially expressed genes, as described in Kittipol et al. (2019).

5.6. Accession numbers

Short read sequence data have been deposited at the Sequence Read Archive under BioProject ID: PRJNA524101

Author contributions

V.K. designed and performed the experiments, analyzed the data and wrote the manuscript.

Z.H. performed root differential expression analysis.

L.W. grown plant material and extracted roots RNA for the differential expression experiment.

T.D-A. helped in method development for glucosinolate quantification and performed some HPLC analysis. S.L. performed some HPLC analysis.

I.B. designed experiments, interpreted results and edited the manuscript.

Funding

This work was supported by UK Biotechnology and Biological Sciences Research Council [grant number BB/L002124/1]. We thank the High-Throughput Genomics Group at the Wellcome Trust Centre for Human Genetics (funded by Wellcome Trust grant reference 090532/Z/09/Z) for the generation of mRNAseq and genomic sequencing data. VK received support from Radhika V Sreedhar Scholarship Fund from the Department of Biology, University of York and Scholarships for Overseas Students from the University of York.

One sentence summary

In *Brassica napus*, orthologues of *HAG1* control variation of aliphatic glucosinolates in leaves and orthologues of *HAG3* control variation of aromatic glucosinolates in roots.

Acknowledgement

We kindly thank Ricardo Gonzaga Pimenta and Martin France for some of the GSL data generated as part of the undergraduate final-year project at the University of York.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.jplph.2019.06.001.

References

Bancroft, I., Morgan, C., Fraser, F., Higgins, J., Wells, R., Clissold, L., Baker, D., Long, Y., Meng, J., Wang, X., et al., 2011. Dissecting the genome of the polyploid crop oilseed rape by transcriptome sequencing. Nat. Biotechnol. 29, 762–766.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. 57, 289.

Bhandari, S., Jo, J., Lee, J., 2015. Comparison of glucosinolate profiles in different tissues of nine Brassica crops. Molecules 20, 15827–15841.

Brown, P.D., Tokuhisa, J.G., Reichelt, M., Gershenzon, J., 2003. Variation of glucosinolate accumulation among different organs and developmental stages of Arabidopsis thaliana. Phytochemistry 62, 471–481.

Celenza, J.L., 2005. The Arabidopsis atr1 myb transcription factor controls indolic glucosinolate homeostasis. Plant Physiol. 137, 253–262.

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I.A.P., Tang, H., Wang, X., Chiquet, J., Belcram, H., Tong, C., Samans, B., et al., 2014. Early allopolyploid evolution in the post-

Neolithic Brassica napus oilseed genome. Science (80-) 345, 950–953.

Doheny-Adams, T., Redeker, K., Kittipol, V., Bancroft, I., Hartley, S.E., 2017. Development of an efficient glucosinolate extraction method. Plant Methods 13, 17.

Dunn, O.J., 1961. Multiple comparisons among means. J. Am. Stat. Assoc. 56, 52–64.

Fahey, J.W., Zalcmann, A.T., Talalay, P., 2001. The chemical diversity and distribution of glucosinolates and isothiocyanates among plants. Phytochemistry 56, 5–51.

Fieldsen, J., Milford, G.F.J., 1994. Changes in glucosinolates during crop development in single- and double-low genotypes of winter oilseed rape (Brassica napus): I. Production and distribution in vegetative tissues and developing pods during development and potential role in the recycling. Ann. Appl. Biol. 124, 531–542.

Frerigmann, H., Gigolashvili, T., 2014. MYB34, MYB51, and MYB122 distinctly regulate indolic glucosinolate biosynthesis in arabidopsis Thaliana. Mol. Plant 7, 814–828.

Gajardo, H.A., Wittkop, B., Soto-Cerda, B., Higgins, E.E., Parkin, I.A.P., Snowdon, R.J., Federico, M.L., Iniguez-Luy, F.L., 2015. Association mapping of seed quality traits in Brassica napus L. Using GWAS and candidate QTL approaches. Mol. Breed. 35, 1–19.

Giamoustaris, A., Mithen, R., 1995. The effect of modifying the glucosinolate content of leaves of oilseed rape (Brassica napus ssp. oleifera) on its interaction with specialist and generalist pests. Ann. Appl. Biol. 126, 347–363.

Gigolashvili, T., Berger, B., Mock, H.P., Müller, C., Weisshaar, B., Flügge, U.I., 2007a. The transcription factor HIG1/MYB51 regulates indolic glucosinolate biosynthesis in Arabidopsis thaliana. Plant J. 50, 886–901.

Gigolashvili, T., Engqvist, M., Yatusevich, R., Müller, C., Flügge, U.I., 2008. HAG2/MYB76 and HAG3/MYB29 exert a specific and coordinated control on the regulation of aliphatic glucosinolate biosynthesis in Arabidopsis thaliana. New Phytol. 177, 627–642.

Gigolashvili, T., Yatusevich, R., Berger, B., Müller, C., Flügge, U.I., 2007b. The R2R3-MYB transcription factor HAG1/MYB28 is a regulator of methionine-derived glucosinolate biosynthesis in Arabidopsis thaliana. Plant J. 51, 247–261.

Gimsing, A.L., Kirkegaard, J.A., 2009. Glucosinolates and biofumigation: fate of glucosinolates and their hydrolysis products in soil. Phytochem. Rev. 8, 299–310.

Glen, D.M., Jones, H., Fieldsend, J.K., 1990. Damage to oilseed rape seedlings by the field slug Deroceras reticulatum in relation to glucosinolate concentration. Ann. Appl. Biol. 117, 197–207.

Griffiths, D.W., Birch, A.N.E., Hillman, J.R., 1998. Antinutritional compounds in the Brassicaceae: analysis, biosynthesis, chemistry and dietary effects. J. Hortic. Sci. Biotechnol. 73, 1–18.

Grubb, C.D., Abel, S., 2006. Glucosinolate metabolism and its control. Trends Plant Sci. 11, 89–100.

Halkier, B.A., Gershenzon, J., 2006. Biology and biochemistry of glucosinolates. Annu. Rev. Plant Biol. 57, 303–333.

Harper, A.L., Trick, M., Higgins, J., Fraser, F., Clissold, L., Wells, R., Hattori, C., Werner, P., Bancroft, I., 2012. Associative transcriptomics of traits in the polyploid crop species Brassica napus. Nat. Biotechnol. 30, 798–802.

Havlickova, L., He, Z., Wang, L., Langer, S., Harper, A.L., Kaur, H., Broadley, M.R., Gegas, V., Bancroft, I., 2018. Validation of an updated Associative Transcriptomics platform for the polyploid crop species Brassica napus by dissection of the genetic architecture of erucic acid and tocopherol isoform variation in seeds. Plant J. 93, 181–192.

He, Z., Cheng, F., Li, Y., Wang, X., Parkin, I.A.P., Chalhoub, B., Liu, S., Bancroft, I., 2015. Construction of Brassica a and C genome-based ordered pan-transcriptomes for use in rapeseed genomic research. Data Br 4, 357–362.

Hirai, M.Y., Sugiyama, K., Sawada, Y., Tohge, T., Obayashi, T., Suzuki, A., Araki, R., Sakurai, N., Suzuki, H., Aoki, K., et al., 2007. Omics-based identification of Arabidopsis myb transcription factors regulating aliphatic glucosinolate biosynthesis. Proc. Natl. Acad. Sci. U. S. A. 104, 6478–6483.

Hopkins, R.J., van Dam, N.M., van Loon, J.J.A., 2009. Role of glucosinolates in insect-plant relationships and multitrophic interactions. Annu. Rev. Entomol. 54, 57–83.

Kittipol V, He Z, Wang L, Doheny-Adams T, Langer S, Bancroft I Data in support of genetic architecture of glucosinolate variations in Brassica napus. Data Br. (2019).

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

Kliebenstein, D.J., Kroymann, J., Brown, P., Figuth, A., Pedersen, D., Gershenzon, J., Mitchell-Olds, T., 2001b. Genetic control of natural variation in Arabidopsis glucosinolate accumulation. Plant Physiol. 126.

Kondra, Z.P., Stefansson, B.R., 1970. Inheritance of the major glucosinolates of Rapesed (Brassica napus) meal. Can. J. Plant Sci. 50, 643–647.

Li, F., Chen, B., Xu, K., Wu, J., Song, W., Bancroft, I., Harper, A.L., Trick, M., Liu, S., Gao, G., et al., 2014. Genome-wide association study dissects the genetic architecture of seed weight and seed quality in rapeseed (Brassica napus L.). DNA Res. 21, 355–367.

Lu, G., Harper, A.L., Trick, M., Morgan, C., Fraser, F., O'Neill, C., Bancroft, I., 2014. Associative transcriptomics study dissects the genetic architecture of seed glucosinolate content in Brassica napus. DNA Res. 21, 613–625.

Mithen, R., 1992. Leaf glucosinolate profiles and their relationships to pest and disease resistance in oilseed rape. Euphytica 63, 71–83.

Naur, P., Petersen, B.L., Mikkelsen, M.D., Bak, S., Rasmussen, H., Olsen, C.E., Halkier, B.A., 2003. CYP83A1 and CYP83B1, two nonredundant cytochrome P450 enzymes metabolizing oximes in the biosynthesis of glucosinolates in Arabidopsis. Plant Physiol. 133, 63–72.

Nour-Eldin, H.H., Andersen, T.G., Burow, M., Madsen, S.R., Jørgensen, M.E., Olsen, C.E., Dreyer, I., Hedrich, R., Geiger, D., Halkier, B.A., 2012. NRT/PTR transporters are essential for translocation of glucosinolate defence compounds to seeds. Nature 488, 531–534.

Porter, A.J.R., Morton, A.M., Kiddle, G., Doughty, K.J., Wallsgrove, R.M., 1991. Variation in the glucosinolate content of oilseed rape (Brassica napus L.) leaves: I. Effect of leaf age and position. Ann. Appl. Biol. 118, 461–468.

Potter, M.J., Vanstone, V.A., Davies, K.A., Rathjen, A.J., 2000. Breeding to increase the concentration of 2-phenylethyl glucosinolate in the roots of Brassica napus. J. Chem. Ecol. 26, 1811–1820.

R core team, 2013. R: a Language and Environment for Statistical Computing. https://doi.org/10.1007/978-3-540-74686-7.

Rask, L., Andréasson, E., Ekbom, B., Eriksson, S., Pontoppidan, B., Meijer, J., 2000. Myrosinase: gene family evolution and herbivore defense in Brassicaceae. Plant Mol. Biol. 42, 93–113.

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2009. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140.

Rosa, E.A.S., Heaney, R.K., Fenwick, G.R., Portas, C.A.M., 1997. Glucosinolates in crop plants. Hortic. Rev. (Am Soc Hortic Sci). https://doi.org/10.1002/9780470650622.ch3.

Rucker, B., Rudloff, E., 1991. Investigations of the inheritance of the glucosinolate content in seeds of winter oilseed rape (Brassica napus L.). Proc 8th Int Rapeseed Congr Saskatoon 191–196.

Sonderby, I.E., Burow, M., Rowe, H.C., Kliebenstein, D.J., Halkier, B.A., 2010. A complex interplay of three R2R3 MYB transcription factors determines the profile of aliphatic glucosinolates in Arabidopsis. Plant Physiol. 153, 348–363.

Sønderby, I.E., Geu-Flores, F., Halkier, B.A., 2010. Biosynthesis of glucosinolates – gene discovery and beyond. Trends Plant Sci. 15, 283–290.

Tayo, T., Dutta, N., Sharma, K., 2012. Effect of feeding canola quality rapeseed mustard meal on animal production - a review. Span. J. Agric. Res. 33, 114–121.

Textor, S., de Kraker, J.-W., Hause, B., Gershenzon, J., Tokuhisa, J.G., 2007. MAM3 catalyzes the formation of all aliphatic glucosinolate chain lengths in Arabidopsis. Plant Physiol. 144, 60–71.

Wittstock, U., Halkier, B.A., 2002. Glucosinolate research in the Arabidopsis era. Trends Plant Sci. 7, 263–270.

Wittstock, U., Halkier, B.A., 2000. Cytochrome P450 CYP79A2 from Arabidopsis thaliana L. Catalyzes the conversion of L -phenylalanine to phenylacetaldoxime in the biosynthesis of benzylglucosinolate. J. Biol. Chem. 275, 14659–14666.

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

https://doi.org/10.1016/j.dib.2019.104402

Data in brief 25 (2019) 104402

Contents lists available at ScienceDirect

Data in brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Data in support of genetic architecture of glucosinolate variations in *Brassica napus*

Varanya Kittipol, Zhesi He, Lihong Wang, Tim Doheny-Adams, Swen Langer, Ian Bancroft[*]

Department of Biology, University of York, Heslington, York, YO10 5DD, UK

ARTICLE INFO

Article history:
Received 10 June 2019
Received in revised form 6 August 2019
Accepted 8 August 2019
Available online 14 August 2019

Keywords:
Glucosinolates
Genetic associations
Associative transcriptomics
SNP markers
Gene expression markers
Population structure
Oilseed rape
Brassica napus

ABSTRACT

The transcriptome-based GWAS approach, Associative Transcriptomics (AT), which was employed to uncover the genetic basis controlling quantitative variation of glucosinolates in *Brassica napus* vegetative tissues is described. This article includes the phenotypic data of leaf and root glucosinolate (GSL) profiles across a diversity panel of 288 *B. napus* genotypes, as well as information on population structure and levels of GSLs grouped by crop types. Moreover, data on genetic associations of single nucleotide polymorphism (SNP) markers and gene expression markers (GEMs) for the major GSL types are presented in detail, while Manhattan plots and QQ plots for the associations of individual GSLs are also included. Root genetic association are supported by differential expression analysis generated from root RNA-seq. For further interpretation and details, please see the related research article entitled *'Genetic architecture of glucosinolate variation in Brassica napus'* (Kittipol et al., 2019).

DOI of original article: https://doi.org/10.1016/j.jplph.2019.06.001.

[*] Corresponding author.
E-mail address: ian.bancroft@york.ac.uk (I. Bancroft).

https://doi.org/10.1016/j.dib.2019.104402

Specifications Table

Subject area	Biology
More specific subject area	*Brassica* secondary metabolite and genetics
Type of data	Figure, Tables (MS Excel spreadsheets)
How data was acquired	Glucosinolate measurements were obtained using HPLC on C18 reverse phase column. SNP identification, transcript quantification, construction of the reference coding DNA sequence and associative transcriptomic analysis platform were developed prior to this publication.
Data format	Raw, processed, analyzed
Experimental factors	Desulfoglucosinolates determined as glucosinolates from leaves and roots of genotyped *B. napus* diversity panel. SNP- and GEM-trait association data were analyzed using R scripts.
Experimental features	Transcriptome-based genome wide association
Data source location	Glucosinolate data was collected at the University of York, York, UK.
Data accessibility	Short read sequence data have been deposited at the Sequence Read Archive under BioProject ID: PRJNA524101 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA524101). Glucosinolate data are provided in Annex spreadsheets.
Related research article	V. Kittipol, Z. He, L. Wang, T. Doheny-Adams, S. Langer, I. Bancroft, Genetic architecture of glucosinolate variation in *Brassica napus*, J. Plant Physiol. 240 (2019) 152988. https://doi.org/10.1016/j.jplph.2019.06.001 [1].

Value of the data

- This data provides comprehensive leaves and roots glucosinolate profiles across a diversity panel of 288 *Brassica napus* (oilseed rape) genotypes with information on the population structure. Glucosinolate trait data can benefit oilseed rape agribusinesses and researchers of this field in the selection of genotypes with desirable profiles or manipulation of profiles to modulate plant-pest interactions.
- The GEM and SNP markers identified in the region of the genome that controls the variation in glucosinolate contents can help accelerate breeding of oilseed rape by marker-assisted selection
- This data could be used for comparison or replication of genetic association markers for the natural glucosinolate variations in other populations and other plant tissues.

1. Data

The data contains information on leaves and roots glucosinolate (GSL) profiles of 288 *Brassica napus* genotypes (Fig. 1). The relatedness of the accessions was analyzed and visualized by the dendrogram (Fig. 1A). The seven assigned crop types shows the expected clustering (Fig. 1B) with the highest likelihood of two differentiated subpopulations ($k = 2$), which separated into the spring or winter oilseed rape crop types or a mixture of the two (Fig. 1C). Full dataset of the GSL profiles are presented as mean from four biological replicates of each accessions (Appendix 1) with distribution of the data displayed as histograms (Appendix 2) and analysis of GSL contents by crop types (Appendix 3).

These phenotypic data were used to generate association data identifying single nucleotide polymorphism (SNP) markers and gene expression markers (GEMs) in transcriptome-based genome wide association studies, Associative Transcriptomics (AT) [2,3]. The Manhattan plots for these associations are shown in Appendix 4 for root traits and Appendix 5 for leaf traits. The significance of the trait associations, shown as $-\log_{10}P$ value, passing both false discovery rate (FDR) threshold at 5% and threshold for Bonferroni significance of 0.05 suggested that the surrounding genomic region has a strong association with the trait. To assess how well the model accounts for population structure and familial relatedness, quantile-quantile (QQ) plots from SNP association analyses have been generated (Appendix 6 & Appendix 7). Appendix 8 summarizes the optimal algorithm showing calculated group kinship matrix, 2*log likelihood function and the estimated heritability for all GSL traits.

As shown in Fig. 1, aliphatic GSLs is the most abundant class of GSL in *B. napus* leaves. SNP-based associations of leaf aliphatic GSL revealed strong associations with markers in the defined regions of chromosome A2, A9, C2, C7 and C9 (Appendix 9). Within these data tables, details of trait associations for genome-assigned markers are provided, including polymorphism, significance of association and the frequency of the minor allele in the population. The same associated regions were shown for total

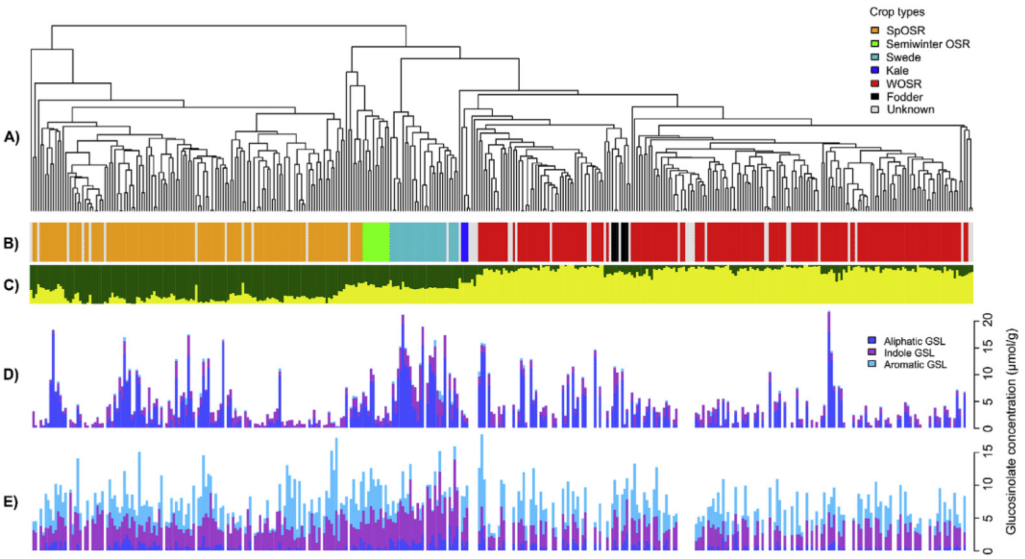


Fig. 1. Population structure and Glucosinolate variation from 288 *B. napus* accessions of the Renewable Industrial Products from Rapeseed (RIPR) Panel. (A) Relatedness of accessions in the panel based on 355 536 scored single-nucleotide polymorphisms (SNPs). (B) Main crop types, color coded: orange for spring oilseed (SpOSR); green for semi-winter oilseed rape; light blue for swede; dark blue for kale; red for winter oilseed rape(WOSR); black for winter fodder and gray for crop type not assigned. (C) Population structure for highest likelihood *k*=2. Variation for glucosinolates content (D) leaf and (E) root of 288 *B.* napus accessions. Individual glucosinolates were grouped according to their structural class as aliphatic (dark blue), indole(margenta) and aromatic(light blue). Panel A, B and C reproduced from Havlickova et al 2018.

seed GSL in *B. napus* (Appendix 10). As presented in [1], orthologues of *HAG1* (AT5G61420), a transcription factor that positively regulated aliphatic GSL biosynthesis, have been discovered within all of these SNP-based associated loci (Appendix 9). In addition, the six GEMs detected above the threshold for the false discovery rate (FDR) at 5% are shown to be involved directly in aliphatic GSL biosynthesis, with orthologues of *HAG1* as the top GEMs (Appendix 11). Presence of GEM association peaks on chromosome A9, C2 and C9 for aliphatic GSL suggested structural genome variation via homoeologous exchange where neighboring genes displayed the same directionality of one genome over-expressed and other genome under-expressed (Appendix 12). The Transcriptome Display Tile Plots [4] was used to visualize the homoeologous exchanges in these regions (Appendix 13).

In *B. napus* roots, aromatic GSL is the dominant GSL class and revealed a clear SNP association peak on chromosome A3 (Appendix 4). As described in [1], an orthologue of *HAG3* was identified in close proximity to the top associated SNP markers within in this region (Appendix 14). To support gene expression analysis in roots, differential expression analysis from root transcriptome-sequence was performed, which compared the expression patterns of 4 accessions with high root aromatic GSLs and 4 accessions with low root aromatic GSLs (Appendix 15). Within the SNP associated region of chromosome A3, *Bna.HAG3.A3* showed the highest significant $\log_2$ fold-change (Appendix 15) with higher expression of *Bna.HAG3.A3* observed in high-root aromatic GSL genotypes and vice versa in the low-root GSL genotypes. To limit potential confounding effect between GSL pathways, further stringent analysis of differential root expression ($p \leq 1 \times 10^{-10}$) was performed between accessions which differs in root aromatic GSLs but are low in aliphatic GSLs (Appendix 16). This analysis revealed insight into genes that had been identified in aliphatic GSL pathway but could have potential roles in the aromatic GSL pathway. This is shown by the significant positive correlations between their expression levels and levels of aromatic GSL (Appendix 17).

To investigate the relationship of GSLs between vegetative tissues and seeds, seed GSL data from [5] was added to the dataset. Correlation analysis between levels of aliphatic GSLs and the transcript abundance of GSL transporters, *GTR1* (AT3G47960) and *GTR2* (AT5G62680), was conducted to investigate the role of transporters on GSL accumulation pattern (Appendix 18), as described in [1]. Finally, correlations between leaf and seed GSLs was analyzed to investigate the basis for the accumulation pattern of GSLs in these tissues (Appendix 19).

2. Experimental design, materials, and methods

2.1. Growth of plant material for glucosinolate content

A subset of 288 *B. napus* accessions from the Renewable Industrial Products from Rapeseed (RIPR) diversity population [2] was grown in long day (16/8 h, 20 °C/14 °C) under controlled glasshouse conditions (University of York, UK). Within this panel, there are 56 Modern Winter OSR, 65 Winter OSR, 6 Winter Fodder, 121 Spring OSR, 26 Swede and 14 Exotic varieties (Appendix 1). Four biological replicates of each accession were grown in root trainers with Terra-Green for ease of root harvesting, supplemented weekly with a half concentration of Murashige and Skoog growth medium [6] adjusted to pH6.5 with KOH. The experiment was arranged as randomized four-block design with one plant per lines in each block. Four weeks after sowing, the third true leaf and the whole root system were harvested from each plant. At harvest, leaves were cut at the base, wrapped in a labelled aluminum foil and immediately frozen in liquid nitrogen. Plants were removed from the tray, had the roots washed, dried with paper towel and cut. All samples were wrapped in labelled aluminum foils and immediately frozen in liquid nitrogen and stored at −80 °C.

2.2. Glucosinolate quantification

As per the recommended quantification method previously tested [7], frozen tissue samples were lyophilized before homogenized to fine powder for 10 min at a frequency of 30 Hz (TissueLyser II, Qiagen). To 50 mg of homogenate, 1975 µl of 80% (v/v) methanol and 25 µl of 5 mM internal standard glucotropaeolinwas added. The sample was mixed and left to stand for 30 min at 20 °C and further mixed with orbital shaker (Vibrax, IKA) at 1200 rpm for 30 min before centrifugation at 8000 rpm for 10 min. Supernatant methanol extract was then transferred to the pre-conditioned Sephadex column in purification step. Purification and desulfation of GSLs was according to [8]. Columns were prepared with 0.5 ml ion-exchange resin (DEAE Sephadex beads in 1:1 ratio with 2 M acetic acid), conditioned with 2 ml imizadoleformate (6 M) and washed twice with 1 ml water. One ml of the extract was transferred to a prepared column and gently washed twice with 1 ml 20 mM sodium acetate (pH 4) before adding 75 µl of purified sulfatase (5 U/ml). Columns were incubated for 24 h and desulfoglucosinolates were eluted with two 1 ml portions of water.

Desulfoglucosinolates were separated by HPLC (Millipore 600E system, Waters) on a reverse phase C18 column at 30 °C (Phenomonex, SphereClone 5µ ODS(2), 150 mm × 4.6 mm) with mobile phase solutions consisting of 100% diH$_2$O and 30% (v/v) acetronitile, as detailed in [7]. Injection was at 10 µl and flow rate was set to 1 ml/min. The absorbance of the eluates was monitored at 229 nm wavelength within the UV spectrum. Samples were separated according to the program described in [7]. Through standard injections, HPLC-MS identification, retention time and photodiode array (PDA) UV spectra, the identity of all major GSLs were confirmed.

2.3. Statistical analysis

Statistical analyses were carried out with R statistical software [9]. One-way ANOVA and Tukey's honest significant difference (HSD) post hoc test were performed on GSL content between crop types (Appendix 3).

2.4. Transcriptome sequencing, SNP identification and transcript quantification

Plant growth conditions, sampling of material, RNA extraction and Illumina transcriptome sequencing was carried out and described previously in [4]. For each genotype, RNA-sequence data was mapped onto recently developed ordered Brassica A and C genome-based pan-transcriptomes as reference sequences [10], using the methods described in [11]. SNP positions were excluded from the alignment if they have a read depth below 10, a base call quality below Q20, missing data below 0.25, and three alleles or more. After rigorous filtering and quality checking parameters to reduce errors in SNP identification and assessment of linkage disequilibrium as detailed in [2], a set of 355 536 SNP markers was generated, of which 256 397 SNP had a minor allele frequency (MAF) > 0.01. Transcript abundance was quantified and normalized as reads per kb per million aligned reads (RPKM) for each sample. Of the 116 098 coding DNA sequence (CDS) models, 53 889 CDS models was detected with significant expression (>0.4 RPKM).

2.5. Associative Transcriptomics

An overview of Associative Transcriptomics (AT) analysis is shown in Fig. 2. The use of transcriptome sequencing in AT allows the discovery of SNP markers in tight linkage disequilibrium with causative genes like conventional GWAS, with the additional feature of finding genes with expression patterns (gene expression markers, GEM) that correlate with the trait variation.

AT was performed using R [9] based on an adaption of the first AT methods [3] with modifications to accommodate for larger dataset, as detailed in [2]. To reduce the risks of false positive associations from undetected population structure that can mimic the signal of association, population structure inference using kernel-PCA and optimization (PSIKO; highest likelihood subpopulation k = 2) [12] was used for Q-matrix generation to correct for population stratification. SNP-based analyses were performed with Genome Association and Prediction Integrated Tool (GAPIT) R package using mixed linear model that includes both fixed and random effects [13]. SNP markers with minor allele frequencies below 0.01 were removed from the SNP dataset leaving 256 397 SNPs for the associations [2]. SNP markers that can be assigned with confidence to the genomic position of the CDS model are rendered dark points and markers that could not be assigned with confidence were rendered pale points. For GEM-based analyses, fixed-effect linear model was calculated in R software, with RPKM values and the Q matrix inferred by PSIKO as explanatory variables, and trait score as the response variable [2]. For each

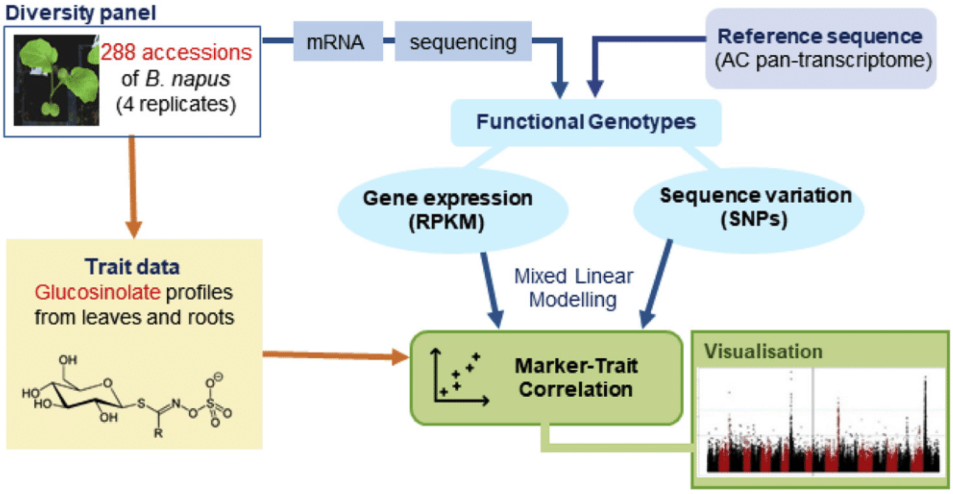


Fig. 2. Overview of associative transcriptomic analysis.

regression, coefficients of determination (R^2), constant, F-value and significance P-values were produced. When genomic inflation factor (λ) was >1, genomic control with P-value adjustment [14] was applied to the GEM analysis to correct for false associations. False discovery rate (FDR) [15] and threshold for Bonferroni [16] corrections were used to set significance threshold at $P < 0.05$. Quantile-Quantile plots all association analyses are included as Appendix 6 for root data and Appendix 7 for leaf data.

2.6. Differential expression analysis of root RNA-seq data

Differential gene expression was analyzed using root transcriptome sequences from four biological replicates (i.e. using root RNA-seq from 4 separate plants of each plant type). The methods in Bioconductor package EdgeR [17] were used to identify the differential expressed genes. In multiple comparisons, both fold change (FC) > 2 and false discovery rate (FDR) < 0.05 were used to flag a gene being differentially expressed. Flags of “1”,“-1” and “0” were used to note positively, and negatively or not significantly expressed genes in the data and can be filtered among comparisons.

Acknowledgments

This work was funded by UK Biotechnology and Biological Sciences Research Council (BB/L002124/1). V.K. received funding from Scholarships for Overseas Students from the University of York and the Radhika V Sreedhar Scholarship Fund from the Department of Biology, University of York. We kindly thank Ricardo Gonzaga Pimenta and Martin France for some of the GSL data generated as part of the undergraduate final-year project at the University of York.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dib.2019.104402.

References

[1] V. Kittipol, Z. He, L. Wang, T. Doheny-Adams, S. Langer, I. Bancroft, Genetic architecture of glucosinolate variation in Brassica napus, J. Plant Physiol. 240 (2019), https://doi.org/10.1016/j.jplph.2019.06.001, 152988.
[2] L. Havlickova, Z. He, L. Wang, S. Langer, A.L. Harper, H. Kaur, M.R. Broadley, V. Gegas, I. Bancroft, Validation of an updated Associative Transcriptomics platform for the polyploid crop species Brassica napus by dissection of the genetic architecture of erucic acid and tocopherol isoform variation in seeds, Plant J. 93 (2018) 181–192, https://doi.org/10.1111/tpj.13767.
[3] A.L. Harper, M. Trick, J. Higgins, F. Fraser, L. Clissold, R. Wells, C. Hattori, P. Werner, I. Bancroft, Associative transcriptomics of traits in the polyploid crop species Brassica napus, Nat. Biotechnol. 30 (2012) 798–802, https://doi.org/10.1038/nbt.2302.
[4] Z. He, L. Wang, A.L. Harper, L. Havlickova, A.K. Pradhan, I.A.P. Parkin, I. Bancroft, Extensive homoeologous genome exchanges in allopolyploid crops revealed by mRNAseq-based visualization, Plant Biotechnol. J. (2016) 1–11, https://doi.org/10.1111/pbi.12657.
[5] G. Lu, A.L. Harper, M. Trick, C. Morgan, F. Fraser, C. O'Neill, I. Bancroft, Associative transcriptomics study dissects the genetic architecture of seed glucosinolate content in Brassica napus, DNA Res. 21 (2014) 613–625, https://doi.org/10.1093/dnares/dsu024.
[6] T. Murashige, F. Skoog, A revised medium for rapid growth and bio assays with tobacco tissue cultures, Physiol. Plant. 15 (1962) 473–497, https://doi.org/10.1111/j.1399-3054.1962.tb08052.x.
[7] T. Doheny-Adams, K. Redeker, V. Kittipol, I. Bancroft, S.E. Hartley, Development of an efficient glucosinolate extraction method, Plant Methods 13 (2017) 17, https://doi.org/10.1186/s13007-017-0164-8.
[8] ISO 9167-1, in: Determination of Glucosinolates Content - Part 1: Method Using High-Performance Liquid Chromatography, Int. Stand., 1992. https://www.evs.ee/products/iso-9167-1-1992.
[9] R core team, R: a Language and Environment for Statistical Computing, 2013, https://doi.org/10.1007/978-3-540-74686-7.
[10] Z. He, F. Cheng, Y. Li, X. Wang, I.A.P. Parkin, B. Chalhoub, S. Liu, I. Bancroft, Construction of Brassica A and C genome-based ordered pan-transcriptomes for use in rapeseed genomic research, Data Br 4 (2015) 357–362, https://doi.org/10.1016/j.dib.2015.06.016.

[11] I. Bancroft, C. Morgan, F. Fraser, J. Higgins, R. Wells, L. Clissold, D. Baker, Y. Long, J. Meng, X. Wang, S. Liu, M. Trick, Dissecting the genome of the polyploid crop oilseed rape by transcriptome sequencing, Nat. Biotechnol. 29 (2011) 762–766, https://doi.org/10.1038/nbt.1926.

[12] A.A. Popescu, A.L. Harper, M. Trick, I. Bancroft, K.T. Huber, A novel and fast approach for population structure inference using Kernel-PCA and optimization, Genetics 198 (2014) 1421–1431, https://doi.org/10.1534/genetics.114.171314.

[13] A.E. Lipka, F. Tian, Q. Wang, J. Peiffer, M. Li, P.J. Bradbury, M.A. Gore, E.S. Buckler, Z. Zhang, GAPIT: genome association and prediction integrated tool, Bioinformatics 28 (2012) 2397–2399, https://doi.org/10.1093/bioinformatics/bts444.

[14] B. Devlin, K. Roeder, Genomic control for association studies, Biometrics 55 (1999) 997–1004, https://doi.org/10.1111/j.0006-341X.1999.00997.x.

[15] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. 57 (1995) 289.

[16] O.J. Dunn, Multiple comparisons among means, J. Am. Stat. Assoc. 56 (1961) 52–64.

[17] M.D. Robinson, D.J. McCarthy, G.K. Smyth, edgeR: A Bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics 26 (2009) 139–140, https://doi.org/10.1093/bioinformatics/btp616.