INVESTIGATING THE GENETIC BASIS AND SELECTION OF DIVERSE PLANT SPECIALIZED METABOLITES IN WILD SOYBEAN, *GLYCINE SOJA*.

by

Farida Yasmin

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Biology

Charlotte

2023

Approved by:

Dr. Bao-Hua Song

Dr. Jun-tao Guo

Dr. Changbao Li

Dr. Xu Li

Dr. Adam M. Reitzel

Dr. Zeba I. Seraj

ABSTRACT

FARIDA YASMIN. Investigating the Genetic Basis and Selection of Diverse Plant Specialized Metabolites in Wild Soybean, *Glycine soja*. (Under the direction of DR. BAO-HUA SONG)

Plant-specialized metabolites, such as glyceollins and soyasaponins, play vital roles in adapting to dynamic environments and promoting human health. Glyceollins, induced phytoalexins derived from the isoflavonoid branch of the phenylpropanoid pathway, and soyasaponins, triterpenoid class compounds naturally abundant in legume species, have particular importance in responding to environmental stresses and contributing to sustainable human nutrition. However, the genetic basis of glyceollin induction and soyasaponin production, especially in wild crop species like wild soybean *(G. soja)*, remains poorly studied. To bridge these knowledge gaps, our study focused on *G. soja*, which has abundant genetic diversity. Our objective was to unravel the genetic basis of glyceollin induction as well as phytochemical diversity with respect to soyasaponin variation. For insights into glyceollin induction, we employed a targeted metabolite-based genome-wide association (mGWA) approach utilizing 264 *G. soja* ecotypes and identified eight significant SNPs associated with glyceollin induction on chromosomes 3, 9, 13, 15, and 20. Among these, six genes near a significant SNP (ss715603454) on chromosome 9 formed two clusters, encoding enzymes of the glycosyltransferase class. We also discovered transcription factor genes, including *MYB* and *WRKY*, within the linkage disequilibrium of the significant SNPs on chromosome 9. Epistasis and strong selection signals were detected for four of the significant SNPs on chromosome 9, indicating their major evolutionary influence on glyceollin induction. For the genetic basis of phytochemical diversity with respect to soyasaponin biosynthesis, we utilized an

untargeted metabolomics approach in an association panel of 190 *G. soja* ecotypes from diverse natural environments. Among the 874 detected metabolite peaks, we annotated 485 metabolites and identified 1155 SNPs significantly associated with 359 metabolites through a genome-wide association study. Clustering analysis revealed eight QTLs, named QTL-multiple metabolite clusters. Mining data within the linkage disequilibrium blocks of these QTLs led to the identification of 612 annotated genes. From this set, we selected 16 candidate genes relevant to the triterpenoid and phenylpropanoid-derived isoflavonoid biosynthetic pathways, with UDP-dependent glycosyltransferase *(UGT)* emerging as a promising candidate gene on chromosome 15. Sequence analysis of the *UGT* gene in 46 different wild soybean ecotypes revealed two haplotypes with three SNPs on exon-1 for 29 ecotypes, resulting in amino acid changes. These haplotypes were significantly associated with varying soyasaponin-producing ecotypes and exhibited notable expression level differences. We also observed the same two haplotypes in different cultivated *G. max* ecotypes. Incidentally, there was a higher frequency of the haplotype associated with relatively low soyasaponin II accumulation in 29 out of 34 *G. max* ecotypes. Our findings provide valuable insights into the genetic basis of glyceollin induction and phytochemical diversity, with a focus on soyasaponin variation. This knowledge will be a good resource for developing phytochemicals-fortified climate-resilient, high-value soybean crops employing metabolic engineering, ultimately benefiting plant and human health.

DEDICATION

To my beloved parents, dearest brother and sister-in-law, and my two adored nephews, Tashfique and Tawsif, I dedicate this dissertation with immense gratitude and heartfelt appreciation. Thank you for being my pillars of strength, no matter the miles that separate us.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my mentors and committee members, Dr. Bao-Hua Song, Dr. Adam M. Reitzel, Dr. Zeba I. Seraj, Dr. Xu Li, Dr. Changbao Li and Dr. Jun-tao Guo for their support and invaluable guidance throughout my dissertation journey. I am grateful to my Ph.D. supervisor Dr. Bao-Hua Song, my co-authors and collaborators, especially Dr. Xu Li and Dr. Hengyou Zhang, for their valuable input and sharing data in my projects and manuscripts. I would also like to acknowledge the incredible support and meaningful discussions provided by my dear friends Dr. Janice Kofsky and Dr. Erik Broemsen. Furthermore, I am deeply thankful to my lab members, undergraduate researchers, graduate cohort, and the entire Department of Biological Sciences, especially our graduate program director Dr. Adam M. Reitzel, for their generous time and unwavering support.

I am indebted to the Schlumberger Foundation Faculty for the Future Fellowship for their indispensable support over the past five years, which has been instrumental in the success of my research. Additionally, I extend my appreciation to the Thomas L. Reynolds Graduate Research Award for their financial assistance, which greatly contributed to the progress of my research endeavors. I extend my sincere appreciation to my mentors and the cohort from the CSHL Frontiers and Techniques in Plant Science course. Their inspiration and guidance have played a significant role in shaping this incredible journey, and I am truly thankful for their contributions to my growth as a researcher and a human being.

Finally, my heartfelt gratitude goes to my beloved parents, brother, sister-in-law, nephews, family and friends for their constant support, unconditional love, and encouragement throughout this transformative journey. While I wish to mention many others who have played a significant role in my journey, the list is truly endless. Their impact on my life and work has been profound, and I am grateful for the amazing support system I have been blessed with. Thank you all so much!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Some of the ideas presented in this chapter have been derived from a review article referenced below, where I was one of the co-authors.

*ZHANG, H., **YASMIN, F.** & SONG, B.-H. 2019. Neglected treasures in the wild — legume wild relatives in food security and human health. Current Opinion in Plant Biology, 49, 17-26.*

The intricacies of plant specialized metabolism

Plants exhibit a remarkable capacity to produce an extensive range of metabolites, including primary metabolites, secondary metabolites, and plant hormones, each serving distinct biological functions. Primary metabolism encompasses essential life activities of plants, involving the production of sugars, amino acids, nucleotides, lipids, and high-energy metabolites through processes like respiration and photosynthesis. Plant hormones, on the other hand, are small compounds that interact with receptor proteins to regulate various organismal processes, such as abscisic acid (ABA) binding to ABA-RESPONSIVE ELEMENT BINDING FACTORS (AREB/ABFs) to control cellular and tissue level responses (Sun et al., 2022).

In addition to these fundamental activities, plants possess an extraordinary ability to synthesize specialized metabolites (also called secondary metabolites or natural products), which utilize the aforementioned primary metabolites as precursors. Unlike primary metabolism, specialized metabolism in plants serves the purpose of producing numerous unique and biologically active substances. Approximately 30% of the carbon fixed through photosynthesis, for example, is allocated towards the biosynthesis of

phenylpropanoids (Rippert and Matringe, 2002). Julius Sachs, a pioneer in modern plant physiology, emphasized that these compounds are not mere "waste products" but rather do not participate in the plant's "inner economy" or are "no longer necessary for the formation of new cells" (Sachs, 1874, Hartmann, 2008). However, the multifunctionality of plant specialized metabolism, as highlighted by Erb and Kliebenstein et al. (2020), demonstrates the integration of specialized metabolites with primary metabolites and plant hormones. Reevaluating this aspect can enhance our comprehension of how these compounds can act as both primary metabolites and regulators, expanding our knowledge and insight into their roles and functions (Erb and Kliebenstein, 2020). The significance of specialized metabolism lies in its role in ensuring plant survival and its potential applications for human use. These compounds have garnered increasing attention due to their diverse biological activities and potential benefits in various fields, including medicine and agriculture (Medema et al., 2021, Weng et al., 2021, Fang et al., 2019, Wurtzel and Kutchan, 2016). Hence, our research endeavors were directed towards investigating the genetic foundations of phytochemical diversity and their induction, specifically targeting isoflavonoids (glyceollin) and triterpenoids (soyasaponins). This choice was driven by our overarching goal to contribute to sustainable agriculture and promote human nutrition through a deeper understanding of these key compounds.

Isoflavonoids and triterpenoids in plants

Plants synthesize a variety of specialized metabolites, which can be categorized into three main classes: phenolic compounds, terpenes, and nitrogen- or sulfur-containing compounds such as alkaloids and glucosinolates. To generate a wide range of specialized metabolites, plants rely on precursors synthesized through fundamental and well-

preserved primary metabolic pathways, including glycolysis, the TCA cycle, aliphatic amino acids, the pentose phosphate pathway, and the shikimate pathway (**Figure 1.1**) (Fernie and Pichersky, 2015, Wang et al., 2022). Several plant species, including Arabidopsis, rice, corn, soybean, and the model legume *Medicago truncatula*, are renowned for their abundance of antimicrobial indole, terpenoid, benzoxazinone, and flavonoid/isoflavonoid natural products. Within the realm of specialized metabolites, isoflavonoids stand out as a notable subclass of flavonoids. They showcase a distinct structural feature, characterized by a phenyl ring fused with a six-membered heterocyclic C-ring, along with another phenyl ring (referred to as the B-ring) positioned at the C3 position. This sets them apart from flavonoids, where the B-ring is substituted at the C2 position. Isoflavonoids exhibit this intriguing structural twist, lending them a unique identity within the family of flavonoids (Han et al., 2009). They are synthesized through the phenylpropanoid pathway and encompass a diverse group of compounds, including phytoalexins, which are predominantly produced by leguminous plant species. Isoflavonoids play a critical role in enhancing plant resistance against biotic stress (Veitch, 2007, Dixon, 2001). Among isoflavonoids, genistein and daidzein serve as the fundamental scaffolds from which thousands of isoflavonoid derivatives are derived through various structural modifications, including hydroxylation, methylation, glycosylation, and molecular rearrangements (Dixon and Steele, 1999, Dixon and Pasinetti, 2010, Sharma and Ramawat, 2013).

In soybean, the predominant isoflavonoids are daidzein, genistein, and glycitein (Graham and Graham, 1991). Furthermore, upon germination and exposure to stress conditions such as soybean cyst nematode infection, seedlings produce phytoalexins such as the

glyceollins (Yasmin et al., 2022). Glyceollin, a pterocarpan phytoalexin, is synthesized in various tissues of soybean plants as a response to pathogenic infection. It has been extensively studied and is recognized as one of the first phytoalexins to be investigated in detail (Paxton, 1975). The accumulation of glyceollin has been linked to its role in enhancing soybean resistance against *Phytophthora sojae*, as evidenced by various studies (Ayers et al., 1976, Albersheim and Valent, 1978, Zähringer et al., 1978, Bhattacharyya and Ward, 1985, Ebel and Grisebach, 1988, Graham and Graham, 1991, Morris et al., 1998).

Triterpenoids, one of the largest and most diverse classes of natural compounds found in plants, have been extensively studied, with over 20,000 triterpenes identified so far (Hill and Connolly, 2013). These compounds exhibit a remarkable range of structural variations and possess significant potential for both plant defense and medicinal applications. While plant is the primary source of triterpene diversity, other organisms also produce triterpenes. Bacteria, for instance, are capable of synthesizing hopene, a simple triterpene derived from squalene (Ourisson and Albrecht, 1992), while sea cucumbers produce triterpene glycosides known for their defensive properties (Van Dyck et al., 2010).

The biosynthesis of triterpenes occurs through the mevalonic acid (MVA) pathway, where the conversion of 2,3-oxidosqualene by oxidosqualene cyclases (OSCs) plays a crucial role in generating a wide variety of triterpene scaffolds. This enzymatic process is considered one of the most intricate reactions in terpene metabolism (Abe, 2007, Phillips et al., 2006, Wendt, 2005). Through cyclization, a diverse range of triterpene structures can be formed, leading to the existence of over 100 distinct triterpene scaffolds in plants

(Xu et al., 2004a). These scaffolds serve as a foundation for further modifications facilitated by triterpene-modifying enzymes, including cytochrome P450s, sugar transferases, and acyltransferases. The interplay between cyclization and enzymatic modifications contributes significantly to the remarkable structural diversity observed in triterpenes (Thimmappa et al., 2014).

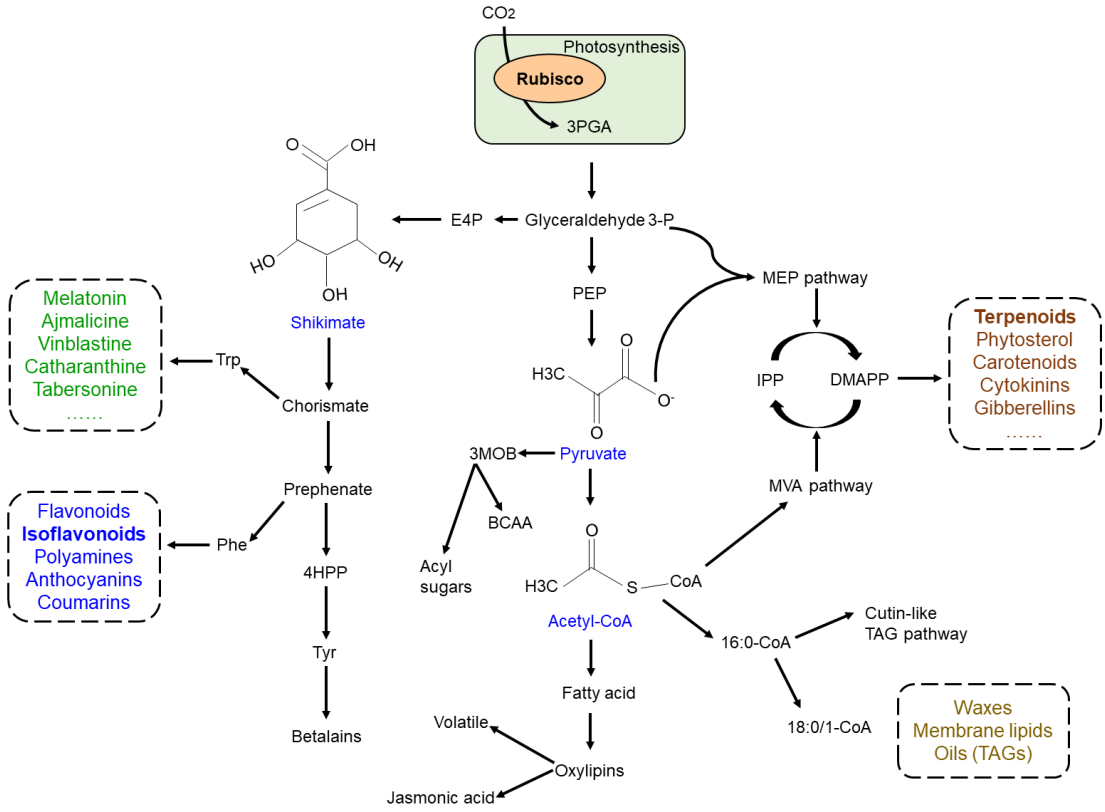


Figure 1.1 The interplay between plant primary and specialized metabolism (adapted from Wang et al. (2022) (Wang et al., 2022). This figure depicts the relationship between plant specialized and primary metabolic pathways, highlighting how the former utilizes products from the latter. The dashed boxes represent isoflavonoids and terpenoids, showcasing specific examples within the realm of specialized metabolism. 3PGA, 3-phosphoglycerate; E4P, Erythrose-4-phosphate; Trp, Tryptophan, Phe, Phenylalanine; 4HPP, 4-hydroxyphenylpyruvate; PEP, Phosphoenolpyruvate; 3MOB, 3-methyl-2-oxobutanoate; BCAA, Branched-chain amino acid; MEP, methylerythritol phosphate; MVA, Mevalonate; TAG, triacylglycerol.

Enzymatic modifiers unleashing phytochemical diversity

The pivotal role of enzymatic reactions in shaping the remarkable diversity of phytochemicals, including isoflavonoids and triterpenes, is evident. Among these reactions, glycosylation stands out as a prominent process facilitated by UDP-glycosyltransferases (UGTs), which belong to the glycosyltransferase superfamily family 1. UGTs utilize uridine diphosphate sugar as a sugar donor, enabling the transfer of glycosyl groups from nucleotide diphosphate-activated sugars to a diverse array of substrates, including specialized metabolites (Brazier-Hicks et al., 2018). This enzymatic pathway plays a vital role in expanding the repertoire of phytochemicals through glycosylation reactions. These modifications have a significant impact on the solubility, chemical properties, bioavailability, stability, and biological activity of these phytochemicals (Pollak et al., 1993). An example is the prevalence of glycosylated forms, such as daidzin, glycitin, and genistin, which are more abundant than their aglycone counterparts. These glycosylated forms play a crucial role in regulating the interactions between legumes and their symbiotic and pathogenic microorganisms, as well as influencing the dietary effects of these flavonoids on human health (Liu et al., 2002, Yu et al., 2000, Yu et al., 2003).

The process of glycosylation alters the physiological, chemical, and biological characteristics of triterpenes, making UDP-glycosyltransferases (UGTs) an intriguing target for metabolic engineering (Bönisch et al., 2014, Rahimi et al., 2019). While a small number of UGT enzymes have been discovered to facilitate the glycosylation of triterpene aglycones, the understanding of UDP-glycosyltransferases (UGTs) in triterpenes lags behind that of isoflavonoids. This is because the majority of UGTs still

lack comprehensive understanding regarding their biochemical functions and substrate specificities (Rahimi et al., 2019). Glycosylation of hydroxyl and/or carboxyl groups in triterpenoids leads to the production of diverse triterpenes. As a result, the expansion of large UGT multigene families reflects the chemical diversification of plants and their adaptations to terrestrial life (Caputi et al., 2012, Yonekura-Sakakibara and Hanada, 2011). However, further research is needed to bridge the knowledge gap and gain deeper insights into UGTs' role in triterpene metabolism.

The extensive diversity of plant specialized metabolites remains largely unexplored, presenting a vast realm of untapped potential (**Figure 1.2**). Even subtle shifts in genetic diversity can have profound implications on specialized metabolism, giving rise to the synthesis of novel molecules whose biological activities are largely unknown (Firn and Jones, 2000). This diversity of metabolites holds significant ecological and practical value for both natural and cultivated ecosystems (Bustos-Segura et al., 2017). Nevertheless, limited research has focused on investigating the broader connections between genetic diversity and chemical diversity within species, particularly examining the variation in chemical compound composition among individuals within a population (Pais et al., 2018). Therefore, it is imperative to elucidate the genetic foundations underlying this diversity.

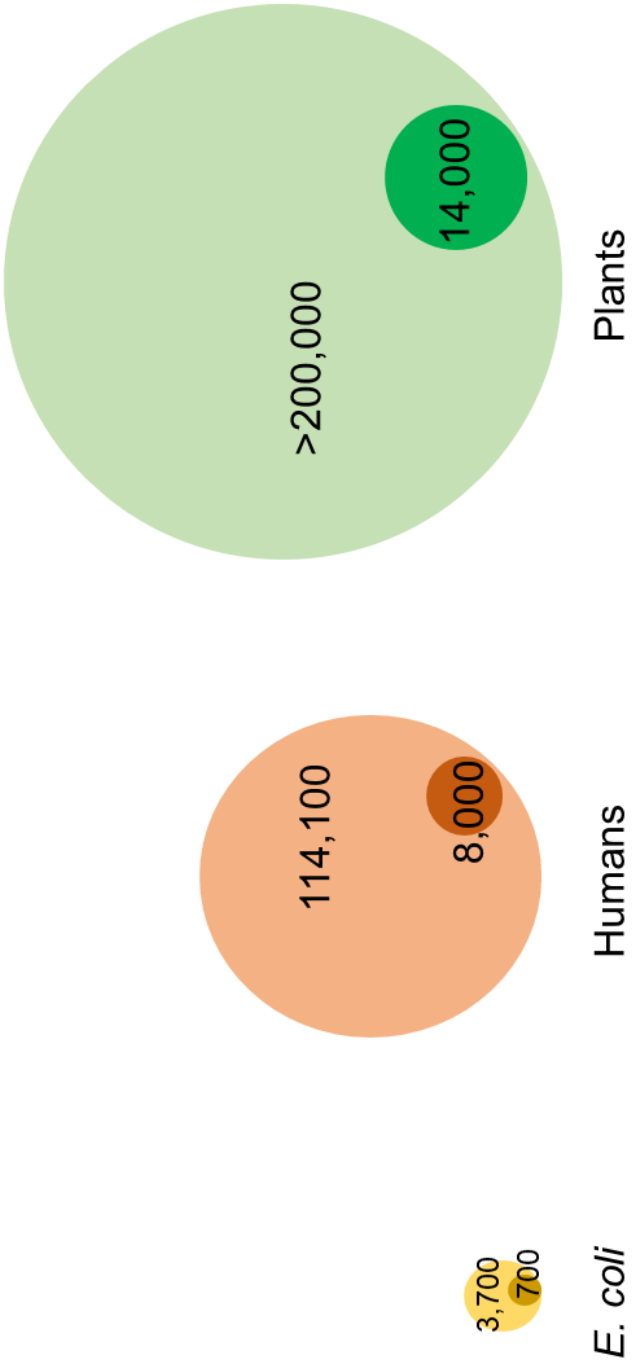


Figure 1.2 A comparative overview of the extensive diversity and scope of plant specialized metabolites. The information has been adapted from Alseekh et al. (2018)

(Alseekh and Fernie, 2018). The larger circles indicate an estimate of the number of metabolites present in *E. coli* (3,700), humans (114,100), and plants (>200,000). Additionally, the smaller circles represent the approximate number of metabolites we can measure (700 for *E. coli*, 8,000 for humans, and 14,000 for plants). The figures regarding metabolite numbers are derived from various studies, including Guo et al. (2012) and Sajed et al. (2016) for *E. coli* (Guo et al., 2012, Sajed et al., 2016), Wishart et al. (2018) for humans (Wishart et al., 2018), and Dixon et al. (2003), Saito and Matsuda (2010), Afendi et al. (2012), Wink et al. (2015), and Rai et al. (2017) for plants (Dixon, 2003, Saito and Matsuda, 2010, Afendi et al., 2012, Wink, 2015, Rai et al., 2017).

Exploring crop wild relatives for a sustainable agriculture and human nutrition

The legume family (Fabaceae) is of great significance, encompassing numerous cultivated species worldwide such as soybeans, peas, lentils, and chickpeas. These plants are rich reservoirs of essential nutrients such as proteins, fats, carbohydrates, dietary fibers, B-group vitamins, and minerals, underscoring their immense nutritional value (Bourion et al., 2018, Fabbri and Crosby, 2016). Legumes are low in fat and devoid of saturated fat, offer a cholesterol-free option. A single serving of legumes, equivalent to half a cup, supplies approximately 115 calories, 20 grams of carbohydrates, 7-9 grams of fiber, 8 grams of protein, and 1 gram of fat. Additionally, they exhibit a low glycemic index, typically falling within the range of 10 to 40 (Polak et al., 2015). During the domestication of legume crops, specific traits related to nutrient content were selectively enhanced. Adzuki beans, for example, exhibit increased starch and fat content (Yang et al., 2015), while soybeans gained attention for their elevated protein and oil content (Zhou et al., 2015). Extensive research has delved into the genetic basis of these traits, unveiling novel loci within the genome (Obala et al., 2019, Patil et al., 2018, Leamy et al., 2017). In soybeans, for instance, stable QTLs (quantitative trait loci) for seed protein and oil on chromosome 20 have been consistently identified across diverse interspecific populations, offering valuable insights (Patil et al., 2018, Lu et al., 2016). A recent study conducted by Goettel et al. (2022) highlighted the role of the *POWR1* gene in influencing a major QTL related to protein and oil content in soybean. The findings of this study revealed that the *POWR1* gene has a substantial impact on enhancing seed quality and yield in soybean (Goettel et al., 2022). Identifying favorable haplotypes/alleles within these regions can significantly accelerate seed improvement efforts.

Legumes are widely recognized for their ability to accumulate a wide range of phytochemicals, making them an invaluable and captivating subject of study. This is especially true due to the abundant genetic diversity found in wild legumes. Thus, wild legumes play a critical role as a substantial source of diverse phytochemicals, which serve as valuable components in improving human health, offering opportunities for disease prevention and the development of alternative medicines (**Figure 1.3**) (Grela et al., 2017). For instance, the isoflavonoid formononetin, commonly found in wild legume species, exhibits anti-hyperglycemic activity, holding promise for diabetes treatment by reducing insulin resistance and hyperglycemia (Qiu et al., 2017, Oza and Kulkarni, 2018). Moreover, wild soybean *(Glycine soja)* demonstrates high levels of phenolics, flavonoids (Takahashi et al., 2016), and saponins (Takahashi et al., 2016, Takahashi et al., 2017), which contribute to its resilience under environmental stresses (**Figure 1.3**). In both soybean and its wild relative, *G. soja*, the induction of phytoalexin glyceollins has been observed (Yasmin et al., 2022), demonstrating notable anti-proliferative effects on breast cancer cells as well as pathogen resistance such as *Phytophthora sojae* and soybean interaction (Lecomte et al., 2017, Jahan et al., 2020). Furthermore, the wild relatives of chickpeas have been found to contain health-beneficial phytochemicals such as polyphenolics and flavonoids, demonstrating their potential value (Von Wettberg et al., 2018).

These findings highlight the exceptional range of phytochemicals found in cultivated legumes, with particular emphasis on their wild relatives. Wild soybeans offer a plethora of beneficial compounds such as soyasaponins, soybean agglutinin, bioactive peptides, lunasin, genistein, and formononetin. These compounds have been found to possess

various health-promoting properties, including anti-inflammatory, antioxidant, vasodilation, and anticancer activities (Jing et al., 2017, Sureda et al., 2017, Zhu et al., 2018a). These compounds showcase remarkable resilience when confronted with dynamic environmental stresses (abiotic and biotic), underscoring their potential not only in enhancing human nutrition but also in promoting sustainable agriculture (Morrissey and Osbourn, 1999, Fujimatsu et al., 2020, Marone et al., 2022). However, under stressful environmental conditions, legume yields often experience significant declines. One crucial factor contributing to this challenge is the loss of genetic diversity during the domestication process, which has limited the adaptability of cultivated crops (Zhang et al., 2017b, Bai and Lindhout, 2007, Gorim and Vandenberg, 2017, Prasanna, 2012). Furthermore, compared to cereals, the legume community faces funding limitations, impeding extensive research efforts for improving legumes. Fortunately, emerging evidence suggests that wild relatives of leguminous plants harbor a wealth of novel alleles with tremendous potential for enhancing crops (Kofsky et al., 2018). This revelation provides an intriguing pathway for further exploration, offering immense potential to advance human health and address the challenges posed by dynamic environmental changes.

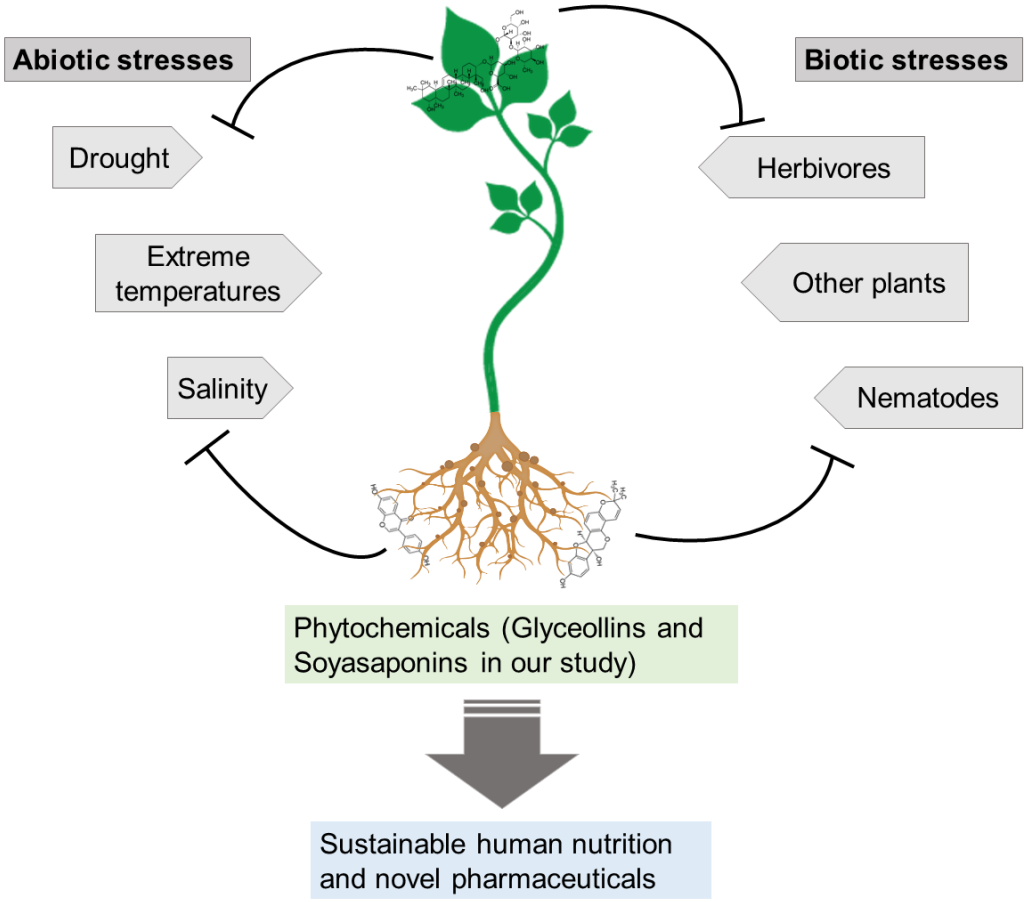


Figure 1.3 The dynamic roles of phytochemicals in sustainable agriculture and human nutrition. When plants experience biotic and abiotic stresses, they have the ability to accumulate a significant number of phytochemicals. This accumulation aids plants in developing resistance to environmental stresses. These phytochemicals also offer various health benefits to human. Thus, the high accumulation of phytochemicals in plants not only contributes to their own resilience but also offers advantages for human well-being.

Approaches to investigate wild relatives for sustainable solutions

It is estimated that global agricultural output must double by 2050 to adequately meet the growing demand for food (Ray et al., 2013). However, the current pace of crop production growth is insufficient to keep up with the pressing issue of food security, primarily due to two key factors. Firstly, modern crop cultivars possess limited capacity to adapt to challenging environments or changing climatic conditions, which can be largely attributed to their low genetic diversity resulting from the domestication bottleneck. Secondly, agricultural land is steadily diminishing as a consequence of urbanization, industrialization, and the increasing demand for animal production, leading to intensified competition for available land resources (Alexander et al., 2015).

On the other hand, crop wild relatives (CWRs), which are the precursors of modern cultivated crops, possess the remarkable ability to thrive in various challenging environments. A substantial proportion of CWRs remains insufficiently investigated, underutilized, and inadequately conserved, however. These wild relatives contain a vast reservoir of genetic diversity that can provide valuable genes and alleles for the breeding and development of phytochemicals-fortified crop varieties with enhanced resilience to demanding growing conditions (Zhang et al., 2019, Zhuang et al., 2022). It is crucial to recognize that only a small portion of the vast array of plant metabolites has been comprehensively investigated regarding their potential for sustainable agriculture and human nutrition. Additionally, the research focused on unraveling the genetic mechanisms responsible for the synthesis or activation of these compounds is currently lacking in quantity. In recent decades, substantial endeavors have been directed towards the exploration and utilization of the genetic variability inherent in CWRs to enhance

crop improvement focusing plant specialized metabolic pathways. **Figure 1.4** illustrates strategies toward plant specialized metabolic pathway gene discovery leveraging omics approaches.

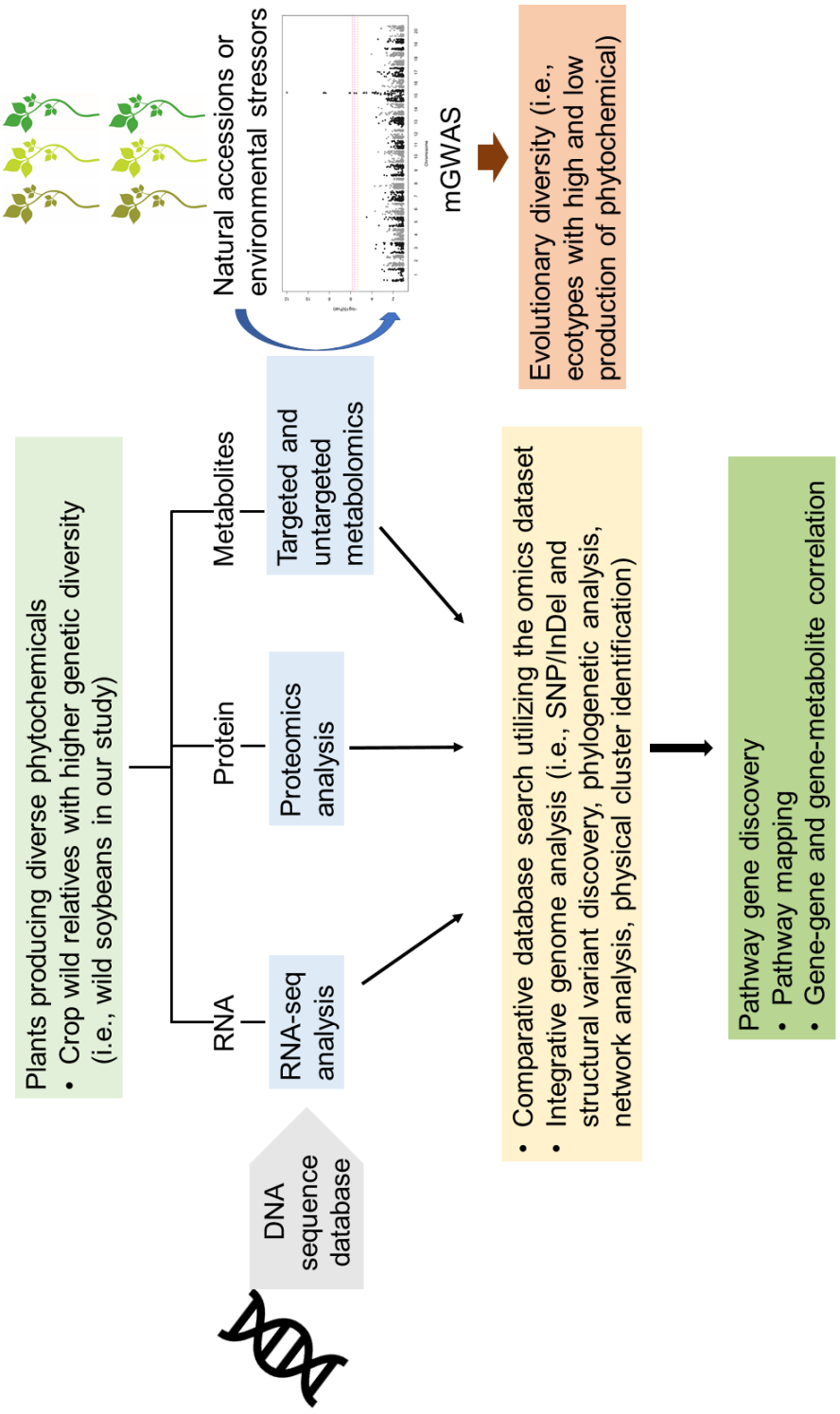


Figure 1.4 Advancing the understanding and manipulation of the genetic foundations underlying a wide range of phytochemicals through the application of systems biology

approaches adapted from Dixon (2001), Kim et al. (2015) and Kangmei et al. (2022) (Dixon, 2001, Zhao and Rhee, 2022, Kim and Buell, 2015).

This study focuses on investigating the genetic underpinnings of phytochemical diversity in natural population and their induction in response to biotic stressors, utilizing the genetic diversity present in wild soybeans. To achieve this, we employed a metabolite-based genome-wide association (mGWA) approach. The specific aims of this research are listed below.

Aim-1: Genetic basis and selection of phytochemical accumulation upon biotic stress.

After encountering biotic elicitors, plants have the ability to synthesize small antimicrobial compounds called phytoalexins as a defense mechanism (VanEtten et al., 1994, Ahuja et al., 2012, Dakora and Phillips, 1996). Within the Leguminosae family, soybeans serve as a prominent source of glyceollins, a type of phytoalexin that provides resistance against pathogens (Keen et al., 1989). The molecular mechanisms by which these elicitors induce the production of glyceollin are still not fully understood. Furthermore, the potential involvement of gene clusters in the synthesis of glyceollin in response to elicitor induction represents a novel and unexplored area of research. In chapter 2, our objective was to investigate the genetic basis and selection for the conversion of daidzein to glyceollin, shedding light on this crucial branch-point in the glyceollin biosynthetic pathway.

Aim-2: Exploring the genetic basis of phytochemical diversity in wild soybeans with a specific focus on soyasaponin biosynthesis pathway.

In order to explore the genetic underpinnings of the natural variation in phytochemical diversity found in wild soybean *(Glycine soja),* we employed a metabolite-based genome-wide association study (mGWAS). In chapter 3, we investigated eight QTLs-multiple

metabolite clusters and identified candidate genes associated with specific plant specialized metabolic pathways. Through this analysis, we aimed to showcase the underexplored yet significant role of higher genetic diversity of wild soybean in shaping phytochemical diversity. Additionally, we focused on the soyasaponin biosynthesis pathway, aiming to uncover the genetic basis behind the accumulation patterns of soyasaponins in wild soybeans. Despite extensive studies on the biosynthetic pathway, structural diversity, composition, and distribution of soyasaponins in soybean, the molecular mechanisms governing the variation in soyasaponin production remain elusive (Kurosawa et al., 2002, Sayama et al., 2012, Shibuya et al., 2010, Yano et al., 2017). Gaining a comprehensive understanding of these molecular mechanisms is crucial as it represents the initial step towards developing soybean crops with enhanced value in the future.

CHAPTER 2: GENETIC BASIS AND SELECTION OF GLYCEOLLIN INDUCTION IN WILD SOYBEAN

***YASMIN, F.[#]**, ZHANG, H.[#], LEAMY, L., WANG, B., WINNIKE, J., REID, R. W., BROUWER, C. R. & SONG, B.-H.[*] 2022. Genetic basis and selection of glyceollin induction in wild soybean. bioRxiv**, 2022.12.17.520864. (submitted to Frontiers in Plant Science; [#]Equal contribution; [*]Corresponding author)*

Author Contributions

BHS conceived the idea and initiated the project. FY, HZ, LL, BW, JW, RR, and CB performed experiments and analyzed data. FY, HZ, LL, BW, and BHS wrote and improved the manuscript. All authors have read and agreed to the published version of the manuscript.

Introduction

Plants produce diverse specialized metabolites (also known as secondary metabolites or phytochemicals), which play a vital role in adapting to changing environments. Phytoalexins are specialized metabolites synthesized *de novo* in response to various biotic and abiotic stresses. Examples include indole alkaloid camalexin in Arabidopsis, phenolic aldehyde gossypol in cotton, phenylpropanoid stilbenes in grapevines, isoflavonoid-derived glyceollins in legume, and momilactones and phytocassanes terpenoids in rice (Jahan et al., 2019, Donnez et al., 2011, Jeandet et al., 2002, Jeandet et al., 2020, Wang et al., 2009, Saga et al., 2012, Yamamura et al., 2015). Isoflavonoids have become a research focal point due to their various pharmacological properties and essential roles in plant defense. The major isoflavones in soybeans are genistein, daidzein, and glycitein, and they make up about 50%, 40%, and 10%, respectively, of the

total isoflavone content. Trace amounts of glyceollins are induced transiently with abiotic and biotic stresses (Jahan et al., 2019, Subramanian et al., 2006). They have multiple effects, including fostering symbiosis between soybean and *Bradyrhizobium japonicum* and inhibiting the growth of various microbes (Subramanian et al., 2006, Graham and Graham, 1996). Moreover, they have anti-cancer, antioxidant, and neuroprotective properties (Kim et al., 2012, Bamji and Corbitt, 2017, Nwachukwu et al., 2013, Seo et al., 2018). However, studies on glyceollins are mainly focused on their medicinal properties, while little is known about how their induction is regulated.

Phytoalexins have been considered the target of natural selection due to their activities in biotic and abiotic stress responses in natural environments (Miyamoto et al., 2016, Qi et al., 2004, Pichersky and Gang, 2000). Therefore, in our study, we chose wild soybean (*Glycine soja*), a wild relative of soybean (*Glycine max)*, to delineate genetic basis and evolution of glyceollin accumulation resulting from biotic stress, i.e., soybean cyst nematode (SCN), the most devastating soybean pest worldwide (Tylka and Marett, 2021). Wild soybeans thrive in diverse habitats and harbor much higher, underexplored genetic diversity than cultivated soybeans (Zhang et al., 2019). Hence, it is an ideal system to understand the genetic basis and evolution of glyceollin variation. Eventually, the essential genes identified in wild soybean can be used for metabolic engineering or in a breeding program to develop nutrition-rich biofortified soybean cultivars as they exhibit similar genome size and content with small reproductive isolation (Singh and Hymowitz, 1999).

A metabolic gene cluster is a group of (two or more) genomically co-localized and potentially coregulated non-homologous genes that encode enzymes involved in a

particular metabolic pathway (Töpfer et al., 2017, Nützmann et al., 2016). They have been a common phenomenon since the early days of microbial genetics (Zheng et al., 2002, Rocha, 2008, Koonin, 2009). However, gene clusters in plant metabolic pathways have been discovered only recently, even though microbes and plants are both extremely rich sources of metabolic diversity. A study by Chae et al. (2014) on metabolic gene clusters in Arabidopsis, soybean, sorghum, and rice suggested that approximately one-third of all the metabolic genes in Arabidopsis, soybean, and sorghum, and one-fifth in rice were rich in gene clusters across primary and specialized metabolic pathways (Chae et al., 2014). There is compelling evidence indicating that the highly plastic plant genome itself generates metabolic gene clusters via gene duplication, neofunctionalization, divergence, and genome reorganization instead of horizontal gene transfer from microbes (Osbourn and Field, 2009). This suggests that plants rewire their genome to gain new adaptive functions driven by the need to survive in distinct environments. Mining and functional validation of the candidate genes in such clusters will facilitate the discovery of new enzymes and chemistries that render pathway prediction. Moreover, metabolic gene clusters are likely to be located within dynamic chromosomal regions, and thus, many identified so far may be due to recent evolution (Qi et al., 2004, Field et al., 2011, Matsuba et al., 2013). If so, investigation of these clusters can provide insights into their evolutionary history. The vast and diverse array of specialized metabolites that are produced through multi-step metabolic pathways play an important role in plant adaptation to various ecological niches. However, the occurrence, prevalence, and evolution of such gene clusters in plants are largely unknown. Thus, the study of plant metabolic gene clusters has implications for molecular biology and evolutionary

genomics (Nützmann et al., 2016, Yeaman and Whitlock, 2011, Takos and Rook, 2012, Chavali and Rhee, 2018).

Due to the extraordinary metabolic diversity, to date, less than 50 plant-specialized metabolic pathways have been biochemically and genetically identified (Nützmann et al., 2016). Metabolomic GWAS (mGWAS) offers an effective approach to understand the genetic basis of metabolites and their associated traits (Luo, 2015, Riedelsheimer et al., 2012, Chan et al., 2011, Chan et al., 2010). mGWAS allows the identification of common polymorphic regions controlling complex metabolic traits by substantially increasing association panel and genome-wide molecular markers. Besides elucidating genetic architecture, mGWAS can also be used to infer gene functions (Luo, 2015). Hence, mGWAS provides a comprehensive approach to discovering candidate genes. Thus far, it has been used to uncover the genetic basis of variations of a number of different metabolites. For example, Chen et al. (2014) carried out a rice mGWA study that identified 36 candidate genes influencing the variation of metabolites with physiological and nutritional importance (Chen et al., 2014b).

The isoflavonoid pathway has been relatively well studied (**Figure 2.1**) (Sukumaran et al., 2018, Yoneyama et al., 2016). However, it is still not clear how glyceollin induction is regulated. This study is the first to employ genomic and evolutionary approaches to understand the genetic basis and selection of glyceollin induction. Our study provides a fundamental basis for the long-term goal of developing glyceollin-fortified soybean cultivars that would improve plant and human health to meet current and future global challenges. In this study, we aim to address these three questions: (1) What is the genetic basis of variation in glyceollin induction by SCN? (2) Are there any gene clusters and

transcription factors involved in glyceollin variation? (3) Are epistatic interactions and natural selection important evolutionary factors influencing the variation of glyceollin induction?

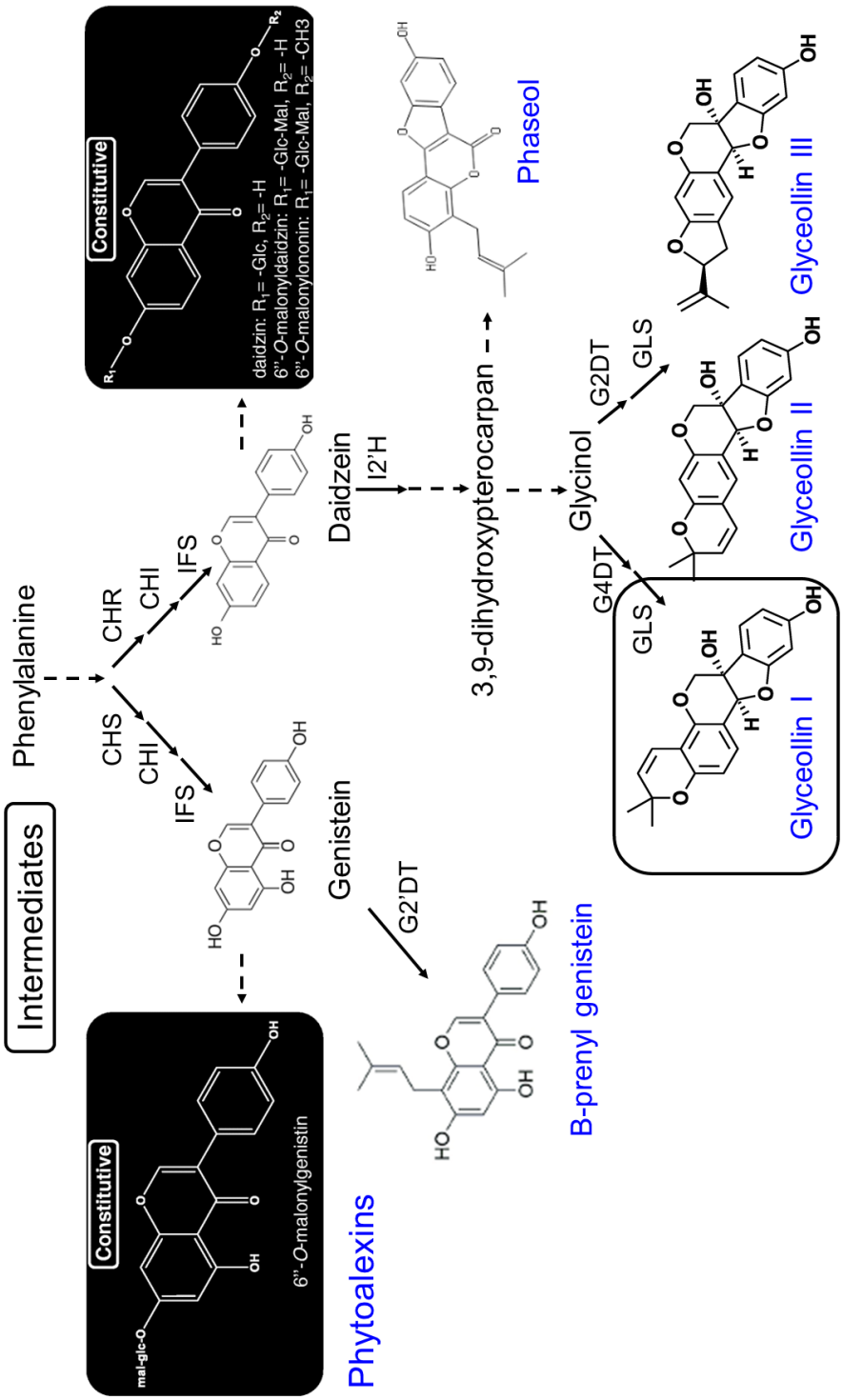


Figure 2.1 The biosynthetic pathway of isoflavonoids in soybeans involves the production of inducible phytoalexins (marked in blue text) and constitutively synthesized

isoflavone conjugates (highlighted within rectangular boxes shaded in black), both of which originate from the isoflavone intermediates daidzein or genistein. The figure has been adapted from the study conducted by Jahan et al. (2009) (Jahan et al., 2019).

Materials and methods

Plant materials

A total of 264 accessions of wild soybean, *Glycine soja,* from a wide geographic range, originally collected from China, Japan, Russia, and South Korea, were utilized (**Supplementary Table 2.1**). The seeds of these ecotypes were obtained from the USDA national germplasm resources laboratory (https://www.ars-grin.gov/).

Plant preparation, SCN inoculation, and sample collection

Seed preparation, germination, transplanting, and soybean cyst nematode (SCN, *Heterodera glycines Ichinohe*, HG type 1.2.5.7) inoculation were performed following a previously developed protocol (Zhang and Song, 2017, Zhang et al., 2017c, Zhang et al., 2017a). Whole root tissues were collected and weighed five days post-infection (dpi). The 5 dpi time point was chosen because our previous study suggested a significant inhibition in SCN development in a resistant genotype compared to normal growth in a susceptible genotype (Zhang et al., 2017a, Zhang et al., 2017c). All samples were flash frozen in liquid nitrogen and stored at -80 °C. Four biological replicates per wild soybean ecotypes were used, eventually a total of 1,020 samples.

Metabolite extraction and quantification

We employed the extraction method of metabolites from root tissue described in (Strauch et al., 2015). The metabolite profiling was provided by the service from David H. Murdock Research Institute at the North Carolina Research Campus employing UPLC-MS/MS (ultraperformance liquid chromatography-tandem mass spectrometry). Peaks that were consistently detected in at least three biological replicates within each ecotype were used for downstream analyses. Each metabolite was confirmed using pure standard

compounds, including daidzein, daidzein-d6, and glyceollin. Due to the low concentrations of these compounds and the small sample masses of the wild soybean root samples that had been collected, we used a signal-to-noise ratio of ≥10 for the measurement of the peaks for glyceollin and daidzein. Our method successfully measured daidzein (μg/g root) and glyceollin (unitless) in 264 accessions of wild soybean *G. soja* roots quantitatively and semi-quantitatively, respectively. Following method development, optimization, and analyses of the test samples, calibration curves were designed using at least six different concentrations of daidzein, created in triplicate to quantify known concentrations of daidzein and glyceollin. A second-degree polynomial was derived from the known concentrations of the standard curve samples and the mass spectrometer response (daidzein/internal standard) from the standard curve data. The resulting polynomial was used to calculate the concentrations of daidzein in the experimental samples. Low, medium, and high QC (quality control) samples were created to assess the accuracy of the calculations. We used the ratio of glyceollin (unitless) to daidzein (μg/g root) (GVSD) as our phenotypic trait. This phenotype henceforth is denoted GVSD.

Genotypic data

Genotype data for the 264 accessions were obtained from SoySNP50K (Song et al., 2013), which included 32,976 genome-wide single nucleotide polymorphic markers (SNPs) with a minor allele frequency (MAF) of at least 5%.

Metabolite-based genome-wide association study and linkage disequilibrium estimation

Our genome-wide association analysis was conducted on GVSD (a ratio of glyceollin mean to daidzein mean) in response to SCN infection on all 264 ecotypes using the BLINK algorithm implemented in the GAPIT R package (2.0) (Tang et al., 2016). To minimize false-positive associations, we controlled population structure among genotypes with four principal components. Heritability estimate and SNP effect were calculated by running GWAS applying CMLM and MLM methods, respectively, implemented in the GAPIT R package (2.0) (Tang et al., 2016).

A conventional Manhattan plot was generated using the qqman R package to visualize the SNPs (Turner, 2014). In addition to the genome-wide significant threshold, we also calculated the chromosome-wide Bonferroni thresholds using independent SNPs estimated on each chromosome following the method of Li and Ji (2005) (Li and Ji, 2005). Linkage disequilibrium (LD) was calculated across the panel with the TASSEL program, version 5 (Bradbury et al., 2007), for the significant SNPs identified from the GWAS analysis. LD was measured using squared correlation R-squared (r^2) of 0.2 (upper right in the LD plot) and P-value < 0.05 (the lower left in the LD plot). A pairwise LD was generated following the R function described by Shin et al. (2006) (Shin et al., 2006). Genes within LD blocks containing significant SNPs were identified as potential sources of candidates for further analyses.

Identification of candidate genes

For extensive gene mining of our identified gene pool, we used an array of bioinformatics tools. Such an approach can improve the accuracy of candidate gene and gene cluster predictions and resolve inconsistencies among the bioinformatics tools (Chavali and Rhee, 2018). Specifically, a pairwise linkage disequilibrium (LD) analysis was initially

used for potential candidate gene identification. Then, genes in each LD block were examined as potential candidate genes, and their annotations were obtained from the Phytozome v13 database (Goodstein et al., 2011). Afterward, a GO enrichment analysis of the identified candidate genes was performed using ShinyGO v0.66: Gene Ontology Enrichment Analysis (P-value cut-off (FDR, false discovery rate) = 0.05) (Ge et al., 2020), SoyBase GO Enrichment Data (Grant et al., 2010). To investigate the involvement of these potential candidate genes in metabolic pathways, a database search was performed through an annotation file from Phytozome v13 (Goodstein et al., 2011), SoyBase (Grant et al., 2010), SoyCyc 10.0 Soybean Metabolic Pathway (Hawkins et al., 2021), and Pathview databases (Luo et al., 2017). Finally, a PMN plant metabolic cluster viewer was applied to categorize enzymes into classes (signature or tailoring) and metabolic domains (Hawkins et al., 2021).

Analysis of epistatic interactions

For any significant SNPs uncovered in the GWAS analysis, it is useful to test whether, beyond their direct effects, they also exhibited interactive effects on GVSD. To accomplish this, we first produced numerically formatted genotypes, in which the homozygous genotype index value is 1 and -1 and the heterozygous 0. This allows us to test for epistasis for each pairwise combination in a simple general linear model with 1 degree of freedom for the additive effects of each of the two SNPs and their interaction. We included the first four principal components from the GAPIT analysis in the model to be consistent with the GWAS scan, where these components were used to adjust for structural relatedness (see below). The significance of all interactions was evaluated with the sequential Bonferroni procedure. To illustrate the interactions of SNP pairs, we also

calculated regressions of GVSD on each SNP, but at each of the three genotypes (using the -1, 0, and 1 index values) of the second SNP involved in the significant interaction.

Extended haplotype homozygosity analysis

To test allele-specific selection patterns of the identified significant SNPs, we analyzed extended haplotype homozygosity (EHH, (Sabeti et al., 2002)) for each significant SNP. The EHH analysis was conducted in SELSCAN v.1.2.0a (Szpiech and Hernandez, 2014) with default parameters, and only SNPs with MAF > 0.05 was used in this analysis.

Results

Genomic dissection of glyceollin accumulation upon biotic interaction

We identified a total of eight significant SNPs, with four located on chromosome 9 and the others on chromosomes 3, 13, 15, and 20 (**Figure 2.2A, Table 2.1**). These SNPs were identified based on both genome-wide Bonferroni threshold of 5.104 and chromosome-wide Bonferroni thresholds that varied narrowly from 3.79 to 3.82 among the 20 chromosomes (3.803 on chromosome 9) (**Figure 2.2A, B**, **Supplementary Table 2.2**). The Manhattan and Q-Q (quantile-quantile) plots are shown in **Figure 2.2A, B, C**. The four significant SNPs on chromosome 9 are located close to each other within a 535 kb region (**Supplementary Table 2.2**). The broad-sense heritability (h^2) was estimated at 35% (**Supplementary Table 2.2**).

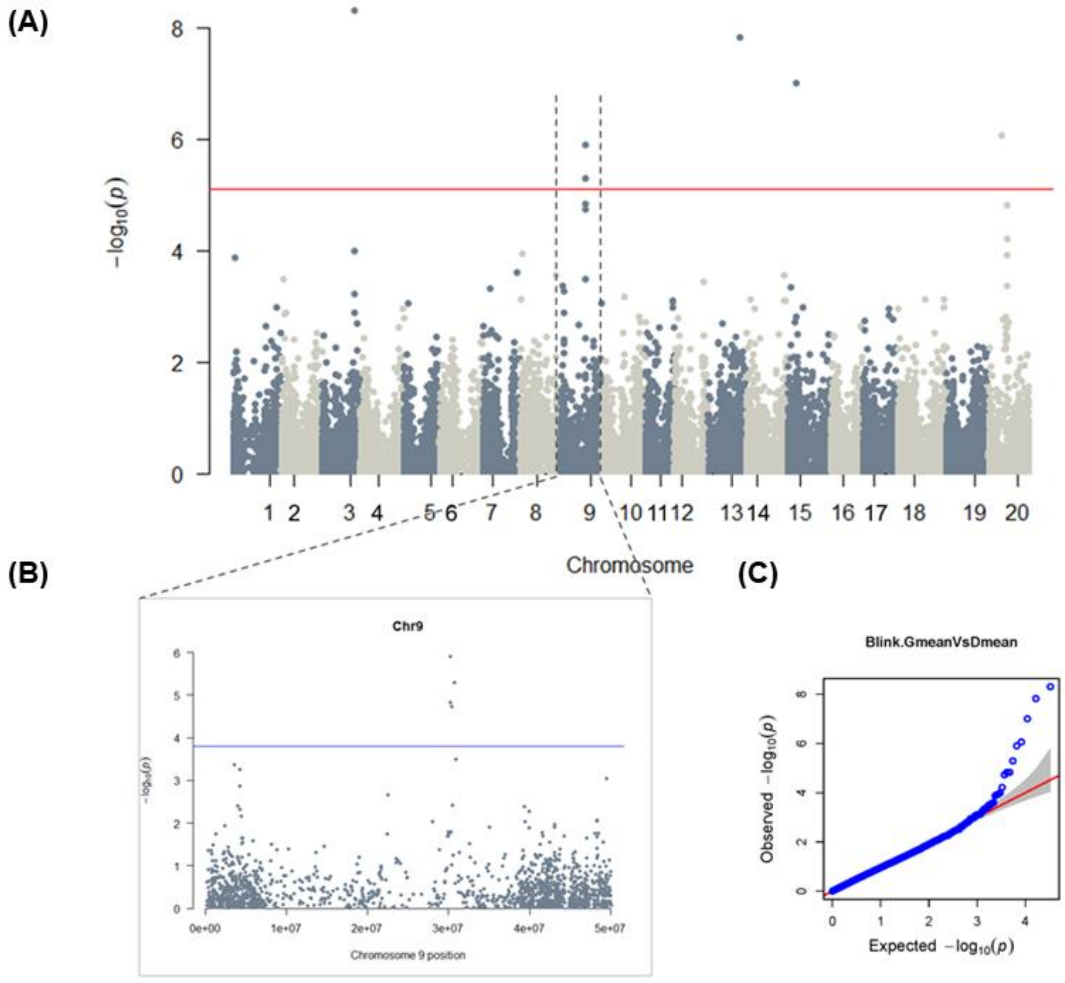


Figure 2.2 GWAS of glyceollin induction with SCN stress: A genome-wide **(A)** and chromosome-wide **(B)** Manhattan plots, with thresholds of 5.104 and 3.803, respectively; **(C)** quantile-quantile (QQ) plot. Significant SNPs are found on chromosomes 3, 9, 13, 15 and 20 at a 5% genome-wide threshold, the probability of 7.86×10^{-6} resulted in a threshold of 5.01 (solid red line in the genome-wide Manhattan plot) **(A)**. The 5% chromosome-wide LOD threshold resulted in significant P-values of 1.57×10^{-4} (threshold 3.803, solid blue line) **(B)**.

Table 2.1 Identification of significant SNPs and functional annotation of the plausible candidate genes.

Significant SNP	**Chromosome**	**Functional annotation of associated genes**
ss715585948	Gm03	*WRKY* family transcription factor family protein
		Zinc fingers superfamily protein
ss715603454	Gm09	UDP-glucosyl transferase *88A1*
ss715603455	Gm09	RING/U-box superfamily protein, RING/FYVE/PHD zinc
ss715603462	Gm09	finger superfamily protein
ss715603471	Gm09	*WRKY* family transcription factor family protein
		MYB domain
		Zinc fingers superfamily protein
		Cytochrome P450 enzyme family
		Zinc finger, RING-type; Transcription factor
		jumonji/aspartyl beta-hydroxylase
ss715615975	Gm13	*bZIP* transcription factor
		RING/U-box superfamily protein, RING/FYVE/PHD zinc
		finger superfamily protein
		Zinc fingers superfamily protein
		NAC transcription factors
		Cytochrome P450 enzyme family
ss715620269	Gm15	RING/U-box superfamily protein, RING/FYVE/PHD zinc
		finger superfamily protein

		WRKY family transcription factor family protein
		MYB domain
ss715636844	Gm20	UDP-Glycosyltransferase superfamily protein
		UDP-glucosyl transferase *85A2*
		hydroxy methylglutaryl CoA reductase 1
		Cytochrome P450, family 71, subfamily B, polypeptide 34
		cytochrome p450 *79a2*
		RING/U-box superfamily protein, RING/FYVE/PHD zinc finger superfamily protein
		Zinc fingers superfamily protein

Linkage disequilibrium analysis and candidate gene identification

We identified a total of 666 possible candidate genes within the linkage disequilibrium (LD) blocks of the eight significant SNPs (soybean reference genome *Glycine max* Wm82.a2.v1) (Goodstein et al., 2011, Zhou et al., 2015). The LD block on chromosome 9 showed the strongest LD with a long range compared to the others (**Figure 2.3B**, **Supplementary Figure 2.1, Supplementary Figure 2.2**). *W*e considered r^2>0.2 as a cutoff for our LD analysis, *where* r^2 *is the extent of allelic association between a pa*ir of sites (Weir, 1990). Candidate gene *Glyma.09G128200* shows the highest level of LD near the significant SNPs on chromosome 9 compared to the LD block for the rest of the significant SNPs on this chromosome (**Figure 2.3B**, **Supplementary Figure 2.1**). The functional annotation of the candidate genes within this block is biosynthetic enzymes involved in isoflavonoid pathway, as well as regulatory genes such as *WRKY* and *MYB*

transcription factors (**Table 2.1, Supplementary Table 2.3,** and **Supplementary Table 2.4**), which may indicate their transcriptional level involvement in glyceollin induction in response to SCN stress (Colinas and Goossens, 2018).

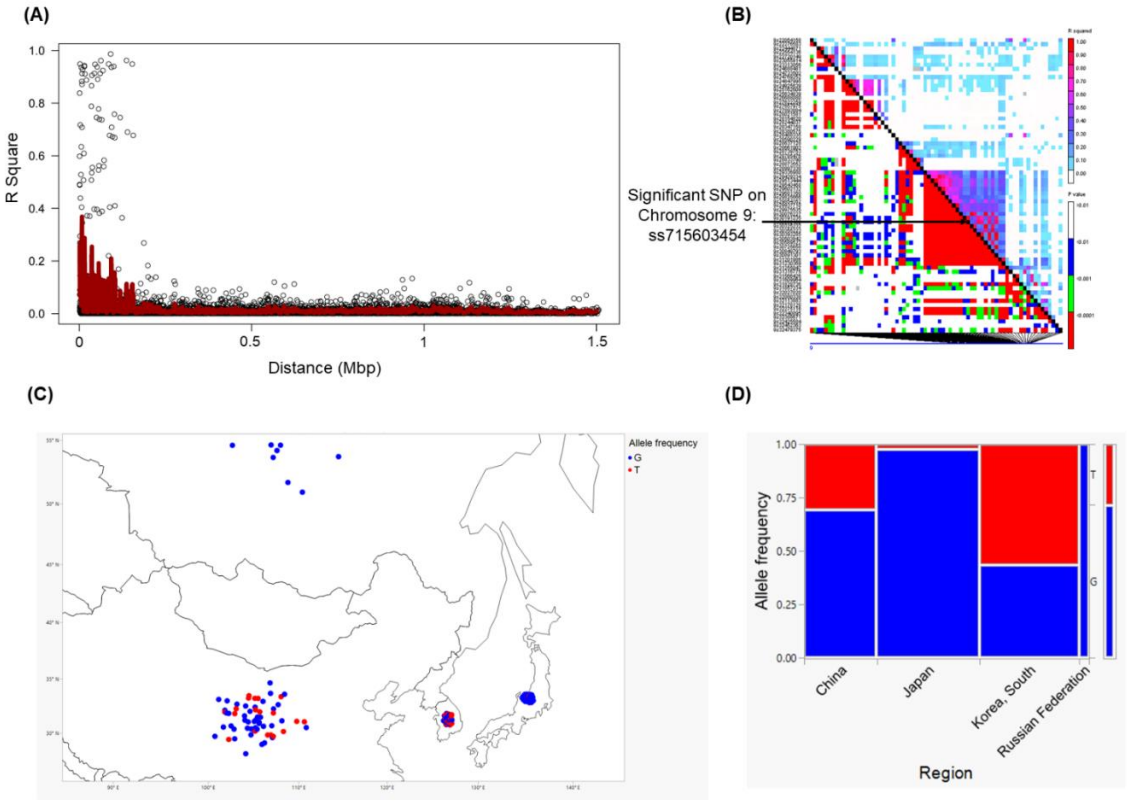


Figure 2.3 Linkage disequilibrium analysis and candidate gene identification. An LD decay measured as R square for pairwise markers and plotted against their distance **(A)** and LD plot for chromosome 9 for significant SNPs **(B)**. The black diagonal denotes LD between each site and itself **(B)**. Geographic range of the alleles of significant SNPs close to the gene clusters on chromosome 9 **(C)**. Allele frequency in each population. Allele frequency in different geographic regions for a significant SNP was generated using JMP®, Version *15*. SAS Institute Inc., Cary, NC, 1989–2021 **(D)**.

We also found putative genes encoding enzymes involved in the specialized metabolic pathways within the LD blocks of the significant SNPs on chromosomes 3, 13, 15, and 20. The enriched GO category includes flavonoid biosynthesis pathway, phenylpropanoid metabolic process, linamarin biosynthesis, and terpenoid biosynthesis (**Supplementary Table 2.5**). Apart from the biosynthetic enzymes on these chromosomes, we also found transcription factor genes, such as *WRKY*, *MYB*, and *NAC* (**Supplementary Table 2.5**).

Metabolic gene clusters identification

We were particularly interested in the candidate genes in the branch from daidzein to glyceollin in the isoflavonoid biosynthesis pathway (Lozovaya et al., 2007). We found that the identified candidate genes on chromosome 9 are clustered together, and they fall into two clusters. These clusters belong to tailoring enzyme glycosyltransferase within phenylpropanoid specialized metabolic domain. And six genes are within the branch of isoflavonoid biosynthesis pathway. Two of these six genes, *Glyma.09G127200* and *Glyma.09G127300*, are called cluster 1, while the rest four (*Glyma.09G127700*, *Glyma.09G128200*, *Glyma.09G128300*, and *Glyma.09G128400*) are called cluster 2 (**Supplementary Table 2.3**).

Following further investigation of annotation of these candidate genes within the gene clusters (**Supplementary Table 2.4**), we found *Glyma.09G127200* gene encodes a glucosyltransferase that may act on 4'-methoxy isoflavones biochanin A, formononetin, 4'-hydroxy isoflavones genistein, and daidzein substrates. However, the enzyme does not act on isoflavanones, flavones, flavanones, flavanols, or coumarins (Köster and Barz, 1981). Within the same cluster, *Glyma.09G127300* has similar annotations and functions as *Glyma.09G127200*. Interestingly, the four genes within cluster 2 have a similar

functional annotation as *Glyma.09G127200* and *Glyma.09G127300* in cluster 1, and all these four genes encode isoenzymes (**Supplementary Table 2.4**). Such a link between these two gene clusters indicates their proximity in the metabolic pathway.

Epistatic interactions among all significant SNPs

The results of the epistasis tests for each of the 28 pairwise combinations of the eight significant SNPs are shown in **Table 2.2**. Three probabilities, all associated with the SNP on chromosome 20, were not estimable (**Table 2.2**). Among the remaining 25 SNP pairs, 21 show statistical significance. Particularly noticeable is the high significance for all interactions of the SNPs on chromosomes 3, 13, and 15. Three of the six pairs among the four SNPs on chromosome 9, all involving ss715603462, are also statistically significant. In general, therefore, this is evidence for substantial epistasis among these SNPs affecting GVSD.

Table 2.2 Epistasis for the eight significant SNPs.

	Ch9a	**Ch9b**	**Ch9c**	**Ch9d**	**Ch13**	**Ch15**	**Ch20**
Ch3	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	0.002*
Ch9a		0.10	0.053	0.007*	<0.001*	<0.001*	0.907
Ch9b			0.012*	0.006*	<0.001*	<0.001*	0.835
Ch9c				<0.0001*	<0.001*	<0.001*	n.e.
Ch9d					<0.001*	<0.001*	n.e.
Ch13						<0.001*	n.e.
Ch15							0.001*

Shown are the probabilities for each pairwise interaction of SNPs. * = $P < 0.05$ from

sequential Bonferroni tests. n.e. = not estimable. Ch3 = ss715585948, Ch9a = ss715603454, Ch9b = ss715603455, Ch9c = ss715603462, Ch9d = ss715603471, Ch13 = ss715615975, Ch15 = ss715620269, Ch20 = ss715636844

These epistatic interactions of the SNP pairs are illustrated in **Figure 2.4** for each of the four chosen combinations. For example, in panel A (**Figure 2.4A**), it can be seen that regression slopes of GVSD on ss715603454 are close to 0 for ss71585948 CC genotype but are positive for TC and especially TT genotypes. In panel D (**Figure 2.4D**), regression slopes of GVSD on ss715603471 are negative for ss715603462 AA and GA genotypes but positive for GG genotypes. With no epistasis, these slopes would be expected to be roughly parallel, but in fact, they diverge considerably from parallelism in these four examples.

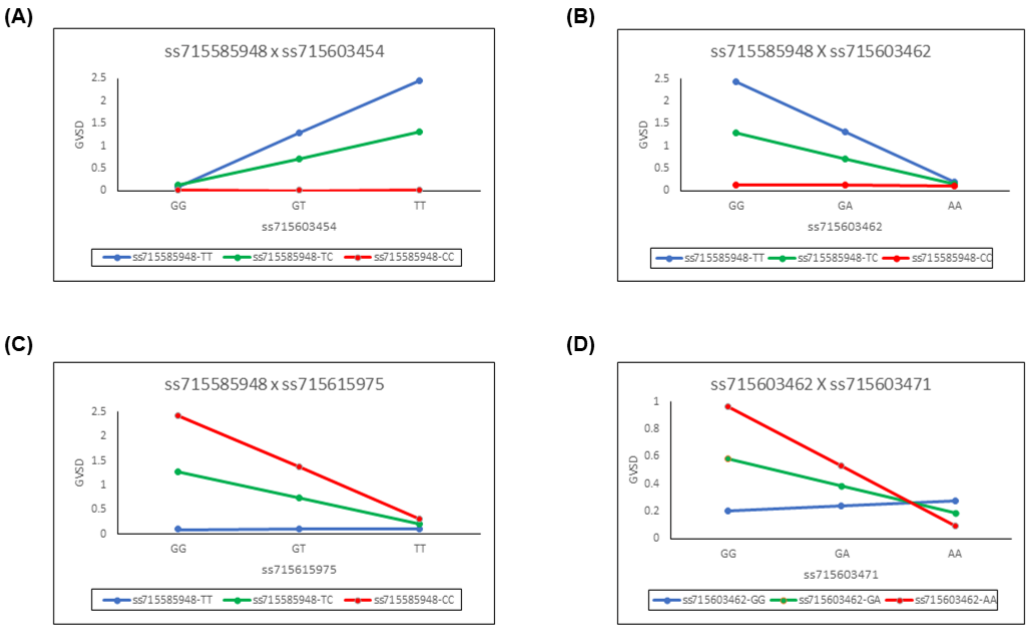


Figure 2.4 Epistatic interactions of the SNP pairs for each of four chosen combinations. Regression slopes of GVSD on ss715603454 are close to 0 for ss715585948 CC genotype but are positive for TC and especially TT genotypes **(A)**. Regression slopes of GVSD on ss715603462 are close to 0 for ss715585948 CC genotype but are negative for TC and especially TT genotypes **(B)**. Regression slopes of GVSD on ss715615975 are close to 0 for ss715585948 TT genotype but are negative for TC and especially CC genotypes **(C)**. Regression slopes of GVSD on ss715603471 are negative in sign for ss715603462 AA and GA genotypes, but positive in sign for GG genotype **(D)**.

Significant SNPs exhibited extended haplotype homozygosity

The extended homozygosity analysis (EHH) analyses revealed allele specific EHH values of the significant SNPs (ss715603454, ss715603455, ss715603462, and ss715603471) on chromosomes 9 (**Figure 2.5**). For example, T allele of ss715603454 showed much higher EHH value than G allele. Alleles of significant SNPs on the other chromosomes showed compatible EHH values (**Supplementary Figure 2.3**).

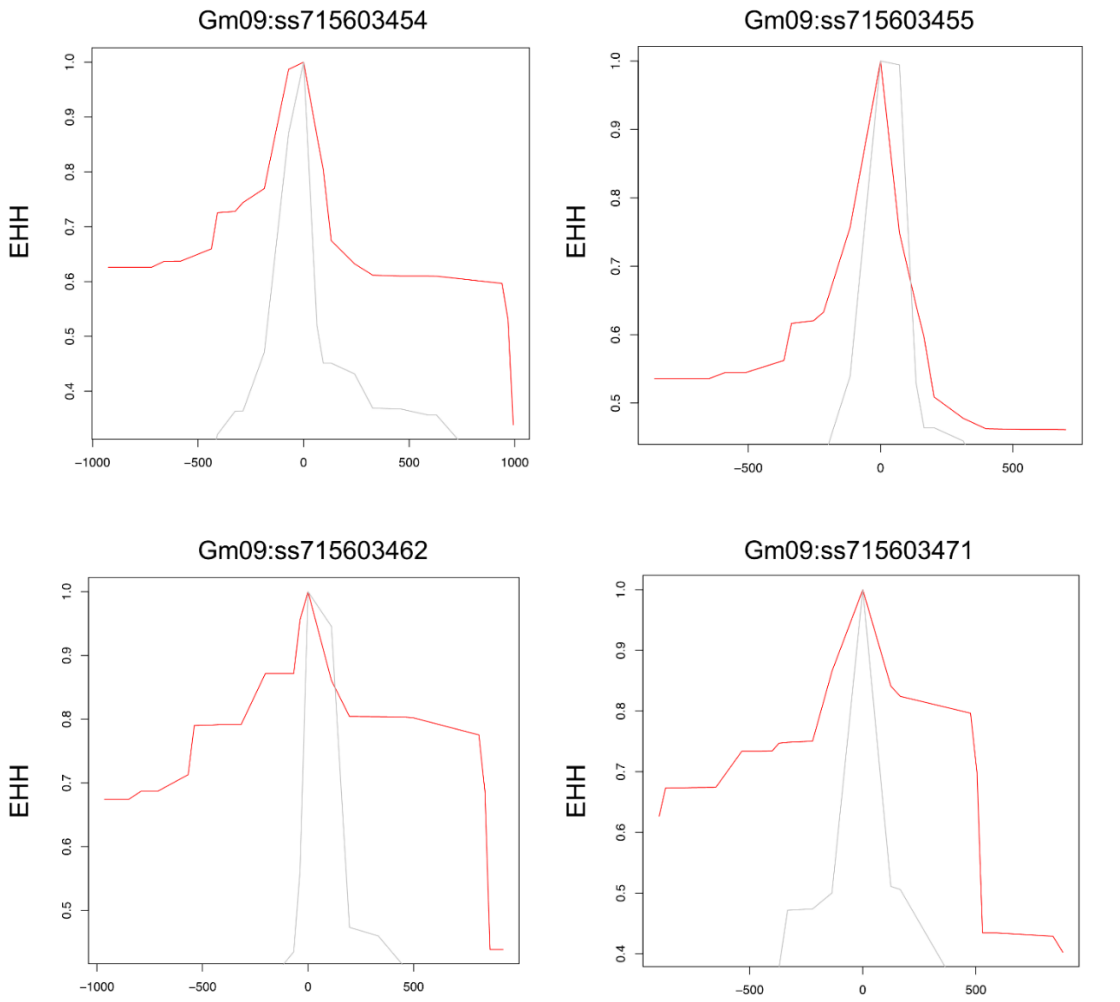


Figure 2.5 Allele-specific Extended Haplotype Homozygosity (EHH) for four significant SNPs on chromosomes 9. Haplotype lengths are shown flanking the T (red) and G (grey) alleles.

Discussion

Metabolic gene clusters in glyceollin induction

Gene clusters have been reported to play important roles in phytochemical diversity in Arabidopsis, sorghum, soybean, tomato and rice (Chae et al., 2014, Fan et al., 2020) as well as their roles in important ecological functions in plants i.e., allelopathic, antibacterial, anti-herbivore, antifungal and insecticidal activities (Polturak and Osbourn, 2021, Polturak et al., 2022). However, their roles in regulating metabolic variation in wild species are relatively less investigated. Even though the isoflavonoid biosynthesis pathway is moderately well studied, the genetic regulation of glyceollin induction is unclear. Particularly, the contribution, prevalence, and occurrence of gene clusters in plant metabolic diversity are largely unclear. Our mGWAS results suggest there are two probable gene clusters with functionally related but non-homologous genes, which may involve in glyceollin induction in wild soybean. Thus far, the genes within these plausible clusters are the first reported candidate genes located on chromosome 9 involved in glyceollin accumulation induced by biotic stimuli in wild soybeans. To date, the reported glyceollin biosynthesis genes are associated with chromosomes Gm01, Gm02, Gm03, Gm04, Gm06, Gm07, Gm10, Gm11, Gm13, Gm15, Gm19 and Gm20 (Jahan et al., 2020, Sukumaran et al., 2018, Yoneyama et al., 2016, Akashi et al., 2009). Our predicted gene clusters suggest that glyceollin may be synthesized where the enzyme-encoding genes are adjacent to each other on the same chromosome (Chavali and Rhee, 2018). Physical clustering of genes with similar functions can facilitate co-inheritance of alleles with favorable combinations and their coordinated regulations at chromatin level (Chu et al., 2011, Osbourn, 2010a). Besides, such clusters incline to locate in the sub-telomeric

regions (Gierl and Frey, 2001, Qi et al., 2004, Sakamoto et al., 2004), near the ends of chromosomes that are known to harbor mutations. For example, an examination of the complete genome sequence revealed that the maize *DIMBOA* cluster is located close to the end of chromosome 4 (Farman, 2007, Jonczyk et al., 2008). Thus, identifying the positions of the genes can contribute to inferences of possible mechanisms underlying chemical diversity in natural populations.

Tailoring enzymes, such as methyltransferases, glycosyltransferases, *CYPs*, dehydrogenases/reductases, and acyltransferases are responsible for modifying the chemical backbone of specialized metabolites (Osbourn, 2010b). The gene clusters we found are associated with tailoring or regulating glycosyltransferase enzymes. A common defense mechanism of plants involves glycosylation of secondary metabolites by involving these enzymes (Mylona et al., 2008). Therefore, the clustering of the genes encoding glycosyltransferase on chromosome 9 indicates the formation of stress-induced (i.e., SCN stress in our study) protective compounds. For example, the cyclic hydroxamic acid *(DIBOA)* in maize (Gierl and Frey, 2001, Frey et al., 1997), the triterpene avenacin in oat (Mugford et al., 2009, Qi et al., 2004, Qi et al., 2006, Field and Osbourn, 2008), and two gene clusters associated with diterpene (momilactone and phytocassane) synthesis in rice, which may be pre-formed or synthesized after stress induction for plant defense. Disruption of such gene clusters may compromise pest and disease resistance and lead to the accumulation of toxic pathway intermediates (Chu et al., 2011). In the multi-step plant specialized metabolic pathways, rapid adaptation to a particular environmental niche could result in highly diverse and rapidly evolving metabolic gene clusters (Osbourn and Field, 2009). Hence, the level of conservation of the identified

gene clusters in this study across different *G. soja* ecotypes can shed light on evolutionary insight of these clusters (Field and Osbourn, 2008). Synthetic biology and functional genetics can further help investigate the organization and contribution of these clusters in metabolite diversity, as well as decipher the mechanism of adaptive evolution and genome plasticity (Osbourn, 2010b, Chu et al., 2011).

Plausible transcriptional factors in glyceollin induction

Advancement of genetics, genomics, and bioinformatic approaches facilitate the prediction and identification of a large number of genes, including transcription factors associated with plant-specialized metabolic pathways (Anarat-Cappillino and Sattely, 2014, Moore et al., 2019). However, the transcriptional regulators of specialized metabolism are less well characterized (Shoji and Yuan, 2021). The regulation of plant specialized metabolic pathways is highly dynamic, responding to the constant changes in the environment. Such regulation generally occurs at transcription level, and thus, it requires coordinated regulation often mediated by transcription factors (TFs) (Colinas and Goossens, 2018, Shoji, 2019). For instance, *MYB* and basic helix-loop-helix *(bHLH)* TF family genes were reported to regulate anthocyanin and related flavonoid biosynthetic pathways in a wide range of species (Chezem and Clay, 2016). Moreover, significant modifications of these regulatory genes give rise to the vast diversity in plant specialized metabolism (Huang et al., 2018, Springer et al., 2019).

It is possible that transcription factors, such as *MYB* and *WRKY* TFs on chromosome 9, may influence glyceollin induction. This indicates regulation of glyceollin induction with SCN stress may involve a highly complex interplay among multiple genes and pathways. Previous studies reported that gene families of transcription factors, such as *NAC*, *MYB*,

bHLH, and *WRKY*, exhibited conservative patterns among Arabidopsis, cotton, grapevine, maize, and rice (Ibraheem et al., 2015, Ogawa et al., 2017, Saga et al., 2012, Xu et al., 2004b, Yamamura et al., 2015, Zheng et al., 2006). These plant species produce various phytoalexins, such as indole alkaloids, terpenoid aldehydes, stilbenoids, deoxyanthocyanidins, and momilactones/ phytocassanes, respectively. This gives rise to the question of whether these TFs are as diversified as the metabolic pathways, or they maintain conservative patterns among species. The investigation of TFs binding promoter regions can give insights if the pathways are co-opted into stress-inducible regulation by the respective TFs (Jahan et al., 2019). The homology of TFs among different plant species can help metabolic engineering a wide variety of crop plants to produce phytoalexins in greater amounts.

In addition to enzyme-encoding genes, TF genes can also be found as gene clusters. For example, the gene cluster of TF *ERF* (jasmonate (JA)- responsive ethylene response factor) consists of five *ERF* genes in tomato (Cárdenas et al., 2016, Thagun et al., 2016), while eight in potato (Cárdenas et al., 2016), two clusters of ten and five in tobacco (Kajikawa et al., 2017), five in *C. roseus* (Singh et al., 2020), four in *Calotropis gigantea* (Singh et al., 2020), and four in *Glesemium sempervirens* (Singh et al., 2020). Besides, TFs involved in plant specialized metabolism can be found in arrays (Zhou et al., 2016, Shoji and Yuan, 2021). So, it is possible that the TFs we identified are located in the same genomic neighborhood as arrays or biosynthetic gene clusters (BGCs). The co-regulation hypothesis of gene clusters poses that clustering of TFs can help co-regulate genes in a pathway. Although co-regulation also exists between un-clustered metabolic

pathways, clustering may accelerate the recruitment of genes into a regulon (Smit and Lichman, 2022, Wisecaver et al., 2017).

Epistasis and plausible selection on glyceollin induction

Metabolic traits have been reported to have low heritability due to environmental effects on their accumulations (Rowe et al., 2008). Recent studies have shown strong epistatic interactions of genes influencing variation of plant specialized metabolites, which may impact fitness in the field (Brachi et al., 2015, Kerwin et al., 2015, Kerwin et al., 2017). For example, numerous epistatic interactions influence the highly complex genetic architecture responsible for Arabidopsis metabolism (Kliebenstein, 2001, Kliebenstein et al., 2001). Moreover, a mixture of positive and negative epistatic interactions can help identify significant QTLs located within a biosynthetic pathway (Rowe et al., 2008). Compared to expression regulations, the power of epistasis in metabolomics is that they can better indicate the interconnectedness of metabolites within the metabolic pathway (Fell and Wagner, 2000, Jeong et al., 2000, Arita, 2004). The widespread interactive effects found among our identified significant SNPs affecting targeted metabolic traits may be a consequence of the interconvertibility between daidzein and glyceollin.

Genes containing causal variation for plant defensive compounds may influence field fitness and thus are likely under natural selection (Kroymann, 2011). For example, Benderoth et al. (2006) detected positive selection in glucosinolate diversification in *Arabidopsis thaliana* and its relatives (Benderoth et al., 2006). Prasad et al. (2012) showed positive selection for a mutation on a metabolic pathway gene could enhance resistance to herbivory in natural populations of a rocky mountain cress species (Prasad et al., 2012). We detected strong signals of selection on the SNPs significantly associated

with glyceollin phenotypes with EHH and LD analyses (**Figure 2.5, Figure 2.3B,** and **Supplementary Figure 2.1**). For example, the LD surrounding the significant SNP ss715603454 that is next to the identified gene clusters is more extensive, suggesting strong selection in this region (**Figure 2.3B, Supplementary Figure 2.1**). Meanwhile, the two alleles of this significant SNP, G and T, showed different EHH values, with T exhibiting much longer haplotype homozygosity. This indicates that this T allele may be under recent positive selection. Interestingly, the T allele is significantly associated with higher induction of glyceollin and has a higher frequency in South Korea (**Figure 2.3C, D**). The allele specific EHH pattern and their geographic distribution may be due to heterogeneous selection pressure in nature.

Perspectives and future directions of our study

Plant specialized metabolites exhibit extreme quantitative and qualitative variation. Therefore, high-throughput metabolite profiling, such as UPLC-MS/MS analysis coupled with GWAS (as applied here) can help better understand the genetic contributions to metabolic diversity in natural populations. A common assumption is that biological variables or traits should show a normal distribution, and skewed data may indicate measurement error. However, the scenario is different in metabolomics, especially in secondary metabolism. For instance, a ratio of two related compounds, rather than their separate values, may provide a comprehensive understanding of the underlying enzymatic process (Kliebenstein, 2007a, Kliebenstein et al., 2001, McMullen et al., 1998, Byrne et al., 1996, Kliebenstein, 2001, Yencho et al., 1998, Chan et al., 2011, Prasad et al., 2012, Petersen et al., 2012). We used a ratio of glyceollin and daidzein concentrations as the phenotypic trait for our association study. The use of a metabolic ratio also may

produce: (1) a reduction in the variability of the data collected for the biological replicates and thus increase statistical power and (2) a reduction in overall noise in the dataset by canceling out systemic experimental errors. Most importantly for our purposes, the glyceollin to daidzein metabolite ratio is correlated to the corresponding reaction rate under optimal steady-state assumptions, as this metabolite pair is connected in the phenylpropanoid biosynthetic pathway (Petersen et al., 2012, Suhre et al., 2011).

The natural world has a lot to offer in tackling diseases and global food scarcity. There is a need to develop new medicines and future value-increased food by unlocking the uncharted gene pools of wild plants. Our chosen study system crop wild relative of soybean poses much higher and underexplored genetic diversity than its domesticated descendants. Given that glyceollin is produced in trace amounts, it is an exciting challenge to define the plant metabolic gene clusters and transcriptional regulators in the glyceollin biosynthesis pathway. Besides complex cancer treatment and therapies, the rise of different types of tumors and tumor subtypes urges the need for new drugs. Along with glyceollin's role in plant defense, it has been well-documented for anti-cancer activities. Our follow-up studies will apply transcriptomics and functional validation of the candidate genes, which can expand our focus to explore associations of genes in clusters to understand their involvement in regulating glyceollin biosynthesis at the systems level. As phytochemical variation can be caused by both structural genes and their expression differences, it will be interesting to explore the role of pathway-specific regulators (i.e., transcription factors) in glyceollin induction (Osbourn, 2010b). Our results suggest that improving our fundamental knowledge of plant specialized metabolic gene clusters and

regulators will facilitate metabolic engineering with improved metabolic traits for sustainable agriculture and novel pharmaceuticals.

CHAPTER 3: INVESTIGATING THE GENETIC BASIS OF PHYTOCHEMICAL DIVERSITY IN WILD SOYBEAN

***Yasmin, F.**, Zhang, H., Chen, H.-Y., Li, C., Piller, K., Reitzel, A.M., Li, X.[*], Song, B.-H.[*]*

Investigating the genetic basis of phytochemical diversity in wild soybean. (In preparation)

[*]Corresponding authors

Author Contributions

BHS and XL conceptualized and designed the project. HZ did experiments and collected tissues for metabolomics analysis. XL and HYC conducted LC-MS analysis and metabolites annotation. XL performed GWA analysis, FY performed the LD analysis, and the candidate gene identification with assistance from XL. FY conducted sequencing and further analyses. FY performed gene expression analysis with assistance from AMR. FY tried gene cloning with assistance from KP and AMR. FY tried virus-induced gene silencing experiments under the guidance of CL. The chapter was drafted and edited by FY and reviewed and edited by BHS and XL. The manuscript in preparation from this chapter is being drafted by FY and will be reviewed and edited by BHS and XL.

Introduction

Plants possess an extraordinary ability to synthesize an extensive array of specialized metabolites, commonly known as secondary metabolites, natural products, or phytochemicals, which exhibit diverse sizes, shapes, and levels of complexity. These metabolites serve various crucial functions in plant survival, ecosystem dynamics, human

nutrition, and the development of novel pharmaceuticals (Medema et al., 2021, Weng et al., 2021, Fang et al., 2019, Wurtzel and Kutchan, 2016). Despite their importance, our understanding of the molecular basis underlying the immense diversity of plant specialized metabolic pathways remains limited (Caspi et al., 2014, Zhao and Rhee, 2022). Currently, only a small fraction (<50) of these pathways have been characterized in terms of their specific genes (Nützmann et al., 2016, Schläpfer et al., 2017).

Phytochemical diversity, both in terms of quantity and quality, exhibits intrinsic variation within and among plant individuals, offering benefits to the plants themselves (Wetzel and Whitehead, 2020, Moore et al., 2014). Genetic variations, such as mutations and gene duplications, are responsible for both qualitative diversity and quantitative variation in plant specialized metabolites (Moore et al., 2014, Zhou and Liu, 2022). Even small modifications to the active site of these enzymes can influence the relative amounts of each product produced (Kollner et al., 2004). Given the significance of phytochemical diversity, it is essential to unravel the genetic foundations that give rise to this remarkable variability within species and environments. However, genetic variation is not the sole driver of phytochemical diversity. Certain enzymes involved in biosynthesis have many functions and catalyze many reactions, like terpene synthases, have the ability to produce different products using the same starting material (Aharoni et al., 2005, Zulak and Bohlmann, 2010). Furthermore, enzymes belonging to extensive gene families exhibit low substrate specificity, leading to functional divergence among these enzymes. This divergence contributes to the generation of metabolic diversity, as seen in the case of carboxyl methyltransferases and acetyltransferases (Negre et al., 2003, Pichersky et al., 2006).

Phytochemical diversity can also arise from plant adaptation to abiotic and biotic stresses (Weng, 2014, Endara et al., 2017, Salazar et al., 2018). Differences in abiotic and biotic conditions among various natural environments have the potential to result in distinct chemical profiles within populations of a given species (Forrister et al., 2023, Thompson, 2019). The existence of phytochemical diversity is considered an adaptive trait (Richards et al., 2015, Salazar et al., 2018, Endara et al., 2022a). Forrister et al. (2023) hypothesized that selection may play a crucial role in generating phytochemical diversity (Forrister et al., 2023). Rather than focusing solely on structurally related compounds, selective pressures arising from abiotic and biotic factors tend to favor structurally unrelated diverse phytochemicals that provide functional advantages (Weng, 2014, Endara et al., 2017, Salazar et al., 2018). Moreover, recent studies have highlighted the importance of preserving a distinct chemical profile for a particular species in comparison to other species within its ecological community. This aspect holds importance alongside the species' ability to produce a wide variety of compounds. For example, certain plant lineages may exhibit specificity towards particular classes of phytochemicals, indicating co-evolutionary dynamics with biotic stressors (Kursar et al., 2009, Forrister et al., 2019, Endara et al., 2022b). Additionally, the growth and development of plants play a significant role in shaping phytochemical diversity (Moore et al., 2014, Weng et al., 2012).

Investigating the genetic, environmental, and ecological drivers of this phytochemical diversity can provide valuable insights into the intricate mechanisms that shape plant chemical defenses and their ecological consequences. This knowledge has practical implications for crop improvement, ecological interactions, and human health. While

numerous compounds identified in this research may contribute to plant defense and human health, deciphering the precise functions of these compounds poses significant challenges in metabolomics studies (Tsugawa et al., 2021). Therefore, in this study, we focus on delineating the genetic basis of the comprehensive chemical profile, which encompasses a range of compounds that are likely selected for various functional purposes.

Soyasaponins are a diverse group of specialized metabolites categorized under the triterpenoid class and found abundantly in legume species, particularly in soybeans. Soybeans are a vital staple crop known for their provision of plant-based protein, oil, and micronutrients. Soyasaponins exhibit well-documented evidence of their effectiveness against both biotic and abiotic stressors, as well as their influence on plant growth and development (Moses et al., 2014, de Costa et al., 2013, Augustin et al., 2011, Morrissey and Osbourn, 1999, Osbourn et al., 2011, Sparg et al., 2004, Tsuno et al., 2018, Berendsen et al., 2012, Hacquard et al., 2015, Andreote and e Silva, 2017, Jacoby et al., 2017). Moreover, saponins have a long history of use as key components in traditional medicines for their cardioprotective effects (Waller and Yamasaki, 2013, Wang et al., 2021). Recent studies have highlighted their antiviral, anti-inflammatory, anticancer, antioxidant, and immunomodulatory properties (Moses et al., 2014, Wu and Kang, 2011, Kang et al., 2010, Zha et al., 2011, Lee et al., 2010, Ahn et al., 2002, Oh et al., 2000, Sun et al., 2014).

The presence and arrangement of sugar moieties in triterpenoid saponins, including soyasaponins, have been found to influence their bioactivities and chemical properties (Bowles et al., 2005). The biosynthesis of soyasaponins involves key enzymes belonging

to multi-gene families, namely oxidosqualene cyclases (OSCs), cytochrome P450-dependent monooxygenases (P450s), and UDP-dependent glycosyltransferases (UGTs) (Augustin et al., 2011). Among these enzymes, UGTs play a crucial role in the final step of soyasaponin biosynthesis, contributing to their structural and chemical diversity. UGTs, which are members of family 1 uridine diphosphate glycosyltransferases, are responsible for conferring biological activity to saponins through glycosylation, likely involving the sequential activity of various enzymes within this enzyme family (Augustin et al., 2011). Specifically, UGTs catalyze the transfer of sugar molecules from UDP-sugars to the soyasapogenol moiety (aglycone), making them an ideal target for investigating enzyme efficiency and sugar donor specificity (Shibuya et al., 2010, Tantry and Khan, 2013).

Considerable research efforts have been devoted to studying the biosynthetic pathway, structural diversity, composition, and distribution of soyasaponins in soybeans. So far, researchers have identified five UGTs responsible for the diversification of soybean saponin sugar moieties (Sayama et al., 2012, Shibuya et al., 2006, Shibuya et al., 2010, Yano et al., 2017). These include *Sg-1* (*UGT73F2*, *Glyma.07G254600*), *Sg-3* (*UGT91H9*, *Glyma.10G104700*), *Sg-4* (*UGT73P10*, *Glyma.01G046300*), *GmSGT2* (*UGT73P2*, *Glyma.11G053400*), and *GmSGT3* (*UGT91H4*, *Glyma.08G181000*). However, many aspects related to UGTs governing DDMP saponin biosynthesis remain relatively unexplored at the molecular level (**Figure 3.1**) (Sundaramoorthy et al., 2019). Therefore, UGTs represent essential targets in our attempt to comprehend the genetic basis that regulates the variation in soyasaponin production.

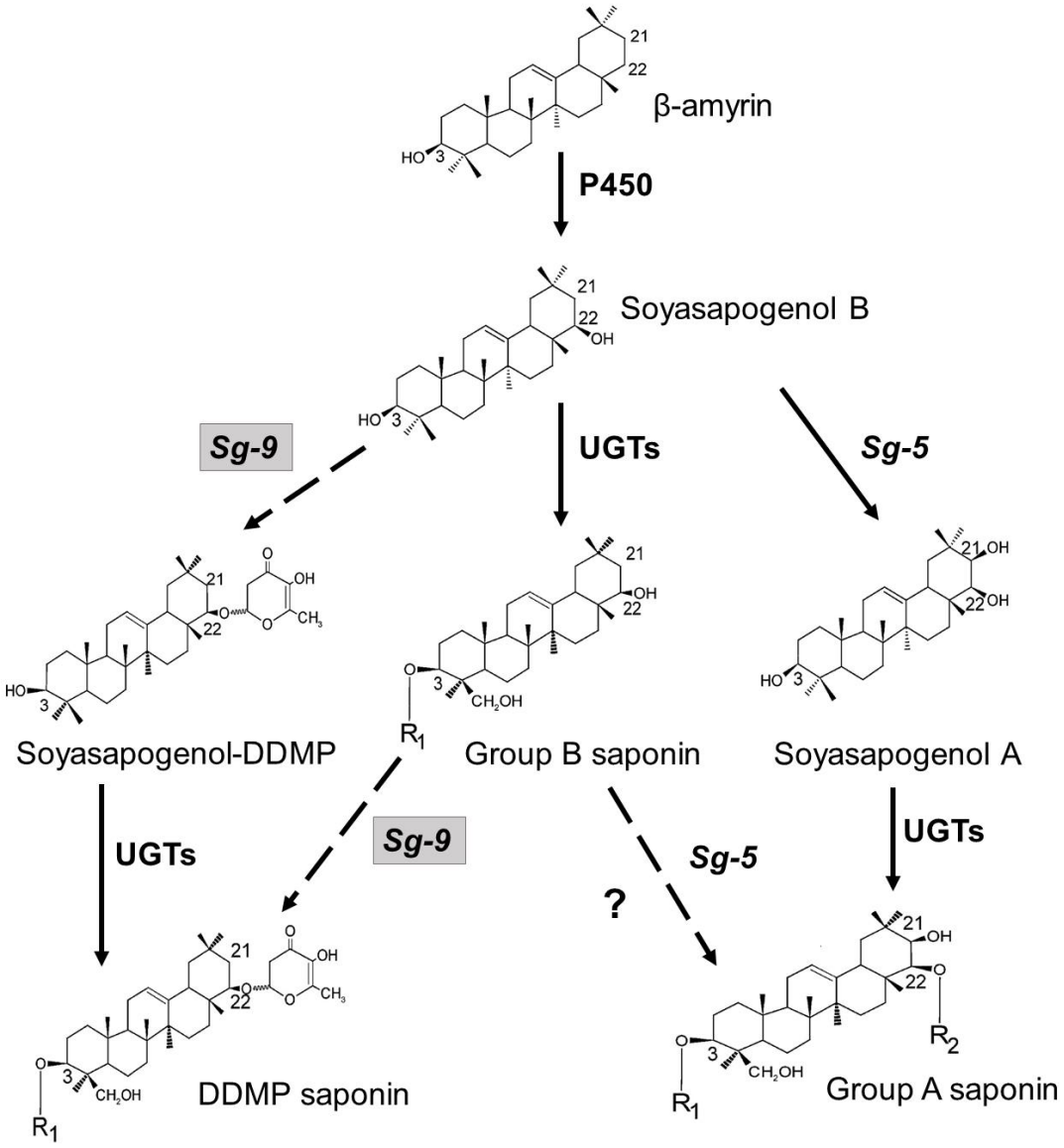


Figure 3.1 A potential biosynthetic pathway of DDMP-conjugated soyasapogenol/saponins in soybean *(Glycine max)*, adapted from Sudaramoorthy et al. (2019) (Sundaramoorthy et al., 2019). The biosynthesis of soyasaponins involves several enzyme families, with the final step being carried out by UDP Glycosyl transferases

(UGTs) that play a crucial role in creating structural and chemical diversity. However, the molecular details of many UGT-dependent steps in soyasaponin biosynthesis remain largely unexplored (dashed arrows indicate knowledge gap). DDMP: 2,3-dihydro-2,5-dihydroxy-6-methyl-4H-pyran-4-one.

In this study, we utilized mGWAS to elucidate the genetic underpinnings of phytochemical diversity. The combination of GWAS with metabolomics (mGWAS) offers a powerful approach in deciphering the genetic contributions to the vast amount of metabolic diversity and hence the complex traits behind those (Luo, 2015, Riedelsheimer et al., 2012, Chan et al., 2011, Chan et al., 2010, Chen et al., 2021). In recent years, metabolite-based genome-wide association studies (mGWAS) have been instrumental in unraveling the genetic mechanisms underlying plant specialized metabolism in major crops such as rice (Dong et al., 2015, Peng et al., 2016, Peng et al., 2017, Chen et al., 2014a, Matsuda et al., 2015, Fang et al., 2016) and maize (Wen et al., 2014). Recent studies on rice using mGWAS showed that metabolic pathways renovation is possible by extracting information from genotype-metabolite associations. Chen et al. (2014) identified candidate genes encoding *O*-methyltransferase, transferase, *UGT*, and another transferase associated with trigonelline, *N*-feruloylagmatine, apigenin 5-*O*-glucoside, and *N*-feruloylputrescine metabolites, respectively, that possess plausible physiological and nutritional importance for rice (Chen et al., 2014b). Moreover, Dong et al. (2014) proposed six possible candidate genes associated with biosynthesis and regulation of flavonoid pathway in rice (Dong et al., 2014). The majority of mGWAS studies focused on model or crop species. However, thus far, little is known about the genetic basis of phytochemical diversity in wild plant species, such as wild soybean.

To comprehensively understand the intraspecific diversity of plant specialized metabolites, it is crucial to leverage crop wild relatives (CWRs) and explore the genetic variations that contribute to the richness and vast diversity of these compounds. In line with our aim to decipher the genetic foundations of phytochemical diversity in crop wild relatives, we have chosen wild soybean as a unique study system that may provide valuable insights into the complexity and variability of these traits. Several factors make wild soybean an ideal choice: (1) Wild soybean exhibits significantly higher genetic diversity compared to domesticated soybeans (Li et al., 2013). Our own data further support this observation, revealing a wide range of diversity in phytochemical production among wild soybean populations. (2) Wild soybean serves as the wild ancestor of cultivated soybean, and there are no breeding barriers between the two. This close relationship allows for direct transfer of essential genes identified in wild soybean to cultivated varieties (Kofsky et al., 2018). These genes can be utilized in metabolic engineering approaches or incorporated into breeding programs to develop nutritionally enriched soybean cultivars with enhanced phytochemical content.

Our study aimed to establish a foundation for future research by tapping into the vast and relatively unexplored genetic diversity of wild soybeans, which gives rise to a wide range of phytochemicals. We sought to address several key questions throughout our research: How does the genetic diversity present in wild soybeans influence the diversity of phytochemicals? Our findings have prompted additional inquiries that can be explored in future studies, such as: Are there any wild soybean genes that have been lost in modern cultivars as a result of domestication? If so, do these genes play a role in the accumulation of essential phytochemicals? Do the geographic origins of wild relatives of

soybean contribute to genetic variability, ultimately shaping the diversity of phytochemicals for local adaptation in the face of changing environments? This study investigated the genetic basis of qualitative phytochemical diversity among different populations of wild soybeans. Additionally, we sought to uncover the quantitative variations of specific phytochemicals, particularly soyasaponin. The insights gained from our research will significantly contribute to the advancement of metabolic engineering, enabling the development of crops fortified with phytochemicals with enhanced health-promoting properties and capable of withstanding future agricultural challenges.

Materials and methods

Plant materials

The seeds were obtained from the USDA National Germplasm Resources Laboratory through a request made via the Germplasm Resources Information Network (GRIN). To ensure a broad representation of genetic diversity, our sample set consisted of 195 samples. This included 190 accessions of *Glycine soja* originating from different regions within the original geographic distribution, namely China, Japan, Russia, and South Korea, covering a range of maturity groups. Additionally, to facilitate a preliminary comparison of metabolite profiles between species, five soybean cultivar samples (*Glycine max)* were included (**Supplementary Table 3.1**).

Plant growth condition and sample collection

Seed preparation, germination, and transplanting procedures were conducted following established protocols (Zhang and Song, 2017, Zhang et al., 2017c). In brief, soybean seeds were subjected to surface sterilization by treating them with 0.5% sodium

hypochlorite for 60 seconds, followed by rinsing with autoclaved water. To enhance germination rates, the seed coat was carefully sliced. The prepared seeds were then placed on a moist sterile filter paper in a petri dish for germination. After 2-3 days, seedlings with healthy roots were individually transplanted into cones (Greenhouse Megastore, Danville, IL, USA) containing sterile sand as the growing medium.

Germination and subsequent growth were carried out in an environmentally controlled chamber set at a temperature of 27 °C with a 16-hour photoperiod (16 hours of light and 8 hours of darkness). For sample collection, trifoliate leaf tissues were carefully harvested and weighed. To preserve the samples' integrity, all collected samples were immediately flash-frozen in liquid nitrogen and stored at -80 °C until further analysis. Each ecotype was represented by four biological replicates to ensure robustness and reliability of the data.

Metabolite profiling

LC-MS (liquid chromatography-mass spectrometry)-based untargeted metabolite profiling was conducted on a total of 606 samples, consisting of 195 soybean accessions, including 190 *G. soja* (wild soybean) and 5 *G. max* (cultivated soybean) samples. Leaf tissue was extracted using a tissue-to-solvent ratio (w/v) of 1:10 with 50% (v/v) methanol at 60 °C for 30 minutes. Prior to analysis on a G6530A Q-TOF LC/MS system (Agilent Technologies, USA), the extract was filtered through a 0.2-µm filter.

Metabolite separation was achieved using an Agilent Eclipse Plus C18 column (3 × 100 mm; 1.8 µm) and a binary gradient of solvent A (0.1% formic acid in water) and solvent B (0.1% formic acid in acetonitrile) at a flow rate of 0.6 mL min−1. Mass spectra acquisition was performed in negative ion mode with the following parameters: drying

gas temperature, 300 °C; drying gas flow rate, 7.0 L min−1; nebulizer pressure, 40 psi; sheath gas temperature, 350 °C; sheath gas flow rate, 10.0 L min−1; Vcap, 3500 V; Nozzle Voltage, 500 V; Fragmentor, 150 V; Skimmer, 65 V; Octopole RF Peak, 750 V. The obtained raw data were processed using Agilent Masshunter Profinder software to generate a peak matrix across all samples. A total of 874 peaks detected in at least one biological replicate of any accession were retained for subsequent analysis. To annotate the peaks, an accurate mass search was performed against multiple databases, including SoyCyc (https://www.plantcyc.org/) and KNApSAcK (http://kanaya.naist.jp/KNApSAcK/). The identification of metabolites was further confirmed by comparison with chemical standards for daidzein, formononetin, genistein, glycitein, daidzin, genistin, and glyceollin.

Metabolite-based genome-wide association study and LD estimation

Genotype data for a diverse collection of 190 wild soybean ecotypes and 5 soybean cultivars were obtained from SoySNP50K dataset, which provided a dataset of 41,896 genome-wide single nucleotide polymorphic markers (SNPs) (Song et al., 2013). To account for missing genotypes, imputation was performed using Beagle 4.0. Following data pre-processing, SNPs with a minor allele frequency of less than 5% were filtered out, resulting in a final genotype dataset of 34,819 SNPs. For the GWAS analysis, to prepare the phenotype data, the peak area values of 874 metabolites were logarithmically (log2) transformed. For the association analysis, three different methods were employed: simple linear regression (LR), linear mixed model (LMM) with a Kinship matrix (K), and LMM with both K and principal component analysis (PCA) (P). To account for multiple testing, the raw P-values were adjusted using the simpleM method, which was more

effective than the overly conservative Bonferroni correction (Gao et al., 2008). Using a corrected P-value threshold of 0.05, statistically significant associations for 727, 426, and 440 metabolite peaks were identified using LR, LMM with K, and LMM with K + P methods, respectively. Consistent with expectations, both LMM methods demonstrated better control over population structure and reduced spurious results compared to LR (**Supplementary Figure 3.2**). The Kinship method appeared to be the most conservative in our dataset, yielding the fewest significant results, with all 426 metabolite traits showing significance in at least one of the other two methods (**Supplementary Figure 3.2**). Further analysis focused on the 359 metabolite traits that were significant across all three methods, with the LMM with K results selected for subsequent investigation (**Supplementary Figure 3.2**).

To examine the linkage disequilibrium (LD) patterns across the panel, the TASSEL program (version 5) was utilized to calculate LD (Bradbury et al., 2007). Our objective was to identify potential candidate genes for further analysis. The LD analysis included all the significant SNPs identified from the GWAS analysis. Our linkage disequilibrium analyses encompassed both patterns: multiple SNPs-multiple metabolites, referred to as QTL-multiple metabolites, and single SNP-multiple metabolites. LD was measured using squared correlations (r2) with a threshold of 0.2 (displayed in the upper right of the LD plot) and a significant threshold of P-value < 0.05 (displayed in the lower left of the LD plot). Pairwise LD was generated following the approach described by Shin et al. (2006) (Shin et al., 2006). In order to identify potential candidate genes for further analysis, genes within LD blocks containing the significant SNPs were determined. These genes

represent promising candidates that may contribute to the observed associations in the GWAS analysis.

Hierarchical clustering

Hierarchical clustering analysis was performed on the metabolite profiles within the QTL1 cluster using a total of 195 samples, which comprised 190 accessions of *G. soja* and 5 accessions of *G. max*. The analysis was conducted using JMP® software, Version 15 (SAS Institute Inc., Cary, NC, 1989–2021).

Gene enrichment analysis

To conduct an extensive gene mining of our identified candidate genes, we performed a comprehensive database search using multiple resources. These included Phytozome v13, ShinyGO v0.66: Gene Ontology Enrichment Analysis, SoyBase GO Enrichment Data, SoyCyc 10.0 Soybean Metabolic Pathway, and the metabolic cluster viewer PMN (Plant Metabolics Network) (Luo et al., 2017, Goodstein et al., 2011, Grant et al., 2010).

To identify potential candidate genes, our pairwise linkage disequilibrium (LD) analysis involved examining a 50-kb window surrounding each significant SNP. Genes located within each LD block were considered as candidate genes, and their annotations were obtained from the Phytozome database, resulting in a total of 612 candidate genes for QTL-multiple metabolite clusters.

To further explore the functional implications of these candidate genes, GO enrichment analysis was performed using ShinyGO v0.66: Gene Ontology Enrichment Analysis, with a significance threshold set at a P-value cut-off (FDR) of 0.05. Additionally, we utilized the SoyBase GO Enrichment Data to gain insights into the gene ontology annotations.

To investigate the involvement of these candidate genes in metabolic pathways, we thoroughly searched through annotation files from Phytozome v13, SoyBase, SoyCyc 10.0 Soybean Metabolic Pathway, and Pathview databases. This comprehensive analysis allowed us to explore the potential roles of the candidate genes in various metabolic pathways.

Finally, we utilized the PMN plant metabolic cluster viewer to further investigate the enzyme class and metabolic domains associated with the identified candidate genes. This analysis provided valuable insights into the functional aspects of the candidate genes within the context of plant metabolism.

DNA sequencing and statistical analysis for association analysis

DNA was extracted from 46 *G. soja* and 34 *G. max* ecotypes using the CTAB method as described by Doyle and Doyle (1990). Gene-specific primers were designed using Primer3plus (see **Supplementary Table 3.5**). Subsequently, PCR (polymerase chain reaction) was conducted using the gene-specific primers, and the PCR products were analyzed by gel electrophoresis. The purified PCR products were subjected to Sanger sequencing using the Sanger sequencing method (Sanger et al., 1977), and the sequencing results were analyzed using SnapGene® software (from Dotmatics; available at snapgene.com).

To investigate the association between high and low-soyasaponin producing ecotypes and haplotypes, we performed a one-way analysis using haplotypes as the factor of interest. An unpaired t-test with Welch's correction was performed using GraphPad Prism version 9.0.0 for Windows, GraphPad Software, San Diego, California USA, www.graphpad.com.

In addition, a subset of 12 ecotypes from the previously mentioned 46 ecotypes was selected for promoter region sequence analysis. For this purpose, gene-specific primers targeting a 3.5 kb region upstream of the gene were designed (see **Supplementary Table 3.6**). The correctness of the PCR products was confirmed by agarose gel electrophoresis. Subsequently, the purified PCR products were sent for Sanger sequencing, and the sequencing results were analyzed using SnapGene® software (from Dotmatics; available at snapgene.com).

Gene expression analysis

Gene expression patterns in wild soybean ecotypes with high and low soyasaponin II production were investigated using quantitative real-time PCR (qPCR). To select the ecotypes, preliminary metabolite data were used to identify those with high and low abundance of soyasaponin. RNA was extracted from the young leaves of these ecotypes using the RNeasy Plant Mini Kit (QIAGEN). Subsequently, cDNA synthesis was performed through reverse transcriptase-polymerase chain reaction (RT-PCR) using the Thermo Scientific RevertAid RT Kit and primers specifically designed for the entire target gene (refer to **Supplementary Table 3.5**). Quantitative real-time PCR (qPCR) was conducted to target the key candidate gene that may involve in soyasaponin synthesis, utilizing gene-specific primers designed through PrimerQuest™ program, IDT, Coralville, Iowa, USA. Accessed 12 December, 2018. https://www.idtdna.com/SciTools (**Supplementary Table 3.7**) (Green et al., 2012). Each qPCR analysis consisted of three biological replicates, and each reaction was repeated twice. The soybean ubiquitin gene *(GmUBI)* was employed as an internal control for normalization purposes.

Designing plasmid constructs for functional analysis of candidate gene

For virus-induced gene silencing, the infectious plasmid DNA used in this study was kindly provided by Dr. Steve Whitham's lab at Iowa State University in Ames, IA, USA. It is engineered to carry a modified viral genome controlled by the 35S promoter, effectively creating a recombinant virus. 300 bp from exon-1 of our target *UGT* gene was cloned to the constructs pBPMV-IA-V1 (sense orientation) and IA-1033_pBPMV-IA-V2 (both sense and antisense orientations).

The CRISPR/Cas9 expression vector was designed following the protocol outlined by Han et al. (2019) (Han et al., 2019). The vector will then be introduced into *Agrobacterium* strain EHA105 using the electroporation technique. For our genetic transformation experiments, the soybean cultivar Williams 82 will be utilized. To achieve genome editing using the CRISPR/Cas9 system, soybean whole-plant transformation will be employed using half-seed explants, as described in the protocol adapted from Paz et al. (2005) (Paz et al., 2006) by Jean-Michel Michno, with any necessary modifications or optimizations.

Results

Exploring the variability of phytochemicals in distinct wild soybean varieties

Through a comprehensive analysis of metabolic traits utilizing untargeted LC-MS, we measured and annotated a total of 874 diverse metabolic traits, resulting in the identification of 485 metabolites through database matches (**Supplementary Table 3.2**). Based on shared significant SNPs, we grouped these metabolites into 74 clusters (**Figure 3.2**). We selected one particular cluster, QTL1, for a comparative analysis between wild and cultivated soybean. Hierarchical clustering of metabolite profiles within the QTL1

cluster revealed no significant differences between wild and cultivated soybean ecotypes. However, when considering the geographic origin, clear patterns in metabolite profiles corresponding to different geographic regions became evident. Particularly, the South Korean population displayed a distinct pattern characterized by higher accumulation of all the metabolites clustered into QTL1 compared to the other three regions (**Supplementary Figure 3.1**).

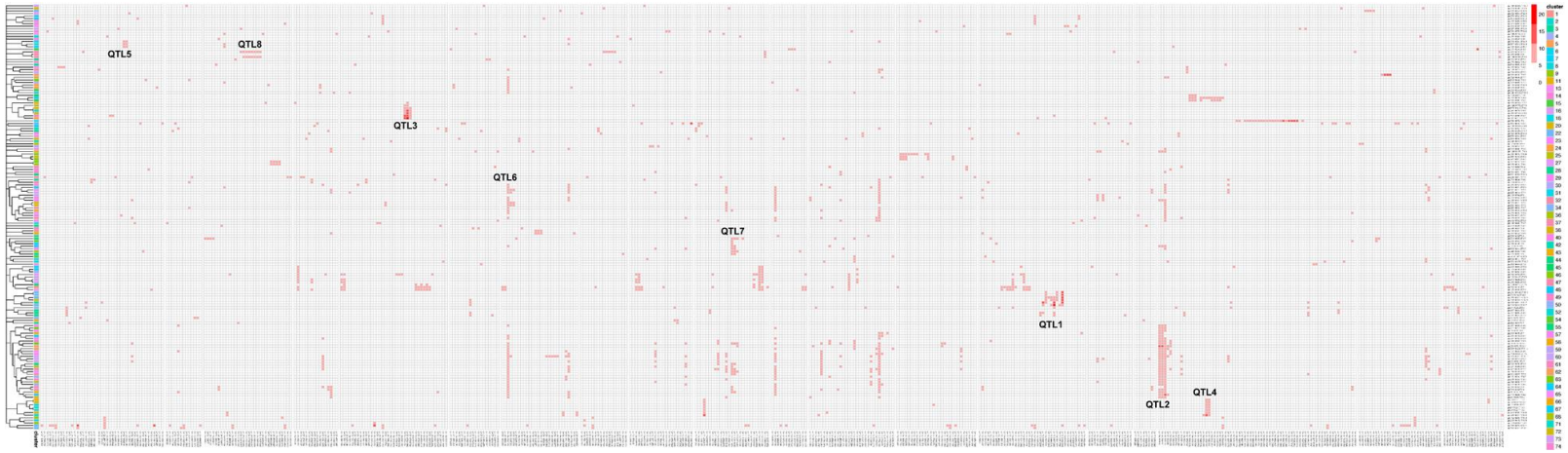


Figure 3.2 The QTL-multiple metabolite clusters, comprising eight distinct QTLs under the multiple-SNP and multiple-metabolites pattern, are of significant interest in our study. These clusters demonstrate the presence of multiple metabolites associated with multiple SNPs within specific genomic regions.

Identification of unique quantitative trait loci (QTLs) associated with varied metabolic pathways

Through genomic analysis, we identified 1155 SNPs that showed significant associations with 359 metabolites, representing a diverse range of phytochemicals. Given the complexity of the genomic regions involved, we employed clustering analysis to focus on specific groups of metabolites. Our clustering analysis revealed two distinct patterns: QTL-multiple metabolite clusters, which involve multiple SNPs and multiple metabolites, and single-SNP and multiple metabolite patterns (**Figure 3.2**). To delve deeper into our investigation, we narrowed our focus to the QTL-multiple metabolite clusters and identified eight QTLs. These QTLs represent multiple metabolites significantly associated with several SNPs within specific genomic regions. The complete list of these eight QTLs, provided in **Supplementary Table 3.3** serves as a valuable resource for functional validation of candidate genes in subsequent studies. **Table 3.1** provides comprehensive information on the QTL1-multiple metabolite clusters, for example grp 70, grp 553, and grp 766. It includes the associated SNPs along with their corresponding P-values and R-squared values.

Table 3.1 Detailed information on the QTLs and their associated SNPs for the QTL1-multiple metabolite cluster.

QTL-multiple metabolite cluster	SNP	P-value	R-squared value
QTL1.grp70.Gm15-38276224	Gm15-38276224	7.63E-08	0.14
QTL1.grp70.Gm15-38575222	Gm15-38575222	2.20E-06	0.11
QTL1.grp70.Gm15-39570425	Gm15-39570425	1.17E-07	0.14
QTL1.grp70.Gm15-39945568	Gm15-39945568	2.72E-15	0.28
QTL1.grp553.Gm15-34552298	Gm15-34552298	5.89E-07	0.12
QTL1.grp553.Gm15-38241517	Gm15-38241517	2.18E-06	0.11
QTL1.grp553.Gm15-38575222	Gm15-38575222	5.96E-07	0.12
QTL1.grp766.Gm15-37622758	Gm15-37622758	7.03E-07	0.12
QTL1.grp766.Gm15-38029609	Gm15-38029609	4.86E-07	0.12
QTL1.grp766.Gm15-38276224	Gm15-38276224	3.50E-10	0.19
QTL1.grp766.Gm15-38380213	Gm15-38380213	1.45E-07	0.14
QTL1.grp766.Gm15-38447399	Gm15-38447399	3.86E-09	0.17
QTL1.grp766.Gm15-38575222	Gm15-38575222	8.07E-08	0.14
QTL1.grp766.Gm15-39570425	Gm15-39570425	3.90E-10	0.19
QTL1.grp766.Gm15-39945568	Gm15-39945568	4.30E-14	0.26

To visually represent our findings, we generated Manhattan plots for QTL1 (**Figure 3.3** and **Supplementary Figure 3.3**), which corresponds to chromosome 15 and includes

triterpenoid derivative soyasaponin, as well as other known and unknown metabolites within the cluster. The QTL1 cluster comprises a total of 31 metabolites, including grp41, grp70, grp87, grp185, grp187, grp311, grp321, grp326, grp371, grp447, grp450, grp492, grp553, grp557, grp573, grp574, grp630, grp641, grp652, grp668, grp718, grp738, grp766, grp799, grp921, grp943, grp960, grp982, grp1005, grp1025, and grp1033. The metabolites within the QTL1 cluster exhibit variations in mass (m/z value) and retention time (RT) (as shown in **Supplementary Table 3.2**). We utilized MS/MS spectra to annotate the exact mass number, retention time and fragmentation pattern, comparing them with the standards mentioned in the methods.

Among these metabolites, we successfully annotated grp 70 and grp 553 (**Table 3.2**). Within the QTL1 cluster, we were able to specifically annotate Grp 70 as Soyasaponin II. The remaining 30 metabolites within this cluster share the same set of SNPs as Grp 70. Furthermore, to enhance our annotation process, we utilized functionally related genes for the putative metabolites that share the same set of SNPs within the QTLs. For instance, the SNPs (ss715621590, ss715621696, ss715621697, ss715621701, ss715621702, ss715581532, ss715621790, ss715621794, and ss715621800) within QTL1 that are located in the chromosome 15 region are associated with a group of metabolites (**Supplementary Figure 3.3**, **Supplementary Table 3.3**). For a comprehensive list of all compound annotations, please refer to **Supplementary Table 3.2**. Both known and unknown metabolites within QTL1 (**Table 3.1 and 3.2**) shared significant SNPs, namely ss715621590, ss715621696, ss715621697, ss715621701, ss715621702, ss715581532, ss715621790, ss715621794, and ss715621800. An example of these significant SNPs on chromosome 15 for soyasaponin, belonging to QTL1, is depicted in **Figure 3.3**. The plot

visually highlights genomic regions potentially influencing soyasaponin variation, aiding in the identification of genetic factors impacting this trait.

Table 3.2 Annotation of the identified peaks in QTL1 including their mass, retention time (RT) and formula.

ID	Mass	RT	Formula	Cpd
70	912.508	12.503	C47H76O17	Soyasaponin II
553	434.2488	7.472	C21H39O7P	1-oleoyl-2-lyso-glycerone phosphate

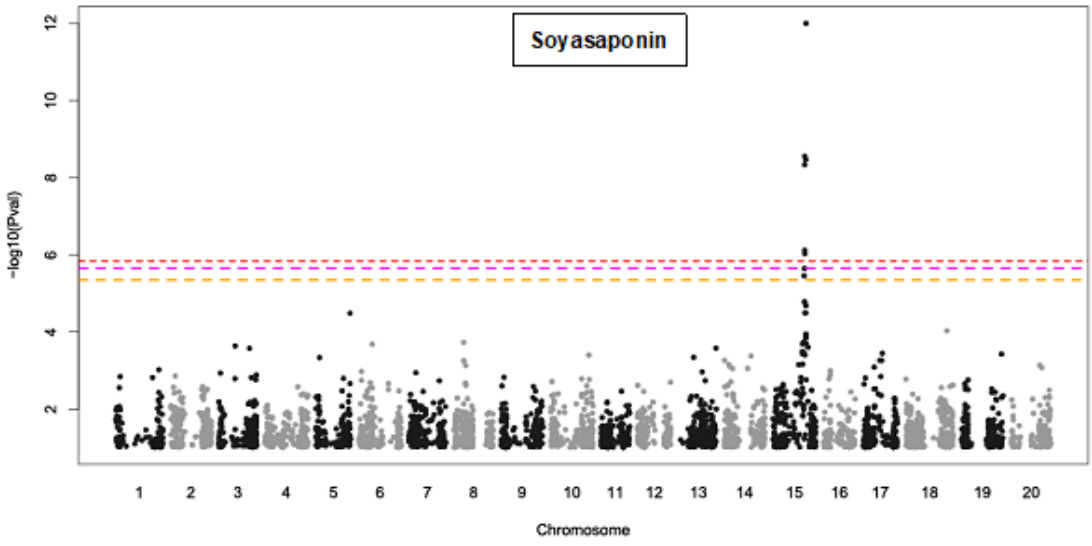


Figure 3.3 A Manhattan plot illustrating significant SNPs associated with soyasaponin variation on chromosome 15. Each point represents a SNP, with the y-axis indicating the statistical significance (-log10 P-value) and the x-axis representing the genomic position. Significant SNPs are indicated above the threshold of significance.

Candidate genes within various quantitative trait loci (QTLs)

Through our gene annotation and enrichment analyses, we successfully identified 16 candidate genes associated with nine significant SNPs (**Table 3.3**). **Figure 3.4** depicts a concise flow chart illustrating our approach for conducting linkage disequilibrium (LD) analysis to identify candidate genes. The aforementioned nine SNPs were found within the LD blocks of the previously mentioned eight QTLs (**Figure 3.5** and **Table 3.3**). In addition, we identified clusters of single SNPs associated with multiple metabolites, as well as candidate genes related to these clusters (**Supplementary Figure 3.4**). However, the primary focus of our study was on the QTL-multiple metabolite clusters and the candidate genes derived from it, with a specific emphasis on the soyasaponin biosynthesis pathway candidate gene.

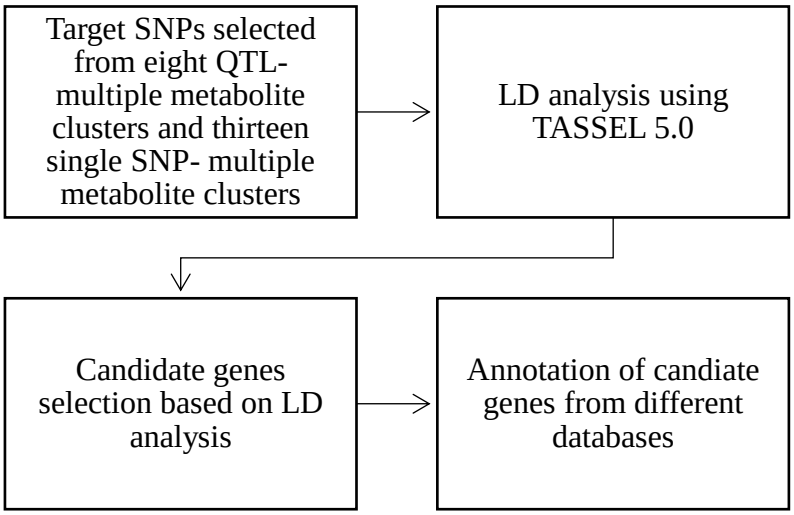
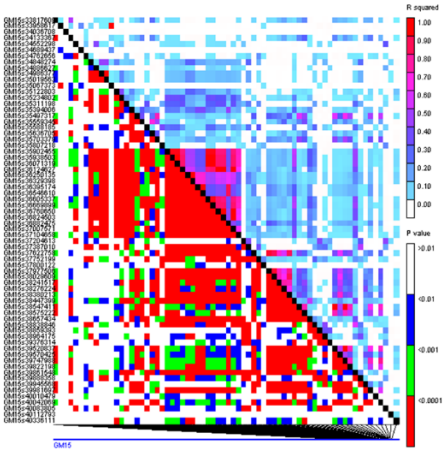
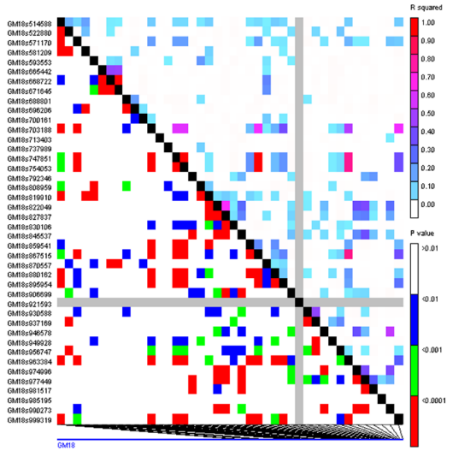


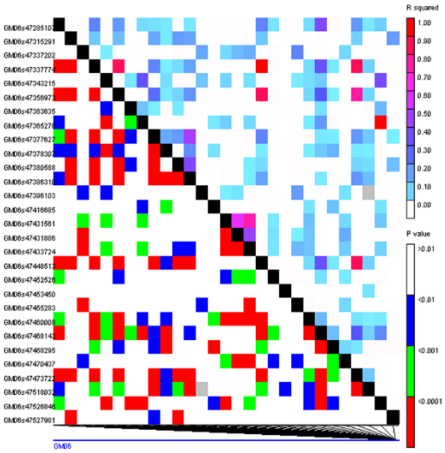
Figure 3.4 A flowchart illustrating the process of candidate gene selection through linkage disequilibrium (LD) analysis.



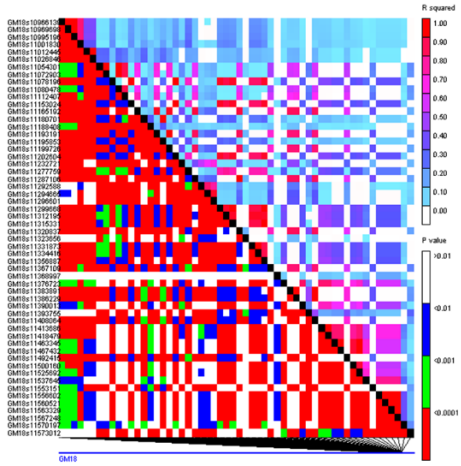
QTL1



QTL2



QTL3



QTL4

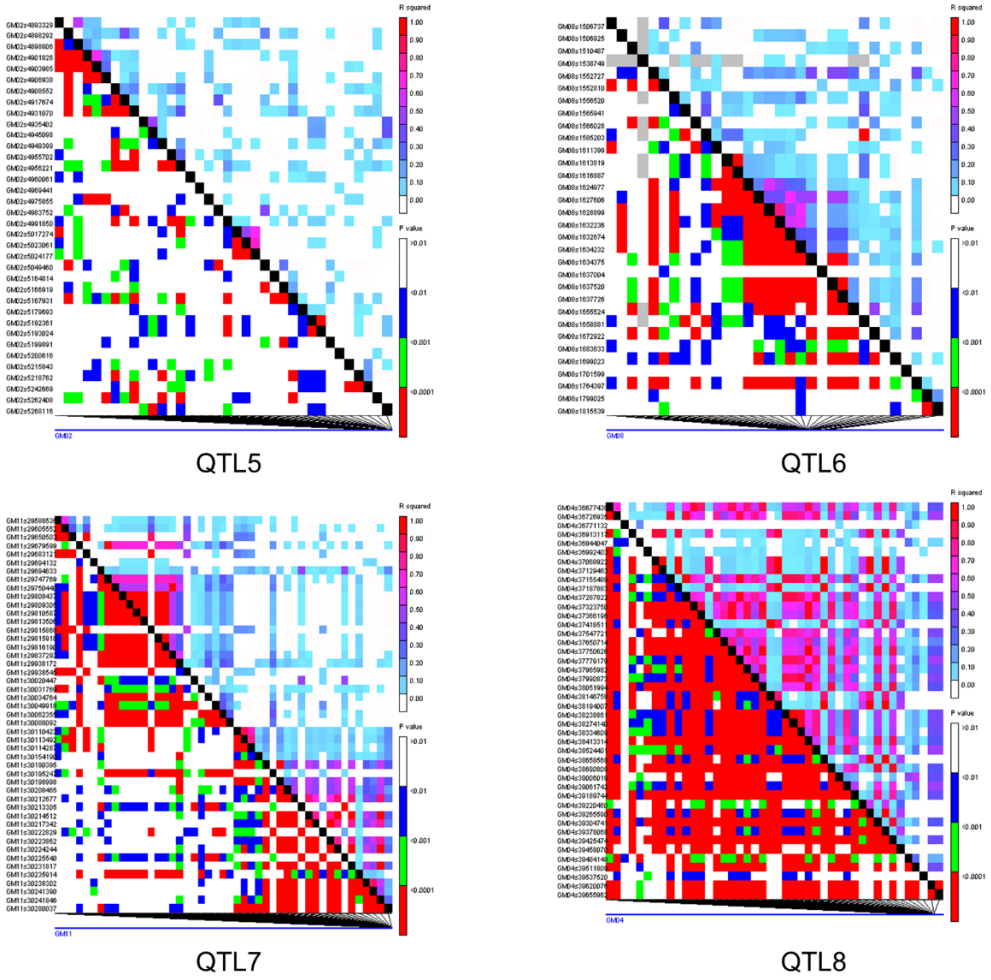


Figure 3.5 Linkage disequilibrium (LD) plots highlight significance of SNPs for eight QTL-multiple metabolite clusters.

Within QTL1, a cluster of SNPs on chromosome 15 exhibited strong linkage disequilibrium (LD), indicating close genetic linkage and limited recombination in this region. Our LD analysis revealed a 5393.3 kb window containing the significant SNP ss715621800, spanning from bp 34552298 to 39945568 on chromosome 15, indicating a high level of linkage within this region. The levels of soyasaponin were significantly associated (P-value of 1.15×10^{-13}) with the SNP ss715621800 located on chromosome 15. Interestingly, the candidate gene *Glyma.15G221300* was found to be in closer proximity to the SNP ss715621800 compared to the other SNPs within QTL1 (as shown in **Figure 3.6**). The marker is located 6.3 kb upstream of the candidate gene and exhibits linkage disequilibrium (LD) (r2 = 0.26) with it. We employed a cutoff of r2>0.2 for our LD analysis, where r2 represents the extent of allelic association between a pair of sites (Weir, 1990).

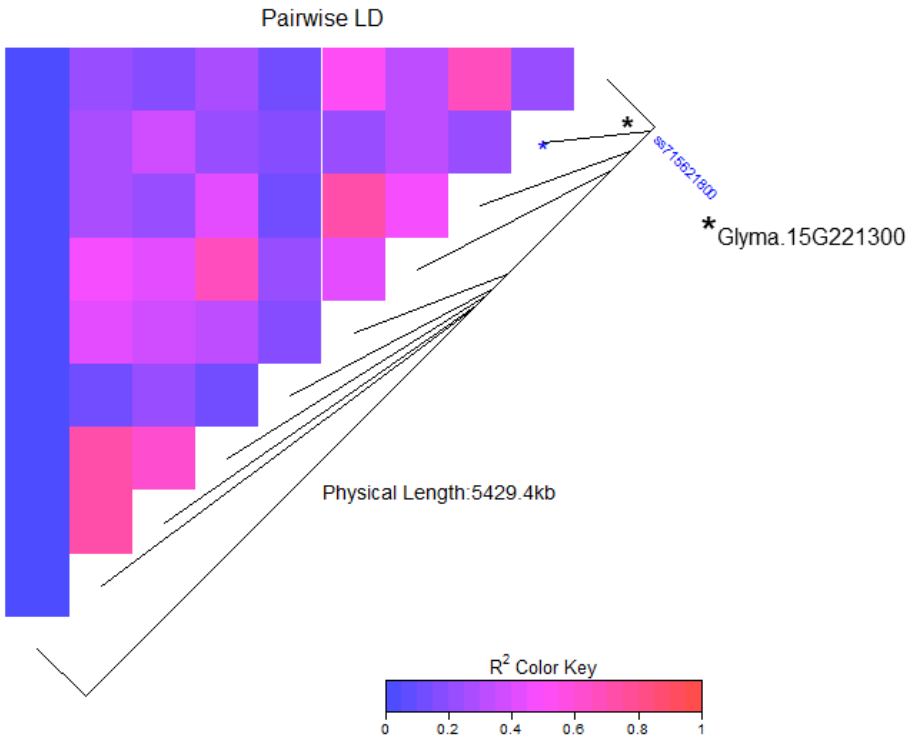


Figure 3.6 A pairwise linkage disequilibrium analysis between the target SNP (ss715621800, indicated in blue text) and the candidate gene *Glyma.15G221300* (represented by a black asterisk). This analysis is conducted within the context of our QTL-multiple metabolite cluster 1 (QTL1). The plot provides a visual representation of the extent of allelic association and LD between the SNP and the candidate gene, indicating their proximity and potential functional relationship within the genomic region of interest.

The identified candidate gene showed enrichment in the triterpenoid pathway, which aligns well with our target metabolites within QTL1. To be specific, the *Glyma.15G221300* gene encodes a protein identified as a soyasapogenol B glucuronide galactosyltransferase, suggesting its potential function as a UDP-glucosyl transferase involved in the critical step of soyasaponin biosynthesis (**Table 3.3**). The *Glyma.15G221300* gene has three homologous UGT genes (*GlysoPI483463.15G190800.1*, *GlymaLee.15G192500.1*, and *GlymaFiskIII.15G208700.1*). The coding sequences of *GlysoPI483463.15G190800.1* and *GlymaLee.15G192500.1* gene are similar to *Glyma.15G221300*, consisting of 1506 bp, while *GlymaFiskIII.15G208700.1* has a coding sequence of 1452 bp.

Table 3.3 The following is a compilation of eight Quantitative Trait Loci (QTLs) along with their respective significant SNPs, candidate genes, and brief descriptions.

Metabolite	**Significant SNP**	**Candidate gene**	**Description**
Triterpenoids (Soyasaponin) QTL1	ss715621800	*Glyma.15G221300*	UDP-glucosyl transferase 73B3/ soyasapogenol B glucuronide galactosyltransferase-like (LOC100810117)
Isoflavonoids QTL2	ss715632654	*Glyma.18G010400*	Protein phosphatase 2A regulatory B subunit family protein/serine/threonine

			protein phosphatase 2A 57 kDa regulatory subunit B' theta isoform-like (LOC100797758)
Epibreynin H QTL3 No annotation	ss715594698	*Glyma.06G285700*	UDP-Glycosyltransferase superfamily protein Metabolic domain: Carbohydrates metabolism; Nucleotides metabolism; Phenylpropanoid derivatives; Specialized metabolism‡
”	”	*Glyma.06G286200*	O-methyltransferase family protein Metabolic domain: Phenylpropanoid derivatives; Specialized metabolism
Phenolic acid/Polyphenols QTL4	ss715628514 ss715628531	*Glyma.18G103400* *Glyma.18G103500* *Glyma.18G103600* *Glyma.18G104000*	spermidine hydroxycinnamoyl transferase

		Glyma.18G104100	Metabolic domain: Phenylpropanoid derivatives; Specialized metabolism
QTL5 No annotation	ss715583518	*Glyma.02G057500*	Cytochrome P450 superfamily protein/cytochrome P450 90B1-like (LOC100800210)
Isoflavonoids QTL6	ss715599702	*Glyma.08G020000*	Disease resistance-responsive (dirigent-like protein) family protein/ disease resistance response protein 206-like (LOC100820276)
Isoflavonoids QTL7	ss715614341	*Glyma.11G209900* *Glyma.11G210300* *Glyma.11G210400* *Glyma.11G210500*	Specialized metabolism; Terpenoids; Phenylpropanoid derivatives
Alkaloids QTL8	ss715587592	*Glyma.04G156700*	biosynthesis of terpenoid indole alkaloids (TIAs)

	Metabolic domain: Nitrogen-containing compounds; Specialized metabolism

QTL (Quantitative Trait Locus): Multiple Metabolites within a cluster and their associated SNPs

‡Primary-specialized interface metabolism

* Please note that these descriptions are summarized and further investigation (i.e., functional validation) is required for a comprehensive understanding of the exact roles and functions of these candidate genes within their respective QTLs.

Genetic variations among ecotypes

Our study identifies the UDP-dependent glycosyltransferase *(UGT)* gene as a potential candidate responsible for the variation in soyasaponin production. Specifically, the gene Glyma.15G221300 consists of exon-1, intron, and exon-2 with lengths of 513 bp, 425 bp, and 993 bp, respectively. Sequence analysis comparing 46 ecotypes with varying soyasaponin production revealed specific coding sequence variations within exon-1 at positions 302, 451, and 500 for 29 ecotypes. These variations led to amino acid changes, namely from proline to glycine, serine to glycine, and serine to leucine, respectively (**Figure 3.7**). The frequency of CAC haplotype was higher in South Korea and Japan, while AGT haplotype frequency was higher in China and Russia (**Supplementary Figure 3.5**).

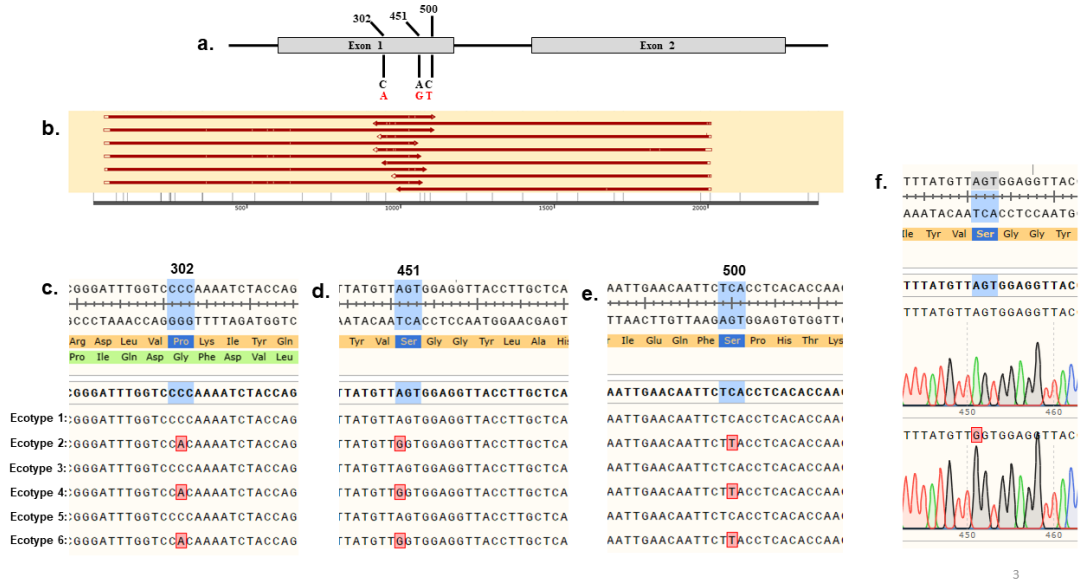


Figure 3.7 Genetic variations within the identified *UGT* gene among different *G. soja* ecotypes. Figure **(a)** depicts a schematic diagram of the UDP-glucosyl transferase *(UGT)* gene, which is a potential candidate involved in soyasaponin variation. In Figure **(b, c, d, e)**, an overview of SNPs in the coding sequence is depicted, highlighting the specific amino acid changes among ecotypes with varying soyasaponin production. Figure **(f)** represents a chromatogram of the sequencing data's peaks.

Further analysis using unpaired t-test with Welch's correction revealed a highly significant difference (P-value < 0.0001, specifically P-value = 0.00002) in soyasaponin accumulation between ecotypes with sequence variation (AGT haplotype blocks) and those without any sequence variation (CAC haplotype blocks, which are similar to the reference G. max Wm82.a4.v1 gene). (**Figure 3.8**).

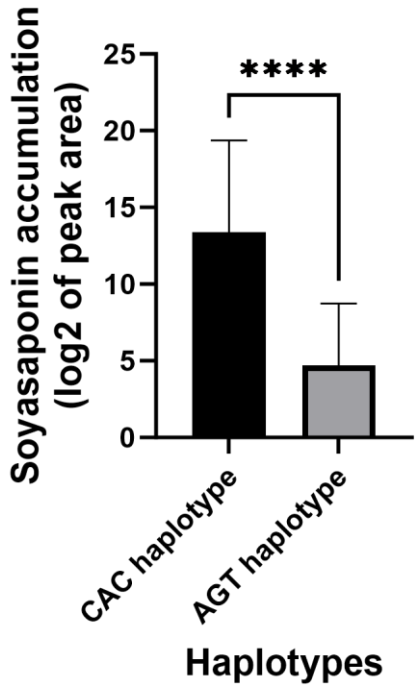


Figure 3.8 Association of two haplotypes with soyasaponin accumulation in wild soybean ecotypes. The X-axis represents haplotypes, with the AGT haplotype associated with sequence variation in wild soybean ecotypes compared to the reference *G. max* Wm82.a4.v1 gene, and the CAC haplotype representing wild soybean ecotypes with no sequence variation compared to the reference gene. The Y-axis corresponds to soyasaponin accumulation (Grp 766 shown here; Grp 70 shows similar significant association), measured on a logarithmic scale (log2 transformed). **** indicates extremely significant P-value (P-value <0.0001).

We conducted a comprehensive analysis of sequence variation among various *G. max* cultivars, which yielded intriguing findings. Upon examining 34 *G. max* cultivars, we observed sequence variations at multiple positions. Specifically, we identified two haplotypes, namely AGT (with sequence variation) and CAC (with no sequence variation), similar to the haplotypes for wild soybean ecotypes mentioned above. Notably, among these cultivars, the 29 ecotypes exhibited sequence variations (AGT haplotype) specifically at positions 302, 451, and 500. These variations closely resembled those observed in *G. soja* ecotypes known for their soyasaponin production variation (**Supplementary Figure 3.6**). Additionally, preliminary data from partial sequencing of the promoter region of 12 *G. soja* ecotypes shows a pattern of sequence variation among different ecotypes associated with soyasaponin production variation (**Supplementary Figure 3.7**). To draw conclusive evidence regarding the influence of the regulatory region on soyasaponin accumulation pattern, a comprehensive sequencing of the promoter region including more *G. soja* ecotypes is necessary.

Variability in gene expression of the candidate gene

Our study sheds light on the gene expression variation observed among different ecotypes, specifically between varying soyasaponin-producing ecotypes (**Figure 3.9a, b and c**). The targeted gene, characterized by exon-1 (513 bp) and exon-2 (993 bp), with an intron spanning 425 bp, displayed intriguing sequence distinctions between the ecotypes. We uncovered significant gene expression variations in both exon-1 with or without SNP and exon-2. The results obtained from the unpaired t-tests with Welch's correction provide additional evidence to support this finding. Specifically, we observed a significant association, indicated by the low P-values, between gene expression variation

and soyasaponin accumulation. For exon-1 (**Figure 3.9a**), the P-value was 0.0007 when considering the SNP at position 302; while for exon-1 without any SNP coverage (**Figure 3.9b**), the P-value was 0.0005. Additionally, for exon-2 (**Figure 3.9c**), the P-value was 0.0009.

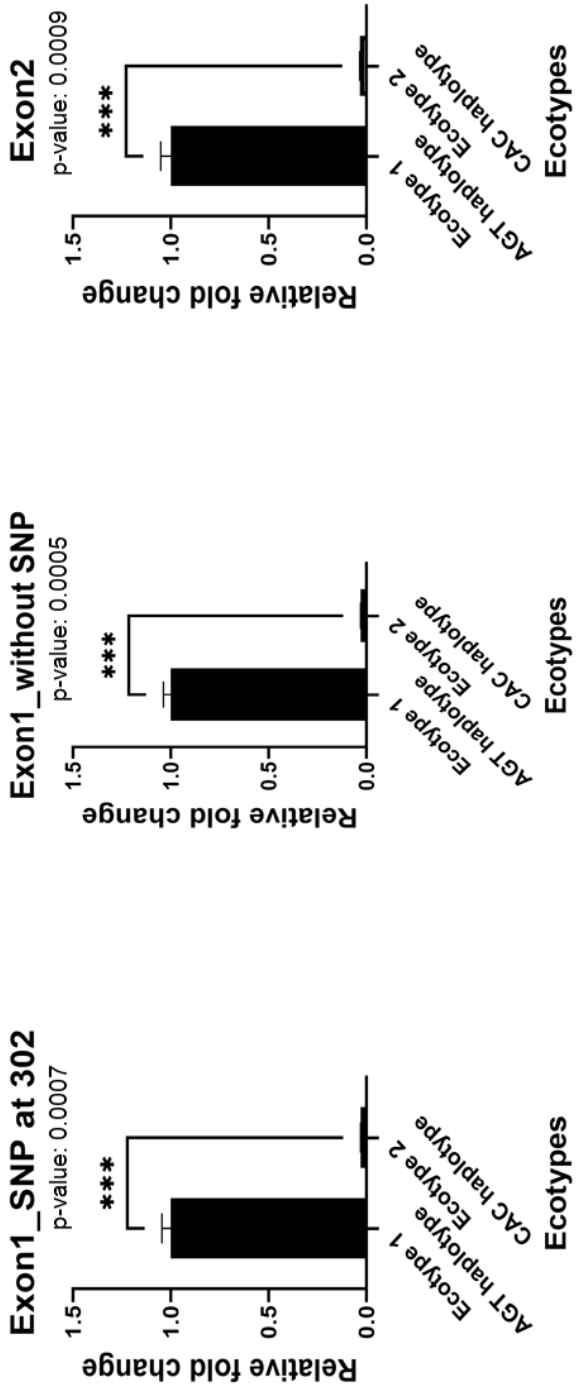


Figure 3.9 Differential expression of the candidate gene *(UGT)* between ecotypes with soyasaponin production variation. The results of the unpaired t-tests with Welch's

correction further support this finding, with a P-value of 0.0007 for exon-1 covering SNP at position 302 **(a)**, a P-value of 0.0005 for exon-1 without covering any SNP **(b)**, and a P-value of 0.0009 for exon-2 **(c)**. *** indicates extremely significant P-value (P-value is in between 0.0001 to 0.001).

These statistical findings reinforce the strong relationship between gene expression patterns and the observed variations in soyasaponin levels in different *G. soja* ecotypes. These findings, coupled with the identified promoter region sequence variations (preliminary data, **Supplementary Figure 3.7**), provide an indication that transcriptional regulation may play a pivotal role in modulating the expression pattern of this gene and thus influencing soyasaponin variations. For further investigation, a comprehensive understanding of the promoter region sequence with an increased number of sample sizes is essential.

Discussion

Understanding the genetic basis of trait variation within a population is crucial for unraveling the complexity of biosynthetic pathways. Recent studies have made significant progress in uncovering the genetic basis of natural variations in metabolic traits across various plant species, including Arabidopsis, tomato, maize, rice, wheat, and potato (Chan et al., 2011, Tieman et al., 2017, Harjes et al., 2008, Riedelsheimer et al., 2012, Wen et al., 2014, Chen et al., 2014a, Matsuda et al., 2015, Itkin et al., 2013, Sue et al., 2011). These studies have provided insights into the genetic mechanisms responsible for the considerable variation in plant metabolomes within a species, highlighting the presence of diverse and distinct metabolic pathways (Fernie and Tohge, 2017). Besides, metabolic diversity could be influenced by the environmental conditions of their habitats (Futuyma and Agrawal, 2009, Li et al., 2015). To comprehensively study the metabolome variation, mGWAS emerges as a suitable approach. It not only provides insights into the genetic basis of phytochemical diversity but also offers potential biochemical and functional understanding of the underlying metabolic pathways (Chen et al., 2014b).

However, the majority of these pathways remain unexplored and our study represents a step in the direction of exploring these pathways.

Motivated by this knowledge gap, our study aimed to investigate the genetic basis underlying the wide array of specialized metabolites found in diverse wild soybean ecotypes. To accomplish this, we employed an untargeted metabolomics approach to assess the abundance and diversity of metabolites present in *G. soja* leaves across North-East Asia, encompassing China, Japan, Russia, and South Korea. It is worth noting that a substantial portion of these metabolites still remains unknown, fueling the researchers' curiosity about the uncharted territory of these compounds.

Genetic basis of phytochemical diversity in wild soybean

In our analysis, we identified a significant number of associations between genetic variations (i.e., 1155 SNPs) and a diverse set of specialized metabolites (359 metabolites) within different metabolic pathways. The higher number of associations identified in our study compared to maize may be attributed to the greater genetic diversity observed in *G. soja* (Riedelsheimer et al., 2012, Zhang et al., 2017b). This observation underscores the complex genetic architecture underlying metabolite production in wild soybean. Furthermore, previous GWA studies conducted on the metabolomes of rice and the model plant Arabidopsis, in addition to our own study, have provided evidence of a complex genetic architecture, indicating intricate genetic regulation of these compounds (Chan et al., 2011, Chen et al., 2014a).

By clustering the metabolites into eight QTLs based on the associated SNPs, we facilitated large-scale gene-metabolite annotation in wild soybeans. This approach along

with linkage disequilibrium analysis enabled us to uncover potential causative genes and their associated SNPs that contribute to the accumulation of specialized metabolites within these ecotypes (**Table 3.3**). Metabolite-based genome-wide association studies that utilize linkage disequilibrium analysis offer a highly precise approach for identifying candidate genes responsible for phenotypic variation. Thus, providing us a comprehensive and detailed understanding of the genetic basis underlying diverse metabolite profiles (Remington et al., 2001). Our findings shed light on the genetic foundations of the wild soybean metabolome, with a particular emphasis on alkaloid, isoflavonoid, polyphenol, and triterpenoid pathways. Moving forward, future research could delve deeper into the molecular mechanisms of these pathways, building upon the insights gained from our study (Chen et al., 2016).

Unveiling a promising candidate gene in soyasaponin biosynthesis pathway

We focused our investigation on a specific QTL associated with the soyasaponin biosynthesis pathway and discovered probable genetic control over soyasaponin production variation. In soybean, the glycosyltransferase (GT) family 1 of the GT superfamily, particularly the UGTs (UDP-dependent glycosyltransferases), play a crucial role in the glycosylation of saponins (Campbell et al., 1997, Lairson et al., 2008). Previous studies have identified five *UGT* genes encoding enzymes involved in the glycosylation process, with four of them (*GmSGT2*, *GmSGT3*, *Sg-3*, and *Sg-4*) responsible for sugar transfer at the C-3 position of soyasaponins, and the *Sg-1* gene encoding enzyme involved in sugar chain formation at the C-22 position of soyasapogenol A (Shibuya et al., 2010, Takada et al., 2012, Chitisankul et al., 2015, Yano et al., 2018, Sayama et al., 2012).

In this study, our identified *UGT* gene *(Glyma.15G221300)* encodes an enzyme named Soyasapogenol B glucuronide galactosyltransferase. Our analysis revealed that *Glyma.15G221300* is a single copy gene with three SNPs (AGT haplotype for low soyasaponin II producing ecotypes) located within exon-1, resulting in amino acid changes. These sequence variations indicate that natural variation may contribute to the variations observed in soyasaponin abundance (Soltis and Kliebenstein, 2015, Huang and Han, 2014). Moreover, the discovery of similar haplotypes in cultivated soybeans highlights the presence of shared genetic traits between *G. max* ecotypes containing AGT haplotype and low-soyasaponin II-producing *G. soja* ecotypes, suggesting potential links in their metabolic pathways. Interestingly, among the 34 different *G. max* cultivars, a higher frequency of AGT haplotypes (29 ecotypes with AGT haplotype) and their incidental association with metabolite groups exhibiting low soyasaponin II accumulation raises questions. Studies have shown an association between soyasaponin and bitter taste in pea seeds (Roland et al., 2017, Munakata, 2021). Our hypothesis suggests that the higher frequency of the AGT haplotype and its incidental association with low soyasaponin II levels might be linked to selection pressures aimed at bitterness control, which in turn could have influenced the prevalence of AGT haplotypes in different cultivated soybeans. Crop domestication and cultivation have significantly impacted the selection of desirable traits in crops, leading to notable changes in their phenotypic variation. These alterations encompass increased biomass, reduced or eliminated toxic compounds, and modifications in flavor profiles (Evans, 1996, Ladizinsky, 2012, Meyer et al., 2012).

Furthermore, our analysis, which includes *UGT* gene expression variations, as well as preliminary data on promoter region sequence variations, suggests a possible involvement of the regulatory region of the gene in shaping the pattern of soyasaponin accumulation. It is important to consider that several factors may contribute to the generation of metabolic variance in this context. These factors encompass: (i) variations in promoter strength caused by variations in methylation or copy number in the promoter region; (ii) single-nucleotide polymorphisms in the coding region affecting enzymatic activity, substrate preference, or both; (iii) polymorphisms leading to a premature stop codon; and (iv) transposons-induced significant gene deletions or insertions (Fernie and Tohge, 2017). For instance, variations in the *Game9* transcription factor gene in tomato and the promoter of *DXS* in kiwifruit lead to changes in the accumulation pattern of steroidal glycoalkaloid and monoterpene metabolites, respectively (Zhu et al., 2018b, Nieuwenhuizen et al., 2015).

At the transcriptional level, gene regulation plays a crucial role in modulating the expression patterns. The binding of transcriptional regulatory factors to promoter regions of target genes, as well as their interactions with DNA-binding proteins, can result in variations in gene expression patterns (Tamura et al., 2018, Yang et al., 2012). This raises the question of whether the difference in soyasaponin abundance is influenced by the combined effects of gene sequence variation and the regulatory region. Secondary metabolism in plants demonstrates a higher capacity to tolerate mutations and adapt to changing environments compared to primary metabolism, which is more evolutionarily constrained (Weng et al., 2012, O'Maille et al., 2008). The observed difference in soyasaponin abundance could be solely attributed to the identified sequence variation in

exon-1, highlighting the crucial role of coding sequence variation that directly affects the enzymatic activity and substrate preference (Fernie and Tohge, 2017). A pairwise linkage analysis revealed that the significant SNP ss715621800 is located 6.3 kb upstream of the *UGT* gene. The presence of the SNP in close proximity to the *UGT* gene suggests a potential genetic linkage or association between the SNP and the gene (Brodie et al., 2016). This proximity indicates the existence of a possible complex genetic control mechanism, as the SNP's location could influence the regulation or expression of the *UGT* gene (Shastry, 2009).

Overall, our study highlights the potential influence of genetic variations, enzymatic activity and regulatory mechanisms in shaping the abundance of soyasaponin, particularly with regards to the identified *UGT* gene *(Glyma.15G221300)*. Exploring the mechanistic impact of selection on determining the optimal combination of haplotypes that enhance metabolic fitness at the individual or species level would be a fascinating area of study (Soltis and Kliebenstein, 2015). Further research is needed to explore the intricate relationship between gene sequence variations, enzymatic activity, regulatory regions, and the resulting soyasaponin accumulation patterns.

Limitations, opportunities and future directions

In our study, we aimed to uncover the molecular basis of phytochemical diversity in wild soybean, specifically focusing on the soyasaponin biosynthesis pathway. To achieve this, we employed a combination of GWAS, metabolomics, and molecular biology techniques. The identification and annotation of these phytochemicals still present some challenges due to their high diversity. For example, in untargeted metabolomics, annotating metabolites involves a mass-based search against databases, followed by manual

verification, which can be both costly and labor-intensive (Xiao et al., 2012). Nonetheless, metabolomics has emerged as a crucial tool for annotating genes' functions and gaining a comprehensive understanding of cellular responses in various biological scenarios (Schauer and Fernie, 2006, Hegeman, 2010, Xiao et al., 2022). Our study encompassed 190 wild soybean ecotypes, representing their original geographic distribution. To enhance the resolution of our study, a sample size of 400-500 would be ideal for mGWAS analysis. Increasing the sample size would provide a more comprehensive understanding of the molecular and evolutionary factors influencing phenotypic variation within a population (Korte and Farlow, 2013, Kliebenstein, 2007b).

Metabolite annotation poses a well-known challenge in untargeted metabolomics studies, and as expected, we encountered numerous unknown metabolites. Synthesizing standard compounds is both time-consuming and cost-intensive. Additionally, in our genotyping process, we utilized 41,896 known SNPs, which, although not densely distributed, still provided a solid foundation for our analysis. However, a study by Katz et al. (2021) revealed that in Arabidopsis, complications arose when the number of genotypes and SNP marker density increased, potentially due to an uneven and sparse sampling across Europe (Katz et al., 2021). Despite this, the known SNPs we employed serve as valuable starting points, laying the groundwork for exploring the unexplored realm of phytochemical diversity. Our findings will greatly contribute to the identification of causal sequence variants, particularly with the assistance of advanced genotyping platforms such as next-generation sequencing (NGS) (Elshire et al., 2011, Davey et al., 2011).

Our study system benefits from the availability of functional study toolkits, providing a valuable advantage. Detailed information on the plasmid constructs that will be used for the functional validation of our identified candidate gene through virus-induced gene silencing and CRISPR/Cas9 technologies is provided in the following sections.

Plasmid constructs for functional validation of candidate genes

To confirm the functional role of the selected candidate genes such as candidate gene *UGT* in soyasaponin biosynthesis pathway, we will utilize virus-induced gene silencing (VIGS) as a potent tool in plant functional genomics. VIGS utilizes the plant's natural defense mechanism called post-transcriptional gene silencing (PTGS) (Kumagai et al., 1995, Robertson, 2004). A crucial step in our methodology will involve the application of *Agrobacterium*-mediated infiltration, also known as agro-infiltration, to introduce an infectious plasmid DNA harboring the Bean pod mottle virus (BPMV) vector into the leaves of soybean *(Glycine max)* (Zhang et al., 2010, Zhang et al., 2013).

Following the delivery of the viral vector (**Figure 3.10**), PTGS initiates the degradation of the target gene's mRNA in a sequence-specific manner, resulting in the downregulation of the *UGT* gene (Burch-Smith et al., 2004). To evaluate the gene function, we will assess the soyasaponin production using liquid chromatography-mass spectrometry (LC-MS). Through VIGS, we anticipate achieving knockdown of the *UGT* gene in different soybean ecotypes, which will validate its role in soyasaponin biosynthesis. Additionally, we will analyze the expression levels of the *UGT* gene in infected plants using RT-qPCR. BPMV-based VIGS vectors are extensively utilized in legumes because of their ability to efficiently silence native genes. They are favored for their ease of use and speedy process (achieving silencing within 3-4 weeks after

infection), eliminating the requirement for establishing stable transformants (Burch-Smith et al., 2004). However, VIGS does have certain limitations that should be taken into account. A notable limitation is its inclination to entirely silence the target gene, posing challenges when conducting comparative analysis between soybean ecotypes exhibiting soyasaponin production variation (Burch-Smith et al., 2004). Complete suppression hinders the observation of the specific phenotype of interest and complicates the determination of the *UGT* gene's involvement in our proposed functional context. Furthermore, VIGS frequently falls short in achieving consistent gene silencing across the entire infected plant and can unintentionally suppress non-target genes (Burch-Smith et al., 2004).

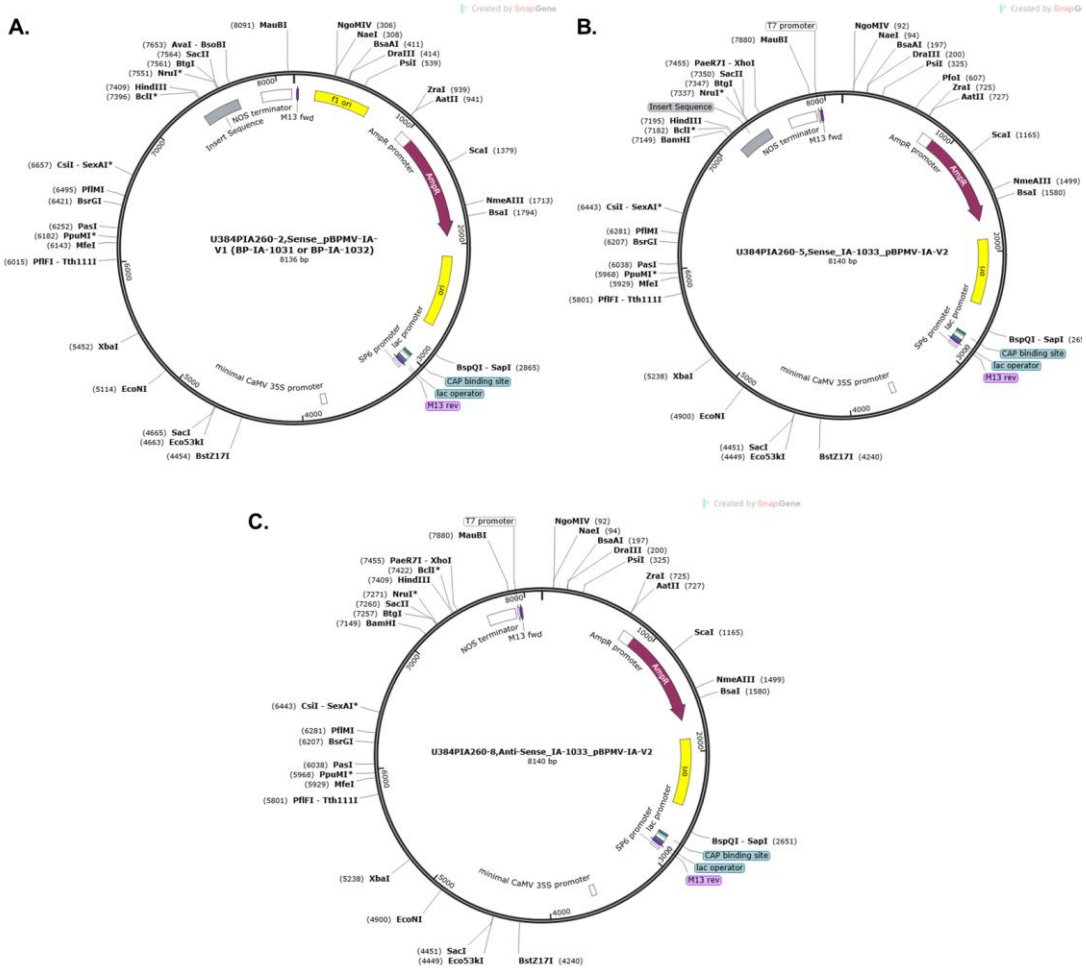


Figure 3.10 Plasmid constructs for virus-induced gene silencing of the target gene. Gene of interest in sense orientation in pBPMV-IA-V1 **(A)**; gene of interest in sense orientation in IA-1033_pBPMV-IA-V2 **(B)**; gene of interest in antisense orientation in IA-1033_pBPMV-IA-V2 **(C)**.

To overcome these limitations and ensure more precise functional validation of the identified *UGT* gene, we intend to utilize state-of-the-art genome editing technology CRISPR/Cas9. It is a widely adopted technique in various crops, including soybeans. This advanced technique will provide us with greater control and specificity in confirming our findings, thereby circumventing the pitfalls associated with VIGS. Quantification of soyasaponin production will be performed using LC-MS on the leaves of T2 plants. We will ensure a sufficient number of independent replicates for robust statistical analysis. Furthermore, the expression levels of the *UGT* gene in the edited plants will be assessed using RT-qPCR. Through gene knockout using the CRISPR/Cas9 system (**Figure 3.11**), we anticipate a complete absence of soyasaponin production, thus confirming the involvement of the identified *UGT* gene in soyasaponin biosynthesis.

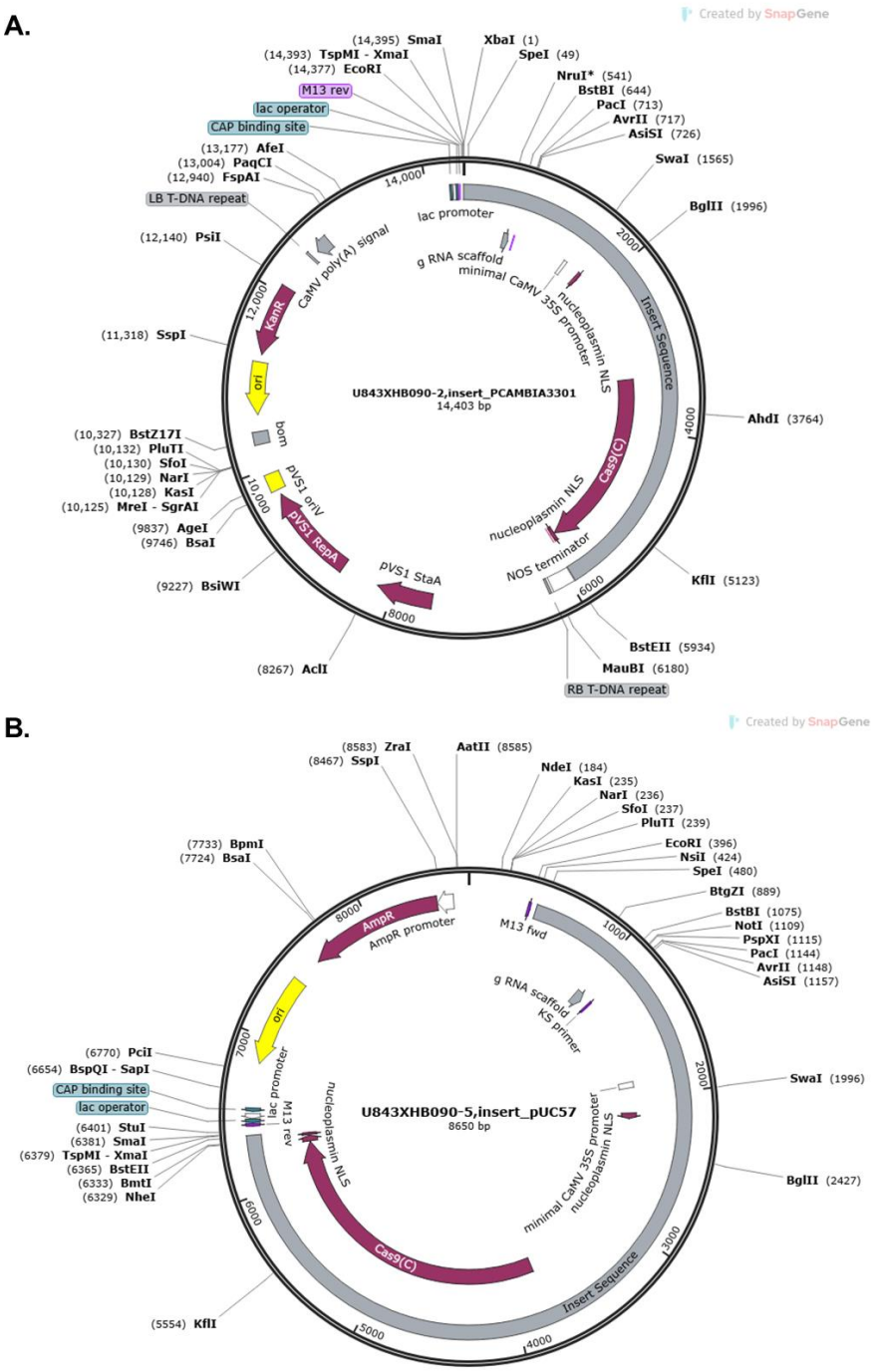


Figure 3.11 CRISPR/Cas9 constructs for gene editing and transformation.

Conclusion

This study represents an application in unraveling the genetic foundations of phytochemical diversity, specifically focusing on the soyasaponin pathway in wild soybean. By delving into the genetic underpinnings of phytochemical diversity, our research lays the groundwork for understanding biosynthetic pathways, enabling the development and integration of genomic tools for applications in metabolic engineering and molecular breeding of desirable plant compounds.

Throughout our investigation, we have identified several candidate genes that are associated with a wide range of metabolic pathways. These findings have significant implications for promoting higher genetic diversity within populations. We utilized a candidate-gene association approach to narrow down the number of candidate genes located within the metabolic clusters, which were then subjected to further evaluation. This strategy proved effective in reducing the scale of genotyping required and minimizing the need for scanning a large number of markers across the entire genome (Flint-Garcia et al., 2003). However, it is important to note that validation experiments for all associations, including 16 candidate genes and 9 SNPs, exceed the scope of a single study. Our research findings highlight the potential of the reported *UGT* gene as a promising candidate responsible for the variation in soyasaponin production among wild soybean ecotypes. In our future endeavors, we intend to conduct experimental characterization of the *Glyma.15G221300* gene, which encodes the UDP-glucosyltransferase enzyme, thereby providing novel biochemical and functional insights. Furthermore, this study has contributed an additional 15 high-confidence candidate genes associated with diverse metabolite contents and elucidated the subdivisions within

metabolic networks (refer to **Table 3.3**). It is evident that both genetic factors and environmental influences play a role in determining metabolite levels. For example, a study on *Arabidopsis thaliana* demonstrated how metabolic profiles are influenced by geographic variability, highlighting the significance of local adaptation strategies in response to a changing environment (Kleessen et al., 2012). By exploring the metabolome, we can expand our understanding of effective defense mechanisms against biotic and abiotic stresses and contribute to improving the nutritional quality of plants, thus promoting sustainable agriculture.

Additionally, our study provides a valuable platform for addressing a relatively neglected question regarding the impact of genetic variation on changes in metabolome levels during the domestication process. We also investigate how the geographical structure of natural populations shapes metabolite profiles. Similar studies conducted on essential crop species such as maize, rice, and durum wheat have utilized metabolite profiling as a molecular phenotyping tool to explore the process of crop domestication (Riedelsheimer et al., 2012, Wen et al., 2014, Skogerson et al., 2010, Chen et al., 2014a, Beleggia et al., 2013). In essence, our research opens up new avenues for examining the soybean crop domestication process and the adaptation of plants in their natural geographic distribution from a metabolic perspective. By investigating into the intricate metabolic aspects, we can gain valuable insights into the evolution and development of this important crop. In future research, phylogenetic comparative methods can be utilized to explore the evolutionary mechanisms that contribute to the development and diversification of phytochemical diversity (Forrister et al., 2023).

CHAPTER 4: OVERALL CONCLUSION

Plant science is instrumental in addressing the urgent challenges faced by floral and faunal habitats, safeguarding the future of our planet. Plants have evolved to produce a wide range of specialized metabolites, which enable them to adapt to their ever-changing environment (i.e., heat, drought, salt stress, flooding, disease outbreaks). With the escalation of climate change's effects on agricultural systems, there is a growing demand to cultivate crops capable of withstanding stress, as a means to address the issue of food insecurity. Hence, it is crucial to understand how plants respond and adapt to various stresses, both biotic (caused by living organisms) and abiotic (caused by environmental factors). Some of these metabolites can potentially improve human health. Therefore, the study of plant-specialized metabolism offers opportunities for advancing agricultural sustainability and holds biomedical relevance, aiming to develop sustainable nutrition sources and novel pharmaceuticals. Unraveling the wide range of specialized metabolites produced by plants constitutes a fundamental aspect of this research endeavor.

The specific focus of this study is twofold. Firstly, in chapter 2, this research investigates the genetic basis and selection processes involved in the glyceollin pathway, mainly when plants interact with biotic stressors in a controlled laboratory environment. Secondly, in chapter 3, it aims to unravel the genetic basis that governs the diversity of phytochemicals and variations in soyasaponin production within natural plant populations. By understanding these mechanisms, we can gain insights into the factors influencing the synthesis of beneficial plant compounds. By undertaking this research, we anticipate to advance our knowledge of plant responses to stresses, the genetic factors influencing specialized metabolite production, and the potential for optimizing plant-

derived compounds to address global challenges. Ultimately, these insights will contribute to the development of sustainable strategies for both environmental conservation and human well-being.

Glyceollins, a type of phytoalexin found in soybean plants, play crucial roles in plant responses to the environment and human health. However, our understanding of the genetic mechanisms and factors governing glyceollin production remains limited. In chapter 2, we investigated the genetic components underlying glyceollin production in wild soybeans exposed to soybean cyst nematode using a metabolite-based genome-wide association (mGWA) approach. Through this analysis, we identified specific genetic variations, known as single nucleotide polymorphisms (SNPs), on chromosomes 3, 9, 13, 15, and 20, strongly associated with glyceollin induction. Notably, we observed a cluster of genes closely located near one of these significant SNPs on chromosome 9, encoding enzymes that may play a vital role in the production of glyceollins.

Additionally, we found transcription factors (genes that regulate the activity of other genes), such as *MYB* and *WRKY*, close to these genetic variations on chromosome 9. Furthermore, our findings revealed interactions between various genetic variations and provided evidence of natural selection, suggesting the potential impact of evolutionary processes on glyceollin production in wild soybeans. Overall, our findings shed light on the key genes and factors in controlling glyceollin induction in wild soybeans, further enhancing our understanding of this critical plant defense mechanism.

Moreover, in order to address the limited understanding of genetic factors contributing to phytochemical diversity in wild plant species, particularly in understudied wild soybeans, we conducted a comprehensive investigation described in chapter 3. Wild soybeans,

known as *Glycine soja*, possess a vast and untapped genetic diversity compared to cultivated soybeans. We employed a metabolite-based genome-wide association (mGWA) approach, which involved analyzing a diverse set of 190 wild soybean ecotypes collected from a range of geographic origins. We successfully identified and annotated 485 out of 874 metabolite peaks through untargeted metabolite profiling using LC-MS analysis. Leveraging 41,896 genome-wide SNPs, we performed a thorough genome-wide association study (GWAS) on these 874 metabolite peaks and discovered significant associations between 1155 SNPs and 359 metabolites. Clustering analysis allowed us to identify eight quantitative trait loci (QTLs) representing clusters of multiple metabolites. Further investigation of these QTLs within linkage disequilibrium blocks led us to identify 612 annotated genes as potential candidate genes.

Among these candidates, we focused on 16 candidate genes relevant to triterpenoid and phenylpropanoid-derived isoflavonoid biosynthesis pathways. Notably, our analysis highlighted the gene responsible for UDP-dependent glycosyltransferase *(UGT)* as a promising candidate that likely plays a key role in soyasaponin production variation, an important phytochemical. Sequence analysis revealed two distinct haplotypes associated with varying soyasaponin-producing ecotypes, with three specific SNPs located at exon-1 resulting in amino acid changes. Furthermore, substantial differences in expression levels were observed between the two haplotypes when comparing varying soyasaponin-producing ecotypes. An intriguing observation is the coincidental connection between the AGT haplotype and low soyasaponin II accumulation, coupled with its higher frequency in various cultivated soybean ecotypes. This puzzling finding warrants further evaluation and investigation to better understand its significance and implications. Our study

represents a significant step toward unraveling the genetic foundations of phytochemical diversity in understudied wild non-model species, specifically wild soybeans. By shedding light on the intricate genetic basis underlying the phytochemical diversity of wild soybeans, our findings contribute to a broader understanding of the genetic landscape of specialized metabolism in plant species.

The discoveries resulting from this study have the potential to revolutionize metabolic engineering and contribute to the development of biofortified crops that are resilient to future challenges, ensuring a sustainable future. As described in chapter 2, understanding the metabolic gene clusters associated with the induction of glyceollins in response to biotic elicitors in wild soybean will shed light on the plant's defense mechanisms. Additionally, the findings from chapter 3 can contribute to unravel the molecular mechanisms underlying the variation in soyasaponin production within genetically diverse wild soybean populations. Our study will enable the discovery of new and exotic genetic resources within natural populations and offer valuable insights into the glyceollin and soyasaponin biosynthesis pathways. Additionally, wild soybean serves as an exceptional study system, facilitating the transfer of identified candidate genes to cultivated soybeans, thus enabling the development of soybean cultivars rich in unique and valuable phytochemicals. Lastly, this project embodies a highly interdisciplinary approach, integrating molecular genetics, metabolomics, and genome-wide association studies (GWAS) to address the research objectives comprehensively.

In future studies, an RNA-seq analysis could be conducted to explore the co-expression of genes within gene clusters, potentially revealing their involvement in glyceollin induction. Furthermore, investigating environmental factors could help elucidate their

contribution to the selection of glyceollin induction. To achieve genomic dissection of phytochemical diversity, the identified candidate genes will undergo functional validation using advanced techniques, such as virus-induced gene silencing and state-of-the-art gene editing technologies like CRISPR/Cas9. These approaches will enable us to gain deeper insights into the functions of these candidate genes and their impact on specialized metabolite production. Additionally, future research may focus on the biochemical characterization of the *UGT* gene, which possibly encodes the enzyme responsible for soyasaponin biosynthesis pathway. By studying this enzyme's characteristics and mechanisms, we can better understand its role in the production of soyasaponins, contributing to our knowledge of plant specialized metabolism. Overall, these proposed future studies will expand our understanding of the genetic and biochemical factors underlying glyceollin induction and soyasaponin biosynthesis, paving the way for advancements in plant science and potential applications in agriculture and human health. The findings derived from our study hold significant potential in developing climate-resilient crops with enhanced value, benefiting both plant and human health. If successful, the resulting biofortified soybean could become a daily staple accessible to people worldwide, including socio-economically disadvantaged regions, addressing nutritional deficiencies and improving overall well-being. Moreover, the insights gained from our study will serve as a foundation for generating and testing new hypotheses, advancing our understanding of complex traits related to plant and human health and paving the way for similar strategies in future research. With its highly interdisciplinary approach integrating molecular genetics, phytochemistry, metabolomics, and genome-

wide association studies (GWAS), our study comprehensively addresses the research objectives while contributing to the broader scientific community.

References

ABE, I. 2007. Enzymatic synthesis of cyclic triterpenes. *Natural product reports,* 24**,** 1311-1331.

AFENDI, F. M., OKADA, T., YAMAZAKI, M., HIRAI-MORITA, A., NAKAMURA, Y., NAKAMURA, K., IKEDA, S., TAKAHASHI, H., ALTAF-UL-AMIN, M. & DARUSMAN, L. K. 2012. KNApSAcK family databases: integrated metabolite–plant species databases for multifaceted plant research. *Plant and Cell Physiology,* 53**,** e1-e1.

AHARONI, A., GAIDUKOV, L., KHERSONSKY, O., GOULD, S. M., ROODVELDT, C. & TAWFIK, D. S. 2005. The'evolvability'of promiscuous protein functions. *Nature genetics,* 37**,** 73-76.

AHN, K.-S., KIM, J.-H., OH, S.-R., MIN, B.-S., KINJO, J. & LEE, H.-K. 2002. Effects of oleanane-type triterpenoids from fabaceous plants on the expression of ICAM-1. *Biological and Pharmaceutical Bulletin,* 25**,** 1105-1107.

AHUJA, I., KISSEN, R. & BONES, A. M. 2012. Phytoalexins in defense against pathogens. *Trends in plant science,* 17**,** 73-90.

AKASHI, T., SASAKI, K., AOKI, T., AYABE, S.-I. & YAZAKI, K. 2009. Molecular cloning and characterization of a cDNA for pterocarpan 4-dimethylallyltransferase catalyzing the key prenylation step in the biosynthesis of glyceollin, a soybean phytoalexin. *Plant physiology,* 149**,** 683-693.

ALBERSHEIM, P. & VALENT, B. S. 1978. Host-pathogen interactions in plants. Plants, when exposed to oligosaccharides of fungal origin, defend themselves by accumulating antibiotics. *The Journal of Cell Biology,* 78**,** 627-643.

ALEXANDER, P., ROUNSEVELL, M. D., DISLICH, C., DODSON, J. R., ENGSTRÖM, K. & MORAN, D. 2015. Drivers for global agricultural land use change: The nexus of diet, population, yield and bioenergy. *Global Environmental Change,* 35**,** 138-147.

ALSEEKH, S. & FERNIE, A. R. 2018. Metabolomics 20 years on: what have we learned and what hurdles remain? *The Plant Journal,* 94**,** 933-942.

ANARAT-CAPPILLINO, G. & SATTELY, E. S. 2014. The chemical logic of plant natural product biosynthesis. *Current opinion in plant biology,* 19**,** 51-58.

ANDREOTE, F. D. & E SILVA, M. D. C. P. 2017. Microbial communities associated with plants: learning from nature to apply it in agriculture. *Current opinion in microbiology,* 37**,** 29-34.

ARITA, M. 2004. The metabolic world of *Escherichia coli* is not small. *Proceedings of the National Academy of Sciences,* 101**,** 1543-1547.

AUGUSTIN, J. M., KUZINA, V., ANDERSEN, S. B. & BAK, S. 2011. Molecular activities, biosynthesis and evolution of triterpenoid saponins. *Phytochemistry,* 72**,** 435-457.

AYERS, A. R., EBEL, J. R., FINELLI, F., BERGER, N. & ALBERSHEIM, P. 1976. Host-pathogen interactions: IX. Quantitative assays of elicitor activity and characterization of the elicitor present in the extracellular medium of cultures of Phytophthora megasperma var. sojae. *Plant physiology,* 57**,** 751-759.

BAI, Y. & LINDHOUT, P. 2007. Domestication and breeding of tomatoes: what have we gained and what can we gain in the future? *Annals of botany,* 100**,** 1085-1094.

BAMJI, S. F. & CORBITT, C. 2017. Glyceollins: Soybean phytoalexins that exhibit a wide range of health-promoting effects. *Journal of Functional Foods,* 34**,** 98-105.
BELEGGIA, R., PLATANI, C., NIGRO, F., DE VITA, P., CATTIVELLI, L. & PAPA, R. 2013. Effect of genotype, environment and genotype-by-environment interaction on metabolite profiling in durum wheat *(Triticum durum Desf.)* grain. *Journal of Cereal Science,* 57**,** 183-192.
BENDEROTH, M., TEXTOR, S., WINDSOR, A. J., MITCHELL-OLDS, T., GERSHENZON, J. & KROYMANN, J. 2006. Positive selection driving diversification in plant secondary metabolism. *Proceedings of the National Academy of Sciences,* 103**,** 9118-9123.
BERENDSEN, R. L., PIETERSE, C. M. & BAKKER, P. A. 2012. The rhizosphere microbiome and plant health. *Trends in plant science,* 17**,** 478-486.
BHATTACHARYYA, M. & WARD, E. 1985. Differential sensitivity of Phytophthora megasperma f. sp. glycinea isolates to glyceollin isomers. *Physiological plant pathology,* 27**,** 299-310.
BÖNISCH, F., FROTSCHER, J., STANITZEK, S., RÜHL, E., WÜST, M., BITZ, O. & SCHWAB, W. 2014. A UDP-Glucose:Monoterpenol Glucosyltransferase Adds to the Chemical Diversity of the Grapevine Metabolome *Plant Physiology,* 165**,** 561-581.
BOURION, V., HEULIN-GOTTY, K., AUBERT, V., TISSEYRE, P., CHABERT-MARTINELLO, M., PERVENT, M., DELAITRE, C., VILE, D., SIOL, M. & DUC, G. 2018. Co-inoculation of a pea core-collection with diverse rhizobial strains shows competitiveness for nodulation and efficiency of nitrogen fixation are distinct traits in the interaction. *Frontiers in Plant Science,* 8**,** 2249.
BOWLES, D., ISAYENKOVA, J., LIM, E.-K. & POPPENBERGER, B. 2005. Glycosyltransferases: managers of small molecules. *Current opinion in plant biology,* 8**,** 254-263.
BRACHI, B., MEYER, C. G., VILLOUTREIX, R., PLATT, A., MORTON, T. C., ROUX, F. & BERGELSON, J. 2015. Coselected genes determine adaptive variation in herbivore resistance throughout the native range of *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences,* 112**,** 4032-4037.
BRADBURY, P. J., ZHANG, Z., KROON, D. E., CASSTEVENS, T. M., RAMDOSS, Y. & BUCKLER, E. S. 2007. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics,* 23**,** 2633-2635.
BRAZIER-HICKS, M., GERSHATER, M., DIXON, D. & EDWARDS, R. 2018. Substrate specificity and safener inducibility of the plant UDP-glucose-dependent family 1 glycosyltransferase super-family. *Plant biotechnology journal,* 16**,** 337-348.
BRODIE, A., AZARIA, J. R. & OFRAN, Y. 2016. How far from the SNP may the causative genes be? *Nucleic acids research,* 44**,** 6046-6054.
BURCH-SMITH, T. M., ANDERSON, J. C., MARTIN, G. B. & DINESH-KUMAR, S. P. 2004. Applications and advantages of virus-induced gene silencing for gene function studies in plants. *The Plant Journal,* 39**,** 734-746.
BUSTOS-SEGURA, C., POELMAN, E. H., REICHELT, M., GERSHENZON, J. & GOLS, R. 2017. Intraspecific chemical diversity among neighbouring plants

correlates positively with plant size and herbivore load but negatively with herbivore damage. *Ecology Letters,* 20**,** 87-97.

BYRNE, P., MCMULLEN, M., SNOOK, M., MUSKET, T., THEURI, J., WIDSTROM, N., WISEMAN, B. & COE, E. 1996. Quantitative trait loci and metabolic pathways: genetic control of the concentration of maysin, a corn earworm resistance factor, in maize silks. *Proceedings of the National Academy of Sciences,* 93**,** 8820-8825.

CAMPBELL, J. A., DAVIES, G. J., BULONE, V. & HENRISSAT, B. 1997. A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. *Biochemical Journal,* 326**,** 929.

CAPUTI, L., MALNOY, M., GOREMYKIN, V., NIKIFOROVA, S. & MARTENS, S. 2012. A genome-wide phylogenetic reconstruction of family 1 UDP-glycosyltransferases revealed the expansion of the family during the adaptation of plants to life on land. *The Plant Journal,* 69**,** 1030-1042.

CÁRDENAS, P. D., SONAWANE, P. D., POLLIER, J., VANDEN BOSSCHE, R., DEWANGAN, V., WEITHORN, E., TAL, L., MEIR, S., ROGACHEV, I. & MALITSKY, S. 2016. *GAME9* regulates the biosynthesis of steroidal alkaloids and upstream isoprenoids in the plant mevalonate pathway. *Nature communications,* 7**,** 1-16.

CASPI, R., ALTMAN, T., BILLINGTON, R., DREHER, K., FOERSTER, H., FULCHER, C. A., HOLLAND, T. A., KESELER, I. M., KOTHARI, A. & KUBO, A. 2014. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic acids research,* 42**,** D459-D471.

CHAE, L., KIM, T., NILO-POYANCO, R. & RHEE, S. Y. 2014. Genomic signatures of specialized metabolism in plants. *science,* 344**,** 510-513.

CHAN, E. K., ROWE, H. C., CORWIN, J. A., JOSEPH, B. & KLIEBENSTEIN, D. J. 2011. Combining genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in *Arabidopsis thaliana*. *PLoS biology,* 9**,** e1001125.

CHAN, E. K., ROWE, H. C., HANSEN, B. G. & KLIEBENSTEIN, D. J. 2010. The complex genetic architecture of the metabolome. *PLoS genetics,* 6**,** e1001198.

CHAVALI, A. K. & RHEE, S. Y. 2018. Bioinformatics tools for the identification of gene clusters that biosynthesize specialized metabolites. *Briefings in bioinformatics,* 19**,** 1022-1034.

CHEN, J., XUE, M., LIU, H., FERNIE, A. R. & CHEN, W. 2021. Exploring the genic resources underlying metabolites through mGWAS and mQTL in wheat: From large-scale gene identification and pathway elucidation to crop improvement. *Plant Communications,* 2**,** 100216.

CHEN, W., GAO, Y., XIE, W., GONG, L., LU, K., WANG, W., LI, Y., LIU, X., ZHANG, H. & DONG, H. 2014a. Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nature genetics,* 46**,** 714-721.

CHEN, W., GAO, Y., XIE, W., GONG, L., LU, K., WANG, W., LI, Y., LIU, X., ZHANG, H., DONG, H., ZHANG, W., ZHANG, L., YU, S., WANG, G., LIAN, X. & LUO, J. 2014b. Genome-wide association analyses provide genetic and

biochemical insights into natural variation in rice metabolism. *Nature Genetics,* 46**,** 714-721.
CHEN, W., WANG, W., PENG, M., GONG, L., GAO, Y., WAN, J., WANG, S., SHI, L., ZHOU, B. & LI, Z. 2016. Comparative and parallel genome-wide association studies for metabolic and agronomic traits in cereals. *Nature communications,* 7**,** 1-10.
CHEZEM, W. R. & CLAY, N. K. 2016. Regulation of plant secondary metabolism and associated specialized cell development by *MYBs* and *bHLHs*. *Phytochemistry,* 131**,** 26-43.
CHITISANKUL, W. T., SHIMADA, K., OMIZU, Y., UEMOTO, Y., VARANYANOND, W. & TSUKAMOTO, C. 2015. Mechanism of DDMP-saponin degradation and maltol production in soymilk preparation. *LWT-Food Science and Technology,* 64**,** 197-204.
CHU, H. Y., WEGEL, E. & OSBOURN, A. 2011. From hormones to secondary metabolism: the emergence of metabolic gene clusters in plants. *The Plant Journal,* 66**,** 66-79.
COLINAS, M. & GOOSSENS, A. 2018. Combinatorial transcriptional control of plant specialized metabolism. *Trends in plant science,* 23**,** 324-336.
DAKORA, F. & PHILLIPS, D. 1996. Diverse functions of isoflavonoids in legumes transcend anti-microbial definitions of phytoalexins. *Physiological and Molecular Plant Pathology,* 49**,** 1-20.
DAVEY, J. W., HOHENLOHE, P. A., ETTER, P. D., BOONE, J. Q., CATCHEN, J. M. & BLAXTER, M. L. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics,* 12**,** 499-510.
DE COSTA, F., YENDO, A. C. A., FLECK, J. D., GOSMANN, G. & FETT-NETO, A. G. 2013. Accumulation of a bioactive triterpene saponin fraction of *Quillaja brasiliensis* leaves is associated with abiotic and biotic stresses. *Plant Physiology and Biochemistry,* 66**,** 56-62.
DIXON, R. A. 2001. Natural products and plant disease resistance. *Nature,* 411**,** 843-847.
DIXON, R. A. 2003. Phytochemistry meets genome analysis, and beyond. *Phytochemistry,* 62**,** 815-816.
DIXON, R. A. & PASINETTI, G. M. 2010. Flavonoids and isoflavonoids: from plant biology to agriculture and neuroscience. *Plant Physiology,* 154**,** 453-457.
DIXON, R. A. & STEELE, C. L. 1999. Flavonoids and isoflavonoids–a gold mine for metabolic engineering. *Trends in plant science,* 4**,** 394-400.
DONG, X., CHEN, W., WANG, W., ZHANG, H., LIU, X. & LUO, J. 2014. Comprehensive profiling and natural variation of flavonoids in rice. *Journal of integrative plant biology,* 56**,** 876-886.
DONG, X., GAO, Y., CHEN, W., WANG, W., GONG, L., LIU, X. & LUO, J. 2015. Spatiotemporal distribution of phenolamides and the genetics of natural variation of hydroxycinnamoyl spermidine in rice. *Molecular Plant,* 8**,** 111-121.
DONNEZ, D., KIM, K.-H., ANTOINE, S., CONREUX, A., DE LUCA, V., JEANDET, P., CLÉMENT, C. & COUROT, E. 2011. Bioproduction of resveratrol and viniferins by an elicited grapevine cell culture in a 2 L stirred bioreactor. *Process Biochemistry,* 46**,** 1056-1062.

EBEL, J. & GRISEBACH, H. 1988. Defense strategies of soybean against the fungus Phytophthora megasperma f. sp. glycinea: a molecular analysis. *Trends in biochemical sciences,* 13**,** 23-27.

ELSHIRE, R. J., GLAUBITZ, J. C., SUN, Q., POLAND, J. A., KAWAMOTO, K., BUCKLER, E. S. & MITCHELL, S. E. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one,* 6**,** e19379.

ENDARA, M.-J., COLEY, P. D., GHABASH, G., NICHOLLS, J. A., DEXTER, K. G., DONOSO, D. A., STONE, G. N., PENNINGTON, R. T. & KURSAR, T. A. 2017. Coevolutionary arms race versus host defense chase in a tropical herbivore–plant system. *Proceedings of the National Academy of Sciences,* 114**,** E7499-E7505.

ENDARA, M.-J., FORRISTER, D., NICHOLLS, J., STONE, G. N., KURSAR, T. & COLEY, P. 2022a. Impacts of plant defenses on host choice by Lepidoptera in Neotropical rainforests. *Caterpillars in the Middle: Tritrophic Interactions in a Changing World.* Springer.

ENDARA, M. J., SOULE, A. J., FORRISTER, D. L., DEXTER, K. G., PENNINGTON, R. T., NICHOLLS, J. A., LOISEAU, O., KURSAR, T. A. & COLEY, P. D. 2022b. The role of plant secondary metabolites in shaping regional and local plant community assembly. *Journal of Ecology,* 110**,** 34-45.

ERB, M. & KLIEBENSTEIN, D. J. 2020. Plant Secondary Metabolites as Defenses, Regulators, and Primary Metabolites: The Blurred Functional Trichotomy. *Plant Physiology,* 184**,** 39-52.

EVANS, L. T. 1996. *Crop evolution, adaptation and yield*, Cambridge university press.

FABBRI, A. D. & CROSBY, G. A. 2016. A review of the impact of preparation and cooking on the nutritional quality of vegetables and legumes. *International Journal of Gastronomy and Food Science,* 3**,** 2-11.

FAN, P., WANG, P., LOU, Y.-R., LEONG, B. J., MOORE, B. M., SCHENCK, C. A., COMBS, R., CAO, P., BRANDIZZI, F. & SHIU, S.-H. 2020. Evolution of a plant gene cluster in Solanaceae and emergence of metabolic diversity. *Elife,* 9**,** e56717.

FANG, C., FERNIE, A. R. & LUO, J. 2019. Exploring the diversity of plant metabolism. *Trends in Plant Science,* 24**,** 83-98.

FANG, C., ZHANG, H., WAN, J., WU, Y., LI, K., JIN, C., CHEN, W., WANG, S., WANG, W. & ZHANG, H. 2016. Control of leaf senescence by an MeOH-jasmonates cascade that is epigenetically regulated by *OsSRT1* in rice. *Molecular Plant,* 9**,** 1366-1378.

FARMAN, M. L. 2007. Telomeres in the rice blast fungus *Magnaporthe oryzae*: the world of the end as we know it. *FEMS microbiology letters,* 273**,** 125-132.

FELL, D. A. & WAGNER, A. 2000. The small world of metabolism. *Nature biotechnology,* 18**,** 1121-1122.

FERNIE, A. R. & PICHERSKY, E. 2015. Focus issue on metabolism: Metabolites, metabolites everywhere. American Society of Plant Biologists.

FERNIE, A. R. & TOHGE, T. 2017. The genetics of plant metabolism. *Annual Review of Genetics,* 51**,** 287-310.

FIELD, B., FISTON-LAVIER, A.-S., KEMEN, A., GEISLER, K., QUESNEVILLE, H. & OSBOURN, A. E. 2011. Formation of plant metabolic gene clusters within

dynamic chromosomal regions. *Proceedings of the National Academy of Sciences,* 108**,** 16116-16121.
FIELD, B. & OSBOURN, A. E. 2008. Metabolic diversification—independent assembly of operon-like gene clusters in different plants. *Science,* 320**,** 543-547.
FIRN, R. D. & JONES, C. G. 2000. The evolution of secondary metabolism–a unifying model. *Molecular microbiology,* 37**,** 989-994.
FLINT-GARCIA, S. A., THORNSBERRY, J. M. & BUCKLER IV, E. S. 2003. Structure of linkage disequilibrium in plants. *Annual review of plant biology,* 54**,** 357-374.
FORRISTER, D. L., ENDARA, M.-J., YOUNKIN, G. C., COLEY, P. D. & KURSAR, T. A. 2019. Herbivores as drivers of negative density dependence in tropical forest saplings. *Science,* 363**,** 1213-1216.
FORRISTER, D. L., ENDARA, M. J., SOULE, A. J., YOUNKIN, G. C., MILLS, A. G., LOKVAM, J., DEXTER, K. G., PENNINGTON, R. T., KIDNER, C. A. & NICHOLLS, J. A. 2023. Diversity and divergence: evolution of secondary metabolism in the tropical tree genus Inga. *New Phytologist,* 237**,** 631-642.
FREY, M., CHOMET, P., GLAWISCHNIG, E., STETTNER, C., GRUN, S., WINKLMAIR, A., EISENREICH, W., BACHER, A., MEELEY, R. B. & BRIGGS, S. P. 1997. Analysis of a chemical plant defense mechanism in grasses. *Science,* 277**,** 696-699.
FUJIMATSU, T., ENDO, K., YAZAKI, K. & SUGIYAMA, A. 2020. Secretion dynamics of soyasaponins in soybean roots and effects to modify the bacterial composition. *Plant direct,* 4**,** e00259-e00259.
FUTUYMA, D. J. & AGRAWAL, A. A. 2009. Macroevolution and the biological diversity of plants and herbivores. *Proceedings of the National Academy of Sciences,* 106**,** 18054-18061.
GAO, X., STARMER, J. & MARTIN, E. R. 2008. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society,* 32**,** 361-369.
GE, S. X., JUNG, D. & YAO, R. 2020. ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics,* 36**,** 2628-2629.
GIERL, A. & FREY, M. 2001. Evolution of benzoxazinone biosynthesis and indole production in maize. *Planta,* 213**,** 493-498.
GOETTEL, W., ZHANG, H., LI, Y., QIAO, Z., JIANG, H., HOU, D., SONG, Q., PANTALONE, V. R., SONG, B.-H., YU, D. & AN, Y.-Q. C. 2022. *POWR1* is a domestication gene pleiotropically regulating seed quality and yield in soybean. *Nature Communications,* 13**,** 3051.
GOODSTEIN, D. M., SHU, S., HOWSON, R., NEUPANE, R., HAYES, R. D., FAZO, J., MITROS, T., DIRKS, W., HELLSTEN, U., PUTNAM, N. & ROKHSAR, D. S. 2011. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research,* 40**,** D1178-D1186.
GORIM, L. Y. & VANDENBERG, A. 2017. Evaluation of wild lentil species as genetic resources to improve drought tolerance in cultivated lentil. *Frontiers in plant science,* 8**,** 1129.

GRAHAM, T. & GRAHAM, M. 1991. Glyceollin elicitors induce major but distinctly different shifts in isoflavonoid metabolism in proximal and distal soybean cell populations. *Molecular plant-microbe interactions: MPMI (USA)*.

GRAHAM, T. L. & GRAHAM, M. Y. 1996. Signaling in soybean phenylpropanoid responses (dissection of primary, secondary, and conditioning effects of light, wounding, and elicitor treatments). *Plant Physiology,* 110**,** 1123-1133.

GRANT, D., NELSON, R. T., CANNON, S. B. & SHOEMAKER, R. C. 2010. SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res,* 38**,** D843-6.

GREEN, M. R., HUGHES, H., SAMBROOK, J. & MACCALLUM, P. 2012. Molecular cloning: a laboratory manual. *Molecular cloning: a laboratory manual.*

GRELA, E. R., SAMOLIŃSKA, W., KICZOROWSKA, B., KLEBANIUK, R. & KICZOROWSKI, P. 2017. Content of minerals and fatty acids and their correlation with phytochemical compounds and antioxidant activity of leguminous seeds. *Biological Trace Element Research,* 180**,** 338-348.

GUO, A. C., JEWISON, T., WILSON, M., LIU, Y., KNOX, C., DJOUMBOU, Y., LO, P., MANDAL, R., KRISHNAMURTHY, R. & WISHART, D. S. 2012. ECMDB: the E. coli Metabolome Database. *Nucleic acids research,* 41**,** D625-D630.

HACQUARD, S., GARRIDO-OTER, R., GONZÁLEZ, A., SPAEPEN, S., ACKERMANN, G., LEBEIS, S., MCHARDY, A. C., DANGL, J. L., KNIGHT, R. & LEY, R. 2015. Microbiota and host nutrition across plant and animal kingdoms. *Cell host & microbe,* 17**,** 603-616.

HAN, J., GUO, B., GUO, Y., ZHANG, B., WANG, X. & QIU, L.-J. 2019. Creation of early flowering germplasm of soybean by CRISPR/Cas9 technology. *Frontiers in Plant science,* 10**,** 1446.

HAN, R.-M., TIAN, Y.-X., LIU, Y., CHEN, C.-H., AI, X.-C., ZHANG, J.-P. & SKIBSTED, L. H. 2009. Comparison of flavonoids and isoflavonoids as antioxidants. *Journal of Agricultural and Food Chemistry,* 57**,** 3780-3785.

HARJES, C. E., ROCHEFORD, T. R., BAI, L., BRUTNELL, T. P., KANDIANIS, C. B., SOWINSKI, S. G., STAPLETON, A. E., VALLABHANENI, R., WILLIAMS, M. & WURTZEL, E. T. 2008. Natural genetic variation in lycopene epsilon cyclase tapped for maize biofortification. *Science,* 319**,** 330-333.

HARTMANN, T. 2008. The lost origin of chemical ecology in the late 19th century. *Proceedings of the National Academy of Sciences,* 105**,** 4541-4546.

HAWKINS, C., GINZBURG, D., ZHAO, K., DWYER, W., XUE, B., XU, A., RICE, S., COLE, B., PALEY, S. & KARP, P. 2021. Plant Metabolic Network 15: A resource of genome-wide metabolism databases for 126 plants and algae. *Journal of integrative plant biology,* 63**,** 1888-1905.

HEGEMAN, A. D. 2010. Plant metabolomics—meeting the analytical challenges of comprehensive metabolite analysis. *Briefings in Functional Genomics,* 9**,** 139-148.

HILL, R. A. & CONNOLLY, J. D. 2013. Triterpenoids. *Natural product reports,* 30**,** 1028-1065.

HUANG, D., WANG, X., TANG, Z., YUAN, Y., XU, Y., HE, J., JIANG, X., PENG, S.-A., LI, L. & BUTELLI, E. 2018. Subfunctionalization of the *Ruby2–Ruby1* gene cluster during the domestication of citrus. *Nature plants,* 4**,** 930-941.

HUANG, X. & HAN, B. 2014. Natural variations and genome-wide association studies in crop plants. *Annual review of plant biology,* 65**,** 531-551.
IBRAHEEM, F., GAFFOOR, I., TAN, Q., SHYU, C.-R. & CHOPRA, S. 2015. A sorghum *MYB* transcription factor induces 3-deoxyanthocyanidins and enhances resistance against leaf blights in maize. *Molecules,* 20**,** 2388-2404.
ITKIN, M., HEINIG, U., TZFADIA, O., BHIDE, A., SHINDE, B., CARDENAS, P., BOCOBZA, S., UNGER, T., MALITSKY, S. & FINKERS, R. 2013. Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. *Science,* 341**,** 175-179.
JACOBY, R., PEUKERT, M., SUCCURRO, A., KOPRIVOVA, A. & KOPRIVA, S. 2017. The role of soil microorganisms in plant mineral nutrition—current knowledge and future directions. *Frontiers in plant science,* 8**,** 1617.
JAHAN, M. A., HARRIS, B., LOWERY, M., COBURN, K., INFANTE, A. M., PERCIFIELD, R. J., AMMER, A. G. & KOVINICH, N. 2019. The *NAC* family transcription factor *GmNAC42–1* regulates biosynthesis of the anticancer and neuroprotective glyceollins in soybean. *BMC Genomics,* 20**,** 149.
JAHAN, M. A., HARRIS, B., LOWERY, M., INFANTE, A. M., PERCIFIELD, R. J. & KOVINICH, N. 2020. Glyceollin transcription factor *GmMYB29A2* regulates soybean resistance to Phytophthora sojae. *Plant physiology,* 183**,** 530-546.
JEANDET, P., DOUILLET-BREUIL, A.-C., BESSIS, R., DEBORD, S., SBAGHI, M. & ADRIAN, M. 2002. Phytoalexins from the Vitaceae: biosynthesis, phytoalexin gene expression in transgenic plants, antifungal activity, and metabolism. *Journal of Agricultural and food chemistry,* 50**,** 2731-2741.
JEANDET, P., SOBARZO-SÁNCHEZ, E., SILVA, A. S., CLÉMENT, C., NABAVI, S. F., BATTINO, M., RASEKHIAN, M., BELWAL, T., HABTEMARIAM, S. & KOFFAS, M. 2020. Whole-cell biocatalytic, enzymatic and green chemistry methods for the production of resveratrol and its derivatives. *Biotechnology advances,* 39**,** 107461.
JEONG, H., TOMBOR, B., ALBERT, R., OLTVAI, Z. N. & BARABÁSI, A.-L. 2000. The large-scale organization of metabolic networks. *Nature,* 407**,** 651-654.
JING, C., YUAN, Y., TANG, Q., ZOU, P., LI, Y. & ZHANG, C. 2017. Extraction optimization, preliminary characterization and antioxidant activities of polysaccharides from Glycine soja. *International journal of biological macromolecules,* 103**,** 1207-1216.
JONCZYK, R., SCHMIDT, H., OSTERRIEDER, A., FIESSELMANN, A., SCHULLEHNER, K., HASLBECK, M., SICKER, D., HOFMANN, D., YALPANI, N. & SIMMONS, C. 2008. Elucidation of the final reactions of DIMBOA-glucoside biosynthesis in maize: characterization of *Bx6* and *Bx7*. *Plant physiology,* 146**,** 1053-1063.
KAJIKAWA, M., SIERRO, N., KAWAGUCHI, H., BAKAHER, N., IVANOV, N. V., HASHIMOTO, T. & SHOJI, T. 2017. Genomic insights into the evolution of the nicotine biosynthesis pathway in tobacco. *Plant Physiology,* 174**,** 999-1011.
KANG, J., BADGER, T. M., RONIS, M. J. & WU, X. 2010. Non-isoflavone phytochemicals in soy and their health effects. *Journal of agricultural and food chemistry,* 58**,** 8119-8133.

KATZ, E., LI, J.-J., JAEGLE, B., ASHKENAZY, H., ABRAHAMS, S. R., BAGAZA, C., HOLDEN, S., PIRES, C. J., ANGELOVICI, R. & KLIEBENSTEIN, D. J. 2021. Genetic variation, environment and demography intersect to shape *Arabidopsis* defense metabolite variation across Europe. *Elife,* 10**,** e67784.

KEEN, N., INGHAM, J., HYMOWLTZ, T., SIMS, J. & MIDLAND, S. 1989. The occurrence of glyceollins in plants related to Glycine max (L.) Merr. *Biochemical systematics and ecology,* 17**,** 395-398.

KERWIN, R., FEUSIER, J., CORWIN, J., RUBIN, M., LIN, C., MUOK, A., LARSON, B., LI, B., JOSEPH, B. & FRANCISCO, M. 2015. Natural genetic variation in *Arabidopsis thaliana* defense metabolism genes modulates field fitness. *Elife,* 4.

KERWIN, R. E., FEUSIER, J., MUOK, A., LIN, C., LARSON, B., COPELAND, D., CORWIN, J. A., RUBIN, M. J., FRANCISCO, M. & LI, B. 2017. Epistasis× environment interactions among *Arabidopsis thaliana* glucosinolate genes impact complex traits and fitness in the field. *new phytologist,* 215**,** 1249-1263.

KIM, H. J., LIM, J.-S., KIM, W.-K. & KIM, J.-S. 2012. Soyabean glyceollins: biological effects and relevance to human health. *Proceedings of the Nutrition Society,* 71**,** 166-174.

KIM, J. & BUELL, C. R. 2015. A revolution in plant metabolism: genome-enabled pathway discovery. *Plant Physiology,* 169**,** 1532-1539.

KLEESSEN, S., ANTONIO, C., SULPICE, R., LAITINEN, R., FERNIE, A. R., STITT, M. & NIKOLOSKI, Z. 2012. Structured patterns in geographic variability of metabolic phenotypes in *Arabidopsis thaliana*. *Nature Communications,* 3**,** 1-7.

KLIEBENSTEIN, D. 2001. Gene duplication and the diversification of secondary metabolism: side chain modification of glucosinolates in *Arabidopsis thaliana*. *Plant cell,* 13**,** 681-693.

KLIEBENSTEIN, D. J. 2007a. Metabolomics and Plant Quantitative Trait Locus Analysis–The optimum genetical genomics platform? *Concepts in plant metabolomics.* Springer.

KLIEBENSTEIN, D. J. Metabolomics and Plant Quantitative Trait Locus Analysis–The optimum genetical genomics platform? Concepts in plant metabolomics, 2007b. Springer, 29-44.

KLIEBENSTEIN, D. J., GERSHENZON, J. & MITCHELL-OLDS, T. 2001. Comparative quantitative trait loci mapping of aliphatic, indolic and benzylic glucosinolate production in *Arabidopsis thaliana* leaves and seeds. *Genetics,* 159**,** 359-370.

KOFSKY, J., ZHANG, H. & SONG, B.-H. 2018. The untapped genetic reservoir: the past, current, and future applications of the wild soybean (Glycine soja). *Frontiers in Plant Science,* 9**,** 949.

KOLLNER, T. G., SCHNEE, C., GERSHENZON, J. & DEGENHARDT, J. 2004. The variability of sesquiterpenes emitted from two Zea mays cultivars is controlled by allelic variation of two terpene synthase genes encoding stereoselective multiple product enzymes. *The Plant Cell,* 16**,** 1115-1131.

KOONIN, E. V. 2009. Evolution of genome architecture. *The international journal of biochemistry & cell biology,* 41**,** 298-306.

KORTE, A. & FARLOW, A. 2013. The advantages and limitations of trait analysis with GWAS: a review. *Plant methods,* 9**,** 1-9.

KÖSTER, J. & BARZ, W. 1981. UDP-glucose: isoflavone 7-O-glucosyltransferase from roots of chick pea *(Cicer arietinum L.)*. *Archives of Biochemistry and Biophysics,* 212**,** 98-104.
KROYMANN, J. 2011. Natural diversity and adaptation in plant secondary metabolism. *Current opinion in plant biology,* 14**,** 246-251.
KUMAGAI, M. H., DONSON, J., DELLA-CIOPPA, G., HARVEY, D., HANLEY, K. & GRILL, L. 1995. Cytoplasmic inhibition of carotenoid biosynthesis with virus-derived RNA. *Proceedings of the National Academy of Sciences,* 92**,** 1679-1683.
KUROSAWA, Y., TAKAHARA, H. & SHIRAIWA, M. 2002. UDP-glucuronic acid:soyasapogenol glucuronosyltransferase involved in saponin biosynthesis in germinating soybean seeds. *Planta,* 215**,** 620-629.
KURSAR, T. A., DEXTER, K. G., LOKVAM, J., PENNINGTON, R. T., RICHARDSON, J. E., WEBER, M. G., MURAKAMI, E. T., DRAKE, C., MCGREGOR, R. & COLEY, P. D. 2009. The evolution of antiherbivore defenses and their contribution to species coexistence in the tropical tree genus Inga. *Proceedings of the National Academy of Sciences,* 106**,** 18073-18078.
LADIZINSKY, G. 2012. *Plant evolution under domestication*, Springer Science & Business Media.
LAIRSON, L., HENRISSAT, B., DAVIES, G. & WITHERS, S. 2008. Glycosyltransferases: structures, functions, and mechanisms. *Annu. Rev. Biochem.,* 77**,** 521-555.
LEAMY, L. J., ZHANG, H., LI, C., CHEN, C. Y. & SONG, B.-H. 2017. A genome-wide association study of seed composition traits in wild soybean *(Glycine soja)*. *BMC genomics,* 18**,** 1-15.
LECOMTE, S., CHALMEL, F., FERRIERE, F., PERCEVAULT, F., PLU, N., SALIGAUT, C., SUREL, C., LELONG, M., EFSTATHIOU, T. & PAKDEL, F. 2017. Glyceollins trigger anti-proliferative effects through estradiol-dependent and independent pathways in breast cancer cells. *Cell Commun Signal,* 15**,** 26.
LEE, I.-A., PARK, Y.-J., YEO, H.-K., HAN, M. J. & KIM, D.-H. 2010. Soyasaponin I attenuates TNBS-induced colitis in mice by inhibiting NF-κB pathway. *Journal of agricultural and food chemistry,* 58**,** 10929-10934.
LI, D., BALDWIN, I. T. & GAQUEREL, E. 2015. Navigating natural variation in herbivory-induced secondary metabolism in coyote tobacco populations using MS/MS structural analysis. *Proceedings of the National Academy of Sciences,* 112**,** E4147-E4155.
LI, J. & JI, L. 2005. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity,* 95**,** 221-227.
LI, Y.-H., ZHAO, S.-C., MA, J.-X., LI, D., YAN, L., LI, J., QI, X.-T., GUO, X.-S., ZHANG, L. & HE, W.-M. J. B. G. 2013. Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. 14**,** 1-12.
LIU, C.-J., BLOUNT, J. W., STEELE, C. L. & DIXON, R. A. 2002. Bottlenecks for metabolic engineering of isoflavone glycoconjugates in Arabidopsis. *Proceedings of the National Academy of Sciences,* 99**,** 14578-14583.
LOZOVAYA, V. V., LYGIN, A. V., ZERNOVA, O. V., ULANOV, A. V., LI, S., HARTMAN, G. L. & WIDHOLM, J. M. 2007. Modification of phenolic

metabolism in soybean hairy roots through down regulation of chalcone synthase or isoflavone synthase. *Planta,* 225**,** 665-679.
LU, X., LI, Q. T., XIONG, Q., LI, W., BI, Y. D., LAI, Y. C., LIU, X. L., MAN, W. Q., ZHANG, W. K. & MA, B. 2016. The transcriptomic signature of developing soybean seeds reveals the genetic basis of seed trait adaptation during domestication. *The Plant Journal,* 86**,** 530-544.
LUO, J. 2015. Metabolite-based genome-wide association studies in plants. *Current opinion in plant biology,* 24**,** 31-38.
LUO, W., PANT, G., BHAVNASI, Y. K., BLANCHARD JR, S. G. & BROUWER, C. 2017. Pathview Web: user friendly pathway visualization and data integration. *Nucleic acids research,* 45**,** W501-W508.
MARONE, D., MASTRANGELO, A. M., BORRELLI, G. M., MORES, A., LAIDÒ, G., RUSSO, M. A. & FICCO, D. B. M. 2022. Specialized metabolites: Physiological and biochemical role in stress resistance, strategies to improve their accumulation, and new applications in crop breeding and management. *Plant Physiology and Biochemistry,* 172**,** 48-55.
MATSUBA, Y., NGUYEN, T. T., WIEGERT, K., FALARA, V., GONZALES-VIGIL, E., LEONG, B., SCHÄFER, P., KUDRNA, D., WING, R. A. & BOLGER, A. M. 2013. Evolution of a complex locus for terpene biosynthesis in *Solanum*. *The Plant Cell,* 25**,** 2022-2036.
MATSUDA, F., NAKABAYASHI, R., YANG, Z., OKAZAKI, Y., YONEMARU, J. I., EBANA, K., YANO, M. & SAITO, K. 2015. Metabolome-genome-wide association study dissects genetic architecture for generating natural variation in rice secondary metabolism. *The Plant Journal,* 81**,** 13-23.
MCMULLEN, M., BYRNE, P., SNOOK, M., WISEMAN, B., LEE, E., WIDSTROM, N. & COE, E. 1998. Quantitative trait loci and metabolic pathways. *Proceedings of the National Academy of Sciences,* 95**,** 1996-2000.
MEDEMA, M. H., DE ROND, T. & MOORE, B. S. 2021. Mining genomes to illuminate the specialized chemistry of life. *Nature Reviews Genetics,* 22**,** 553-571.
MEYER, R. S., DUVAL, A. E. & JENSEN, H. R. 2012. Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. *New Phytologist,* 196**,** 29-48.
MIYAMOTO, K., FUJITA, M., SHENTON, M. R., AKASHI, S., SUGAWARA, C., SAKAI, A., HORIE, K., HASEGAWA, M., KAWAIDE, H. & MITSUHASHI, W. 2016. Evolutionary trajectory of phytoalexin biosynthetic gene clusters in rice. *The plant journal,* 87**,** 293-304.
MOORE, B. D., ANDREW, R. L., KÜLHEIM, C. & FOLEY, W. J. 2014. Explaining intraspecific diversity in plant secondary metabolites in an ecological context. *New Phytologist,* 201**,** 733-750.
MOORE, B. M., WANG, P., FAN, P., LEONG, B., SCHENCK, C. A., LLOYD, J. P., LEHTI-SHIU, M. D., LAST, R. L., PICHERSKY, E. & SHIU, S.-H. 2019. Robust predictions of specialized metabolism genes through machine learning. *Proceedings of the National Academy of Sciences,* 116**,** 2344-2353.
MORRIS, P. F., BONE, E. & TYLER, B. M. 1998. Chemotropic and contact responses of Phytophthora sojae hyphae to soybean isoflavonoids and artificial substrates. *Plant physiology,* 117**,** 1171-1178.

MORRISSEY, J. P. & OSBOURN, A. E. 1999. Fungal resistance to plant antibiotics as a mechanism of pathogenesis. *Microbiology and Molecular Biology Reviews,* 63**,** 708-724.
MOSES, T., PAPADOPOULOU, K. K. & OSBOURN, A. 2014. Metabolic and functional diversity of saponins, biosynthetic intermediates and semi-synthetic derivatives. *Critical reviews in biochemistry and molecular biology,* 49**,** 439-462.
MUGFORD, S. T., QI, X., BAKHT, S., HILL, L., WEGEL, E., HUGHES, R. K., PAPADOPOULOU, K., MELTON, R., PHILO, M. & SAINSBURY, F. 2009. A serine carboxypeptidase-like acyltransferase is required for synthesis of antimicrobial compounds and disease resistance in oats. *The Plant Cell,* 21**,** 2473-2484.
MUNAKATA, R. 2021. Pulses without the Characteristic Distasteful Bitterness: Pea TILLING Lines Lacking the Major β-Amyrin Synthase in Soyasaponin Biosynthesis. *Plant and Cell Physiology,* 62**,** 749-751.
MYLONA, P., OWATWORAKIT, A., PAPADOPOULOU, K., JENNER, H., QIN, B., FINDLAY, K., HILL, L., QI, X., BAKHT, S. & MELTON, R. 2008. *Sad3* and *Sad4* are required for saponin biosynthesis and root development in oat. *The Plant Cell,* 20**,** 201-212.
NEGRE, F., KISH, C. M., BOATRIGHT, J., UNDERWOOD, B., SHIBUYA, K., WAGNER, C., CLARK, D. G. & DUDAREVA, N. 2003. Regulation of methylbenzoate emission after pollination in snapdragon and petunia flowers. *The Plant Cell,* 15**,** 2992-3006.
NIEUWENHUIZEN, N. J., CHEN, X., WANG, M. Y., MATICH, A. J., PEREZ, R. L., ALLAN, A. C., GREEN, S. A. & ATKINSON, R. G. 2015. Natural variation in monoterpene synthesis in kiwifruit: transcriptional regulation of terpene synthases by NAC and ETHYLENE-INSENSITIVE3-like transcription factors. *Plant Physiology,* 167**,** 1243-1258.
NÜTZMANN, H. W., HUANG, A. & OSBOURN, A. 2016. Plant metabolic clusters–from genetics to genomics. *New phytologist,* 211**,** 771-789.
NWACHUKWU, I. D., LUCIANO, F. B. & UDENIGWE, C. C. 2013. The inducible soybean glyceollin phytoalexins with multifunctional health-promoting properties. *Food research international,* 54**,** 1208-1216.
O'MAILLE, P. E., MALONE, A., DELLAS, N., ANDES HESS JR, B., SMENTEK, L., SHEEHAN, I., GREENHAGEN, B. T., CHAPPELL, J., MANNING, G. & NOEL, J. P. 2008. Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases. *Nature chemical biology,* 4**,** 617-623.
OBALA, J., SAXENA, R. K., SINGH, V. K., KUMAR, C. S., SAXENA, K., TONGOONA, P., SIBIYA, J. & VARSHNEY, R. K. 2019. Development of sequence-based markers for seed protein content in pigeonpea. *Molecular Genetics and Genomics,* 294**,** 57-68.
OGAWA, S., MIYAMOTO, K., NEMOTO, K., SAWASAKI, T., YAMANE, H., NOJIRI, H. & OKADA, K. 2017. *OsMYC2*, an essential factor for JA-inductive sakuranetin production in rice, interacts with MYC2-like proteins that enhance its transactivation ability. *Scientific reports,* 7**,** 1-11.
OH, S.-R., KINJO, J., SHII, Y., IKEDA, T., NOHARA, T., AHN, K. S., KIM, J. H. & LEE, H.-K. 2000. Effects of triterpenoids from Pueraria lobata on

immunohemolysis: β-D-glucuronic acid plays an active role in anticomplementary activity *in vitro*. *Planta Medica,* 66**,** 506-510.
OSBOURN, A. 2010a. Gene clusters for secondary metabolic pathways: an emerging theme in plant biology. *Plant physiology,* 154**,** 531-535.
OSBOURN, A. 2010b. Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation. *Trends in Genetics,* 26**,** 449-457.
OSBOURN, A., GOSS, R. J. & FIELD, R. A. 2011. The saponins–polar isoprenoids with important and diverse biological activities. *Natural product reports,* 28**,** 1261-1268.
OSBOURN, A. E. & FIELD, B. 2009. Operons. *Cellular and Molecular Life Sciences,* 66**,** 3755-3775.
OURISSON, G. & ALBRECHT, P. 1992. Hopanoids. 1. Geohopanoids: the most abundant natural products on Earth? *Accounts of Chemical Research,* 25**,** 398-402.
OZA, M. J. & KULKARNI, Y. A. 2018. Formononetin treatment in type 2 diabetic rats reduces insulin resistance and hyperglycemia. *Frontiers in pharmacology,* 9**,** 739.
PAIS, A. L., LI, X. & XIANG, Q. Y. 2018. Discovering variation of secondary metabolite diversity and its relationship with disease resistance in Cornus florida L. *Ecology and Evolution,* 8**,** 5619-5636.
PATIL, G., VUONG, T. D., KALE, S., VALLIYODAN, B., DESHMUKH, R., ZHU, C., WU, X., BAI, Y., YUNGBLUTH, D. & LU, F. 2018. Dissecting genomic hotspots underlying seed protein, oil, and sucrose content in an interspecific mapping population of soybean using high-density linkage mapping. *Plant Biotechnology Journal,* 16**,** 1939-1953.
PAXTON, J. 1975. Phytoalexins, phenolics, and other antibiotics in roots resistant to soil-borne fungi. *Phytoalexins, phenolics, and other antibiotics in roots resistant to soil-borne fungi.***,** 185-192.
PAZ, M. M., MARTINEZ, J. C., KALVIG, A. B., FONGER, T. M. & WANG, K. 2006. Improved cotyledonary node method using an alternative explant derived from mature seed for efficient Agrobacterium-mediated soybean transformation. *Plant Cell Reports,* 25**,** 206-213.
PENG, M., GAO, Y., CHEN, W., WANG, W., SHEN, S., SHI, J., WANG, C., ZHANG, Y., ZOU, L. & WANG, S. 2016. Evolutionarily distinct BAHD N-acyltransferases are responsible for natural variation of aromatic amine conjugates in rice. *The Plant Cell,* 28**,** 1533-1550.
PENG, M., SHAHZAD, R., GUL, A., SUBTHAIN, H., SHEN, S., LEI, L., ZHENG, Z., ZHOU, J., LU, D. & WANG, S. 2017. Differentially evolved glucosyltransferases determine natural variation of rice flavone accumulation and UV-tolerance. *Nature communications,* 8**,** 1-12.
PETERSEN, A.-K., KRUMSIEK, J., WÄGELE, B., THEIS, F. J., WICHMANN, H.-E., GIEGER, C. & SUHRE, K. 2012. On the hypothesis-free testing of metabolite ratios in genome-wide and metabolome-wide association studies. *BMC bioinformatics,* 13**,** 1-7.
PHILLIPS, D. R., RASBERY, J. M., BARTEL, B. & MATSUDA, S. P. 2006. Biosynthetic diversity in plant triterpene cyclization. *Current opinion in plant biology,* 9**,** 305-314.

PICHERSKY, E. & GANG, D. R. 2000. Genetics and biochemistry of secondary metabolites in plants: an evolutionary perspective. *Trends in plant science,* 5**,** 439-445.
PICHERSKY, E., NOEL, J. P. & DUDAREVA, N. 2006. Biosynthesis of plant volatiles: nature's diversity and ingenuity. *Science,* 311**,** 808-811.
POLAK, R., PHILLIPS, E. M. & CAMPBELL, A. 2015. Legumes: health benefits and culinary approaches to increase intake. *Clinical Diabetes,* 33**,** 198-205.
POLLAK, P. E., VOGT, T., MO, Y. & TAYLOR, L. P. 1993. Chalcone synthase and flavonol accumulation in stigmas and anthers of Petunia hybrida. *Plant physiology,* 102**,** 925-932.
POLTURAK, G., DIPPE, M., STEPHENSON, M. J., CHANDRA MISRA, R., OWEN, C., RAMIREZ-GONZALEZ, R. H., HAIDOULIS, J. F., SCHOONBEEK, H.-J., CHARTRAIN, L. & BORRILL, P. 2022. Pathogen-induced biosynthetic pathways encode defense-related molecules in bread wheat. *Proceedings of the National Academy of Sciences,* 119**,** e2123299119.
POLTURAK, G. & OSBOURN, A. 2021. The emerging role of biosynthetic gene clusters in plant defense and plant interactions. *PLoS Pathogens,* 17**,** e1009698.
PRASAD, K. V., SONG, B.-H., OLSON-MANNING, C., ANDERSON, J. T., LEE, C.-R., SCHRANZ, M. E., WINDSOR, A. J., CLAUSS, M. J., MANZANEDA, A. J. & NAQVI, I. 2012. A gain-of-function polymorphism controlling complex traits and fitness in nature. *science,* 337**,** 1081-1084.
PRASANNA, B. 2012. Diversity in global maize germplasm: characterization and utilization. *Journal of biosciences,* 37**,** 843-855.
QI, X., BAKHT, S., LEGGETT, M., MAXWELL, C., MELTON, R. & OSBOURN, A. 2004. A gene cluster for secondary metabolism in oat: implications for the evolution of metabolic diversity in plants. *Proceedings of the National Academy of Sciences,* 101**,** 8233-8238.
QI, X., BAKHT, S., QIN, B., LEGGETT, M., HEMMINGS, A., MELLON, F., EAGLES, J., WERCK-REICHHART, D., SCHALLER, H. & LESOT, A. 2006. A different function for a member of an ancient and highly conserved cytochrome P450 family: from essential sterols to plant defense. *Proceedings of the National Academy of Sciences,* 103**,** 18848-18853.
QIU, G., TIAN, W., HUAN, M., CHEN, J. & FU, H. 2017. Formononetin exhibits anti-hyperglycemic activity in alloxan-induced type 1 diabetic mice. *Exp Biol Med (Maywood),* 242**,** 223-230.
RAHIMI, S., KIM, J., MIJAKOVIC, I., JUNG, K.-H., CHOI, G., KIM, S.-C. & KIM, Y.-J. 2019. Triterpenoid-biosynthetic UDP-glycosyltransferases from plants. *Biotechnology advances,* 37**,** 107394.
RAI, A., SAITO, K. & YAMAZAKI, M. 2017. Integrated omics analysis of specialized metabolism in medicinal plants. Wiley Online Library.
RAY, D. K., MUELLER, N. D., WEST, P. C. & FOLEY, J. A. 2013. Yield Trends Are Insufficient to Double Global Crop Production by 2050. *PLoS One,* 8**,** e66428.
REMINGTON, D. L., THORNSBERRY, J. M., MATSUOKA, Y., WILSON, L. M., WHITT, S. R., DOEBLEY, J., KRESOVICH, S., GOODMAN, M. M. & BUCKLER IV, E. S. 2001. Structure of linkage disequilibrium and phenotypic

associations in the maize genome. *Proceedings of the national academy of sciences,* 98**,** 11479-11484.
RICHARDS, L. A., DYER, L. A., FORISTER, M. L., SMILANICH, A. M., DODSON, C. D., LEONARD, M. D. & JEFFREY, C. S. 2015. Phytochemical diversity drives plant–insect community diversity. *Proceedings of the National Academy of Sciences,* 112**,** 10973-10978.
RIEDELSHEIMER, C., LISEC, J., CZEDIK-EYSENBERG, A., SULPICE, R., FLIS, A., GRIEDER, C., ALTMANN, T., STITT, M., WILLMITZER, L. & MELCHINGER, A. E. 2012. Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. *Proceedings of the National Academy of Sciences,* 109**,** 8872-8877.
RIPPERT, P. & MATRINGE, M. 2002. Molecular and biochemical characterization of an Arabidopsis thaliana arogenate dehydrogenase with two highly similar and active protein domains. *Plant molecular biology,* 48**,** 361-368.
ROBERTSON, D. 2004. VIGS vectors for gene silencing: many targets, many tools. *Annu. Rev. Plant Biol.,* 55**,** 495-519.
ROCHA, E. P. 2008. The organization of the bacterial genome. *Annual review of genetics,* 42**,** 211-233.
ROLAND, W. S., POUVREAU, L., CURRAN, J., VAN DE VELDE, F. & DE KOK, P. M. 2017. Flavor aspects of pulse ingredients. *Cereal Chemistry,* 94**,** 58-65.
ROWE, H. C., HANSEN, B. G., HALKIER, B. A. & KLIEBENSTEIN, D. J. 2008. Biochemical networks and epistasis shape the *Arabidopsis thaliana* metabolome. *The Plant Cell,* 20**,** 1199-1216.
SABETI, P. C., REICH, D. E., HIGGINS, J. M., LEVINE, H. Z., RICHTER, D. J., SCHAFFNER, S. F., GABRIEL, S. B., PLATKO, J. V., PATTERSON, N. J. & MCDONALD, G. J. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature,* 419**,** 832-837.
SACHS, J. 1874. *Lehrbuch der botanik,* Engelmann.
SAGA, H., OGAWA, T., KAI, K., SUZUKI, H., OGATA, Y., SAKURAI, N., SHIBATA, D. & OHTA, D. 2012. Identification and characterization of *ANAC042*, a transcription factor family gene involved in the regulation of camalexin biosynthesis in *Arabidopsis*. *Molecular plant-microbe interactions,* 25**,** 684-696.
SAITO, K. & MATSUDA, F. 2010. Metabolomics for functional genomics, systems biology, and biotechnology. *Annual review of plant biology,* 61**,** 463-489.
SAJED, T., MARCU, A., RAMIREZ, M., PON, A., GUO, A. C., KNOX, C., WILSON, M., GRANT, J. R., DJOUMBOU, Y. & WISHART, D. S. 2016. ECMDB 2.0: A richer resource for understanding the biochemistry of E. coli. *Nucleic acids research,* 44**,** D495-D501.
SAKAMOTO, T., MIURA, K., ITOH, H., TATSUMI, T., UEGUCHI-TANAKA, M., ISHIYAMA, K., KOBAYASHI, M., AGRAWAL, G. K., TAKEDA, S. & ABE, K. 2004. An overview of gibberellin metabolism enzyme genes and their related mutants in rice. *Plant physiology,* 134**,** 1642-1653.
SALAZAR, D., LOKVAM, J., MESONES, I., VÁSQUEZ PILCO, M., AYARZA ZUÑIGA, J. M., DE VALPINE, P. & FINE, P. V. 2018. Origin and maintenance

of chemical diversity in a species-rich tropical tree lineage. *Nature Ecology & Evolution,* 2**,** 983-990.
SANGER, F., NICKLEN, S. & COULSON, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America,* 74**,** 5463-5467.
SAYAMA, T., ONO, E., TAKAGI, K., TAKADA, Y., HORIKAWA, M., NAKAMOTO, Y., HIROSE, A., SASAMA, H., OHASHI, M., HASEGAWA, H., TERAKAWA, T., KIKUCHI, A., KATO, S., TATSUZAKI, N., TSUKAMOTO, C. & ISHIMOTO, M. 2012. The *Sg-1* glycosyltransferase locus regulates structural diversity of triterpenoid saponins of soybean. *The Plant cell,* 24**,** 2123-2138.
SCHAUER, N. & FERNIE, A. R. 2006. Plant metabolomics: towards biological function and mechanism. *Trends in Plant Science,* 11**,** 508-516.
SCHLÄPFER, P., ZHANG, P., WANG, C., KIM, T., BANF, M., CHAE, L., DREHER, K., CHAVALI, A. K., NILO-POYANCO, R. & BERNARD, T. 2017. Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. *Plant physiology,* 173**,** 2041-2059.
SEO, J. Y., KIM, B. R., OH, J. & KIM, J.-S. 2018. Soybean-derived phytoalexins improve cognitive function through activation of *Nrf2/HO-1* signaling pathway. *International journal of molecular sciences,* 19**,** 268.
SHARMA, V. & RAMAWAT, K. G. 2013. Isoflavonoids. *In:* RAMAWAT, K. G. & MÉRILLON, J.-M. (eds.) *Natural Products: Phytochemistry, Botany and Metabolism of Alkaloids, Phenolics and Terpenes.* Berlin, Heidelberg: Springer Berlin Heidelberg.
SHASTRY, B. S. 2009. SNPs: impact on gene function and phenotype. *Single nucleotide polymorphisms: Methods and protocols***,** 3-22.
SHIBUYA, M., HOSHINO, M., KATSUBE, Y., HAYASHI, H., KUSHIRO, T. & EBIZUKA, Y. 2006. Identification of β-amyrin and sophoradiol 24-hydroxylase by expressed sequence tag mining and functional expression assay. *The FEBS journal,* 273**,** 948-959.
SHIBUYA, M., NISHIMURA, K., YASUYAMA, N. & EBIZUKA, Y. 2010. Identification and characterization of glycosyltransferases involved in the biosynthesis of soyasaponin I in *Glycine max*. *FEBS letters,* 584**,** 2258-2264.
SHIN, J.-H., BLAY, S., MCNENEY, B. & GRAHAM, J. 2006. LDheatmap: an R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *Journal of statistical software,* 16**,** 1-10.
SHOJI, T. 2019. The recruitment model of metabolic evolution: jasmonate-responsive transcription factors and a conceptual model for the evolution of metabolic pathways. *Frontiers in Plant Science,* 10**,** 560.
SHOJI, T. & YUAN, L. 2021. *ERF* gene clusters: working together to regulate metabolism. *Trends in Plant Science,* 26**,** 23-32.
SINGH, R. J. & HYMOWITZ, T. 1999. Soybean genetic resources and crop improvement. *Genome,* 42**,** 605-616.
SINGH, S. K., PATRA, B., PAUL, P., LIU, Y., PATTANAIK, S. & YUAN, L. 2020. Revisiting the *ORCA* gene cluster that regulates terpenoid indole alkaloid biosynthesis in *Catharanthus roseus*. *Plant Science,* 293**,** 110408.

SKOGERSON, K., HARRIGAN, G. G., REYNOLDS, T. L., HALLS, S. C., RUEBELT, M., IANDOLINO, A., PANDRAVADA, A., GLENN, K. C. & FIEHN, O. 2010. Impact of genetics and environment on the metabolite composition of maize grain. *Journal of Agricultural and Food Chemistry,* 58**,** 3600-3610.

SMIT, S. J. & LICHMAN, B. R. 2022. Plant biosynthetic gene clusters in the context of metabolic evolution. *Natural Product Reports*.

SOLTIS, N. E. & KLIEBENSTEIN, D. J. 2015. Natural variation of plant metabolism: genetic mechanisms, interpretive caveats, and evolutionary and mechanistic insights. *Plant Physiology,* 169**,** 1456-1468.

SONG, Q., HYTEN, D. L., JIA, G., QUIGLEY, C. V., FICKUS, E. W., NELSON, R. L. & CREGAN, P. B. 2013. Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *PloS one,* 8**,** e54985.

SPARG, S., LIGHT, M. & VAN STADEN, J. 2004. Biological activities and distribution of plant saponins. *Journal of ethnopharmacology,* 94**,** 219-243.

SPRINGER, N., DE LEÓN, N. & GROTEWOLD, E. 2019. Challenges of translating gene regulatory information into agronomic improvements. *Trends in plant science,* 24**,** 1075-1082.

STRAUCH, R. C., SVEDIN, E., DILKES, B., CHAPPLE, C. & LI, X. 2015. Discovery of a novel amino acid racemase through exploration of natural variation in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences of the United States of America,* 112**,** 11726-11731.

SUBRAMANIAN, S., STACEY, G. & YU, O. 2006. Endogenous isoflavones are essential for the establishment of symbiosis between soybean and *Bradyrhizobium japonicum*. *The Plant Journal,* 48**,** 261-273.

SUE, M., NAKAMURA, C. & NOMURA, T. 2011. Dispersed benzoxazinone gene cluster: molecular characterization and chromosomal localization of glucosyltransferase and glucosidase genes in wheat and rye. *Plant physiology,* 157**,** 985-997.

SUHRE, K., SHIN, S.-Y., PETERSEN, A.-K., MOHNEY, R. P., MEREDITH, D., WÄGELE, B., ALTMAIER, E., DELOUKAS, P., ERDMANN, J. & GRUNDBERG, E. 2011. Human metabolic individuality in biomedical and pharmaceutical research. *Nature,* 477**,** 54-60.

SUKUMARAN, A., MCDOWELL, T., CHEN, L., RENAUD, J. & DHAUBHADEL, S. 2018. Isoflavonoid-specific prenyltransferase gene family in soybean: *GmPT01*, a pterocarpan 2-dimethylallyltransferase involved in glyceollin biosynthesis. *The Plant Journal,* 96**,** 966-981.

SUN, T., YAN, X., GUO, W. & ZHAO, D. 2014. Evaluation of cytotoxicity and immune modulatory activities of soyasaponin Ab: An *in vitro* and *in vivo* study. *Phytomedicine,* 21**,** 1759-1766.

SUN, Y., OH, D.-H., DUAN, L., RAMACHANDRAN, P., RAMIREZ, A., BARTLETT, A., TRAN, K.-N., WANG, G., DASSANAYAKE, M. & DINNENY, J. R. 2022. Divergence in the *ABA* gene regulatory network underlies differential growth control. *Nature Plants,* 8**,** 549-560.

SUNDARAMOORTHY, J., PARK, G. T., KOMAGAMINE, K., TSUKAMOTO, C., CHANG, J. H., LEE, J. D., KIM, J. H., SEO, H. S. & SONG, J. T. 2019.

Biosynthesis of DDMP saponins in soybean is regulated by a distinct UDP-glycosyltransferase. *New Phytologist,* 222**,** 261-274.
SUREDA, A., SILVA, A. S., SÁNCHEZ-MACHADO, D. I., LÓPEZ-CERVANTES, J., DAGLIA, M., NABAVI, S. F. & NABAVI, S. M. 2017. Hypotensive effects of genistein: From chemistry to medicine. *Chemico-biological interactions,* 268**,** 37-46.
SZPIECH, Z. A. & HERNANDEZ, R. D. 2014. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Molecular biology and evolution,* 31**,** 2824-2827.
TAKADA, Y., TAYAMA, I., SAYAMA, T., SASAMA, H., SARUTA, M., KIKUCHI, A., ISHIMOTO, M. & TSUKAMOTO, C. 2012. Genetic analysis of variations in the sugar chain composition at the C-3 position of soybean seed saponins. *Breeding science,* 61**,** 639-645.
TAKAHASHI, Y., LI, X.-H., TSUKAMOTO, C. & WANG, K.-J. 2017. Categories and components of soyasaponin in the Chinese wild soybean (Glycine soja) genetic resource collection. *Genetic Resources and Crop Evolution,* 64**,** 2161-2171.
TAKAHASHI, Y., LI, X. H., TSUKAMOTO, C. & WANG, K. J. 2016. Identification of a novel variant lacking group A soyasaponin in a Chinese wild soybean (Glycine soja Sieb. & Zucc.): implications for breeding significance. *Plant Breeding,* 135**,** 607-613.
TAKOS, A. M. & ROOK, F. 2012. Why biosynthetic genes for chemical defense compounds cluster. *Trends in plant science,* 17**,** 383-388.
TAMURA, K., YOSHIDA, K., HIRAOKA, Y., SAKAGUCHI, D., CHIKUGO, A., MOCHIDA, K., KOJOMA, M., MITSUDA, N., SAITO, K. & MURANAKA, T. 2018. The basic helix–loop–helix transcription factor *GubHLH3* positively regulates soyasaponin biosynthetic genes in *Glycyrrhiza uralensis*. *Plant and Cell Physiology,* 59**,** 783-796.
TANG, Y., LIU, X., WANG, J., LI, M., WANG, Q., TIAN, F., SU, Z., PAN, Y., LIU, D. & LIPKA, A. E. 2016. GAPIT version 2: an enhanced integrated tool for genomic association and prediction. *The plant genome,* 9**,** plantgenome2015.11.0120.
TANTRY, M. A. & KHAN, I. A. 2013. Saponins from *Glycine max Merrill* (soybean). *Fitoterapia,* 87**,** 49-56.
THAGUN, C., IMANISHI, S., KUDO, T., NAKABAYASHI, R., OHYAMA, K., MORI, T., KAWAMOTO, K., NAKAMURA, Y., KATAYAMA, M. & NONAKA, S. 2016. Jasmonate-responsive *ERF* transcription factors regulate steroidal glycoalkaloid biosynthesis in tomato. *Plant and Cell Physiology,* 57**,** 961-975.
THIMMAPPA, R., GEISLER, K., LOUVEAU, T., O'MAILLE, P. & OSBOURN, A. 2014. Triterpene biosynthesis in plants. *Annual review of plant biology,* 65**,** 225-257.
THOMPSON, J. N. 2019. *The geographic mosaic of coevolution*, University of Chicago Press.
TIEMAN, D., ZHU, G., RESENDE JR, M. F., LIN, T., NGUYEN, C., BIES, D., RAMBLA, J. L., BELTRAN, K. S. O., TAYLOR, M. & ZHANG, B. 2017. A chemical genetic roadmap to improved tomato flavor. *Science,* 355**,** 391-394.
TÖPFER, N., FUCHS, L. M. & AHARONI, A. 2017. The PhytoClust tool for metabolic gene clusters discovery in plant genomes. *Nucleic Acids Res,* 45**,** 7049-7063.

TSUGAWA, H., RAI, A., SAITO, K. & NAKABAYASHI, R. 2021. Metabolomics and complementary techniques to investigate the plant phytochemical cosmos. *Natural Product Reports,* 38**,** 1729-1759.

TSUNO, Y., FUJIMATSU, T., ENDO, K., SUGIYAMA, A. & YAZAKI, K. 2018. Soyasaponins: A new class of root exudates in soybean *(Glycine max). Plant and Cell Physiology,* 59**,** 366-375.

TURNER, S. D. 2014. qqman: an R package for visualizing GWAS results using QQ and manhattan plots. *Biorxiv***,** 005165.

TYLKA, G. L. & MARETT, C. C. 2021. Known distribution of the soybean cyst nematode, *Heterodera glycines*, in the United States and Canada in 2020. *Plant Health Progress,* 22**,** 72-74.

VAN DYCK, S., GERBAUX, P. & FLAMMANG, P. 2010. Qualitative and quantitative saponin contents in five sea cucumbers from the Indian Ocean. *Marine Drugs,* 8**,** 173-189.

VANETTEN, H. D., MANSFIELD, J. W., BAILEY, J. A. & FARMER, E. E. 1994. Two classes of plant antibiotics: phytoalexins versus" phytoanticipins". *The Plant Cell,* 6**,** 1191.

VEITCH, N. C. 2007. Isoflavonoids of the Leguminosae. *Natural product reports,* 24**,** 417-464.

VON WETTBERG, E. J., CHANG, P. L., BAŞDEMIR, F., CARRASQUILA-GARCIA, N., KORBU, L. B., MOENGA, S. M., BEDADA, G., GREENLON, A., MORIUCHI, K. S. & SINGH, V. 2018. Ecology and genomics of an important crop wild relative as a prelude to agricultural innovation. *Nature communications,* 9**,** 649.

WALLER, G. R. & YAMASAKI, K. 2013. *Saponins used in traditional and modern medicine*, Springer Science & Business Media.

WANG, R., WANG, M., ZHOU, J., WU, D., YE, J., SUN, G. & SUN, X. 2021. Saponins in Chinese herbal medicine exerts protection in myocardial ischemia–reperfusion injury: Possible mechanism and target analysis. *Frontiers in Pharmacology,* 11**,** 570867.

WANG, S., LI, Y., HE, L., YANG, J., FERNIE, A. R. & LUO, J. 2022. Natural variance at the interface of plant primary and specialized metabolism. *Current Opinion in Plant Biology,* 67**,** 102201.

WANG, X., HOWELL, C. P., CHEN, F., YIN, J. & JIANG, Y. 2009. Gossypol-a polyphenolic compound from cotton plant. *Advances in food and nutrition research,* 58**,** 215-263.

WEIR, B. S. 1990. *Genetic data analysis. Methods for discrete population genetic data*, Sinauer Associates, Inc. Publishers.

WEN, W., LI, D., LI, X., GAO, Y., LI, W., LI, H., LIU, J., LIU, H., CHEN, W. & LUO, J. 2014. Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. Nat Commun 5: 3438.

WENDT, K. U. 2005. Enzyme mechanisms for triterpene cyclization: new pieces of the puzzle. *Angewandte Chemie International Edition,* 44**,** 3966-3971.

WENG, J.-K., LYNCH, J. H., MATOS, J. O. & DUDAREVA, N. 2021. Adaptive mechanisms of plant specialized metabolism connecting chemistry to function. *Nature Chemical Biology,* 17**,** 1037-1045.

WENG, J.-K., PHILIPPE, R. N. & NOEL, J. P. 2012. The rise of chemodiversity in plants. *Science,* 336**,** 1667-1670.
WENG, J. K. 2014. The evolutionary paths towards complexity: a metabolic perspective. *New Phytologist,* 201**,** 1141-1149.
WETZEL, W. C. & WHITEHEAD, S. R. 2020. The many dimensions of phytochemical diversity: linking theory to practice. *Ecology letters,* 23**,** 16-32.
WINK, M. 2015. Modes of action of herbal medicines and plant secondary metabolites. *Medicines,* 2**,** 251-286.
WISECAVER, J. H., BOROWSKY, A. T., TZIN, V., JANDER, G., KLIEBENSTEIN, D. J. & ROKAS, A. 2017. A global coexpression network approach for connecting genes to specialized metabolic pathways in plants. *The Plant Cell,* 29**,** 944-959.
WISHART, D. S., FEUNANG, Y. D., MARCU, A., GUO, A. C., LIANG, K., VÁZQUEZ-FRESNO, R., SAJED, T., JOHNSON, D., LI, C. & KARU, N. 2018. HMDB 4.0: the human metabolome database for 2018. *Nucleic acids research,* 46**,** D608-D617.
WU, X. & KANG, J. 2011. Phytochemicals in soy and their health effects. *Phytochemicals-Bioactivities and Impact on Health.* IntechOpen.
WURTZEL, E. T. & KUTCHAN, T. M. 2016. Plant metabolism, the diverse chemistry set of the future. *Science,* 353**,** 1232-1236.
XIAO, J. F., ZHOU, B. & RESSOM, H. W. 2012. Metabolite identification and quantitation in LC-MS/MS-based metabolomics. *Trends in analytical chemistry : TRAC,* 32**,** 1-14.
XIAO, Q., MU, X., LIU, J., LI, B., LIU, H., ZHANG, B. & XIAO, P. 2022. Plant metabolomics: a new strategy and tool for quality evaluation of Chinese medicinal materials. *Chinese Medicine,* 17**,** 45.
XU, R., FAZIO, G. C. & MATSUDA, S. P. 2004a. On the origins of triterpenoid skeletal diversity. *Phytochemistry,* 65**,** 261-291.
XU, Y.-H., WANG, J.-W., WANG, S., WANG, J.-Y. & CHEN, X.-Y. 2004b. Characterization of *GaWRKY1,* a cotton transcription factor that regulates the sesquiterpene synthase gene (+)-δ-cadinene synthase-A. *Plant physiology,* 135**,** 507-515.
YAMAMURA, C., MIZUTANI, E., OKADA, K., NAKAGAWA, H., FUKUSHIMA, S., TANAKA, A., MAEDA, S., KAMAKURA, T., YAMANE, H. & TAKATSUJI, H. 2015. Diterpenoid phytoalexin factor, a *bHLH* transcription factor, plays a central role in the biosynthesis of diterpenoid phytoalexins in rice. *The Plant Journal,* 84**,** 1100-1113.
YANG, C. Q., FANG, X., WU, X. M., MAO, Y. B., WANG, L. J. & CHEN, X. Y. 2012. Transcriptional regulation of plant secondary metabolism F. *Journal of integrative plant biology,* 54**,** 703-712.
YANG, K., TIAN, Z., CHEN, C., LUO, L., ZHAO, B., WANG, Z., YU, L., LI, Y., SUN, Y. & LI, W. 2015. Genome sequencing of adzuki bean (Vigna angularis) provides insight into high starch and low fat accumulation and domestication. *Proceedings of the National Academy of Sciences,* 112**,** 13213-13218.
YANO, R., TAKAGI, K., TAKADA, Y., MUKAIYAMA, K., TSUKAMOTO, C., SAYAMA, T., KAGA, A., ANAI, T., SAWAI, S. & OHYAMA, K. 2017.

Metabolic switching of astringent and beneficial triterpenoid saponins in soybean is achieved by a loss-of-function mutation in cytochrome P450 *72A69*. *The Plant Journal,* 89**,** 527-539.

YANO, R., TAKAGI, K., TOCHIGI, S., FUJISAWA, Y., NOMURA, Y., TSUCHINAGA, H., TAKAHASHI, Y., TAKADA, Y., KAGA, A. & ANAI, T. 2018. Isolation and characterization of the soybean *Sg-3* gene that is involved in genetic variation in sugar chain composition at the C-3 position in soyasaponins. *Plant and Cell Physiology,* 59**,** 797-810.

YASMIN, F., ZHANG, H., LEAMY, L., WANG, B., WINNIKE, J., REID, R. W., BROUWER, C. R. & SONG, B.-H. 2022. Genetic basis and selection of glyceollin induction in wild soybean. *bioRxiv***,** 2022.12.17.520864.

YEAMAN, S. & WHITLOCK, M. C. 2011. The genetic architecture of adaptation under migration–selection balance. *Evolution: International Journal of Organic Evolution,* 65**,** 1897-1911.

YENCHO, G., KOWALSKI, S., KOBAYASHI, R., SINDEN, S., BONIERBALE, M. & DEAHL, K. 1998. QTL mapping of foliar glycoalkaloid aglycones in *Solanum tuberosum*× *S. berthaultii* potato progenies: quantitative variation and plant secondary metabolism. *Theoretical and applied genetics,* 97**,** 563-574.

YONEKURA-SAKAKIBARA, K. & HANADA, K. 2011. An evolutionary view of functional diversity in family 1 glycosyltransferases. *The Plant Journal,* 66**,** 182-193.

YONEYAMA, K., AKASHI, T. & AOKI, T. 2016. Molecular characterization of soybean pterocarpan 2-dimethylallyltransferase in glyceollin biosynthesis: local gene and whole-genome duplications of prenyltransferase genes led to the structural diversity of soybean prenylated isoflavonoids. *Plant and Cell Physiology,* 57**,** 2497-2509.

YU, O., JUNG, W., SHI, J., CROES, R. A., FADER, G. M., MCGONIGLE, B. & ODELL, J. T. 2000. Production of the isoflavones genistein and daidzein in non-legume dicot and monocot tissues. *Plant physiology,* 124**,** 781-794.

YU, O., SHI, J., HESSION, A. O., MAXWELL, C. A., MCGONIGLE, B. & ODELL, J. T. 2003. Metabolic engineering to increase isoflavone biosynthesis in soybean seed. *Phytochemistry,* 63**,** 753-763.

ZÄHRINGER, U., EBEL, J. & GRISEBACH, H. 1978. Induction of phytoalexin synthesis in soybean: Elicitor-induced increase in enzyme activities of flavonoid biosynthesis and incorporation of mevalonate into glyceollin. *Archives of Biochemistry and Biophysics,* 188**,** 450-455.

ZHA, L.-Y., MAO, L.-M., LU, X.-C., DENG, H., YE, J.-F., CHU, X.-W., SUN, S.-X. & LUO, H.-J. 2011. Anti-inflammatory effect of soyasaponins through suppressing nitric oxide production in LPS-stimulated RAW 264.7 cells by attenuation of NF-κB-mediated nitric oxide synthase expression. *Bioorganic & medicinal chemistry letters,* 21**,** 2415-2418.

ZHANG, C., BRADSHAW, J. D., WHITHAM, S. A. & HILL, J. H. 2010. The development of an efficient multipurpose bean pod mottle virus viral vector set for foreign gene expression and RNA silencing. *Plant physiology,* 153**,** 52-65.

ZHANG, C., WHITHAM, S. A. & HILL, J. H. 2013. Virus-induced gene silencing in soybean and common bean. *Virus-Induced Gene Silencing.* Springer.

ZHANG, H., KJEMTRUP-LOVELACE, S., LI, C., LUO, Y., CHEN, L. P. & SONG, B.-H. 2017a. Comparative RNA-seq analysis uncovers a complex regulatory network for soybean cyst nematode resistance in wild soybean *(Glycine soja)*. *Scientific reports,* 7**,** 1-14.

ZHANG, H., MITTAL, N., LEAMY, L. J., BARAZANI, O. & SONG, B. H. 2017b. Back into the wild—Apply untapped genetic diversity of wild relatives for crop improvement. *Evolutionary Applications,* 10**,** 5-24.

ZHANG, H. & SONG, B. H. 2017. RNA-seq data comparisons of wild soybean genotypes in response to soybean cyst nematode (*Heterodera glycines*). *Genom Data,* 14**,** 36-39.

ZHANG, H., YASMIN, F. & SONG, B.-H. 2019. Neglected treasures in the wild — legume wild relatives in food security and human health. *Current Opinion in Plant Biology,* 49**,** 17-26.

ZHANG, H. Y., KJEMTRUP-LOVELACE, S., LI, C. B., LUO, Y., CHEN, L. P. & SONG, B. H. 2017c. Comparative RNA-seq analysis uncovers a complex regulatory network for soybean cyst nematode resistance in wild soybean (*Glycine soja*). *Scientific Reports,* 7**,** 9699.

ZHAO, K. & RHEE, S. Y. 2022. Omics-guided metabolic pathway discovery in plants: Resources, approaches, and opportunities. *Current Opinion in Plant Biology,* 67**,** 102222.

ZHENG, Y., SZUSTAKOWSKI, J. D., FORTNOW, L., ROBERTS, R. J. & KASIF, S. 2002. Computational identification of operons in microbial genomes. *Genome research,* 12**,** 1221-1230.

ZHENG, Z., QAMAR, S. A., CHEN, Z. & MENGISTE, T. 2006. *Arabidopsis WRKY33* transcription factor is required for resistance to necrotrophic fungal pathogens. *The Plant Journal,* 48**,** 592-605.

ZHOU, X. & LIU, Z. 2022. Unlocking plant metabolic diversity: A (pan)-genomic view. *Plant Communications***,** 100300.

ZHOU, Y., MA, Y., ZENG, J., DUAN, L., XUE, X., WANG, H., LIN, T., LIU, Z., ZENG, K. & ZHONG, Y. 2016. Convergence and divergence of bitterness biosynthesis and regulation in Cucurbitaceae. *Nature plants,* 2**,** 1-8.

ZHOU, Z., JIANG, Y., WANG, Z., GOU, Z., LYU, J., LI, W., YU, Y., SHU, L., ZHAO, Y. & MA, Y. 2015. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nature biotechnology,* 33**,** 408-414.

ZHU, F., DU, B. & XU, B. 2018a. Anti-inflammatory effects of phytochemicals from fruits, vegetables, and food legumes: A review. *Critical reviews in food science and nutrition,* 58**,** 1260-1270.

ZHU, G., WANG, S., HUANG, Z., ZHANG, S., LIAO, Q., ZHANG, C., LIN, T., QIN, M., PENG, M. & YANG, C. 2018b. Rewiring of the fruit metabolome in tomato breeding. *Cell,* 172**,** 249-261. e12.

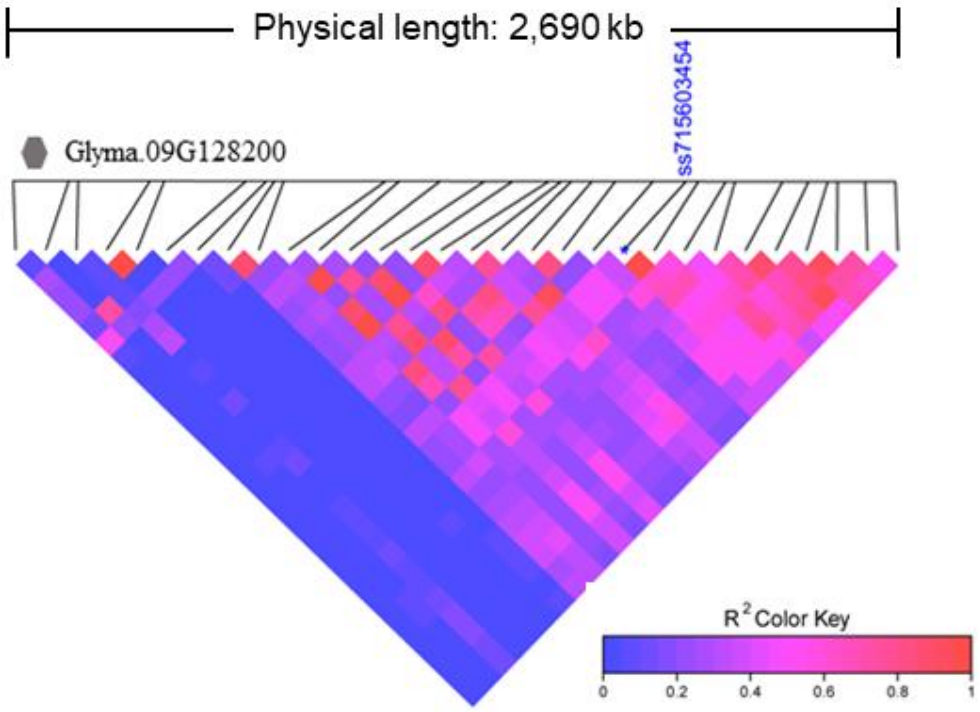
ZHUANG, Y., LI, X., HU, J., XU, R. & ZHANG, D. 2022. Expanding the gene pool for soybean improvement with its wild relatives. *aBIOTECH,* 3**,** 115-125.

ZULAK, K. G. & BOHLMANN, J. 2010. Terpenoid biosynthesis and specialized vascular cells of conifer defense. *Journal of Integrative Plant Biology,* 52**,** 86-97.

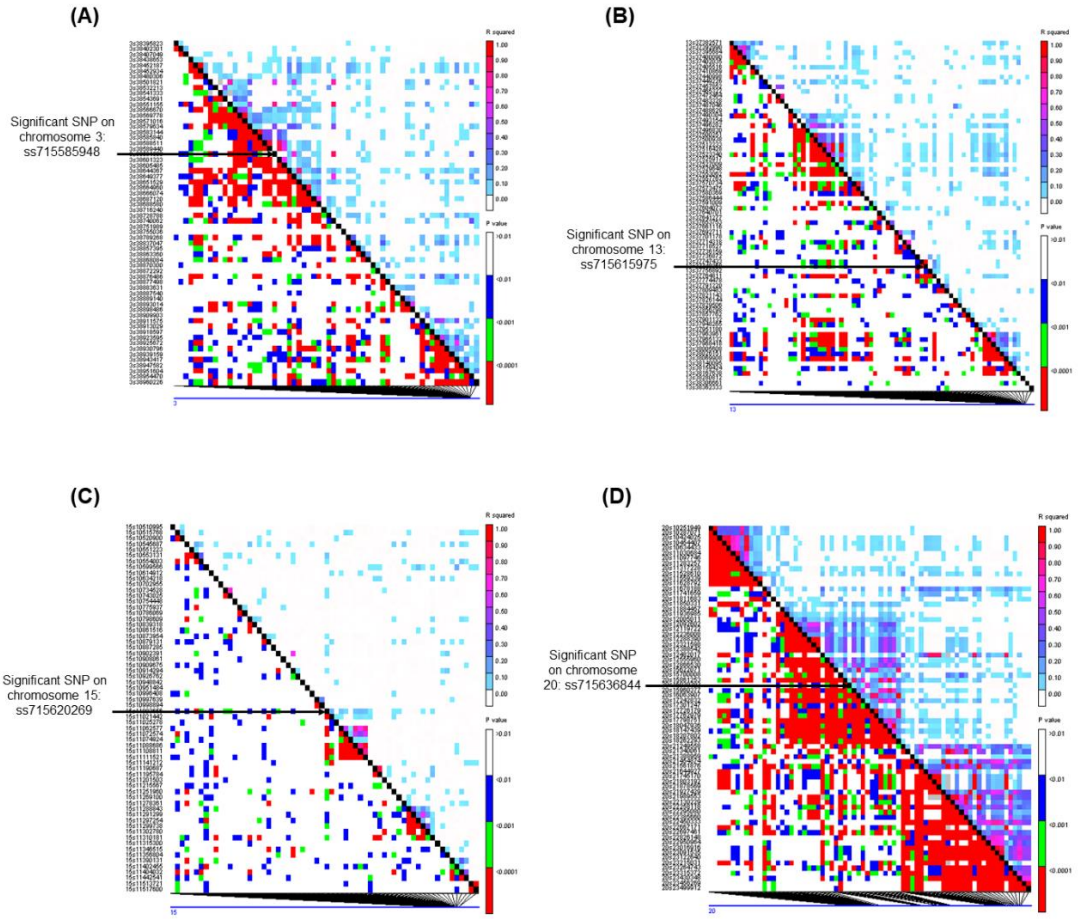
APPENDIX A: CHAPTER 2 SUPPLEMENTARY FIGURES AND TABLES

Supplementary Figures

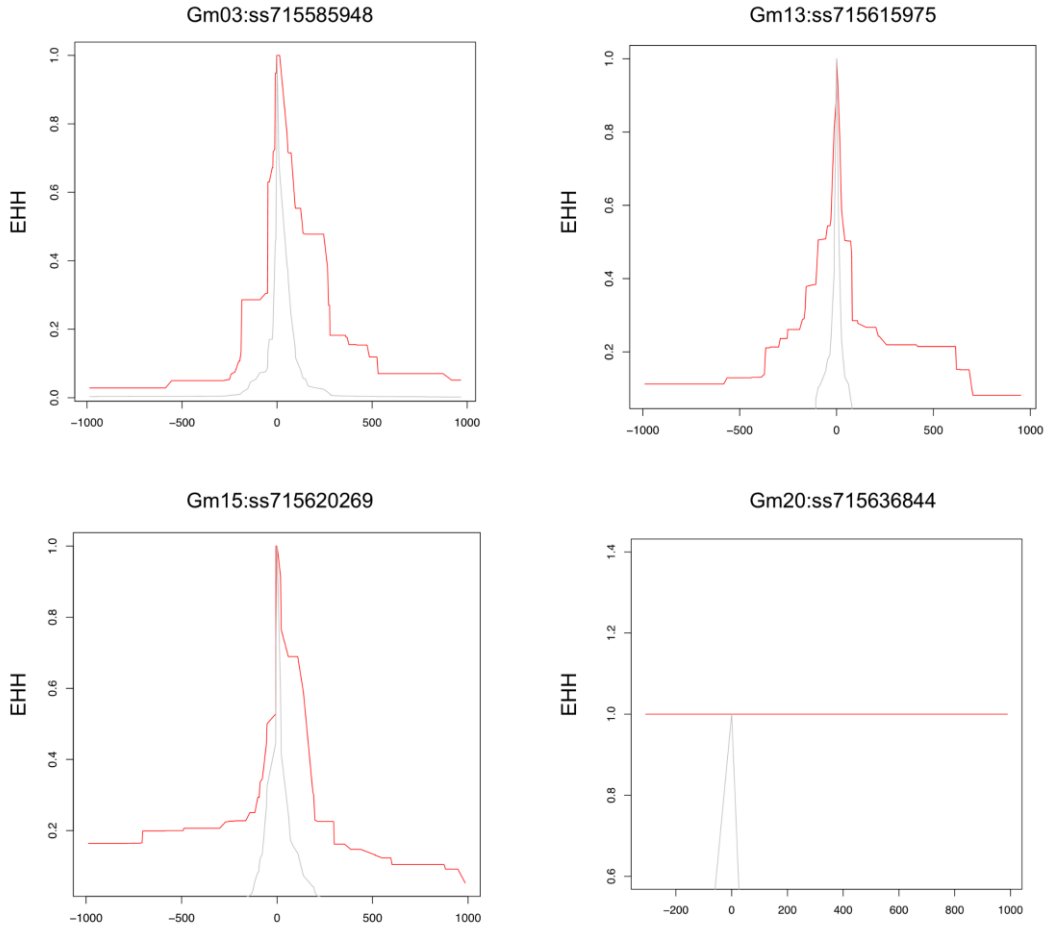
Supplementary Figure 2.1 A pairwise linkage disequilibrium between target SNP (ss715603454) and candidate gene *Glyma.09G128200* (location of the gene is indicated as a gray hexagon) from our metabolic gene cluster of interest.



Supplementary Figure 2.2 Significant SNPs ss715585948 **(A)**, ss715615975 **(B)**, ss715620269 **(C)**, and ss715636844 **(D)** on chromosomes 3, 13, 15, and 20, respectively, show narrow LD blocks.



Supplementary Figure 2.3 Allele-specific Extended Haplotype Homozygosity (EHH) for significant SNPs on chromosomes 3, 13, 15 and 20. Haplotype lengths are shown flanking the T (red) and G (grey) allele.



Supplementary Tables

Supplementary Table 2.1 Wild soybean ecotypes and metabolite data used in this study.

PI#	Daidzein (?g/g root)	Daidzein (?g/g root)	Daidzein (?g/g root)	Daidzein (?g/g root)	Daidzein	Glyceollin	Glyceollin	Glyceollin	Glyceollin	Glyceollin	GLYmean: DZNmean (GVSD)
	T1	T2	T3	T4	mean	T1	T2	T3	T4	mean	
PI 101404 B	383.27	592.04	574.57	376.76	481.66		9.74	12.77	14.33	12.28	0.025
PI 339732	120.39	79.79		160.04	120.07		29.69	44.85	40.13	38.23	0.318
PI 366120	158.44	140.07		162.14	153.55	7.06	10.57	9.49		9.04	0.059
PI 366122		479.48	486.39	323.98	429.95	33.47	47.69		59.8	46.99	0.109
PI 366123	413.58	450.62	494.51		452.9	112.96	36.45	98.95	22.14	67.63	0.149
PI 406684	119.74	22.19	60.02	257.04	114.75	8.76		10.24	10.95	9.98	0.087
PI 407037	172.28	327.4	375.64		291.78		12.53	6.14	7.63	8.76	0.03
PI 407047	325.11	189.88		225.25	246.75	6.64		2.67	2.33	3.88	0.016
PI 407050	156.73		95.96	121.66	124.78	25.95	27.71	12.28	5.34	17.82	0.143
PI 407053	52.79	266.17	125.33	67.5	127.95	13.56	14.39	14.73		14.23	0.111
PI 407089		143.66	182.14	265.2	197		13.01	9.93	13.76	12.23	0.062
PI 407097	431.12	291.61	287.07		336.6	15.71	17.88	30.89		21.49	0.064
PI 407120	265.52	369.54		357.44	330.83		5.74	10.14	3.94	6.6	0.02
PI 407167	297.14	177.74		157.72	210.86	23.57	34.36	16.85		24.92	0.118
PI 407174	306.96	226.41	185.78		239.72	10.45	23.92		18.5	17.63	0.074
PI 407198	166.55		170.29	103.12	146.65		8.16	8.04	6.98	7.72	0.053
PI 407201	276.46	361.51	280.38	255.9	293.56		11.12	5.11	7.16	7.8	0.027
PI 407202	289.8	358.94	239.4		296.04	8.72	4.05	9.7		7.49	0.025
PI 407217	360.99	320.66	217.49		299.71	11.06	4.61	36.14	65.45	29.32	0.098
PI 407221	204.08	534.68	391.65	504.82	408.81		6.66	13.69	4.24	8.2	0.02
PI 407246	188.39	312.71	248.87	198.58	237.14	2.59	1.32	1.79		1.9	0.008
PI 407249	484.36	370.37	528.12		460.95	5.26	7.35		11.48	8.03	0.017
PI 407254	616.11	463.03		456.93	512.02	3.06	6.12	4.68		4.62	0.009
PI 407267	336.6		441.75	318.53	365.62	14.64		14.04	19.83	16.17	0.044
PI 407271	92.43	49.27	155.9	33.88	82.87	21.13	10.05	9.9		13.69	0.165
PI 407275	204.98	267.1	237.06		236.38		7.78	13.22	8.24	9.75	0.041
PI 407278	240.08		283.77	248.15	257.34	2.42		2.12	2.4	2.31	0.009
PI 407302		256.25	188.91	241.47	228.88	4.52	4.57	8.61	12.25	7.49	0.033
PI 407304	306.58	274.6	146.31	129.91	214.35		8.5	13.95	7.8	10.08	0.047
PI 424059 B	177.58	77.43	178.21	79.42	128.16	7.72		1.52	7.87	5.71	0.045
PI 424063		222.15	386.14	301.84	303.37		1.71	4.04	4.24	3.33	0.011
PI 424064	278.15	302.85		377.86	319.62	4.29		2.96	4.65	3.97	0.012
PI 424088	312.28		393.73	252.41	319.47		4.09	10.62	8.87	7.86	0.025
PI 424093	205.54		154.22	381.2	246.99	5.11	5.76		13.5	8.13	0.033
PI 424102 A	302.11	408.96	294.42		335.17	11.83		11.15	5.36	9.45	0.028
PI 424117	288.15	191.82	199.54		226.5	3.62	3.53	3.79		3.65	0.016
PI 468396 B	497.02	471.44		332.81	433.76	21.25	21.84	10.3		17.8	0.041
PI 468397 A		281.27	167.09	189.64	212.67	28.63	30.96	36.24		31.95	0.15
PI 468398 B	20.38	200.54	42.46	101.3	91.17		51	34.99	69.09	51.69	0.567
PI 468399 B	192.42		128.05	135.58	152.02	58.99	27.19	5.62	12.54	26.08	0.172
PI 483466	401.92	507.26	414.84	591.31	478.83	24.79		16.34	15.86	19	0.04
PI 483468 A	346.27	261.14	494.63		367.35	8.27	16.39		16.3	13.65	0.037
PI 487430	206.11	214.76	285.26	264.83	242.74	6.31	8.86		8.35	7.84	0.032
PI 507582		355.54	348.96	440	381.5		7.88	11.74	5.22	8.28	0.022
PI 507632	187.64		347.63	163.38	232.88	5.75		14.8	11.38	10.64	0.046
PI 507644	148.95	205.46	203.36	125.58	170.84	8.56	10.21	11.79	7.93	9.62	0.056
PI 508067	128.78	267.22	112.23	147.42	163.91	41.96		61.05		42.16	0.257
PI 522179		369.46	208.57	248.69	275.57	41.25	33.27	63.64		46.06	0.167
PI 522180		213.28	139.48		191.98			60.98	68.29	63.71	0.332
PI 549037		297.35	237.95	287.13	274.15	28.64	20.86	35.66		28.39	0.104
PI 549046	147.55	93.55		130.12	123.74	12.13		8.7	11.93	10.92	0.088
PI 562544	265.23	232.64	382.27		293.38	8.68	12.63	15.83		12.38	0.042
PI 562550	393.54	352.86	267.69		338.03	46.11	17.94	45.89		36.65	0.108
PI 578345		416.6	361.25	406.05	394.63	5.43		4.51	12.83	7.59	0.019
PI 163453	65.82	119.63	67.38	157.15	102.5	5.49	15.41	15.95		12.28	0.12
PI 366121		300.1	241.41	189.32	243.61	3.65		10.23	6.34	6.74	0.028
PI 378695 A	324.5	433.88	309.84		356.08	15.66	5.17	3.98		8.27	0.023
PI 378701 A	82.52	123.05	184.34		129.97		2.26	28.07	53.17	27.83	0.214
PI 407030	178.1	53.99	118.81	420.19	192.77	11.19		10.62	26.46	16.09	0.083
PI 407034	430.6	180.95	175.91		262.49	12.64		10.44	12.38	11.82	0.045
PI 407124		106.98	118.77	99.4	108.38	18.58		24.85	20.47	21.3	0.197
PI 407200	462.23		463.74	408.52	444.83	3.04	12.24		6.14	7.14	0.016

PI#	Daidzein (?g/g root)	Daidzein (?g/g root)	Daidzein (?g/g root)	Daidzein (?g/g root)	Daidzein	Glyceollin	Glyceollin	Glyceollin	Glyceollin	Glyceollin	GLYmean: DZNmean (GVSD)
	T1	T2	T3	T4	mean	T1	T2	T3	T4	mean	
PI 407279	425.93	320.9		248.36	331.73	4.5	6.11	2.22		4.28	0.013
PI 407296	284.57		341	404.82	343.46	9.95		7.13	6.15	7.74	0.023
PI 423993	358.8	488.04		385.19	410.68	24	24.4	19.84		22.75	0.055
PI 423996	179.91	368.4	327.21	207.54	270.76		4.74	4.22	4.43	4.46	0.016
PI 464934	237.67		219.58	253.75	237		6.67	5.36	2.88	4.97	0.021
PI 464936 B	351.1	475.66	317.82		381.53	5.99	5.94	4.38	4.6	5.23	0.014
PI 464937 A	568.6	232.66	317.26	600.8	429.83	19.66	86.05	16.13	45.08	41.73	0.097
PI 423988	432.52	478.55	214.13	182.6	326.95	4.03	4.06		5.55	4.55	0.014
PI 479749		85.83	33.43	92.19	70.48	14.34	8.09		19.88	14.11	0.2
PI 479751	113.52	269.1	248.46		210.36	13.75	10.35	11.52		11.88	0.056
PI 483466	556.66		267.21	297.6	373.82	10.24		12.07	8.16	10.16	0.027
PI 504287 A	279.95	590.42		403.97	424.78		8.75	8.6	8.94	8.76	0.021
PI 507727	294.8	575.34	355.01	502.98	432.03	22.87	19.91		22.04	21.61	0.05
PI 507787		301.28	337.48	162.33	267.03		19.84	12.09	17.97	16.63	0.062
PI 507798	222.4	193.88	480.17	620.69	379.29	8.78		10.44	11.59	10.27	0.027
PI 508066	29.55	44.36		30.76	34.89		8.58	21.14	20.42	16.71	0.479
PI 508069	21.32	6.89		9.86	12.69	4.05	11.31		3.8	6.39	0.503
PI 522211 B		307.97	335.41	312.09	318.49	12.49		11.37	17.67	13.84	0.043
PI 522234	256.85	274.96	398.62	324.68	313.78	9.47	21.09	10.39	5.67	11.66	0.037
PI 135624	699.08	420		697.32	605.46	21.8	11.29	12.77	7.58	13.36	0.022
PI 339731	131.8	291.67		451.6	291.69	13.57	13.35		7.31	11.41	0.039
PI 339735 A	215.62	361.42	191.26	410.67	294.74	8.51	28.04	16.95	40.8	23.58	0.08
PI 366119	164.38	123.86	241.27	99.08	157.15	15.1		15.42	15.6	15.37	0.098
PI 378685	321.31	120.61	286.09		242.67	9.18	44.57	3.38		19.04	0.078
PI 378687 A	199.53	179.79	370.99		250.1	3.66	7.53	2.89		4.69	0.019
PI 378689	211.93	252.25	66.83		177	12.28	22.36	13.48		16.04	0.091
PI 378700	56.19	157.97	109.18		107.78	32.19	7.05	17.9		19.05	0.177
PI 378702		131.39	111.85	100.35	114.53	4.62	28.66	5.49	18.15	14.23	0.124
PI 407020	124.04	96.44	82.7	139.86	110.76		21.17	20.93	22.63	21.58	0.195
PI 407026		215.04	262.97	264.37	247.46	5.2		6.12	15.58	8.97	0.036
PI 407029	113.38	87.37	302.1		167.62	11.4	13.96	10.12		11.83	0.071
PI 407032 B	150.41	222.87	30.56	60.73	116.14		19.23	27.93	28.23	25.13	0.216
PI 407044		201.28	137.9	135.07	158.08	20.1	6.16	9.06	17.1	13.11	0.083
PI 407048	231.92	177.2	165.43	288.49	215.76	9.27	3.89	8.66		7.27	0.034
PI 407055	136.01	111.18		169.12	138.77		28.16	15.64	19.1	20.97	0.151
PI 407060	247.07	91.71		35.17	124.65		6.96	6.07	2.78	5.27	0.042
PI 407063	125.16	374.4	67.13	312.63	219.83	8.28		12.87	15.74	12.3	0.056
PI 407069	80.55	190.4	120.27	63.16	113.59		25.09	28.91	21.04	25.01	0.22
PI 407072		377.45	91.83	202.98	224.09	12.68	36	31.08	11.15	22.73	0.101
PI 407085	0	120.2	123.33		81.18	3.96	12.52	7.33		7.94	0.098
PI 407092	185.26	56.8	2.25	6.59	62.73	0.47	6.92	3.88	23.82	8.77	0.14
PI 407094	184.99	126.61	358.89	61.32	182.95	9.76	8.53		8.74	9.01	0.049
PI 407096	27.5	123.16	112.47	28.13	72.82	29	13.3	10.23		17.51	0.24
PI 407099	136.42	9.07	17.73	49.19	53.1	2.17	0.33	0.52		1.01	0.019
PI 407100	393.72	139.6	44.89	20.73	149.74	17.66	15.74	13.4		15.6	0.104
PI 407102	328.8	70.52	225.24	67.39	172.99	22.86	20.03	25.07		22.66	0.131
PI 407107		157.25	237.91	275.16	223.44		10.58	25.13	59.08	31.6	0.141
PI 407109	206.2	153.09	111.29		156.86	25.67		25.2	38.72	29.86	0.19
PI 407113	140.67	79.33		93.07	104.36	7.33	4.29	12.72	5.34	7.42	0.071
PI 407119	176.67	172.8	62.69		137.38	16.33		15	3.99	11.78	0.086
PI 407121	475.92	317.61	108.33	172.02	268.47	23.95	41.73	63.54	67.76	49.25	0.183
PI 407126	383.76	336.44	183.33	196.22	274.94	14.68		14.98	14.39	14.69	0.053
PI 407131	127.6	197.76	71.36		132.24	6.2	7.09	6.79		6.7	0.051
PI 407137	475.88	341.05	304.56	422.65	386.04		12.62	9.83	11.49	11.31	0.029
PI 407142	341.05	198.67	250.98	167.43	239.53	3.24	5.74		3.03	4.01	0.017
PI 407145		268.14	252.57	280.96	267.22	13.17	14.99	8.26		12.14	0.045
PI 407147	279.96		295.6	248.26	274.61	6.45	15.71	8.56		10.24	0.037
PI 407149	305.12		339.75	260.63	301.83	7.62		6.76	4.49	6.29	0.021
PI 407151		362.86	295.98	385.7	348.18	5.63	13.95		8.74	9.44	0.027
PI 407153	518.01	344.52	448.28	299.53	402.59		11.58	8.94	7.11	9.21	0.023
PI 407155	116.5	241.6	206.51	100.13	166.19		9.24	5.8		6.77	0.041
PI 407159	203.87	150.82		177	177.23	26.59	47.3		13.5	29.13	0.164
PI 407172	234.93	357.82	111.7	126.5	207.74	16.57	12.44	12.33		13.78	0.066
PI 407178	194.27	201.11	306.08	278.28	244.94	13.42	4.88	23.6	6.52	12.1	0.049
PI 407187	131.35	232.29	249.81	133.03	186.62	14.52		9.48	11.31	11.77	0.063

PI#	Daidzein (?g/g root)	Daidzein (?g/g root)	Daidzein (?g/g root)	Daidzein (?g/g root)	Daidzein	Glyceollin	Glyceollin	Glyceollin	Glyceollin	Glyceollin	GLYmean: DZNmean (GVSD)
	T1	T2	T3	T4	mean	T1	T2	T3	T4	mean	
PI 407191	161.93	253.07	177.99		197.66	10.38		14.76	13.38	12.84	0.065
PI 407193	201.39	241.84		378.33	273.85		10.3	17.11	7.97	11.79	0.043
PI 407194		297.71	379.1	266.95	314.59		17.07	18.06	17.43	17.52	0.056
PI 407203		180.81	167.34	345.64	254.93		131.65		83.85	87.28	0.342
PI 407205	154.52	247.34	212.39	160.47	193.68	14.27	36.63	13.45		21.45	0.111
PI 407207	401.49	444.2	405.68		417.13	22.65	29.44	17.03		23.04	0.055
PI 407208	215.57	274.63	349.46		279.89	13.65		11.98	15.04	13.56	0.048
PI 407211	260.62	88.16	56.51	322.89	182.04		0.02	20.92	50.84	23.93	0.131
PI 407214	272.03	225.79	90.16	137.86	181.46	7.73	12.01	5.58		8.44	0.047
PI 407222	118.15	63.5	7.19	0	47.21	20.04	17.65	11.03		16.24	0.344
PI 407223	316.45	342.96	753.55	42.38	363.84	32.1	9.78		12.94	18.27	0.05
PI 407233	170.73	277.6	18.81	58.68	131.46	14.3	14.09		15.62	14.67	0.112
PI 407236	11.83		17.92	29.7	19.81	25.97	56.58	86.96	20.54	47.51	2.398
PI 407252	118.41	191.72		110.52	140.21	0	2.18		0.29	0.82	0.006
PI 407256	139.02	107.77	13.93		86.91	12.52	10.45	3.17		8.71	0.1
PI 407258	62.86	148.51	182.85		131.4	15.04	7.52		17.2	13.25	0.101
PI 407260	425.52	140.49	417.45	766.46	437.48	95.24	71.37	23.57	21.59	52.94	0.121
PI 407261	62.28	312.69	195.79		190.26	69.53	20.37	7.72		32.54	0.171
PI 407272		234.5	308.19	143.33	228.67		50.2	26.82	17.03	31.35	0.137
PI 407274	169.19	155.88	310.48		211.85	19.82	6.63	5.01		10.49	0.05
PI 407285	231.74	300.41	6.33	46.33	146.2	28.85	24.39		37.85	30.36	0.208
PI 407290	276.63		394.77	200.5	290.63		8.32	6.02	4.95	6.43	0.022
PI 407292	266.72	177.22	135.64	224.72	201.08	5.7		5.23	10.36	7.1	0.035
PI 407294		286.87	350.27	191.6	276.25	10.77	5.81		9.26	8.61	0.031
PI 407298	390.66		466.7	313.47	390.28	66.66	42.4		35.59	48.21	0.124
PI 407300		39.01	125.75	135.57	100.11		91.24	64.65	59.33	71.74	0.717
PI 407310	388.57	223.84	74.94	131.93	204.82	28.19	87.71	16.12	82.63	53.66	0.262
PI 407312		394.99		204.26	321.37		44.79		55.97	43.39	0.135
PI 407319		153.72	167.81	294.39	205.31	79.08	43.37		72.23	64.89	0.316
PI 407320	39.82	135.77	63.08		79.55	17.94	11.79	14.18		14.64	0.184
PI 407322	493.87	502.7	492.6		496.39	8.19	10	25.31		14.5	0.029
PI 424006 A	190.32		388.93	136.78	238.68	20.12	15.26	23.61		19.66	0.082
PI 424010	104.9	266.19	564.47	433.75	342.33	17.6	48.98	38.98	14	29.89	0.087
PI 424012	587.27	238.44	337.7	410.59	393.5	24.72		44.7	30.51	33.31	0.085
PI 424017 A		143.65	143.89	144.69	144.08		22.99	9.86	17.78	16.88	0.117
PI 424022 A		47.74	45.19	75.29	56.07	28.36	23.21	30.45		27.34	0.488
PI 424023	139.97		147.92	273.59	187.16	13.89		17.52	40.01	23.81	0.127
PI 424027 B	57.71	148.58	139.71		115.34	0.55	0	130.53		43.69	0.379
PI 424039 A	395.66	156.16	296.53	187.71	259.01	13.83	16.56		10.69	13.69	0.053
PI 424040	178.51	150.51		250.88	193.3	28.74	25.43		36.43	30.2	0.156
PI 424042	36.25	205.03	291.04		177.44	28.05	61.8	17.31		35.72	0.201
PI 424043 A	266.06		119.16	283.88	223.03	22.13		23	16.94	20.69	0.093
PI 424044		238.4	296.68	154.77	229.95	21.45	24.44		26.49	24.13	0.105
PI 424046 A	197.98	261.77		276.27	245.34		42.05	59.2	41.17	47.47	0.194
PI 424047		181.6	208.1	249.43	213.04		27.76	19.05	35.39	27.4	0.129
PI 424051 B	200.82	238.42	332.52		257.25	24.78	21.97	23.56		23.44	0.091
PI 424052	123.76	175.66	163.49	104.39	141.83	34.71	36.1	10.2	14.14	23.79	0.168
PI 424056	53.98	296.65	149.71		166.78	10.71	9.57	0.3		6.86	0.041
PI 424058	202.2		168.11	122.81	164.37	30.52	23.06		31.75	28.44	0.173
PI 424061 A	50.87	71.36			61.11	9.3	10.54			9.92	0.162
PI 424067	130.05	137.96	79.72	52.44	100.04	7.52	4.94	2.58		5.01	0.05
PI 424068	87.43	40.18	167.61	176.03	117.81	5.63	4.12		12.56	7.44	0.063
PI 424070 A	81.48	231.61	256.27	81.26	162.66	26.16	18.13		32.16	25.48	0.157
PI 424072 A	384.86		238.42	376.7	333.33	31.82	43.06		24.2	33.03	0.099
PI 424074	153.97		142.11	132.57	142.88	9.18	1.61		6.26	5.68	0.04
PI 424080	296.97	202.48	273.12		257.52	1.69	0.03	3.4		1.71	0.007
PI 424085 A		241.85		250.81	259.77	28.27		25.6		28.51	0.11
PI 424086		340.8	366.57	185.88	265.98	26.61	37.18	35.25		29.75	0.112
PI 424091 A	96.3	51.82	185.36		111.16	34.68	13.46	9.05		19.06	0.171
PI 424094	639.71	478.01	440.36		519.36	72.64	66.08	346.17		161.63	0.311
PI 424095	29.62	106.98	97.56	39.03	68.3	5.15	5.55		7.08	5.93	0.087
PI 424100 A	82.73	38.91	62.83		61.49	26.64	41.07		14.24	27.32	0.444
PI 424101		92.04	96.22	96.05	94.77	23.99	20.75		12.47	19.07	0.201
PI 424104	535.19		422.18	225.39	394.26	45.15		171.16	33.13	83.15	0.211
PI 424105	147.75	155	47.33	50.46	100.14	7.74	19.74	3.18		10.22	0.102

PI#	Daidzein (?g/g root)	Daidzein (?g/g root)	Daidzein (?g/g root)	Daidzein (?g/g root)	Daidzein	Glyceollin	Glyceollin	Glyceollin	Glyceollin	Glyceollin	GLYmean: DZNmean (GVSD)
	T1	T2	T3	T4	mean	T1	T2	T3	T4	mean	
PI 424110		329.61	604.14	549.57	494.44		124.56	111.81	68.23	101.53	0.205
PI 424112	174.31	89.35	295.43		186.37	12.82	27.46	23.46		21.25	0.114
PI 424115 A	121.43	362.58	350.45		278.15	39.39	163.66	34.48		79.17	0.285
PI 424123	259.17	292.18	259.02		270.12		41.9		62.5	46.39	0.172
PI 424125	44.1	79.91	9.69		44.57	33.72	31.67	74.99		46.79	1.05
PI 447004	389.7		463.37	354.77	402.61	29.56	31.01	79.28	50.82	47.67	0.118
PI 458535		229.24	396.23	194.95	273.47	52.01	23.01	69.34	36.78	45.29	0.166
PI 458536	192.96	181.31	181.13		185.13	25.25	21.41		27.02	24.56	0.133
PI 458539 A	159.42	291.87	313.99		255.09	20.9	35.21		29.59	28.57	0.112
PI 458540 A	33.65		75.32	55.85	54.94	55.98	52.7	105.84	175.61	97.53	1.775
PI 464867	191.56	179.02	227.06	220.92	204.64	32.99	48.57		55.24	45.6	0.223
PI 464868 A		176.61	159.05	128.8	154.82	68.4	25.75	63.4	34.3	47.96	0.31
PI 464869 A	203.08	337.5		256.1	265.56	31.38	17.85	26.25	14.78	22.57	0.085
PI 464870	140.47	156.17		80.82	125.82	43.42	32.29	23.28		32.99	0.262
PI 464871 A	60.38		19.59	32.28	37.42	36.71		23.91	5.84	22.15	0.592
PI 464889 A		327.38	303.74	209.76	280.3	23.19	32.96		23.07	26.41	0.094
PI 464890 B		292.78	215	257.72	255.16	49.65	20.78	43.29	23.8	34.38	0.135
PI 464892	235.29	388.78	181.63	386.21	297.98	30.97		3.67	31.01	21.88	0.073
PI 464929 A	370.88	189.22	467.01	172.31	299.85	72.78	53.81	28.31	130.87	71.44	0.238
PI 464939 A	64.63	25.65	28.37		39.55	19.15	9.57	6.61	23.05	14.6	0.369
PI 468398 A	122.85	173.52	81.68		126.02	8.51	15.08	2.26		8.62	0.068
PI 468400 A		24.74	74.76	38.16	45.89		12.74	0.02	0	4.25	0.093
PI 468916	25.1	91.92		112.7	76.58	2.57	6.66	6.33	3.29	4.71	0.062
PI 479746 A	285.38	148.5		149.07	194.32	42.99	16.73		23.35	27.69	0.143
PI 479747	383.14		466.45	360.44	403.34		15.36	12.92	11.06	13.11	0.033
PI 479748	10.44	12.59	136.99		53.34	5.75	52.91	18.45		25.7	0.482
PI 479750	102.18	343.72	342.83	65.4	213.53	6.8		5.96	3.08	5.28	0.025
PI 479753 A	626.24	306.15	607.82	106.6	411.71	100.3	78.9	52.53		77.25	0.188
PI 479767	253.5	234.9	283.31		257.23	37.83	6.19	13.69	44.77	25.62	0.1
PI 483071 A	547.66	580.31		615.65	581.21	25.17	28.12		37.92	30.4	0.052
PI 483461		707.75	622.13	439.53	589.8	100.32		116.56	94.31	103.73	0.176
PI 483463		21.27	27.03	54.18	34.16	59.61	15.36	13	38.88	31.71	0.928
PI 483464 A	408.33	496.19	294.83	243.05	360.6	19.18		19.61	17.5	18.76	0.052
PI 483465	396.52	517.71	58.34	112.01	271.14	33.6	27.65	24.19		28.48	0.105
PI 504289	227.3	48.25	112.61		129.38	42.28	50.84	15.87		36.33	0.281
PI 507581	43.33	98.7		34.01	58.68	10.74	13.24		17.69	13.89	0.237
PI 507584	303.25		320.02	359.42	327.56		66.24	49.76	23.36	46.45	0.142
PI 507585		54.25	61.34	91.19	68.93	23.63		22.99	16.42	21.01	0.305
PI 507590 A	276.82	210.69	416.26		301.26	14.16	72.41	26.3		37.62	0.125
PI 507590 B	55.4	56.85	271.4		127.89	1.36	7.77	10.59		6.57	0.051
PI 507591	151.46	199.04		165.14	171.88	14.34		10.65	8.3	11.09	0.065
PI 507599	184.49	370.1	582.79		379.13	0.02	0.02	47.47		15.83	0.042
PI 507600	241.43	146.57	109.22		165.74	105.29	16.04	6.78		42.7	0.258
PI 507601 A	110.6	182.67	225		172.76	36.51	8.84	15.05		20.13	0.117
PI 507605	366.24	220.18	107.26		231.23	12.08	11.11	8.69		10.63	0.046
PI 507606	17.74	346.28	206.54		190.19	0.21	32.69	0.06		10.99	0.058
PI 507607	36.78		183.56	98.48	106.27	20.23	102.6	104.98	50.29	69.53	0.654
PI 507617	325.92	484.69	237.26		349.29	56.11		17.36	10.87	28.11	0.08
PI 507618		383.89		287.94	324.33		32.27		28.47	24.72	0.076
PI 507619 A			200.13		234.91	22.8				26.12	0.111
PI 507621	216.29		279.49	133.03	209.61	10.22	7.31	8.51		8.68	0.041
PI 507626	321.41	280.28	326.69		309.46		4.5	6.25	8.67	6.47	0.021
PI 507627	317.4		349.33	242.42	303.05	5.78	7.28	9.98		7.68	0.025
PI 507628	143.1	40.17		44.04	75.77	8.71	5.82		9.64	8.06	0.106
PI 507631	175.12	240.86	367.92		261.3	6.26	11.19	3.73		7.06	0.027
PI 507633		65.52	107.64	86.32	86.49	4.32	3.93		7.56	5.27	0.061
PI 507639	131.54	72.28	227.63		143.81	2.91	1.92	25.76		10.2	0.071
PI 507640	146.94	75.64	64.47	166.99	113.51	7.6		13.24	9.05	9.97	0.088
PI 507641	83.87	20.52	51.74		52.04	11.58		16.73	8.31	12.21	0.235
PI 507653	313.44	75.06	249.59	507.74	286.46	11.88	13.66	7.81		11.11	0.039
PI 507655	177.6	132.02		286	198.54	6.86		9.87	8.18	8.3	0.042
PI 507658	240.12	290.16		219.09	249.79	2.74		8.07	8.67	6.49	0.026
PI 507662	86.57	151.16		118.65	118.79	2.78	2.76		4.29	3.28	0.028
PI 507665	268.51	278.55	232.21		259.76	1.92	2.23		9.29	4.48	0.017
PI 522181	145.27	84.19	133.25		120.9	31.63	10.21	28.22		23.36	0.193
PI 522183 A	240.83	211	540.23	704.86	424.23	8.63	9.82		14.89	11.12	0.026
PI 522184		92.35	155.03	88.36	111.91		19.29	31.28	12.5	21.02	0.188
PI 532449			284.79	296.8	313.32	28.38		19.75	14.25	20.79	0.066
PI 532452 A	418.02	200.72	372.63	209.16	300.13	4.27		8.04	6.16	6.16	0.021
PI 549032	76.52	137.82	93.09	215.21	130.66	2.94	3.01		4.09	3.35	0.026
PI 549035 A	99.29		139.56	389.28	209.38	7.38		7.36	9.21	7.98	0.038
PI 549036	409.65	247.56	325.22	294.1	319.13	19.57	12.79	12.72		15.03	0.047
PI 549039	243.26	237.39	472.34	360.63	328.41		5.1	4.6	1.96	3.89	0.012

Supplementary Table 2.2 SNP effect and heritability estimate for the significant SNPs related to plant specialized metabolic pathways.

Significant SNP	Chromosome	Position	LOD score	SNP effect	% R-squared value	h^2 (%)
ss715585948	Gm03	38591888	3.80	-0.31	10.58	35
ss715603454	Gm09	30262482	3.80	0.06	4.22	
ss715603455	Gm09	30191235	3.80	0.05	3.32	
ss715603462	Gm09	30393285	3.80	-0.06	3.96	
ss715603471	Gm09	30725658	3.80	-0.06	3.86	
ss715615975	Gm13	37748250	3.81	0.22	12.59	
ss715620269	Gm15	11003656	3.81	-0.20	11.42	
ss715636844	Gm20	15930392	3.79	-0.13	4.16	

Locations of the significant SNPs shown in base pairs, bp; standardized SNP effects, and the percentage (%) of the total phenotypic variation (glyceollin induction) explained by significant SNPs (% R- squared values) on different chromosomes. LOD scores represent the chromosome-wide significant level from 3.79 to 3.82. Heritability (h^2) of glyceollin induction was calculated with all SNPs.

Supplementary Table 2.3 Gene of interest with enzyme class and associated metabolic domain for chromosome 9.

Gene Cluster	Gene Name	Enzyme Class	Signature or Tailoring?	Metabolic Domain
Cluster 1	*Glyma.09G127200* *Glyma.09G127300*	glycosyltransferase	tailoring	Phenylpropanoid Derivatives; Specialized Metabolism
Cluster 2	*Glyma.09G127700* *Glyma.09G128200* *Glyma.09G128300* *Glyma.09G128400*	glycosyltransferase	tailoring	Phenylpropanoid Derivatives; Specialized Metabolism

Supplementary Table 2.4 Annotation of candidate genes of gene clusters 1 and 2.

Genes	Annotation
Glyma.09G127200 (cluster 1)	*UGT88A1* (AT3G16520.3); isoflavone 7-O-glucosyltransferase (2.4.1.170); metabolic process (GO:0008152); hexosyltransferase activity (GO:0016758), glucosyl/glucuronosyl transferases (PTHR11926); UDP-glucoronosyl and UDP-glucosyl transferase (PF00201)
Glyma.09G127300 (cluster 1)	*UGT88A1* (AT3G16520.3); isoflavone 7-O-glucosyltransferase (2.4.1.170); metabolic process (GO:0008152); hexosyltransferase activity (GO:0016758); glucosyl/glucuronosyl transferases (PTHR11926); UDP-glucoronosyl and UDP-glucosyl transferase (PF00201); glycosyltransferases (K08237)
Glyma.09G127700 (cluster 2)	*UGT88A1* (AT3G16520.3); isoflavone 7-O-glucosyltransferase (2.4.1.170); metabolic process (GO:0008152); hexosyltransferase activity (GO:0016758), glucosyl/glucuronosyl transferases (PTHR11926); UDP-glucoronosyl and UDP-glucosyl transferase (PF00201)
Glyma.09G128300 (cluster 2)	*UGT88A1* (AT3G16520.3); isoflavone 7-O-glucosyltransferase (2.4.1.170); metabolic process (GO:0008152); hexosyltransferase activity (GO:0016758), glucosyl/glucuronosyl transferases (PTHR11926); UDP-glucoronosyl and UDP-glucosyl transferase (PF00201)

Glyma.09G128200 (cluster 2)	UDP-glucosyl transferase *88A1* (AT3G16520.3); metabolic process (GO:0008152); intracellular membrane-bounded organelle (GO:0043231); transferase activity; transferring hexosyl groups (GO:0016758); quercetin 3-O-glucosyltransferase activity (GO:0080043); quercetin 7-O-glucosyltransferase activity (PWY-2345); biochanin A conjugates interconversion (PWY-2861); isoflavone 7-O-glucosyltransferase (GN7V-57685)
Glyma.09G128400 (cluster 2)	*UGT88A1* (AT3G16520.3), metabolic process (GO:0008152), hexosyltransferase activity (GO:0016758), glucosyl/glucuronosyl transferases (PTHR11926), and UDP-glucoronosyl and UDP-glucosyl transferase (PF00201)

Supplementary Table 2.5 Annotation of candidate genes other than genes within cluster 1 and 2.

Genes	**Annotation**
Glyma.20G052000, *Glyma.20G052400*	UDP-Glycosyltransferase superfamily protein
Glyma.20G057500	UDP-glucosyl transferase *85A2*
Glyma.20G058000	hydroxy methylglutaryl CoA reductase 1 (Mevalonate pathway I; isoprenoid biosynthetic process; sterol biosynthetic process; coumarin biosynthetic process; oxidoreductase activity)
Glyma.20G053900	cytochrome P450, family 71, subfamily B, polypeptide 34 (Oxidoreductase activity)
Glyma.20G065000, *Glyma.20G065100*	cytochrome p450 *79a2*
Glyma.13G272500	*bZIP* transcription factor
Glyma.09G125400, *Glyma.13G272700*, *Glyma.13G273500*, *Glyma.13G277600*, *Glyma.15G134700*, *Glyma.20G064600*	RING/U-box superfamily protein, RING/FYVE/PHD zinc finger superfamily protein

Glyma.03G176600, *Glyma.09G129100*, *Glyma.09G127100* *Glyma.15G139000*, *Glyma.15G135600*	*WRKY* family transcription factor family protein
Glyma.09G113000, *Glyma.09G113100*, *Glyma.15G134100*	*myb* domain
Glyma.03G173100, *Glyma.03G173200*, *Glyma.03G173300*, *Glyma.09G128700*, *Glyma.09G115100*, *Glyma.09G107400* *Glyma.13G274600*, *Glyma.20G059300*	zinc fingers superfamily protein
Glyma.13G274300, *Glyma.13G279900*, *Glyma.13G280000*	*NAC* transcription factors
Glyma.09G115900, *Glyma.09G117400*, *Glyma.09G123200*, *Glyma.13G277100*,	cytochrome P450 enzyme family

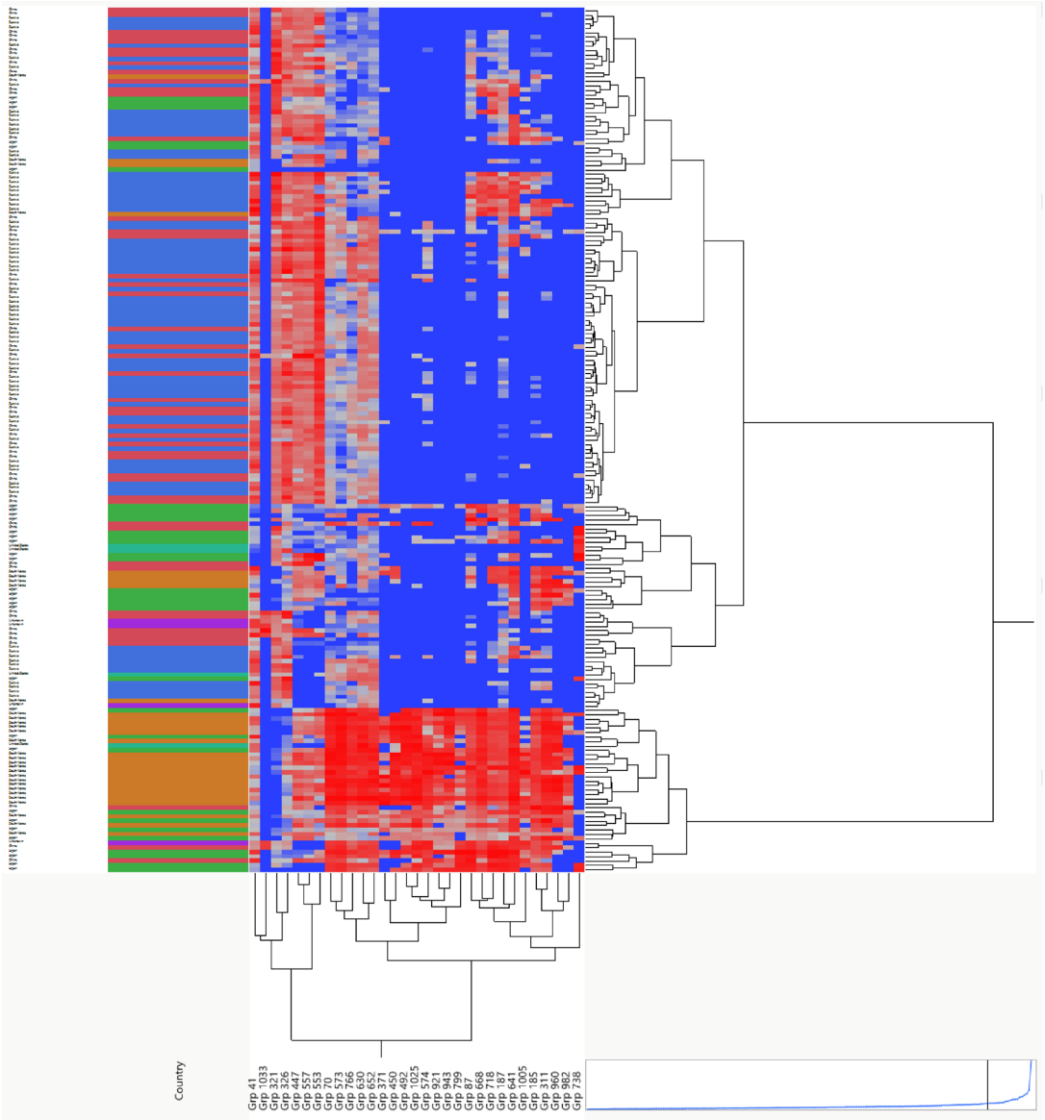
Glyma.20G053900, *Glyma.20G065000*, *Glyma.20G065100*	
Glyma.20G065000, *Glyma.20G065100*	cytochrome p450 *79a2* (oxidoreductase activity; Linamarin biosynthesis**)**
Glyma.09G108800	Zinc finger, RING-type; Transcription factor jumonji/aspartyl beta-hydroxylase
Glyma.09G109500, *Glyma.09G122500*	Terpenoid cyclases family protein, terpene synthase 03
Glyma.03G176300	Glutathione S-transferase family protein (glucosinolate biosynthetic process)
Glyma.09G126500 *Glyma.09G126900* *Glyma.15G139200*	phenylpropanoid metabolic process

APPENDIX B: CHAPTER 3 SUPPLEMENTARY FIGURES AND TABLES

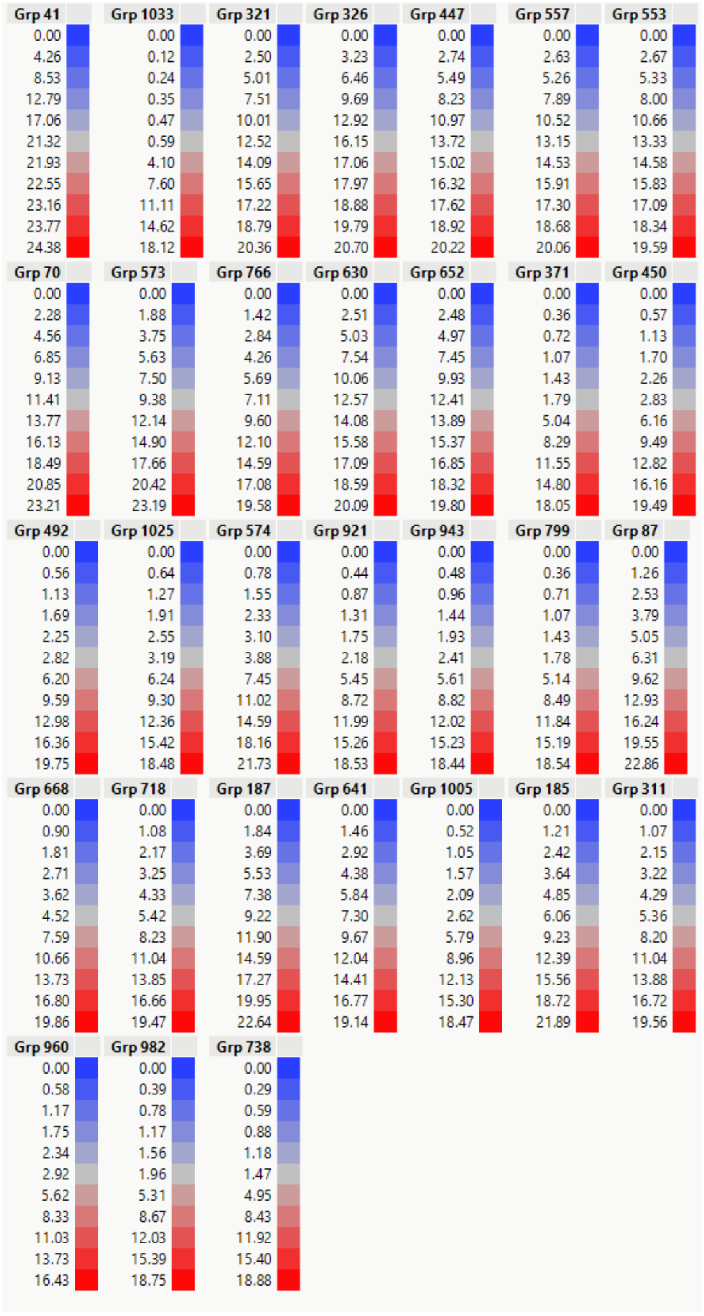
Supplementary Figures

Supplementary Figure 3.1 Hierarchical clustering of QTL1 metabolite traits. A two-way clustering of QTL1 metabolite traits and ecotypes based on the geographic region **(A)**. Metabolite accumulation pattern represented by a color scale, with high values shown in red and low values shown in blue **(B).**

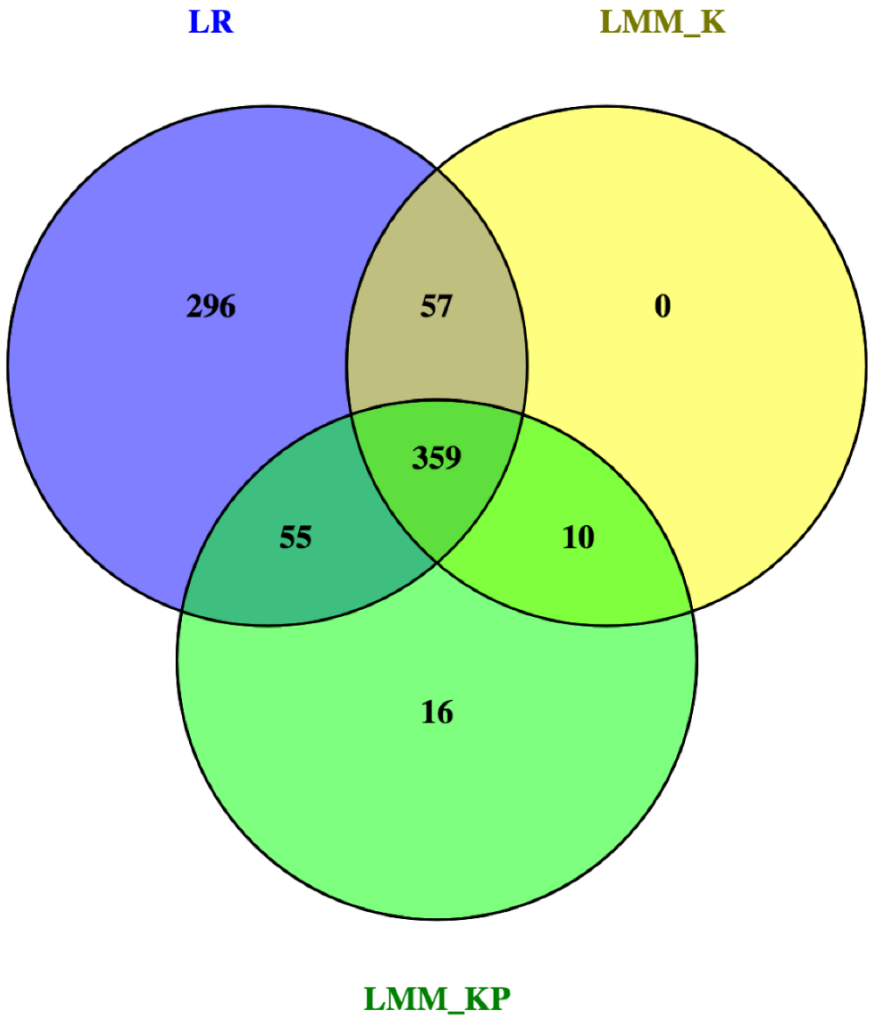
A.



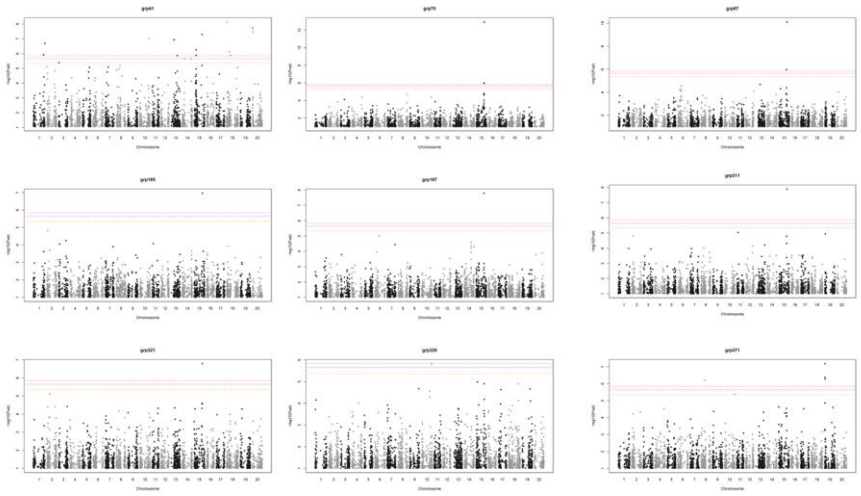
B.

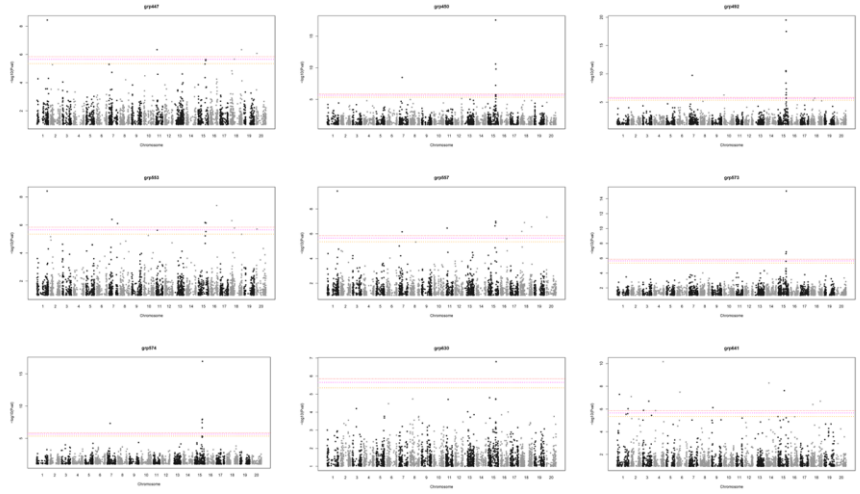

Grp 41
0.00
4.26
8.53
12.79
17.06
21.32
21.93
22.55
23.16
23.77
24.38
Grp 1033
0.00
0.12
0.24
0.35
0.47
0.59
4.10
7.60
11.11
14.62
18.12
Grp 321
0.00
2.50
5.01
7.51
10.01
12.52
14.09
15.65
17.22
18.79
20.36
Grp 326
0.00
3.23
6.46
9.69
12.92
16.15
17.06
17.97
18.88
19.79
20.70
Grp 447
0.00
2.74
5.49
8.23
10.97
13.72
15.02
16.32
17.62
18.92
20.22
Grp 557
0.00
2.63
5.26
7.89
10.52
13.15
14.53
15.91
17.30
18.68
20.06
Grp 553
0.00
2.67
5.33
8.00
10.66
13.33
14.58
15.83
17.09
18.34
19.59
Grp 70
0.00
2.28
4.56
6.85
9.13
11.41
13.77
16.13
18.49
20.85
23.21
Grp 573
0.00
1.88
3.75
5.63
7.50
9.38
12.14
14.90
17.66
20.42
23.19
Grp 766
0.00
1.42
2.84
4.26
5.69
7.11
9.60
12.10
14.59
17.08
19.58
Grp 630
0.00
2.51
5.03
7.54
10.06
12.57
14.08
15.58
17.09
18.59
20.09
Grp 652
0.00
2.48
4.97
7.45
9.93
12.41
13.89
15.37
16.85
18.32
19.80
Grp 371
0.00
0.36
0.72
1.07
1.43
1.79
5.04
8.29
11.55
14.80
18.05
Grp 450
0.00
0.57
1.13
1.70
2.26
2.83
6.16
9.49
12.82
16.16
19.49
Grp 492
0.00
0.56
1.13
1.69
2.25
2.82
6.20
9.59
12.98
16.36
19.75
Grp 1025
0.00
0.64
1.27
1.91
2.55
3.19
6.24
9.30
12.36
15.42
18.48
Grp 574
0.00
0.78
1.55
2.33
3.10
3.88
7.45
11.02
14.59
18.16
21.73
Grp 921
0.00
0.44
0.87
1.31
1.75
2.18
5.45
8.72
11.99
15.26
18.53
Grp 943
0.00
0.48
0.96
1.44
1.93
2.41
5.61
8.82
12.02
15.23
18.44
Grp 799
0.00
0.36
0.71
1.07
1.43
1.78
5.14
8.49
11.84
15.19
18.54
Grp 87
0.00
1.26
2.53
3.79
5.05
6.31
9.62
12.93
16.24
19.55
22.86
Grp 668
0.00
0.90
1.81
2.71
3.62
4.52
7.59
10.66
13.73
16.80
19.86
Grp 718
0.00
1.08
2.17
3.25
4.33
5.42
8.23
11.04
13.85
16.66
19.47
Grp 187
0.00
1.84
3.69
5.53
7.38
9.22
11.90
14.59
17.27
19.95
22.64
Grp 641
0.00
1.46
2.92
4.38
5.84
7.30
9.67
12.04
14.41
16.77
19.14
Grp 1005
0.00
0.52
1.05
1.57
2.09
2.62
5.79
8.96
12.13
15.30
18.47
Grp 185
0.00
1.21
2.42
3.64
4.85
6.06
9.23
12.39
15.56
18.72
21.89
Grp 311
0.00
1.07
2.15
3.22
4.29
5.36
8.20
11.04
13.88
16.72
19.56
Grp 960
0.00
0.58
1.17
1.75
2.34
2.92
5.62
8.33
11.03
13.73
16.43
Grp 982
0.00
0.39
0.78
1.17
1.56
1.96
5.31
8.67
12.03
15.39
18.75
Grp 738
0.00
0.29
0.59
0.88
1.18
1.47
4.95
8.43
11.92
15.40
18.88

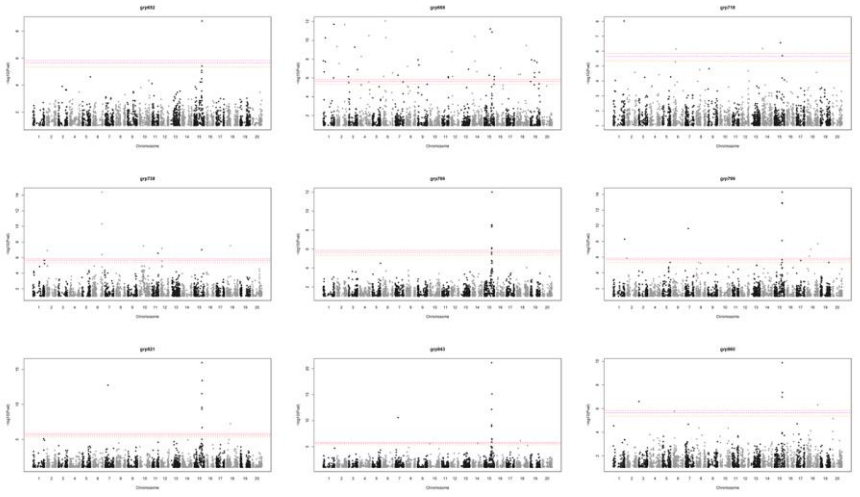
Supplementary Figure 3.2 The Venn diagram illustrates the results obtained from three different methods: LR, LMM with K, and LMM with K + P. A total of 727 metabolite peaks were identified using the LR method, 426 peaks using LMM with K, and 440 peaks using LMM with K + P. There were 359 metabolite traits that showed significance across all three methods, indicating a robust and consistent association.

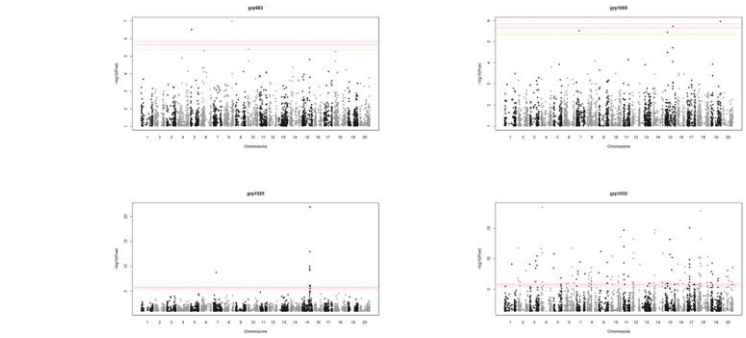


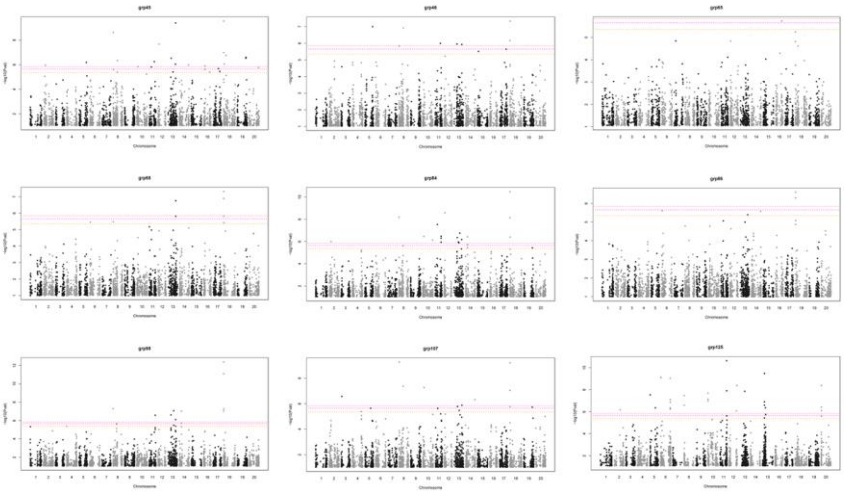
Supplementary Figure 3.3 The Manhattan plots depict the significant loci identified for the QTL-multiple metabolite cluster 1 (QTL1). QTL1 encompasses a diverse set of metabolite groups, including Grp 41, Grp 70, Grp 87, Grp 185, Grp 187, Grp 311, Grp 321, Grp 326, Grp 371, Grp 447, Grp 450, Grp 492, Grp 553, Grp 557, Grp 573, Grp 574, Grp 630, Grp 641, Grp 652, Grp 668, Grp 718, Grp 738, Grp 766, Grp 799, Grp 921, Grp 943, Grp 960, Grp 982, Grp 1005, Grp 1025, and Grp 1033. These plots provide a visual representation of the genetic association results, highlighting the significance and distribution of the identified loci within the QTL1 cluster.

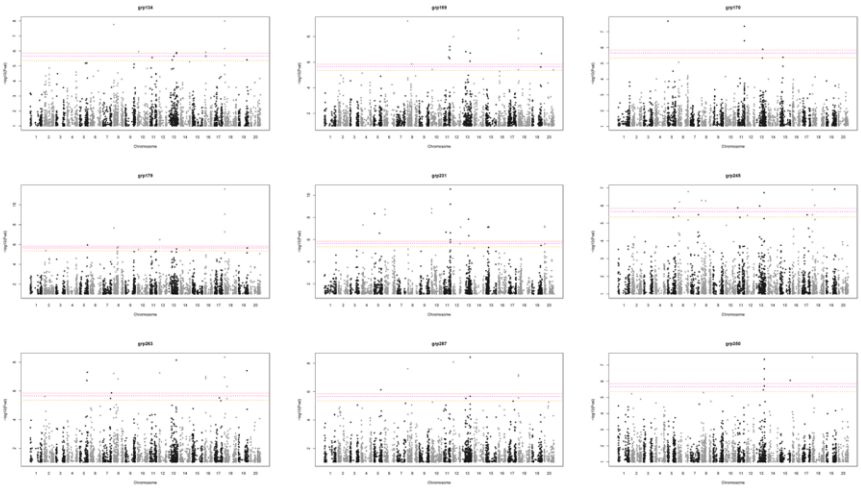


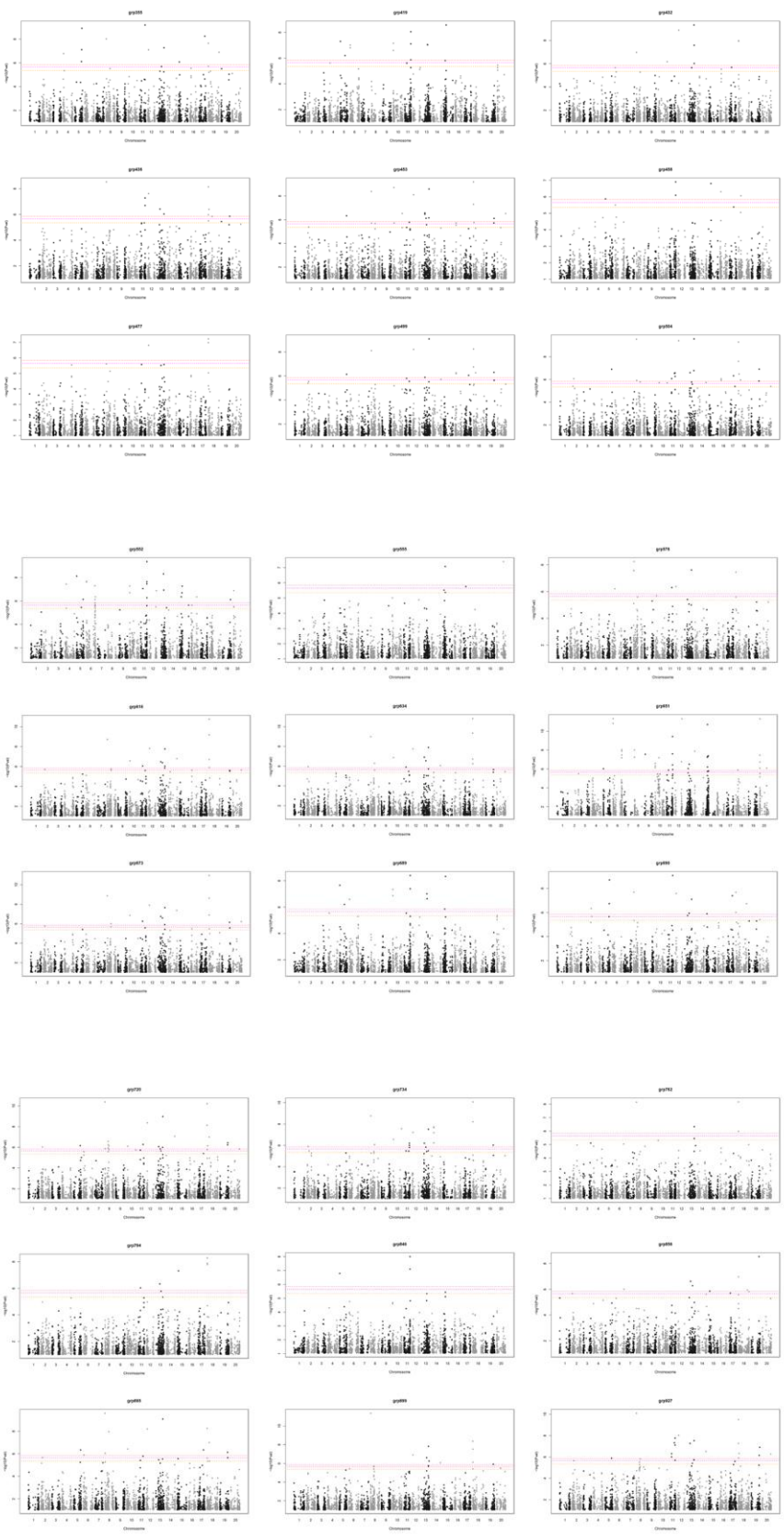


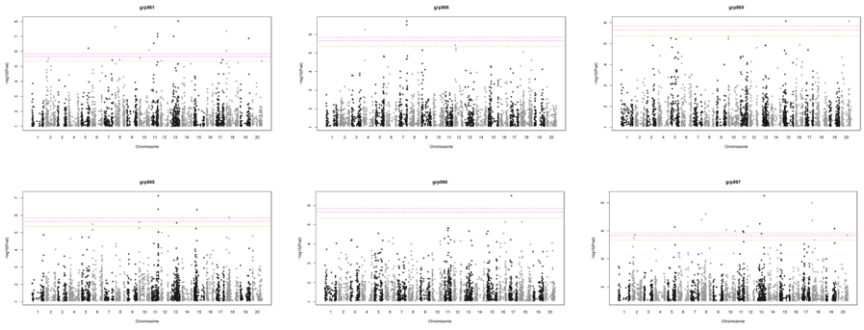


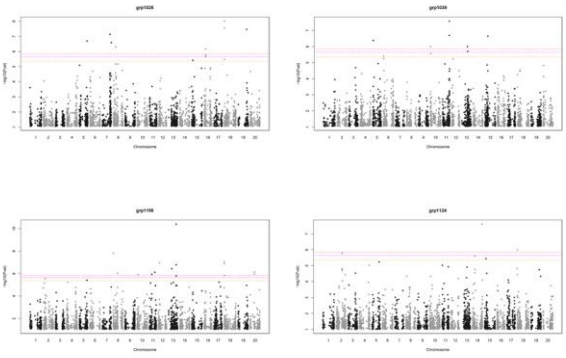


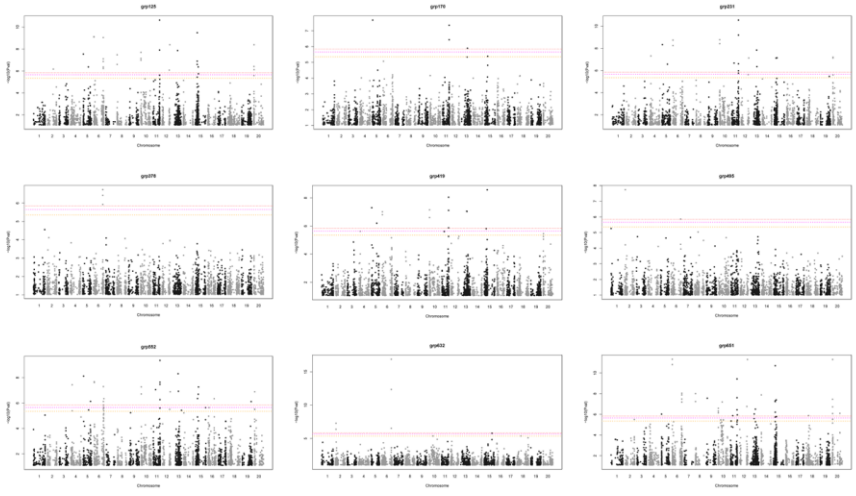


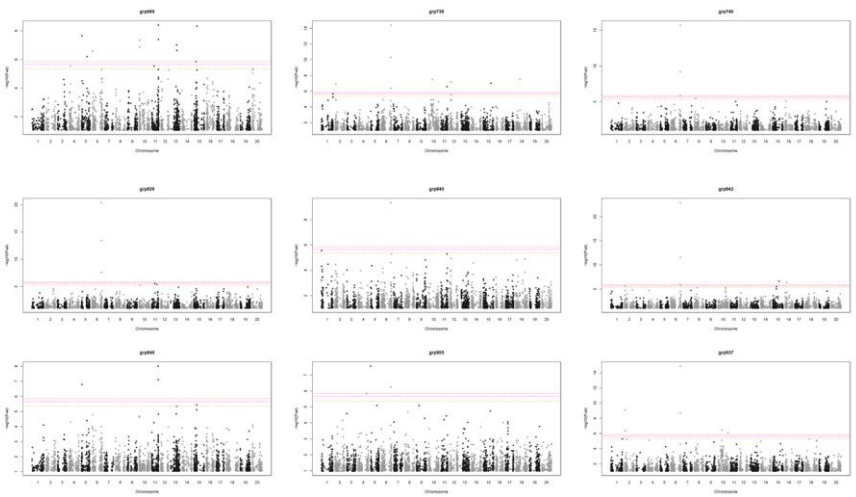


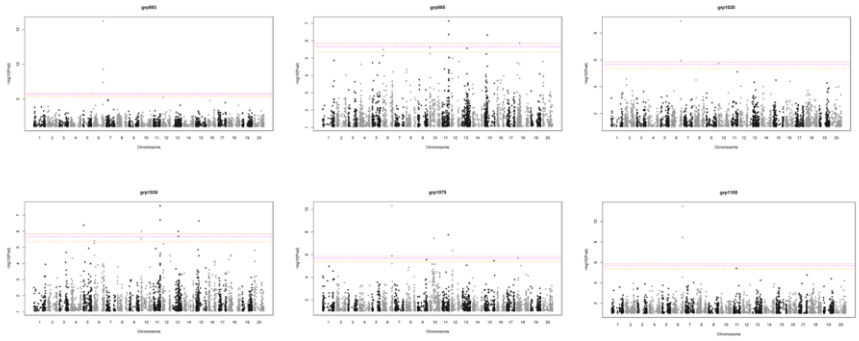


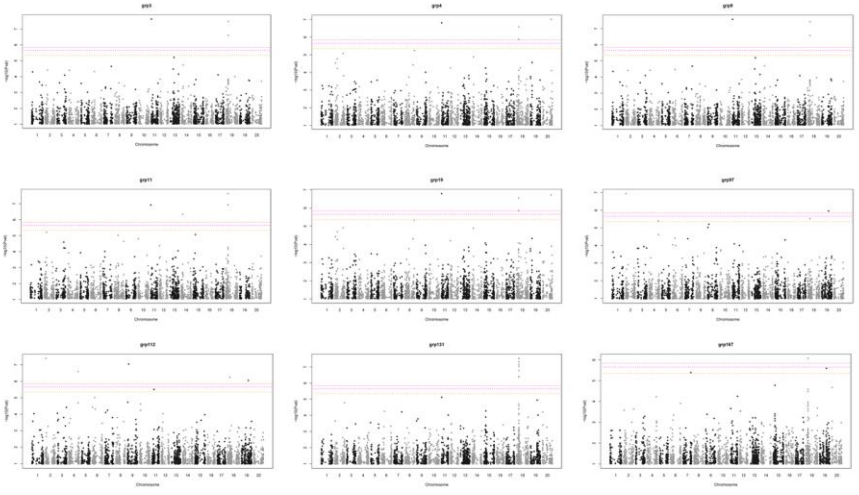


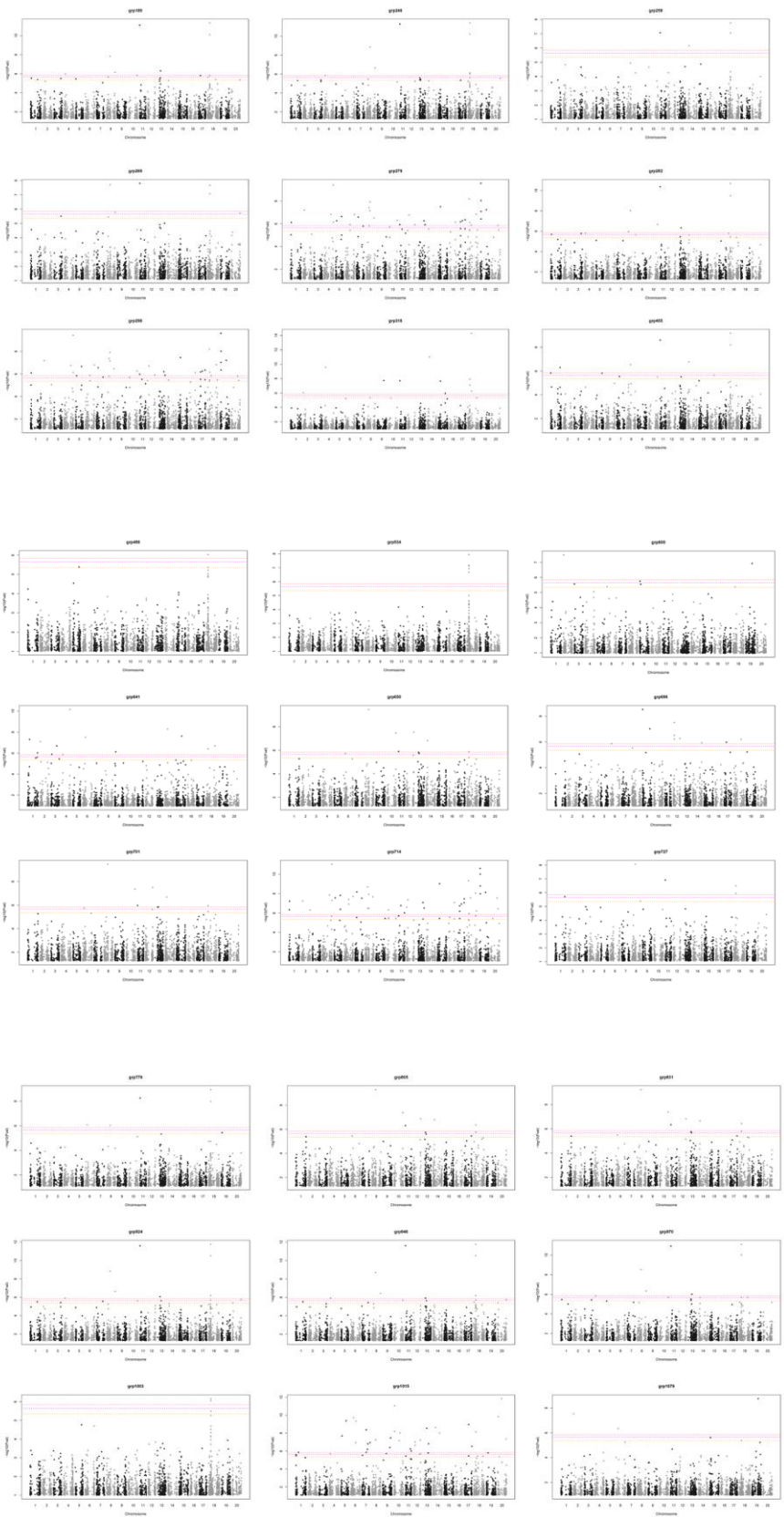


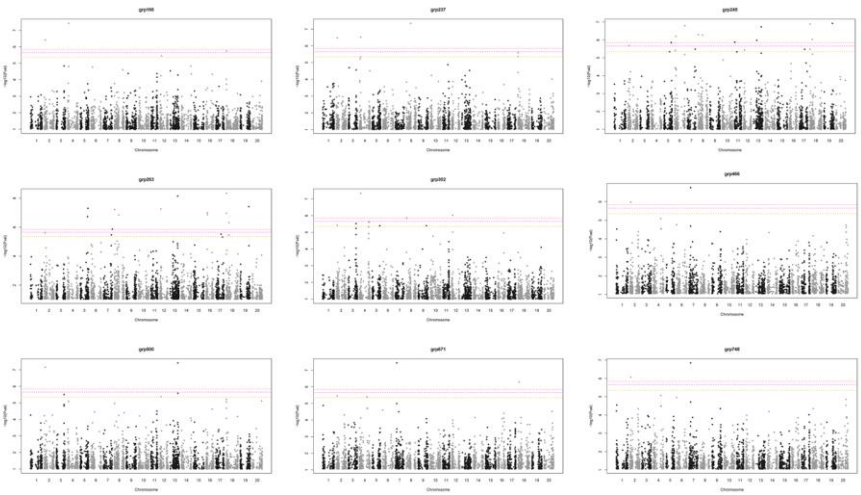


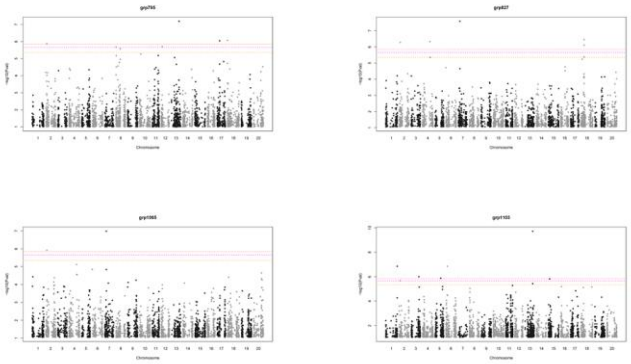


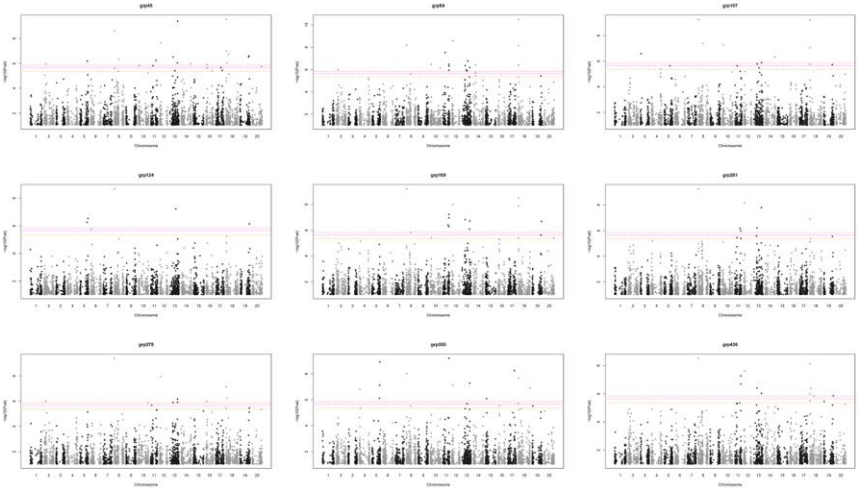


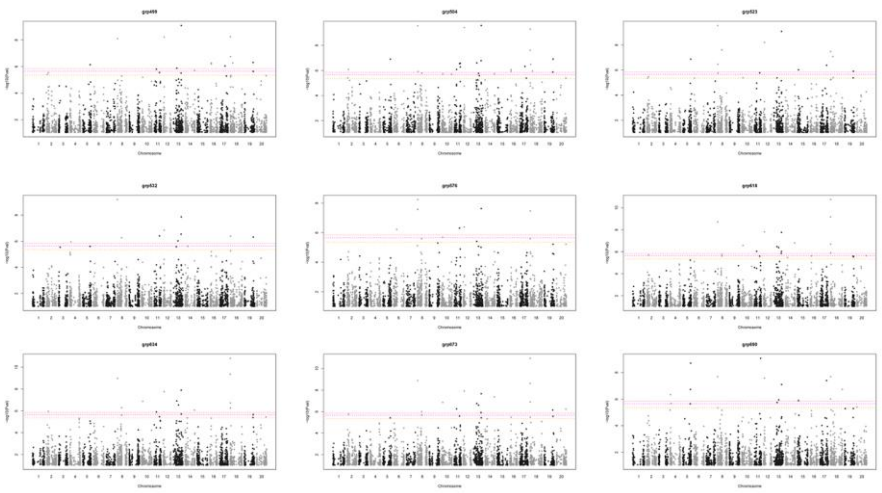


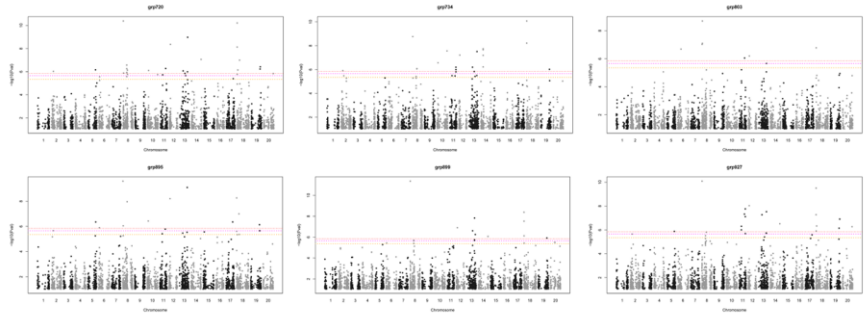


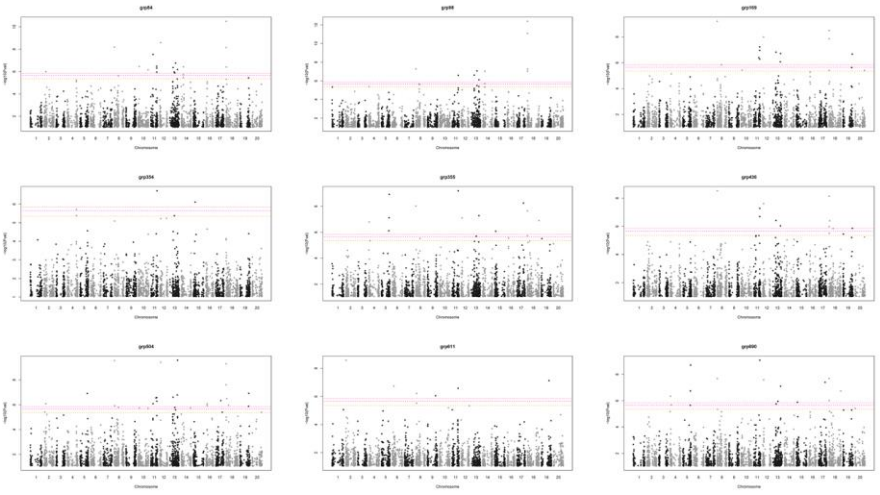


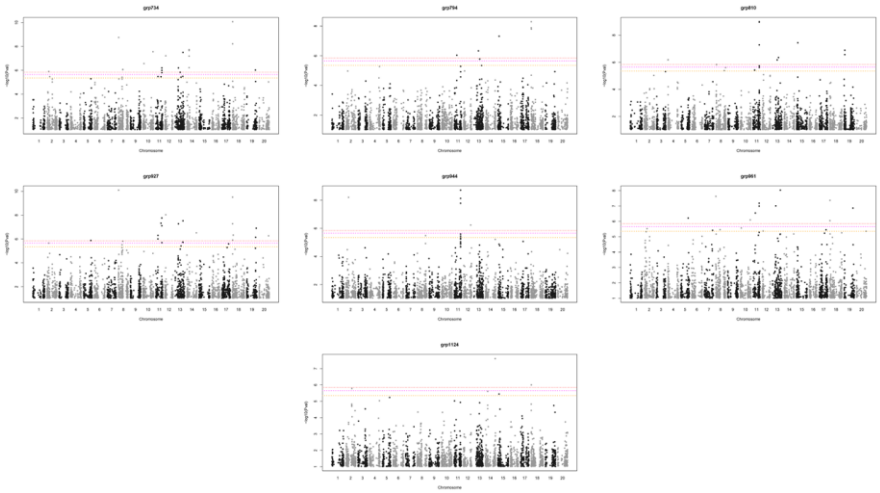


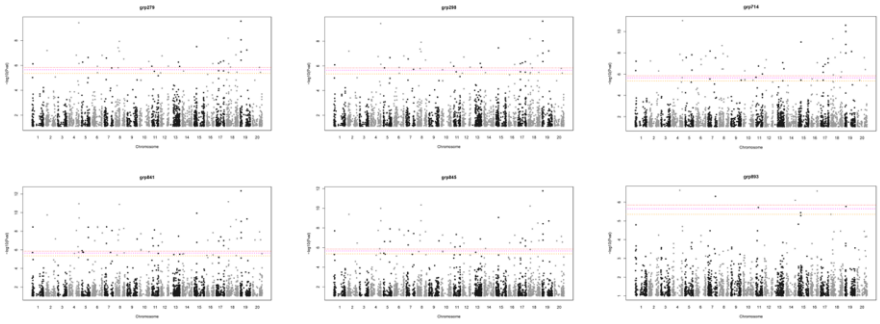


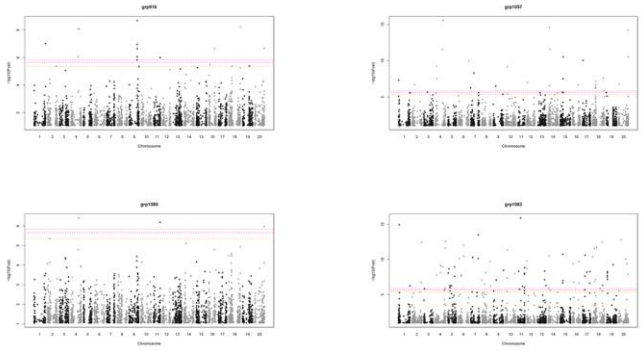




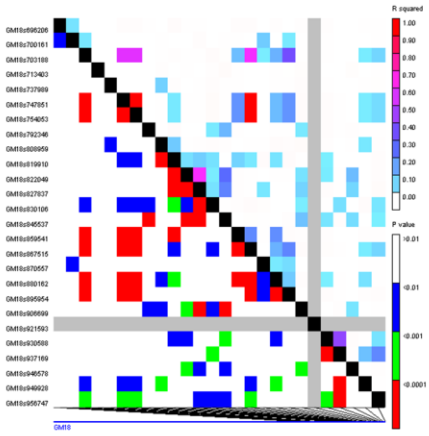



grp734
grp794
grp810
grp827
grp844
grp861
grp1124
Chromosome

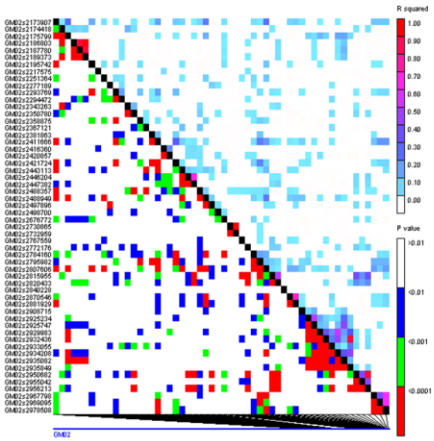

grp379
grp398
grp714
grp841
grp845
grp863
Chromosome



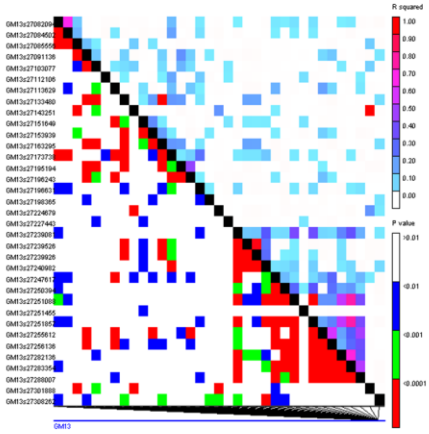
Supplementary Figure 3.4 Linkage disequilibrium (LD) plots highlight significance of SNPs for single SNP-multiple metabolite clusters.



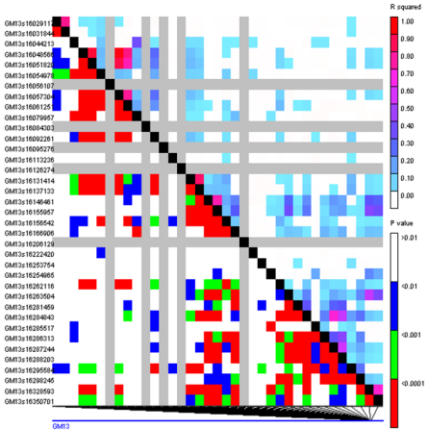
Gm18s867515



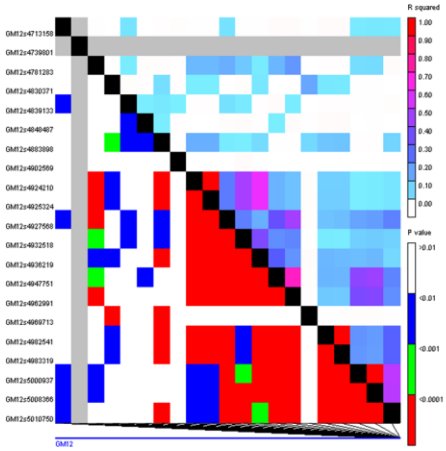
Gm13s37591009



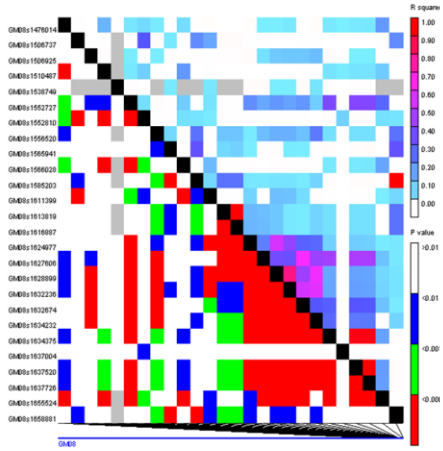
Gm13s27173738



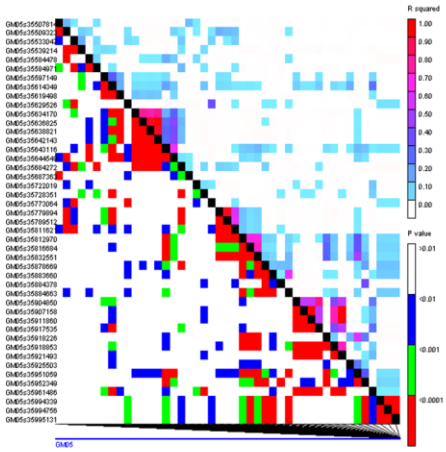
Gm13s16155957



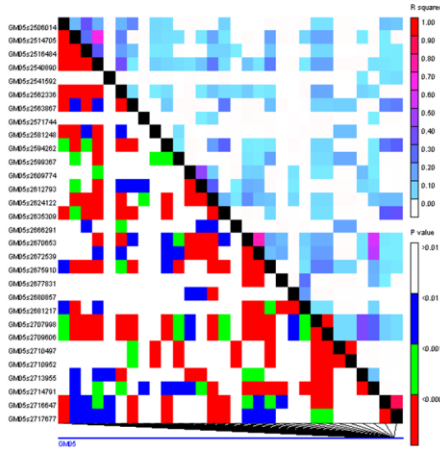
Gm12s4962991



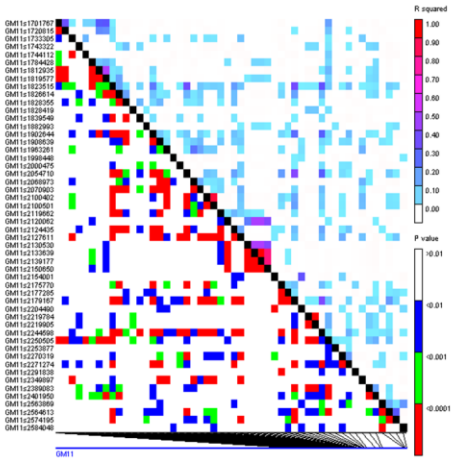
Gm08s1552810



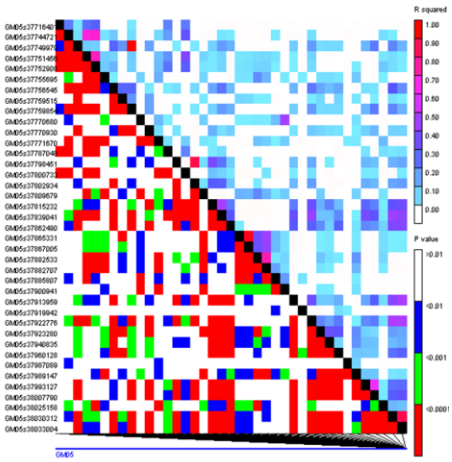
Gm05s35812970



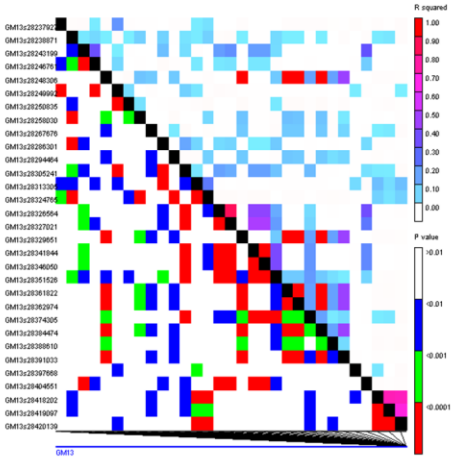
Gm05s2563867



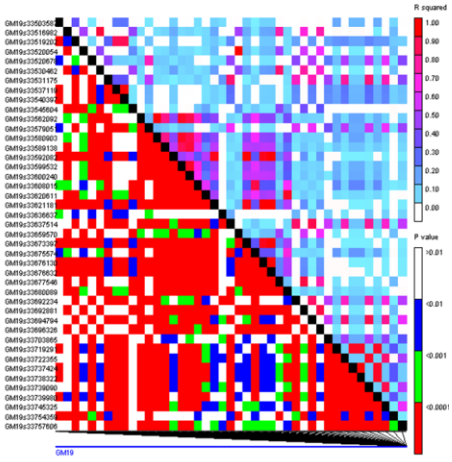
Gm11s2000475



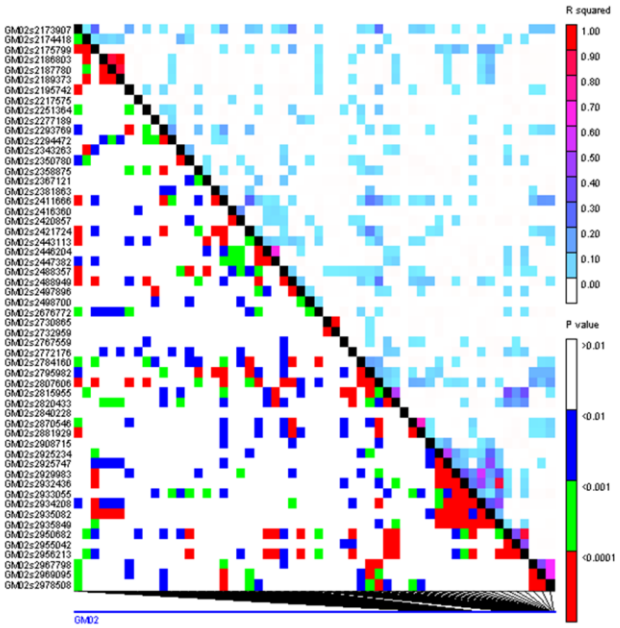
Gm05s37919942



Gm13s28326564

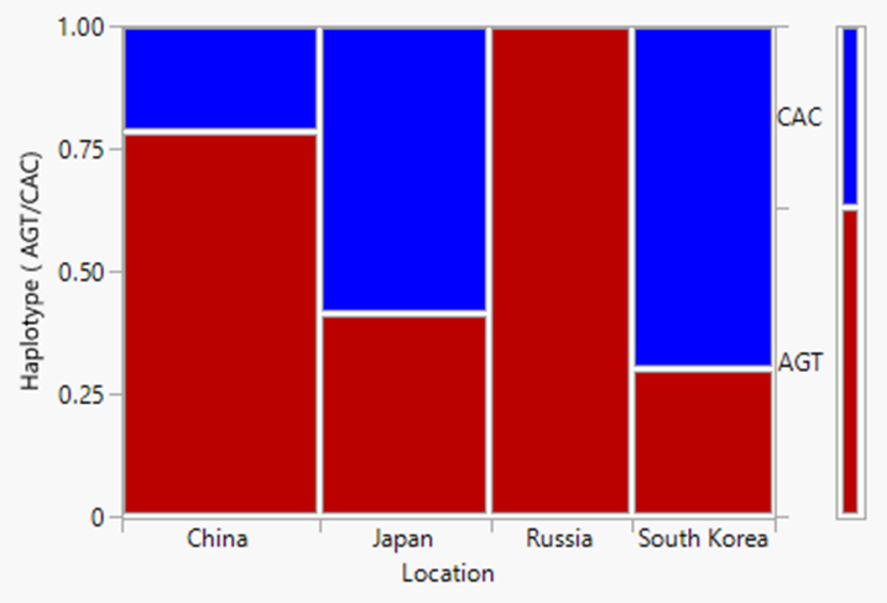


Gm19s33692234

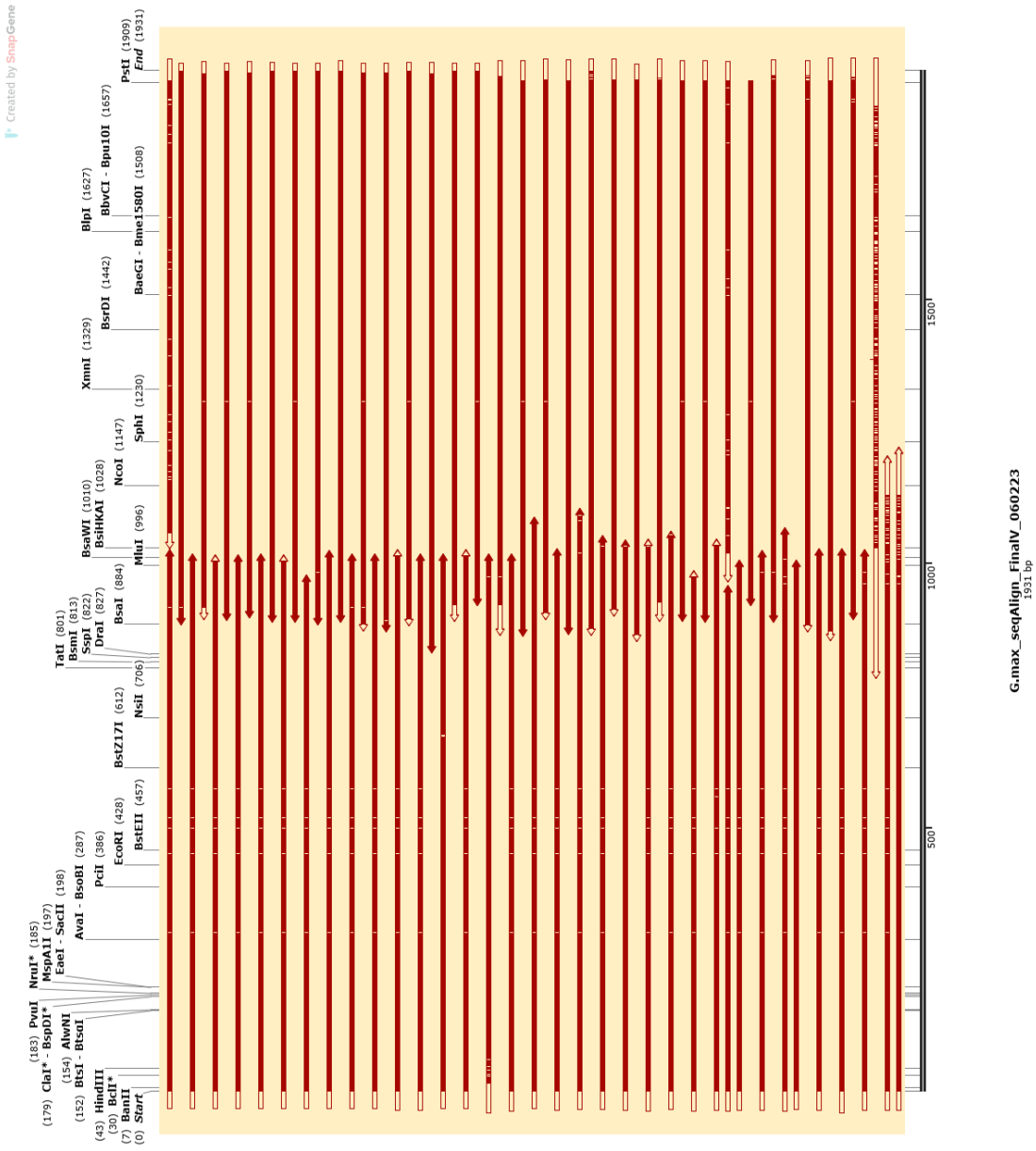


Gm02s2411666

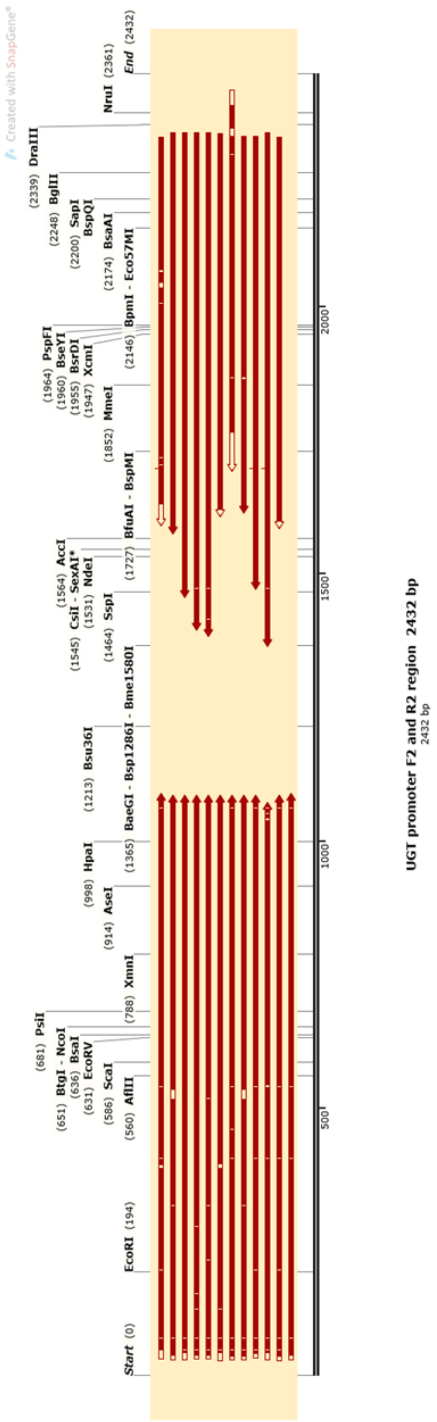
Supplementary Figure 3.5 Haplotype frequency of different *G. soja* ecotypes in different geographic regions.



Supplementary Figure 3.6 The sequence alignment map in this figure illustrates the variation observed in the UDP-glucosyl transferase *(UGT)* gene among 34 different *G. max* cultivars. The map highlights the sequence variations present within the *UGT* gene across these cultivars.



Supplementary Figure 3.7 A sequence alignment map was generated for the promoter region of the *UGT* gene, comparing 12 *G. soja* ecotypes and revealing sequence variations.



Supplementary Tables

Supplementary Table 3.1 The plant ID and geographic origins used for the mGWAS study.

ID	PI #	Country	Maturity group
Gs_1	PI101404 B	China	II
Gs_10	PI366122	Japan	IV
Gs_1001	PI479752	Jilin Sheng, China	x
Gs_101	PI464868B	Heilongjiang Sheng, China	x
Gs_103	PI464889 B	Jilin Sheng, China	x
Gs_104	PI464889 C	China	II
Gs_105	PI464890 A	China	II
Gs_106	PI464891 B	China	II
Gs_107	PI464925 C	China	I
Gs_108	PI464926	China	0
Gs_109	PI464927 A	China	0
Gs_11	PI366123	Japan	IV
Gs_110	PI464927 B	China	0
Gs_111	PI464928	China	0
Gs_113	PI468396 B	China	IV
Gs_114	PI468397 A	China	IV
Gs_116	PI464938	Jiangsu Sheng, China	x
Gs_117	PI464936 B	China	VI
Gs_118	PI464937 A	China	VI
Gs_119	PI468398 B	China	IV
Gs_120	PI468399 A	Shandong Sheng, China	x
Gs_121	PI468399 B	China	IV
Gs_122	PI468918	China	III
Gs_123	PI479745	China	I
Gs_124	PI479746 B	China	II
Gs_125	PI479749	China	III
Gs_126	PI479750	China	I
Gs_127	PI479751	China	III
Gs_128	PI483466	China	V
Gs_129	PI483467	Henan Sheng, China	x
Gs_130	PI483468 A	China	V
Gs_132	PI487428	Japan	V
Gs_133	PI487430	Japan	V
Gs_134	PI487431	Japan	IX
Gs_135	PI504287 A	Japan	IV
Gs_136	PI507582	Japan	V
Gs_139	PI507632	Japan	VII
Gs_14	PI378686 A	Japan	VII
Gs_140	PI507644	Japan	VI
Gs_143	PI507667	Japan	VI
Gs_146	PI507722	Russian Federation	0
Gs_147	PI507723 B	Russian Federation	II
Gs_148	PI507725 B	Russian Federation	0
Gs_149	PI507727	Russian Federation	0
Gs_15	PI378691	Japan	VII
Gs_150	PI507729	Russian Federation	0
Gs_151	PI507730	Russian Federation	0
Gs_152	PI507731	Russian Federation	0

ID	PI #	Country	Maturity group
Gs_153	PI507734	Russian Federation	0
Gs_154	PI507735	Russian Federation	0
Gs_155	PI507736	Amur, Russian Federation	
Gs_156	PI507738	Russian Federation	0
Gs_157	PI507739 B	Russian Federation	0
Gs_158	PI507740	Russian Federation	0
Gs_159	PI507742	Russian Federation	0
Gs_16	PI378695 A	Japan	VI
Gs_161	PI507749	Russian Federation	0
Gs_163	PI507757	Russian Federation	0
Gs_164	PI507759	Russian Federation	0
Gs_165	PI507760	Russian Federation	0
Gs_167	PI507764	Russian Federation	0
Gs_168	PI507774	Russian Federation	II
Gs_169	PI507776	Russian Federation	I
Gs_17	PI378698	Japan	VI
Gs_170	PI507777	Russian Federation	I
Gs_171	PI507780	Primorye, Russian Federation	
Gs_172	PI507782	Russian Federation	0
Gs_174	PI507784	Russian Federation	II
Gs_175	PI507787	Russian Federation	II
Gs_176	PI507788	Russian Federation	III
Gs_177	PI507794	Russian Federation	I
Gs_178	PI507798	Russian Federation	II
Gs_179	PI507799	Russian Federation	I
Gs_18	PI378701 A	Japan	V
Gs_180	PI507803	Russian Federation	0
Gs_181	PI507805	Russian Federation	0
Gs_182	PI507806 A	Russian Federation	0
Gs_183	PI507808	Russian Federation	
Gs_184	PI507814	Russian Federation	0
Gs_185	PI507815	Amur, Russian Federation	
Gs_186	PI507816	Russian Federation	0
Gs_187	PI507818 B	Russian Federation	0
Gs_188	PI507821	Russian Federation	0
Gs_189	PI507826	Russian Federation	0
Gs_190	PI507830 B	Russian Federation	0
Gs_191	PI507833	Russian Federation	0
Gs_192	PI507836	Russian Federation	0
Gs_193	PI507839	Russian Federation	0
Gs_194	PI507841 B	Amur, Russian Federation	
Gs_195	PI507847	Russian Federation	II
Gs_197	PI508066	Japan	IV
Gs_199	PI508069	Japan	IV
Gs_2	PI163453	China	VII
Gs_20	PI406684	Japan	III
Gs_201	PI522179	China	0
Gs_202	PI522180	China	0
Gs_203	PI522182 A	China	0
Gs_204	PI522182 B	China	I

ID	PI #	Country	Maturity group
Gs_205	PI522193	Russian Federation	0
Gs_206	PI522194 B	Primorye, Russian Federation	
Gs_207	PI522196 A	Russian Federation	0
Gs_208	PI522198 A	Russian Federation	I
Gs_209	PI522200 A	Russian Federation	II
Gs_209332	PI209332	Hokkaidô, Japan	
Gs_21	PI407030	Akita, Japan	
Gs_210	PI522211 B	Russian Federation	III
Gs_211	PI522212 B	Primorye, Russian Federation	
Gs_213	PI522217	Russian Federation	II
Gs_214	PI522223	Russian Federation	II
Gs_215	PI522226	Russian Federation	II
Gs_216	PI522227	Russian Federation	0
Gs_217	PI522229	Russian Federation	0
Gs_218	PI522230 A	Russian Federation	0
Gs_219	PI522234	Russian Federation	I
Gs_220	PI522235 A	Russian Federation	I
Gs_221	PI532453 A	China	III
Gs_222	PI538411 A	Amur, Russian Federation	
Gs_223	PI549037	China	III
Gs_224	PI549046	China	IV
Gs_226	PI562544	Korea, South	Unkown
Gs_227	PI562550	Korea, South	Unkown
Gs_23	PI407034	Japan	V
Gs_231	PI578338 A	Russian Federation	Unkown
Gs_232	PI578345	Russian Federation	Unkown
Gs_25	PI407037	Japan	V
Gs_27	PI407047	Japan	V
Gs_28	PI407050	Japan	V
Gs_29	PI407053	Japan	VI
Gs_3	PI326581	Russian Federation	
Gs_32	PI407089	Japan	VI
Gs_33	PI407097	Japan	VI
Gs_34	PI407115	Hyôgo, Japan	
Gs_35	PI407120	Japan	VII
Gs_36	PI407124	Hyôgo, Japan	
Gs_38	PI407162	Kyonggi, Korea, South	
Gs_39	PI407167	Korea, South	V
Gs_4	PI326582 A	Russian Federation	II
Gs_40	PI407174	Korea, South	V
Gs_43	PI407198	Korea, South	V
Gs_437654	PI437654	China	
Gs_44	PI407200	Korea, South	IV
Gs_45	PI407201	Korea, South	V
Gs_46	PI407202	Korea, South	V
Gs_48	PI407217	Korea, South	IV
Gs_49	PI407221	Korea, South	V
Gs_5	PI339732	Korea, South	IV
Gs_50	PI407229	Korea, South	V
Gs_53	PI407246	Korea, South	V

ID	PI #	Country	Maturity group
Gs_54	PI407249	Korea, South	V
Gs_548316	PI548316	Zhejiang Sheng, China	
Gs_55	PI407254	Korea, South	VI
Gs_57	PI407271	Korea, South	V
Gs_58	PI407275	Korea, South	IV
Gs_59	PI407278	Korea, South	IV
Gs_6	PI339871 A	Korea, South	V
Gs_60	PI407279	Jeju-teukbyeoljachido, Korea, South	
Gs_62	PI407296	China	II
Gs_64	PI407298	China	II
Gs_65	PI407302	China	V
Gs_66	PI407304	China	VI
Gs_68	PI423988	Amur, Russian Federation	
Gs_69	PI423990 A	Russian Federation	0
Gs_7	PI342621 C	Russian Federation	0
Gs_72	PI423993	Russian Federation	0
Gs_73	PI423995	Russian Federation	0
Gs_75	PI423996	Amur, Russian Federation	
Gs_76	PI423997	Russian Federation	0
Gs_78	PI423999 B	Russian Federation	0
Gs_8	PI366120	Japan	IV
Gs_80	PI424000	Russian Federation	0
Gs_81	PI424001	Russian Federation	0
Gs_83	PI424002	Russian Federation	0
Gs_85	PI424032	Korea, South	IV
Gs_86	PI424059 B	Korea, South	V
Gs_87	PI424063	Korea, South	IV
Gs_88	PI424064	Korea, South	V
Gs_88788	PI088788	Liaoning Sheng, China	
Gs_89	PI424088	Korea, South	IV
Gs_89772	PI089772	China	
Gs_9	PI366121	Hukusima, Japan	
Gs_90763	PI090763	Beijing Shi, China	
Gs_91	PI424093	Korea, South	V
Gs_92	PI424096	Korea, South	V
Gs_93	PI424102 A	Korea, South	V
Gs_94	PI424117	Korea, South	V
Gs_95	PI458537 A	China	0
Gs_96	PI458538	Heilongjiang Sheng, China	
Gs_97	PI458539 B	Heilongjiang Sheng, China	
Gs_98	PI458540 D	China	0
Gs_99	PI464866 A	China	0
Gs_essex	PI548667	Virginia, United States 'Essex'	
Gs_hut	PI518664		
Gs_lee74	PI548658	Arkansas, United States 'Lee 74'	
Gs_peking	PI548402	Beijing Shi, China 'Peking'	
Gs_w82	PI518671	Illinois, United States 'Williams 82'	

Supplementary Table 3.2 Detailed information of the peak annotation for all the compounds.

ID	Mass	RT	hits	delta_ppm	Formula	NeutralMass	Cpd	KEGG	CAS	ChemSpider
4	382.126	10.665	1	-13.1531568	C15H28O7P2	382.1310262	(2E,6E)-farnesyl diphosphate	C00448	############	11633047
7	356.0733	4.726	1	-0.0411786	C9H17N4O9P	356.0733147	5-amino-6-(5-phospho-D-ribitylamino)uracil	C04454		NA
11	368.1106	9.631	2	-0.3592362	C17H20O9	368.1107322	O-feruloylquinate	C02572		NA
11	368.1106	9.631	2	17.3670479	C16H20N2O6S	368.1042071	indolylmethyl-desulfoglucosinolate	C16517		NA
12	626.1489	7.409	4	0.95898033	C27H30O17	626.1482995	delphinidin 3,5-di-O-beta-D-glucoside	C16312		NA
12	626.1489	7.409	4	0.95898033	C27H30O17	626.1482995	delphinidin 3-O-sophoroside	C16307		NA
12	626.1489	7.409	4	0.95898033	C27H30O17	626.1482995	quercetin 3-O-sophoroside	C12667	18609-17-1	NA
12	626.1489	7.409	4	0.95898033	C27H30O17	626.1482995	Quercetin 3-gentiobioside		7431-83-6	NA
13	610.1538	7.834	8	0.68029812	C27H30O16	610.1533849	cyanidin 3,5-di-O-beta-D-glucoside	C08639		NA
13	610.1538	7.834	8	0.68029812	C27H30O16	610.1533849	delphinidin 3-O-rutinoside			NA
13	610.1538	7.834	8	0.68029812	C27H30O16	610.1533849	cyanidin 3,7-di-O-beta-D-glucoside			NA
13	610.1538	7.834	8	0.68029812	C27H30O16	610.1533849	cyanidin 3-O-sophoroside	C16306		NA
13	610.1538	7.834	8	0.68029812	C27H30O16	610.1533849	quercetin 3-O-rhamnoside-7-O-glucoside			NA
13	610.1538	7.834	8	0.68029812	C27H30O16	610.1533849	kaempferol 3-O-beta-D-glucosyl-(1->2)-glucoside	C12634	19895-95-5	NA
13	610.1538	7.834	8	0.68029812	C27H30O16	610.1533849	Kaempferol 3-O-gentiobioside		22149-35-5	NA
13	610.1538	7.834	8	0.68029812	C27H30O16	610.1533849	Quercetin 3-O-neohesperidoside		32453-36-4	NA
17	356.0745	6.507	1	3.32891388	C9H17N4O9P	356.0733147	5-amino-6-(5-phospho-D-ribitylamino)uracil	C04454		NA
19	382.1257	10.744	1	-13.9382278	C15H28O7P2	382.1310262	(2E,6E)-farnesyl diphosphate	C00448	############	11633047
20	594.1593	8.299	4	1.39644311	C27H30O15	594.1584703	pelargonidin-3,5-di-O-beta-D-glucoside	C08725		NA
20	594.1593	8.299	4	1.39644311	C27H30O15	594.1584703	pelargonidin 3-O-sophoroside	C16305		NA
20	594.1593	8.299	4	1.39644311	C27H30O15	594.1584703	kaempferol 3-O-rhamnoside-7-O-glucoside			NA
20	594.1593	8.299	4	1.39644311	C27H30O15	594.1584703	Kaempferol 3-neohesperidoside		32602-81-6	NA
22	772.2065	6.918	4	0.37769147	C33H40O21	772.2062083	delphinidin 3-O-rutinoside-7-O-glucoside			NA
22	772.2065	6.918	4	0.37769147	C33H40O21	772.2062083	Kaempferol 3-glucosyl-(1->2)-gentiobioside		55696-59-8	NA
22	772.2065	6.918	4	0.37769147	C33H40O21	772.2062083	Quercetin 3-(2G-glucosylrutinoside)		55696-55-4	NA
22	772.2065	6.918	4	0.37769147	C33H40O21	772.2062083	Quercetin 3-(2G-rhamnosylgentiobioside)		55780-29-5	NA
23	464.0958	8.077	5	0.69790467	C21H20O12	464.0954761	quercetin 4'-O-glucoside			NA
23	464.0958	8.077	5	0.69790467	C21H20O12	464.0954761	quercetin 3'-O-glucoside			NA
23	464.0958	8.077	5	0.69790467	C21H20O12	464.0954761	delphinidin-3-O-beta-D-glucoside	C12138		NA
23	464.0958	8.077	5	0.69790467	C21H20O12	464.0954761	quercetin-3-glucoside	C05623		NA
23	464.0958	8.077	5	0.69790467	C21H20O12	464.0954761	quercetin 7-O-glucoside			NA
25	356.074	6.243	1	1.92470868	C9H17N4O9P	356.0733147	5-amino-6-(5-phospho-D-ribitylamino)uracil	C04454		NA
27	740.2177	7.563	2	1.78447862	C33H40O19	740.2163791	pelargonidin-3-O-rutinoside-5-O-beta-D-glucoside	C12645		NA
27	740.2177	7.563	2	1.78447862	C33H40O19	740.2163791	Clitorin		55804-74-5	NA
33	942.5207	12.516	1	1.99922842	C48H78O18	942.5188157	soyasaponin I		51330-27-9	NA
34	796.4605	12.925	1	-0.51086173	C42H68O14	796.4609069	soyasaponin III			NA
36	190.083	7.584	1	-5.91084662	C8H14O5	190.0841236	(R)-3-((R)-3-hydroxybutanoyloxy)-butanoate			NA
46	448.0996	8.559	6	-2.14568644	C21H20O11	448.1005615	isoorientin			NA
46	448.0996	8.559	6	-2.14568644	C21H20O11	448.1005615	cyanidin-3-O-beta-D-galactoside			NA
46	448.0996	8.559	6	-2.14568644	C21H20O11	448.1005615	cyanidin-3-O-beta-D-glucoside	C08604	7084-24-4	NA
46	448.0996	8.559	6	-2.14568644	C21H20O11	448.1005615	orientin			NA
46	448.0996	8.559	6	-2.14568644	C21H20O11	448.1005615	quercetin 3-O-rhamnoside			NA
46	448.0996	8.559	6	-2.14568644	C21H20O11	448.1005615	kaempferol-3-glucoside	C12249		NA
51	594.1589	8.147	4	0.72322204	C27H30O15	594.1584703	pelargonidin-3,5-di-O-beta-D-glucoside	C08725		NA
51	594.1589	8.147	4	0.72322204	C27H30O15	594.1584703	pelargonidin 3-O-sophoroside	C16305		NA
51	594.1589	8.147	4	0.72322204	C27H30O15	594.1584703	kaempferol 3-O-rhamnoside-7-O-glucoside			NA
51	594.1589	8.147	4	0.72322204	C27H30O15	594.1584703	Kaempferol 3-neohesperidoside		32602-81-6	NA
55	356.0729	3.483	1	-1.16454276	C9H17N4O9P	356.0733147	5-amino-6-(5-phospho-D-ribitylamino)uracil	C04454		NA
57	268.0723	12.233	2	-4.69599466	C16H12O4	268.0735589	isoformononetin	C12125	486-63-5	3632
57	268.0723	12.233	2	-4.69599466	C16H12O4	268.0735589	formononetin	C00858	485-72-3	4444070
60	326.0993	8.109	4	-2.66038686	C15H18O8	326.1001676	cis-coumarinic acid-beta-D-glucoside	C05839		NA
60	326.0993	8.109	4	-2.66038686	C15H18O8	326.1001676	4-O-beta-D-glucosyl-4-hydroxycinnamate			NA
60	326.0993	8.109	4	-2.66038686	C15H18O8	326.1001676	trans-beta-D-glucosyl-2-hydroxycinnamate			NA
60	326.0993	8.109	4	-2.66038686	C15H18O8	326.1001676	1-O-(4-coumaroyl)-beta-D-glucose			NA
70	912.508	12.503	1	-0.27506984	C47H76O17	912.508251	Soyasaponin II		55319-36-3	NA
72	240.063	6.49	1	-1.6167255	C11H12O6	240.0633881	(1R,6R)-6-hydroxy-2-succinylcyclohexa-2,4-diene-1-carboxylate	C05817		NA
80	788.1965	6.574	3	-5.86521189	C33H40O22	788.201123	delphinidin 3,3',5-tri-O-beta-D-glucoside	C16314		NA
80	788.1965	6.574	3	-5.86521189	C33H40O22	788.201123	quercetin 3-O-beta-D-glucosyl-(1->2)-glucosyl-(1->2)-beta-D-glucoside			NA
80	788.1965	6.574	3	-5.86521189	C33H40O22	788.201123	Quercetin 3-(2G-glucosylgentiobioside)		55696-56-5	NA
81	300.0837	5.963	4	-2.72419519	C13H16O8	300.0845175	4-(beta-D-glucosyloxy)benzoate			NA
81	300.0837	5.963	4	-2.72419519	C13H16O8	300.0845175	salicylate beta-D-glucose ester			NA
81	300.0837	5.963	4	-2.72419519	C13H16O8	300.0845175	salicylate 2-O-beta-D-glucoside			NA
81	300.0837	5.963	4	-2.72419519	C13H16O8	300.0845175	1-O-4-hydroxybenzoyl-beta-D-glucose			NA

ID	Mass	RT	hits	delta_ppm	Formula	NeutralMass	Cpd	KEGG	CAS	ChemSpider
83	316.08	3.313	1	-2.948977	C13H16O9	316.0794321	2,5-dihydroxybenzoate 5-O-beta-D-glucoside			NA
84	432.1	8.261	7	-2.422697	C21H20O10	432.1056469	kaempferol-3-rhamnoside			NA
84	432.1	8.261	7	-2.422697	C21H20O10	432.1056469	isovitexin			NA
84	432.1	8.261	7	-2.422697	C21H20O10	432.1056469	luteolinidin 5-O-glucoside			NA
84	432.1	8.261	7	-2.422697	C21H20O10	432.1056469	pelargonidin-3-O-beta-D-glucoside	C12137		NA
84	432.1	8.261	7	-2.422697	C21H20O10	432.1056469	vitexin		3681-93-4	NA
84	432.1	8.261	7	-2.422697	C21H20O10	432.1056469	genistin	C09126	529-59-9	19265428
84	432.1	8.261	7	-2.422697	C21H20O10	432.1056469	Demethyltexasin 4'-O-glucoside		34307-23-8	NA
86	270.05	9.096	11	4.7271104	C15H10O5	270.0528234	2'-hydroxydaidzein	C02495		4444153
86	270.05	9.096	11	4.7271104	C15H10O5	270.0528234	6,7,4'-trihydroxyisoflavone	C14314		4447693
86	270.05	9.096	11	4.7271104	C15H10O5	270.0528234	quinol vinyl ether			NA
86	270.05	9.096	11	4.7271104	C15H10O5	270.0528234	phenoxy radical VII			NA
86	270.05	9.096	11	4.7271104	C15H10O5	270.0528234	luteolinidin			NA
86	270.05	9.096	11	4.7271104	C15H10O5	270.0528234	pelargonidin	C05904	0134-04-03	NA
86	270.05	9.096	11	4.7271104	C15H10O5	270.0528234	apigenin	C01477	520-36-5	NA
86	270.05	9.096	11	4.7271104	C15H10O5	270.0528234	baicalein	C10023		4444924
86	270.05	9.096	11	4.7271104	C15H10O5	270.0528234	genistein	C06563	446-72-0	NA
86	270.05	9.096	11	4.7271104	C15H10O5	270.0528234	Demethyltexasin		17817-31-1	NA
86	270.05	9.096	11	4.7271104	C15H10O5	270.0528234	8-Hydroxydaidzein		75187-63-2	NA
97	460.1	11.394	1	-2.741756	C22H20O11	460.1005615	wogonin 7-O-beta-D-glucuronate			NA
98	270.05	11.116	11	-4.530338	C15H10O5	270.0528234	2'-hydroxydaidzein	C02495		4444153
98	270.05	11.116	11	-4.530338	C15H10O5	270.0528234	6,7,4'-trihydroxyisoflavone	C14314		4447693
98	270.05	11.116	11	-4.530338	C15H10O5	270.0528234	quinol vinyl ether			NA
98	270.05	11.116	11	-4.530338	C15H10O5	270.0528234	phenoxy radical VII			NA
98	270.05	11.116	11	-4.530338	C15H10O5	270.0528234	luteolinidin			NA
98	270.05	11.116	11	-4.530338	C15H10O5	270.0528234	pelargonidin	C05904	0134-04-03	NA
98	270.05	11.116	11	-4.530338	C15H10O5	270.0528234	apigenin	C01477	520-36-5	NA
98	270.05	11.116	11	-4.530338	C15H10O5	270.0528234	baicalein	C10023		4444924
98	270.05	11.116	11	-4.530338	C15H10O5	270.0528234	genistein	C06563	446-72-0	NA
98	270.05	11.116	11	-4.530338	C15H10O5	270.0528234	Demethyltexasin		17817-31-1	NA
98	270.05	11.116	11	-4.530338	C15H10O5	270.0528234	8-Hydroxydaidzein		75187-63-2	NA
110	286.04	10.175	4	-16.91345	C15H10O6	286.0477381	2'-hydroxygenistein	C12134		NA
110	286.04	10.175	4	-16.91345	C15H10O6	286.0477381	luteolin	C01514	491-70-3	NA
110	286.04	10.175	4	-16.91345	C15H10O6	286.0477381	cyanidin	C05905		NA
110	286.04	10.175	4	-16.91345	C15H10O6	286.0477381	kaempferol	C05903		NA
112	460.1	10.893	1	-3.393787	C22H20O11	460.1005615	wogonin 7-O-beta-D-glucuronate			NA
118	594.16	8.147	4	0.3866115	C27H30O15	594.1584703	pelargonidin-3,5-di-O-beta-D-glucoside	C08725		NA
118	594.16	8.147	4	0.3866115	C27H30O15	594.1584703	pelargonidin 3-O-sophoroside	C16305		NA
118	594.16	8.147	4	0.3866115	C27H30O15	594.1584703	kaempferol 3-O-rhamnoside-7-O-glucoside			NA
118	594.16	8.147	4	0.3866115	C27H30O15	594.1584703	Kaempferol 3-neohesperidoside		32602-81-6	NA
122	416.11	7.385	5	-0.077476	C21H20O9	416.1107322	apigeninidin 5-O-glucoside			NA
122	416.11	7.385	5	-0.077476	C21H20O9	416.1107322	8C-hexosyl chrysin			NA
122	416.11	7.385	5	-0.077476	C21H20O9	416.1107322	6C-hexosyl chrysin			NA
122	416.11	7.385	5	-0.077476	C21H20O9	416.1107322	daidzin	C10216	552-66-9	10164919
122	416.11	7.385	5	-0.077476	C21H20O9	416.1107322	Hispidol 6-glucoside		############	NA
128	578.15	5.387	1	6.873225	C30H26O12	578.1424263	pelargonidin 3-O-beta-D-p-coumaroylglucosid	C16368		NA
130	338.11	14.956	4	-5.393678	C20H18O5	338.1154237	glyceollin II	C10422	67314-98-1	158203
130	338.11	14.956	4	-5.393678	C20H18O5	338.1154237	glyceollin I	C01701	57103-57-8	142931
130	338.11	14.956	4	-5.393678	C20H18O5	338.1154237	glyceollin III	C15511	61080-23-7	10128488
130	338.11	14.956	4	-5.393678	C20H18O5	338.1154237	Canescacarpin		79082-46-5	NA
134	450.12	7.077	2	1.0851706	C21H22O11	450.1162115	8C-glucosyl-2-hydroxynaringenin			NA
134	450.12	7.077	2	1.0851706	C21H22O11	450.1162115	6C-glucosyl-2-hydroxynaringenin			NA
135	608.17	9.675	1	-0.362323	C28H32O15	608.1741204	6-C-Glucopyranosyl-8-C-galactopyranosylgenkwanin			NA
149	356.07	4.002	1	-0.041179	C9H17N4O9	356.0733147	5-amino-6-(5-phospho-D-ribitylamino)uracil	C04454		NA
156	208.15	13.831	2	-1.584859	C13H20O2	208.1463299	3-hydroxy-9-apo-delta-caroten-9-one			NA
156	208.15	13.831	2	-1.584859	C13H20O2	208.1463299	3-hydroxy-beta-ionone			NA
159	756.21	7.268	3	0.5372546	C33H40O20	756.2112937	Kaempferol 3-(2G-glucosylrutinoside)		55696-58-7	NA
159	756.21	7.268	3	0.5372546	C33H40O20	756.2112937	Kaempferol 3-(2G-rhamnosylgentiobioside)		55780-30-8	NA
159	756.21	7.268	3	0.5372546	C33H40O20	756.2112937	Manghaslin		55696-57-6	NA
162	336.1	14.648	2	-5.872131	C20H16O5	336.0997736	Sojagol		18979-00-5	NA
162	336.1	14.648	2	-5.872131	C20H16O5	336.0997736	Clandestacarpin		79002-16-7	NA
163	210.04	3.022	1	-3.177038	C6H10O8	210.0375673	2-carboxy-L-threo-pentonate			NA
169	138.03	9.238	4	-4.303782	C7H6O3	138.0316941	4-hydroxybenzoate	C00156	99-96-7	132
169	138.03	9.238	4	-4.303782	C7H6O3	138.0316941	salicylate	C00805	69-72-7	4964
169	138.03	9.238	4	-4.303782	C7H6O3	138.0316941	3-hydroxybenzoate			NA

ID	Mass	RT	hits	delta_ppm	Formula	NeutralMass	Cpd	KEGG	CAS	ChemSpider
169	138.03	9.238	4	-4.3037819	C7H6O3	138.0316941	protocatechualdehyde	C16700		8438
179	448.1	8.399	6	-2.5920148	C21H20O11	448.1005615	isoorientin			NA
179	448.1	8.399	6	-2.5920148	C21H20O11	448.1005615	cyanidin-3-O-beta-D-galactoside			NA
179	448.1	8.399	6	-2.5920148	C21H20O11	448.1005615	cyanidin-3-O-beta-D-glucoside	C08604	7084-24-4	NA
179	448.1	8.399	6	-2.5920148	C21H20O11	448.1005615	orientin			NA
179	448.1	8.399	6	-2.5920148	C21H20O11	448.1005615	quercetin 3-O-rhamnoside			NA
179	448.1	8.399	6	-2.5920148	C21H20O11	448.1005615	kaempferol-3-glucoside	C12249		NA
180	296.05	7.239	1	-2.0853048	C13H12O8	296.0532174	phaselate	C10483		NA
186	912.51	12.869	1	-2.2476547	C47H76O17	912.508251	Soyasaponin II		55319-36-3	NA
189	164.05	10.406	5	-6.9743409	C9H8O3	164.0473441	enol-phenylpyruvate	C02763	5801-57-0	20058468
189	164.05	10.406	5	-6.9743409	C9H8O3	164.0473441	4-coumarate	C00811	7400-08-0	4450678
189	164.05	10.406	5	-6.9743409	C9H8O3	164.0473441	2-coumarate			NA
189	164.05	10.406	5	-6.9743409	C9H8O3	164.0473441	coumarinate	C05838		20118034
189	164.05	10.406	5	-6.9743409	C9H8O3	164.0473441	2-oxo-3-phenylpropanoate	C00166	0156-06-09	3784710
191	338.11	11.405	4	-5.0979209	C20H18O5	338.1154237	glyceollin II	C10422	67314-98-1	158203
191	338.11	11.405	4	-5.0979209	C20H18O5	338.1154237	glyceollin I	C01701	57103-57-8	142931
191	338.11	11.405	4	-5.0979209	C20H18O5	338.1154237	glyceollin III	C15511	61080-23-7	10128488
191	338.11	11.405	4	-5.0979209	C20H18O5	338.1154237	Canescacarpin		79082-46-5	NA
192	268.04	11.154	2	-7.3622874	C15H8O5	268.0371734	6,7-dehydrobaicalein			NA
192	268.04	11.154	2	-7.3622874	C15H8O5	268.0371734	Coumestrol		479-13-0	NA
193	434.12	8.748	3	-3.2178214	C21H22O10	434.1212969	6C-glucosyl-2,5,7-trihydroxyflavanone			NA
193	434.12	8.748	3	-3.2178214	C21H22O10	434.1212969	8C-glucosyl-2,5,7-trihydroxyflavanone			NA
193	434.12	8.748	3	-3.2178214	C21H22O10	434.1212969	Dihydrogenistin		441045-21-2	NA
194	240.06	6.657	1	-4.9491787	C11H12O6	240.0633881	(1R,6R)-6-hydroxy-2-succinylcyclohexa-2,4-diene-1-carboxylate	C05817		NA
205	154.03	6.709	2	-5.899504	C7H6O4	154.0266087	protocatechuate			NA
205	154.03	6.709	2	-5.899504	C7H6O4	154.0266087	gentisate			NA
208	164.05	7.779	5	-2.0976999	C9H8O3	164.0473441	enol-phenylpyruvate	C02763	5801-57-0	20058468
208	164.05	7.779	5	-2.0976999	C9H8O3	164.0473441	4-coumarate	C00811	7400-08-0	4450678
208	164.05	7.779	5	-2.0976999	C9H8O3	164.0473441	2-coumarate			NA
208	164.05	7.779	5	-2.0976999	C9H8O3	164.0473441	coumarinate	C05838		20118034
208	164.05	7.779	5	-2.0976999	C9H8O3	164.0473441	2-oxo-3-phenylpropanoate	C00166	0156-06-09	3784710
210	268.07	10.138	2	0.52645252	C16H12O4	268.0735589	isoformononetin	C12125	486-63-5	3632
210	268.07	10.138	2	0.52645252	C16H12O4	268.0735589	formononetin	C00858	485-72-3	4444070
211	594.16	8.297	4	-0.4549148	C27H30O15	594.1584703	pelargonidin-3,5-di-O-beta-D-glucoside	C08725		NA
211	594.16	8.297	4	-0.4549148	C27H30O15	594.1584703	pelargonidin 3-O-sophoroside	C16305		NA
211	594.16	8.297	4	-0.4549148	C27H30O15	594.1584703	kaempferol 3-O-rhamnoside-7-O-glucoside			NA
211	594.16	8.297	4	-0.4549148	C27H30O15	594.1584703	Kaempferol 3-neohesperidoside		32602-81-6	NA
212	254.06	9.825	3	-6.3324468	C15H10O4	254.0579088	hispidol		5786-54-9	4824659
212	254.06	9.825	3	-6.3324468	C15H10O4	254.0579088	apigeninidin			NA
212	254.06	9.825	3	-6.3324468	C15H10O4	254.0579088	daidzein	C10208	486-66-8	4445025
215	756.21	7.235	3	1.19844547	C33H40O20	756.2112937	Kaempferol 3-(2G-glucosylrutinoside)		55696-58-7	NA
215	756.21	7.235	3	1.19844547	C33H40O20	756.2112937	Kaempferol 3-(2G-rhamnosylgentiobioside)		55780-30-8	NA
215	756.21	7.235	3	1.19844547	C33H40O20	756.2112937	Manghaslin		55696-57-6	NA
224	204.1	8.999	1	-16.922346	C7FH13N4O2	204.1022539	(E)-alpha-monofluoromethyldehydroarginine			4943693
226	298.08	10.11	2	-6.1175955	C17H14O5	298.0841236	apigenin-7,4'-dimethyl ether	C10019	5128-44-9	NA
226	298.08	10.11	2	-6.1175955	C17H14O5	298.0841236	Afrormosin		550-79-8	NA
240	594.16	7.9	4	0.3866115	C27H30O15	594.1584703	pelargonidin-3,5-di-O-beta-D-glucoside	C08725		NA
240	594.16	7.9	4	0.3866115	C27H30O15	594.1584703	pelargonidin 3-O-sophoroside	C16305		NA
240	594.16	7.9	4	0.3866115	C27H30O15	594.1584703	kaempferol 3-O-rhamnoside-7-O-glucoside			NA
240	594.16	7.9	4	0.3866115	C27H30O15	594.1584703	Kaempferol 3-neohesperidoside		32602-81-6	NA
241	306.08	6.724	2	7.66872607	C15H14O7	306.0739528	gallocatechin	C12127		NA
241	306.08	6.724	2	7.66872607	C15H14O7	306.0739528	leucocyanidin	C05906		389677
246	194.06	10.663	2	-5.7138006	C10H10O4	194.0579088	5-hydroxy-coniferaldehyde	C12204		4445308
246	194.06	10.663	2	-5.7138006	C10H10O4	194.0579088	ferulate	C01494	1135-24-6	4573888
251	210.04	5.678	1	-3.6531436	C6H10O8	210.0375673	2-carboxy-L-threo-pentonate			NA
258	368.11	9.63	2	-4.1624396	C17H20O9	368.1107322	O-feruloylquinate	C02572		NA
258	368.11	9.63	2	13.5637771	C16H20N2O6S	368.1042071	indolylmethyl-desulfoglucosinolate	C16517		NA
261	330.24	11.612	1	-0.6788871	C18H34O5	330.2406242	(9Z)-12,13,17-trihydroxyoctadeca-9-enoate			NA
269	180.04	9.626	6	-4.2142573	C9H8O4	180.0422587	(E)-2,4-dihydroxycinnamate			NA
269	180.04	9.626	6	-4.2142573	C9H8O4	180.0422587	(Z)-2,4-dihydroxycinnamate			NA
269	180.04	9.626	6	-4.2142573	C9H8O4	180.0422587	trans-caffeate	C01197		4450294
269	180.04	9.626	6	-4.2142573	C9H8O4	180.0422587	cis-caffeate			NA
269	180.04	9.626	6	-4.2142573	C9H8O4	180.0422587	4-hydroxyphenylpyruvate	C01179	156-39-8	5341947
269	180.04	9.626	6	-4.2142573	C9H8O4	180.0422587	Caffeic acid		501-16-6	NA
273	314.06	6.66	2	-17.276407	C10H20O7P2	314.068426	geranyl diphosphate	C00341		14211068

ID	Mass	RT	hits	delta_ppm	Formula	NeutralMass	Cpd	KEGG	CAS	ChemSpider
273	314.06	6.66	2	-17.27641	C10H20O7P2	314.068426	(+)-bornyl-diphosphate			NA
277	450.11	7.608	2	-4.024621	C21H22O11	450.1162115	8C-glucosyl-2-hydroxynaringenin			NA
277	450.11	7.608	2	-4.024621	C21H22O11	450.1162115	6C-glucosyl-2-hydroxynaringenin			NA
278	246.1	8.238	3	11.704406	C8H14N4O5	246.0964196	3,6,8-trimethyl-2-oxo-4-hydroxy-4-carboxy-5-ureidoimidazoline			NA
278	246.1	8.238	3	-4.641716	C13H14N2O3	246.1004423	indole-3-acetyl-alanine			NA
278	246.1	8.238	3	-18.33855	C10H18N2O3S	246.1038132	9-mercaptodethiobiotin			NA
284	448.1	7.438	6	-1.922522	C21H20O11	448.1005615	isoorientin			NA
284	448.1	7.438	6	-1.922522	C21H20O11	448.1005615	cyanidin-3-O-beta-D-galactoside			NA
284	448.1	7.438	6	-1.922522	C21H20O11	448.1005615	cyanidin-3-O-beta-D-glucoside	C08604	7084-24-4	NA
284	448.1	7.438	6	-1.922522	C21H20O11	448.1005615	orientin			NA
284	448.1	7.438	6	-1.922522	C21H20O11	448.1005615	quercetin 3-O-rhamnoside			NA
284	448.1	7.438	6	-1.922522	C21H20O11	448.1005615	kaempferol-3-glucoside	C12249		NA
291	192.03	5.098	5	-2.617399	C6H8O7	192.0270026	citrate	C00158	77-92-9	29081
291	192.03	5.098	5	-2.617399	C6H8O7	192.0270026	2,3-dioxo-L-gulonate	C04575		20015966
291	192.03	5.098	5	-2.617399	C6H8O7	192.0270026	dehydroascorbate (bicyclic form)			NA
291	192.03	5.098	5	-2.617399	C6H8O7	192.0270026	2-carboxy-L-xylonolactone			NA
291	192.03	5.098	5	-2.617399	C6H8O7	192.0270026	D-threo-isocitrate	C00451	320-77-4	4573553
292	298.04	11.035	1	-9.522142	C16H10O6	298.0477381	2'-hydroxypseudobaptigenin	C16226		NA
303	328.08	6.838	1	-3.145918	C14H16O9	328.0794321	2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylate	C16519		NA
304	594.16	8.147	4	0.2183062	C27H30O15	594.1584703	pelargonidin-3,5-di-O-beta-D-glucoside	C08725		NA
304	594.16	8.147	4	0.2183062	C27H30O15	594.1584703	pelargonidin 3-O-sophoroside	C16305		NA
304	594.16	8.147	4	0.2183062	C27H30O15	594.1584703	kaempferol 3-O-rhamnoside-7-O-glucoside			NA
304	594.16	8.147	4	0.2183062	C27H30O15	594.1584703	Kaempferol 3-neohesperidoside		32602-81-6	NA
308	594.16	7.906	4	0.3866115	C27H30O15	594.1584703	pelargonidin-3,5-di-O-beta-D-glucoside	C08725		NA
308	594.16	7.906	4	0.3866115	C27H30O15	594.1584703	pelargonidin 3-O-sophoroside	C16305		NA
308	594.16	7.906	4	0.3866115	C27H30O15	594.1584703	kaempferol 3-O-rhamnoside-7-O-glucoside			NA
308	594.16	7.906	4	0.3866115	C27H30O15	594.1584703	Kaempferol 3-neohesperidoside		32602-81-6	NA
313	284.07	13.826	9	-5.891165	C16H12O5	284.0684735	2-hydroxyformononetin	C02920	1890-99-9	4444180
313	284.07	13.826	9	-5.891165	C16H12O5	284.0684735	prunetin	C10521		4445116
313	284.07	13.826	9	-5.891165	C16H12O5	284.0684735	(+)-maackiain	C16229		141688
313	284.07	13.826	9	-5.891165	C16H12O5	284.0684735	(-)-maackiain	C10502		82631
313	284.07	13.826	9	-5.891165	C16H12O5	284.0684735	wogonin			NA
313	284.07	13.826	9	-5.891165	C16H12O5	284.0684735	genkwanin			NA
313	284.07	13.826	9	-5.891165	C16H12O5	284.0684735	calycosin	C01562	20575-57-9	4444104
313	284.07	13.826	9	-5.891165	C16H12O5	284.0684735	biochanin-A	C00814	491-80-5	NA
313	284.07	13.826	9	-5.891165	C16H12O5	284.0684735	Glycitein		40957-83-3	NA
316	164.05	10.564	5	-6.974341	C9H8O3	164.0473441	enol-phenylpyruvate	C02763	5801-57-0	20058468
316	164.05	10.564	5	-6.974341	C9H8O3	164.0473441	4-coumarate	C00811	7400-08-0	4450678
316	164.05	10.564	5	-6.974341	C9H8O3	164.0473441	2-coumarate			NA
316	164.05	10.564	5	-6.974341	C9H8O3	164.0473441	coumarinate	C05838		20118034
316	164.05	10.564	5	-6.974341	C9H8O3	164.0473441	2-oxo-3-phenylpropanoate	C00166	0156-06-09	3784710
322	436.14	6.876	1	-2.859168	C21H24O10	436.136947	phlorizin	C01604		NA
329	210.04	5.246	1	-4.605354	C6H10O8	210.0375673	2-carboxy-L-threo-pentonate			NA
331	164.05	7.582	5	-6.364761	C9H8O3	164.0473441	enol-phenylpyruvate	C02763	5801-57-0	20058468
331	164.05	7.582	5	-6.364761	C9H8O3	164.0473441	4-coumarate	C00811	7400-08-0	4450678
331	164.05	7.582	5	-6.364761	C9H8O3	164.0473441	2-coumarate			NA
331	164.05	7.582	5	-6.364761	C9H8O3	164.0473441	coumarinate	C05838		20118034
331	164.05	7.582	5	-6.364761	C9H8O3	164.0473441	2-oxo-3-phenylpropanoate	C00166	0156-06-09	3784710
338	356.11	6.751	1	-7.39163	C16H20O9	356.1107322	1-O-feruloyl-beta-D-glucose			NA
341	338.11	14.048	4	-5.393678	C20H18O5	338.1154237	glyceollin II	C10422	67314-98-1	158203
341	338.11	14.048	4	-5.393678	C20H18O5	338.1154237	glyceollin I	C01701	57103-57-8	142931
341	338.11	14.048	4	-5.393678	C20H18O5	338.1154237	glyceollin III	C15511	61080-23-7	10128488
341	338.11	14.048	4	-5.393678	C20H18O5	338.1154237	Canescacarpin		79082-46-5	NA
352	326.1	6.653	4	-5.113621	C15H18O8	326.1001676	cis-coumarinic acid-beta-D-glucoside	C05839		NA
352	326.1	6.653	4	-5.113621	C15H18O8	326.1001676	4-O-beta-D-glucosyl-4-hydroxycinnamate			NA
352	326.1	6.653	4	-5.113621	C15H18O8	326.1001676	trans-beta-D-glucosyl-2-hydroxycinnamate			NA
352	326.1	6.653	4	-5.113621	C15H18O8	326.1001676	1-O-(4-coumaroyl)-beta-D-glucose			NA
353	300.06	9.264	5	-1.959975	C16H12O6	300.0633881	(-)-sophorol	C16228		NA
353	300.06	9.264	5	-1.959975	C16H12O6	300.0633881	chrysoeriol	C04293		NA
353	300.06	9.264	5	-1.959975	C16H12O6	300.0633881	pratensein	C10520		NA
353	300.06	9.264	5	-1.959975	C16H12O6	300.0633881	(+)-6a-hydroxymaackiain	C16230		7822419
353	300.06	9.264	5	-1.959975	C16H12O6	300.0633881	scutellarein 7-methyl ether			NA
354	176.07	6.043	2	-4.393145	C7H12O5	176.0684735	(2S)-2-isopropylmalate	C02504		4925359
354	176.07	6.043	2	-4.393145	C7H12O5	176.0684735	(2R,3S)-3-isopropylmalate	C04411		5256741
355	192.03	5.759	5	-0.534359	C6H8O7	192.0270026	citrate	C00158	77-92-9	29081

ID	Mass	RT	hits	delta_p	Formula	NeutralMas	Cpd	KEGG	CAS	ChemSpider
355	192.03	5.759	5	-0.534	C6H8O7	192.027	2,3-dioxo-L-gulonate	C04575		20015966
355	192.03	5.759	5	-0.534	C6H8O7	192.027	dehydroascorbate (bicyclic form)			NA
355	192.03	5.759	5	-0.534	C6H8O7	192.027	2-carboxy-L-xylonolactone			NA
355	192.03	5.759	5	-0.534	C6H8O7	192.027	D-threo-isocitrate	C00451	320-77-4	4573553
357	432.1	8.648	7	-1.728	C21H20O10	432.10565	kaempferol-3-rhamnoside			NA
357	432.1	8.648	7	-1.728	C21H20O10	432.10565	isovitexin			NA
357	432.1	8.648	7	-1.728	C21H20O10	432.10565	luteolinidin 5-O-glucoside			NA
357	432.1	8.648	7	-1.728	C21H20O10	432.10565	pelargonidin-3-O-beta-D-glucoside	C12137		NA
357	432.1	8.648	7	-1.728	C21H20O10	432.10565	vitexin		3681-93-4	NA
357	432.1	8.648	7	-1.728	C21H20O10	432.10565	genistin	C09126	529-59-9	19265428
357	432.1	8.648	7	-1.728	C21H20O10	432.10565	Demethyltexasin 4'-O-glucoside		34307-23-8	NA
358	594.16	8.302	4	-1.128	C27H30O15	594.15847	pelargonidin-3,5-di-O-beta-D-glucoside	C08725		NA
358	594.16	8.302	4	-1.128	C27H30O15	594.15847	pelargonidin 3-O-sophoroside	C16305		NA
358	594.16	8.302	4	-1.128	C27H30O15	594.15847	kaempferol 3-O-rhamnoside-7-O-glucoside			NA
358	594.16	8.302	4	-1.128	C27H30O15	594.15847	Kaempferol 3-neohesperidoside		32602-81-6	NA
359	316.08	5.377	1	-3.582	C13H16O9	316.07943	2,5-dihydroxybenzoate 5-O-beta-D-glucoside			NA
360	434.25	8.464	1	15.107	C21H39O7P	434.24334	1-oleoyl-2-lyso-glycerone phosphate			NA
362	314.06	5.739	2	-17.28	C10H20O7P2	314.06843	geranyl diphosphate	C00341		14211068
362	314.06	5.739	2	-17.28	C10H20O7P2	314.06843	(+)-bornyl-diphosphate			NA
365	356.07	5.711	1	4.4523	C9H17N4O9P	356.07331	5-amino-6-(5-phospho-D-ribitylamino)uracil	C04454		NA
389	310.21	14.682	3	-3.576	C18H30O4	310.21441	2-R-hydroperoxy-linolenate			NA
389	310.21	14.682	3	-3.576	C18H30O4	310.21441	13(S)-HPOTE	C04785		NA
389	310.21	14.682	3	-3.576	C18H30O4	310.21441	9(S)-HPOTE	C16321		NA
390	356.07	5.36	1	2.4864	C9H17N4O9P	356.07331	5-amino-6-(5-phospho-D-ribitylamino)uracil	C04454		NA
398	300.06	11.335	5	-4.959	C16H12O6	300.06339	(-)-sophorol	C16228		NA
398	300.06	11.335	5	-4.959	C16H12O6	300.06339	chrysoeriol	C04293		NA
398	300.06	11.335	5	-4.959	C16H12O6	300.06339	pratensein	C10520		NA
398	300.06	11.335	5	-4.959	C16H12O6	300.06339	(+)-6a-hydroxymaackiain	C16230		7822419
398	300.06	11.335	5	-4.959	C16H12O6	300.06339	scutellarein 7-methyl ether			NA
407	314.06	5.05	2	-14.41	C10H20O7P2	314.06843	geranyl diphosphate	C00341		14211068
407	314.06	5.05	2	-14.41	C10H20O7P2	314.06843	(+)-bornyl-diphosphate			NA
409	354.11	12.581	2	-7.168	C20H18O6	354.11034	2,3-dehydrokievitone			NA
409	354.11	12.581	2	-7.168	C20H18O6	354.11034	Glyceofuran		78873-52-6	NA
410	130.06	5.575	2	-5.337	C6H10O3	130.06299	4-methyl-2-oxopentanoate	C00233	816-66-0	2766269
410	130.06	5.575	2	-5.337	C6H10O3	130.06299	(S)-3-methyl-2-oxopentanoate	C00671	1460-34-0	19951083
413	634.41	14.018	1	-2.338	C36H58O9	634.40808	soyasapogenol B-3-O-beta-glucuronide			NA
415	210.04	6.501	1	-1.749	C6H10O8	210.03757	2-carboxy-L-threo-pentonate			NA
417	298.05	11.599	1	-8.851	C16H10O6	298.04774	2'-hydroxypseudobaptigenin	C16226		NA
420	224.14	13.172	3	-4.214	C13H20O3	224.14124	5,6-epoxy-3-hydroxy-9-apo-beta-caroten-9-one			NA
420	224.14	13.172	3	-4.214	C13H20O3	224.14124	grasshopper ketone			NA
420	224.14	13.172	3	-4.214	C13H20O3	224.14124	methyl jasmonate	C11512		4519204
434	234.1	8.49	1	-3.826	C10H19O4P	234.1021	geranyl monophosphate			NA
444	194.06	8.115	2	-6.229	C10H10O4	194.05791	5-hydroxy-coniferaldehyde	C12204		4445308
444	194.06	8.115	2	-6.229	C10H10O4	194.05791	ferulate	C01494	1135-24-6	4573888
453	192.03	6.241	5	-1.576	C6H8O7	192.027	citrate	C00158	77-92-9	29081
453	192.03	6.241	5	-1.576	C6H8O7	192.027	2,3-dioxo-L-gulonate	C04575		20015966
453	192.03	6.241	5	-1.576	C6H8O7	192.027	dehydroascorbate (bicyclic form)			NA
453	192.03	6.241	5	-1.576	C6H8O7	192.027	2-carboxy-L-xylonolactone			NA
453	192.03	6.241	5	-1.576	C6H8O7	192.027	D-threo-isocitrate	C00451	320-77-4	4573553
459	608.17	9.802	1	-1.349	C28H32O15	608.17412	6-C-Glucopyranosyl-8-C-galactopyranosylgenkwanin			NA
462	290.09	8.621	1	-8.175	C14H14N2O5	290.09027	indole-3-acetyl-L-aspartate			NA
464	251.08	7.142	1	-0.687	C12H13NO5	251.07937	N-benzoyl-L-glutamate			11310035
467	210.04	5.828	1	-3.653	C6H10O8	210.03757	2-carboxy-L-threo-pentonate			NA
481	294.22	16.267	6	-7.46	C18H30O3	294.21949	colneleate			NA
481	294.22	16.267	6	-7.46	C18H30O3	294.21949	(9Z,12Z)-15,16-epoxyoctadeca-9,12-dienoate			NA
481	294.22	16.267	6	-7.46	C18H30O3	294.21949	(9Z,15Z)-12,13-epoxy-octadeca-9,15-dienoate			NA
481	294.22	16.267	6	-7.46	C18H30O3	294.21949	(12Z,15Z)-9,10-epoxyoctadeca-12,15-dienoate			NA
481	294.22	16.267	6	-7.46	C18H30O3	294.21949	9,10-epoxy-10,12Z-octadecadienoate			NA
481	294.22	16.267	6	-7.46	C18H30O3	294.21949	3-oxo-2-(cis-2'-pentenyl)-cyclopentane-1-octanoate	C04780		NA
491	356.07	5.078	1	3.6098	C9H17N4O9P	356.07331	5-amino-6-(5-phospho-D-ribitylamino)uracil	C04454		NA
493	116.08	5.455	1	-4.562	C6H12O2	116.08373	hexanoate	C01585	142-62-1	3599616
504	210.04	6.251	1	-4.605	C6H10O8	210.03757	2-carboxy-L-threo-pentonate			NA
516	204.09	5.335	1	-4.3	C11H12N2O2	204.08988	L-tryptophan	C00078	6912-86-3	NA
531	356.11	8.403	1	-2.899	C16H20O9	356.11073	1-O-feruloyl-beta-D-glucose			NA
534	580.14	8.177	3	0.9994	C26H28O15	580.14282	cyanidin 3-O-[2''-O-(xylosyl) glucoside			NA

ID	Mass	RT	hits	delta_ppm	Formula	NeutralMass	Cpd	KEGG	CAS	ChemSpider
534	580.14	8.177	3	0.99936133	C26H28O15	580.1428202	cyanidin 3-O-(beta-D-xylosyl-(1->2)-beta-D-galactoside)			NA
534	580.14	8.177	3	0.99936133	C26H28O15	580.1428202	Isocarlinoside		83151-90-0	NA
549	251.08	7.773	1	-2.28028927	C12H13NO5	251.0793725	N-benzoyl-L-glutamate			11310035
553	434.25	7.472	1	12.5733875	C21H39O7P	434.2433401	1-oleoyl-2-lyso-glycerone phosphate			NA
554	234.1	9.157	1	-2.97129036	C10H19O4P	234.1020956	geranyl monophosphate			NA
561	326.1	7.511	4	-3.88700383	C15H18O8	326.1001676	cis-coumarinic acid-beta-D-glucoside	C05839		NA
561	326.1	7.511	4	-3.88700383	C15H18O8	326.1001676	4-O-beta-D-glucosyl-4-hydroxycinnamate			NA
561	326.1	7.511	4	-3.88700383	C15H18O8	326.1001676	trans-beta-D-glucosyl-2-hydroxycinnamate			NA
561	326.1	7.511	4	-3.88700383	C15H18O8	326.1001676	1-O-(4-coumaroyl)-beta-D-glucose			NA
582	192.03	5.832	5	-4.17967936	C6H8O7	192.0270026	citrate	C00158	77-92-9	29081
582	192.03	5.832	5	-4.17967936	C6H8O7	192.0270026	2,3-dioxo-L-gulonate	C04575		20015966
582	192.03	5.832	5	-4.17967936	C6H8O7	192.0270026	dehydroascorbate (bicyclic form)			NA
582	192.03	5.832	5	-4.17967936	C6H8O7	192.0270026	2-carboxy-L-xylonolactone			NA
582	192.03	5.832	5	-4.17967936	C6H8O7	192.0270026	D-threo-isocitrate	C00451	320-77-4	4573553
590	338.11	10.538	4	-3.91489298	C20H18O5	338.1154237	glyceollin II	C10422	67314-98-1	158203
590	338.11	10.538	4	-3.91489298	C20H18O5	338.1154237	glyceollin I	C01701	57103-57-8	142931
590	338.11	10.538	4	-3.91489298	C20H18O5	338.1154237	glyceollin III	C15511	61080-23-7	10128488
590	338.11	10.538	4	-3.91489298	C20H18O5	338.1154237	Canescacarpin		79082-46-5	NA
614	138.03	5.96	4	-5.75272444	C7H6O3	138.0316941	4-hydroxybenzoate	C00156	99-96-7	132
614	138.03	5.96	4	-5.75272444	C7H6O3	138.0316941	salicylate	C00805	69-72-7	4964
614	138.03	5.96	4	-5.75272444	C7H6O3	138.0316941	3-hydroxybenzoate			NA
614	138.03	5.96	4	-5.75272444	C7H6O3	138.0316941	protocatechualdehyde	C16700		8438
616	474.12	9.097	1	-1.71170502	C23H22O11	474.1162115	Genistin 6''-O-acetate		73566-30-0	NA
617	502.11	8.751	1	-5.23025534	C24H22O12	502.1111262	malonyldaidzin	C16191		NA
619	326.1	6.373	4	-1.43376989	C15H18O8	326.1001676	cis-coumarinic acid-beta-D-glucoside	C05839		NA
619	326.1	6.373	4	-1.43376989	C15H18O8	326.1001676	4-O-beta-D-glucosyl-4-hydroxycinnamate			NA
619	326.1	6.373	4	-1.43376989	C15H18O8	326.1001676	trans-beta-D-glucosyl-2-hydroxycinnamate			NA
619	326.1	6.373	4	-1.43376989	C15H18O8	326.1001676	1-O-(4-coumaroyl)-beta-D-glucose			NA
625	284.07	11.115	9	-6.59522078	C16H12O5	284.0684735	2-hydroxyformononetin	C02920	1890-99-9	4444180
625	284.07	11.115	9	-6.59522078	C16H12O5	284.0684735	prunetin	C10521		4445116
625	284.07	11.115	9	-6.59522078	C16H12O5	284.0684735	(+)-maackiain	C16229		141688
625	284.07	11.115	9	-6.59522078	C16H12O5	284.0684735	(-)-maackiain	C10502		82631
625	284.07	11.115	9	-6.59522078	C16H12O5	284.0684735	wogonin			NA
625	284.07	11.115	9	-6.59522078	C16H12O5	284.0684735	genkwanin			NA
625	284.07	11.115	9	-6.59522078	C16H12O5	284.0684735	calycosin	C01562	20575-57-9	4444104
625	284.07	11.115	9	-6.59522078	C16H12O5	284.0684735	biochanin-A	C00814	491-80-5	NA
625	284.07	11.115	9	-6.59522078	C16H12O5	284.0684735	Glycitein		40957-83-3	NA
626	302.17	10.614	1	-1.78230123	C15H26O6	302.1729386	tributyrin	C13870		13849665
631	254.06	8.337	3	-3.57716949	C15H10O4	254.0579088	hispidol		5786-54-9	4824659
631	254.06	8.337	3	-3.57716949	C15H10O4	254.0579088	apigeninidin			NA
631	254.06	8.337	3	-3.57716949	C15H10O4	254.0579088	daidzein	C10208	486-66-8	4445025
634	474.12	9.089	1	-1.71170502	C23H22O11	474.1162115	Genistin 6''-O-acetate		73566-30-0	NA
645	356.08	7.144	1	14.0008734	C9H17N4O9P	356.0733147	5-amino-6-(5-phospho-D-ribitylamino)uracil	C04454		NA
647	284.07	10.045	9	-5.89116518	C16H12O5	284.0684735	2-hydroxyformononetin	C02920	1890-99-9	4444180
647	284.07	10.045	9	-5.89116518	C16H12O5	284.0684735	prunetin	C10521		4445116
647	284.07	10.045	9	-5.89116518	C16H12O5	284.0684735	(+)-maackiain	C16229		141688
647	284.07	10.045	9	-5.89116518	C16H12O5	284.0684735	(-)-maackiain	C10502		82631
647	284.07	10.045	9	-5.89116518	C16H12O5	284.0684735	wogonin			NA
647	284.07	10.045	9	-5.89116518	C16H12O5	284.0684735	genkwanin			NA
647	284.07	10.045	9	-5.89116518	C16H12O5	284.0684735	calycosin	C01562	20575-57-9	4444104
647	284.07	10.045	9	-5.89116518	C16H12O5	284.0684735	biochanin-A	C00814	491-80-5	NA
647	284.07	10.045	9	-5.89116518	C16H12O5	284.0684735	Glycitein		40957-83-3	NA
648	310.21	14.462	3	-4.2211005	C18H30O4	310.2144094	2-R-hydroperoxy-linolenate			NA
648	310.21	14.462	3	-4.2211005	C18H30O4	310.2144094	13(S)-HPOTE	C04785		NA
648	310.21	14.462	3	-4.2211005	C18H30O4	310.2144094	9(S)-HPOTE	C16321		NA
662	354.11	11.553	2	-7.16812735	C20H18O6	354.1103383	2,3-dehydrokievitone			NA
662	354.11	11.553	2	-7.16812735	C20H18O6	354.1103383	Glyceofuran		78873-52-6	NA
672	532.08	8.074	1	-7.56745518	C16H26N2O14	532.0859265	dTDP-D-oliose			NA
680	740.22	8.87	2	-0.64724277	C33H40O19	740.2163791	pelargonidin-3-O-rutinoside-5-O-beta-D-glucoside	C12645		NA
680	740.22	8.87	2	-0.64724277	C33H40O19	740.2163791	Clitorin		55804-74-5	NA
690	192.03	5.825	5	-0.5343587	C6H8O7	192.0270026	citrate	C00158	77-92-9	29081
690	192.03	5.825	5	-0.5343587	C6H8O7	192.0270026	2,3-dioxo-L-gulonate	C04575		20015966
690	192.03	5.825	5	-0.5343587	C6H8O7	192.0270026	dehydroascorbate (bicyclic form)			NA
690	192.03	5.825	5	-0.5343587	C6H8O7	192.0270026	2-carboxy-L-xylonolactone			NA
690	192.03	5.825	5	-0.5343587	C6H8O7	192.0270026	D-threo-isocitrate	C00451	320-77-4	4573553

ID	Mass	RT	hits	delta_ppm	Formula	NeutralMass	Cpd	KEGG	CAS	ChemSpider
692	342.09	5.827	1	-2.8710582	C15H18O9	342.0950822	1-O-caffeoyl-beta-D-glucose			NA
693	294.22	16.111	6	-9.8389942	C18H30O3	294.2194948	colneleate			NA
693	294.22	16.111	6	-9.8389942	C18H30O3	294.2194948	(9Z,12Z)-15,16-epoxyoctadeca-9,12-dienoate			NA
693	294.22	16.111	6	-9.8389942	C18H30O3	294.2194948	(9Z,15Z)-12,13-epoxy-octadeca-9,15-dienoate			NA
693	294.22	16.111	6	-9.8389942	C18H30O3	294.2194948	(12Z,15Z)-9,10-epoxyoctadeca-12,15-dienoate			NA
693	294.22	16.111	6	-9.8389942	C18H30O3	294.2194948	9,10-epoxy-10,12Z-octadecadienoate			NA
693	294.22	16.111	6	-9.8389942	C18H30O3	294.2194948	3-oxo-2-(cis-2'-pentenyl)-cyclopentane-1-octanoate	C04780		NA
696	356.08	7.59	1	5.5756422	C9H17N4O9P	356.0733147	5-amino-6-(5-phospho-D-ribitylamino)uracil	C04454		NA
715	356.11	7.529	1	-3.4602684	C16H20O9	356.1107322	1-O-feruloyl-beta-D-glucose			NA
719	338.06	6.335	1	-0.7394382	C9H15N4O8P	338.06275	5-amino-1-(5-phospho-D-ribosyl)imidazole-4-carboxamide	C04677	3031-94-5	NA
735	458.12	8.339	1	1.31640944	C23H22O10	458.1212969	Daidzein 6''-O-acetate		71385-83-6	NA
745	270.05	8.261	11	-4.1600398	C15H10O5	270.0528234	2'-hydroxydaidzein	C02495		4444153
745	270.05	8.261	11	-4.1600398	C15H10O5	270.0528234	6,7,4'-trihydroxyisoflavone	C14314		4447693
745	270.05	8.261	11	-4.1600398	C15H10O5	270.0528234	quinol vinyl ether			NA
745	270.05	8.261	11	-4.1600398	C15H10O5	270.0528234	phenoxy radical VII			NA
745	270.05	8.261	11	-4.1600398	C15H10O5	270.0528234	luteolinidin			NA
745	270.05	8.261	11	-4.1600398	C15H10O5	270.0528234	pelargonidin	C05904	0134-04-03	NA
745	270.05	8.261	11	-4.1600398	C15H10O5	270.0528234	apigenin	C01477	520-36-5	NA
745	270.05	8.261	11	-4.1600398	C15H10O5	270.0528234	baicalein	C10023		4444924
745	270.05	8.261	11	-4.1600398	C15H10O5	270.0528234	genistein	C06563	446-72-0	NA
745	270.05	8.261	11	-4.1600398	C15H10O5	270.0528234	Demethyltexasin		17817-31-1	NA
745	270.05	8.261	11	-4.1600398	C15H10O5	270.0528234	8-Hydroxydaidzein		75187-63-2	NA
765	238.12	12.603	1	-2.1378394	C13H18O4	238.1205091	3-[(3aS,4S,7aS)-7a-methyl-1,5-dioxo-octahydro-1H-inden-4-yl]propanoate			NA
768	284.07	8.505	9	-6.9472486	C16H12O5	284.0684735	2-hydroxyformononetin	C02920	1890-99-9	4444180
768	284.07	8.505	9	-6.9472486	C16H12O5	284.0684735	prunetin	C10521		4445116
768	284.07	8.505	9	-6.9472486	C16H12O5	284.0684735	(+)-maackiain	C16229		141688
768	284.07	8.505	9	-6.9472486	C16H12O5	284.0684735	(-)-maackiain	C10502		82631
768	284.07	8.505	9	-6.9472486	C16H12O5	284.0684735	wogonin			NA
768	284.07	8.505	9	-6.9472486	C16H12O5	284.0684735	genkwanin			NA
768	284.07	8.505	9	-6.9472486	C16H12O5	284.0684735	calycosin	C01562	20575-57-9	4444104
768	284.07	8.505	9	-6.9472486	C16H12O5	284.0684735	biochanin-A	C00814	491-80-5	NA
768	284.07	8.505	9	-6.9472486	C16H12O5	284.0684735	Glycitein		40957-83-3	NA
786	338.06	11.403	1	-3.6974689	C9H15N4O8P	338.06275	5-amino-1-(5-phospho-D-ribosyl)imidazole-4-carboxamide	C04677	3031-94-5	NA
794	508.12	8.473	1	-0.5724131	C23H24O13	508.1216909	eupatolitin 3-glucoside			NA
798	298.08	12.445	2	-8.4659259	C17H14O5	298.0841236	apigenin-7,4'-dimethyl ether	C10019	5128-44-9	NA
798	298.08	12.445	2	-8.4659259	C17H14O5	298.0841236	Afrormosin		550-79-8	NA
817	606.25	7.565	1	3.73634654	C35H34N4O6	606.2478348	pheophorbide b			NA
839	342.09	7.028	1	-2.2864254	C15H18O9	342.0950822	1-O-caffeoyl-beta-D-glucose			NA
842	354.11	9.729	2	-3.4969552	C20H18O6	354.1103383	2,3-dehydrokievitone			NA
842	354.11	9.729	2	-3.4969552	C20H18O6	354.1103383	Glyceofuran		78873-52-6	NA
851	356.11	7.004	1	-5.4259491	C16H20O9	356.1107322	1-O-feruloyl-beta-D-glucose			NA
854	594.16	8.305	4	-1.6330517	C27H30O15	594.1584703	pelargonidin-3,5-di-O-beta-D-glucoside	C08725		NA
854	594.16	8.305	4	-1.6330517	C27H30O15	594.1584703	pelargonidin 3-O-sophoroside	C16305		NA
854	594.16	8.305	4	-1.6330517	C27H30O15	594.1584703	kaempferol 3-O-rhamnoside-7-O-glucoside			NA
854	594.16	8.305	4	-1.6330517	C27H30O15	594.1584703	Kaempferol 3-neohesperidoside		32602-81-6	NA
857	614.15	6.415	1	2.16973155	C20H31N4O16P	614.1472675	CMP-N-acetyl-beta-neuraminate			NA
884	532.12	8.974	2	-2.2379383	C25H24O13	532.1216909	(-)-maackiain-3-O-glucoside-6''-malonate	C16231		NA
884	532.12	8.974	2	-2.2379383	C25H24O13	532.1216909	biochanin A-7-O-glucoside-6''-malonate	C12625		NA
890	958.51	10.408	1	0.49001708	C48H78O19	958.5137303	Soyasaponin V		114590-20-4	NA
918	338.06	10.895	1	-6.0638935	C9H15N4O8P	338.06275	5-amino-1-(5-phospho-D-ribosyl)imidazole-4-carboxamide	C04677	3031-94-5	NA
919	326.1	7.894	4	-1.4337699	C15H18O8	326.1001676	cis-coumarinic acid-beta-D-glucoside	C05839		NA
919	326.1	7.894	4	-1.4337699	C15H18O8	326.1001676	4-O-beta-D-glucosyl-4-hydroxycinnamate			NA
919	326.1	7.894	4	-1.4337699	C15H18O8	326.1001676	trans-beta-D-glucosyl-2-hydroxycinnamate			NA
919	326.1	7.894	4	-1.4337699	C15H18O8	326.1001676	1-O-(4-coumaroyl)-beta-D-glucose			NA
925	756.22	10.654	3	14.6867391	C33H40O20	756.2112937	Kaempferol 3-(2G-glucosylrutinoside)		55696-58-7	NA
925	756.22	10.654	3	14.6867391	C33H40O20	756.2112937	Kaempferol 3-(2G-rhamnosylgentiobioside)		55780-30-8	NA
925	756.22	10.654	3	14.6867391	C33H40O20	756.2112937	Manghaslin		55696-57-6	NA
934	446.12	9.705	4	-2.0104954	C22H22O10	446.1212969	(-)-maackiain-3-O-glucoside	C10538		NA
934	446.12	9.705	4	-2.0104954	C22H22O10	446.1212969	wogonin 7-O-beta-D-glucoside			NA
934	446.12	9.705	4	-2.0104954	C22H22O10	446.1212969	biochanin A-7-O-glucoside	C05376		16498850
934	446.12	9.705	4	-2.0104954	C22H22O10	446.1212969	Glycitein 7-O-glucoside		##########	NA
946	450.11	10.658	2	-9.1344124	C21H22O11	450.1162115	8C-glucosyl-2-hydroxynaringenin			NA
946	450.11	10.658	2	-9.1344124	C21H22O11	450.1162115	6C-glucosyl-2-hydroxynaringenin			NA
959	314.08	11.182	3	-8.7181253	C17H14O6	314.0790382	(+)-pisatin	C10516		91879
959	314.08	11.182	3	-8.7181253	C17H14O6	314.0790382	ladanein			NA

ID	Mass	RT	hits	delta_ppm	Formula	NeutralMass	Cpd	KEGG	CAS	ChemSpider
959	314.08	11.182	3	-8.718125	C17H14O6	314.0790382	cirsimaritin			NA
976	286.04	10.145	4	-13.41753	C15H10O6	286.0477381	2'-hydroxygenistein	C12134		NA
976	286.04	10.145	4	-13.41753	C15H10O6	286.0477381	luteolin	C01514	491-70-3	NA
976	286.04	10.145	4	-13.41753	C15H10O6	286.0477381	cyanidin	C05905		NA
976	286.04	10.145	4	-13.41753	C15H10O6	286.0477381	kaempferol	C05903		NA
978	594.13	12.893	3	-7.979501	C30H26O13	594.1373409	cyanidin 3-(p-coumaroyl)-glucoside	C12095		NA
978	594.13	12.893	3	-7.979501	C30H26O13	594.1373409	pelargonidin 3-O-beta-D-caffeoylglucoside	C16297		NA
978	594.13	12.893	3	-7.979501	C30H26O13	594.1373409	cyanidin 3-O-p-coumaroylglucoside			NA
992	354.11	10.541	2	-2.93216	C20H18O6	354.1103383	2,3-dehydrokievitone			NA
992	354.11	10.541	2	-2.93216	C20H18O6	354.1103383	Glyceofuran		78873-52-6	NA
1031	508.12	8.733	1	-2.737249	C23H24O13	508.1216909	eupatolitin 3-glucoside			NA
1055	958.51	12.438	1	-0.970578	C48H78O19	958.5137303	Soyasaponin V		114590-20-4	NA
1062	362.17	14.93	6	-9.218146	C20H26O6	362.1729386	(-)-secoisolariciresinol	C18167		58845
1062	362.17	14.93	6	-9.218146	C20H26O6	362.1729386	(+)-secoisolariciresinol	C20456		NA
1062	362.17	14.93	6	-9.218146	C20H26O6	362.1729386	gibberellin A36	C11862		NA
1062	362.17	14.93	6	-9.218146	C20H26O6	362.1729386	gibberellin A38			NA
1062	362.17	14.93	6	-9.218146	C20H26O6	362.1729386	gibberellin A19	C02034	6980-44-5	NA
1062	362.17	14.93	6	-9.218146	C20H26O6	362.1729386	gibberellin A25			NA
1063	562.26	7.576	1	2.4799942	C34H34N4O4	562.2580056	protoporphyrin IX	C02191	0553-12-8	20171337
1104	610.15	7.654	8	0.3525117	C27H30O16	610.1533849	cyanidin 3,5-di-O-beta-D-glucoside	C08639		NA
1104	610.15	7.654	8	0.3525117	C27H30O16	610.1533849	delphinidin 3-O-rutinoside			NA
1104	610.15	7.654	8	0.3525117	C27H30O16	610.1533849	cyanidin 3,7-di-O-beta-D-glucoside			NA
1104	610.15	7.654	8	0.3525117	C27H30O16	610.1533849	cyanidin 3-O-sophoroside	C16306		NA
1104	610.15	7.654	8	0.3525117	C27H30O16	610.1533849	quercetin 3-O-rhamnoside-7-O-glucoside			NA
1104	610.15	7.654	8	0.3525117	C27H30O16	610.1533849	kaempferol 3-O-beta-D-glucosyl-(1->2)-glucoside	C12634	19895-95-5	NA
1104	610.15	7.654	8	0.3525117	C27H30O16	610.1533849	Kaempferol 3-O-gentiobioside		22149-35-5	NA
1104	610.15	7.654	8	0.3525117	C27H30O16	610.1533849	Quercetin 3-O-neohesperidoside		32453-36-4	NA
1119	532.12	8.924	2	-2.989646	C25H24O13	532.1216909	(-)-maackiain-3-O-glucoside-6''-malonate	C16231		NA
1119	532.12	8.924	2	-2.989646	C25H24O13	532.1216909	biochanin A-7-O-glucoside-6''-malonate	C12625		NA
1120	532.12	9.383	2	-1.674157	C25H24O13	532.1216909	(-)-maackiain-3-O-glucoside-6''-malonate	C16231		NA
1120	532.12	9.383	2	-1.674157	C25H24O13	532.1216909	biochanin A-7-O-glucoside-6''-malonate	C12625		NA
1122	730.17	9.591	1	-0.019565	C34H34O18	730.1745143	cyanidin 3-O-glucoside-7-O-(6-O-(p-hydroxybenzoyl)-glucoside)			NA

Supplementary Table 3.3 A comprehensive information on the eight QTL-multiple metabolite clusters, including their associated single nucleotide polymorphisms (SNPs) along with corresponding P-values and R-squared values.

QTL-multiple metabolite cluster	SNP	pval	Rsq
QTL01.grp41.Gm15-38575222	Gm15-38575222	1.89E-07	0.132807498
QTL01.grp70.Gm15-38276224	Gm15-38276224	7.63E-08	0.140744196
QTL01.grp70.Gm15-38575222	Gm15-38575222	2.20E-06	0.110969809
QTL01.grp70.Gm15-39570425	Gm15-39570425	1.17E-07	0.137024514
QTL01.grp70.Gm15-39945568	Gm15-39945568	2.72E-15	0.279524997
QTL01.grp87.Gm15-38241517	Gm15-38241517	1.23E-07	0.136566043
QTL01.grp87.Gm15-38276224	Gm15-38276224	1.78E-06	0.112889854
QTL01.grp87.Gm15-39945568	Gm15-39945568	5.09E-13	0.239551221
QTL01.grp185	Gm15-39945568	5.67E-11	0.201765253
QTL01.grp187	Gm15-39945568	4.18E-10	0.185236979
QTL01.grp311.Gm15-38241517	Gm15-38241517	1.44E-06	0.1147904
QTL01.grp311.Gm15-38575222	Gm15-38575222	1.05E-07	0.137958612
QTL01.grp311.Gm15-39945568	Gm15-39945568	3.26E-12	0.224855717
QTL01.grp321.Gm15-39945568	Gm15-39945568	9.30E-11	0.197698665
QTL01.grp326	Gm15-39945568	5.34E-07	0.123613537
QTL01.grp371.Gm15-38276224	Gm15-38276224	1.44E-06	0.114786919
QTL01.grp447.Gm15-38241517	Gm15-38241517	1.10E-06	0.117174787
QTL01.grp447.Gm15-38575222	Gm15-38575222	1.38E-06	0.115161673
QTL01.grp450.Gm15-38241517	Gm15-38241517	9.11E-13	0.234971476
QTL01.grp450.Gm15-38276224	Gm15-38276224	6.19E-08	0.142573968
QTL01.grp450.Gm15-38380213	Gm15-38380213	7.16E-07	0.121010282
QTL01.grp450.Gm15-38447399	Gm15-38447399	3.70E-07	0.126863384
QTL01.grp450.Gm15-38547411	Gm15-38547411	1.01E-07	0.138276314
QTL01.grp450.Gm15-38575222	Gm15-38575222	2.12E-20	0.362193025
QTL01.grp450.Gm15-39520837	Gm15-39520837	1.01E-07	0.138276314
QTL01.grp450.Gm15-39570425	Gm15-39570425	4.57E-07	0.125001032
QTL01.grp450.Gm15-39747988	Gm15-39747988	5.50E-08	0.143599806
QTL01.grp450.Gm15-39945568	Gm15-39945568	1.80E-13	0.247653269
QTL01.grp492.Gm15-38241517	Gm15-38241517	1.89E-12	0.229205601
QTL01.grp492.Gm15-38276224	Gm15-38276224	1.31E-10	0.194856531
QTL01.grp492.Gm15-38380213	Gm15-38380213	2.03E-06	0.111708869
QTL01.grp492.Gm15-38447399	Gm15-38447399	1.05E-10	0.196666635
QTL01.grp492.Gm15-38547411	Gm15-38547411	3.01E-08	0.148828614
QTL01.grp492.Gm15-38575222	Gm15-38575222	2.87E-22	0.390037308
QTL01.grp492.Gm15-39520837	Gm15-39520837	3.01E-08	0.148828614
QTL01.grp492.Gm15-39570425	Gm15-39570425	1.78E-07	0.133304004
QTL01.grp492.Gm15-39747988	Gm15-39747988	1.17E-08	0.156964698
QTL01.grp492.Gm15-39945568	Gm15-39945568	1.46E-21	0.37965756
QTL01.grp492.Gm15-39981697	Gm15-39981697	4.79E-09	0.164649411
QTL01.grp553.Gm15-34552298	Gm15-34552298	5.89E-07	0.122742343
QTL01.grp553.Gm15-38241517	Gm15-38241517	2.18E-06	0.111067627
QTL01.grp553.Gm15-38575222	Gm15-38575222	5.96E-07	0.1226315
QTL01.grp557.Gm15-34552298	Gm15-34552298	5.57E-07	0.123240707
QTL01.grp557.Gm15-38241517	Gm15-38241517	5.50E-08	0.143601108
QTL01.grp557.Gm15-38575222	Gm15-38575222	1.14E-07	0.137229191
QTL01.grp573.Gm15-37752199	Gm15-37752199	2.02E-06	0.111768898
QTL01.grp573.Gm15-38241517	Gm15-38241517	1.86E-07	0.132930397
QTL01.grp573.Gm15-38276224	Gm15-38276224	9.15E-09	0.159106631
QTL01.grp573.Gm15-38575222	Gm15-38575222	1.29E-06	0.115743563
QTL01.grp573.Gm15-39570425	Gm15-39570425	4.60E-09	0.165002368
QTL01.grp573.Gm15-39945568	Gm15-39945568	2.36E-18	0.33024404
QTL01.grp574.Gm15-38241517	Gm15-38241517	7.83E-07	0.120208845
QTL01.grp574.Gm15-38276224	Gm15-38276224	3.94E-10	0.185739031
QTL01.grp574.Gm15-38380213	Gm15-38380213	1.14E-06	0.116860361
QTL01.grp574.Gm15-38447399	Gm15-38447399	2.44E-08	0.150659167
QTL01.grp574.Gm15-38575222	Gm15-38575222	2.09E-08	0.152011173
QTL01.grp574.Gm15-39570425	Gm15-39570425	1.15E-09	0.17674683
QTL01.grp574.Gm15-39945568	Gm15-39945568	1.75E-20	0.363436928

QTL-multiple metabolite cluster	SNP	pval	Rsq
QTL01.grp574.Gm15-40010479	Gm15-40010479	1.63E-06	0.11366
QTL01.grp630	Gm15-39945568	5.18E-08	0.14412
QTL01.grp641.Gm15-39945568	Gm15-39945568	6.03E-08	0.14279
QTL01.grp652.Gm15-39570425	Gm15-39570425	1.11E-06	0.1171
QTL01.grp652.Gm15-39945568	Gm15-39945568	1.50E-09	0.17453
QTL01.grp668.Gm15-39945568	Gm15-39945568	1.83E-13	0.24753
QTL01.grp718.Gm15-39945568	Gm15-39945568	1.70E-07	0.13375
QTL01.grp738.Gm15-38575222	Gm15-38575222	3.31E-07	0.12785
QTL01.grp766.Gm15-37622758	Gm15-37622758	7.03E-07	0.12117
QTL01.grp766.Gm15-38029609	Gm15-38029609	4.86E-07	0.12445
QTL01.grp766.Gm15-38276224	Gm15-38276224	3.50E-10	0.18672
QTL01.grp766.Gm15-38380213	Gm15-38380213	1.45E-07	0.13509
QTL01.grp766.Gm15-38447399	Gm15-38447399	3.86E-09	0.16649
QTL01.grp766.Gm15-38575222	Gm15-38575222	8.07E-08	0.14025
QTL01.grp766.Gm15-39570425	Gm15-39570425	3.90E-10	0.18581
QTL01.grp766.Gm15-39945568	Gm15-39945568	4.30E-14	0.25869
QTL01.grp799.Gm15-38241517	Gm15-38241517	4.60E-14	0.25818
QTL01.grp799.Gm15-38276224	Gm15-38276224	5.93E-10	0.18231
QTL01.grp799.Gm15-38380213	Gm15-38380213	1.14E-06	0.11688
QTL01.grp799.Gm15-38575222	Gm15-38575222	1.48E-15	0.28405
QTL01.grp799.Gm15-39945568	Gm15-39945568	2.78E-15	0.27937
QTL01.grp921.Gm15-38241517	Gm15-38241517	1.70E-13	0.24809
QTL01.grp921.Gm15-38276224	Gm15-38276224	5.29E-12	0.22099
QTL01.grp921.Gm15-38380213	Gm15-38380213	1.06E-06	0.11756
QTL01.grp921.Gm15-38447399	Gm15-38447399	9.57E-10	0.17829
QTL01.grp921.Gm15-38575222	Gm15-38575222	1.78E-18	0.3322
QTL01.grp921.Gm15-39570425	Gm15-39570425	1.55E-08	0.1546
QTL01.grp921.Gm15-39945568	Gm15-39945568	6.62E-17	0.30669
QTL01.grp943.Gm15-38241517	Gm15-38241517	3.74E-14	0.25974
QTL01.grp943.Gm15-38276224	Gm15-38276224	1.68E-11	0.21166
QTL01.grp943.Gm15-38380213	Gm15-38380213	1.24E-06	0.1161
QTL01.grp943.Gm15-38447399	Gm15-38447399	3.00E-09	0.16864
QTL01.grp943.Gm15-38547411	Gm15-38547411	1.06E-07	0.13788
QTL01.grp943.Gm15-38575222	Gm15-38575222	4.74E-24	0.41551
QTL01.grp943.Gm15-39520837	Gm15-39520837	1.06E-07	0.13788
QTL01.grp943.Gm15-39570425	Gm15-39570425	2.55E-08	0.15026
QTL01.grp943.Gm15-39747988	Gm15-39747988	3.12E-08	0.14853
QTL01.grp943.Gm15-39945568	Gm15-39945568	7.35E-19	0.33828
QTL01.grp943.Gm15-39981697	Gm15-39981697	1.91E-06	0.11226
QTL01.grp960.Gm15-38241517	Gm15-38241517	2.62E-09	0.16979
QTL01.grp960.Gm15-38575222	Gm15-38575222	5.08E-13	0.23956
QTL01.grp960.Gm15-39945568	Gm15-39945568	5.37E-11	0.2022
QTL01.grp982.Gm15-38276224	Gm15-38276224	7.50E-07	0.1206
QTL01.grp982.Gm15-39945568	Gm15-39945568	1.91E-06	0.11226
QTL01.grp1005.Gm15-39945568	Gm15-39945568	1.73E-07	0.13356
QTL01.grp1025.Gm15-38029609	Gm15-38029609	2.11E-06	0.11137
QTL01.grp1025.Gm15-38241517	Gm15-38241517	3.29E-11	0.20621
QTL01.grp1025.Gm15-38276224	Gm15-38276224	1.52E-11	0.21249
QTL01.grp1025.Gm15-38380213	Gm15-38380213	6.76E-08	0.1418
QTL01.grp1025.Gm15-38447399	Gm15-38447399	1.70E-10	0.19271
QTL01.grp1025.Gm15-38547411	Gm15-38547411	1.90E-06	0.11231
QTL01.grp1025.Gm15-38575222	Gm15-38575222	3.94E-15	0.27676
QTL01.grp1025.Gm15-39520837	Gm15-39520837	1.90E-06	0.11231
QTL01.grp1025.Gm15-39570425	Gm15-39570425	1.50E-07	0.1348
QTL01.grp1025.Gm15-39747988	Gm15-39747988	9.84E-07	0.11818
QTL01.grp1025.Gm15-39945568	Gm15-39945568	3.14E-26	0.44523
QTL01.grp1025.Gm15-39981697	Gm15-39981697	1.14E-07	0.13727
QTL01.grp1025.Gm15-40010479	Gm15-40010479	1.11E-06	0.11709
QTL01.grp1033.Gm15-34848274	Gm15-34848274	2.05E-07	0.13209
QTL02.grp45.Gm18-747851	Gm18-747851	9.58E-07	0.11842

QTL-multiple metabolite cluster	SNP	pval	Rsq
QTL02.grp45.Gm18-754053	Gm18-754053	9.58E-07	0.118418
QTL02.grp45.Gm18-867515	Gm18-867515	1.06E-08	0.157882
QTL02.grp46.Gm18-867515	Gm18-867515	1.55E-06	0.114103
QTL02.grp65.Gm18-747851	Gm18-747851	9.99E-07	0.118043
QTL02.grp65.Gm18-754053	Gm18-754053	9.99E-07	0.118043
QTL02.grp65.Gm18-867515	Gm18-867515	4.90E-07	0.124375
QTL02.grp68.Gm18-747851	Gm18-747851	4.03E-07	0.12611
QTL02.grp68.Gm18-754053	Gm18-754053	4.03E-07	0.12611
QTL02.grp68.Gm18-867515	Gm18-867515	1.28E-06	0.11584
QTL02.grp84.Gm18-747851	Gm18-747851	7.15E-08	0.141316
QTL02.grp84.Gm18-754053	Gm18-754053	7.15E-08	0.141316
QTL02.grp84.Gm18-867515	Gm18-867515	1.37E-09	0.175291
QTL02.grp86.Gm18-747851	Gm18-747851	1.28E-06	0.115806
QTL02.grp86.Gm18-754053	Gm18-754053	1.28E-06	0.115806
QTL02.grp86.Gm18-867515	Gm18-867515	1.15E-06	0.11681
QTL02.grp98.Gm18-747851	Gm18-747851	9.14E-11	0.197841
QTL02.grp98.Gm18-754053	Gm18-754053	9.14E-11	0.197841
QTL02.grp98.Gm18-867515	Gm18-867515	1.27E-09	0.175939
QTL02.grp98.Gm18-963384	Gm18-963384	5.90E-07	0.122729
QTL02.grp107.Gm18-747851	Gm18-747851	5.13E-07	0.123977
QTL02.grp107.Gm18-754053	Gm18-754053	5.13E-07	0.123977
QTL02.grp107.Gm18-867515	Gm18-867515	1.28E-08	0.156238
QTL02.grp125.Gm18-703188	Gm18-703188	1.40E-08	0.155472
QTL02.grp134.Gm18-867515	Gm18-867515	3.25E-07	0.128003
QTL02.grp169.Gm18-747851	Gm18-747851	6.34E-08	0.142357
QTL02.grp169.Gm18-754053	Gm18-754053	6.34E-08	0.142357
QTL02.grp169.Gm18-867515	Gm18-867515	2.86E-08	0.149289
QTL02.grp170.Gm18-703188	Gm18-703188	6.47E-07	0.121913
QTL02.grp170.Gm18-747851	Gm18-747851	2.90E-07	0.129022
QTL02.grp170.Gm18-754053	Gm18-754053	2.90E-07	0.129022
QTL02.grp170.Gm18-867515	Gm18-867515	4.49E-08	0.145373
QTL02.grp179.Gm18-747851	Gm18-747851	1.64E-07	0.134051
QTL02.grp179.Gm18-754053	Gm18-754053	1.64E-07	0.134051
QTL02.grp179.Gm18-867515	Gm18-867515	3.43E-09	0.167491
QTL02.grp179.Gm18-963384	Gm18-963384	8.30E-07	0.119697
QTL02.grp231.Gm18-703188	Gm18-703188	2.46E-10	0.189659
QTL02.grp231.Gm18-867515	Gm18-867515	6.97E-07	0.121246
QTL02.grp245.Gm18-867515	Gm18-867515	1.32E-06	0.115585
QTL02.grp263.Gm18-867515	Gm18-867515	3.52E-07	0.127312
QTL02.grp287.Gm18-747851	Gm18-747851	1.42E-06	0.114913
QTL02.grp287.Gm18-754053	Gm18-754053	1.42E-06	0.114913
QTL02.grp287.Gm18-867515	Gm18-867515	1.45E-06	0.114724
QTL02.grp350.Gm18-747851	Gm18-747851	5.13E-07	0.123965
QTL02.grp350.Gm18-754053	Gm18-754053	5.13E-07	0.123965
QTL02.grp355.Gm18-867515	Gm18-867515	9.10E-07	0.118879
QTL02.grp419.Gm18-703188	Gm18-703188	6.45E-09	0.162105
QTL02.grp419.Gm18-867515	Gm18-867515	1.96E-06	0.112037
QTL02.grp432.Gm18-747851	Gm18-747851	7.65E-07	0.120424
QTL02.grp432.Gm18-754053	Gm18-754053	7.65E-07	0.120424
QTL02.grp432.Gm18-867515	Gm18-867515	1.24E-06	0.116148
QTL02.grp436.Gm18-867515	Gm18-867515	4.34E-07	0.125443
QTL02.grp453.Gm18-747851	Gm18-747851	1.45E-06	0.114736
QTL02.grp453.Gm18-754053	Gm18-754053	1.45E-06	0.114736
QTL02.grp453.Gm18-867515	Gm18-867515	1.09E-07	0.137605
QTL02.grp458.Gm18-703188	Gm18-703188	1.84E-06	0.112576
QTL02.grp458.Gm18-747851	Gm18-747851	6.74E-07	0.121545
QTL02.grp458.Gm18-754053	Gm18-754053	6.74E-07	0.121545
QTL02.grp458.Gm18-867515	Gm18-867515	3.00E-07	0.128728
QTL02.grp477.Gm18-747851	Gm18-747851	1.11E-06	0.117084
QTL02.grp477.Gm18-754053	Gm18-754053	1.11E-06	0.117084

QTL-multiple metabolite cluster	SNP	pval	Rsq
QTL02.grp997.Gm18-867515	Gm18-867515	1.93E-07	0.132601
QTL02.grp1026.Gm18-747851	Gm18-747851	4.44E-07	0.125248
QTL02.grp1026.Gm18-754053	Gm18-754053	4.44E-07	0.125248
QTL02.grp1026.Gm18-867515	Gm18-867515	1.37E-06	0.115242
QTL02.grp1039.Gm18-703188	Gm18-703188	9.86E-08	0.138506
QTL02.grp1039.Gm18-747851	Gm18-747851	2.51E-08	0.150414
QTL02.grp1039.Gm18-754053	Gm18-754053	2.51E-08	0.150414
QTL02.grp1039.Gm18-867515	Gm18-867515	2.11E-09	0.171607
QTL02.grp1039.Gm18-963384	Gm18-963384	7.70E-07	0.120357
QTL02.grp1106.Gm18-747851	Gm18-747851	1.17E-06	0.116611
QTL02.grp1106.Gm18-754053	Gm18-754053	1.17E-06	0.116611
QTL02.grp1106.Gm18-867515	Gm18-867515	1.80E-06	0.112807
QTL02.grp1124.Gm18-747851	Gm18-747851	1.64E-06	0.113627
QTL02.grp1124.Gm18-754053	Gm18-754053	1.64E-06	0.113627
QTL03.grp125.Gm06-47356973	Gm06-47356973	4.83E-07	0.124504
QTL03.grp125.Gm06-47473722	Gm06-47473722	1.66E-07	0.133932
QTL03.grp170.Gm06-47473722	Gm06-47473722	1.86E-06	0.112506
QTL03.grp231.Gm06-47356973	Gm06-47356973	7.29E-09	0.161058
QTL03.grp231.Gm06-47473722	Gm06-47473722	2.04E-09	0.171897
QTL03.grp378.Gm06-47431561	Gm06-47431561	3.76E-07	0.12673
QTL03.grp378.Gm06-47431886	Gm06-47431886	1.83E-08	0.153165
QTL03.grp378.Gm06-47433724	Gm06-47433724	8.67E-07	0.11931
QTL03.grp419.Gm06-47356973	Gm06-47356973	1.26E-07	0.136354
QTL03.grp419.Gm06-47473722	Gm06-47473722	1.90E-08	0.15281
QTL03.grp495.Gm06-47431886	Gm06-47431886	2.58E-07	0.130048
QTL03.grp552.Gm06-47356973	Gm06-47356973	2.43E-09	0.170414
QTL03.grp552.Gm06-47473722	Gm06-47473722	2.72E-10	0.18881
QTL03.grp632.Gm06-47431561	Gm06-47431561	3.03E-14	0.261357
QTL03.grp632.Gm06-47431886	Gm06-47431886	3.25E-20	0.359353
QTL03.grp632.Gm06-47433724	Gm06-47433724	4.31E-08	0.145715
QTL03.grp651.Gm06-47356973	Gm06-47356973	1.31E-08	0.156012
QTL03.grp651.Gm06-47473722	Gm06-47473722	5.91E-09	0.162849
QTL03.grp689.Gm06-47356973	Gm06-47356973	5.06E-08	0.144326
QTL03.grp689.Gm06-47473722	Gm06-47473722	1.98E-08	0.152481
QTL03.grp738.Gm06-47431561	Gm06-47431561	5.23E-12	0.221073
QTL03.grp738.Gm06-47431886	Gm06-47431886	1.28E-17	0.318377
QTL03.grp738.Gm06-47433724	Gm06-47433724	9.94E-08	0.138431
QTL03.grp746.Gm06-47431561	Gm06-47431561	1.31E-09	0.175622
QTL03.grp746.Gm06-47431886	Gm06-47431886	9.81E-17	0.303868
QTL03.grp746.Gm06-47433724	Gm06-47433724	1.33E-06	0.115502
QTL03.grp826.Gm06-47431561	Gm06-47431561	3.78E-13	0.241883
QTL03.grp826.Gm06-47431886	Gm06-47431886	1.65E-20	0.363821
QTL03.grp826.Gm06-47433724	Gm06-47433724	4.14E-08	0.146069
QTL03.grp840.Gm06-47431561	Gm06-47431561	1.29E-06	0.115735
QTL03.grp840.Gm06-47431886	Gm06-47431886	1.56E-11	0.212252
QTL03.grp843.Gm06-47431561	Gm06-47431561	1.68E-13	0.248173
QTL03.grp843.Gm06-47431886	Gm06-47431886	7.08E-26	0.440511
QTL03.grp843.Gm06-47433724	Gm06-47433724	1.71E-07	0.133662
QTL03.grp846.Gm06-47473722	Gm06-47473722	1.76E-06	0.112985
QTL03.grp905.Gm06-47431886	Gm06-47431886	1.09E-06	0.117236
QTL03.grp937.Gm06-47431561	Gm06-47431561	3.36E-10	0.187046
QTL03.grp937.Gm06-47431886	Gm06-47431886	4.52E-18	0.325702
QTL03.grp993.Gm06-47431561	Gm06-47431561	1.76E-09	0.173149
QTL03.grp993.Gm06-47431886	Gm06-47431886	8.23E-17	0.305129
QTL03.grp993.Gm06-47433724	Gm06-47433724	3.52E-08	0.147472
QTL03.grp995.Gm06-47356973	Gm06-47356973	2.58E-07	0.130051
QTL03.grp995.Gm06-47473722	Gm06-47473722	1.54E-07	0.134591
QTL03.grp1020.Gm06-47431561	Gm06-47431561	4.98E-07	0.124224
QTL03.grp1020.Gm06-47431886	Gm06-47431886	8.25E-10	0.179542
QTL03.grp1039.Gm06-47356973	Gm06-47356973	1.02E-06	0.117887

QTL-multiple metabolite cluster	SNP	pval	Rsq
QTL03.grp1039.Gm06-47473722	Gm06-47473722	3.94E-07	0.12631
QTL03.grp1075.Gm06-47431561	Gm06-47431561	2.04E-07	0.13211
QTL03.grp1075.Gm06-47431886	Gm06-47431886	3.39E-13	0.24273
QTL03.grp1100.Gm06-47431561	Gm06-47431561	2.63E-09	0.16977
QTL03.grp1100.Gm06-47431886	Gm06-47431886	3.02E-12	0.22548
QTL04.grp3.Gm18-11294665	Gm18-11294665	3.11E-08	0.14856
QTL04.grp3.Gm18-11296601	Gm18-11296601	9.79E-08	0.13857
QTL04.grp4.Gm18-11294665	Gm18-11294665	1.40E-07	0.13541
QTL04.grp4.Gm18-11296601	Gm18-11296601	3.48E-07	0.1274
QTL04.grp9.Gm18-11294665	Gm18-11294665	3.40E-08	0.14778
QTL04.grp9.Gm18-11296601	Gm18-11296601	1.08E-07	0.13774
QTL04.grp11.Gm18-11294665	Gm18-11294665	1.52E-08	0.15475
QTL04.grp11.Gm18-11296601	Gm18-11296601	3.72E-08	0.147
QTL04.grp19.Gm18-11294665	Gm18-11294665	1.49E-07	0.13488
QTL04.grp19.Gm18-11296601	Gm18-11296601	3.67E-07	0.12693
QTL04.grp97.Gm18-11553151	Gm18-11553151	8.21E-08	0.14011
QTL04.grp112.Gm18-11553151	Gm18-11553151	7.52E-09	0.16079
QTL04.grp131.Gm18-11054301	Gm18-11054301	8.46E-07	0.11952
QTL04.grp131.Gm18-11072903	Gm18-11072903	2.07E-06	0.11152
QTL04.grp131.Gm18-11080478	Gm18-11080478	8.86E-07	0.11912
QTL04.grp131.Gm18-11112407	Gm18-11112407	1.88E-07	0.13285
QTL04.grp131.Gm18-11165192	Gm18-11165192	9.56E-11	0.19747
QTL04.grp131.Gm18-11193191	Gm18-11193191	2.53E-10	0.18943
QTL04.grp131.Gm18-11199726	Gm18-11199726	2.53E-10	0.18943
QTL04.grp131.Gm18-11287106	Gm18-11287106	5.85E-10	0.18243
QTL04.grp131.Gm18-11292588	Gm18-11292588	2.38E-08	0.15089
QTL04.grp131.Gm18-11294665	Gm18-11294665	6.30E-10	0.1818
QTL04.grp131.Gm18-11296601	Gm18-11296601	2.17E-09	0.1714
QTL04.grp131.Gm18-11320837	Gm18-11320837	5.85E-10	0.18243
QTL04.grp131.Gm18-11367109	Gm18-11367109	8.65E-08	0.13965
QTL04.grp131.Gm18-11383891	Gm18-11383891	9.52E-11	0.1975
QTL04.grp131.Gm18-11386229	Gm18-11386229	1.44E-10	0.19409
QTL04.grp131.Gm18-11408064	Gm18-11408064	9.52E-11	0.1975
QTL04.grp131.Gm18-11492415	Gm18-11492415	8.29E-07	0.1197
QTL04.grp131.Gm18-11553151	Gm18-11553151	2.13E-10	0.19083
QTL04.grp167.Gm18-11165192	Gm18-11165192	1.30E-06	0.11572
QTL04.grp167.Gm18-11193191	Gm18-11193191	1.58E-06	0.11396
QTL04.grp167.Gm18-11199726	Gm18-11199726	1.58E-06	0.11396
QTL04.grp167.Gm18-11294665	Gm18-11294665	9.92E-07	0.11811
QTL04.grp167.Gm18-11296601	Gm18-11296601	1.24E-06	0.11612
QTL04.grp167.Gm18-11383891	Gm18-11383891	1.63E-06	0.11368
QTL04.grp167.Gm18-11386229	Gm18-11386229	6.08E-07	0.12246
QTL04.grp167.Gm18-11408064	Gm18-11408064	1.63E-06	0.11368
QTL04.grp167.Gm18-11553151	Gm18-11553151	1.44E-09	0.17484
QTL04.grp189.Gm18-11292588	Gm18-11292588	1.94E-07	0.13257
QTL04.grp189.Gm18-11294665	Gm18-11294665	5.71E-12	0.22037
QTL04.grp189.Gm18-11296601	Gm18-11296601	2.67E-11	0.2079
QTL04.grp246.Gm18-11232721	Gm18-11232721	4.02E-07	0.12613
QTL04.grp246.Gm18-11277759	Gm18-11277759	6.88E-07	0.12136
QTL04.grp246.Gm18-11292588	Gm18-11292588	1.37E-07	0.13564
QTL04.grp246.Gm18-11294665	Gm18-11294665	4.94E-12	0.22154
QTL04.grp246.Gm18-11296601	Gm18-11296601	2.07E-11	0.20999
QTL04.grp258.Gm18-11294665	Gm18-11294665	1.29E-08	0.15614
QTL04.grp258.Gm18-11296601	Gm18-11296601	3.08E-08	0.14864
QTL04.grp269.Gm18-11294665	Gm18-11294665	1.51E-08	0.1548
QTL04.grp269.Gm18-11296601	Gm18-11296601	2.78E-08	0.14953
QTL04.grp279.Gm18-11294665	Gm18-11294665	4.30E-07	0.12553
QTL04.grp279.Gm18-11296601	Gm18-11296601	1.10E-06	0.11715
QTL04.grp282.Gm18-11292588	Gm18-11292588	4.61E-07	0.12491
QTL04.grp282.Gm18-11294665	Gm18-11294665	2.09E-11	0.20991

QTL-multiple metabolite cluster	SNP	pval	Rsq
QTL04.grp282.Gm18-11296601	Gm18-11296601	9.71E-11	0.197342
QTL04.grp298.Gm18-11294665	Gm18-11294665	4.02E-07	0.126135
QTL04.grp298.Gm18-11296601	Gm18-11296601	1.03E-06	0.117771
QTL04.grp316.Gm18-11294665	Gm18-11294665	1.02E-08	0.158168
QTL04.grp316.Gm18-11296601	Gm18-11296601	2.89E-08	0.149194
QTL04.grp455.Gm18-11292588	Gm18-11292588	1.15E-06	0.116751
QTL04.grp455.Gm18-11294665	Gm18-11294665	4.71E-10	0.184246
QTL04.grp455.Gm18-11296601	Gm18-11296601	1.72E-09	0.173352
QTL04.grp486.Gm18-11112407	Gm18-11112407	2.14E-06	0.111216
QTL04.grp486.Gm18-11165192	Gm18-11165192	4.57E-08	0.145218
QTL04.grp486.Gm18-11193191	Gm18-11193191	1.87E-07	0.132892
QTL04.grp486.Gm18-11199726	Gm18-11199726	1.87E-07	0.132892
QTL04.grp486.Gm18-11287106	Gm18-11287106	3.20E-07	0.128148
QTL04.grp486.Gm18-11320837	Gm18-11320837	3.20E-07	0.128148
QTL04.grp486.Gm18-11383891	Gm18-11383891	4.88E-07	0.124411
QTL04.grp486.Gm18-11386229	Gm18-11386229	8.23E-07	0.119769
QTL04.grp486.Gm18-11408064	Gm18-11408064	4.88E-07	0.124411
QTL04.grp534.Gm18-11054301	Gm18-11054301	1.41E-06	0.114944
QTL04.grp534.Gm18-11112407	Gm18-11112407	1.55E-06	0.114121
QTL04.grp534.Gm18-11165192	Gm18-11165192	1.21E-10	0.195529
QTL04.grp534.Gm18-11193191	Gm18-11193191	7.35E-10	0.180513
QTL04.grp534.Gm18-11199726	Gm18-11199726	7.35E-10	0.180513
QTL04.grp534.Gm18-11287106	Gm18-11287106	2.18E-09	0.171333
QTL04.grp534.Gm18-11296601	Gm18-11296601	1.42E-06	0.1149
QTL04.grp534.Gm18-11320837	Gm18-11320837	2.18E-09	0.171333
QTL04.grp534.Gm18-11367109	Gm18-11367109	4.73E-07	0.124691
QTL04.grp534.Gm18-11383891	Gm18-11383891	7.44E-10	0.180407
QTL04.grp534.Gm18-11386229	Gm18-11386229	1.36E-09	0.175308
QTL04.grp534.Gm18-11408064	Gm18-11408064	7.44E-10	0.180407
QTL04.grp534.Gm18-11492415	Gm18-11492415	7.60E-07	0.120484
QTL04.grp600.Gm18-11553151	Gm18-11553151	1.60E-07	0.134283
QTL04.grp641.Gm18-11553151	Gm18-11553151	4.95E-07	0.124295
QTL04.grp650.Gm18-11294665	Gm18-11294665	7.69E-07	0.120375
QTL04.grp650.Gm18-11296601	Gm18-11296601	1.25E-06	0.116061
QTL04.grp696.Gm18-11356887	Gm18-11356887	1.98E-06	0.111919
QTL04.grp701.Gm18-11294665	Gm18-11294665	6.54E-07	0.121814
QTL04.grp701.Gm18-11296601	Gm18-11296601	1.08E-06	0.117313
QTL04.grp714.Gm18-11294665	Gm18-11294665	5.98E-07	0.122608
QTL04.grp714.Gm18-11296601	Gm18-11296601	1.53E-06	0.114253
QTL04.grp727.Gm18-11294665	Gm18-11294665	1.87E-07	0.132904
QTL04.grp727.Gm18-11296601	Gm18-11296601	3.28E-07	0.127918
QTL04.grp776.Gm18-11294665	Gm18-11294665	1.01E-09	0.177809
QTL04.grp776.Gm18-11296601	Gm18-11296601	3.60E-09	0.167086
QTL04.grp805.Gm18-11294665	Gm18-11294665	2.64E-07	0.129846
QTL04.grp805.Gm18-11296601	Gm18-11296601	5.38E-07	0.123543
QTL04.grp831.Gm18-11294665	Gm18-11294665	2.28E-07	0.131128
QTL04.grp831.Gm18-11296601	Gm18-11296601	5.08E-07	0.124053
QTL04.grp924.Gm18-11277759	Gm18-11277759	5.95E-07	0.122654
QTL04.grp924.Gm18-11292588	Gm18-11292588	8.91E-08	0.139385
QTL04.grp924.Gm18-11294665	Gm18-11294665	2.24E-12	0.227862
QTL04.grp924.Gm18-11296601	Gm18-11296601	1.10E-11	0.215136
QTL04.grp946.Gm18-11277759	Gm18-11277759	5.72E-07	0.123002
QTL04.grp946.Gm18-11292588	Gm18-11292588	8.69E-08	0.139605
QTL04.grp946.Gm18-11294665	Gm18-11294665	2.23E-12	0.227869
QTL04.grp946.Gm18-11296601	Gm18-11296601	1.04E-11	0.215564
QTL04.grp970.Gm18-11277759	Gm18-11277759	2.15E-06	0.111176
QTL04.grp970.Gm18-11292588	Gm18-11292588	2.75E-07	0.129488
QTL04.grp970.Gm18-11294665	Gm18-11294665	9.95E-12	0.215907
QTL04.grp970.Gm18-11296601	Gm18-11296601	3.39E-11	0.20596
QTL04.grp1003.Gm18-11054301	Gm18-11054301	6.66E-07	0.12165

QTL-multiple metabolite cluster	SNP	pval	Rsq
QTL04.grp1003.Gm18-11072903	Gm18-11072903	6.97E-07	0.1212428
QTL04.grp1003.Gm18-11080478	Gm18-11080478	2.19E-06	0.1110229
QTL04.grp1003.Gm18-11112407	Gm18-11112407	3.87E-07	0.1264725
QTL04.grp1003.Gm18-11165192	Gm18-11165192	3.87E-09	0.1664705
QTL04.grp1003.Gm18-11193191	Gm18-11193191	2.58E-09	0.1699083
QTL04.grp1003.Gm18-11199726	Gm18-11199726	2.58E-09	0.1699083
QTL04.grp1003.Gm18-11287106	Gm18-11287106	1.27E-07	0.1362508
QTL04.grp1003.Gm18-11320837	Gm18-11320837	1.27E-07	0.1362508
QTL04.grp1003.Gm18-11383891	Gm18-11383891	2.83E-08	0.149387
QTL04.grp1003.Gm18-11386229	Gm18-11386229	1.48E-08	0.1549915
QTL04.grp1003.Gm18-11408064	Gm18-11408064	2.83E-08	0.149387
QTL04.grp1003.Gm18-11553151	Gm18-11553151	2.53E-07	0.1302242
QTL04.grp1015.Gm18-11232721	Gm18-11232721	5.95E-09	0.1627913
QTL04.grp1079.Gm18-11553151	Gm18-11553151	8.41E-08	0.1398978
QTL05.grp166.Gm02-5192361	Gm02-5192361	9.64E-07	0.118365
QTL05.grp166.Gm02-5193024	Gm02-5193024	9.64E-07	0.118365
QTL05.grp237.Gm02-5192361	Gm02-5192361	5.04E-07	0.1241272
QTL05.grp237.Gm02-5193024	Gm02-5193024	5.04E-07	0.1241272
QTL05.grp245.Gm02-5192361	Gm02-5192361	1.47E-06	0.1145768
QTL05.grp245.Gm02-5193024	Gm02-5193024	1.47E-06	0.1145768
QTL05.grp263.Gm02-5192361	Gm02-5192361	1.08E-06	0.1173432
QTL05.grp263.Gm02-5193024	Gm02-5193024	1.08E-06	0.1173432
QTL05.grp302.Gm02-5192361	Gm02-5192361	1.17E-06	0.1166
QTL05.grp302.Gm02-5193024	Gm02-5193024	1.17E-06	0.1166
QTL05.grp466.Gm02-5166919	Gm02-5166919	4.73E-07	0.1246799
QTL05.grp500.Gm02-5192361	Gm02-5192361	1.09E-07	0.1375936
QTL05.grp500.Gm02-5193024	Gm02-5193024	1.09E-07	0.1375936
QTL05.grp671.Gm02-5166919	Gm02-5166919	1.56E-06	0.1140826
QTL05.grp748.Gm02-5166919	Gm02-5166919	4.16E-07	0.1258197
QTL05.grp795.Gm02-5179693	Gm02-5179693	1.55E-06	0.1141125
QTL05.grp827.Gm02-5166919	Gm02-5166919	3.33E-07	0.127801
QTL05.grp1065.Gm02-5166919	Gm02-5166919	5.60E-07	0.123187
QTL05.grp1103.Gm02-5192361	Gm02-5192361	2.82E-07	0.1292537
QTL05.grp1103.Gm02-5193024	Gm02-5193024	2.82E-07	0.1292537
QTL06.grp45.Gm08-1552810	Gm08-1552810	4.85E-07	0.1244749
QTL06.grp84.Gm08-1552810	Gm08-1552810	3.17E-07	0.1282195
QTL06.grp107.Gm08-1552810	Gm08-1552810	4.85E-07	0.1244609
QTL06.grp124.Gm08-1552810	Gm08-1552810	3.56E-07	0.127199
QTL06.grp169.Gm08-1552810	Gm08-1552810	3.86E-07	0.1264927
QTL06.grp261.Gm08-1552810	Gm08-1552810	5.91E-07	0.1227101
QTL06.grp278.Gm08-1552810	Gm08-1552810	1.17E-06	0.1166279
QTL06.grp355.Gm08-1552810	Gm08-1552810	7.43E-07	0.1206748
QTL06.grp436.Gm08-1552810	Gm08-1552810	2.03E-06	0.1117106
QTL06.grp499.Gm08-1552810	Gm08-1552810	1.62E-06	0.1137461
QTL06.grp504.Gm08-1552810	Gm08-1552810	4.11E-08	0.1461375
QTL06.grp523.Gm08-1552810	Gm08-1552810	3.81E-08	0.1468004
QTL06.grp532.Gm08-1552810	Gm08-1552810	4.77E-07	0.124619
QTL06.grp576.Gm08-1552810	Gm08-1552810	6.30E-07	0.1221444
QTL06.grp616.Gm08-1552810	Gm08-1552810	6.46E-08	0.1422035
QTL06.grp634.Gm08-1552810	Gm08-1552810	7.80E-08	0.1405514
QTL06.grp673.Gm08-1552810	Gm08-1552810	5.86E-08	0.143042
QTL06.grp690.Gm08-1552810	Gm08-1552810	1.36E-06	0.1152706
QTL06.grp720.Gm08-1552810	Gm08-1552810	1.01E-08	0.1582399
QTL06.grp734.Gm08-1552810	Gm08-1552810	2.29E-07	0.13109
QTL06.grp803.Gm08-1552810	Gm08-1552810	1.69E-06	0.1133537
QTL06.grp895.Gm08-1552810	Gm08-1552810	5.35E-08	0.1438353
QTL06.grp899.Gm08-1552810	Gm08-1552810	7.38E-09	0.1609587
QTL06.grp927.Gm08-1552810	Gm08-1552810	1.51E-08	0.154787
QTL07.grp84.Gm11-30110423	Gm11-30110423	5.95E-07	0.1226556
QTL07.grp84.Gm11-30113492	Gm11-30113492	3.55E-07	0.1272237

QTL-multiple metabolite cluster	SNP	pval	Rsq
QTL07.grp84.Gm11-30114287	Gm11-30114287	7.82E-07	0.120228
QTL07.grp98.Gm11-30114287	Gm11-30114287	4.30E-07	0.125544
QTL07.grp169.Gm11-30110423	Gm11-30110423	3.42E-08	0.147744
QTL07.grp169.Gm11-30113492	Gm11-30113492	5.47E-08	0.143649
QTL07.grp169.Gm11-30114287	Gm11-30114287	1.71E-07	0.133687
QTL07.grp354.Gm11-30110423	Gm11-30110423	4.78E-07	0.12459
QTL07.grp355.Gm11-30110423	Gm11-30110423	2.99E-09	0.168668
QTL07.grp436.Gm11-30110423	Gm11-30110423	2.94E-07	0.12889
QTL07.grp436.Gm11-30113492	Gm11-30113492	1.19E-06	0.116518
QTL07.grp504.Gm11-30110423	Gm11-30110423	1.67E-06	0.113459
QTL07.grp504.Gm11-30113492	Gm11-30113492	7.18E-07	0.120983
QTL07.grp611.Gm11-30154190	Gm11-30154190	5.04E-08	0.144367
QTL07.grp690.Gm11-30110423	Gm11-30110423	4.11E-09	0.16596
QTL07.grp734.Gm11-30110423	Gm11-30110423	9.34E-07	0.118644
QTL07.grp734.Gm11-30113492	Gm11-30113492	1.24E-06	0.116123
QTL07.grp734.Gm11-30114287	Gm11-30114287	1.23E-06	0.116198
QTL07.grp794.Gm11-30114287	Gm11-30114287	1.36E-06	0.115295
QTL07.grp810.Gm11-30110423	Gm11-30110423	8.21E-08	0.140106
QTL07.grp810.Gm11-30113492	Gm11-30113492	8.12E-08	0.140198
QTL07.grp810.Gm11-30114287	Gm11-30114287	1.07E-06	0.117446
QTL07.grp927.Gm11-30110423	Gm11-30110423	1.02E-07	0.138207
QTL07.grp927.Gm11-30113492	Gm11-30113492	3.33E-07	0.127793
QTL07.grp944.Gm11-30110423	Gm11-30110423	3.57E-08	0.147361
QTL07.grp944.Gm11-30113492	Gm11-30113492	6.08E-09	0.162608
QTL07.grp944.Gm11-30114287	Gm11-30114287	8.41E-09	0.159835
QTL07.grp944.Gm11-30214512	Gm11-30214512	2.08E-06	0.111493
QTL07.grp961.Gm11-30110423	Gm11-30110423	1.32E-07	0.135954
QTL07.grp961.Gm11-30113492	Gm11-30113492	2.01E-07	0.132233
QTL07.grp1124.Gm11-30114287	Gm11-30114287	1.22E-06	0.116258
QTL08.grp279.Gm04-37366196	Gm04-37366196	2.17E-07	0.131583
QTL08.grp298.Gm04-37366196	Gm04-37366196	2.52E-07	0.130244
QTL08.grp714.Gm04-37366196	Gm04-37366196	2.07E-08	0.152084
QTL08.grp841.Gm04-37366196	Gm04-37366196	1.02E-08	0.15816
QTL08.grp845.Gm04-37366196	Gm04-37366196	2.20E-08	0.151556
QTL08.grp893.Gm04-37366196	Gm04-37366196	3.31E-07	0.127855
QTL08.grp916.Gm04-37068922	Gm04-37068922	1.03E-06	0.117813
QTL08.grp916.Gm04-37129463	Gm04-37129463	1.03E-06	0.117813
QTL08.grp916.Gm04-37187883	Gm04-37187883	1.03E-06	0.117813
QTL08.grp916.Gm04-38194007	Gm04-38194007	1.03E-06	0.117813
QTL08.grp916.Gm04-38658568	Gm04-38658568	1.03E-06	0.117813
QTL08.grp916.Gm04-39265590	Gm04-39265590	1.03E-06	0.117813
QTL08.grp916.Gm04-39378068	Gm04-39378068	1.03E-06	0.117813
QTL08.grp1057.Gm04-36913117	Gm04-36913117	7.65E-09	0.160649
QTL08.grp1057.Gm04-37068922	Gm04-37068922	1.42E-10	0.194224
QTL08.grp1057.Gm04-37129463	Gm04-37129463	1.42E-10	0.194224
QTL08.grp1057.Gm04-37187883	Gm04-37187883	1.42E-10	0.194224
QTL08.grp1057.Gm04-38194007	Gm04-38194007	1.42E-10	0.194224
QTL08.grp1057.Gm04-38658568	Gm04-38658568	1.42E-10	0.194224
QTL08.grp1057.Gm04-39265590	Gm04-39265590	1.42E-10	0.194224
QTL08.grp1057.Gm04-39378068	Gm04-39378068	1.42E-10	0.194224
QTL08.grp1080.Gm04-37068922	Gm04-37068922	2.19E-06	0.111049
QTL08.grp1080.Gm04-37129463	Gm04-37129463	2.19E-06	0.111049
QTL08.grp1080.Gm04-37187883	Gm04-37187883	2.19E-06	0.111049
QTL08.grp1080.Gm04-38194007	Gm04-38194007	2.19E-06	0.111049
QTL08.grp1080.Gm04-38658568	Gm04-38658568	2.19E-06	0.111049
QTL08.grp1080.Gm04-39265590	Gm04-39265590	2.19E-06	0.111049
QTL08.grp1080.Gm04-39378068	Gm04-39378068	2.19E-06	0.111049
QTL08.grp1083.Gm04-37366196	Gm04-37366196	4.11E-10	0.185385

Supplementary Table 3.4 *G. max* ecotypes used for sequence analysis of the *UGT* gene.

PI	Name	Origin
PI 89772	7193	China
PI 438496 A	Peking	United States
PI 548658	Lee 74'	Arkansas, United States
PI 548656	Lee'	Mississippi, United States
DNC	Dunphy NC	
PI 437654	Er-hej-jan	China
PI 548982	Pickett 71'	Mississippi, United States
PI 567548	Hua li hu zi	Shandong Sheng, China
PI 417500	Escura A	Brazil
PI 90763	7570	Beijing Shi, China
PI 641156	NC-Raleigh'	North Carolina, United States
PI 88788	5913	Liaoning Sheng, China
PI 84973	Takiya	Saitama, Japan
PI 548402	Peking'	Beijing Shi, China
PI 86006	Kiio Shokuzu	Hokkaidô, Japan
PI 84631	S-56	Kyonggi, Korea, South
PI 423926	Tousan 72	Nagano, Japan
PI 58955	Common Yellow Va	Shandong Sheng, China
PI 70080	6908	Jilin Sheng, China
PI 81785	Chusei Hadaka	Hokkaidô, Japan
PI 84656	S-81	Kyonggi, Korea, South
PI 92651	7846	Jilin Sheng, China
PI 404166	Krasnoarmejskaja	China
PI 438258	VIR 4714	China
PI 240664	Bilomi No. 3	Luzon, Philippines
PI 84946 -2	(Kandokon)	Busan-gwangyeoksi, Korea, South
PI 200508	Natsu Daizu	Japan
PI 88468	Iganzu	Liaoning Sheng, China
PI 438323	Grignon 53-F-3	France
PI 86904 -1	Fukota	Chungcheongbuk-do, Korea, South
PI 86972 -2	(Pakute)	Jeollabuk-do, Korea, South
PI 361093	Novosadska Br. 1	Serbia
PI 417479	Yougetsu	Japan
PI 83881	Orukon	Kangweonto, Korea, North
PI 84656	S-81	Kyonggi, Korea, South

Supplementary Table 3.5 A list of primers (5’ to 3’) for sequence analysis of the *UGT* gene.

Forward Primer	Reverse Primer
GCAAGCATTCCAATCGCTCCCA	CACAGTGCTATGTCCTGAATTGATTGC
ATGGGCTCCTTGATTGTTCCAGGT	TGTGGTTGGCCTTCTGCAGTTTGA
ATGGGCTCCTTGATTGTTCCAGG	CTAACTTTTGTGGTTGGCCTTCTG
ATGGGCTCCTTGATTGTTCCAGG	GAGAATTGTTCAATTGTGTTCTGA
TTCTCACCTCACACCAAGGTA	CTTCACTTTCACCCTTTTTCCTA
TGAAGACGGGGAAGGCAATGACT	CTAACTTTTGTGGTTGGCCTTCTG

Supplementary Table 3.6 A list of primers (5’ to 3’) for sequence analysis of the *UGT* gene promoter region.

Forward Primer	Reverse Primer
GTGAAACTTCATAAAGGGCCCAACGC	GGGAGCGATTGGAATGCTTGCAC
TGAAAGGTCCACCCTGAAAGCCA	AACCACGCGCTTTGTCGGGA
GCTCAAGAGACCACACCATGGCA	
GGTTACCTTGCTCACTCTTCTC	GCTCTTTCTCTCCGAGTCTTTC
TCCCAAGGTGTCGTTTATGG	GGAATGCTTGCACACTGATTT
TAAGGCGAGGAGTGGACATA	CATGTTCACCTGGAACAATCAAG
CACCATAACCTGCCACAGTATAA	TTGCTTCCACTCCACCATAAA
AGCGCGTGGTTCCCATGTGT	ACCTGGAACAATCAAGGAGCCCA
TGCATCCAGCTCTTCGTCCACCT	GGGAGCGATTGGAATGCTTGCAC

Supplementary Table 3.7 A list of primers (5’ to 3’) for the *UGT* gene expression analysis.

Forward Primer	Reverse Primer
GTGTAATGTTGGATGTGTTCCC	ACACAATTGAGTTCAACACAAACCG
CACCAAGGTACTGCTGGATT	GGAAATTCTACCTCTGTCGAATATGA
AGGTACTGCTGGTTACATGAC	ATAATTGGAAATTCTACCTCTGTCG
CCATGTGAACAAGTTGGGTTG	ATGGTGAGTCCCTGGTAGAT
AGTCACCTCATTCCCGTAGTA	GTGGCTGTGGTGGTGATTAT
TACCAGGGACTCACCATTCT	GGTAGAACATGTCGGTGAACA
TAGCTAGGCTCTTTGCCATTC	ACTTGACAACGTGGGTTCTAATA

GAGGGAGATTATGAGGAGCATTAC	CCTCTATCGGCCTTATCCAAAG
TGAATTGAAGATGACGCGTTTG	GCTCTTTCTCTCCGAGTCTTTC
GAGTTTGGAGATGAGGTGGTAAA	GAAGACCCTCCAACCTGAATAG