bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# Title: Genetic basis and selection of glyceollin induction in wild soybean

Farida Yasmin[1#], Hengyou Zhang[1#&], Larry Leamy[1], Baosheng Wang[2,3], Jason Winnike[4], Robert W. Reid[5], Cory R. Brouwer[5] and Bao-Hua Song[1*]

ORCIDs:

F.Y.: https://orcid.org/0000-0002-4621-7796

H.Z.: https://orcid.org/0000-0003-4103-8980

B-H. S.: https://orcid.org/0000-0003-3537-7783

[1]Department of Biological Sciences, The University of North Carolina at Charlotte, NC 28223, USA; [2]Key Laboratory of Plant Resources Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, China; [3]Center of Conservation Biology, Core Botanical Gardens, Chinese Academy of Sciences, Guangzhou 510650, China; [4]David H. Murdock Research Institute, NC Research Campus, Kannapolis, NC 28081, US; [5]Department of Bioinformatics and Genomics, The University of North Carolina at Charlotte, NC 28223, USA.

[&]Current Address: Key Laboratory of Soybean Molecular Design Breeding, Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Harbin 150081, China

[#]These authors contributed equally to this work.

[*]Author for correspondence:

*Bao-Hua Song*

*704.687.5465*

*Email: bsong5@uncc.edu*

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

| | | | |
|---|---|---|---|
| Total word count (excluding summary, references and legends): | 4915 | No. of figures: | 4 (Figs 1-4 in color) |
| Summary: | 180 | No. of Tables: | 2 |
| Introduction: | 999 | No. of Supporting Information files: | 7 (Figs S1-S2 in color; Tables S1-S5) |
| Materials and Methods: | 1086 | | |
| Results: | 828 | | |
| Discussion: | 1839 | | |
| Acknowledgements: | 101 | | |
| Author contributions: | 62 | | |

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

## Summary

- Glyceollins, a family of phytoalexin induced in legume species, play essential roles in responding to environmental stresses and in human health. However, little is known about the genetic basis and selection of glyceollin induction.
- We employed a metabolite-based genome-wide association (mGWA) approach to identify candidate genes involved in glyceollin induction from genetically diverse and understudied wild soybeans subjected to soybean cyst nematode stress.
- Eight SNPs on chromosomes 3, 9, 13, 15, and 20 showed significant association with glyceollin induction. Six genes close to one of the significant SNPs (ss715603454) on chromosome 9 fell into two clusters, and they encode enzymes in the glycosyltransferase class within the phenylpropanoid pathway. Transcription factors (TFs) genes, such as *MYB* and *WRKY* were also found within the linkage disequilibrium of the significant SNPs on chromosome 9. Epistasis and a strong selection signal were detected on the four significant SNPs on chromosome 9.
- Gene clusters and transcription factors may play important roles in regulating glyceollin induction in wild soybeans. Additionally, as major evolutionary factors, epistatic interactions and selection may influence glyceollin variation in natural populations.

## Keywords

Epistasis, Gene cluster, mGWAS, phytoalexin, Plant and human health, Selection, Transcription factors, Wild soybean.

## Abbreviations list:

| | |
|---|---|
| bp | base pair |
| BLINK | bayesian-information and linkage-disequilibrium iteratively nested keyway |
| dpi | days post infection |

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

| | |
|---|---|
| FDR | false discovery rate |
| Fig. (Figs) | figure (figures) |
| LD | linkage disequilibrium |
| LOD | logarithm of the odds |
| Mbp | megabase pair |
| mGWAS | metabolite-based genome-wide association study |
| SNP | single nucleotide polymorphism |
| µm | micromolar |
| µg/g | microgram/gram |

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

## 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 (Donnez et al., 2011, Jahan et al., 2019, Jeandet et al., 2002, Jeandet et al., 2020, Saga et al., 2012, Wang et al., 2009, Yamamura et al., 2015). Isoflavonoids have become a research hot spot 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 (Graham and Graham, 1996, Subramanian et al., 2006). Moreover, they have anti-cancer, antioxidant, and neuroprotective properties (Bamji and Corbitt, 2017, Kim et al., 2012, 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, Pichersky and Gang, 2000, Qi et al., 2004). 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 soybean (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).

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (Nützmann et al., 2016, Töpfer et al., 2017). They have been a common phenomenon since the early days of microbial genetics (Koonin, 2009, Rocha, 2008, Zheng et al., 2002). 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. Systematic 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 (Field et al., 2011, Matsuba et al., 2013, Qi et al., 2004). 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 (Chavali and Rhee, 2018, Nützmann et al., 2016, Takos and Rook, 2012, Yeaman and Whitlock, 2011).

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 (Chan et al., 2010, Chan et al., 2011, Luo, 2015, Riedelsheimer et al., 2012). mGWAS allows the identification of common polymorphic regions controlling complex metabolic traits by substantially increasing association panel and genome-

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 mGWAS study that identified 36 candidate genes influencing the variation of metabolites with physiological and nutritional importance  (Chen et al., 2014).

The isoflavonoid pathway has been relatively well studied (Sukumaran et al., 2018). 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?

## 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 (Table S1). 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 et al., 2017a, Zhang et al., 2017b, Zhang and Song, 2017). Whole root tissues were collected and weighed five days post-infection (dpi). The 5 dpi time point was chosen because

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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., 2017b). All samples were flash frozen in liquid nitrogen and stored at -80 °C. Four biological replicates per wild soybean genotype 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. Peaks that were consistently detected in at least three biological replicates within each genotype 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 $\geqslant 10$ 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%.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

## Metabolite-based genome-wide association study (mGWAS) 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 [6], 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 cutoff (FDR, false discovery rate) = 0.05) (Ge et al.,

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 analyses

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 induction

We identified a total of eight significant SNPs, with four located on chromosome 9 and the others on chromosomes 3, 13, 15, and 20 (Fig. **1a**, Table 1). These SNPs were identified based on both genome-wide Bonferroni threshold of 5.104 and chromosome-wide Bonferroni thresholds that

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

varied narrowly from 3.79 to 3.82 among the 20 chromosomes (3.803 on chromosome 9) (Figs **1a,b**, Table S2). The manhattan and Q-Q (quantile-quantile)  plots are shown in Fig.**1a,b,c**. The four significant SNPs on chromosome 9 are located close to each other within a 535 kb region (Table S2). The broad-sense heritability ($h^2$) was estimated 35% (Table S2).

### 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 (Figs **2b**, **S1, S2**). We considered $r^2$>0.2 as a cutoff for our LD analysis, where $r^2$ is the extent of allelic association between a pair 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 (Figs **2b**, **S1**). 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 (Tables 1, S3, and S4), which may indicate their transcriptional level involvement in glyceollin induction in response to SCN stress (Colinas and Goossens, 2018).

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 (Table S5). Apart from the biosynthetic enzymes on these chromosomes, we also found transcription factor genes, such as *WRKY*, *MYB*, and *NAC* (Table S5).

### 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 fell into two clusters. Both of these two clusters belong to tailoring enzyme glycosyltransferase within phenylpropanoid specialized metabolic domain. And six genes are within the branch of isoflavonoid biosynthesis

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (Table S3).

Further investigation of annotation of these candidate genes within the gene clusters (Table S4), 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 (Table S4). 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. Three probabilities, all associated with the SNP on chromosome 20, were not estimable (Table 2). Among the remaining 25 SNP pairs, 20 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, also are statistically significant. In general, therefore, this is evidence for substantial epistasis among these SNPs affecting GVSD.

These epistatic interactions of the SNP pairs are illustrated in Fig. **3** for each of the four chosen combinations. For example, in panel **a** (Fig. **3a**), 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 (Fig. **3d**), 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.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

### 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 (**Fig. 4**). 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 (**Fig. 4**).

## Discussion

### Metabolic gene clusters in glyceollin induction

Gene clusters have been reported to play important roles in phytochemical diversity in *Arabidopsis*, sorghum, soybean, and rice (Chae et al., 2014). However, their roles in regulating metabolic variation in wild species are relatively less investigated. Even though the isoflavonoid biosynthesis pathway is relatively 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 gene clusters with functionally related but non-homologous genes, which may involve in glyceollin induction in wild soybean. Thus far, these are the first reported plausible gene clusters involved in glyceollin accumulation induced by biotic stimuli. These 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

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (Frey et al., 1997, Gierl and Frey, 2001), the triterpene avenacin in oat (Field and Osbourn, 2008, Mugford et al., 2009, Qi et al., 2004, Qi et al., 2006), 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 *Glycine soja* genotypes 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 (Chu et al., 2011, Osbourn, 2010b).

## 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 highly diverse plant specialized metabolic pathways is dynamic given the ever-changing 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).

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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., 2004, 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 (Shoji and Yuan, 2021, Zhou et al., 2016). 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 coregulate 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**

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (Arita, 2004, Fell and Wagner, 2000, Jeong et al., 2000). 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 (Figs **4, 2b,** and **S1**). 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 (Figs **2b, S1**). 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 (Fig. **2c,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 LC-MS analysis coupled with GWAS (as applied

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (Byrne et al., 1996, Chan et al., 2011, Kliebenstein, 2001, Kliebenstein et al., 2001, Kliebenstein, 2007, McMullen et al., 1998, Petersen et al., 2012, Prasad et al., 2012, Yencho et al., 1998). 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

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

knowledge of plant specialized metabolic gene clusters and regulators will facilitate metabolic engineering with improved metabolic traits for sustainable agriculture and novel pharmaceuticals.

## Acknowledgements:

Research in B-H.S. lab was supported by the National Institute of General Medical Sciences, Award Number: R15GM122029, R15AT011603; North Carolina Biotechnology Center, Award Number: 2019-BIG-6507, 2020-FLG-3806; The North Carolina Soybean Producers Association, Award Number: 18-0252; and the University of North Carolina at Charlotte, Award Number: 111239. Work by F.Y. was (partially) supported by the Schlumberger Foundation Faculty for the Future program.

We thank Thomas Mitchell-Olds for the constructive comments and suggestions. We thank Xu Li and Janice Kofsky for their help with methods and discussions. We also thank Song lab members Melissa Hatley, Judy Janiol, Janice Kofsky and Neha Mittal for their help with planting and sample collection.

## Author contributions

B-H.S. conceived the study and designed the experiment. H.Z., J.W., R.R., C.B. and F.Y. conducted the experiments and collected the data. F.Y., H.Z., L.L., and B.W. performed data analysis. F.Y. drafted the original manuscript with input from . F.Y., H.Z., L.L., B.W., J.W. and B-H.S. reviewed and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

## References

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

ANDERSON, C. J. The role of standing genetic variation in adaptation of digital organisms to a new environment. ALIFE 2012: The Thirteenth International Conference on the Synthesis and Simulation of Living Systems, 2012. MIT Press, 3-10.

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

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.

BARGHI, N., HERMISSON, J. & SCHLÖTTERER, C. 2020. Polygenic adaptation: a unifying framework to understand positive selection. *Nature Reviews Genetics,* 21**,** 769-781.

BARRETT, R. D. & SCHLUTER, D. 2008. Adaptation from standing genetic variation. *Trends in ecology & evolution,* 23**,** 38-44.

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.

BONT, Z., ZÜST, T., ARCE, C. C., HUBER, M. & ERB, M. 2020. Heritable variation in root secondary metabolites is associated with recent climate. *Journal of ecology,* 108**,** 2611-2624.

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.

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.

CÁRDENAS, P. D., SONAWANE, P. D., POLLIER, J., VANDEN BOSSCHE, R., DEWANGAN, V., WEITHORN, E., TAL, L., MEIR, S., ROGACHEV, I. &

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

MALITSKY, S. 2016. *GAME9* regulates the biosynthesis of steroidal alkaloids and upstream isoprenoids in the plant mevalonate pathway. *Nature communications,* 7**,** 1-16.

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., HANSEN, B. G. & KLIEBENSTEIN, D. J. 2010. The complex genetic architecture of the metabolome. *PLoS genetics,* 6**,** e1001198.

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.

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, 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. 2014. Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nature Genetics,* 46**,** 714-721.

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.

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.

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.

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.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

FIELD, B. & OSBOURN, A. E. 2008. Metabolic diversification—independent assembly of operon-like gene clusters in different plants. *Science,* 320**,** 543-547.

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.

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.

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.

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.

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.

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.

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.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

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.

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.

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.

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.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

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., 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.

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

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

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.

LANDER, E. & KRUGLYAK, L. 1995. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature genetics,* 11**,** 241-247.

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.

LI, J. & JI, L. 2005. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity,* 95**,** 221-227.

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.

LUO, J. 2015. Metabolite-based genome-wide association studies in plants. *Current opinion in plant biology,* 24**,** 31-38.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

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.

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.

MESSER, P. W. & PETROV, D. A. 2013. Population genomics of rapid adaptation by soft selective sweeps. *Trends in ecology & evolution,* 28**,** 659-669.

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. 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.

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.

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.

NÜTZMANN, H. W., HUANG, A. & OSBOURN, A. 2016. Plant metabolic clusters–from genetics to genomics. *New phytologist,* 211**,** 771-789.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

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.

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. E. & FIELD, B. 2009. Operons. *Cellular and Molecular Life Sciences,* 66**,** 3755-3775.

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.

PICHERSKY, E. & GANG, D. R. 2000. Genetics and biochemistry of secondary metabolites in plants: an evolutionary perspective. *Trends in plant science,* 5**,** 439-445.

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.

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.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

ROCHA, E. P. 2008. The organization of the bacterial genome. *Annual review of genetics,* 42**,** 211-233.

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.

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.

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.

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.

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.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

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

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.

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.

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.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

TAKOS, A. M. & ROOK, F. 2012. Why biosynthetic genes for chemical defense compounds cluster. *Trends in plant science,* 17**,** 383-388.

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.

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.

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.

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.

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.

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.

XU, Y.-H., WANG, J.-W., WANG, S., WANG, J.-Y. & CHEN, X.-Y. 2004. 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.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

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.

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. & 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. 2017b. Comparative RNA-seq analysis uncovers a complex regulatory network for soybean cyst nematode resistance in wild soybean (*Glycine soja*). *Scientific Reports,* 7**,** 9699.

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, 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.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

## Figure legends

**Fig. 1** 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 $7.86 \times 10^{-6}$ 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 $1.57 \times 10^{-4}$ (threshold 3.803, solid blue line) **(b)**.

**Fig. 2** 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. 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)**.

**Fig. 3** Epistatic interactions of the SNP pairs for each of four chosen combinations. Regression slopes of GVSD on ss715603454 are close to 0 for ss715603454 CC genotypes but are positive for TC and especially TT genotypes **(a)**. Regression slopes of GVSD on ss715603462 are close to 0 for ss715585948 CC genotypes but are negative for TC and especially TT genotypes **(b)**. Regression slopes of GVSD on ss715615975 are close to 0 for ss715585948 TT genotypes 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 genotypes **(d)**.

**Fig. 4** Allele-specific Extended Haplotype Homozygosity (EHH) for four significant SNPs on chromosomes 9.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

## Tables

**Table 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<br>Zinc fingers superfamily protein |
| ss715603454 | Gm09 | UDP-glucosyl transferase 88A1 |
| ss715603455 | Gm09 | *RING/U-box* superfamily protein, *RING/FYVE/PHD* zinc finger superfamily protein |
| ss715603462 | Gm09 | *WRKY* family transcription factor family protein<br>*MYB* domain |
| ss715603471 | Gm09 | Zinc fingers superfamily protein<br>*Cytochrome P450* enzyme family<br>Zinc finger, *RING-type*; Transcription factor jumonji/aspartyl beta-hydroxylase |
| ss715615975 | Gm13 | *bZIP* transcription factor<br>*RING/U-box* superfamily protein, *RING/FYVE/PHD* zinc finger superfamily protein<br>Zinc fingers superfamily protein<br>*NAC* transcription factors<br>*Cytochrome P450* enzyme family |
| ss715620269 | Gm15 | *RING/U-box* superfamily protein, *RING/FYVE/PHD* zinc finger superfamily protein<br>*WRKY* family transcription factor family protein<br>*MYB* domain |

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

**Table 2** Epistasis for the eight significant SNPs.

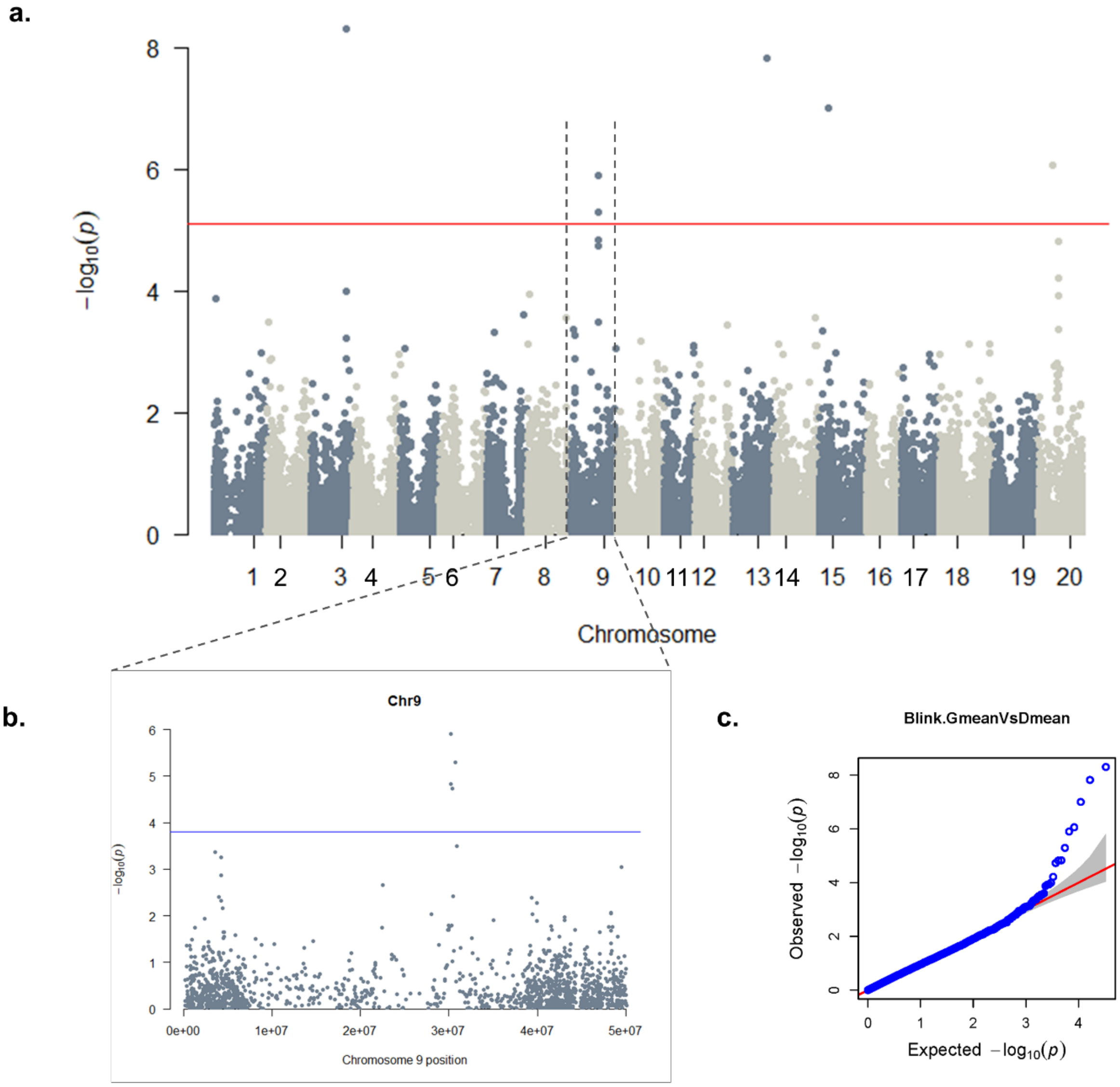
| | **Ch9a** | **Ch9b** | **Ch9c** | **Ch9d** | **Ch13** | **Ch15** | **Ch20** |
|---|---|---|---|---|---|---|---|
| **Ch3** | <0.001* | <0.001* | <0.001* | <0.001* | <0.001* | <0.001* | 0.002* |
| **Ch9a** | | 0.10 | 0.053 | 0.007* | <0.001* | <0.001* | 0.907 |
| **Ch9b** | | | 0.012 | 0.006* | <0.001* | <0.001* | 0.835 |
| **Ch9c** | | | | <0.000* | <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

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

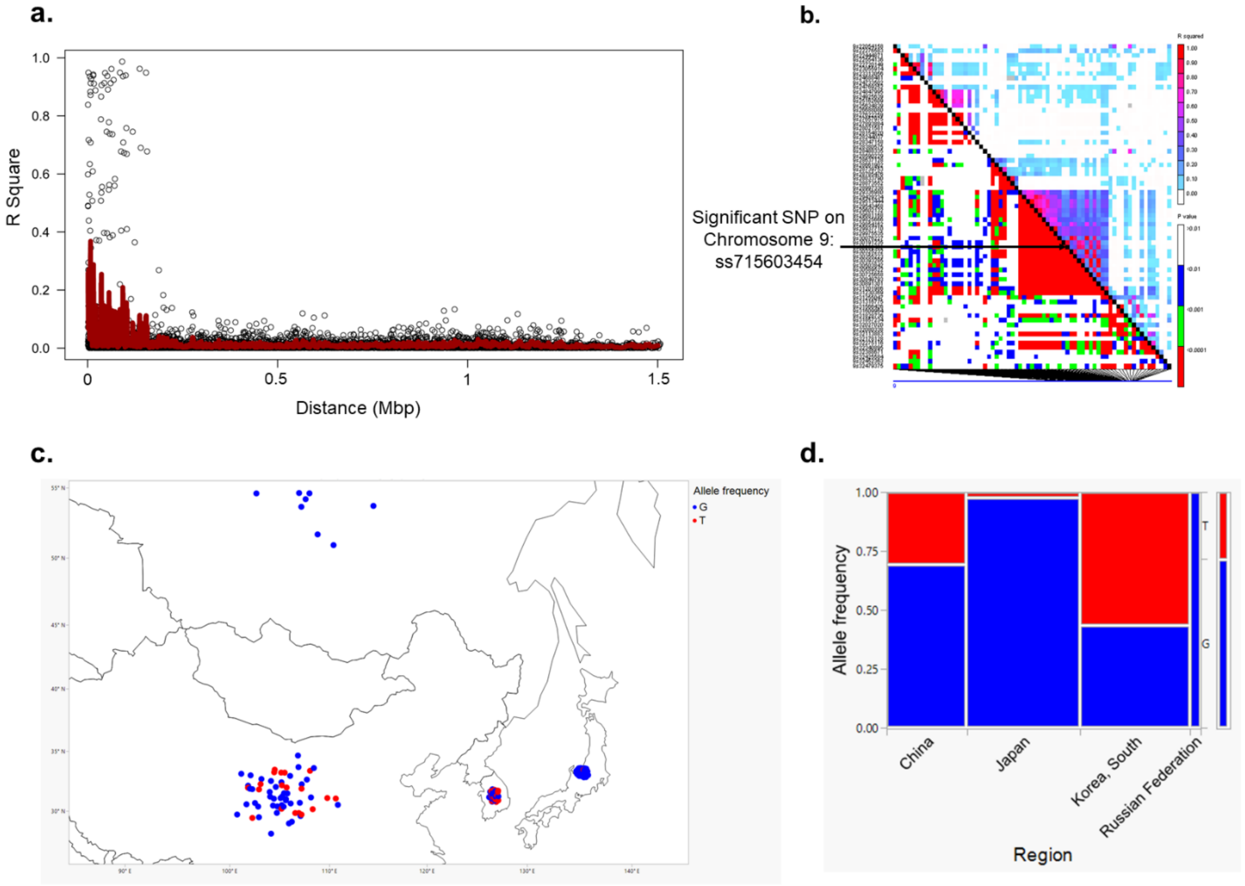
## Figures

## Fig. 1



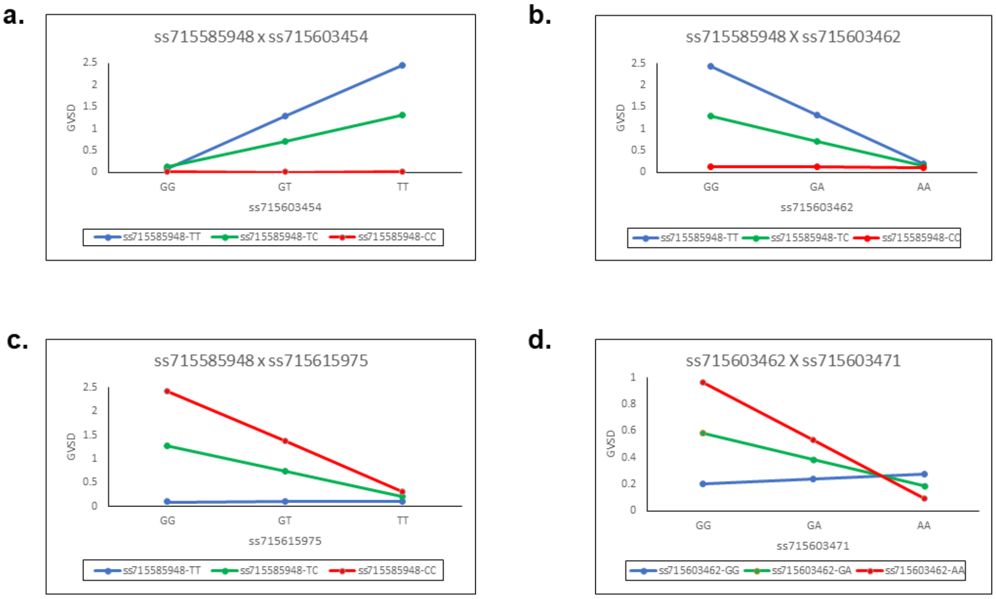
bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

**Fig. 2**



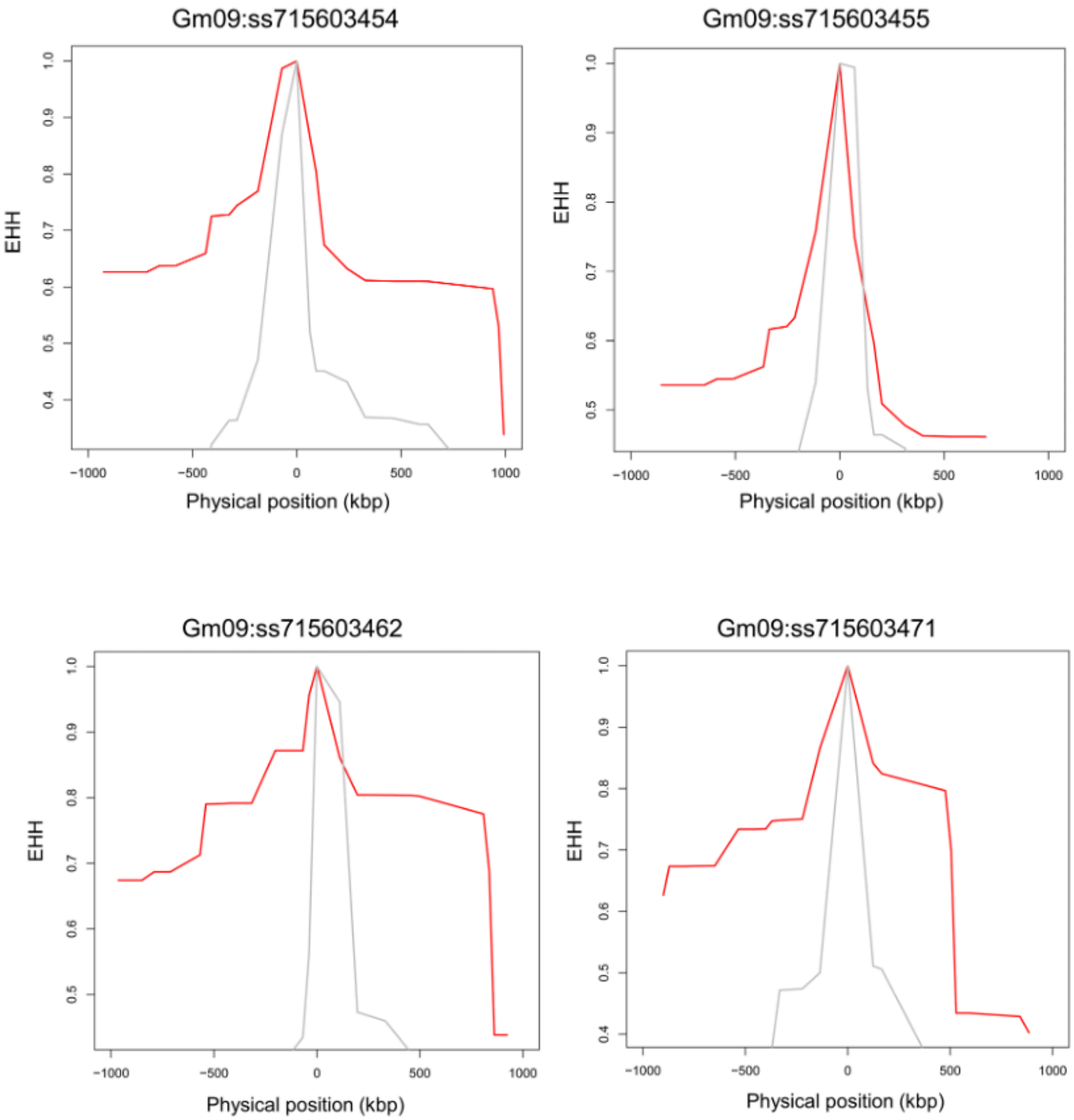
bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

**Fig. 3**



bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520864; this version posted December 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

879 **Fig. 4**



880

881

882