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

27

Total word count (excluding summary, references and legends):	4975	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:	1146		
Results:	828		
Discussion:	1839		
Acknowledgements:	101		
Author contributions:	62		

28

29

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
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

52

53

54 **Introduction**

55 Plants produce diverse specialized metabolites (also known as secondary metabolites or
56 phytochemicals), which play a vital role in adapting to changing environments. Phytoalexins are
57 specialized metabolites synthesized *de novo* in response to various biotic and abiotic stresses.
58 Examples include indole alkaloid camalexin in *Arabidopsis,* phenolic aldehyde gossypol in cotton,
59 phenylpropanoid stilbenes in grapevines, isoflavonoid-derived glyceollins in legume, and
60 momilactones and phytocassanes terpenoids in rice (Donnez et al., 2011, Jahan et al., 2019, Jeandet
61 et al., 2002, Jeandet et al., 2020, Saga et al., 2012, Wang et al., 2009, Yamamura et al., 2015).
62 Isoflavonoids have become a research hot spot due to their various pharmacological properties and
63 essential roles in plant defense. The major isoflavones in soybeans are genistein, daidzein, and
64 glycitein, and they make up about 50%, 40%, and 10%, respectively, of the total isoflavone content.
65 Trace amounts of glyceollins are induced transiently with abiotic and biotic stresses (Jahan et al.,
66 2019, Subramanian et al., 2006). They have multiple effects, including fostering symbiosis
67 between soybean and *Bradyrhizobium japonicum* and inhibiting the growth of various microbes
68 (Graham and Graham, 1996, Subramanian et al., 2006). Moreover, they have anti-cancer,
69 antioxidant, and neuroprotective properties (Bamji and Corbitt, 2017, Kim et al., 2012,
70 Nwachukwu et al., 2013, Seo et al., 2018). However, studies on glyceollins are mainly focused on
71 their medicinal properties, while little is known about how their induction is regulated.

72

73 Phytoalexins have been considered the target of natural selection due to their activities in biotic
74 and abiotic stress responses in natural environments (Miyamoto et al., 2016, Pichersky and Gang,
75 2000, Qi et al., 2004). Therefore, in our study, we chose wild soybean (*Glycine soja*), a wild
76 relative of soybean (*Glycine max)*, to delineate genetic basis and evolution of glyceollin
77 accumulation resulting from biotic stress, i.e., soybean cyst nematode (SCN), the most devastating
78 soybean pest worldwide (Tylka and Marett, 2021). Wild soybeans thrive in diverse habitats and
79 harbor much higher, underexplored genetic diversity than cultivated soybean (Zhang et al., 2019).
80 Hence, it is an ideal system to understand the genetic basis and evolution of glyceollin variation.
81 Eventually, the essential genes identified in wild soybean can be used for metabolic engineering
82 or in a breeding program to develop nutrition-rich biofortified soybean cultivars as they exhibit
83 similar genome size and content with small reproductive isolation (Singh and Hymowitz, 1999).

84

85 A metabolic gene cluster is a group of (two or more) genomically co-localized and potentially
86 coregulated non-homologous genes that encode enzymes involved in a particular metabolic
87 pathway (Nützmann et al., 2016, Töpfer et al., 2017). They have been a common phenomenon
88 since the early days of microbial genetics (Koonin, 2009, Rocha, 2008, Zheng et al., 2002).
89 However, gene clusters in plant metabolic pathways have been discovered only recently, even
90 though microbes and plants are both extremely rich sources of metabolic diversity. A study by
91 Chae et al. (2014) on metabolic gene clusters in *Arabidopsis,* soybean, sorghum, and rice suggested
92 that approximately one-third of all the metabolic genes in *Arabidopsis,* soybean, and sorghum, and
93 one-fifth in rice were rich in gene clusters across primary and specialized metabolic pathways
94 (Chae et al., 2014). There is compelling evidence indicating that the highly plastic plant genome
95 itself generates metabolic gene clusters via gene duplication, neofunctionalization, divergence, and
96 genome reorganization instead of horizontal gene transfer from microbes (Osbourn and Field,
97 2009). This suggests that plants rewire their genome to gain new adaptive functions driven by the
98 need to survive in distinct environments. Systematic mining and functional validation of the
99 candidate genes in such clusters will facilitate the discovery of new enzymes and chemistries that
100 render pathway prediction. Moreover, metabolic gene clusters are likely to be located within
101 dynamic chromosomal regions, and thus, many identified so far may be due to recent evolution
102 (Field et al., 2011, Matsuba et al., 2013, Qi et al., 2004). If so, investigation of these clusters can
103 provide insights into their evolutionary history. The vast and diverse array of specialized
104 metabolites that are produced through multi-step metabolic pathways play an important role in
105 plant adaptation to various ecological niches. However, the occurrence, prevalence, and evolution
106 of such gene clusters in plants are largely unknown. Thus, the study of plant metabolic gene
107 clusters has implications for molecular biology and evolutionary genomics (Chavali and Rhee,
108 2018, Nützmann et al., 2016, Takos and Rook, 2012, Yeaman and Whitlock, 2011).

109 Due to the extraordinary metabolic diversity, to date, less than 50 plant-specialized metabolic
110 pathways have been biochemically and genetically identified (Nützmann et al., 2016).
111 Metabolomic GWAS (mGWAS) offers an effective approach to understand the genetic basis of
112 metabolites and their associated traits (Chan et al., 2010, Chan et al., 2011, Luo, 2015,
113 Riedelsheimer et al., 2012). mGWAS allows the identification of common polymorphic regions
114 controlling complex metabolic traits by substantially increasing association panel and genome-

115 wide molecular markers. Besides elucidating genetic architecture, mGWAS can also be used to
116 infer gene functions (Luo, 2015). Hence, mGWAS provides a comprehensive approach to
117 discovering candidate genes. Thus far, it has been used to uncover the genetic basis of variations
118 of a number of different metabolites. For example, Chen et al. (2014) carried out a rice mGWAS
119 study that identified 36 candidate genes influencing the variation of metabolites with physiological
120 and nutritional importance (Chen et al., 2014).

121

122 The isoflavonoid pathway has been relatively well studied (Sukumaran et al., 2018). However, it
123 is still not clear how glyceollin induction is regulated. This study is the first to employ genomic
124 and evolutionary approaches to understand the genetic basis and selection of glyceollin induction.
125 Our study provides a fundamental basis for the long-term goal of developing glyceollin-fortified
126 soybean cultivars that would improve plant and human health to meet current and future global
127 challenges. In this study, we aim to address these three questions: (1) What is the genetic basis of
128 variation in glyceollin induction by SCN? (2) Are there any gene clusters and transcription factors
129 involved in glyceollin variation? (3) Are epistatic interactions and natural selection important
130 evolutionary factors influencing the variation of glyceollin induction?

131

132 **Materials and Methods**

133 **Plant materials**

134 A total of 264 accessions of wild soybean, *Glycine soja,* from a wide geographic range, originally
135 collected from China, Japan, Russia, and South Korea, were utilized (Table S1). The seeds of these
136 ecotypes were obtained from the USDA national germplasm resources laboratory
137 (https://www.ars-grin.gov/).

138

139 **Plant preparation, SCN inoculation, and sample collection**

140 Seed preparation, germination, transplanting, and soybean cyst nematode (SCN, *Heterodera*
141 *glycines Ichinohe*, HG type 1.2.5.7) inoculation were performed following a previously developed
142 protocol (Zhang et al., 2017a, Zhang et al., 2017b, Zhang and Song, 2017). Whole root tissues
143 were collected and weighed five days post-infection (dpi). The 5 dpi time point was chosen because

144 our previous study suggested a significant inhibition in SCN development in a resistant genotype
145 compared to normal growth in a susceptible genotype (Zhang et al., 2017a, Zhang et al., 2017b).
146 All samples were flash frozen in liquid nitrogen and stored at -80 °C. Four biological replicates
147 per wild soybean genotype were used, eventually a total of 1,020 samples.

148

149 **Metabolite extraction and quantification**

150 Metabolites were extracted from each root sample and processed as previously described (Strauch
151 et al., 2015). Briefly, root tissues were extracted with 50% (v/v) methanol in a 60 °C water bath
152 for 30 min. A tissue-to-solvent ratio (w/v) of 1:10 was used constantly across all samples. The
153 extracts were filtered using a 0.2-µm filter prior to LC-MS (liquid chromatography-mass
154 spectrometry) profiling on a G6530A Q-TOF LC/MS (Agilent Technologies, USA). Peaks that
155 were consistently detected in at least three biological replicates within each genotype were used
156 for downstream analyses. The metabolites were confirmed using pure standard compounds,
157 including daidzein, daidzein-d6, and glyceollin. Due to the low concentrations of these compounds
158 and the small sample masses of the wild soybean root samples that had been collected, we used a
159 signal-to-noise ratio of $\geqslant 10$ for the measurement of the peaks for glyceollin and daidzein. Our
160 method successfully measured daidzein (µg/g root) and glyceollin (unitless) in 264 accessions of
161 wild soybean *G. soja* roots quantitatively and semi-quantitatively, respectively. Following method
162 development, optimization, and analyses of the test samples, calibration curves were designed
163 using at least six different concentrations of daidzein, created in triplicate to quantify known
164 concentrations of daidzein and glyceollin. A second-degree polynomial was derived from the
165 known concentrations of the standard curve samples and the mass spectrometer response
166 (daidzein/internal standard) from the standard curve data. The resulting polynomial was used to
167 calculate the concentrations of daidzein in the experimental samples. Low, medium, and high QC
168 (quality control) samples were created to assess the accuracy of the calculations. We used the ratio
169 of glyceollin (unitless) to daidzein (µg/g root) (GVSD) as our phenotypic trait. This phenotype
170 henceforth is denoted GVSD.

171

172 **Genotypic data**

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

176 **Metabolite-based genome-wide association study (mGWAS) and linkage disequilibrium**
177 **estimation**

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

184

185 A conventional Manhattan plot was generated using the qqman R package to visualize the SNPs
186 (Turner, 2014). In addition to the genome-wide significant threshold, we also calculated the
187 chromosome-wide Bonferroni thresholds using independent SNPs estimated on each chromosome
188 following the method of Li and Ji (2005) (Li and Ji, 2005). Linkage disequilibrium (LD) was
189 calculated across the panel with the TASSEL program, version 5 [6], for the significant SNPs
190 identified from the GWAS analysis. LD was measured using squared correlation R-squared (r^2) of
191 0.2 (upper right in the LD plot) and p-value < 0.05 (the lower left in the LD plot). A pairwise LD
192 was generated following the R function described by Shin et al. (2006) (Shin et al., 2006). Genes
193 within LD blocks containing significant SNPs were identified as potential sources of candidates
194 for further analyses.

195

196 **Identification of candidate genes**

197 For extensive gene mining of our identified gene pool, we used an array of bioinformatics tools.
198 Such an approach can improve the accuracy of candidate gene and gene cluster predictions and
199 resolve inconsistencies among the bioinformatics tools (Chavali and Rhee, 2018). Specifically, a
200 pairwise linkage disequilibrium (LD) analysis was initially used for potential candidate gene
201 identification. Then, genes in each LD block were examined as potential candidate genes, and their

202 annotations were obtained from the Phytozome v13 database (Goodstein et al., 2011). Afterward,
203 a GO enrichment analysis of the identified candidate genes was performed using ShinyGO v0.66:
204 Gene Ontology Enrichment Analysis (*p*-value cutoff (FDR, false discovery rate) = 0.05) (Ge et al.,
205 2020), Soybase GO Enrichment Data (Grant et al., 2010). To investigate the involvement of these
206 potential candidate genes in metabolic pathways, a database search was performed through an
207 annotation file from Phytozome v13 (Goodstein et al., 2011), Soybase (Grant et al., 2010), SoyCyc
208 10.0 Soybean Metabolic Pathway (Hawkins et al., 2021), and Pathview databases (Luo et al., 2017).
209 Finally, a PMN plant metabolic cluster viewer was applied to categorize enzymes into classes
210 (signature or tailoring) and metabolic domains (Hawkins et al., 2021).

211

212 **Analysis of epistatic interactions**

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

224

225 **Extended haplotype homozygosity analyses**

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

230 **Results**

231 **Genomic dissection of glyceollin accumulation upon biotic induction**

232 We identified a total of eight significant SNPs, with four located on chromosome 9 and the others
233 on chromosomes 3, 13, 15, and 20 (Fig. **1a**, Table 1). These SNPs were identified based on both
234 genome-wide Bonferroni threshold of 5.104 and chromosome-wide Bonferroni thresholds that
235 varied narrowly from 3.79 to 3.82 among the 20 chromosomes (3.803 on chromosome 9) (Figs
236 **1a,b**, Table S2). The manhattan and Q-Q (quantile-quantile) plots are shown in Fig.**1a,b,c**. The
237 four significant SNPs on chromosome 9 are located close to each other within a 535 kb region
238 (Table S2). The broad-sense heritability (h^2) was estimated 35% (Table S2).

239 **Linkage disequilibrium analysis and candidate gene identification**

240 We identified a total of 666 possible candidate genes within the linkage disequilibrium (LD) blocks
241 of the eight significant SNPs (soybean reference genome *Glycine max* Wm82.a2.v1) (Goodstein
242 et al., 2011, Zhou et al., 2015). The LD block on chromosome 9 showed the strongest LD with a
243 long range compared to the others (Figs **2b**, **S1, S2**). We considered r^2>0.2 as a cutoff for our LD
244 analysis, where r^2 is the extent of allelic association between a pair of sites (Weir, 1990). Candidate
245 gene *Glyma.09G128200* shows the highest level of LD near the significant SNPs on chromosome
246 9 compared to the LD block for the rest of the significant SNPs on this chromosome (Figs **2b**, **S1**).
247 The functional annotation of the candidate genes within this block is biosynthetic enzymes
248 involved in isoflavonoid pathway, as well as regulatory genes such as *WRKY* and *MYB*
249 transcription factors (Tables 1, S3, and S4), which may indicate their transcriptional level
250 involvement in glyceollin induction in response to SCN stress (Colinas and Goossens, 2018).

251 We also found putative genes encoding enzymes involved in the specialized metabolic pathways
252 within the LD blocks of the significant SNPs on chromosomes 3, 13, 15, and 20. The enriched GO
253 category includes flavonoid biosynthesis pathway, phenylpropanoid metabolic process, linamarin
254 biosynthesis, and terpenoid biosynthesis (Table S5). Apart from the biosynthetic enzymes on these
255 chromosomes, we also found transcription factor genes, such as *WRKY*, *MYB*, and *NAC* (Table
256 S5).

257

258 **Metabolic gene clusters identification**

259 We were particularly interested in the candidate genes in the branch from daidzein to glyceollin in
260 the isoflavonoid biosynthesis pathway (Lozovaya et al., 2007). We found that the identified

261 candidate genes on chromosome 9 are clustered together, and they fell into two clusters. Both of
262 these two clusters belong to tailoring enzyme glycosyltransferase within phenylpropanoid
263 specialized metabolic domain. And six genes are within the branch of isoflavonoid biosynthesis
264 pathway. Two of these six genes, *Glyma.09G127200* and *Glyma.09G127300,* are called cluster 1,
265 while the rest four (*Glyma.09G127700*, *Glyma.09G128200*, *Glyma.09G128300*, and
266 *Glyma.09G128400*) are called cluster 2 (Table S3).
267
268 Further investigation of annotation of these candidate genes within the gene clusters (Table S4),
269 we found *Glyma.09G127200* gene encodes a glucosyltransferase that may act on 4'-methoxy
270 isoflavones biochanin A, formononetin, 4'-hydroxy isoflavones genistein, and daidzein substrates.
271 However, the enzyme does not act on isoflavanones, flavones, flavanones, flavanols, or coumarins
272 (Köster and Barz, 1981). Within the same cluster, *Glyma.09G127300* has similar annotations and
273 functions as *Glyma.09G127200*. Interestingly, the four genes within cluster 2 have a similar
274 functional annotation as *Glyma.09G127200 and Glyma.09G127300* in cluster 1, and all these four
275 genes encode isoenzymes (Table S4). Such a link between these two gene clusters indicates their
276 proximity in the metabolic pathway.
277
278 **Epistatic interactions among all significant SNPs**
279 The results of the epistasis tests for each of the 28 pairwise combinations of the eight significant
280 SNPs are shown in Table 2. Three probabilities, all associated with the SNP on chromosome 20,
281 were not estimable (Table 2). Among the remaining 25 SNP pairs, 20 show statistical significance.
282 Particularly noticeable is the high significance for all interactions of the SNPs on chromosomes 3,
283 13, and 15. Three of the six pairs among the four SNPs on chromosome 9, all involving
284 ss715603462, also are statistically significant. In general, therefore, this is evidence for substantial
285 epistasis among these SNPs affecting GVSD.
286
287 These epistatic interactions of the SNP pairs are illustrated in Fig. **3** for each of the four chosen
288 combinations. For example, in panel **a** (Fig. **3a**), it can be seen that regression slopes of GVSD on
289 ss715603454 are close to 0 for ss71585948 CC genotype but are positive for TC and especially
290 TT genotypes. In panel d (Fig. **3d**), regression slopes of GVSD on ss715603471 are negative for
291 ss715603462 AA and GA genotypes but positive for GG genotypes. With no epistasis, these slopes

292 would be expected to be roughly parallel, but in fact, they diverge considerably from parallelism
293 in these four examples.
294
295 **Significant SNPs exhibited extended haplotype homozygosity**
296 The extended homozygosity analysis (EHH) analyses revealed allele-specific EHH values of the
297 significant SNPs (ss715603454, ss715603455, ss715603462, and ss715603471) on chromosomes
298 9 (Fig. 4). For example, T allele of ss715603454 showed much higher EHH value than G allele.
299 Alleles of significant SNPs on the other chromosomes showed compatible EHH values (Fig. 4).
300
301 **Discussion**
302 **Metabolic gene clusters in glyceollin induction**
303 Gene clusters have been reported to play important roles in phytochemical diversity in *Arabidopsis*,
304 sorghum, soybean, and rice (Chae et al., 2014). However, their roles in regulating metabolic
305 variation in wild species are relatively less investigated. Even though the isoflavonoid biosynthesis
306 pathway is relatively well studied, the genetic regulation of glyceollin induction is unclear.
307 Particularly, the contribution, prevalence, and occurrence of gene clusters in plant metabolic
308 diversity are largely unclear. Our mGWAS results suggest there are two gene clusters with
309 functionally related but non-homologous genes, which may involve in glyceollin induction in wild
310 soybean. Thus far, these are the first reported plausible gene clusters involved in glyceollin
311 accumulation induced by biotic stimuli. These gene clusters suggest that glyceollin may be
312 synthesized where the enzyme-encoding genes are adjacent to each other on the same chromosome
313 (Chavali and Rhee, 2018). Physical clustering of genes with similar functions can facilitate co-
314 inheritance of alleles with favorable combinations and their coordinated regulations at chromatin
315 level (Chu et al., 2011, Osbourn, 2010a). Besides, such clusters incline to locate in the sub-
316 telomeric regions (Gierl and Frey, 2001, Qi et al., 2004, Sakamoto et al., 2004), near the ends of
317 chromosomes that are known to harbor mutations. For example, an examination of the complete
318 genome sequence revealed that the maize *DIMBOA* cluster is located close to the end of
319 chromosome 4 (Farman, 2007, Jonczyk et al., 2008). Thus, identifying the positions of the genes
320 can contribute to inferences of possible mechanisms underlying chemical diversity in natural
321 populations.
322

323 Tailoring enzymes, such as methyltransferases, glycosyltransferases, *CYPs*,
324 dehydrogenases/reductases, and acyltransferases are responsible for modifying the chemical
325 backbone of specialized metabolites (Osbourn, 2010b). The gene clusters we found are associated
326 with tailoring or regulating glycosyltransferase enzymes. A common defense mechanism of plants
327 involves glycosylation of secondary metabolites by involving these enzymes (Mylona et al., 2008).
328 Therefore, the clustering of the genes encoding glycosyltransferase on chromosome 9 indicates the
329 formation of stress-induced (i.e., SCN stress in our study) protective compounds. For example, the
330 cyclic hydroxamic acid *(DIBOA)* in maize (Frey et al., 1997, Gierl and Frey, 2001), the triterpene
331 avenacin in oat (Field and Osbourn, 2008, Mugford et al., 2009, Qi et al., 2004, Qi et al., 2006),
332 and two gene clusters associated with diterpene (momilactone and phytocassane) synthesis in rice,
333 which may be pre-formed or synthesized after stress induction for plant defense. Disruption of
334 such gene clusters may compromise pest and disease resistance and lead to the accumulation of
335 toxic pathway intermediates (Chu et al., 2011). In the multi-step plant specialized metabolic
336 pathways, rapid adaptation to a particular environmental niche could result in highly diverse and
337 rapidly evolving metabolic gene clusters (Osbourn and Field, 2009). Hence, the level of
338 conservation of the identified gene clusters in this study across different *Glycine soja* genotypes
339 can shed light on evolutionary insight of these clusters (Field and Osbourn, 2008). Synthetic
340 biology and functional genetics can further help investigate the organization and contribution of
341 these clusters in metabolite diversity, as well as decipher the mechanism of adaptive evolution and
342 genome plasticity (Chu et al., 2011, Osbourn, 2010b).
343
344 **Plausible transcriptional factors in glyceollin induction**
345 Advancement of genetics, genomics, and bioinformatic approaches facilitate the prediction and
346 identification of a large number of genes, including transcription factors associated with plant-
347 specialized metabolic pathways (Anarat-Cappillino and Sattely, 2014, Moore et al., 2019).
348 However, the transcriptional regulators of specialized metabolism are less well characterized
349 (Shoji and Yuan, 2021). The regulation of highly diverse plant specialized metabolic pathways is
350 dynamic given the ever-changing environment. Such regulation generally occurs at transcription
351 level, and thus, it requires coordinated regulation often mediated by transcription factors (TFs)
352 (Colinas and Goossens, 2018, Shoji, 2019). For instance, *MYB* and basic helix-loop-helix *(bHLH)*
353 TF family genes were reported to regulate anthocyanin and related flavonoid biosynthetic

354 pathways in a wide range of species (Chezem and Clay, 2016). Moreover, significant
355 modifications of these regulatory genes give rise to the vast diversity in plant specialized
356 metabolism (Huang et al., 2018, Springer et al., 2019).
357
358 It is possible that transcription factors, such as *MYB* and *WRKY* TFs on chromosome 9, may
359 influence glyceollin induction. This indicates regulation of glyceollin induction with SCN stress
360 may involve a highly complex interplay among multiple genes and pathways. Previous studies
361 reported that gene families of transcription factors, such as *NAC*, *MYB*, *bHLH*, and *WRKY*,
362 exhibited conservative patterns among *Arabidopsis*, cotton, grapevine, maize, and rice (Ibraheem
363 et al., 2015, Ogawa et al., 2017, Saga et al., 2012, Xu et al., 2004, Yamamura et al., 2015, Zheng
364 et al., 2006). These plant species produce various phytoalexins, such as indole alkaloids, terpenoid
365 aldehydes, stilbenoids, deoxyanthocyanidins, and momilactones/ phytocassanes,
366 respectively. This gives rise to the question of whether these TFs are as diversified as the metabolic
367 pathways, or they maintain conservative patterns among species. The investigation of TFs binding
368 promoter regions can give insights if the pathways are co-opted into stress-inducible regulation by
369 the respective TFs (Jahan et al., 2019). The homology of TFs among different plant species can
370 help metabolic engineering a wide variety of crop plants to produce phytoalexins in greater
371 amounts.
372
373 In addition to enzyme-encoding genes, TF genes can also be found as gene clusters. For example,
374 the gene cluster of TF *ERF* (jasmonate (JA)- responsive ethylene response factor) consists of five
375 *ERF* genes in tomato (Cárdenas et al., 2016, Thagun et al., 2016), while eight in potato (Cárdenas
376 et al., 2016), two clusters of ten and five in tobacco (Kajikawa et al., 2017), five in *C. roseus* (Singh
377 et al., 2020), four in *Calotropis gigantea* (Singh et al., 2020), and four in *Glesemium sempervirens*
378 (Singh et al., 2020). Besides, TFs involved in plant specialized metabolism can be found in arrays
379 (Shoji and Yuan, 2021, Zhou et al., 2016). So, it is possible that the TFs we identified are located
380 in the same genomic neighborhood as arrays or biosynthetic gene clusters (BGCs). The co-
381 regulation hypothesis of gene clusters poses that clustering of TFs can help coregulate genes in a
382 pathway. Although co-regulation also exists between un-clustered metabolic pathways, clustering
383 may accelerate the recruitment of genes into a regulon (Smit and Lichman, 2022, Wisecaver et al.,
384 2017).

385

386 **Epistasis and plausible selection on glyceollin induction**

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

399

400 Genes containing causal variation for plant defensive compounds may influence field fitness and
401 thus are likely under natural selection (Kroymann, 2011). For example, Benderoth et al. (2006)
402 detected positive selection in glucosinolate diversification in *Arabidopsis thaliana* and its relatives
403 (Benderoth et al., 2006). Prasad et al. (2012) showed positive selection for a mutation on a
404 metabolic pathway gene could enhance resistance to herbivory in natural populations of a rocky
405 mountain cress species (Prasad et al., 2012). We detected strong signals of selection on the SNPs
406 significantly associated with glyceollin phenotypes with EHH and LD analyses (Figs **4, 2b,** and
407 **S1**). For example, the LD surrounding the significant SNP ss715603454 that is next to the
408 identified gene clusters is more extensive, suggesting strong selection in this region (Figs **2b, S1**).
409 Meanwhile, the two alleles of this significant SNP, G and T, showed different EHH values, with
410 T exhibiting much longer haplotype homozygosity. This indicates that this T allele may be under
411 recent positive selection. Interestingly, the T allele is significantly associated with higher induction
412 of glyceollin and has a higher frequency in South Korea (Fig. **2c,d**). The allele-specific EHH
413 pattern and their geographic distribution may be due to heterogeneous selection pressure in nature.

414

415 **Perspectives and future directions of our study**

416 Plant specialized metabolites exhibit extreme quantitative and qualitative variation. Therefore,
417 high-throughput metabolite profiling, such as LC-MS analysis coupled with GWAS (as applied
418 here) can help better understand the genetic contributions to metabolic diversity in natural
419 populations. A common assumption is that biological variables or traits should show a normal
420 distribution, and skewed data may indicate measurement error. However, the scenario is different
421 in metabolomics, especially in secondary metabolism. For instance, a ratio of two related
422 compounds, rather than their separate values, may provide a comprehensive understanding of the
423 underlying enzymatic process (Byrne et al., 1996, Chan et al., 2011, Kliebenstein, 2001,
424 Kliebenstein et al., 2001, Kliebenstein, 2007, McMullen et al., 1998, Petersen et al., 2012, Prasad
425 et al., 2012, Yencho et al., 1998). We used a ratio of glyceollin and daidzein concentrations as the
426 phenotypic trait for our association study. The use of a metabolic ratio also may produce: (1) a
427 reduction in the variability of the data collected for the biological replicates and thus increase
428 statistical power and (2) a reduction in overall noise in the dataset by canceling out systemic
429 experimental errors. Most importantly for our purposes, the glyceollin to daidzein metabolite ratio
430 is correlated to the corresponding reaction rate under optimal steady-state assumptions, as this
431 metabolite pair is connected in the phenylpropanoid biosynthetic pathway (Petersen et al., 2012,
432 Suhre et al., 2011).
433
434 The natural world has a lot to offer in tackling diseases and global food scarcity. There is a need
435 to develop new medicines and future value-increased food by unlocking the uncharted gene pools
436 of wild plants. Our chosen study system crop wild relative of soybean poses much higher and
437 underexplored genetic diversity than its domesticated descendants. Given that glyceollin is
438 produced in trace amounts, it is an exciting challenge to define the plant metabolic gene clusters
439 and transcriptional regulators in the glyceollin biosynthesis pathway. Besides complex cancer
440 treatment and therapies, the rise of different types of tumors and tumor subtypes urges the need
441 for new drugs. Along with glyceollin's role in plant defense, it has been well-documented for anti-
442 cancer activities. Our follow-up studies will apply transcriptomics and functional validation of the
443 candidate genes, which can expand our focus to explore associations of genes in clusters to
444 understand their involvement in regulating glyceollin biosynthesis at the systems level. As
445 phytochemical variation can be caused by both structural genes and their expression differences,
446 it will be interesting to explore the role of pathway-specific regulators (i.e., transcription factors)

447 in glyceollin induction (Osbourn, 2010b). Our results suggest that improving our fundamental
448 knowledge of plant specialized metabolic gene clusters and regulators will facilitate metabolic
449 engineering with improved metabolic traits for sustainable agriculture and novel pharmaceuticals.
450

451 **Acknowledgements:**

452 Research in B-H.S. lab was supported by the National Institute of General Medical Sciences,
453 Award Number: R15GM122029, R15AT011603; North Carolina Biotechnology Center, Award
454 Number: 2019-BIG-6507, 2020-FLG-3806; The North Carolina Soybean Producers Association,
455 Award Number: 18-0252; and the University of North Carolina at Charlotte, Award Number:
456 111239. Work by F.Y. was (partially) supported by the Schlumberger Foundation Faculty for the
457 Future program.

458 We thank Thomas Mitchell-Olds for the constructive comments and suggestions and Janice
459 Kofsky for the helpful discussion. We thank Song lab members Melissa Hatley, Judy Janiol, Janice
460 Kofsky and Neha Mittal for their help with planting and sample collection.

461

462 **Author contributions**

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

468

469 **References**

470 ANARAT-CAPPILLINO, G. & SATTELY, E. S. 2014. The chemical logic of plant natural
471 product biosynthesis. *Current opinion in plant biology,* 19**,** 51-58.
472 ANDERSON, C. J. The role of standing genetic variation in adaptation of digital organisms to a
473 new environment. ALIFE 2012: The Thirteenth International Conference on the
474 Synthesis and Simulation of Living Systems, 2012. MIT Press, 3-10.
475 ARITA, M. 2004. The metabolic world of *Escherichia coli* is not small. *Proceedings of the*
476 *National Academy of Sciences,* 101**,** 1543-1547.
477 BAMJI, S. F. & CORBITT, C. 2017. Glyceollins: Soybean phytoalexins that exhibit a wide
478 range of health-promoting effects. *Journal of Functional Foods,* 34**,** 98-105.
479 BARGHI, N., HERMISSON, J. & SCHLÖTTERER, C. 2020. Polygenic adaptation: a unifying
480 framework to understand positive selection. *Nature Reviews Genetics,* 21**,** 769-781.
481 BARRETT, R. D. & SCHLUTER, D. 2008. Adaptation from standing genetic variation. *Trends*
482 *in ecology & evolution,* 23**,** 38-44.
483 BENDEROTH, M., TEXTOR, S., WINDSOR, A. J., MITCHELL-OLDS, T., GERSHENZON,
484 J. & KROYMANN, J. 2006. Positive selection driving diversification in plant secondary
485 metabolism. *Proceedings of the National Academy of Sciences,* 103**,** 9118-9123.
486 BONT, Z., ZÜST, T., ARCE, C. C., HUBER, M. & ERB, M. 2020. Heritable variation in root
487 secondary metabolites is associated with recent climate. *Journal of ecology,* 108**,** 2611-
488 2624.
489 BRACHI, B., MEYER, C. G., VILLOUTREIX, R., PLATT, A., MORTON, T. C., ROUX, F. &
490 BERGELSON, J. 2015. Coselected genes determine adaptive variation in herbivore
491 resistance throughout the native range of *Arabidopsis thaliana*. *Proceedings of the*
492 *National Academy of Sciences,* 112**,** 4032-4037.
493 BYRNE, P., MCMULLEN, M., SNOOK, M., MUSKET, T., THEURI, J., WIDSTROM, N.,
494 WISEMAN, B. & COE, E. 1996. Quantitative trait loci and metabolic pathways: genetic
495 control of the concentration of maysin, a corn earworm resistance factor, in maize silks.
496 *Proceedings of the National Academy of Sciences,* 93**,** 8820-8825.
497 CÁRDENAS, P. D., SONAWANE, P. D., POLLIER, J., VANDEN BOSSCHE, R.,
498 DEWANGAN, V., WEITHORN, E., TAL, L., MEIR, S., ROGACHEV, I. &

499 MALITSKY, S. 2016. *GAME9* regulates the biosynthesis of steroidal alkaloids and
500 upstream isoprenoids in the plant mevalonate pathway. *Nature communications,* 7**,** 1-16.
501 CHAE, L., KIM, T., NILO-POYANCO, R. & RHEE, S. Y. 2014. Genomic signatures of
502 specialized metabolism in plants. *science,* 344**,** 510-513.
503 CHAN, E. K., ROWE, H. C., HANSEN, B. G. & KLIEBENSTEIN, D. J. 2010. The complex
504 genetic architecture of the metabolome. *PLoS genetics,* 6**,** e1001198.
505 CHAN, E. K., ROWE, H. C., CORWIN, J. A., JOSEPH, B. & KLIEBENSTEIN, D. J. 2011.
506 Combining genome-wide association mapping and transcriptional networks to identify
507 novel genes controlling glucosinolates in *Arabidopsis thaliana*. *PLoS biology,* 9**,**
508 e1001125.
509 CHAVALI, A. K. & RHEE, S. Y. 2018. Bioinformatics tools for the identification of gene
510 clusters that biosynthesize specialized metabolites. *Briefings in bioinformatics,* 19**,** 1022-
511 1034.
512 CHEN, W., GAO, Y., XIE, W., GONG, L., LU, K., WANG, W., LI, Y., LIU, X., ZHANG, H.,
513 DONG, H., ZHANG, W., ZHANG, L., YU, S., WANG, G., LIAN, X. & LUO, J. 2014.
514 Genome-wide association analyses provide genetic and biochemical insights into natural
515 variation in rice metabolism. *Nature Genetics,* 46**,** 714-721.
516 CHEZEM, W. R. & CLAY, N. K. 2016. Regulation of plant secondary metabolism and
517 associated specialized cell development by *MYBs* and *bHLHs*. *Phytochemistry,* 131**,** 26-
518 43.
519 CHU, H. Y., WEGEL, E. & OSBOURN, A. 2011. From hormones to secondary metabolism: the
520 emergence of metabolic gene clusters in plants. *The Plant Journal,* 66**,** 66-79.
521 COLINAS, M. & GOOSSENS, A. 2018. Combinatorial transcriptional control of plant
522 specialized metabolism. *Trends in plant science,* 23**,** 324-336.
523 DONNEZ, D., KIM, K.-H., ANTOINE, S., CONREUX, A., DE LUCA, V., JEANDET, P.,
524 CLÉMENT, C. & COUROT, E. 2011. Bioproduction of resveratrol and viniferins by an
525 elicited grapevine cell culture in a 2 L stirred bioreactor. *Process Biochemistry,* 46**,** 1056-
526 1062.
527 FARMAN, M. L. 2007. Telomeres in the rice blast fungus *Magnaporthe oryzae*: the world of the
528 end as we know it. *FEMS microbiology letters,* 273**,** 125-132.

529 FELL, D. A. & WAGNER, A. 2000. The small world of metabolism. *Nature biotechnology,* 18**,**
530 1121-1122.
531 FIELD, B. & OSBOURN, A. E. 2008. Metabolic diversification—independent assembly of
532 operon-like gene clusters in different plants. *Science,* 320**,** 543-547.
533 FIELD, B., FISTON-LAVIER, A.-S., KEMEN, A., GEISLER, K., QUESNEVILLE, H. &
534 OSBOURN, A. E. 2011. Formation of plant metabolic gene clusters within dynamic
535 chromosomal regions. *Proceedings of the National Academy of Sciences,* 108**,** 16116-
536 16121.
537 FREY, M., CHOMET, P., GLAWISCHNIG, E., STETTNER, C., GRUN, S., WINKLMAIR, A.,
538 EISENREICH, W., BACHER, A., MEELEY, R. B. & BRIGGS, S. P. 1997. Analysis of
539 a chemical plant defense mechanism in grasses. *Science,* 277**,** 696-699.
540 GE, S. X., JUNG, D. & YAO, R. 2020. ShinyGO: a graphical gene-set enrichment tool for
541 animals and plants. *Bioinformatics,* 36**,** 2628-2629.
542 GIERL, A. & FREY, M. 2001. Evolution of benzoxazinone biosynthesis and indole production
543 in maize. *Planta,* 213**,** 493-498.
544 GOODSTEIN, D. M., SHU, S., HOWSON, R., NEUPANE, R., HAYES, R. D., FAZO, J.,
545 MITROS, T., DIRKS, W., HELLSTEN, U., PUTNAM, N. & ROKHSAR, D. S. 2011.
546 Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research,*
547 40**,** D1178-D1186.
548 GRAHAM, T. L. & GRAHAM, M. Y. 1996. Signaling in soybean phenylpropanoid responses
549 (dissection of primary, secondary, and conditioning effects of light, wounding, and
550 elicitor treatments). *Plant Physiology,* 110**,** 1123-1133.
551 GRANT, D., NELSON, R. T., CANNON, S. B. & SHOEMAKER, R. C. 2010. SoyBase, the
552 USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res,* 38**,** D843-6.
553 HAWKINS, C., GINZBURG, D., ZHAO, K., DWYER, W., XUE, B., XU, A., RICE, S., COLE,
554 B., PALEY, S. & KARP, P. 2021. Plant Metabolic Network 15: A resource of genome-
555 wide metabolism databases for 126 plants and algae. *Journal of integrative plant biology,*
556 63**,** 1888-1905.
557 HUANG, D., WANG, X., TANG, Z., YUAN, Y., XU, Y., HE, J., JIANG, X., PENG, S.-A., LI,
558 L. & BUTELLI, E. 2018. Subfunctionalization of the *Ruby2–Ruby1* gene cluster during
559 the domestication of citrus. *Nature plants,* 4**,** 930-941.

560 IBRAHEEM, F., GAFFOOR, I., TAN, Q., SHYU, C.-R. & CHOPRA, S. 2015. A sorghum *MYB*
561 transcription factor induces 3-deoxyanthocyanidins and enhances resistance against leaf
562 blights in maize. *Molecules,* 20**,** 2388-2404.
563 JAHAN, M. A., HARRIS, B., LOWERY, M., COBURN, K., INFANTE, A. M., PERCIFIELD,
564 R. J., AMMER, A. G. & KOVINICH, N. 2019. The *NAC* family transcription factor
565 *GmNAC42–1* regulates biosynthesis of the anticancer and neuroprotective glyceollins in
566 soybean. *BMC Genomics,* 20**,** 149.
567 JEANDET, P., DOUILLET-BREUIL, A.-C., BESSIS, R., DEBORD, S., SBAGHI, M. &
568 ADRIAN, M. 2002. Phytoalexins from the Vitaceae: biosynthesis, phytoalexin gene
569 expression in transgenic plants, antifungal activity, and metabolism. *Journal of*
570 *Agricultural and food chemistry,* 50**,** 2731-2741.
571 JEANDET, P., SOBARZO-SÁNCHEZ, E., SILVA, A. S., CLÉMENT, C., NABAVI, S. F.,
572 BATTINO, M., RASEKHIAN, M., BELWAL, T., HABTEMARIAM, S. & KOFFAS,
573 M. 2020. Whole-cell biocatalytic, enzymatic and green chemistry methods for the
574 production of resveratrol and its derivatives. *Biotechnology advances,* 39**,** 107461.
575 JEONG, H., TOMBOR, B., ALBERT, R., OLTVAI, Z. N. & BARABÁSI, A.-L. 2000. The
576 large-scale organization of metabolic networks. *Nature,* 407**,** 651-654.
577 JONCZYK, R., SCHMIDT, H., OSTERRIEDER, A., FIESSELMANN, A., SCHULLEHNER,
578 K., HASLBECK, M., SICKER, D., HOFMANN, D., YALPANI, N. & SIMMONS, C.
579 2008. Elucidation of the final reactions of DIMBOA-glucoside biosynthesis in maize:
580 characterization of *Bx6* and *Bx7*. *Plant physiology,* 146**,** 1053-1063.
581 KAJIKAWA, M., SIERRO, N., KAWAGUCHI, H., BAKAHER, N., IVANOV, N. V.,
582 HASHIMOTO, T. & SHOJI, T. 2017. Genomic insights into the evolution of the nicotine
583 biosynthesis pathway in tobacco. *Plant Physiology,* 174**,** 999-1011.
584 KERWIN, R., FEUSIER, J., CORWIN, J., RUBIN, M., LIN, C., MUOK, A., LARSON, B., LI,
585 B., JOSEPH, B. & FRANCISCO, M. 2015. Natural genetic variation in *Arabidopsis*
586 *thaliana* defense metabolism genes modulates field fitness. *Elife,* 4.
587 KERWIN, R. E., FEUSIER, J., MUOK, A., LIN, C., LARSON, B., COPELAND, D., CORWIN,
588 J. A., RUBIN, M. J., FRANCISCO, M. & LI, B. 2017. Epistasis× environment
589 interactions among *Arabidopsis thaliana* glucosinolate genes impact complex traits and
590 fitness in the field. *new phytologist,* 215**,** 1249-1263.

591 KIM, H. J., LIM, J.-S., KIM, W.-K. & KIM, J.-S. 2012. Soyabean glyceollins: biological effects
592 and relevance to human health. *Proceedings of the Nutrition Society,* 71**,** 166-174.
593 KLIEBENSTEIN, D. 2001. Gene duplication and the diversification of secondary metabolism:
594 side chain modification of glucosinolates in *Arabidopsis thaliana*. *Plant cell,* 13**,** 681-
595 693.
596 KLIEBENSTEIN, D. J., GERSHENZON, J. & MITCHELL-OLDS, T. 2001. Comparative
597 quantitative trait loci mapping of aliphatic, indolic and benzylic glucosinolate production
598 in *Arabidopsis thaliana* leaves and seeds. *Genetics,* 159**,** 359-370.
599 KLIEBENSTEIN, D. J. 2007. Metabolomics and Plant Quantitative Trait Locus Analysis–The
600 optimum genetical genomics platform? *Concepts in plant metabolomics.* Springer.
601 KOONIN, E. V. 2009. Evolution of genome architecture. *The international journal of*
602 *biochemistry & cell biology,* 41**,** 298-306.
603 KÖSTER, J. & BARZ, W. 1981. UDP-glucose: isoflavone 7-O-glucosyltransferase from roots of
604 chick pea *(Cicer arietinum L.)*. *Archives of Biochemistry and Biophysics,* 212**,** 98-104.
605 KROYMANN, J. 2011. Natural diversity and adaptation in plant secondary metabolism. *Current*
606 *opinion in plant biology,* 14**,** 246-251.
607 LANDER, E. & KRUGLYAK, L. 1995. Genetic dissection of complex traits: guidelines for
608 interpreting and reporting linkage results. *Nature genetics,* 11**,** 241-247.
609 LEAMY, L. J., ZHANG, H., LI, C., CHEN, C. Y. & SONG, B.-H. 2017. A genome-wide
610 association study of seed composition traits in wild soybean *(Glycine soja)*. *BMC*
611 *genomics,* 18**,** 1-15.
612 LI, J. & JI, L. 2005. Adjusting multiple testing in multilocus analyses using the eigenvalues of a
613 correlation matrix. *Heredity,* 95**,** 221-227.
614 LOZOVAYA, V. V., LYGIN, A. V., ZERNOVA, O. V., ULANOV, A. V., LI, S., HARTMAN,
615 G. L. & WIDHOLM, J. M. 2007. Modification of phenolic metabolism in soybean hairy
616 roots through down regulation of chalcone synthase or isoflavone synthase. *Planta,* 225**,**
617 665-679.
618 LUO, J. 2015. Metabolite-based genome-wide association studies in plants. *Current opinion in*
619 *plant biology,* 24**,** 31-38.

620 LUO, W., PANT, G., BHAVNASI, Y. K., BLANCHARD JR, S. G. & BROUWER, C. 2017.
621 Pathview Web: user friendly pathway visualization and data integration. *Nucleic acids*
622 *research,* 45**,** W501-W508.
623 MATSUBA, Y., NGUYEN, T. T., WIEGERT, K., FALARA, V., GONZALES-VIGIL, E.,
624 LEONG, B., SCHÄFER, P., KUDRNA, D., WING, R. A. & BOLGER, A. M. 2013.
625 Evolution of a complex locus for terpene biosynthesis in *Solanum*. *The Plant Cell,* 25**,**
626 2022-2036.
627 MCMULLEN, M., BYRNE, P., SNOOK, M., WISEMAN, B., LEE, E., WIDSTROM, N. &
628 COE, E. 1998. Quantitative trait loci and metabolic pathways. *Proceedings of the*
629 *National Academy of Sciences,* 95**,** 1996-2000.
630 MESSER, P. W. & PETROV, D. A. 2013. Population genomics of rapid adaptation by soft
631 selective sweeps. *Trends in ecology & evolution,* 28**,** 659-669.
632 MIYAMOTO, K., FUJITA, M., SHENTON, M. R., AKASHI, S., SUGAWARA, C., SAKAI,
633 A., HORIE, K., HASEGAWA, M., KAWAIDE, H. & MITSUHASHI, W. 2016.
634 Evolutionary trajectory of phytoalexin biosynthetic gene clusters in rice. *The plant*
635 *journal,* 87**,** 293-304.
636 MOORE, B. M., WANG, P., FAN, P., LEONG, B., SCHENCK, C. A., LLOYD, J. P., LEHTI-
637 SHIU, M. D., LAST, R. L., PICHERSKY, E. & SHIU, S.-H. 2019. Robust predictions of
638 specialized metabolism genes through machine learning. *Proceedings of the National*
639 *Academy of Sciences,* 116**,** 2344-2353.
640 MUGFORD, S. T., QI, X., BAKHT, S., HILL, L., WEGEL, E., HUGHES, R. K.,
641 PAPADOPOULOU, K., MELTON, R., PHILO, M. & SAINSBURY, F. 2009. A serine
642 carboxypeptidase-like acyltransferase is required for synthesis of antimicrobial
643 compounds and disease resistance in oats. *The Plant Cell,* 21**,** 2473-2484.
644 MYLONA, P., OWATWORAKIT, A., PAPADOPOULOU, K., JENNER, H., QIN, B.,
645 FINDLAY, K., HILL, L., QI, X., BAKHT, S. & MELTON, R. 2008. *Sad3* and *Sad4* are
646 required for saponin biosynthesis and root development in oat. *The Plant Cell,* 20**,** 201-
647 212.
648 NÜTZMANN, H. W., HUANG, A. & OSBOURN, A. 2016. Plant metabolic clusters–from
649 genetics to genomics. *New phytologist,* 211**,** 771-789.

650 NWACHUKWU, I. D., LUCIANO, F. B. & UDENIGWE, C. C. 2013. The inducible soybean
651 glyceollin phytoalexins with multifunctional health-promoting properties. *Food research*
652 *international,* 54**,** 1208-1216.
653 OGAWA, S., MIYAMOTO, K., NEMOTO, K., SAWASAKI, T., YAMANE, H., NOJIRI, H. &
654 OKADA, K. 2017. *OsMYC2*, an essential factor for JA-inductive sakuranetin production
655 in rice, interacts with MYC2-like proteins that enhance its transactivation ability.
656 *Scientific reports,* 7**,** 1-11.
657 OSBOURN, A. 2010a. Gene clusters for secondary metabolic pathways: an emerging theme in
658 plant biology. *Plant physiology,* 154**,** 531-535.
659 OSBOURN, A. 2010b. Secondary metabolic gene clusters: evolutionary toolkits for chemical
660 innovation. *Trends in Genetics,* 26**,** 449-457.
661 OSBOURN, A. E. & FIELD, B. 2009. Operons. *Cellular and Molecular Life Sciences,* 66**,** 3755-
662 3775.
663 PETERSEN, A.-K., KRUMSIEK, J., WÄGELE, B., THEIS, F. J., WICHMANN, H.-E.,
664 GIEGER, C. & SUHRE, K. 2012. On the hypothesis-free testing of metabolite ratios in
665 genome-wide and metabolome-wide association studies. *BMC bioinformatics,* 13**,** 1-7.
666 PICHERSKY, E. & GANG, D. R. 2000. Genetics and biochemistry of secondary metabolites in
667 plants: an evolutionary perspective. *Trends in plant science,* 5**,** 439-445.
668 PRASAD, K. V., SONG, B.-H., OLSON-MANNING, C., ANDERSON, J. T., LEE, C.-R.,
669 SCHRANZ, M. E., WINDSOR, A. J., CLAUSS, M. J., MANZANEDA, A. J. & NAQVI,
670 I. 2012. A gain-of-function polymorphism controlling complex traits and fitness in
671 nature. *science,* 337**,** 1081-1084.
672 QI, X., BAKHT, S., LEGGETT, M., MAXWELL, C., MELTON, R. & OSBOURN, A. 2004. A
673 gene cluster for secondary metabolism in oat: implications for the evolution of metabolic
674 diversity in plants. *Proceedings of the National Academy of Sciences,* 101**,** 8233-8238.
675 QI, X., BAKHT, S., QIN, B., LEGGETT, M., HEMMINGS, A., MELLON, F., EAGLES, J.,
676 WERCK-REICHHART, D., SCHALLER, H. & LESOT, A. 2006. A different function
677 for a member of an ancient and highly conserved cytochrome P450 family: from essential
678 sterols to plant defense. *Proceedings of the National Academy of Sciences,* 103**,** 18848-
679 18853.

680 RIEDELSHEIMER, C., LISEC, J., CZEDIK-EYSENBERG, A., SULPICE, R., FLIS, A.,
681 GRIEDER, C., ALTMANN, T., STITT, M., WILLMITZER, L. & MELCHINGER, A. E.
682 2012. Genome-wide association mapping of leaf metabolic profiles for dissecting
683 complex traits in maize. *Proceedings of the National Academy of Sciences,* 109**,** 8872-
684 8877.
685 ROCHA, E. P. 2008. The organization of the bacterial genome. *Annual review of genetics,* 42**,**
686 211-233.
687 ROWE, H. C., HANSEN, B. G., HALKIER, B. A. & KLIEBENSTEIN, D. J. 2008. Biochemical
688 networks and epistasis shape the *Arabidopsis thaliana* metabolome. *The Plant Cell,* 20**,**
689 1199-1216.
690 SABETI, P. C., REICH, D. E., HIGGINS, J. M., LEVINE, H. Z., RICHTER, D. J.,
691 SCHAFFNER, S. F., GABRIEL, S. B., PLATKO, J. V., PATTERSON, N. J. &
692 MCDONALD, G. J. 2002. Detecting recent positive selection in the human genome from
693 haplotype structure. *Nature,* 419**,** 832-837.
694 SAGA, H., OGAWA, T., KAI, K., SUZUKI, H., OGATA, Y., SAKURAI, N., SHIBATA, D. &
695 OHTA, D. 2012. Identification and characterization of *ANAC042*, a transcription factor
696 family gene involved in the regulation of camalexin biosynthesis in *Arabidopsis*.
697 *Molecular plant-microbe interactions,* 25**,** 684-696.
698 SAKAMOTO, T., MIURA, K., ITOH, H., TATSUMI, T., UEGUCHI-TANAKA, M.,
699 ISHIYAMA, K., KOBAYASHI, M., AGRAWAL, G. K., TAKEDA, S. & ABE, K. 2004.
700 An overview of gibberellin metabolism enzyme genes and their related mutants in rice.
701 *Plant physiology,* 134**,** 1642-1653.
702 SEO, J. Y., KIM, B. R., OH, J. & KIM, J.-S. 2018. Soybean-derived phytoalexins improve
703 cognitive function through activation of *Nrf2/HO-1* signaling pathway. *International*
704 *journal of molecular sciences,* 19**,** 268.
705 SHIN, J.-H., BLAY, S., MCNENEY, B. & GRAHAM, J. 2006. LDheatmap: an R function for
706 graphical display of pairwise linkage disequilibria between single nucleotide
707 polymorphisms. *Journal of statistical software,* 16**,** 1-10.
708 SHOJI, T. 2019. The recruitment model of metabolic evolution: jasmonate-responsive
709 transcription factors and a conceptual model for the evolution of metabolic pathways.
710 *Frontiers in Plant Science,* 10**,** 560.

711 SHOJI, T. & YUAN, L. 2021. *ERF* gene clusters: working together to regulate metabolism.
712 *Trends in Plant Science,* 26**,** 23-32.
713 SINGH, R. J. & HYMOWITZ, T. 1999. Soybean genetic resources and crop improvement.
714 *Genome,* 42**,** 605-616.
715 SINGH, S. K., PATRA, B., PAUL, P., LIU, Y., PATTANAIK, S. & YUAN, L. 2020. Revisiting
716 the *ORCA* gene cluster that regulates terpenoid indole alkaloid biosynthesis in
717 *Catharanthus roseus*. *Plant Science,* 293**,** 110408.
718 SMIT, S. J. & LICHMAN, B. R. 2022. Plant biosynthetic gene clusters in the context of
719 metabolic evolution. *Natural Product Reports*.
720 SONG, Q., HYTEN, D. L., JIA, G., QUIGLEY, C. V., FICKUS, E. W., NELSON, R. L. &
721 CREGAN, P. B. 2013. Development and evaluation of SoySNP50K, a high-density
722 genotyping array for soybean. *PloS one,* 8**,** e54985.
723 SPRINGER, N., DE LEÓN, N. & GROTEWOLD, E. 2019. Challenges of translating gene
724 regulatory information into agronomic improvements. *Trends in plant science,* 24**,** 1075-
725 1082.
726 STRAUCH, R. C., SVEDIN, E., DILKES, B., CHAPPLE, C. & LI, X. 2015. Discovery of a
727 novel amino acid racemase through exploration of natural variation in *Arabidopsis*
728 *thaliana*. *Proceedings of the National Academy of Sciences of the United States of*
729 *America,* 112**,** 11726-11731.
730 SUBRAMANIAN, S., STACEY, G. & YU, O. 2006. Endogenous isoflavones are essential for
731 the establishment of symbiosis between soybean and *Bradyrhizobium japonicum*. *The*
732 *Plant Journal,* 48**,** 261-273.
733 SUHRE, K., SHIN, S.-Y., PETERSEN, A.-K., MOHNEY, R. P., MEREDITH, D., WÄGELE,
734 B., ALTMAIER, E., DELOUKAS, P., ERDMANN, J. & GRUNDBERG, E. 2011.
735 Human metabolic individuality in biomedical and pharmaceutical research. *Nature,* 477**,**
736 54-60.
737 SUKUMARAN, A., MCDOWELL, T., CHEN, L., RENAUD, J. & DHAUBHADEL, S. 2018.
738 Isoflavonoid-specific prenyltransferase gene family in soybean: *GmPT01*, a pterocarpan
739 2-dimethylallyltransferase involved in glyceollin biosynthesis. *The Plant Journal,* 96**,**
740 966-981.

741 SZPIECH, Z. A. & HERNANDEZ, R. D. 2014. selscan: an efficient multithreaded program to
742 perform EHH-based scans for positive selection. *Molecular biology and evolution,* 31**,**
743 2824-2827.
744 TAKOS, A. M. & ROOK, F. 2012. Why biosynthetic genes for chemical defense compounds
745 cluster. *Trends in plant science,* 17**,** 383-388.
746 TANG, Y., LIU, X., WANG, J., LI, M., WANG, Q., TIAN, F., SU, Z., PAN, Y., LIU, D. &
747 LIPKA, A. E. 2016. GAPIT version 2: an enhanced integrated tool for genomic
748 association and prediction. *The plant genome,* 9**,** plantgenome2015.11.0120.
749 THAGUN, C., IMANISHI, S., KUDO, T., NAKABAYASHI, R., OHYAMA, K., MORI, T.,
750 KAWAMOTO, K., NAKAMURA, Y., KATAYAMA, M. & NONAKA, S. 2016.
751 Jasmonate-responsive *ERF* transcription factors regulate steroidal glycoalkaloid
752 biosynthesis in tomato. *Plant and Cell Physiology,* 57**,** 961-975.
753 TÖPFER, N., FUCHS, L. M. & AHARONI, A. 2017. The PhytoClust tool for metabolic gene
754 clusters discovery in plant genomes. *Nucleic Acids Res,* 45**,** 7049-7063.
755 TURNER, S. D. 2014. qqman: an R package for visualizing GWAS results using QQ and
756 manhattan plots. *Biorxiv***,** 005165.
757 TYLKA, G. L. & MARETT, C. C. 2021. Known distribution of the soybean cyst nematode,
758 *Heterodera glycines*, in the United States and Canada in 2020. *Plant Health Progress,* 22**,**
759 72-74.
760 WANG, X., HOWELL, C. P., CHEN, F., YIN, J. & JIANG, Y. 2009. Gossypol-a polyphenolic
761 compound from cotton plant. *Advances in food and nutrition research,* 58**,** 215-263.
762 WEIR, B. S. 1990. *Genetic data analysis. Methods for discrete population genetic data*, Sinauer
763 Associates, Inc. Publishers.
764 WISECAVER, J. H., BOROWSKY, A. T., TZIN, V., JANDER, G., KLIEBENSTEIN, D. J. &
765 ROKAS, A. 2017. A global coexpression network approach for connecting genes to
766 specialized metabolic pathways in plants. *The Plant Cell,* 29**,** 944-959.
767 XU, Y.-H., WANG, J.-W., WANG, S., WANG, J.-Y. & CHEN, X.-Y. 2004. Characterization of
768 *GaWRKY1*, a cotton transcription factor that regulates the sesquiterpene synthase gene
769 *(+)-δ-cadinene synthase-A*. *Plant physiology,* 135**,** 507-515.
770 YAMAMURA, C., MIZUTANI, E., OKADA, K., NAKAGAWA, H., FUKUSHIMA, S.,
771 TANAKA, A., MAEDA, S., KAMAKURA, T., YAMANE, H. & TAKATSUJI, H. 2015.

772 Diterpenoid phytoalexin factor, a *bHLH* transcription factor, plays a central role in the
773 biosynthesis of diterpenoid phytoalexins in rice. *The Plant Journal,* 84**,** 1100-1113.
774 YEAMAN, S. & WHITLOCK, M. C. 2011. The genetic architecture of adaptation under
775 migration–selection balance. *Evolution: International Journal of Organic Evolution,* 65**,**
776 1897-1911.
777 YENCHO, G., KOWALSKI, S., KOBAYASHI, R., SINDEN, S., BONIERBALE, M. &
778 DEAHL, K. 1998. QTL mapping of foliar glycoalkaloid aglycones in *Solanum*
779 *tuberosum× S. berthaultii* potato progenies: quantitative variation and plant secondary
780 metabolism. *Theoretical and applied genetics,* 97**,** 563-574.
781 ZHANG, H., KJEMTRUP-LOVELACE, S., LI, C., LUO, Y., CHEN, L. P. & SONG, B.-H.
782 2017a. Comparative RNA-seq analysis uncovers a complex regulatory network for
783 soybean cyst nematode resistance in wild soybean *(Glycine soja)*. *Scientific reports,* 7**,** 1-
784 14.
785 ZHANG, H. & SONG, B. H. 2017. RNA-seq data comparisons of wild soybean genotypes in
786 response to soybean cyst nematode (*Heterodera glycines*). *Genom Data,* 14**,** 36-39.
787 ZHANG, H., YASMIN, F. & SONG, B.-H. 2019. Neglected treasures in the wild — legume
788 wild relatives in food security and human health. *Current Opinion in Plant Biology,* 49**,**
789 17-26.
790 ZHANG, H. Y., KJEMTRUP-LOVELACE, S., LI, C. B., LUO, Y., CHEN, L. P. & SONG, B.
791 H. 2017b. Comparative RNA-seq analysis uncovers a complex regulatory network for
792 soybean cyst nematode resistance in wild soybean (*Glycine soja*). *Scientific Reports,* 7**,**
793 9699.
794 ZHENG, Y., SZUSTAKOWSKI, J. D., FORTNOW, L., ROBERTS, R. J. & KASIF, S. 2002.
795 Computational identification of operons in microbial genomes. *Genome research,* 12**,**
796 1221-1230.
797 ZHENG, Z., QAMAR, S. A., CHEN, Z. & MENGISTE, T. 2006. *Arabidopsis WRKY33*
798 transcription factor is required for resistance to necrotrophic fungal pathogens. *The Plant*
799 *Journal,* 48**,** 592-605.
800 ZHOU, Y., MA, Y., ZENG, J., DUAN, L., XUE, X., WANG, H., LIN, T., LIU, Z., ZENG, K. &
801 ZHONG, Y. 2016. Convergence and divergence of bitterness biosynthesis and regulation
802 in Cucurbitaceae. *Nature Plants,* 2**,** 1-8.

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

806

807

808 **Figure legends**

809 **Fig. 1** GWAS of Glyceollin induction with SCN stress: A genome-wide **(a)** and chromosome-
810 wide **(b)** Manhattan plots, with thresholds of 5.104 and 3.803, respectively; **(c)** quantile-quantile
811 (QQ) plot. Significant SNPs are found on chromosomes 3, 9, 13, 15 and 20 at a 5% genome-wide
812 threshold, the probability of 7.86×10^{-6} resulted in a threshold of 5.01 (solid red line in the genome-
813 wide Manhattan plot) **(a)**. The 5% chromosome-wide LOD threshold resulted in significant p-
814 values of 1.57×10^{-4} (threshold 3.803, solid blue line) **(b)**.

815

816 **Fig. 2** An LD decay measured as R square for pairwise markers and plotted against their distance
817 **(a)** and LD plot for chromosome 9 for significant SNPs. The black diagonal denotes LD between
818 each site and itself **(b)**. Geographic range of the alleles of significant SNPs close to the gene
819 clusters on chromosome 9 **(c)**. Allele frequency in each population. Allele frequency in different
820 geographic regions for a significant SNP was generated using JMP®, Version 15. SAS Institute
821 Inc., Cary, NC, 1989–2021. **(d)**.

822

823 **Fig. 3** Epistatic interactions of the SNP pairs for each of four chosen combinations. Regression
824 slopes of GVSD on ss715603454 are close to 0 for ss715603454 CC genotypes but are positive
825 for TC and especially TT genotypes **(a)**. Regression slopes of GVSD on ss715603462 are close to
826 0 for ss715585948 CC genotypes but are negative for TC and especially TT genotypes **(b)**.
827 Regression slopes of GVSD on ss715615975 are close to 0 for ss715585948 TT genotypes but are
828 negative for TC and especially CC genotypes **(c)**. Regression slopes of GVSD on ss715603471
829 are negative in sign for ss715603462 AA and GA genotypes, but positive in sign for GG genotypes
830 **(d)**.

831

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

834

835 **Tables**

836 **Table 1.** Identification of significant SNPs and functional annotation of the plausible candidate
837 genes.
838

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 finger superfamily protein
ss715603462	Gm09	*WRKY* family transcription factor family protein *MYB* domain
ss715603471	Gm09	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

839

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

841

	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*

842

843 Shown are the probabilities for each pairwise interaction of SNPs. * = $P < 0.05$ from
844 sequential Bonferroni tests. n.e. = not estimable. Ch3 = ss715585948, Ch9a = ss715603454, Ch9b
845 = ss715603455, Ch9c = ss715603462, Ch9d = ss715603471, Ch13 = ss715615975, Ch15 =
846 ss715620269, Ch20 = ss715636844

847

848

849

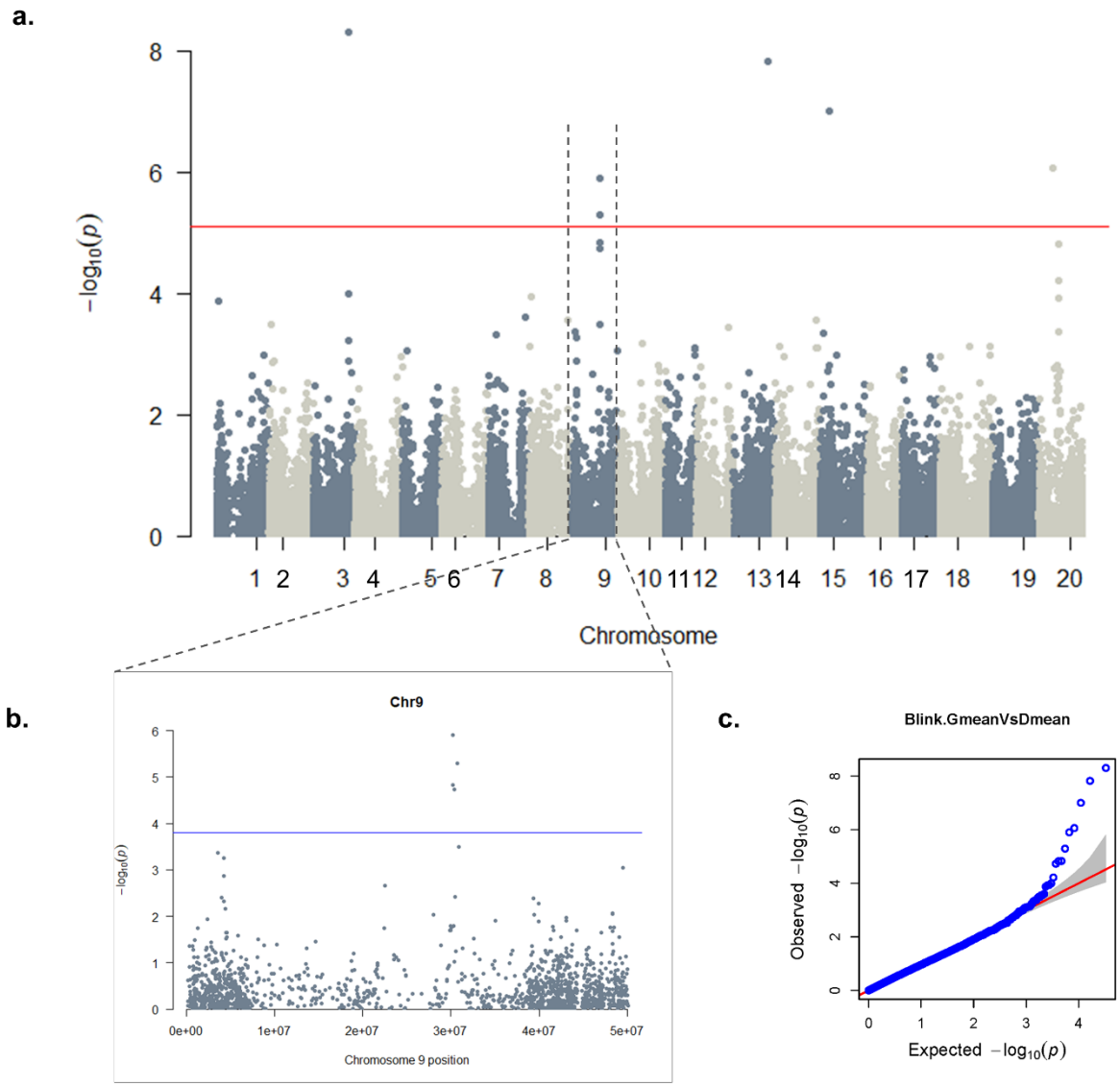
850

851

852 **Figures**

853 **Fig. 1**

854



855

856

857

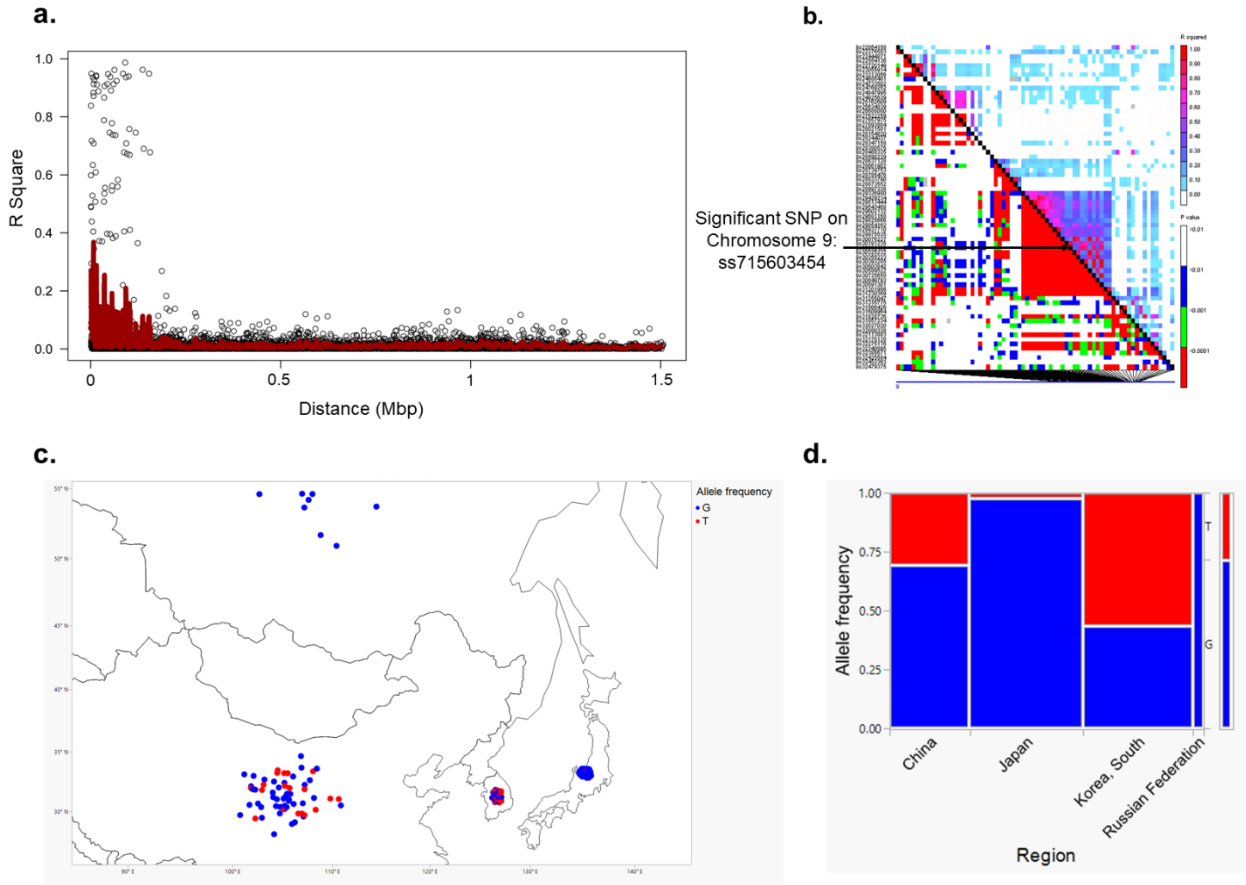
858

859

860

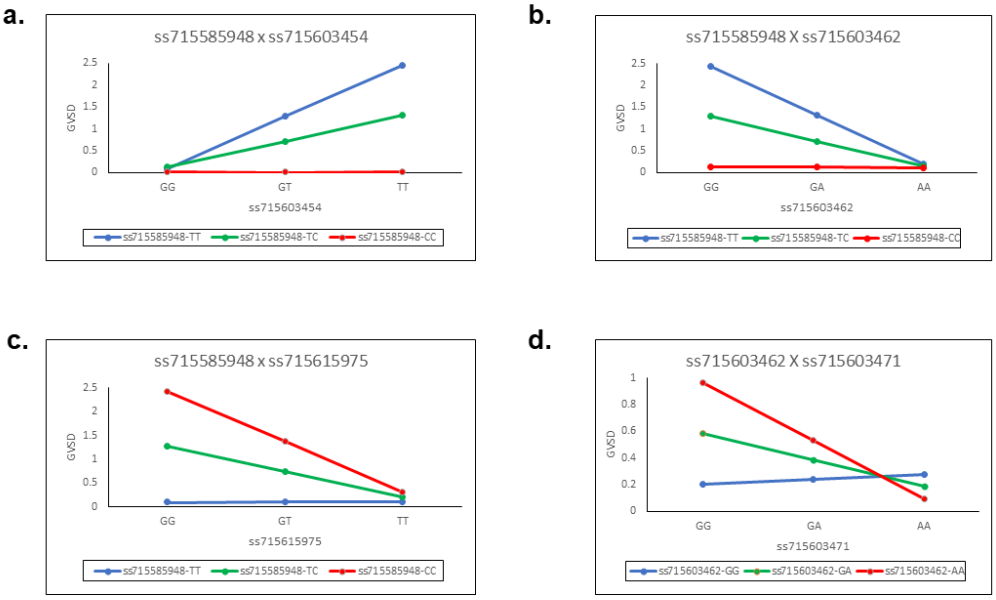
861

862 **Fig. 2**



863
864
865
866
867
868
869
870
871
872
873
874

875 **Fig. 3**



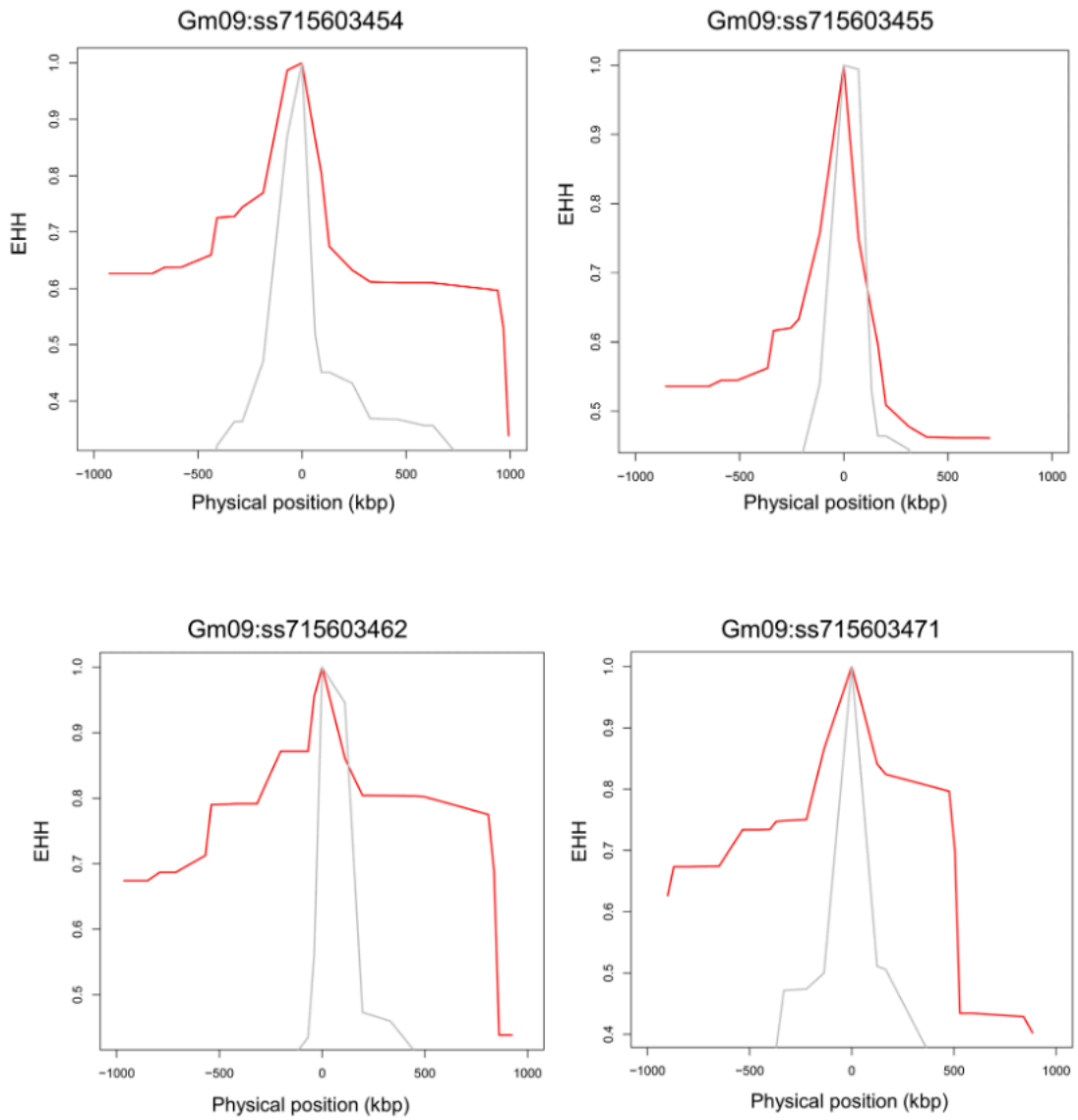
876

877

878

879

880 **Fig. 4**



881

882

883