Resequencing 250 soybean accessions: new insights into genes associated with agronomic traits and genetic networks

Chunming Yang[1#], Jun Yan[2#], Shuqin Jiang[2], Xia Li[3], Haowei Min[4*], Xiangfeng Wang[2*], Dongyun Hao[1*]

[1]*Key Laboratory for Agricultural Biotechnology of Jilin Provincial, Institute of Agricultural Biotechnology, Jilin Academy of Agricultural Sciences (JAAS), Jilin 130033, China*

[2]*Department of Crop Genomics and Bioinformatics, College of Agronomy and Biotechnology, China Agricultural University, Beijing 100094, China*

[3]*Key Laboratory of Molecular Cytogenetics and Genetic Breeding of Heilongjiang Province, College of Life Science and Technology, Harbin Normal University, Harbin 150025, Heilongjiang, China*

[4]*BioTrust Technology. Inc., Beijing 100094, China*

\# Equal contribution

* Corresponding authors

E-mail:

dyhao@cjaas.com (Hao D);

xwang@cau.edu.cn (Wang X);

biotrust_st@163.com (Min H);

Running title: Chunming Yang *et al.* / Soybean resequence

30

31

Abstract

33 Limited knowledge on genomic diversity and the functional genes associated with

34 soybean variety traits has resulted in slow breeding progress. We sequenced the

35 genome of 250 soybean landraces and cultivars from China, America and Europe, and

36 investigated their population structure, genetic diversity and architecture and selective

37 sweep regions of accessions. We identified five novel agronomically important genes

38 and studied the effects of functional mutations in respective genes. We found

39 candidate genes *GSTT1*, *GL3* and *GSTL3* associated with isoflavone content, *CKX3*

40 associated with yield traits, and *CYP85A2* associated with both architecture and yield

41 traits. Our phenotype-gene network analysis revealed that hub nodes play a role in

42 complex phenotypic associations. In this work, we describe novel agronomic trait

43 associated genes and a complex genetic network, providing a valuable resource for

44 future soybean molecular breeding.

45

Introduction

47 Soybean *Glycine max* [L.] Merr. is one of the most important crops worldwide of

48 vegetable oil and proteins source for human and livestock feed etc. Soybean

49 originated in China and its wild species (*G. soja* Sieb. & Zucc.) were domesticated in

50 approximately 3,000 B.C. before introduced to Korea and Japan about 3,000 years

51 later. It was brought to Europe and North America in the 18th century, and extensively

52 cultivated on a global scale since the 19th century[1].

53   With the rapid development of modern molecular biology and the

54 high-throughput sequencing technologies, whole-genome resequencing and genome

55 wide association studies (GWAS) have become common methods used to study

56 population genetic diversity and locating phenotypic related quantitative trait loci

57 (QTL) or genes. This has improved our knowledge extensively in crop genomes and

58 selective breeding. In recent years, for example, there are increasing number of

59 reports on the domestication and improvement of soybean at the genome-wide level.

60 This includes genes and genetic networks related to soybean agronomic traits and

61 functions[2-4]. However, due to the diversity of soybean varieties and their complex

62 genetic background, our knowledge of the soybean genome and functional genes is

63 still limited in comparison to rice and maize. A greater number of soybean varieties

64 need further exploration at the genomic level, particularly in relation to molecular

65 traits associated with edible quality, soybean "ideotype" and the underlying genetic

66 network of high-yielding varieties.

67   In this study, we collected 250 soybean varieties from the core Northeast China

68 soybean germplasm pool, which consisted of 134 accessions of landrace and cultivar

69 from Northeast and Northwest China, as well as 116 accessions from European and

70 North American cultivars. The genomes of the most accessions are not sequenced

71 previously. We performed the high-depth whole-genome resequencing and

72 comprehensive analyses of this 250 soybean population. The generated dataset

73 revealed valuable new information on soybean genome structure, novel genes

74 associated with important agronomic traits and the genetic networks. These genetic

75 resources provide unique references into molecular breeding and evolution study in

76 soybean.

77

Results

Genome resequencing and variation calling

80 High-depth whole genome resequencing was performed on 250 soybean accessions,

81 including 51 landraces and 83 cultivars originating from provinces in Northeast China

82 (i.e. Heilongjiang, Jilin, Liaoning and Northeast Inner Mongolia) and Northwest

83 China (i.e. Xinjiang, Ningxia and Gansu), as well as 116 cultivars originating from

84 Europe and North America **(Figure 1a, Supplementary Table 3).** In total, we

85 obtained approximately 10G of pair-end reads and 3T bases. The maximum

86 sequencing depth of a single accession was 22.5x, with the average depth at 11x. After

87   filtering out the raw sequencing data (see methods), the remaining high-quality

88   cleaned data were compared with the soybean reference genome *G. max v2.0[5]*. The

89   effective mapping rates ranged from 74.8% to 87.6%, while the genome coverage

90   ranged from 94.8% to 97.0% (**Supplementary Table 1**). The high mapping rates and

91   coverage guarantee that the sequenced data is reliable and of high quality.

92   Through standard variation detection, genotype filtering and imputation steps

93   (see methods), we detected in total 6,333,721 single nucleotide polymorphisms (SNPs)

94   and 2,565,797 insertion & deletions (Indels). This includes 244,360 SNPs and 62,714

95   Indels located in the exon regions. The ratio of non-synonymous SNP to synonymous

96   SNP substitution was 1.37. There are 4,311,814 SNPs with a minor allele frequency

97   (MAF) larger than 0.05 (**Supplementary Table 2 & Supplementary Figure 1**). In

98   summary, we achieved over 6M high-density and high-quality genotype data from

99   250 soybean accessions with a density of one SNP per 15 bases.

100

Population structure of soybean landraces and cultivars

102   Using the 6M SNP genotype dataset, we constructed a phylogenetic tree using the

103   neighbour joining (NJ) method. This resulted in the classification of the 250 soybean

104   accessions into four groups (**Figure 1b**). Among them, *Group 1* included 65 Chinese

105   varieties, four European and six American cultivars, whereas *Group 2* contained 56

106   Chinese varieties, one European and 13 American varieties. In *Group 3*, there were 21

107   European, two Chinese and six American cultivars, while 65 North American

108   cultivars and 11 Chinese varieties were clustered within *Group 4* (**Figure 1c**).

109   Principal Component Analysis (PCA) results were consistent with the phylogenetic

110   tree results. Three groups, *Group 1*, *3*, and *4* radiated away from *Group 2* within the

111   rectangular coordinate system projected using eigenvector 1 and eigenvector 2 data on

112   X and Y axes, respectively. Concurrently, the distribution of varieties in the four

113   groups had continuity, indicating varieties located in different groups also have

114   genetic similarities (**Figure 1e**). A Bayesian clustering algorithm based on a mixed

115   model was used to estimate the proportion of ancestors in each accession. That is,

116   when *K = 2*, the main ancestor component (yellow) of *Group 4* was split, indicating

117   that *Group 4* has the highest level of selection. When $K = 3$, the main ancestor

118   component (blue) of *Group 3* was split, indicating that *Group 3* has the second level

119   of selection. However, when $K = 4$ and $K = 5$, *Group 1 and Group 2* exhibit complex

120   differentiated mixed ancestor components, indicating a higher genetic diversity and

121   lower selection level in *Groups 1* and *2* (**Supplementary Table 3, Figure 1d**).

122   These results indicate that group classification of the 250 soybean accessions is

123   closely related to their geographical distribution. That is, varieties with similar

124   geographical distribution have similar genetic backgrounds. Generally speaking, the

125   group classification was also related to the level of domestication. Landraces have a

126   lower level of domestication, while cultivars have higher levels of domestication.

127   Varieties with similar domestication levels tend to have a higher similarity in genetic

128   backgrounds. However, there are still differences in geographical distribution and

129   domestication level among breeds with similar genetic backgrounds, indicating that

130   gene exchange may have occurred between accessions of different groups. This

131   observation reflected the complexity of soybean domestication history.

132

133   **Genetic diversity and selective sweep analysis**

134   Linkage disequilibrium (LD) analysis showed that the overall LD decay distance was

135   more than 100 kb, and the LD decay distance of the landraces was smaller than that of

136   the cultivars (**Figure 2A**). Further LD decay analysis of the four groups showed that

137   the LD decay distance of *Group 1* was the smallest, followed by *Groups 2* and *3*,

138   while *Group 4* had the largest LD decay distance (**Figure 2B**). In addition, the LD

139   levels varied for different chromosomes or different regions across one chromosome.

140   Identical by state (IBS) analysis can reflect the degree of relatedness among

141   individuals by calculating the consistency of all genetic markers. The IBS values of

142   all comparisons in each group were calculated, and it was found that the average IBS

143   values of landraces were less than that of cultivars (**Figure 2C**). The IBS values of

144   *Groups 1-4* followed the same trends as that of LD decay distance. In particular, the

145   IBS values of *Group 1* were the lowest and the IBS values of *Group 4* were the

146   highest of all groups (**Figure 2D**). $\theta\pi$ values can reflect the genetic diversity within a

147   population by calculating the number of different sites between any two sequences or

148   individuals within a population. *Fst* is a calculation used to measure the

149   differentiation and genetic distance between two populations. $\theta\pi$ values were

150   calculated for landraces, cultivars, all accessions, and *Groups 1-4*. *Fst* values were

151   calculated between landraces and cultivars, and between each comparison of the four

152   Groups. Results show that a population with a higher level of LD decay distance or

153   higher IBS values correlate with a smaller $\theta\pi$ (**Figure 2E, F**). This pattern is opposite

154   to that of the LD decay and IBS values. The lowest *Fst* value was for *Group 1* versus

155   *Group 2*, while the highest value was for *Group 3* versus *Group 4*. We also observed

156   that the *Fst* value of *Group 2* versus *Group 3* was higher than that of *Group 1* versus

157   *Group 3* (**Figure 2G**). The *Fst* value of *Group 2* versus *Group 4* was smaller than that

158   of *Group 1* versus *Group 4*. In addition, the results of allele frequency distribution

159   (AFD) analysis, as an alternative population similarity measurement, were consistent

160   with the *Fst* results (**Supplementary Figure 2**). The population diversity analysis

161   results, when combined with population structure and geographical distribution

162   information, infer that the European and American soybean varieties may have

163   originated from different Chinese ancestors before undergoing independent selection.

164   Results indicate that European cultivars and the Chinese landrace group (*Group 1*) in

165   our study have a more recent common ancestor, while North American cultivars and

166   the Chinese cultivar group (*Group 2*) have a more recent common ancestor.

167   *Tajima' D* (based on a neutral test), $\theta\pi$ (based on genetic diversity within a

168   population), and *Fst* (based on genetic diversity between two populations), have

169   provided us with highly effective tools that screen selective sweep signals across a

170   genome[6]. We combined methods in pairs for mining potential selective sweep

171   regions in the soybean genome that may have underwent artificial selection. One pair

172   was *Tajima' D* combined with $\theta\pi$ for the whole population. Another pair was *Fst*

173   combined with $\theta\pi$ ratios between two subpopulations, landrace and cultivar. We used

174   a sliding window method to calculate the values of *Tajima' D, $\theta\pi$, and Fst* in each

175   window across the whole genome, and selected the top 5% most significant windows

176   as potential selective sweep regions (**Supplementary Figure 3A, B**). A total of 148

177   and 222 potential selective sweep regions were screened by the two methods, and they

178   covered 36.09 Mb and 88.15 Mb genome regions, respectively (**Supplementary**

179   **Table 4**). These potential selective sweep regions covered 9,128 genes, accounting for

180   approximately one sixth of all soybean genes. A total of 1,876 genes were screened by

181   both methods (**Supplementary Figure 3C**). A runs of homozygosity (ROH) region is

182   a continuous homozygous chromosome region in a genome, which may relate to

183   domestication or artificial selection[7]. Through ROH analysis, 71 ROH regions

184   larger than 300 kb were obtained from all 250 accessions, with a total length of 27.84

185   Mb. The longest ROH regions up to 911 kb were located at the beginning of

186   chromosome 10 (**Supplementary Table 5**). There were 3,397 genes located in these

187   ROH regions, 924 of which were also located in the potential selected sweep region

188   (**Supplementary Figure 3D**).

189

190   **Phenotype-related loci and genes identified using GWAS analysis**

191   We measured 50 agronomic traits in 250 soybean accessions from three geographic

192   locations for three years, and then integrated them using best linear unbiased

193   prediction (BLUP). The 50 traits included traits related to architecture (15), colour (5),

194   isoflavone (1), oil (4), protein (18) and yield (7) and were classified into six categories

195   (**Supplementary Table 6**). We calculated Pearson correlation coefficients for traits so

196   as to compare within and between categories, and found that traits in the same

197   categories were more strongly correlated than traits in different categories. For

198   example, there were strong positive or negative correlations between almost all

199   protein-related traits, oil-related traits and yield-related traits. Linoleic acid content

200   was positively correlated with linolenic acid content, but negatively correlated with

201   oleic acid content. Stem intension was negatively correlated with lodging

202   (**Supplementary Figure 4**). Some traits were evenly distributed, while others were

203   ranked (**Supplementary Figure 5-54**).

204   Using 4,311,814 SNPs with a MAF > 0.05 as an input, we performed GWAS

205   analysis using the mixed linear model (MLM) method for 50 agronomic traits. For

206   each trait, we used a clump based method[8] and defined a significant associated loci

207   (SAL) at a chromosome region with a substantial amount of SNPs associated with the

208   trait. A total of 203 SALs were detected in 43 traits (**Supplementary Figure 5-54,**

209   **Supplementary Table 7**). Since each SAL may contain dozens of genes, we used a

210   functional mutation based haplotype test method for further mining of the most

211   reliable candidate trait associated genes[9]. In particular, we considered only the

212   non-synonymous SNPs, frameshift Indels, mutations within a gene that happened on a

213   start or stop codon, splice site or transcription start sites as effective functional

214   mutations. We used these mutations to classify a gene into different haplotypes, and

215   subsequently tested the phenotypic differences of the accessions belonging to each

216   haplotype. A gene with significant phenotypic differences was defined as significant

217   associated gene (SAG), and 3,165 SAGs were screened in 43 traits. These SAGs

218   include some QTL or genes that have been previously identified, such as: the flower

219   colour related *chr13:16551728-19506795*; pubescence colour related

220   *chr6:16930159-19168772*; seed coat lustre related *chr15:8910798-10281804*;

221   palmitic acid content related *chr5:879095-1682551[4]*; isoflavone content related

222   *chr5:38880530-39142565[10]*; plant height related *Dt1[4]*; and oil content related

223   *FAD2* and *SAT1[11]*, among others. They also contain genes that we have identified

224   for the first time in soybean, such as: the isoflavone content related *GL3* and *GSTL3*;

225   the yield traits related *CKX3*; and the architecture and yield traits related *CYP85A2*.

226

227   **Novel genes related to isoflavone content**

228   Isoflavone content is an important quality-related trait in soybean, but its molecular

229   mechanism is still unclear. Here we identified four SALs related to isoflavone content,

230   namely *chr3:38590023-38728718*, *chr5:3888053-39142565*,

231   *chr13:18342836-18541809*, and *chr5:24726091-24852447*. Only one SAL

232   *chr5:24726091-24852447* overlaps with a previously reported QTL that contains a

233   *GST* (Glutathione S-transferase) gene *GSTT1*[10]. All other SALs are newly identified.

234   There are 48 genes located within these SALs (**Supplementary Table 8**), and three

235   genes (*GSTT1, GL3* and *GSTL3)* may be related to isoflavone content (**Figure 3A, B**).

236   There are two functional mutation sites at *c5s38936266* and *c5s38940717*, forming

237　　two haplotypes for *GSTT1a* and *GSTT1b*, respectively. For each *GSTT1* gene, soybean

238　　accessions with a different haplotype have significantly different isoflavone contents.

239　　Since *GSTT1a* and *GSTT1b* are approximately only 1 kb apart from each other in the

240　　same genome region, we considered the two genes to be one in further analysis. Three

241　　haplotypes were formed by two functional mutation sites when the two *GSTT1* genes

242　　were analysed as one. *Haplotype 1* versus *Haplotype 2*, as well as *Haplotype 2* versus

243　　*Haplotype 3* showed significant differences in isoflavone content, while *Haplotype 1*

244　　versus *Haplotype 3* showed no significant differences (analysed using Tukey's test).

245　　This suggests that *GSTT1a* is associated with isoflavone content due to its linkage

246　　with *GSTT1b*. However, *c5s38936266* did not contribute to the isoflavone content

247　　difference. Thus, only *c5s38940717* on *GSTT1b* was associated with isoflavone

248　　content (**Figure 3C**).

249　　　　We found that two functional mutations, *c5s39035509* and *c5s39036346*

250　　producing two haplotypes in *GL3*, were associated with different isoflavone contents

251　　in soybean accessions (**Figure 3D**). We also identified another *GST* gene, *GSTL3,*

252　　which was located in chromosome 13. Two functional mutations produced three

253　　haplotypes, and significant associations between the different haplotypes and

254　　isoflavone content was detected for each comparison (**Figure 3E**). Based on the

255　　above results, we drew a schematic diagram of the roles of candidate genes according

256　　to their biological functions where we indicate that *GL3* regulates isoflavone synthesis,

257　　whilst *GSTT1* and *GSTL3* participate in isoflavone transport (**Figure 3F**).

258

259　　**Yield related traits and the artificial selection of *CKX3***

260　　Four yield related traits (pod number per plant, seed number per plant, one hundred

261　　seed weight and seed size) have a common SAL located at the ~4.0 Mb to ~4.2 Mb

262　　region of chromosome 17 (**Figure 4A**). Further analysis revealed that this SAL

263　　contains two tandem repeat *CKX* (cytokinin oxidase/dehydrogenase) genes named

264　　*CKX3* and *CKX4*, approximately 15 kb apart from each other (**Figure 4F**).

265　　　　We further analysed the relationship between functional mutations in *CKX3* and

266　　*CKX4*. There were three non-synonymous SNPs on *CKX3* and two non-synonymous

267   SNPs on *CKX4*. As these two genes are only approximately 3 kb apart from each

268   other in the same genome region, we analysed the two genes separately as well as

269   combined as one in relation to their association with haplotypes and traits (**Figure**

270   **4B-E**). Results showed that the functional mutations can both form three haplotypes

271   for *CKX3* and *CKX4 separately*, and four haplotypes for *CKX3+4* combined. For all

272   comparisons in all traits, *Haplotype 1* always showed significant differences

273   compared with the other haplotypes. The relationship between different haplotypes in

274   terms of pod or seed number per plant showed a consistent trend, while that of one

275   hundred seed weight or seed size showed a consistent but opposite trend. There was a

276   phenotypic correlation between pod number per plant and seed number per plant

277   (0.92), and between one hundred seed weight and seed size (0.67) (**Figure 4F**).

278   Furthermore, we observed that *CKX3* and *CKX4* were located in different strands of

279   the same chromosome, suggesting they are more likely to have independent functions.

280   However, we did not detect expression of *CKX4* in the subsequent qPCR validation.

281   Thus, only *CKX3* is regarded as a real candidate gene, while the role of *CKX4* need

282   further study.

283   When we compared the soybean accession information of each haplotype for the

284   four yield-related traits, we observed that most accessions with the *Haplotype 1*

285   genotype had dominant traits (lower pod or seed numbers and larger seeds and seed

286   weights), and were more associated with cultivars. The other haplotypes were mainly

287   landrace-specific haplotypes, and these accessions all belong to *Group 1*. We found

288   that *CKX3* was also located on a strong selective sweep region. This indicates that the

289   functional mutation sites in *CKX3* experienced strong directed artificial selection,

290   resulting in genotype differences and affecting yield related traits. Furthermore, we

291   compared all SAGs and selective sweep regions for all traits, and found that

292   approximately 12% of genes are located in the selected sweep regions, which have

293   experienced artificial selection (**Supplementary Table 9**). It is interesting to note that

294   of all the SAGs located in selective sweep regions, about 55% are related to yield

295   traits, 36% related to protein traits, and less than 10% are related to other traits

296   (**Figure 4G, H**).

297

CYP85A2 is associated with architecture and yield traits

There is one SAL on chromosome 18 which is associated with six traits including plant height, main stem number, stem strength, lodging, podding habit, and seed weight per plant. Interestingly, these traits include both architecture and yield related traits. A cytochrome P450 family gene named *CYP85A2* is located within a 4.37 kb region of this SAL (**Figure 5A, B**). The association of the *CPY85A2* gene with architecture and yield traits in soybean is a novel finding. We also observed that a non-synonymous mutation site *c18s55526062* is involved in producing two haplotypes. The haplotype with a *CC* genotype has a dwarf plant height, a low main stem node number, a high stem strength and a low lodging rate. When plants produced mostly limited or semi-limited pods, their seed weight per plant was also found to increase (**Figure 5C**). Phenotypically, plant height was positively correlated with main stem node number (0.95), while stem strength and lodging were negatively correlated (-0.81), showing a trend consistent with the genotype (**Figure 5D**). The *CC* genotype of *c18s55526062* is a dominant genotype, which is useful when designing an ideal plant type and increasing soybean yield.

314

Different phenotypes coupled through hub gene modules form a complex phenotype-gene network

Based on in-depth exploration of the GWAS results, we observed that one trait is associated with multiple genes and vice versa. At the same time, due to widespread protein-level interactions between genes, a complex network was also found between various phenotypes and genes. In order to explore this further, we used a functional mutation-based haplotype test to screen SAGs in all SALs for all traits. We then constructed a phenotype-gene network which included 34 traits and 853 SAGs (**Figure 6**). At the trait level, they were divided into six categories, namely architecture, colour, oil, isoflavone, protein, and yield. At the gene level, besides the six categories, there emerged a mixed category with which genes associated more than with any one trait category. We found that traits in the same category were

11

327 closely linked within the entire network. However, some trait categories were also

328 linked with each other, such as yield, oil, protein and colour, and they were all closely

329 linked to architecture through common SAGs. This suggests that there are subtle

330 relationships between architecture and other trait categories. In this genetic network,

331 six trait categories were linked through 15 hub nodes containing a total of 367 genes

332 (**Supplementary Table 10**). The largest hub was *HA1* (short form for *Hub*

333 *Architecture 1*). The genes in this hub were only associated with two or more

334 architecture-related traits. Unlike *HA1*, the genes in the *HA2* node were associated

335 with two or more architecture-related traits, but also had protein interactions with

336 other genes. Multiple yield traits were associated with hub *HY1*, containing *CKX3*,

337 while hub *HM4*, which contains *CYP85A2*, was connected with architecture and yield

338 traits.

339

Discussion

341 In this study, we deeply sequenced 250 representative landrace and cultivar soybean

342 accessions. Through population genetics and GWAS analyses, the genetic structure of

343 European soybean varieties was analysed for the first time. Novel candidate genes

344 related to seed isoflavone content, yield and architecture traits were identified.

345 Moreover, we constructed a soybean phenotype-gene interaction network, and found

346 evidence of the improvement of soybean yield related traits at molecular level.

347 A total of ~3T bases and 6M SNPs were obtained, the maximum sequencing

348 depth of a single accession was 22.5x, with the average depth at 11x (higher than

349 previous soybean resequencing studies[2-4]. Eighty-four percent of the accessions

350 with their genome were sequenced for the first time, which provides new data for

351 soybean genome research. Previous soybean research mainly focused on varieties

352 from Asia and North America, but not Europe[3, 4]. This study completed the

353 resequencing of 26 European accessions and, for the first time, outlines a breeding

354 history of European soybean. It was found that European soybean cultivars had higher

355 genetic diversities and lower breeding levels compared to North American cultivars.

12

356 Both European and American soybean cultivars may have been introduced from
357 different ancestors in China. This theory is based on the following findings: there is a
358 small population difference between European varieties and Chinese landraces, and
359 between American varieties and Chinese cultivars, whilst there is a large population
360 difference between American and European varieties (**Figure 2G, Supplementary**
361 **Figure 2**). Our findings are consistent with the current hypothesis that soybean
362 originated in China, and they show that ancestral components from the area of origin
363 are the most complex. This study showed that the heterozygosity rates of most
364 accessions are less than 0.2, except four accessions with a higher heterozygosity
365 which may be caused by their complex ancestral compositions (**Supplementary**
366 **Table S3**). Further combination analysis of the selective sweep and GWAS revealed
367 that artificial selection of soybean at the phenotypic level is consistent with the
368 genome level. Genomic regions associated with yield and quality traits are more
369 likely to experience artificial selection. This may be a reflection of yield- and
370 quality-directed artificial selection of soybean breeding at the genetic level.
371 Furthermore, evidence of functional mutations under artificial selection for a
372 candidate gene *CKX3* related to multiple yield traits were identified. The results of
373 this study provide valuable information for marker-assisted selection, which is vital in
374 the improvement of soybean breeding.

375 Isoflavone is a secondary metabolite produced via phenylpropane metabolic
376 pathways in higher plants. Isoflavone is associated with plant stress resistance,
377 defence against microbial and insect infection, promotion of rhizobium chemotaxis,
378 and the development of rhizome and nitrogen fixation in plants. It also provides health
379 benefits to human, such as in reducing the incidence of cancer and cardiovascular
380 diseases, and regulating the immune response[12]. Therefore, increasing the seed
381 isoflavone content of soybean can improve its nutritional and health benefits.
382 However, few genome-wide studies have investigated the molecular mechanism of
383 soybean's isoflavone content. Isoflavone is synthesized in the cytoplasm, but due to
384 cell cytotoxicity, it cannot accumulate in the cytoplasm and must be continuously
385 transported to vacuoles for storage. Therefore, isoflavone content mainly depends on

386 two factors: synthesis efficiency and transport efficiency[13]. The transcript factor

387 *GL3* is a *bHLH* gene family member which can form the MYB-bHLH-WD40 (MBW)

388 complex with two other transcription factors (*MYB* and *WD40*) to jointly regulate the

389 synthesis of flavonoids and anthocyanin in plants[14]. *GST* can bind with glutathione

390 (GSH) to form an ABC transporter to transport and catalyse the entry of flavonoids

391 into vacuoles for accumulation[13]. In this study, we identified four novel genes that

392 may be associated with isoflavone content. These genes include transcription factors

393 *GL3*, which participate in the regulation of multi-enzyme systems from

394 phenylpropanoid to isoflavone biosynthesis pathways, and two *GST* genes, *GSTT1*

395 and *GSTL3*, which facilitate the transporting of isoflavone from the cytoplasm to

396 vacuoles (**Figure 3F**). In addition, we observed many other genes in the SALs, such

397 as cation/H+ exchanger (*CHX20*), pyrophosphorylase 4 (*PPa4*), an actin

398 depolymerizing factor 7 (*ADF7*), a mitochondrial substrate carrier family protein, a

399 myosin heavy chain-related protein, an *ATP* synthase alpha/beta family protein, and a

400 protein kinase superfamily protein, among others (**Supplementary Table 8**) are all

401 related to isoflavone transport. This over-representation of transport-related genes

402 further suggests that the accumulation of soybean isoflavone is related to its transport

403 to the vacuole. In conclusion, soybean isoflavone content is not merely determined by

404 one or several genes or loci, but by a multiple gene system involved in synthesis,

405 regulation, transport, and storage.

406 　　We observed that other novel candidate genes, such as *CKX3,* is associated with

407 multiple yield traits. We also observed, for the first time, that *CYP85A2* is associated

408 with multiple architecture and yield traits in soybean. It is well-known that cytokinin

409 promotes cell division and plant growth, and *CKX* is one of the key enzymes in

410 cytokinin metabolism. A functional variation in the *CKX* gene may affect the

411 cytokinin metabolism, thus affecting grain yield and related traits. A number of

412 studies on *Arabidopsis thaliana*, rice and other crops have shown that mutation or

413 reduced expression levels of *CKX* family genes are related to a decrease in seed

414 setting rate and an increase in seed weight[15, 16]. *CYP85A2* is involved in the

415 brassinosteroid biosynthesis pathway in *Arabidopsis thaliana* and it converts

14

416   6-deoxocastasterone to castasterone, which is followed by the conversion of

417   castasterone to brassinolide[17]. Brassinosteroids (BRs) are broad-spectrum plant

418   growth regulators, playing an important role in plant growth and development, as well

419   as in biological and abiotic stress responses[18]. Mutations in the *CYP85A2* gene have

420   led to an increased production of the dwarf phenotype[19], and an overexpression of

421   the *CYP85A* family gene resulting in increased BR content, biomass, plant height,

422   plant fresh weight and fruit yield[20]. These results showed that, *CKX3* and *CYP85A2*

423   may affect soybean yield and architecture related traits through different molecular

424   mechanisms. The potential effect of functional mutations in these genes on the

425   phenotypes was further confirmed by our haplotype tests. However, to verify whether

426   these candidate genes and functional mutations are the true cause of the phenotypic

427   differences, further functional verification of these genes is necessary. Multiple

428   methods such as the construction of isolated populations, transgene, gene knockout,

429   gene editing, and expression verification could be used for this purpose. In this study,

430   we   performed   expression   verification   in   seedlings   with   different

431   haplotypes/phenotypes for six genes *GL3*, *GSTL3*, *GSTT1b*, *CKX3*, *CKX4* and

432   *CYP85A2*. The results show that, except for no expression were detected for *CKX4*,

433   all the other five genes were expressed differently for different haplotypes/phenotypes

434   in seedlings; the expression levels of *GL3*, *GSTL3* and *GSTT1b* related to isoflavone

435   content in the strains with high isoflavone content values was significantly higher than

436   that in the strains with low isoflavone content values (T-test, $P < 0.05$); the expression

437   level of *CKX3* in the strains with high yield phenotype values was significantly higher

438   than that in the strains with low yield phenotype values (T-test, $P < 0.05$)

439   (**Supplementary Figure 55**).

440   The highest goal of plant breeding is to aggregate many desired traits into a

441   single genome. Breeders need to simultaneously select and improve multiple related

442   traits. However, because multiple traits are interrelated, it is possible that when

443   screening for a favourable trait one also selects an unfavourable one. Understanding

444   the genetic network behind different traits can help breeders increase breeding

445   efficiency. Although soybean genetic networks for multiple agronomic traits have

15

446   been established at the loci level[4], we built a new phenotype-gene network which

447   includes 34 traits and 853 genes. This network reflects the relationship between

448   phenotypes and genes more directly than the previous phenotype-SAL network, and is

449   more conducive to the discovery of important candidate genes. For example, the *Hub*

450   *Mixed 1* (*HM1*) node was associated with two or more trait types (architecture, colour

451   or protein), while the *HBT* gene in the *HM1* node was associated with six architectural

452   traits (branch number, main stem number, plant height, stem strength, lodging, and

453   podding habit) and four protein-related traits (phenylalanine content, isoleucine

454   content, tyrosine content, and glycine content). It is known that the *HBT* gene belongs

455   to the *CDC27b* gene family and is involved in cell cycle regulation, which is related

456   to cell development and division[21]. Therefore, soybean architecture is likely

457   affected by *HBT*, despite its relationship with amino acid content is unclear. There are

458   also many other interesting examples of the above. Leaf shape is known to affect

459   photosynthesis efficiency, followed by carbohydrate accumulation, and, as a

460   consequence, oil accumulation, while Hub *HM2*, containing *FAD2*, connects oil

461   content and leaf shape[22]. It has also been reported that oil traits and seed coat lustre

462   traits experienced parallel selection during bean domestication[23]. The hub *HM3*

463   node connects oil-related traits and seed coat lustre. Anthocyanin synthesis and

464   isoflavone synthesis share part of their metabolic pathways and hub *HM5* connects

465   colour traits and isoflavone content. Our phenotype-gene network may surpass the

466   phenotype-SAL network in terms of candidate gene selection, which is also beneficial

467   to polymerization breeding programs. For example, breeders can achieve

468   polymerization breeding by directly selecting a favourable gene (such as *CYP85A2*) in

469   Hub *HM4*, which is related to both yield and architecture traits, and eliminate the

470   confusion of other adverse genes located in the same SAL. Furthermore, it is worth

471   noting that the architecture related traits, which centrally connect various other trait

472   categories, have the most extensive connectivity. In other words, there are numerous

473   relationships between architecture related traits and other trait categories in the

474   phenotype-gene network (**Figure 6**), suggesting that some candidate genes related to

475   architecture traits may also be related to other trait types. This may provide theoretical

16

476 support and practical guidance for parallel selection breeding and promote "ideotype"

477 breeding in soybean. The next step is to conduct more in-depth functional

478 investigations on genes with a potential application value, such as *CKX3* and

479 *CYP85A2*. This would help promote the design and breeding process of soybean

480 varieties with a higher yield and quality. Overall, our work is conducive to promoting

481 soybean genome functional research and genomic breeding.

482

483 **Materials and Methods**

484 **Plant materials and phenotyping**

485 A total of 250 soybean varieties were analysed in this study, which were provided by

486 the National Crop Germplasm Resources Platform, Institute of Crop Genetics,

487 Institute of Crop Sciences, Chinese Academy of Agricultural Sciences. All materials

488 were planted and phenotyped at three locations: the Gongzhuling experimental site in

489 the Jilin Academy of Agricultural Sciences (latitude 43.51°, east longitude 124.80°),

490 the Harbin experimental site in the Heilongjiang Academy of Agricultural Sciences

491 (45.68° north latitude, 126.61° east longitude), and the Chifeng experimental site un

492 the Agricultural Science Institute in Inner Mongolia (42.27° north latitude, 118.90°

493 east longitude) in late April of 2008, 2009 and 2010, respectively. Grain protein

494 content was measured using the Kjeldahl method from National Food Safety Standard

495 GB5009.5-2010 China[24], while the grain fatty acid contents were determined using

496 the Soxhlet extraction method from National Food Safety Standard GB/T5512-2008,

497 China[25]. Amino acid content was determined using high performance liquid

498 chromatography (S433D, Seckam, Germany) following a previous amino acid

499 determination method from National Food Safety Standard GB/T 18246-2000,

500 China[26]. Grain isoflavone content was determined using high performance liquid

501 chromatography from National Food Safety Standard GB/T23788-2009, China[27].

502 Finally, the phenotypic data were integrated using the BLUP (Best Linear Unbiased

503 Prediction) method using R[28] in order to remove environmental effects and obtain

504 stable genetic phenotypes. Seeds were planted in (CLC-BIV-M/CLC404-TV, MMM,

505   Germany) at 20°C (with 12-h day/12-h night) and a relative humidity of 60–80% till

506   six leaves stage (about two-week-old). Two-week-old seedlings (24°C, 12-h day/12-h

507   night cycle) were used in this qPCR validation.

508

509   **DNA preparation and sequencing**

510   The genomic DNA for all soybean accessions were extracted from soybean leaves

511   after three weeks of growth. DNA extraction was performed using the

512   cetyltrimethylammonium bromide (CTAB) method[29]. The library for each

513   accession was constructed with an insert size of approximately 500 base pairs,

514   following manufacturer's instructions (Illumina Inc., San Diego, CA, USA). All

515   soybean accessions were sequenced and paired-end 150 bp reads were produced using

516   an Illumina NovaSeq 6000 sequencer at the BerryGenomics Company

517   (http://www.berrygenomics.com/ Beijing, China).

518

519   **Total RNA extraction, cDNA synthesis and qRT-PCR analysis**

520   Total RNA was isolated for each sample using TRIzol Reagent (Invitrogen,

521   Nottingham, UK) according to the manufacturer's instructions. The purified RNA was

522   stored at -80°C until subsequent analyses. According to the manufacturer's

523   instructions (Takara, Shiga, Japan), first-strand cDNA synthesis was performed using

524   M-MLV reverse transcriptase. Quantitative real-time PCR (qRT-PCR) was performed

525   using a SYBR Premix Ex Taq Kit (Takara) and a real-time PCR machine (CFX96;

526   Bio-Rad, Hercules, CA, USA), following the manufacturer's instructions. The

527   procedure used for qRT-PCR was 95°C at 10 minutes, followed by 38 cycles of 15 s at

528   95°C and 60 s at 61-62°C. β-actin was used as a reference gene for analysis of relative

529   expression patterns of mRNA. The reactions were carried out with three biological

530   replicates, with at least two technical replicates for each sample. The data were

531   analyzed using the method according to the previous study[30], and the means ±

532   standard errors (SE) of three biological replicates are presented.

533

534   **Mapping, variant calling and annotation**

535 Raw paired-end resequencing reads were first cleaned by removing reads with

536 adaptors, reads of low quality and reads with "N"s. The high-quality clean reads were

537 then mapped to the soybean reference genome (Williams 82 assembly v2.1) with

538 BWA[31]. Statistical analyses of mapping rate and genomic coverage of clean reads

539 were performed using in-house scripts. The Speedseq pipeline[32] was used for SNP

540 and Indel calling, and vcftools[33] was used for genotype filtering. Missing genotypes

541 were imputed and phased through a localized haplotype clustering algorithm

542 implemented using Beagle v3.0[34]. Variant annotation was performed using

543 ANNOVAR[35] against the soybean gene model set v2.1.42. After annotation, SNPs

544 and Indels were categorized into exonic, intronic, intergenic, splicing, 5'UTRs,

545 3'UTRs, upstream, and downstream. Exonic SNPs were further categorized into

546 synonymous, nonsynonymous, stop gain, and stop loss. Exonic Indels were further

547 categorized into frameshift, non-frameshift, stop gain, and stop loss.

548

Population structure analysis

549

550 Approximately 6M SNPs from the 250 soybean accessions were concatenated for the

551 construction of a phylogenetic tree. Using a neighbour joining algorithm with a

552 pairwise gap deletion method for 100 bootstrap replications, a phylogenetic tree was

553 constructed with MegaCC[36]. The output was displayed using the iTOL[37] web

554 tool. With the whole genome genotype as the input, a principal component analysis

555 (PCA) was done using flashPCA [38] and the first two eigenvectors were plotted. A

556 population admixture analysis with $k = 2$ to $k = 5$ parameters were set to infer the

557 admixture of ancestors using fastSTRUCTURE.

558

Genetic diversity analysis

559

560 Linkage disequilibrium analyses for each subpopulation were performed using

561 PLINK[39] by calculating the correlation coefficient (r^2) of any two SNP pairs in one

562 chromosome. An LD decay plot was drawn using the average r^2 value for the distance

563 from 0 to 1,000 kb. Pairwise IBS calculations were also performed using PLINK and

564 a distance matrix was generated for each subpopulation. Population genetic diversities

565 were measured using VCFtools[33] by calculating $\theta\pi$ and Fst. $\theta\pi$ was used to measure

566 the genetic diversity of each subpopulation, while Fst, plus the allele frequency

567 distribution (AFD) plot (which was generated by in-house scripts), were used to

568 measure genetic diversity between subpopulations. In addition, sliding window

569 calculations of r^2, $\theta\pi$, Fst and $Tajima'D$ values were also performed for genome-wide

570 displays of soybean genetic diversities with a 100 kb window and a 10 kb step.

571

Selective sweep analysis

573 We used two methods to detect selective sweep regions across the soybean genome:

574 $Tajima'D$ combined $\theta\pi$ and Fst combined $\theta\pi$ ratios. Firstly, a genome-wide sliding

575 window calculation of $\theta\pi$, Fst, and $Tajima'D$ values (with a 100 kb window and a 10

576 kb step) were performed on landraces, cultivars, and the whole population,

577 respectively. Secondly, the top 5% of the $Tajima'D$ and $\theta\pi$ windows for the whole

578 population were selected. In addition, the top 5% of the Fst and $\theta\pi$ ratio windows for

579 the landraces versus cultivars were also selected. Thirdly, the selected windows from

580 these two methods were merged together to become the final selective sweep regions.

581 ROH analyses for each accession were performed using PLINK[39] with the

582 parameters of a minimum ROH length set to 300 kb.

583

GWAS and significantly associated loci

585 Association analysis for each trait on each SNP with an MAF larger than 0.05 was

586 performed using a single-locus mixed linear model (MLM) implemented in

587 GEMMA[40] (which corrects confounding by population structure and the

588 relatedness matrix). The GWAS results were displayed using a Manhattan plot and a

589 QQ-plot created with the R package CMplot[41] . A clump based method

590 implemented in PLINK[39] was used to reduce a false peak and to detect real SALs.

591 The P-value cut-off was set to 10^{-5} so as to, firstly, uncover significant associated

592 SNPs. Following this, for each significantly associated SNP, if there were more than

593 10 SNPs within a 100 kb distance that had P-values smaller than 10^{-4}, then the region

594 was regarded as a potential SAL. Finally, all overlapping SALs were merged to

595   generate final SAL sets and the SNP with the smallest P-value in a SAL was defined

596   as a peak.

597

598   **Detection of significantly associated genes**

599   There are usually tens of genes in a SAL, and it is difficult to determine which genes

600   are truly associated with traits and which are irrelevant. We improved a functional

601   mutation-based haplotype test method for SAG discovery in SAL. As most variants

602   within a gene are non-functional, the gene's amino acid sequence and its function will

603   not change. Only a few variants have the potential to change a gene's amino acid

604   sequence, such as nonsynonymous SNPs, frameshift Indels, variants in splicing sites,

605   promoter regions, start codons, and stop codons. These combined functional

606   mutations can only produce two or three different gene haplotypes. It is possible to

607   test the relationship between gene haplotypes and traits. If they are significantly

608   associated, then the gene is also most likely associated with the trait, which is how

609   SAG is defined. In this study, Welch's test was used for a two-group haplotype test

610   and a Tukey's test was used for a multiple group haplotype test to detect SAGs.

611   Functional annotation of SAGs was directly retrieved from SoyBase[42].

612

613   **Network construction**

614   Of all the genes located in the SALs, the most significant SAGs with a P-value

615   smaller than 10^{-5}, and their corresponding traits, were retained to build the

616   phenotype-gene network for soybean. Protein-protein interaction information for

617   soybean was retrieved form the String database[43] and mapped to the soybean genes

618   using BLAST[44]. Construction, visualization and exploration of the network was

619   performed using Cytoscape[45].

620

621   **Data availability**

622   The raw sequence data reported in this paper have been deposited in the Genome

623   Sequence Archive[46] in BIG Data Center[47], Beijing Institute of Genomics (BIG),

624 Chinese Academy of Sciences, under accession numbers CRA002552 that are

625 publicly accessible at https://bigd.big.ac.cn/gsa. The variation data reported in this

626 paper have been deposited in the Genome Variation Map [48] under accession number

627 GVM000076 that can be publicly accessible at

628 http://bigd.big.ac.cn/gvm/getProjectDetail?project=GVM000076. The bioinformatics

629 analysis scripts used in this paper can be download through

630 https://github.com/yjthu/GPB_250SoyReseq.

631

632

CRediT statement

634 **Chunming Yang:** Resources, Investigation, Validation. **Jun Yan:** Methodology,

635 Formal analysis, Writing, Revision. **Shuqin Jiang:** Formal analysis. **Xia Li:**

636 Investigation, Validation. **Haowei Min:** Conceptualization, Supervision, Formal

637 analysis, Writing, Revision. **Xiangfeng Wang:** Conceptualization, Supervision,

638 Writing, Revision. **Dongyun Hao:** Conceptualization, Supervision, Writing, Revision.

639 All authors read and approved the final manuscript.

640

Competing interests

642 The authors declare no competing financial interests.

643

Acknowledgements

References

658 **References**

659 [1] Sedivy, E.J., F. Wu, and Y. Hanzawa. Soybean domestication: the origin, genetic
660   architecture and molecular bases. New Phytol 2017;214:539-553.

661 [2] Lam, H.M., X. Xu, X. Liu, W. Chen, G. Yang, F.L. Wong, et al. Resequencing of
662   31 wild and cultivated soybean genomes identifies patterns of genetic diversity
663   and selection. Nat Genet 2010;42:1053-9.

664 [3] Zhou, Z., Y. Jiang, Z. Wang, Z. Gou, J. Lyu, W. Li, et al. Resequencing 302 wild
665   and cultivated accessions identifies genes related to domestication and
666   improvement in soybean. Nat Biotechnol 2015;33:408-14.

667 [4] Fang, C., Y. Ma, S. Wu, Z. Liu, Z. Wang, R. Yang, et al. Genome-wide association
668   studies dissect the genetic networks underlying agronomical traits in soybean.
669   Genome Biol 2017;18:161.

670 [5] Schmutz, J., S.B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, et al. Genome
671   sequence of the palaeopolyploid soybean. Nature 2010;463:178-83.

672 [6] Nielsen, R. Molecular signatures of natural selection. Annu Rev Genet
673   2005;39:197-218.

674 [7] Purfield, D.C., S. McParland, E. Wall, and D.P. Berry. The distribution of runs of
675   homozygosity and selection signatures in six commercial meat sheep breeds.
676   PLoS One 2017;12:e0176780.

677 [8] Crowell, S., P. Korniliev, A. Falcao, A. Ismail, G. Gregorio, J. Mezey, et al.
678   Genome-wide association and high-resolution phenotyping link Oryza sativa
679   panicle traits to numerous trait-specific QTL clusters. Nat Commun
680   2016;7:10527.

681 [9] Yano, K., E. Yamamoto, K. Aya, H. Takeuchi, P.C. Lo, L. Hu, et al. Genome-wide
682   association study using whole-genome sequencing rapidly identifies new
683   genes influencing agronomic traits in rice. Nat Genet 2016;48:927-34.

684 [10] Meng, S., J. He, T. Zhao, G. Xing, Y. Li, S. Yang, et al. Detecting the QTL-allele
685   system of seed isoflavone content in Chinese soybean landrace population for
686   optimal cross design and gene system exploration. Theor Appl Genet
687   2016;129:1557-76.

688 [11] Zhang, J., X. Wang, Y. Lu, S.J. Bhusal, Q. Song, P.B. Cregan, et al.
689   Genome-wide Scan for Seed Composition Provides Insights into Soybean
690   Quality Improvement and the Impacts of Domestication and Breeding. Mol
691   Plant 2018;11:460-472.

692 [12] Messina, M. A brief historical overview of the past two decades of soy and
693   isoflavone research. J Nutr 2010;140:1350S-4S.

694 [13] Braidot, E., M. Zancani, E. Petrussa, C. Peresson, A. Bertolini, S. Patui, et al.
695   Transport and accumulation of flavonoids in grapevine (Vitis vinifera L.).
696   Plant Signal Behav 2008;3:626-32.

697 [14] Xu, W., C. Dubos, and L. Lepiniec. Transcriptional control of flavonoid
698   biosynthesis by MYB-bHLH-WDR complexes. Trends Plant Sci
699   2015;20:176-85.

700 [15] Li, J., X. Nie, J.L. Tan, and F. Berger. Integration of epigenetic and genetic
701   controls of seed size by cytokinin in Arabidopsis. Proc Natl Acad Sci U S A
702   2013;110:15479-84.

703 [16] Ashikari, M., H. Sakakibara, S. Lin, T. Yamamoto, T. Takashi, A. Nishimura, et al.
704   Cytokinin oxidase regulates rice grain production. Science 2005;309:741-5.

705 [17] Kim, T.W., J.Y. Hwang, Y.S. Kim, S.H. Joo, S.C. Chang, J.S. Lee, et al.
706   Arabidopsis CYP85A2, a cytochrome P450, mediates the Baeyer-Villiger
707   oxidation of castasterone to brassinolide in brassinosteroid biosynthesis. Plant
708   Cell 2005;17:2397-412.

709 [18] De Bruyne, L., M. Hofte, and D. De Vleesschauwer. Connecting growth and
710   defense: the emerging roles of brassinosteroids and gibberellins in plant innate
711   immunity. Mol Plant 2014;7:943-959.

712 [19] Kwon, M.F., Shozo & Ji, H.J. & Kim, Ho Bang & Takatsuto, Suguru & Yoshida,
713   Shigeo & An, Chung & Choe, Sunghwa. A double mutant for the CYP85A1
714   and CYP85A2 genes of Arabidopsis exhibits a brassinosteroid dwarf
715   phenotype. Journal of Plant Biology 2005;48:237-244.

716 [20] Northey, J.G., S. Liang, M. Jamshed, S. Deb, E. Foo, J.B. Reid, et al.
717   Farnesylation mediates brassinosteroid biosynthesis to regulate abscisic acid
718   responses. Nat Plants 2016;2:16114.

719 [21] Perez-Perez, J.M., O. Serralbo, M. Vanstraelen, C. Gonzalez, M.C. Criqui, P.
720   Genschik, et al. Specialization of CDC27 function in the Arabidopsis thaliana
721   anaphase-promoting complex (APC/C). Plant J 2008;53:78-89.

722 [22] Dar, A.A., A.R. Choudhury, P.K. Kancharla, and N. Arumugam. The FAD2 Gene
723   in Plants: Occurrence, Regulation, and Role. Front Plant Sci 2017;8:1789.

724 [23] Zhang, D., L. Sun, S. Li, W. Wang, Y. Ding, S.A. Swarm, et al. Elevation of
725   soybean seed oil content through selection for seed coat shininess. Nat Plants
726   2018;4:30-35.

727 [24] GB/T 5009.5-2010 National food safety standard Determination of protein in
728   foods.

729 [25] GB/T 5512-2008 Inspect of grain and oilseeds—Determination of crude fat
730   content in grain.

731 [26] GB/T 18246-2000 Determination of amino acids in feeds.

732 [27] GB/T 23788-2009 Determination of soybean isoflavone in health-care
733   food-High-performance liquid chromatography.

734 [28] Liu X., Rong J., Liu X. Best linear unbiased prediction for linear combinations in
735   general mixed linear models. Journal of Multivariate Analysis 2008;99 (8):
736   1503–1517.

737 [29] Murray, M.G. and W.F. Thompson. Rapid isolation of high molecular weight

738 plant DNA. Nucleic Acids Res 1980;8:4321-5.

[30] Livak, K.J. and T.D. Schmittgen. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 2001;25:402-8.

[31] Li, H. and R. Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010;26:589-95.

[32] Chiang, C., R.M. Layer, G.G. Faust, M.R. Lindberg, D.B. Rose, E.P. Garrison, et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat Methods 2015;12:966-8.

[33] Danecek, P., A. Auton, G. Abecasis, C.A. Albers, E. Banks, M.A. DePristo, et al. The variant call format and VCFtools. Bioinformatics 2011;27:2156-8.

[34] Browning, S.R. and B.L. Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 2007;81:1084-97.

[35] Wang, K., M. Li, and H. Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38:e164.

[36] Kumar, S., G. Stecher, D. Peterson, and K. Tamura. MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. Bioinformatics 2012;28:2685-6.

[37] Letunic, I. and P. Bork. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res 2019;47:W256-W259.

[38] Abraham, G. and M. Inouye. Fast principal component analysis of large-scale genome-wide data. PLoS One 2014;9:e93766.

[39] Chang, C.C., C.C. Chow, L.C. Tellier, S. Vattikuti, S.M. Purcell, and J.J. Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 2015;4:7.

[40] Zhou, X. and M. Stephens. Genome-wide efficient mixed-model analysis for association studies. Nat Genet 2012;44:821-4.

[41] Yin L., Zhang H., Tang Z., Xu J., Yin D., Zhang Z., Yuan X., Zhu M., Zhao S., Li X. et al. rMVP: A Memory-efficient, Visualization-enhanced, and Parallel-accelerated tool for Genome-Wide Association Study. bioRxiv 2020; doi:10.1101/2020.08.20.258491.

[42] Grant, D., R.T. Nelson, S.B. Cannon, and R.C. Shoemaker. SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res 2010;38:D843-6.

[43] Szklarczyk, D., J.H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res 2017;45:D362-D368.

[44] Madden, T. The BLAST sequence analysis tool. The NCBI Handbook [Internet]. 2nd edition 2013;

[45] Shannon, P., A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13:2498-504.

782 [46] Wang, Y., F. Song, J. Zhu, S. Zhang, Y. Yang, T. Chen, et al. GSA: Genome
783   Sequence   Archive<sup/>.   Genomics   Proteomics   Bioinformatics
784   2017;15:14-18.
785 [47] National Genomics Data Center, M. and Partners. Database Resources of the
786   National   Genomics   Data   Center   in   2020.   Nucleic   Acids   Res
787   2020;48:D24-D33.
788 [48] Song, S., D. Tian, C. Li, B. Tang, L. Dong, J. Xiao, et al. Genome Variation Map:
789   a data repository of genome variations in BIG Data Center. Nucleic Acids Res
790   2018;46:D944-D949.

791

792

793

794

795

796

797

798

799

800

801 **Tables**

802 **Table1.** Functional variants of representative significant associated genes

Variant ID	Chrom	Positon	Ref	Alt	Variant type	Gene ID	Gene symbol
c5s38936266	5	38936266	C	T	nonsynonymous SNV	GLYMA_05G206900	*GSTT1a*
c5s38940717	5	38940717	C	T	nonsynonymous SNV	GLYMA_05G207000	*GSTT1b*
c5s39035509	5	39035509	G	C	nonsynonymous SNV	GLYMA_05G208300	*GL3*
c5s39036346	5	39036346	T	C	nonsynonymous SNV	GLYMA_05G208300	*GL3*
c13s24804891	13	24804891	C	T	nonsynonymous SNV	GLYMA_13G135600	*GSTL3*
c13s24805363	13	24805363	A	T	splicing SNV	GLYMA_13G135600	*GSTL3*
c17s4143663	17	4143663	C	T	nonsynonymous SNV	GLYMA_17G054500	*CKX3*
c17s4143832	17	4143832	T	C	nonsynonymous SNV	GLYMA_17G054500	*CKX3*
c17s4146922	17	4146922	G	T	nonsynonymous SNV	GLYMA_17G054500	*CKX3*

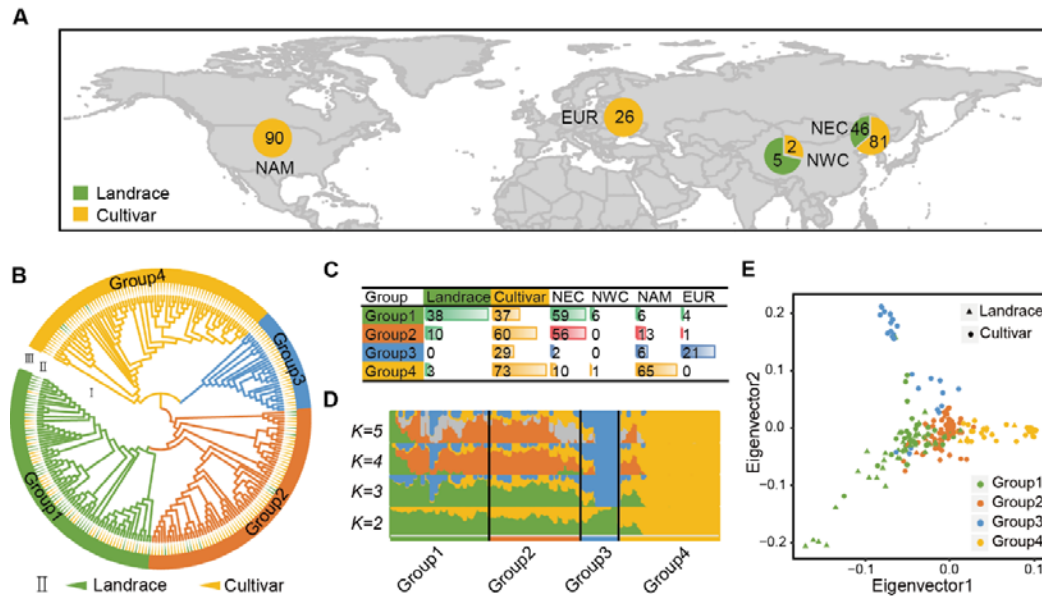| c17s4151713 | 17 | 4151713 | C | A | nonsynonymous SNV | GLYMA_17G054600 | *CKX4* |
| c17s4151752 | 17 | 4151752 | T | C | nonsynonymous SNV | GLYMA_17G054600 | *CKX4* |
| c18s55526062 | 18 | 55526062 | C | T | nonsynonymous SNV | GLYMA_18G272300 | *CYP85A2* |

803

804

805

806

807

808

809

810

811

812

813

814

Figures

Figure 1

Figure 1 Population structure of 250 soybean accessions. **A.** Geographic distribution of 250 soybean landraces and cultivars. Landraces are shown with green color and cultivars are shown with yellow color. **B.** Phylogenetic tree constructed for all soybean accessions. Group1-4 are shown with different colors, Landraces are labeled with green triangles and cultivars are labeled with yellow triangles. **C.** Statistics of the geographic origin for each subpopulation. **D.** Mixed ancestors analysis for soybean subpopulations. Each color represents an ancestral component. K from 2 to 5 are set to trace different ancestral components. **E.** PCA plot of the first two eigenvectors for all soybean accessions. Landraces and cultivars are shown will different shape, while groups are shown with different colors.
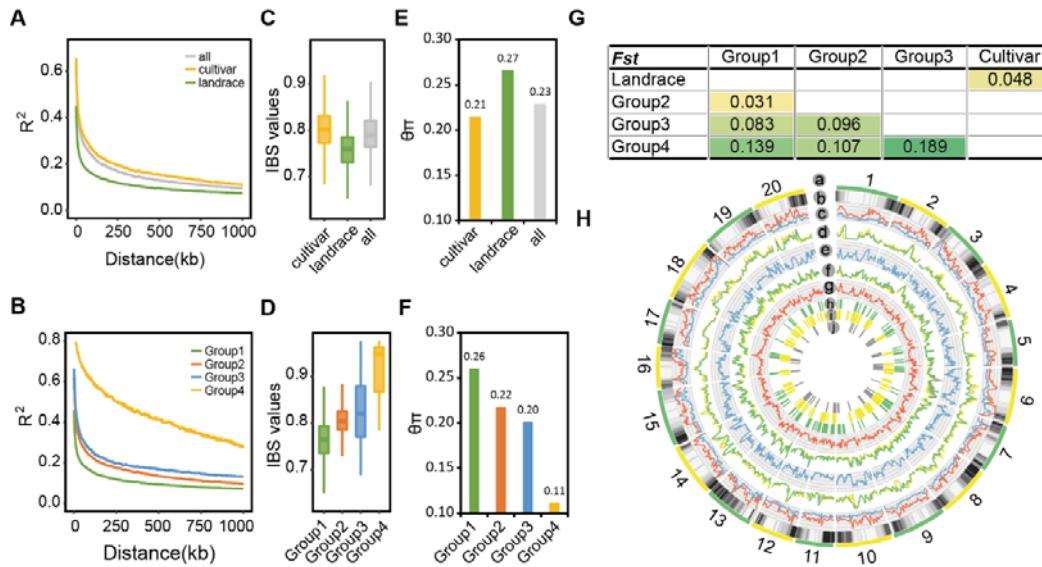
835 **Figure 2**

836

837 **Figure 2** Genetic diversity of soybean subpopulations. **A.** LD decay plots for landrace

838 (green), cultivars (yellow) and all soybean accessions (grey). **B.** LD decay plots for

839 soybean subpopulations. **C.** IBS values distribution for landrace (green), cultivars

840 (yellow) and all soybean accessions (grey). **D.** IBS values distribution for soybean

841 subpopulations.**E.** Comparison of $\theta\pi$ values for landrace (green), cultivars (yellow)

842 and all soybean accessions (grey). **F.** Comparison of $\theta\pi$ values for soybean

843 subpopulations. **G.** Comparison of *Fst* values between subpopulations. **H.** Landscape

844 of soybean genetic diversity across the whole genome. (a) Chromosomes. (b) Density

845 of genes (c) Density of SNPs (red) and Indels (blue). (d) LD values distribution for

846 landraces(green), cultivars(yellow) and all accessions(grey). (e) *Fst* values

847 distribution of landraces versus cultivars (f) $\theta\pi$ values distribution for

848 landraces(green), cultivars(yellow) and all accessions(grey). (g) *Tajima'D* values

849 distribution of all accessions. (h) Putative selective sweep regions detected by

850 *Tajima'D* combine $\theta\pi$. (i) Putative selective sweep regions detected by *Fst* combine

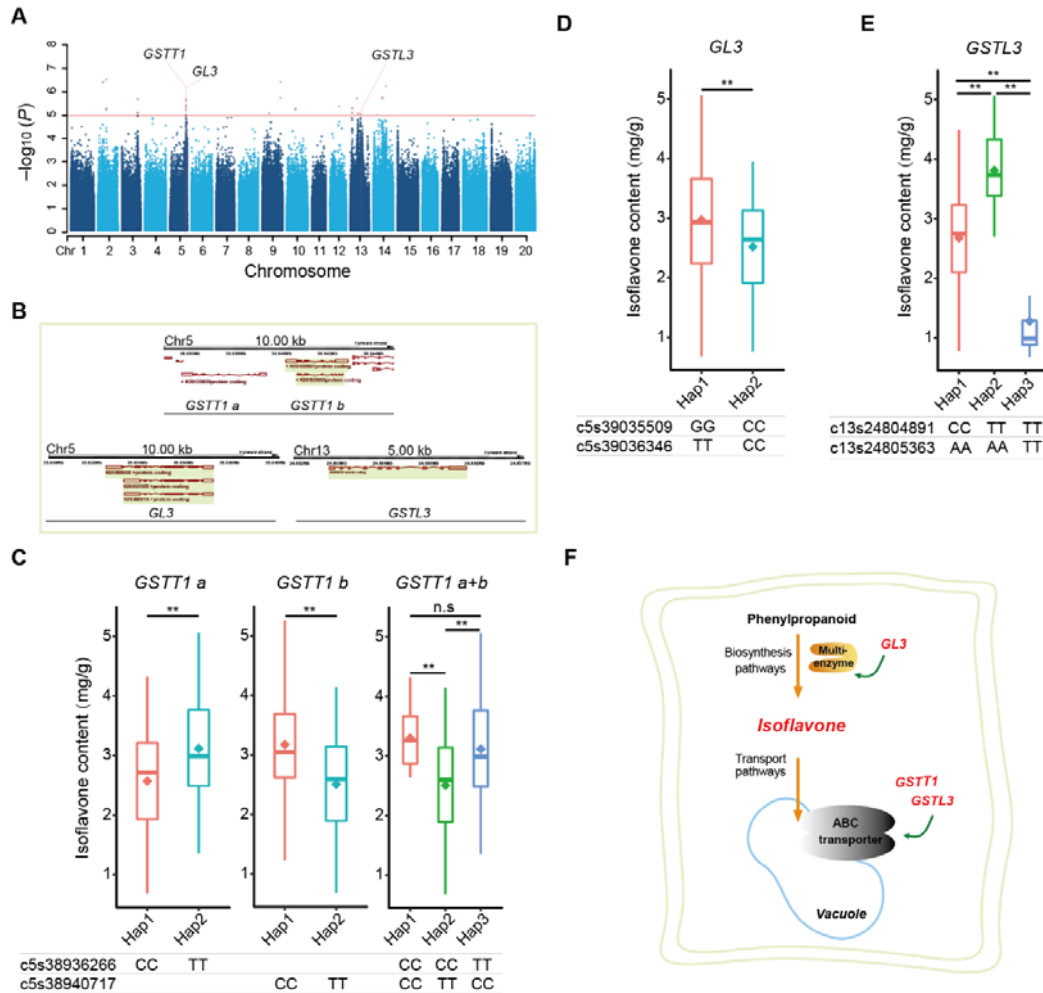851 $\theta\pi$ ratios. (j) ROH region larger than 300 Kb.

852

853

854

855

29

856

857 **Figure 3**

858

859

860 **Figure 3** GWAS of soybean isoflavone content. **A.** Manhantan plot and four candidate
861 genes for soybean isoflavone content**. B.** Chromosome location and transcripts
862 structure of the candidate genes. **C.** Soybean isoflavone content distribution for the
863 haplotypes of gene *GSTT1*. **D.** Soybean isoflavone content distribution for the
864 haplotypes of gene *GL3*. **E.** Soybean isoflavone content distribution for the
865 haplotypes of gene *GSTL3*. **F.** Diagram of soybean isoflavone synthesis and transport,
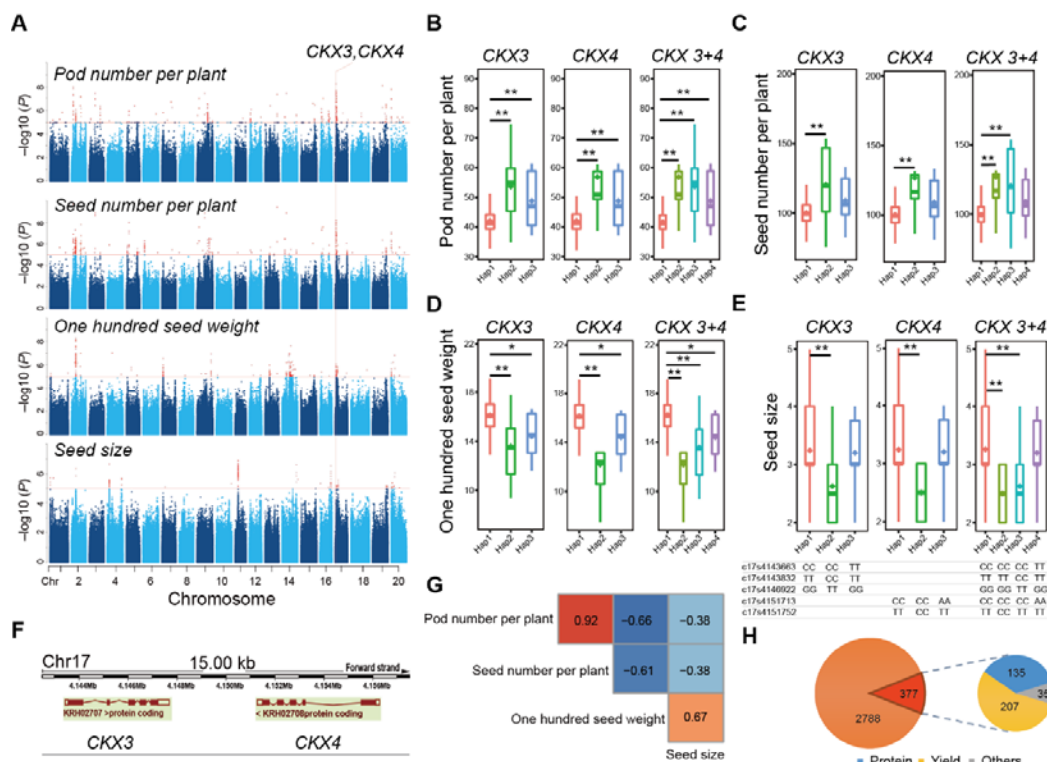866 and the roles of candidate genes detected by GWAS. (*$P < 0.05$; **$P < 0.01$; n.s., not
867 significant)

868

869

870 **Figure 4**

871

872

873 **Figure 4** Association of *CKX* and yield related traits in soybean. **A.** Manhantan plot

874 of four yield related traits pod number per plant, seed number per plant, one hundred

875 seed weight and seed size, and the candidate *CKX* genes. **B.** Pod number per plant

876 distribution for the haplotypes of *CKX* genes.**C.** Seed number per plant distribution

877 for the haplotypes of *CKX* genes. **D.** One hundred seed weight distribution for the

878 haplotypes of *CKX* genes. **E.** Seed size distribution for the haplotypes of *CKX* genes.

879 **F.** Chromosome location and transcripts structure of *CKX3* and *CKX4*. **G.** Phenotype

880 correlation of four traits. **H.** Statistics of SAGs located in selective sweep regions and

881 their percentage for trait categories. (*$P < 0.05$; **$P < 0.01$)

882

883

884

885

886

887   **Figure 5**

888
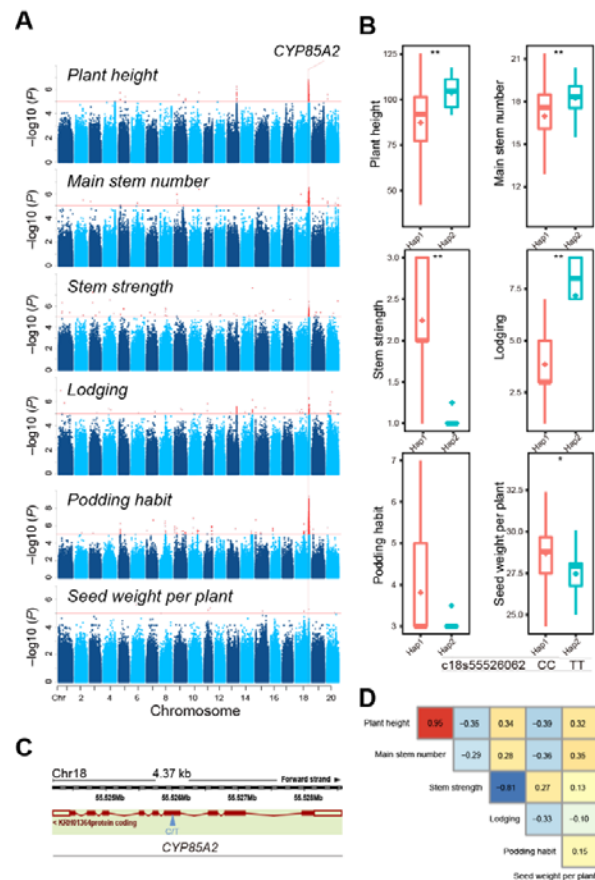
889

890   **Figure 5** Association of *CYP85A2* and architecture or yield related traits in soybean.

891   **A.** Manhantan plot of plant height, main stem number, stem strength, lodging,

892   podding habit, seed weight per plant, and the candidate gene *CYP85A2*. **B.** Traits

893   distribution for the haplotypes of gene *CYP85A2*. **C.** Chromosome location and

894   transcripts structure of *CYP85A2*. **D.** Phenotype correlation of the six traits. (*$P <$*

895   0.05; **$P < 0.01$*)

896

897

898

899 **Figure 6**

900

901

902 **Figure 6** Phenotype-gene genetic network in soybean. Traits are solid rhombuses,
903 genes are solid circles, and hubs are hollow ellipses. Six trait categories, their
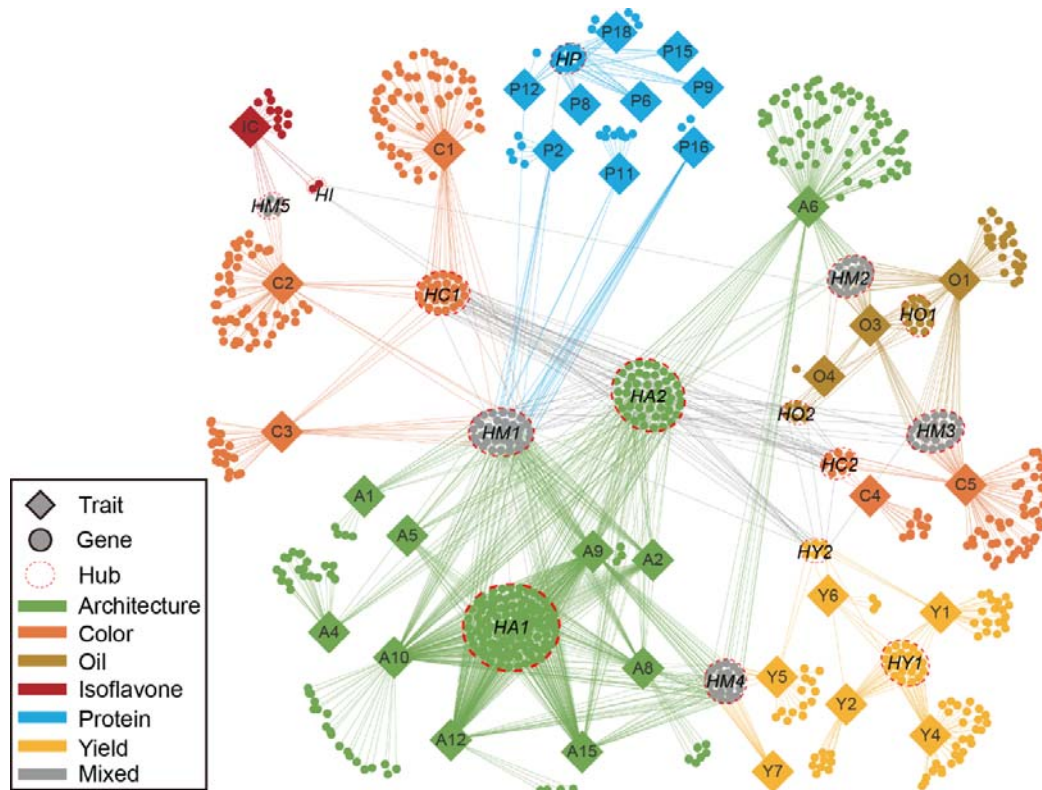904 associated genes and links between them are colored accordingly, genes associated
905 with more than one trait categories are colored grey. Genes with protein-protein
906 interaction are linked with gray lines.

907

908

909

910

911

912

913

914

915

Supplementary material

Supplementary Tables 1-10; Supplementary Figures 1-55.

Supplementary Table 1 Summary of mapping and coverage

Supplementary Table 2 Summary of SNPs and Indels

Supplementary Table 3 The ancestry proportion estimates for each accession

Supplementary Table 4 Putative regions experiencing selective sweeps

Supplementary Table 5 Summary of ROH regions in soybean varieties

Supplementary Table 6 Information of 50 agronomic traits

Supplementary Table 7 Summary of significant associated loci detected by GWAS analysis

Supplementary Table 8 Genes located in significant associated loci for isoflavone content

Supplementary Table 9 Genes located in both significant associated loci and putative selective sweep regions

Supplementary Table 10 Summary of hub genes in soybean agronomic traits networks

Figure S1 SNP density distribution across soybean chromosomes.

Figure S2 Allele frequency distribution between soybean subpopulations.

Figure S3 Selective sweep analysis for 250 soybean accessions. **A.** Selective sweep analysis by *Tajima'D* combine $\theta\pi$. **B.** Selective sweep analysis by *Fst* combine $\theta\pi$ ratios. Red dots present the top 5% selected windows. **C.** Venn diagram of genes screened by two selective sweep analysis methods. **D.** Venn diagram of genes screened by two selective sweep analysis methods and ROH analysis.

Figure S4 Phenotype correlations between 50 soybean traits.

942   **Figure S5** GWAS of pod height at bottom using MLM. A. Density distribution of pod

943   height at bottom. B. Manhattan plots for pod height at bottom. Negative log10

944   P-values from a genome-wide scan are plotted against SNP positions of 20

945   chromosomes. C. Quantile-quantile plot for pod height at bottom. The horizontal red

946   line indicates the significant threshold (10-5). Trait-associated SNPs above the

947   significant threshold are colored in red.

948   **Figure S6** GWAS of effective branch number using MLM. A. Density distribution of

949   effective branch number. B. Manhattan plots for effective branch number. Negative

950   log10 P-values from a genome-wide scan are plotted against SNP positions of 20

951   chromosomes. C. Quantile-quantile plot for effective branch number. The horizontal

952   red line indicates the significant threshold (10-5). Trait-associated SNPs above the

953   significant threshold are colored in red.

954   **Figure S7** GWAS of pubescence density using MLM. **A.** Density distribution of

955   pubescence density. **B.** Manhattan plots for pubescence density. Negative $\log_{10}$

956   P-values from a genome-wide scan are plotted against SNP positions of 20

957   chromosomes. **C.** Quantile-quantile plot for pubescence density. The horizontal red

958   line indicates the significant threshold (10^{-5}). Trait-associated SNPs above the

959   significant threshold are colored in red.

960   **Figure S8** GWAS of defollation using MLM. **A.** Density distribution of defollation. **B.**

961   Manhattan plots for defollation. Negative $\log_{10}$ P-values from a genome-wide scan are

962   plotted against SNP positions of 20 chromosomes. **C.** Quantile-quantile plot for

963   defollation. The horizontal red line indicates the significant threshold (10^{-5}).

964   Trait-associated SNPs above the significant threshold are colored in red.

965   **Figure S9** GWAS of inflorenscence length using MLM. **A.** Density distribution of

966   inflorenscence length. **B.** Manhattan plots for inflorenscence length. Negative $\log_{10}$

967   P-values from a genome-wide scan are plotted against SNP positions of 20

968   chromosomes. **C.** Quantile-quantile plot for inflorenscence length. The horizontal red

969   line indicates the significant threshold (10^{-5}). Trait-associated SNPs above the

970   significant threshold are colored in red.

971   **Figure S10** GWAS of leaf shape using MLM. **A.** Density distribution of leaf shape. **B.**

972   Manhattan plots for leaf shape. Negative $\log_{10}$ P-values from a genome-wide scan are

973   plotted against SNP positions of 20 chromosomes. **C.** Quantile-quantile plot for leaf

974   shape. The horizontal red line indicates the significant threshold (10^{-5}).

975   Trait-associated SNPs above the significant threshold are colored in red.

976   **Figure S11** GWAS of leaflet size using MLM. **A.** Density distribution of leaflet size.

977   **B.** Manhattan plots for leaflet size. Negative $\log_{10}$ P-values from a genome-wide scan

978   are plotted against SNP positions of 20 chromosomes. **C.** Quantile-quantile plot for

979   leaflet size. The horizontal red line indicates the significant threshold (10^{-5}).

980   Trait-associated SNPs above the significant threshold are colored in red.

981   **Figure S12** GWAS of lodging using MLM. **A.** Density distribution of lodging. **B.**

982   Manhattan plots for lodging. Negative $\log_{10}$ P-values from a genome-wide scan are

983   plotted against SNP positions of 20 chromosomes. **C.** Quantile-quantile plot for

984   lodging. The horizontal red line indicates the significant threshold (10^{-5}).

985   Trait-associated SNPs above the significant threshold are colored in red.

986   **Figure S13** GWAS of number of nodes on main stem using MLM. **A.** Density

987   distribution of number of nodes on main stem. **B.** Manhattan plots for number of

988   nodes on main stem. Negative $\log_{10}$ P-values from a genome-wide scan are plotted

989   against SNP positions of 20 chromosomes. **C.** Quantile-quantile plot for number of

990   nodes on main stem. The horizontal red line indicates the significant threshold (10^{-5}).

991   Trait-associated SNPs above the significant threshold are colored in red.

992   **Figure S14** GWAS of plant height using MLM. **A.** Density distribution of plant

993   height. **B.** Manhattan plots for plant height. Negative $\log_{10}$ P-values from a

994   genome-wide scan are plotted against SNP positions of 20 chromosomes. **C.**

995   Quantile-quantile plot for plant height. The horizontal red line indicates the significant

996   threshold (10^{-5}). Trait-associated SNPs above the significant threshold are colored in

997   red.

998   **Figure S15** GWAS of plant type using MLM. **A.** Density distribution of plant type. **B.**

999   Manhattan plots for plant type. Negative $\log_{10}$ P-values from a genome-wide scan are

1000   plotted against SNP positions of 20 chromosomes. **C.** Quantile-quantile plot for plant

1001   type. The horizontal red line indicates the significant threshold (10^{-5}). Trait-associated

1002   SNPs above the significant threshold are colored in red.

1003   **Figure S16** GWAS of stem termination using MLM. **A.** Density distribution of

1004   podding habit. **B.** Manhattan plots for stem termination. Negative $\log_{10}$ P-values from

1005   a genome-wide scan are plotted against SNP positions of 20 chromosomes. **C.**

1006   Quantile-quantile plot for stem termination. The horizontal red line indicates the

1007   significant threshold (10^{-5}). Trait-associated SNPs above the significant threshold are

1008   colored in red.

1009   **Figure S17** GWAS of seed crack using MLM. **A.** Density distribution of seed crack.

1010   **B.** Manhattan plots for seed crack. Negative $\log_{10}$ P-values from a genome-wide scan

1011   are plotted against SNP positions of 20 chromosomes. **C.** Quantile-quantile plot for

1012   seed crack. The horizontal red line indicates the significant threshold (10^{-5}).

1013   Trait-associated SNPs above the significant threshold are colored in red.

1014   **Figure S18** GWAS of stem diameter using MLM. **A.** Density distribution of stem

1015   diameter. **B.** Manhattan plots for stem diameter. Negative $\log_{10}$ P-values from a

1016   genome-wide scan are plotted against SNP positions of 20 chromosomes. **C.**

1017   Quantile-quantile plot for stem diameter. The horizontal red line indicates the

1018   significant threshold (10^{-5}). Trait-associated SNPs above the significant threshold are

1019   colored in red.

1020   **Figure S19** GWAS of stem intension using MLM. **A.** Density distribution of stem

1021   intension. **B.** Manhattan plots for stem intension. Negative $\log_{10}$ P-values from a

1022   genome-wide scan are plotted against SNP positions of 20 chromosomes. **C.**

1023   Quantile-quantile plot for stem intension. The horizontal red line indicates the

1024   significant threshold (10^{-5}). Trait-associated SNPs above the significant threshold are

1025   colored in red.

1026   **Figure S20** GWAS of pubescence color using MLM. **A.** Density distribution of

1027   pubescence color. **B.** Manhattan plots for pubescence color. Negative $\log_{10}$ P-values

1028   from a genome-wide scan are plotted against SNP positions of 20 chromosomes. **C.**

1029   Quantile-quantile plot for pubescence color. The horizontal red line indicates the

1030   significant threshold (10^{-5}). Trait-associated SNPs above the significant threshold are

1031   colored in red.

1032   **Figure S21** GWAS of flower color using MLM. **A.** Density distribution of flower

1033   color. **B.** Manhattan plots for flower color. Negative $\log_{10}$ P-values from a

1034   genome-wide scan are plotted against SNP positions of 20 chromosomes. **C.**

1035   Quantile-quantile plot for flower color. The horizontal red line indicates the

1036   significant threshold (10^{-5}). Trait-associated SNPs above the significant threshold are

1037   colored in red.

1038   **Figure S22** GWAS of leaf color using MLM. **A.** Density distribution of leaf color. **B.**

1039   Manhattan plots for leaf color. Negative $\log_{10}$ P-values from a genome-wide scan are

1040   plotted against SNP positions of 20 chromosomes. **C.** Quantile-quantile plot for leaf

1041   color. The horizontal red line indicates the significant threshold (10^{-5}).

1042   Trait-associated SNPs above the significant threshold are colored in red.

1043   **Figure S23** GWAS of mature pod color using MLM. **A.** Density distribution of

1044   mature pod color. **B.** Manhattan plots for mature pod color. Negative $\log_{10}$ P-values

1045   from a genome-wide scan are plotted against SNP positions of 20 chromosomes. **C.**

1046   Quantile-quantile plot for mature pod color. The horizontal red line indicates the

1047   significant threshold (10^{-5}). Trait-associated SNPs above the significant threshold are

1048   colored in red.

1049   **Figure S24** GWAS of seed coat luster using MLM. **A.** Density distribution of seed

1050   coat luster. **B.** Manhattan plots for seed coat luster. Negative $\log_{10}$ P-values from a

1051   genome-wide scan are plotted against SNP positions of 20 chromosomes. **C.**

1052   Quantile-quantile plot for seed coat luster. The horizontal red line indicates the

1053   significant threshold (10^{-5}). Trait-associated SNPs above the significant threshold are

1054   colored in red.

1055   **Figure S25** GWAS of isoflavone content using MLM. **A.** Density distribution of

1056   isoflavone content. **B.** Manhattan plots for isoflavone content. Negative $\log_{10}$ P-values

1057   from a genome-wide scan are plotted against SNP positions of 20 chromosomes. **C.**

1058   Quantile-quantile plot for isoflavone content. The horizontal red line indicates the

1059   significant threshold (10^{-5}). Trait-associated SNPs above the significant threshold are

1060   colored in red.

1061   **Figure S26** GWAS of linoleic acid content using MLM. **A.** Density distribution of

1062   linoleic acid content. **B.** Manhattan plots for linoleic acid content. Negative $\log_{10}$

1063   P-values from a genome-wide scan are plotted against SNP positions of 20

1064   chromosomes. **C.** Quantile-quantile plot for linoleic acid content. The horizontal red

1065   line indicates the significant threshold (10^{-5}). Trait-associated SNPs above the

1066   significant threshold are colored in red.

1067   **Figure S27** GWAS of linolenic acid content using MLM. **A.** Density distribution of

1068   linolenic acid content. **B.** Manhattan plots for linolenic acid content. Negative $\log_{10}$

1069   P-values from a genome-wide scan are plotted against SNP positions of 20

1070   chromosomes. **C.** Quantile-quantile plot for linolenic acid content. The horizontal red

1071   line indicates the significant threshold (10^{-5}). Trait-associated SNPs above the

1072   significant threshold are colored in red.

1073   **Figure S28** GWAS of oleic acid content using MLM. **A.** Density distribution of oleic

1074 acid content. **B.** Manhattan plots for oleic acid content. Negative $\log_{10}$ P-values from a

1075 genome-wide scan are plotted against SNP positions of 20 chromosomes. **C.**

1076 Quantile-quantile plot for oleic acid content. The horizontal red line indicates the

1077 significant threshold (10^{-5}). Trait-associated SNPs above the significant threshold are

1078 colored in red.

1079 **Figure S29** GWAS of palmitic acid content using MLM. **A.** Density distribution of

1080 palmitic acid content. **B.** Manhattan plots for palmitic acid content. Negative $\log_{10}$

1081 P-values from a genome-wide scan are plotted against SNP positions of 20

1082 chromosomes. **C.** Quantile-quantile plot for palmitic acid content. The horizontal red

1083 line indicates the significant threshold (10^{-5}). Trait-associated SNPs above the

1084 significant threshold are colored in red.

1085 **Figure S30** GWAS of crude protein content using MLM. **A.** Density distribution of

1086 crude protein content. **B.** Manhattan plots for crude protein content. Negative $\log_{10}$

1087 P-values from a genome-wide scan are plotted against SNP positions of 20

1088 chromosomes. **C.** Quantile-quantile plot for crude protein content. The horizontal red

1089 line indicates the significant threshold (10^{-5}). Trait-associated SNPs above the

1090 significant threshold are colored in red.

1091 **Figure S31** GWAS of alanine content using MLM. **A.** Density distribution of alanine

1092 content. **B.** Manhattan plots for alanine content. Negative $\log_{10}$ P-values from a

1093 genome-wide scan are plotted against SNP positions of 20 chromosomes. **C.**

1094 Quantile-quantile plot for alanine content. The horizontal red line indicates the

1095 significant threshold (10^{-5}). Trait-associated SNPs above the significant threshold are

1096   colored in red.

1097   **Figure S32** GWAS of arginine content using MLM. **A.** Density distribution of

1098   arginine content. **B.** Manhattan plots for arginine content. Negative $\log_{10}$ P-values

1099   from a genome-wide scan are plotted against SNP positions of 20 chromosomes. **C.**

1100   Quantile-quantile plot for arginine content. The horizontal red line indicates the

1101   significant threshold (10^{-5}). Trait-associated SNPs above the significant threshold are

1102   colored in red.

1103   **Figure S33** GWAS of aspartic acid content using MLM. **A.** Density distribution of

1104   aspartic acid content. **B.** Manhattan plots for aspartic acid content. Negative $\log_{10}$

1105   P-values from a genome-wide scan are plotted against SNP positions of 20

1106   chromosomes. **C.** Quantile-quantile plot for aspartic acid content. The horizontal red

1107   line indicates the significant threshold (10^{-5}). Trait-associated SNPs above the

1108   significant threshold are colored in red.

1109   **Figure S34** GWAS of glutamate content using MLM. **A.** Density distribution of

1110   glutamate content. **B.** Manhattan plots for glutamate content. Negative $\log_{10}$ P-values

1111   from a genome-wide scan are plotted against SNP positions of 20 chromosomes. **C.**

1112   Quantile-quantile plot for glutamate content. The horizontal red line indicates the

1113   significant threshold (10^{-5}). Trait-associated SNPs above the significant threshold are

1114   colored in red.

1115   **Figure S35** GWAS of glycine content using MLM. **A.** Density distribution of glycine

1116   content. **B.** Manhattan plots for glycine content. Negative $\log_{10}$ P-values from a

1117   genome-wide scan are plotted against SNP positions of 20 chromosomes. **C.**

1118   Quantile-quantile plot for glycine content. The horizontal red line indicates the

1119   significant threshold (10^{-5}). Trait-associated SNPs above the significant threshold are

1120   colored in red.

1121   **Figure S36** GWAS of histidine content using MLM. **A.** Density distribution of

1122   histidine content. **B.** Manhattan plots for histidine content. Negative $\log_{10}$ P-values

1123   from a genome-wide scan are plotted against SNP positions of 20 chromosomes. **C.**

1124   Quantile-quantile plot for histidine content. The horizontal red line indicates the

1125   significant threshold (10^{-5}). Trait-associated SNPs above the significant threshold are

1126   colored in red.

1127   **Figure S37** GWAS of isoleucine content using MLM. **A.** Density distribution of

1128   isoleucine content. **B.** Manhattan plots for isoleucine content. Negative $\log_{10}$ P-values

1129   from a genome-wide scan are plotted against SNP positions of 20 chromosomes. **C.**

1130   Quantile-quantile plot for isoleucine content. The horizontal red line indicates the

1131   significant threshold (10^{-5}). Trait-associated SNPs above the significant threshold are

1132   colored in red.

1133   **Figure S38** GWAS of leucine content using MLM. **A.** Density distribution of leucine

1134   content. **B.** Manhattan plots for leucine content. Negative $\log_{10}$ P-values from a

1135   genome-wide scan are plotted against SNP positions of 20 chromosomes. **C.**

1136   Quantile-quantile plot for leucine content. The horizontal red line indicates the

1137   significant threshold (10^{-5}). Trait-associated SNPs above the significant threshold are

1138   colored in red.

1139   **Figure S39** GWAS of lysine content using MLM. **A.** Density distribution of lysine

1140   content. **B.** Manhattan plots for lysine content. Negative $\log_{10}$ P-values from a

1141   genome-wide scan are plotted against SNP positions of 20 chromosomes. **C.**

1142   Quantile-quantile plot for lysine content. The horizontal red line indicates the

1143   significant threshold (10^{-5}). Trait-associated SNPs above the significant threshold are

1144   colored in red.

1145   **Figure S40** GWAS of methionine content using MLM. **A.** Density distribution of

1146   methionine content. **B.** Manhattan plots for methionine content. Negative $\log_{10}$

1147   P-values from a genome-wide scan are plotted against SNP positions of 20

1148   chromosomes. **C.** Quantile-quantile plot for methionine content. The horizontal red

1149   line indicates the significant threshold (10^{-5}). Trait-associated SNPs above the

1150   significant threshold are colored in red.

1151   **Figure S41** GWAS of phenylalanine content using MLM. **A.** Density distribution of

1152   phenylalanine content. **B.** Manhattan plots for phenylalanine content. Negative $\log_{10}$

1153   P-values from a genome-wide scan are plotted against SNP positions of 20

1154   chromosomes. **C.** Quantile-quantile plot for phenylalanine content. The horizontal red

1155   line indicates the significant threshold (10^{-5}). Trait-associated SNPs above the

1156   significant threshold are colored in red.

1157   **Figure S42** GWAS of proline content using MLM. **A.** Density distribution of proline

1158   content. **B.** Manhattan plots for proline content. Negative $\log_{10}$ P-values from a

1159   genome-wide scan are plotted against SNP positions of 20 chromosomes. **C.**

1160   Quantile-quantile plot for proline content. The horizontal red line indicates the

1161   significant threshold (10^{-5}). Trait-associated SNPs above the significant threshold are

1162 colored in red.

1163 **Figure S43** GWAS of serine content using MLM. **A.** Density distribution of serine

1164 content. **B.** Manhattan plots for serine content. Negative $\log_{10}$ P-values from a

1165 genome-wide scan are plotted against SNP positions of 20 chromosomes. **C.**

1166 Quantile-quantile plot for serine content. The horizontal red line indicates the

1167 significant threshold (10^{-5}). Trait-associated SNPs above the significant threshold are

1168 colored in red.

1169 **Figure S44** GWAS of threonine content using MLM. **A.** Density distribution of

1170 threonine content. **B.** Manhattan plots for threonine content. Negative $\log_{10}$ P-values

1171 from a genome-wide scan are plotted against SNP positions of 20 chromosomes. **C.**

1172 Quantile-quantile plot for threonine content. The horizontal red line indicates the

1173 significant threshold (10^{-5}). Trait-associated SNPs above the significant threshold are

1174 colored in red.

1175 **Figure S45** GWAS of tyrosine content using MLM. **A.** Density distribution of

1176 tyrosine content. **B.** Manhattan plots for tyrosine content. Negative $\log_{10}$ P-values

1177 from a genome-wide scan are plotted against SNP positions of 20 chromosomes. **C.**

1178 Quantile-quantile plot for tyrosine content. The horizontal red line indicates the

1179 significant threshold (10^{-5}). Trait-associated SNPs above the significant threshold are

1180 colored in red.

1181 **Figure S46** GWAS of valine content using MLM. **A.** Density distribution of valine

1182 content. **B.** Manhattan plots for valine content. Negative $\log_{10}$ P-values from a

1183 genome-wide scan are plotted against SNP positions of 20 chromosomes. **C.**

1184   Quantile-quantile plot for valine content. The horizontal red line indicates the

1185   significant threshold (10^{-5}). Trait-associated SNPs above the significant threshold are

1186   colored in red.

1187   **Figure S47** GWAS of total amino acids content using MLM. **A.** Density distribution

1188   of total amino acids content. **B.** Manhattan plots for total amino acids content.

1189   Negative $\log_{10}$ P-values from a genome-wide scan are plotted against SNP positions

1190   of 20 chromosomes. **C.** Quantile-quantile plot for total amino acids content. The

1191   horizontal red line indicates the significant threshold (10^{-5}). Trait-associated SNPs

1192   above the significant threshold are colored in red.

1193   **Figure S48** GWAS of hundred grain weight using MLM. **A.** Density distribution of

1194   one hundred seed weight. **B.** Manhattan plots for hundred grain weight. Negative

1195 $\log_{10}$ P-values from a genome-wide scan are plotted against SNP positions of 20

1196   chromosomes. **C.** Quantile-quantile plot for hundred grain weight. The horizontal red

1197   line indicates the significant threshold (10^{-5}). Trait-associated SNPs above the

1198   significant threshold are colored in red.

1199   **Figure S49** GWAS of pod number per plant using MLM. **A.** Density distribution of

1200   pod number per plant. **B.** Manhattan plots for pod number per plant. Negative $\log_{10}$

1201   P-values from a genome-wide scan are plotted against SNP positions of 20

1202   chromosomes. **C.** Quantile-quantile plot for pod number per plant. The horizontal red

1203   line indicates the significant threshold (10^{-5}). Trait-associated SNPs above the

1204   significant threshold are colored in red.

1205   **Figure S50** GWAS of pod size using MLM. **A.** Density distribution of pod size. **B.**

1206   Manhattan plots for pod size. Negative $\log_{10}$ P-values from a genome-wide scan are

1207   plotted against SNP positions of 20 chromosomes. **C.** Quantile-quantile plot for pod

1208   size. The horizontal red line indicates the significant threshold (10^{-5}). Trait-associated

1209   SNPs above the significant threshold are colored in red.

1210   **Figure S51** GWAS of seed number per plant using MLM. **A.** Density distribution of

1211   seed number per plant. **B.** Manhattan plots for seed number per plant. Negative $\log_{10}$

1212   P-values from a genome-wide scan are plotted against SNP positions of 20

1213   chromosomes. **C.** Quantile-quantile plot for seed number per plant. The horizontal red

1214   line indicates the significant threshold (10^{-5}). Trait-associated SNPs above the

1215   significant threshold are colored in red.

1216   **Figure S52** GWAS of seed number per pod using MLM. **A.** Density distribution of

1217   seed number per pod. **B.** Manhattan plots for seed number per pod. Negative $\log_{10}$

1218   P-values from a genome-wide scan are plotted against SNP positions of 20

1219   chromosomes. **C.** Quantile-quantile plot for seed number per pod. The horizontal red

1220   line indicates the significant threshold (10^{-5}). Trait-associated SNPs above the

1221   significant threshold are colored in red.

1222   **Figure S53** GWAS of seed size using MLM. **A.** Density distribution of seed size. **B.**

1223   Manhattan plots for seed size. Negative $\log_{10}$ P-values from a genome-wide scan are

1224   plotted against SNP positions of 20 chromosomes. **C.** Quantile-quantile plot for seed

1225   size. The horizontal red line indicates the significant threshold (10^{-5}). Trait-associated

1226   SNPs above the significant threshold are colored in red.

1227   **Figure S54** GWAS of seed weight per plant using MLM. **A.** Density distribution of

1228   seed weight per plant. **B.** Manhattan plots for seed weight per plant. Negative $\log_{10}$

1229   P-values from a genome-wide scan are plotted against SNP positions of 20

1230   chromosomes. **C.** Quantile-quantile plot for seed weight per plant. The horizontal red

1231   line indicates the significant threshold (10^{-5}). Trait-associated SNPs above the

1232   significant threshold are colored in red.

1233   **Figure S55** Gene expression validation of different haplotypes/phenotypes for five

1234   candidate genes. **A.** Phenotype distribution (left) and expression level (right) of *GL3*

1235   for different haplotypes. **B.** Phenotype distribution (left) and expression level (right)

1236   of *GSTL3* for different haplotypes. **C.** Phenotype distribution (left) and expression

1237   level (right) of *GSTT1b* for different haplotypes. **D.** Phenotype distribution (left) and

1238   expression level (right) of *CKX3* for different haplotypes. **E.** Phenotype distribution

1239   (left) and expression level (right) of *CYP85A2* for different haplotypes.

1240

1241