1

2 **SoyFGB v2.0: a unique access to variations of Chinese Soybean Gene**
 3 **Bank (CNSGB) germplasm**

4

5 **Running title: SoyFGB v2.0: a soybean variation source**

6 Tianqing Zheng^{1,†}, Yinghui Li^{1,†}, Yanfei Li¹, Shengrui Zhang¹, Chunchao Wang¹, Fan
 7 Zhang¹, Lina Zhang¹, Xiangyun Wu¹, Yu Tian¹, Shan Jiang¹, Jianlong Xu¹, Lijuan Qiu¹,

8 *

9 ¹. Institute of Crop Sciences/National Key Facility for Crop Gene Resources and
 10 Genetic Improvement, Chinese Academy of Agricultural Sciences, No. 12 South
 11 Zhong-Guan-Cun Street, Beijing 100081, P.R. China

12 [†]These authors contributed equally to this work.

13 * To whom correspondence should be addressed: Lijuan Qiu, Email: qiulijuan@caas.cn

14 **Summary**

15 In Chinese National Soybean GeneBank (CNSGB), we have collected more than 30,000
 16 soybean accessions. However, data sharing for soybean remains an especially ‘sensitive’
 17 question, and how to share the genome variations within rule frame has been bothering
 18 the soybean germplasm workers for a long time.

19 Here we release a big data source named Soybean Functional Genomics &
 20 Breeding database (SoyFGB v2.0) (<https://sfgb.rmbreeding.cn/>), which embed a core
 21 collection of 2,214 soybean resequencing genome (2K-SG) from the CNSGB
 22 germplasm. This source presents a unique example which may help elucidating the
 23 following three major functions for multiple genome data mining with general interests
 24 for plant researchers. 1) On-line analysis tools are provided by the ‘Analysis’ module
 25 for haplotype mining in high-throughput genotyped germplasms with different methods.
 26 2) Variations for 2K-SG are provided in SoyFGB v2.0 by Browse module which
 27 contains two functions of ‘SNP’ and ‘InDel’. Together with the ‘Gene (SNP & InDel)’
 28 function embedded in Search module, the genotypic information of 2K-SG for targeting
 29 gene / region is accessible. 3) Scaled phenotype data of 42 traits, including 9 quality and
 30 33 quantitative traits are provided by SoyFGB v2.0. With the scaled-phenotype data
 31 search and seed request tools under a control list, the germplasm information could be
 32 shared without direct downloading the unpublished phenotypic data or information of
 33 sensitive germplasms.

34 In a word, the mode of data mining and sharing underlies SoyFGB v2.0 may inspire
 35 more ideas for works on genome resources of not only soybean but also the other plants.

36 **Key words**

37 Soybean; Variation mining; Data analysis; Data share; 2,214 soybean resequencing genome
38 (2K-SG)

39

40 **Introduction**

41 In an era of genomic information moving forward from theory to application, two
42 key barriers still exist to prevent the widespread sharing of big data: (1) Phenotype is a
43 leading factor for both breeding and genetic analysis. When handling big datasets,
44 deciding how to share germplasms with unpublished phenotypic data, especially
45 quantitative trait data, remains difficult; (2) Genotypic data typically require large
46 storage capacity and relatively infrequent access, whereas germplasm data with
47 phenotypic information require relatively little storage space but frequent access. Thus,
48 how to balance between efficiency and cost is challenging for such databases, especially
49 for researcher-maintained data sources.

50 Soybean (*Glycine max*) is one of the most important plant sources of protein and oil,
51 and a model crop for legume genome research. Worldwide genebanks such as the
52 Chinese National Soybean GeneBank (CNSGB) and the USDA-ARS Soybean
53 Germplasm Collection contain more than 170,000 soybean accessions, including
54 cultivated soybean (*G. max*) and its progenitor *G. soja*. However, in this era of
55 genomics-based breeding, the sharing of big data, especially the genome sequencing
56 data for multiple soybean accessions remains a bottleneck.

57 Phytozome (Goodstein et al., 2012) is a popular online resource for plant
58 researchers, and it makes available different versions of reference genomes. Soybase
59 (Grant et al., 2010) provides a unique source of soybean genetic information for
60 multiple soybean genomes based on chip (SoySNP50K) data. In recent years, numerous

61 studies have reported re-sequenced soybean genomes (Li et al., 2014; Liu et al., 2020b;
62 Torkamaneh et al., 2020). The MBKbase (Peng et al., 2020) is also going to release a
63 set of germplasm sequencing data (<http://www.mbkbase.org/soybean>) based on a recent
64 pan-genome report(Liu et al., 2020a), of which germplasm list for 522 accessions is
65 recently accessible through National Genomic Data Center (Li et al., 2020). LegumeIP,
66 an integrative database for comparative genomics and transcriptomics of model legumes,
67 which has recently been updated to its third version (Dai et al., 2020).

68 Even so, comparing to other crop species such as rice, access to soybean multiple
69 genome data, especially for haplotype mining with phenotypic data still remains quite
70 limited. Furthermore, soybean is always on the control list for germplasm share. Thus,
71 how to share global soybean collections, especially the core collection with both
72 phenotype and genome data, has become an urgent request by plant researchers.

73

74 **Results**

75 In SoyFGB v2.0, we included three major modules, which are Search, Browse, and
76 Analysis. The Search module contains four functions of ‘Germplasm’, ‘Phenotype’,
77 ‘Gene (SNP & InDel)’, and ‘Knowledge’. Users can select favourable germplasms by
78 phenotype scaling in ‘Phenotype’ or by target gene variations embedded in ‘Gene (SNP
79 & InDel)’ module. More information about 2K-SG or soybean was supplied by
80 ‘Germplasm’ and ‘Knowledge’ functions. With the Browse module, the SNP or InDel
81 variations were accessible in a view of genome browser embedded in ‘SNP’ and ‘InDel’
82 functions, respectively. In the Analysis module, three functions named ‘Hap-GWAS’,
83 ‘Soy_Haplotype’, and ‘Intro_Hap’ were supplied. With these tools, user may carry out a
84 deep mining for haplotypes in genotyped soybean germplasm.

85 Typical use of SoyFGB v2.0 are demonstrated with followed user cases as example:

86 1) Haplotype mining with embedded/user-owned 2K-SG phenotypic data.

87 In the ‘Soy_Haplotype’ function, user may define a target region with gene name,
88 physical range or a set of SNPs, to mine the possible haplotype variations from the 2K-
89 SG. With the embedded or user input phenotypic data, the mean values of target traits
90 for different haplotypes are available. The possible donor lists are provided with a
91 straight-forward statistical analysis based on ANOVA protected t-test as reference for
92 users.

93 The candidate genes for isoflavone content in soybean were identified by a joint
94 work of bulk segregant analysis (BSA) with a natural population and weighted gene co-
95 expression network analysis (WGCNA) using the transcriptome of different seed
96 development stages. SoyFGB v2.0 provided haplotype analysis function for the
97 candidate genes. Firstly, locus number of one candidate gene ID and the phenotypic
98 data of isoflavone content of 2K-SG from user were submitted to the ‘Soy_Haplotype’
99 function embedded in Analysis module of SoyFGB v2.0. Then, all the haplotypes of
100 this gene were presented. Subsequently, With the aid of straight-forward statistical
101 analysis between different haplotypes, germplasms harbouring different haplotypes
102 were found to be significantly distinct from each other in isoflavone content. This
103 implies the possible contribution of the candidate gene in regulating isoflavone content
104 of soybean grain. Finally, the haplotype variations including and the germplasm list for
105 the candidate gene were also downloaded for further lab works (Figure 1).

106 2) Finding candidate gene / region with Hap-GWAS function.

107 As shown in Figure 2, an enhanced correlation between the phenotypes and haplotypes
108 would be mined with Hap-GWAS function, which adopted the methodology raised
109 recently (Zhang et al., 2021). In order to save the possible waiting time for this analysis,
110 an email-remind system was adopted. Once the analysis results were ready, a remind-
111 email with a direct access link to the output would be sent to the mailbox defined by
112 user. Together with the instant screening with the previous ‘Soy_Haplotype’ function,

113 correlations between the phenotype and haplotypes may be fully mined at different
114 depths.

115 3) Intro_Hap function to mine haplotypes with SNP chip data.

116 Considering more and more data sets accumulated by relatively lower density
117 genotyping methods, e.g. SNP chip, an analysis tool for haplotype in target region using
118 this type of data were also provided with 'Intro_Hap' function for both populations
119 with/without known parents. Since the request for this module is recently raised by
120 users during out indoor testing period, it is still looking forward to more users'
121 responses within a 6-month open period since this release of SoyFGB v2.0.

122 4) Exploring germplasms based on scaled-phenotype or accession information.

123 A typical pre-breeding / forward genetics scheme starts with phenotyping. In SoyFGB
124 v2.0, a set of scaled data covering 42 traits, including 9 quality and 33 quantitative traits
125 (maturity time, 100-seed weight, height, protein content, oil content, virus index, virus
126 level, average SCN amount 1, SCN infection level 1, average SCN amount 2, SCN
127 infection level 2, average SCN amount 3, SCN infection level 3, average SCN amount 4,
128 SCN infection level 4, average SCN amount 5, SCN infection level 5, cystine acid
129 content, methionine acid content, stearic acid content, palmitic acid content, oleic acid
130 content, linoleic acid content, linolenic acid content, salt tolerance during germination,
131 salt tolerance of seedlings, sprouting drought tolerance, drought tolerance of adult plants,
132 cold tolerance during germination, cold tolerance of seedlings, resistance to soybean
133 rust, type of reaction to soybean rust, soybean rust infection level) were embedded in
134 Search based on 'Phenotype' function. The user can screen 2K-SG germplasm with
135 scaled phenotypic data. A three-step route can be followed for exploring elite donors for
136 a breeding scheme: (a) target trait scaling, demonstrated herein by screening the top 30%
137 protein content as an example, which includes 13 samples; (b) from these samples,
138 favourable early mature (top 50%) samples were further screened, and favourable
139 samples may be added to create a list of candidate germplasms (three samples); (c) the

140 user can then export a list of candidate donors for different breeding schemes based on
141 the two grouping levels. An easy way to ‘Seed Request’ function is available through
142 just one click on a key named ‘Request Germplasm’.

143 5) Searching for variation within candidate genes for favourable germplasm.
144 A route for shortlisting candidate genes using SoyFGB v2.0 is shown in Figure 3, and
145 involves the following: (a) get a target region by mapping methods such as GWAS or
146 sorting accessions with favourable target traits, (b) using ‘SNP’ or ‘InDel’ functions in
147 Browse module, variations within target gene / region can be explored, (c) inputting
148 target region or gene locus ID into the ‘Gene(SNP/InDel)’ function of Search module,
149 (d) with the genotype information (SNP or InDel) downloaded, users can carry out
150 further work with primer design and wet-lab confirm.

151 Discussion

152 SoyFGB v2.0 is accessible through the following URL: <https://sfgb.rmbreeding.cn/>.
153 It is designed to be adaptive and responsive to the overwhelming quantity of genomic
154 data and phenotypic data resulting from functional genomic breeding materials,
155 including 2K-SG data. The FGB data sharing mode in SoyFGB v2.0 has characteristics
156 as followings:

157 (i) Instead of providing a direct download link to raw phenotypic data, a scaled-
158 based phenotypic data-led germplasm sharing mode is employed by SoyFGB v2.0. On
159 the other hand, the correlations between phenotypic and genotypic data, such as GWAS
160 results are commonly provided by search function of other websites (Li et al., 2020;
161 Zhao et al., 2021). SoyFGB v2.0 make an attempt to the online analysis with ‘Hap-
162 GWAS’, ‘Soy_Haplotype’, and ‘Intro_Hap’. Mining elite donors with favourable
163 haplotype are of high value for breeders. This has provided a model which is more
164 conducive to data contributors sharing their own unpublished data with public users.

(ii) To keep up with the development of multiple-client ends, a development framework different from previous FGB website (Wang et al., 2020) has been employed in SoyFGB v2.0. The front end was developed by the Vue-Element-Axios tool, and the back end of was developed by using the java-based Spring Boot tool. The website is driven by Nginx. The RESTful API facilitates easy data access through different client platforms. Additionally, in order to balance efficiency and cost, a distributed database structure was designed for SoyFGB v2.0 (Figure 4). Phenotypic and genotypic data are stored in servers with different capacities. Phenotypic data are managed by an instant-response server with a relatively small storage capacity using the MySQL database. Genotypic data are stored in a server with a slower response but a larger storage capacity. All these attempts are going to meet the possible developing trends including decentralization and multiple accessing ways to biological data in near future.

(iii) Searching plant omics databases for functional information has grown in popularity (Gui et al., 2020). Accordingly, in SoyFGB v2.0 the Search function is important for helping users to search for useful information inside and/or outside SoyFGB. With all three major modules embedded, users can access 2K-SG data in an effective and efficient manner.

In summary, SoyFGB v2.0 provides a unique example of platform for sharing big datasets (both phenotypic, genotypic, and mining data) from of multiple soybean sequenced accessions. With the development of SoyFGB, the FGB has now evolved into a phenotype-led re-sequencing data sharing mode. This may inspire new ideas for mining complex traits in soybean and other plants. With more and more 2K-SG data published by users who query and use this resources, SoyFGB v2.0 will continue to acquire more and more data, especially phenotypic data, that will be available for users. In future updates, we will focus on data integrating with more dimensions (for both genotypic and phenotypic data) to collate variations in 2K-SG germplasm data from different groups in China and across the world. With progress in genotypic data analysis

192 and phenotypic data collection from re-sequenced genomes, more open links between
193 these datasets is hopefully going to be constructed in progressing versions of SoyFGB.
194 Further development of the FGB and other databases will facilitate a unique access to
195 soybean and other plant collections in genebanks.

196

197 **Materials and Methods**

198 As shown in Figure 5, SoyFGB v2.0 includes 2,214 accessions (2K-SG) from four
199 major soybean production and distribution areas (Asia, America, Europe and Africa)
200 based on the core collection strategy of the CNSGB. The 2K-SG dataset comprises four
201 major classes of soybean species; cultivated species (1,993 *G. max*), annual wild species
202 (218 *G. soja*), perennial wild species (2 *G. tomentella*), and others (1 *G. tabacine*).
203 Among them, *G. tomentella* and *G. tabacine* are the only two perennial wild species
204 found in China. Of the 218 *G. soja* accessions, 99.5% were collected from native
205 sources (East Asia). This includes China (179), Korea (10), Japan (19) and Russia (9),
206 providing broad species diversity. Among the 1,993 *G. max* accessions, more than half
207 (56.7%) are landraces primarily collected from China and applied to core collections to
208 represent the diversity of the 23,587 cultivated soybean accessions preserved in the
209 CNSGB. The 862 improved cultivars were collected from 17 countries, especially major
210 soybean producing countries including the USA and China.

211 Whole-genome resequencing was carried out according to a standard procedure.
212 Specifically, a genomic library was constructed using a TruseqNano DNA HT sample
213 preparation Kit (Illumina, San Diego, California, USA), purified using an AMPure XP
214 bead system (Beckman Coulter, Brea, California, USA), and the size distribution was
215 analysed by an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara,
216 California, USA) and quantified using real-time PCR. An Illumina Hiseq X platform

217 was then employed to generate ~10.58 Tb of raw sequences with a read length of 150
218 bp. Removal of low-quality paired reads resulted in 16.41 Tb of high-quality paired-end
219 reads, including 96.05% with phred quality $\geq$ Q20 and 90.98% with phred quality $\geq$ Q30.

220 The obtained fastq files were submitted to a pipeline composed by BWA (v.
221 0.7.17-r1188), SAMtools (v.1.39), Sambamba (v.0.6.8), picard (v.2.18.15,
222 <http://broadinstitute.github.io/picard>), and GATK (version v4.1.2.0) and screened
223 against Williams 82 assembly V2.0 (<http://www.phytozome.net/soybean>). This yielded
224 65,374,688 single-nucleotide polymorphisms (SNPs; 60,153,828 bi-allelic) and
225 10,952,749 InDels (8,349,613 small insertions and deletions <15 bp and <50% missing).
226 Using a standard SNP screening procedure, 8,785,134 highly-credible biallelic SNPs
227 were obtained.

228 Based on this set of SNPs, two different levels of grouping were carried out and
229 presented in SoyFGB v2.0. Level one (Group 1), the SNP-only level, includes 1507
230 cultivated, 313 wild, and 394 admixture accessions. Level two (Group 2), based on the
231 output of a two-step grouping, includes a CGCC-based subgrouping and an SNP-based
232 subgrouping within each group. In Group 2, the cultivar group was divided into five
233 subgroups; the Chinese Southern Region (C_SR), the Chinese central region
234 surrounding the middle area downstream of the Yellow River valley (C_CR), the
235 Chinese northern region plus Japan, the Korean peninsula and the Russian far east
236 region (C_NR), America (C_Am), and admixture (C_AD) subgroups. The wild group
237 was divided into four subgroups; the Chinese southern region (W_SR), the Chinese
238 central region surrounding the middle area downstream of the Yellow River valley
239 (W_CR), the Chinese northern region plus Japan, the Korean peninsula and the Russian
240 far east region (W_NR), and admixture (W_AD) subgroups.

241 It's known that, InDel and SNP tend to be gathering throughout the genome
242 (Montgomery et al., 2013). Thus, in present release of haplotype analysis, SNP is still
243 the main points, the InDel presented by “-” was also taken into consideration during the

analysis. Additionally, heterozygotes were regarded as one type. In the future versions, analysis with models considering more variations including InDel and more reference genomes would be taken into accounts. In ‘Soy_Haplotype’ function, a straight-forward statistical analysis based on ANOVA protected t-test is provided. In the ‘Hap_GWAS’ function, a linear model for GWAS (Zhang et al., 2021) was adopted.

Phenotypic data were obtained from CNSGB accumulated data. Quantitative data were then transformed into scaled form.

Author Contributions

QLJ, YHL and TQZ conceived the study and drafted the manuscript. YFL, SRZ, YHL, YT, SJ and TQZ contributed to data sharing. CCW and FZ contributed to raw code. LNZ and XYW contributed to database maintenance. JLX contributed to writing and revision.

Acknowledgements

This work was partially supported by the National Key Research and Development Plan (2016YFD0100201) from MOST, the National Natural Science Foundation of China (31871715), the Central Public-interest Scientific Institution Basal Research Fund (Y2020PT24) and the Agricultural Science and Technology Innovation Program (CAASTIPS, Y2020YJ09) from the Chinese Academy of Agricultural Sciences, the Phenomics project from the Institute of Crop Sciences (ICS2020YJ07BX), and the ‘Green Super Rice’ project from the Bill & Melinda Gates Foundation (OPP1130530).

266 **Declaration of Interests**

267 Authors declare no conflict of interests.

268 **Figures Legends**

269 **Figure 1. User case of ‘Soy_Haplotype’ function embedded Analysis module of**
270 **SoyFGB v2.0.**

271 **Figure 2. Find target gene / region with ‘Hap-GWAS’ function embedded in**
272 **Analysis module of SoyFGB v2.0.**

273 **Figure 3. User case of variation mining for flowering time with Search module of**
274 **SoyFGB v2.0.**

275 **Figure 4. Database structure of SoyFGB v2.0.**

276 **Figure 5. The 2,214 sequenced soybean genome (2K-SG) germplasm embedded in**
277 **SoyFGB v2.0.**

278

279

280

281

282 References

- 283 Dai, X., Zhuang, Z., Boschiero, C., Dong, Y. and Zhao, P.X. (2020) LegumeIP V3: from
284 models to crops-an integrative gene discovery platform for translational genomics in
285 legumes. *Nucleic Acids Res.*
- 286 Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks,
287 W., Hellsten, U., Putnam, N. and Rokhsar, D.S. (2012) Phytozome: a comparative
288 platform for green plant genomics. *Nucleic Acids Res* **40**, D1178-1186.
- 289 Grant, D., Nelson, R.T., Cannon, S.B. and Shoemaker, R.C. (2010) SoyBase, the USDA-ARS
290 soybean genetics and genomics database. *Nucleic Acids Res* **38**, D843-846.
- 291 Gui, S., Yang, L., Li, J., Luo, J., Xu, X., Yuan, J., Chen, L., Li, W., Yang, X., Wu, S., Li, S.,
292 Wang, Y., Zhu, Y., Gao, Q., Yang, N. and Yan, J. (2020) ZEAMAP, a comprehensive
293 database adapted to the maize multi-omics era. *iScience* **23**, 101241.
- 294 Li, C., Tian, D., Tang, B., Liu, X., Teng, X., Zhao, W., Zhang, Z. and Song, S. (2020) Genome
295 Variation Map: a worldwide collection of genome variations across multiple species.
296 *Nucleic Acids Res* **49**, D1186-D1191.
- 297 Li, Y.H., Zhou, G., Ma, J., Jiang, W., Jin, L.G., Zhang, Z., Guo, Y., Zhang, J., Sui, Y., Zheng,
298 L., Zhang, S.S., Zuo, Q., Shi, X.H., Li, Y.F., Zhang, W.K., Hu, Y., Kong, G., Hong,
299 H.L., Tan, B., Song, J., Liu, Z.X., Wang, Y., Ruan, H., Yeung, C.K., Liu, J., Wang, H.,
300 Zhang, L.J., Guan, R.X., Wang, K.J., Li, W.B., Chen, S.Y., Chang, R.Z., Jiang, Z.,
301 Jackson, S.A., Li, R. and Qiu, L.J. (2014) De novo assembly of soybean wild relatives
302 for pan-genome analysis of diversity and agronomic traits. *Nat Biotechnol* **32**, 1045-
303 1052.
- 304 Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.-A., Zhang, H., Liu, Z., Shi, M.,
305 Huang, X., Li, Y., Zhang, M., Wang, Z., Zhu, B., Han, B., Liang, C. and Tian, Z.
306 (2020a) Pan-Genome of Wild and Cultivated Soybeans. *Cell* **182**, 162-176.e113.
- 307 Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.A., Zhang, H., Liu, Z., Shi, M.,
308 Huang, X., Li, Y., Zhang, M., Wang, Z., Zhu, B., Han, B., Liang, C. and Tian, Z.
309 (2020b) Pan-genome of wild and cultivated soybeans. *Cell* **182**, 162-176 e113.
- 310 Montgomery, S.B., Goode, D.L., Kvikstad, E., Albers, C.A., Zhang, Z.D., Mu, X.J., Ananda, G.,
311 Howie, B., Karczewski, K.J., Smith, K.S., Anaya, V., Richardson, R., Davis, J.,
312 Genomes Project, C., MacArthur, D.G., Sidow, A., Duret, L., Gerstein, M., Makova,
313 K.D., Marchini, J., McVean, G. and Lunter, G. (2013) The origin, evolution, and
314 functional impact of short insertion-deletion variants identified in 179 human genomes.
315 *Genome research* **23**, 749-761.
- 316 Peng, H., Wang, K., Chen, Z., Cao, Y., Gao, Q., Li, Y., Li, X., Lu, H., Du, H., Lu, M., Yang, X.
317 and Liang, C. (2020) MBKbase for rice: an integrated omics knowledgebase for
318 molecular breeding in rice. *Nucleic Acids Res* **48**, D1085-D1092.

319 Torkamaneh, D., Laroche, J., Valliyodan, B., O'Donoghue, L., Cober, E., Rajcan, I., Vilela
320 Abdelnoor, R., Sreedasyam, A., Schmutz, J., Nguyen, H.T. and Belzile, F. (2020)
321 Soybean (*Glycine max*) Haplotype Map (GmHapMap): a universal resource for soybean
322 translational and functional genomics. *Plant Biotechnol J*.
323 Wang, C.C., Yu, H., Huang, J., Wang, W.S., Faruquee, M., Zhang, F., Zhao, X.Q., Fu, B.Y.,
324 Chen, K., Zhang, H.L., Tai, S.S., Wei, C., McNally, K.L., Alexandrov, N., Gao, X.Y.,
325 Li, J., Li, Z.K., Xu, J.L. and Zheng, T.Q. (2020) Towards a deeper haplotype mining of
326 complex traits in rice with RFGB v2.0. *Plant Biotechnol J* **18**, 14-16.
327 Zhang, F., Wang, C., Li, M., Cui, Y., Shi, Y., Wu, Z., Hu, Z., Wang, W., Xu, J. and Li, Z. (2021)
328 The landscape of gene–CDS–haplotype diversity in rice: Properties, population
329 organization, footprints of domestication and breeding, and implications for genetic
330 improvement. *Molecular Plant* **14**, 787-804.
331 Zhao, H., Li, J., Yang, L., Qin, G., Xia, C., Xu, X., Su, Y., Liu, Y., Ming, L., Chen, L.-L.,
332 Xiong, L. and Xie, W. (2021) An inferred functional impact map of genetic variants in
333 rice. *Molecular Plant* **14**, 1584-1599.

334

335

Figure 1, Analysis – Soy_Haplotype - Isoflavone

Figure 2, Analysis – Hap-GWAS

Figure 3, SNP & InDel Marker development

Figure 4,

I: Private domain with larger storage but slower response;

II: Public domain with faster response but smaller storage.

Figure 5,