**Framework of a novel fine mapping strategy from the view of epistasis in large-scale GWAS panel**

Yanchen Zhou1†, Yajun Xiong1†, Jian Song2†, Zhan Li1, Tingxuan Liu1, Zhihao Zhang3, Yongzhe Gu3, Xiaoxue Liu1, Fan Zhang2, Jun Wang1\*, Yinghui Li3\*, Lijuan Qiu3\*

*1 College of Agriculture, Yangtze University, Jingzhou, 434025, Hubei, P.R. China;*

*2 College of Life Sciences, Yangtze University, Jingzhou, 434025, Hubei, P.R. China;*

*3 National Key Facility for Gene Resources and Genetic Improvement (NFCRI)/Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, P.R.China*

*4 Department of Computer Science, Yangtze University, Jingzhou, 434025, Hubei, P.R. China;*

**\***Corresponding author: Jun Wang, wangjagri@yangtzeu.edu.cn; Yinghui Li, liyinghui@caas.cn; Lijuan Qiu, qiulijuan@caas.cn; †These authors contributed equally to this work.

# Novelty statement

We proposed epistasis based fine mapping (EBFP) strategy, and provided a general framework describing the basic concepts concerning how to perform EBFP analysis. It is the first attempt to fine map association region of GWAS from the view of epistasis. The EBFP strategy for fine mapping would be of great interest to wide readership to whom working on genetics of complex traits using GWAS, no matter in animals, plants or microbes.

**Abstract**

Current statistical fine mapping strategies inferred association regions mainly via significance test of each SNP separately or SNPs in combination within a single-handed region under GWAS scenario, epistasis between loci has not been fully explored up to date. In this study we proposed a novel pipeline termed as Epistasis Based Fine Mapping (EBFP) in large GWAS panel. In EBFP strategy, impact of each SNP in epistatic region on each SNP in hypostatic region are evaluated in terms of corrected P value and Epistatic Effect Index (EEI). The application of EBFP strategy in soybean leaf traits (leaf length to width ratio, LLWR) efficiently narrowed down association regions to almost 1/10 of the preliminary result, and a novel genetic locus was identified as well. Specifically, Ln locus, which has already cloned and functionally characterized previously, showed the highest EEI value, suggested that the EBFP strategy was reliable. However, EBFP strategy is only suitable for GWAS analysis with large scale sample size with dense-genotyping for the sake of high statistical power. With the decreasing trend of sequencing price, wider application of EBFP could be expected now and in the future.

**Keywords:** Genome-wide association study, Epistasis, Fine mapping

Based on omnigenic model, only partial phenotypic variation is contributed by a given genetic locus alone for complex traits, and the epistasis responding to the “missing heritability” is commonly existed (Boyle et al., 2017; McKinney and Pajewski, 2011). In the post-GWAS era, statistical fine mapping strategies inferred association regions mainly via significance test of each SNP within a single-handed region, e.g., triaging variants based on *P*-values or linkage disequilibrium (LD) to the lead SNP, heuristic LD approach, penalized regression, Bayesian and trans-ethnic fine mapping, etc. (Schaid et al., 2018). However, the epistasis between genetic loci has not been fully explored in the fine mapping up to date.

In many cases, the causal variant of GWAS could explain phenotypic variance most (or close to) significantly compared to that of non-causal variants if dense-genotyping was applied in large scale GWAS panel with high statistical power (Schaid et al., 2018), the farther the non-causal variants to the causal variant was the lower *P* value would be due to LD, and so does it in the case that epistatic effect on the causal variant. Based on these premises, we proposed a novel strategy, hereafter termed as *Epistasis Based Fine Mapping* (EBFM), for both credible association region narrowing down and novel genetic loci identification.

To begin with EBFM, we assume that there are two significant regions (hereafter designated as *region 1* and *region 2*) associated to a given trait and *region 1* (epistatic) shows epistatic effect on *region 2* (hypostatic), and each region contains *m* and *n* significant SNPs respectively. To fine map *region 2*, each significant SNP within *region 1* is firstly fixed by *AA* and *aa* separately with heterozygous (*Aa*) omitted. By performing so the epistatic effect of *region 1* will be eliminated in either AA or aa genotyped sample panel. Then subset sample panels for each genotype (*AA* and *aa*) are used for association analysis on *region 2* by EMMAX (Efficient Mixed-Model Association eXpedited) (Kang et al., 2010). After that two *P* value *matrixes*, hereafter designated as *Ph1* (*Ph*, *P* value of hypostatic region, AA) and *Ph2* (aa)respectively, would be obtained. As we do not know which genotype (AA or aa?) shows profounder epistatic effect on *region 2*, the *matrix* *Pa* is then generated by , to ensure that each *P* value reflects maximum impact of *region 1* on *region 2*. As each SNP within *region 1* is independent with each other, different sample panels generated by each SNP genotype fixed are actually samples with different population structures, which could be regarded as different biological replicates to a broader sense. Then mean (corrected) *P* value of each SNP of *region 2 in Pa matrix* (each row) wascalculated to represent the real significance level (Figure 1a). The significant threshold was determined by a stringent modified Bonferroni correction of 0.001/(SNP number of *region 1* × SNP number of *region 2*).

To fine map *region 1*, *Epistatic Effect Index* (*EEI*) were proposed to portray the epistatic effect impact of *region 1* on *region 2*, which is defined by the formula of

Where *Pej* represents the *P* value for the association between trait and each significant SNP of *region 1* with full sample panel, represents the lower *P* value of *matrixes Ph1* and *Ph2* as described above. The higher the *EEI* indicates higher impact of epistatic SNP on the hypostatic SNP.

To evaluate to what extent a SNP could be a lead SNP (peak on Manhattan plot), the fitness coefficient of each significant SNP is calculated. Firstly, the LD coefficient (*r2*) is set to constant within limited region to simplify the situation, *Si* and *Pi* denote the physical location, and significant–log10(P) value respectively, the relation between *Pi* and *Si* could be defined as:

When , the formula could be transformed into if the *Pi* value was overturned vertically (Figure 1b), then a linear regression could be defined between *P* and *S* (*P* value and physical location distance between *SNPi* and *SNPcausal* respectively). The fitness to what extent a significant SNP could be a peak can then be described as the correlation coefficient (R) between and . Theoretically, the higher positive R value indicates higher possibility of a given SNP to be the lead SNP, which could be adopted as a reference index for causal variant judgement.

To better illustrate EBFM strategy, leaf traits of soybean was taken as an example. Dataset employed in this study including 1061 soybean accessions collected across China, leaf length to width ratio (LLWR) was phenotyped and GWAS analysis revealed that two associated regions on chromosome 19 and 20 were identified, hereafter designated as *qLLWR19-1* and *qLLWR20-1* respectively (detailed GWAS analysis of all soybean leaf traits will be displayed in another article). The imputated SNP density of these tworegions were 4.4 and 8.2 SNP/kb respectively. *qLLWR20-1* (Chr20:34,752,555..37,061,502) was overlapped with the *Ln* mapping region (Fang et al., 2013; Jeong et al., 2011; Jeong et al., 2012), *qLLWR19-1* (Chr19:44,426,228..45,399,939) embraced the region (Chr19:45,143,539..45,150,769) responsible for leaf area, and *Ln* showed epistatic effect on *qLLWR19-1* by SNP-fixing method (Fang et al., 2017).

To fine map these two regions, *qLLWR20-1* and *qLLWR19-1* were designated as *region 1* and *region 2* respectively.There were 592 and 700 significant SNPs identified (by EMMAX) for *qLLWR20-1* and *qLLWR19-1*, defining the association regions of 2.3Mb and 973.7kb respectively (Figure 1c and d). By EBFM analysis, the average subset sample panel size was 806, and two apparent peaks with distance of 331.97kb within *qLLWR19-1* were identified using a threshold of 8.62 (-log10(592×700)) (Figure 1e, g). Both of these two peaks showed fitness coefficient of larger than 0.7 (Figure 1h), indicated high possibility to be two distinct QTLs (*qLLWR19-1.1* and *qLLWR19-1.2* respectively). Taken both corrected *P* value and fitness into consideration, *qLLWR19-1.1* and *qLLWR19-1.2* were defined as 10.7kb and 51.6kb respectively considering the LD block size, which reduced the mapping region from 973.7kb to 62.3kb (sum of *qLLWR19-1.1* and *qLLWR19-1.2* regions). Notably, *qLLWR19-1.2* (Chr19:45,132,093..45,183,701) was in agreement with the region identified by Fang et al. (2017), but *qLLWR19-1.1* was obviously a distinct novel genetic locus.

For the *qLLWR20-1* region, mean *EEI* of each SNP showed roughly normal distribution, and a physical span of 276.7kb (Chr20:35,819,645..36,096,416) was defined by fitness coefficient larger than 0.7 (Figure 1f, i, and j). The *EEI* of *qLLWR20-1* peaked within the second exon of the *Ln* causal gene (*Glyma.20G116200*, Chr20:35,828,042) (R=0.86) (Figure 1j, k). This locus had been reported previously (Jeong et al., 2012; Sayama et al., 2017). The transversion of G (*Ln*: broad leaf) to C (*ln*: narrow leaf) on this locus changed Asp into His at amino acid level, and then increased LLWR in *ln* significantly (Figure 1k). Comparison of results before and after EBFM analysis demonstrated that association significance of *qLLWR19-1.1* increased a lot, however that of *qLLWR19-1.2* decreased to limited extent when *Ln* (Chr20:35,828,042) was fixed (GG) (Figure 1l), suggested that contribution of *qLLWR19-1.1* to *LLWR* might be concealed by *ln.* This was then confirmed by interaction analysis between *qLLWR20-1* and *qLLWR19-1.1* using Two-way ANOVA (Figure 1m, P < 0.001), however no interaction was observed between *qLLWR20-1* and *qLLWR19-1.2* (Figure 1m, P > 0.05).

Based on gene models annotated in *Glycine max* Wm82.a2.v1 (<https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Gmax>) (Schmutz et al., 2010), there was only 1 gene (*Glyma.19G190300*) located within *qLLWR19-1.1* (10.7kb), which encoded a protein with unknown function. For this unknown protein, there were two domains, namely DUF3475 and DUF668, were identified by InterProScan (Jones et al., 2014). These two domains were reported to be existed in PSI as well, which was functionally characterized to regulate growth, including the leaf area, in *Arabidopsis thaliana* (Stührwohldt et al., 2014), strongly implied that *Glyma.19G190300* would be possibly one of the causal genes of LLWR. Further analysis revealed that the lead SNP (Chr19:44,797,369) within *qLLWR19-1.1* region was located within promoter (-315) of *Glyma.19G190300*, and this substitution site was presumably the AT-hook, and as well as TBP (TATA Binding Protein) and SEF1 (Soybean Embryo Factor 1) binding site (featured with taTTTATg and ATATTtatg respectively) predicted by PlantPAN 2.0 (<http://PlantPAN2.itps.ncku.edu.tw>) (Chow et al., 2016) (Figure 1n). When *Ln* (GG) was fixed, the average LLWR of samples with TT (Chr19:44,797,369) was significantly higher than that of sample with genotype of GG (Student’s *t*-test, P < 1×10-40), whereas in the *ln* (CC) subsample panel, no significant difference was observed in terms of LLWR between TT and GG (Figure 1n) (Student’s *t*-test, P > 0.05). Taken all those findings together, *Glyma.19G190300* was highly persuadable to be regulated by *Ln* locus, and confirmation of the role of this gene in leaf trait is under investigation.

In this study, EBFM pipeline provided genetic information from the view of epistasis, and simultaneously reduced two associated regions for leaf traits in soybean to a great extent (~1/10), of which *Ln* has been previously confirmed and candidate gene of a novel genetic locus was analyzed as well, indicated a high reliability of EBFM strategy in fine mapping. With the trend of decreased sequencing price, large-scale GWAS panel has becoming easier to access, and wider application of EBFM could be expected.

**Acknowledgements**

This work was supported by grants of National Key R&D Program of China (Grant No.: 2016YFD0100304-6; 2016YFD0100201-14), the Open Fund of Engineering Research Center of Ecology and Agricultural Use f Wetland, Ministry of Education (Grant KF201808).

**Author contributions**

J.W. and Y.L., and L.Q. conceived the strategy and supervised the project. Y.Z., J.S., Y. X., Z.L., T.L., Z.Z., Y.G., X.L., and F.Z. performed the experiments. Y.Z. and J.S. analyzed the data and R script coding, J. W. wrote the manuscript.

**Conflicts of interest**

The authors declare no competing interests.

**Reference**

Boyle, E.A., Li, Y.I. and Pritchard, J.K. (2017) An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177-1186.

Chow, C.N., Zheng, H.Q., Wu, N.Y., Chien, C.H., Huang, H.D., Lee, T.Y., Chiang-Hsieh, Y.F., Hou, P.F., Yang, T.Y. and Chang, W.C. (2016) PlantPAN 2.0: an update of plant promoter analysis navigator for reconstructing transcriptional regulatory networks in plants. *Nucleic Acids Res* **44**, D1154-1160.

Fang, C., Li, W.Y., Li, G.Q., Wang, Z., Zhou, Z.K., Ma, Y.M., Shen, Y.T., Li, C.C., Wu, Y.S., Zhu, B.G., Yang, W.C. and Tian, Z.X. (2013) Cloning of Ln Gene Through Combined Approach of Map-based Cloning and Association Study in Soybean. *Journal of Genetics and Genomics* **40**, 93-96.

Fang, C., Ma, Y.M., Wu, S.W., Liu, Z., Wang, Z., Yang, R., Hu, G.H., Zhou, Z.K., Yu, H., Zhang, M., Pan, Y., Zhou, G.A., Ren, H.X., Du, W.G., Yan, H.R., Wang, Y.P., Han, D.Z., Shen, Y.T., Liu, S.L., Liu, T.F., Zhang, J.X., Qin, H., Yuan, J., Yuan, X.H., Kong, F.J., Liu, B.H., Li, J.Y., Zhang, Z.W., Wang, G.D., Zhu, B.G. and Tian, Z.X. (2017) Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. *Genome Biology* **18**.

He, F., Ding, S., Wang, H. and Qin, F. (2020) IntAssoPlot: An R Package for Integrated Visualization of Genome-Wide Association Study Results With Gene Structure and Linkage Disequilibrium Matrix. *Front Genet* **11**, 260.

Jeong, N., Moon, J.K., Kim, H.S., Kim, C.G. and Jeong, S.C. (2011) Fine genetic mapping of the genomic region controlling leaflet shape and number of seeds per pod in the soybean. *Theor Appl Genet* **122**, 865-874.

Jeong, N., Suh, S.J., Kim, M.H., Lee, S., Moon, J.K., Kim, H.S. and Jeong, S.C. (2012) Ln is a key regulator of leaflet shape and number of seeds per pod in soybean. *Plant Cell* **24**, 4807-4818.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A.F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.Y., Lopez, R. and Hunter, S. (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240.

Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.-y., Freimer, N.B., Sabatti, C. and Eskin, E. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* **42**, 348-354.

McKinney, B.A. and Pajewski, N.M. (2011) Six Degrees of Epistasis: Statistical Network Models for GWAS. *Front Genet* **2**, 109.

Sayama, T., Tanabata, T., Saruta, M., Yamada, T., Anai, T., Kaga, A. and Ishimoto, M. (2017) Confirmation of the pleiotropic control of leaflet shape and number of seeds per pod by the Ln gene in induced soybean mutants. *Breed Sci* **67**, 363-369.

Schaid, D.J., Chen, W. and Larson, N.B. (2018) From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet* **19**, 491-504.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., Xu, D., Hellsten, U., May, G.D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M.K., Sandhu, D., Valliyodan, B., Lindquist, E., Peto, M., Grant, D., Shu, S., Goodstein, D., Barry, K., Futrell-Griggs, M., Abernathy, B., Du, J., Tian, Z., Zhu, L., Gill, N., Joshi, T., Libault, M., Sethuraman, A., Zhang, X.-C., Shinozaki, K., Nguyen, H.T., Wing, R.A., Cregan, P., Specht, J., Grimwood, J., Rokhsar, D., Stacey, G., Shoemaker, R.C. and Jackson, S.A. (2010) Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178-183.

Stührwohldt, N., Hartmann, J., Dahlke, R.I., Oecking, C. and Sauter, M. (2014) The PSI family of nuclear proteins is required for growth in arabidopsis. *Plant Molecular Biology* **86**, 289-302.

**Figure Legend**

Figure 1. EBFM analysis on soybean leaf length to width ratio. (a) Flow chart of EBFM strategy. (b) Rationale of fitness coefficient calculation. (c)-(d) Manhattan plot of *qLLWR19-1* and *qLLWR20-1* respectively, red, blue, and black dash line indicate threshold of -log10(0.1, 0.2, 0.5/SNP number) respectively. (e)-(f) Heatmap of *Pa matrix* and *EEI matrix* (transformed by -log10)obtained in EBFM analysis. (g) Manhattan plot of corrected *P* value of each significant SNPs within *qLLWR19-1*. Two peaks in red indicating *qLLWR19-1.1* and *qLLWR19-1.2*. (h) Fitness coefficient of significant SNPs within *qLLWR19-1*. (i) *EEI* of significant SNPs within *qLLWR20-1*, *Ln* is indicated in red. (j) Fitness coefficient scatter plot of *EEI* of significant SNPs within *qLLWR20-1*, *Ln* is indicated by red. (k) Regional display of -log10(P), LD heatmap, gene structure of *Ln* and flanking region (He et al., 2020). (l) Comparison of association significance of each SNP within *qLLWR19-1* with *Ln* fixed (GG). (m) Interaction analysis between *qLLWR20-1* and *qLLWR19-1* using two-way ANOVA. (n) Regional display of -log10(P), LD heatmap, gene structure of *Glyma.19G190300* and flanking region.

