RIL-StEp: epistasis analysis of recombinant inbred lines (RILs) reveals candidate interacting genes that control rice seed hull color

Toshiyuki Sakai[1,2]*, Akira Abe[3], Motoki Shimizu[3], Ryohei Terauchi[1,3]*

[1]Laboratory of Crop Evolution, Graduate School of Agriculture, Kyoto University, Kyoto, Japan

[2]The Sainsbury Laboratory, University of East Anglia, Norwich Research Park, NR4 7UH, Norwich, UK

[3]Iwate Biotechnology Research Center, Kitakami, Japan

*Corresponding author: toshiyuki.sakai@tsl.ac.uk, terauchi@ibrc.or.jp

Summary

Studying epistatic gene interactions is important in understanding genetic architecture of complex traits in organisms. However, due to an enormous number of gene combinations to be analyzed, detection of epistatic gene-gene interactions has been computationally demanding. Here, we show a simple approach RIL-StEp, specialized to Recombinant Inbred Lines (RILs), to study epistasis using single nucleotide polymorphisms (SNPs) information of the genome. We applied the method to reveal epistasis affecting rice seed hull color phenotype, and successfully identified gene pairs that presumably control seed hull color. This method has a potential to enhancing our understanding of genetic architecture of various traits.

Introduction

Understanding the links between genes and phenotypes of organisms is one of the most important subjects in biology. Non-additive gene interactions is called epistasis (Fisher, 1919; Phillips, 2008), and is important for crop improvement through cross breeding (Cordell, 2002; Carlborg and Haley, 2004; Xu and Crouch, 2008; Heffner *et al.*, 2009; Wang *et al.*, 2012).

In recent years, genome-wide association studies (GWAS) came to be widely employed to elucidate genetic variations that affect complex phenotypic traits, allowing identification of candidate loci controlling crop phenotypes (Huang *et al.*, 2012; Sukumaran *et al.*, 2014; Zhou *et al.*, 2015). However, phenotype is affected by biological pathways that involve interactions of multiple genes (Mackay, 2014). GWAS approach has been conventionally used to identifying major Quantitative Trait Loci (QTL) associated with a phenotype of interest. In most cases, these QTL were considered as contributing additive effects to the trait values independent of the effects of other loci. If there are strong phenotypic effects of gene-gene interaction, however, GWAS approach potentially misses important loci that control the trait in combination with other loci. In such cases, additive QTL may not explain the whole phenotypic variations (Carlborg and Haley, 2004; Mackay and Moore, 2014). Therefore, it is necessary to take epistasis into account for better understanding of the genetic factors controlling

phenotypic variations.

Identification of epistatic gene pairs is challenging, since one needs to consider a large number of combinations of genotypes, which incurs a heavy computational load and low statistical power due to multiple test correction. Despite these difficulties, a number of methods have been developed (Wei *et al.*, 2014; Niel *et al.*, 2015) that are classified into the two major approaches; the exhaustive and the non-exhaustive approach.

The exhaustive approach is designed to test all combinations of genetic variants including SNPs (Wan *et al.*, 2010; Hemani *et al.*, 2011; Li, 2017). The advantage of exhaustive approach is its lower risk of failure in detecting epistasis. However, the exhaustive search requires a higher computational inputs but nevertheless tends to have a lower statistical power due to multiple tests resulting from studying a large number of combinations of all pairwise genetic variations (Wei *et al.*, 2014). Therefore, reduction of search space is needed to mitigate the computational burden. There have been several studies that attempted to reduce the search space by incorporating information of candidate genes based on metabolic pathways, gene ontology, and protein-protein interactions (Ritchie, 2011; Sun *et al.*, 2014). However, these approaches are prone to ignoring unknown, but important, genes affecting the phenotype.

The non-exhaustive approach as represented by machine learning algorithms attempts to make non-parametric models to detect epistasis. Non-exhaustive approach is useful to detect higher-order epistatic relationships thanks to a low computational cost. However, this approach tends to generate highly complex models that sometimes suffer from a local optimality problem (Wei *et al.*, 2014; Tuo, 2018). Especially when the sample size is small, the complexity of models easily becomes too large as compared to the sample size. This complexity leads to overfitting of the model to the sample dataset (Niel *et al.*, 2015). Therefore, non-exhaustive approach is not appropriate in the samples with small sizes.

Recombinant Inbred Lines (RILs) are generated by first performing an intercross of genetically distinct inbred parents to obtain the F1 progeny. The F1 plants are self-pollinated to obtain the F2 plants, and each of the F2 progeny is self-pollinated several times by single seed descent (SSD) method to obtain further generations (Bailey, 1971). Each self-pollination reduces heterozygosity by half, so that after substantial number of generations (e.g. > F6), the genotypes of RILs become random mosaics of parental genotypes and the majority of genomes of RILs become homozygous. Therefore, using RILs enables one to remove the effects of heterozygous genotypes, which contributes to reducing the complexity of models used for the detection of epistasis. In addition, RILs allows phenotyping of multiple individuals from the same genotype, increasing the reliability of phenotype measurements.

In this study, we report a new approach named RIL-StEp specialized to RILs to detect epistasis in a pair of genetic variations based on the comparison of simple linear models. Bayes factor value is used to evaluating a model with epistasis against a null model without epistasis, and if this value is larger than a certain threshold, we assume that there is epistasis. This model considers the additive effects of significant QTL as well as epistatic effects between two

selected SNPs. Therefore, the model is simple and easy to interpret. We applied the method to study epistatic relationships of loci that affect seed hull color of rice (*Oryza sativa*). RIL-StEp identified three candidate genes, a gene for major QTL and two epistatically interacting genes, that may control seed hull color. We suggest RIL-StEp would lead to enhancing our understanding of the genetic architecture of phenotypes of important crops as well as other organisms.

Results

Phenotyping of seed hull color of rice RILs

In order to quantify rice seed hull color phenotype, we converted the color to numeric values based on the CIE XYZ color space. We then measured color values of the seeds of 235 RILs of F7 generation derived from a cross between the rice cultivar “Hitomebore” (japonica type rice) and “Kaluheenati” (aus type rice). Seed hull color of RILs showed a gradation, and was not categorized into the two discrete parental phenotypes, beige and black for Hitomebore and Kaluheenati, respectively (Figure 1, Table S1). Frequency distribution of color values of the 235 RILs is skewed toward the higher phenotypic value (Figure 1); approximately one-third of RILs were whitish brown seeds (the higher phenotypic values) whereas the rest were darker brown seeds (the lower phenotypic values). From these data, we conclude that seed hull color is not controlled by a single gene, but by multiple genes.

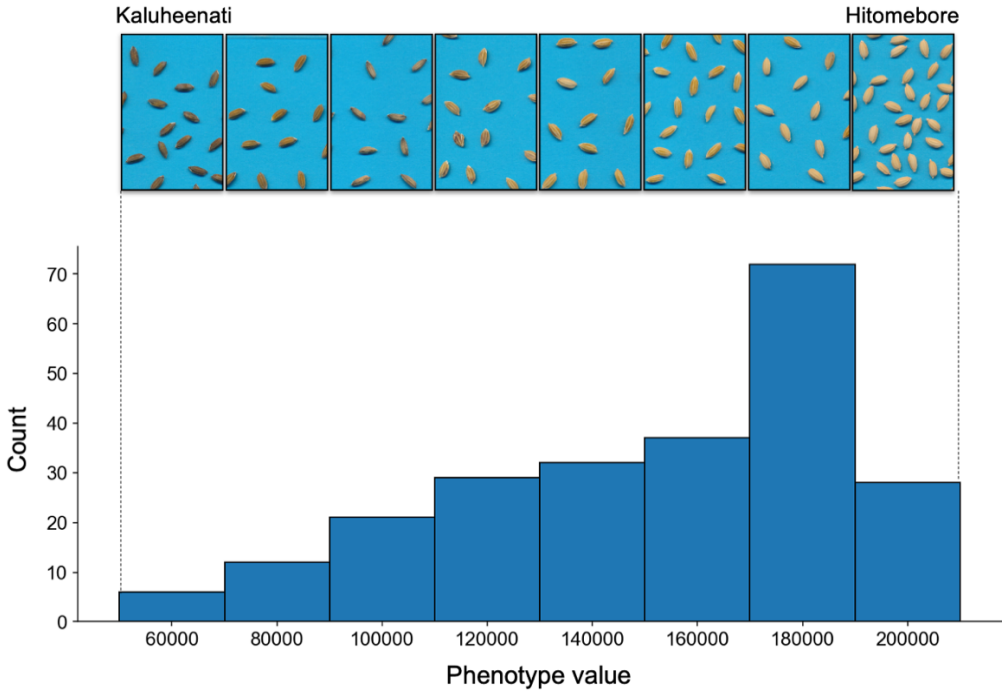

Figure 1 Seed hull color variation among the RILs and the distribution of phenotypic values. A histogram showing the distribution of phenotypic values. X-axis shows the range of phenotypic values of CIE XYZ color space. Y-axis shows the number of lines with phenotypic

values in each range. The panels at the top show the representative images of seeds in each range of the phenotypic values.

QTL analysis of seed hull color

We first carried out conventional QTL analysis to identify SNPs to be included in the models of RIL-StEp. Between the genomes of the two parents Hitomebore and Kaluheenati, we identified a total of 1,046,779 SNPs. We selected one SNP per 5,000bp interval and used 59,287 SNPs for subsequent QTL analysis and RIL-StEp. QTL analysis was carried out using 235 RILs by an R package "GWASpoly" (Rosyara *et al.*, 2016) to detect SNPs associated with the seed hull color phenotypes. As a result, we extracted two genomic regions showing statistical significance after the Bonferroni correction, *i.e.* $-\log_{10}(p) > 6.07$ on chromosome 4 and 9 (Figure 2, Table S2). Then, we selected two SNPs showing the highest $-\log_{10}(p)$ values in each region. These SNPs were located at chr04:23121877 and chr09: 6953870, respectively. We incorporated these two SNP values into the RIL-StEp models as the QTL variables. In order to study the possibility of epistasis of these two loci, we examined the effects of their genotypes on the phenotype. When the genotype of the SNP located in chr04:23121877 is Kaluheenati genotype, phenotype values tended to be lower (Figure S1A). The SNP located in chr09:6953870 also showed a similar tendency (Figure S1B), indicating there is no epistatic interaction between the two loci. However, when we focus on the SNP at chr04:23121877, the trait value variance of RILs with Kaluheenati genotype was larger than that of Hitomebore genotype (Figure S1C). Thus, the two QTLs do not fully explain the phenotypic variance and other genomic regions apart from these possibly affect the seed hull color.

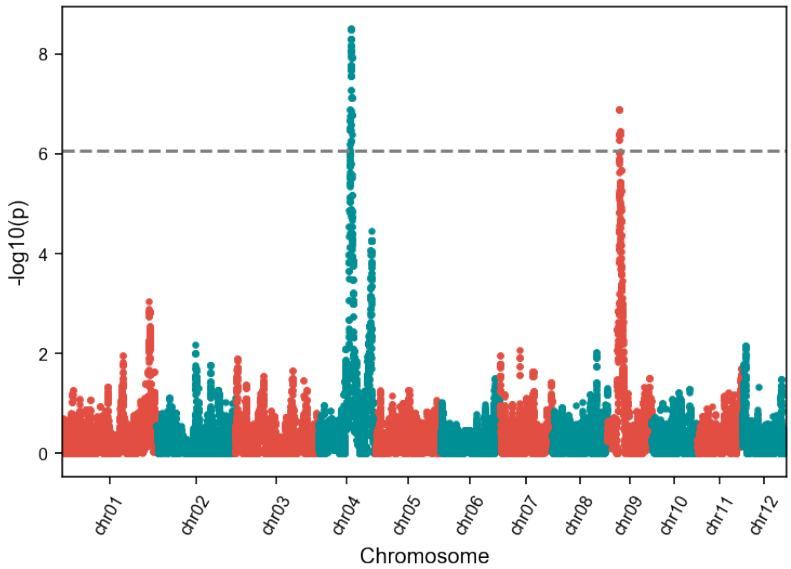

Figure 2 QTL analysis of rice seed hull color. A Manhattan plot showing the significant association of SNPs with seed hull color phenotype as calculated by GWASpoly (Rosyara *et al.*, 2016). Y-axis shows the $-\log_{10}(p)$ value of each SNP. X-axis shows the genomic position. Dashed line indicates the significance threshold after Bonferroni correction of multiple tests. Only

SNPs located near chr04:23121877 and chr09:6953870 exceeded the threshold.

RIL-StEp (*R*ecombinant *I*nbred *L*ines *St*epwise *Ep*istasis detection)

To detect genomic regions of RILs that are epistatically interacting, we developed a simple approach named RIL-StEp. In RIL-StEp, we generate linear models incorporating major QTLs as well as two SNPs at a time that are sampled from the entire genome. Two models, one with epistasis between the two SNPs and the other without epistasis, are compared by using Bayes factor. Specifically, we consider the following two linear models:

$$Model_1 : \boldsymbol{y} = \mu + \sum_{i=1}^{q} \boldsymbol{Q_i}\alpha_i + \boldsymbol{S_1}\beta_1 + \boldsymbol{S_2}\beta_2 + \boldsymbol{e} \quad (1)$$

$$Model_2 : \boldsymbol{y} = \mu + \sum_{i=1}^{q} \boldsymbol{Q_i}\alpha_i + \boldsymbol{S_1}\beta_1 + \boldsymbol{S_2}\beta_2 + \boldsymbol{E_1}\beta_3 + \boldsymbol{E_2}\beta_4 + \boldsymbol{E_3}\beta_5 + \boldsymbol{e} \quad (2)$$

$$\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

$\boldsymbol{y}$ is an *n*-vector of phenotypic values for *n* samples; μ is an intercept term; α_i is the additive effect of each SNP detected by QTL analysis; q is the number of QTLs; β_1 is the effect of first SNP and β_2 is the effect of second SNP. $\beta_{3\sim5}$ are the interaction effects of the alleles from the two SNPs; β_3: P1 (Parent 1) allele and P2 (Parent 2) allele; β_4: P2 allele and P1 allele, β_5: P2 allele and P2 allele, for the SNP1 and SNP2, respectively. One combination of alleles (P1-P1) is not included to escape multicollinearity (Table S3). $\boldsymbol{Q_i}, \boldsymbol{S_{1,2}}$ are the *n*-dimensional genotype vectors of 1s and 0s for each QTL and the two selected SNPs. $\boldsymbol{E_{1\sim3}}$ are *n*-dimensional vectors with 1s for samples with the specific combination of alleles of selected SNPs and 0s for the rest. $\boldsymbol{e}$ is an *n*-vector of residual error and σ^2 is residual error variance.

The $Model_1$ only includes QTLs and two selected SNPs as the variables. In the $Model_2$, we incorporated the variables of epistasis effects between the two selected SNPs in addition to the $Model_1$. We compared the $Model_1$ and the $Model_2$ based on Bayes factor. The Bayes factor is a ratio of the marginal likelihoods of the two models of hypotheses. To measure the better fit of $Model_2$ as compared to $Model_1$, we use the Bayes factor K given by:

$$K = \frac{Pr(\boldsymbol{y}|Model_2)}{Pr(\boldsymbol{y}|Model_1)} \quad (3)$$

$Pr(\boldsymbol{y}|Model)$ is the likelihood representing the probability that phenotypic data are produced under the assumption of the *Model*. The Bayes factor $K > 1$ means the $Model_2$ is more strongly supported by the phenotype dataset compared to $Model_1$, indicating that the model with epistasis effects is better supported. We considered the value of K larger than 100 as the evidence of epistasis, following the interpretation table (Jarosz and Wiley, 2014).

Application of RIL-StEp to rice RILs

We used RIL-StEp to detect SNP pairs showing significant genetic interactions in rice seed hull color trait. In this analysis, we incorporated two major QTLs in chromosome 4 and 9, respectively (Figure 2). To detect loci showing epistasis, we first selected one SNP every 10 SNPs out of 5,9287 SNPs across the genome, resulting in 5,929 SNPs to be considered. We applied RIL-StEp to the all pairs of the 5,929 SNPs. After calculating the Bayes factors for SNP pairs (Table S4), we focused on the genomic regions with SNP combinations showing the Bayes factor values > 100. After finding approximate positions of the loci showing possible epistasis, we applied RIL-StEp again to the combinations of all SNPs in the two regions (Figure 3, Table S5). As a result, we identified two genomic regions, chr04:22350619~25534998 and chr04:31048756~33482737 as the candidate regions showing epistatic interactions. The first region matched the position of the SNP detected by QTL analysis (chr04:23121877). The second region was not detected as a significant QTL; however, this corresponded to a peak with –log10p = 4.26 (Figure 2). SNP pairs between these regions showed a large Bayes factors values (Table S5). Thus, we hypothesized that the genes located in these two regions are interacting to each other. To validate this finding, we selected a SNP pair with the highest Bayes factor, and plotted the phenotype values for the combination of genotypes for the SNP pair (Figure 4). When the genotypes of SNPs located in chr04:23048862 and chr04:31581963 are both Kaluheenati types, the phenotype values tend to be low. On the other hand, if the genotypes are in other combinations, the color values were higher and similar to each other (Figure 4). This result suggested that both these regions should be Kaluheenati types to make seed hull color black. Therefore, it is assumed that two genes close to these SNPs are functioning together to determine seed hull color. To confirm this hypothesis, we surveyed candidate genes located in these two regions.

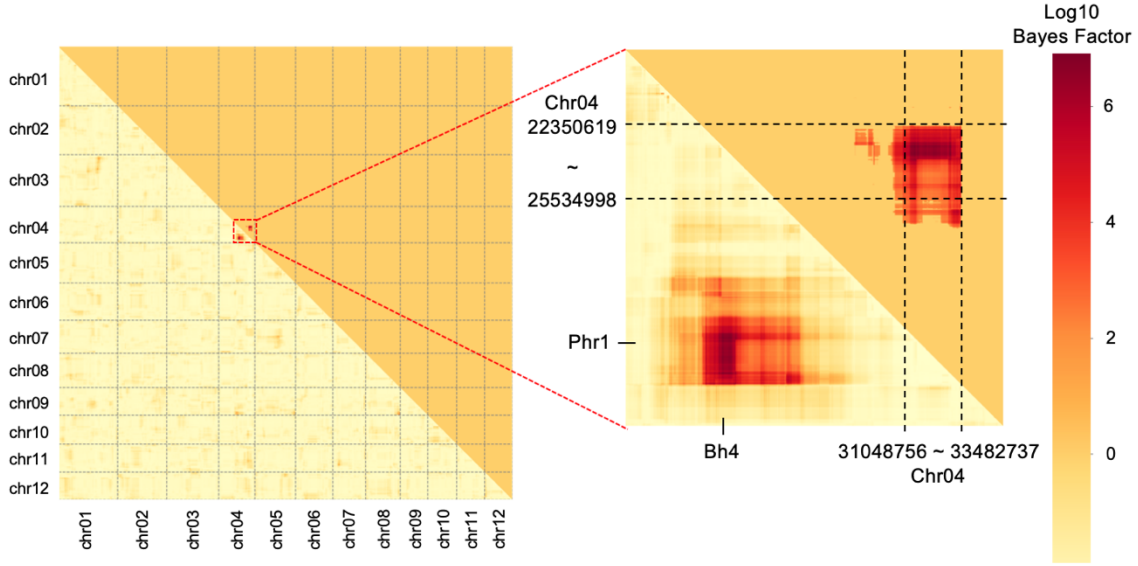

Figure 3 Heatmap showing Bayes factors for combinations of SNPs as revealed by RIL-StEp. The left heatmap shows the Bayes factor of SNP combinations over the whole genome.

The right heatmap magnifies genomic regions with the high Bayes factors. The lower triangle shows Bayes factors of all SNP combinations. The upper triangle highlights only combinations with Bayes factors > 100. Bayes factors of all combinations of SNPs located between chr04:22350619~25534998 and chr04:31048756~33482737, respectively, were higher than 100.

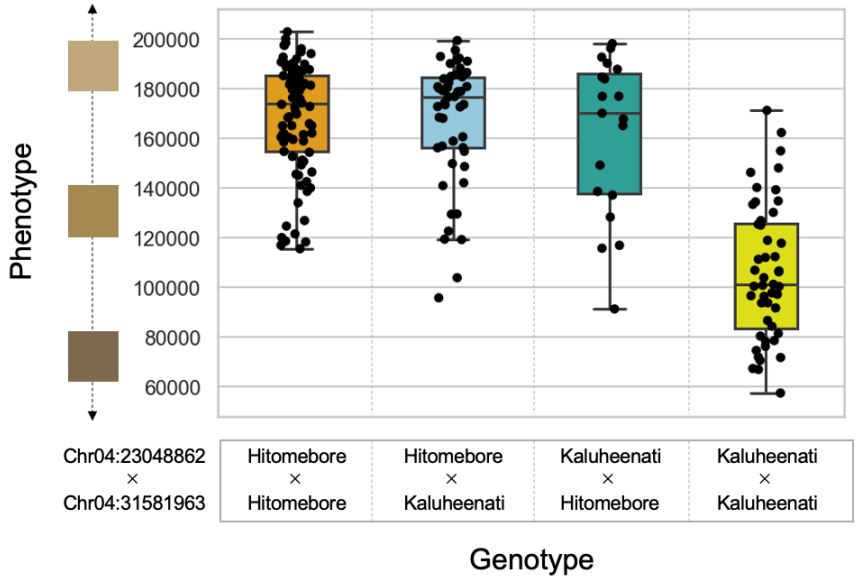

Figure 4 Relationships between phenotypic values and genotypes of the two epistatic SNPs as identified by RIL-StEp. A boxplot showing the phenotypic values of RILs with different combinations of genotypes of SNPs at chr04:23048862 and chr04:31581963. X-axis shows the combinations of genotypes. Y-axis shows phenotypic values. When genotypes of SNPs at chr04:23048862 and chr04:31581963 are both Kaluheenati genotype, phenotypic values tended to be low, whereas in other combinations, the values were higher and similar.

Identifying candidate genes involved in seed hull color epistasis

We surveyed genes located in the two regions as detected by RIL-StEp and tried to identify genes that may affect seed hull color. In the region chr04:22350619~25534998, there was *Black Hull 4* gene (*BH4* :chr04:22969845~22971859). In the region chr04:31048756~33482737, we found a gene called *Phenol reaction 1* (*Phr1*: chr04:31749141~31751604). A previous study showed that the loss of function of *Bh4* changed the black hull phenotype of wild rice species to white hull of cultivated rice (Zhu *et al.*, 2011). *Phr1* is known as the gene related to phenol reaction (Yu *et al.*, 2008). It was reported that brown hull color of *indica* rice is caused by the presence of *Phr1* (Yu *et al.*, 2008). RIL-StEp identified a pair of SNPs showing a high Bayes factor (Figure 3) and two genes close to the SNPs have been previously reported to control seed hull color. Therefore, we hypothesize that these genes are the major factors epistatically affecting seed hull color in our RILs.

We compared the nucleotide sequences of *BH4* and *Phr1* from the parental cultivars

Hitomebore and Kaluheenati used for generating the RILs. Kaluheenati had intact *BH4* and *Phr1* genes, whereas Hitomebore had a 22bp deletion in *BH4* and a 18bp deletion in *Phr1* (Figure S2A, S2B). These deletions are identical to those reported in other japonica cultivars (Fukuda *et al.*, 2012). In addition, these deletions were reported to cause loss-of-function in the respective genes (Yu *et al.*, 2008; Zhu *et al.*, 2011). Thus, we conclude that Kaluheenati maintains the function of *BH4* and *Phr1*, and Hitomebore cultivar probably lost their functions.

Using a crossed line between an indica type cultivar "Habataki" and japonica type cultivar "Arroz da Terra", Fukuda et al. (2012) reported that both *BH4* and *Phr1* are necessary for maintaining black hull phenotype. *BH4* encodes a tyrosine transporter and *Phr1* encodes a polyphenol oxidase of the tyrosinase family (Yu *et al.*, 2008; Zhu *et al.*, 2011). Tyrosine is converted by the tyrosinase to melanin, the main black pigment (Riley, 1997). Thus, it is assumed that *BH4* is required for transportation of tyrosine and *Phr1* for melanin biosynthesis (Figure 5). This suggests that the melanin biosynthesis pathway does not operate if either of these two genes does not function. It is consistent with the result that seed hull color tends to be beige when even one of the two SNPs is Hitomebore genotype (Figure 4).

In addition, we surveyed genes located near the SNP chr09:6953870 as identified by the QTL analysis to address its contribution to seed hull color in combination with *BH4* and *Phr1*. We found *Inhibitor for brown furrows1* (*IBF1*) located in chr09:6873236~6874612. The previous study showed *ibf1* mutants of japonica and indica type cultivars accumulate brown pigments during seed maturation. Thus, *IBF1* is a suppressor of brown pigment deposition in rice hull furrows (Shao *et al.*, 2012). We compared the sequences of *IBF1* for the two parental cultivars. Kaluheenati had 19bp deletion in *IBF1*, whereas Hitomebore had an intact protein coding region (Figure S2C). This result suggests that 19-bp deletion in Kaluheenati caused loss of function of *IBF1,* which no more suppresses the accumulation of brown pigmentation of rice hull furrows. This is in line with the lower phenotypic value (brown color) of RILs with Kaluheenati-type genotype near the *IBF1* gene (Figure S1B). *IBF1* is reported to be involved in flavonoids biosynthesis (Shao *et al.*, 2012).

Taken together, the relationship between seed hull color phenotype and genotypes of the three SNPs located near *BH4*, *Phr1*, and *IBF1* showed that the effect of *IBF1* is clearly independent of that of *BH4* and *Phr1* (Figure S3). Thus, the pathway involving *BH4* and *Phr1* and that of *IBF1* are probably functioning independently (Figure 5).

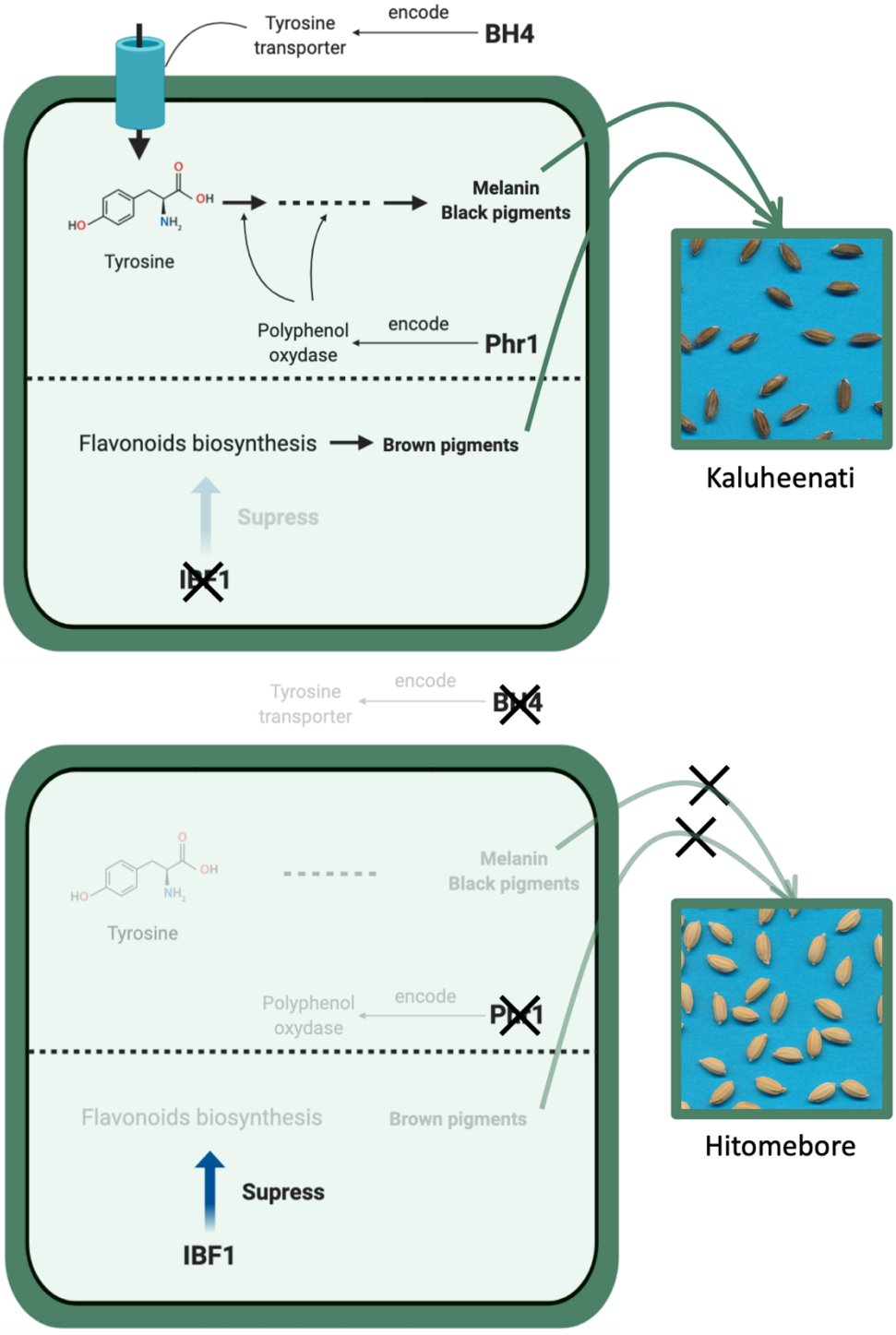

Figure 5 A simplified scheme of the pathways related to rice seed hull color as hypothesized in the present study. This figure shows the summary of biological function of *BH4*, *Phr1*, and *IBF1*. *BH4* encodes a tyrosine transporter (Zhu *et al.*, 2011) and *Phr1* encodes a polyphenol oxidase (Yu *et al.*, 2008). These genes are related to melanin biosynthesis pathway. *IBF1* inhibits flavonoids biosynthesis as a suppressor (Shao *et al.*, 2012).

Discussion

In this study, we describe a new approach "RIL-StEp" for detecting epistatic relationships of genes. This approach is specialized to RIL population and based on Bayes factors for comparison of simple linear models. Using RIL-StEp we successfully detected a likely gene pair showing epistasis that affect seed hull color. The advantage of RIL-StEp is its high interpretability as compared to other approaches that consider many variables at once. Our model includes in variables only significant QTLs as well as epistasis of two SNPs at a time without considering heterozygous genotype. Thus, our model has a low complexity without any possibility of overfitting as seen in the complex model. Therefore, we believe that RIL-StEp is a recommended option to detect epistasis in any traits when the RILs are used. Additionally, our approach adopted Bayes factors. It is known that Bayes Factors have flexibility to combine prior information on each effect of genetic variants (Wakefield, 2009; Runcie and Crawford, 2019). Thus, although we specified prior distribution according to the paper (Liang *et al.*, 2008), our approach is capable to incorporate any prior information. The disadvantage of our approach is the difficulty in detecting higher-order (e.g. 3 loci) epistatic relationships. Detection of high-order relationships using our exhaustive approach increases computational cost explosively and decreases the interpretability of the models (Taylor and Ehrenreich, 2015). Therefore, the non-exhaustive approach may be more appropriate to identify the high-order epistasis.

We succeeded in identifying genomic regions that show epistasis. However, these regions contained multiple genes and we could not specify the responsible genes only by the genetic analysis. It is a challenging problems of GWAS to fill the gap between identification of the genomic regions and identification of the causative genes responsible for the phenotype (Gallagher and Chen-Plotkin, 2018). It could be possible to pin down to much smaller genomic regions by applying a more strict threshold in the epistasis analysis. However, this has a risk of missing true positive SNPs. It is known to be difficult to decide the proper balance between type I and type II errors (Todorov and Rao, 1997). Thus, the approach to identify genes using only statistical significance threshold is usually not possible and not appropriate.

In our case, we have successfully specified strong candidate genes presumably controlling the phenotype using the knowledge about the candidate genes and the sequence analysis. However, this approach may not be applicable in every case. In epistasis analysis, there may be several approaches to validate the epistatic relationship between the genes. For example, co-expression analysis explore genes in the same biological processes (Aoki *et al.*, 2007; Mao and Chen, 2012; van Dam *et al.*, 2018). eQTL analysis identify genetic variants regulated by specific genes (Gilad *et al.*, 2008; Feltus, 2014). Therefore, combining information from other sources of evidence to RIL-StEp may enhance our capability of identifying interacting genes.

We quantitatively measured seed hull color to use as phenotypes. Seed hull color can be treated not only as categorical traits, but also as quantitative traits (Shao *et al.*, 2011). In our study, the difference in seed hull color between the two parental lines is most likely controlled

by the three genes. Two genes of them seem interacting to each other. Thus, seed hull color exhibited gradual change according to the genotypes of these genes.

The main factor of seed hull color is probably black and brown pigments. Thus, it is assumed that measuring seed hull colors based on the brightness in CIE XYZ color space was appropriate. However, when various colors are included, it is difficult to convert colors to 1-dimensional values and we have to consider other methods of measurement. Some of other traits are also difficult to assess, like virulence and resistance response (Stewart and McDonald, 2014; Stewart *et al.*, 2016). Thus, developing methods to quantitatively measure the trait is one of the critical steps in understanding genetic architecture that affects traits.

To summarize, we propose a novel approach based on simple linear models to detect epistatic interaction for quantitative traits in the RIL population. By applying RIL-StEp to rice seed hull color, we succeeded in identifying three genomic regions related to seed hull color. Incorporating additional information allowed us to identify candidate genes involved in seed hull color variation. Thus, our approach has the potential to identify epistasis in various biological traits.

Experimental procedures

Materials

A rice (*Oryza sativa*) cultivar Hitomebore belonging to the japonica rice group shows white seed hull color. On the other hand, a cultivar Kaluheenati, which is one of the NARO World Rice Core Collection (WRC) (Kojima *et al.*, 2005), belonging to the aus rice group shows black seed hull color (Figure 1). Hitomebore and Kaluheenati were crossed and RILs of the F9 generations consisting of 235 lines were generated by the single seed descent (SSD) method. Images of seeds of each line were scanned and saved as pictures for phenotyping of seed hull color.

Methods

Genotyping of RILs by whole genome resequencing

To obtain the genotypes of all RILs, we performed the whole genome resequencing of the parents and 235 RILs. We filtered and trimmed these sequences using prinseq (Schmieder and Edwards, 2011) and FaQCs (Lo and Chain, 2014). Then, the quality-trimmed Illumina short read data were aligned against the reference genome using BWA (Li and Durbin, 2009). We used genome sequence of Os-Nipponbare-Reference-IRGSP-1.0 as the reference (Kawahara *et al.*, 2013). After mapping, we sorted and added index to bam files using samtools(Li *et al.*, 2009). These bam files were subjected to variant calling with bcftools (Narasimhan *et al.*, 2016). Finally, we imputed the variants based on Hitomebore and Kaluheenati genotypes using LB-impute (Fragoso *et al.*, 2016). For biallelic SNPs in our RILs, there are three genotypic classes Hitomebore-Hitomebore, Hitomebore-Kaluheenati and Kaluheenati-Kaluheenati. These genotype classes were parameterized to {0, 1, 2}. We used 59,287 SNPs for the analysis

that were selected from a total of 1,046,779 SNPs found between the two rice parents. For selection of SNP, we used only one SNP per 5,000bp interval.

Phenotyping and quantification of seed hull color

In the RILs, seed hull color showed gradation between beige and black colors (Figure 1). It is known that quantitation of phenotypes tend to improve statistical power and interpretability of relationships between genetic variants and phenotypes (Bush and Moore, 2012). Therefore, in order to convert seed hull color phenotypes to quantitative values, we measured the brightness of the seed hull color. First, we extracted seed image from the original picture and constructed the matrix of RGB values of the image. Then, we applied Principal Component Analysis(PCA) to extract the RGB values to pick up representative color of all seeds in the image (Figure S4). We applied this process to each RIL and obtained the representative RGB value of seed hull color for each line. Finally, we converted these representative RGB values to CIE XYZ color space. Y axis value showed the brightness in CIE XYZ color space. Thus, we used y-axis values as quantitative phenotypes. The larger y-axis values express brighter color as compared to the lower y-axis values corresponding to darker color (Figure S4).

QTL analysis

To identify SNPs corresponding to major QTLs and include these SNPs in our linear models, we used genome-wide association study (GWAS) approach based on the mixed linear model (Yu *et al.*, 2006). We utilized the R package "GWASpoly" (Rosyara *et al.*, 2016) to identify genomic regions that show a significant association with the phenotypic effect. We used the Bonferroni method to determine the QTL significance threshold. Then, we selected a SNP with the largest values of $-log10p$ for each genomic region that exceeds the Bonferroni threshold. These selected representative SNPs will be included in the $Model_1$ and $Model_2$ as the major QTLs.

Calculating Bayes factors in RIL-StEp

Bayes factors are computed by integrating the likelihood with respect to the priors on parameters. We estimated Bayes factors based on Monte Carlo sampling for the integration of parameters. The equation (1) and (2) can be expressed as:

$$\boldsymbol{y} = \mu + \boldsymbol{X\theta} + \boldsymbol{e}, \boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}) \quad (4)$$

$\boldsymbol{X}$ is a $n \times r$ design matrix of genotypes for QTL or epistasis variables. $\boldsymbol{\theta}$ is a $r \times 1$ vector of QTL and epistasis effects. r is sum of the number of QTLs and epistasis variables used in the model. In Monte Carlo sampling, we specified the prior distribution of $\boldsymbol{\theta}$ as given by:

$$\boldsymbol{\theta} \sim N(\boldsymbol{0}, g\sigma^2(\boldsymbol{X^T X^{-1}})), g \sim InverseGamma\left(1/2, \sqrt{2}/8\right) \quad (5)$$

The number of iterations to estimate Bayes factor was 10,000. We used the R package "BayesFactor" (Morey *et al.*, 2018) to compute Bayes factors. We applied these processes to a total of 17,573,556 combinations of SNPs. When we calculated Bayes factor, we did not consider RILs showing heterozygous genotypes at the QTL or the selected SNPs.

The source codes and detailed usage instructions of RIL-StEp are freely available from GitHub (https://github.com/slt666666/RILStEp) under MIT license.

Data availability

The genotype dataset, seed images of RILs, and detail of supporting information (Table S4, S5) were deposited in the Zenodo (10.5281/zenodo.3882105). All other relevant data are within the paper and the supplemental files. RIL-StEp package source codes and a user manual are freely available through GitHub (https://github.com/slt666666/RILStEp) under MIT License. The scripts used in phenotyping process are also deposited in GitHub (https://github.com/slt666666/Seed_phenotyping)

Author Contributions

R.T. and T.S. conceptualized the study. T.S., A.A., and M.S. performed research. The original draft was written by T.S., and reviewed by R.T., A.A., and M.S.

Acknowledgements

We thank the National Agriculture and Food Research Organization (NARO) gene bank, Japan for providing the WRC seed. This study was supported by grants from the Project of the NARO Bio-oriented Technology Research Advancement Institution (Research program on development of innovative technology), and by grant JSPS KAKENHI 15H05779 and 20H00421 to RT, 17H03752 and 20H02962 to AA. We thank Sophien Kamoun for valuable comments. The authors declare that there are no conflicts of interest.

Short legends for Supporting Information

Table S1 Phenotypic values of seed hull color.

Table S2 The results of GWAS analysis using GWASpoly.

Table S3 Values of RIL-StEp coefficients used for each genotype.

Table S4 Results of RIL-StEp for combinations in whole genome regions.

Table S5 Results of RIL-StEp for combinations in strong candidate genomic regions.

References

Aoki, K., Ogata, Y. and Shibata, D. (2007) Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant Cell Physiol.*, **48**, 381–390.

Bailey, D.W. (1971) Recombinant-inbred strains: An aid to finding identity, linkage, and function op histocompatibility and other genes. *Transplantation*, **11**, 325–327.

Bush, W.S. and Moore, J.H. (2012) Chapter 11: Genome-Wide Association Studies. *PLoS Comput. Biol.*, **8**.

Carlborg, Ö. and Haley, C.S. (2004) Epistasis: Too often neglected in complex trait studies? *Nat. Rev. Genet.*, **5**, 618–625.

Cordell, H.J. (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.*, **11**, 2463–2468.

Dam, S. van, Võsa, U., Graaf, A. van der, Franke, L. and Magalhães, J.P. de (2018) Gene co-expression analysis for functional classification and gene-disease predictions. *Brief. Bioinform.*, **19**, 575–592.

Feltus, F.A. (2014) Systems genetics: A paradigm to improve discovery of candidate genes and mechanisms underlying complex traits. *Plant Sci.*, **223**, 45–48. Available at: http://dx.doi.org/10.1016/j.plantsci.2014.03.003.

Fisher, R.A. (1919) XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Trans. R. Soc. Edinburgh*, **52**, 399–433. Available at: https://www.cambridge.org/core/product/identifier/S0080456800012163/type/journal_article.

Fragoso, C.A., Heffelfinger, C., Zhao, H. and Dellaporta, S.L. (2016) Imputing genotypes in biallelic populations from low-coverage sequence data. *Genetics*, **202**, 487–495.

Fukuda, A., Shimizu, H., Shiratsuchi, H., Yamaguchi, H., Ohdaira, Y. and Mochida, H. (2012) Complementary Genes That Cause Black Ripening Hulls in F 1 Plants of Crosses between Indica and Japonica Rice Cultivars. *Plant Prod. Sci.*, **15**, 270–273. Available at: https://www.tandfonline.com/doi/full/10.1626/pps.15.270.

Gallagher, M.D. and Chen-Plotkin, A.S. (2018) The Post-GWAS Era: From Association to Function. *Am. J. Hum. Genet.*, **102**, 717–730. Available at: https://doi.org/10.1016/j.ajhg.2018.04.002.

Gilad, Y., Rifkin, S.A. and Pritchard, J.K. (2008) Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.*, **24**, 408–415.

Heffner, E.L., Sorrells, M.E. and Jannink, J.L. (2009) Genomic selection for crop improvement. *Crop Sci.*, **49**, 1–12.

Hemani, G., Theocharidis, A., Wei, W. and Haley, C. (2011) EpiGPU: Exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. *Bioinformatics*, **27**, 1462–1465.

Huang, X., Zhao, Y., Wei, X., et al. (2012) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat. Genet.*, **44**, 32–39.

Jarosz, A.F. and Wiley, J. (2014) What Are the Odds? A Practical Guide to Computing and Reporting Bayes Factors. *J. Probl. Solving*, **7**, 2–9. Available at: https://docs.lib.purdue.edu/jps/vol7/iss1/2.

Kawahara, Y., la Bastide, M. de, Hamilton, J.P., et al. (2013) Improvement of the Oryza

sativa Nipponbare reference genome using next generation sequence and optical map data. *Rice*, **6**, 1–10.

Kojima, Y., Ebana, K., Fukuoka, S., Nagamine, T. and Kawase, M. (2005) Development of an RFLP-based rice diversity research set of germplasm. *Breed. Sci.*, **55**, 431–440.

Li, H. and Durbin, R. (2009) Making the Leap: Maq to BWA. *Mass Genomics*, **25**, 1754–1760. Available at: http://massgenomics.org/2009/12/making-the-leap-maq-to-bwa.html.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li, X. (2017) A fast and exhaustive method for heterogeneity and epistasis analysis based on multi-objective optimization. *Bioinformatics*, **33**, 2829–2836.

Liang, F., Paulo, R., Molina, G., Clyde, M.A. and Berger, J.O. (2008) Mixtures of g priors for Bayesian variable selection. *J. Am. Stat. Assoc.*, **103**, 410–423.

Lo, C.C. and Chain, P.S.G. (2014) Rapid evaluation and quality control of next generation sequencing data with FaQCs. *BMC Bioinformatics*, **15**, 1–8.

Mackay, T.F.C. (2014) Epistasis and quantitative traits: Using model organisms to study gene-gene interactions. *Nat. Rev. Genet.*, **15**, 22–33. Available at: http://dx.doi.org/10.1038/nrg3627.

Mackay, T.F.C. and Moore, J.H. (2014) Why epistasis is important for tackling complex human disease genetics. *Genome Med.*, **6**, 6–8.

Mao, D. and Chen, C. (2012) Colinearity and Similar Expression Pattern of Rice DREB1s Reveal Their Functional Conservation in the Cold-Responsive Pathway. *PLoS One*, **7**.

Morey, R.D., Rouder, J.N., Jamil, T., Urbanek, S., Forner, K. and Ly, A. (2018) BayesFactor 0.9.12-4.2. *Compr. R Arch. Netw.*

Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C. and Durbin, R. (2016) BCFtools/RoH: A hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics*, **32**, 1749–1751.

Niel, C., Sinoquet, C., Dina, C. and Rocheleau, G. (2015) A survey about methods dedicated to epistasis detection. *Front. Genet.*, **6**.

Phillips, P.C. (2008) Epistasis - The essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.*, **9**, 855–867.

Riley, P.A. (1997) Melanin. *Int. J. Biochem. Cell Biol.*, **29**, 1235–1239. Available at: https://linkinghub.elsevier.com/retrieve/pii/S1357272597000137.

Ritchie, M.D. (2011) Using biological knowledge to uncover the mystery in the search for epistasis in genome-wide association studies. *Ann. Hum. Genet.*, **75**, 172–182.

Rosyara, U.R., Jong, W.S. de, Douches, D.S. and Endelman, J.B. (2016) Software for genome-wide association studies in autopolyploids and its application to potato. *Plant Genome*, **9**, 1–10.

Runcie, D.E. and Crawford, L. (2019) Fast and flexible linear mixed models for genome-wide

genetics. *PLoS Genet.*, **15**, 1–24.

Schmieder, R. and Edwards, R. (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**, 863–864.

Shao, T., Qian, Q., Tang, D., Chen, J., Li, M., Cheng, Z. and Luo, Q. (2012) A novel gene IBF1 is required for the inhibition of brown pigment deposition in rice hull furrows. *Theor. Appl. Genet.*, **125**, 381–390.

Shao, Y., Jin, L., Zhang, G., Lu, Y., Shen, Y. and Bao, J. (2011) Association mapping of grain color, phenolic content, flavonoid content and antioxidant capacity in dehulled rice. *Theor. Appl. Genet.*, **122**, 1005–1016.

Stewart, E.L., Hagerty, C.H., Mikaberidze, A., Mundt, C.C., Zhong, Z. and McDonald, B.A. (2016) An improved method for measuring quantitative resistance to the wheat pathogen Zymoseptoria tritici using high-throughput automated image analysis. *Phytopathology*, **106**, 782–788.

Stewart, E.L. and McDonald, B.A. (2014) Measuring quantitative virulence in the wheat pathogen zymoseptoria tritici using high-throughput automated image analysis. *Phytopathology*, **104**, 985–992.

Sukumaran, S., Dreisigacker, S., Lopes, M., Chavez, P. and Reynolds, M.P. (2014) Genome-wide association study for grain yield and related traits in an elite spring wheat population grown in temperate irrigated environments. *Theor. Appl. Genet.*, **128**, 353–363.

Sun, X., Lu, Q., Mukheerjee, S., Crane, P.K., Elston, R. and Ritchie, M.D. (2014) Analysis pipeline for the epistasis search - statistical versus biological filtering. *Front. Genet.*, **5**, 1–7.

Taylor, M.B. and Ehrenreich, I.M. (2015) Higher-order genetic interactions and their contribution to complex traits. *Trends Genet.*, **31**, 34–40. Available at: http://dx.doi.org/10.1016/j.tig.2014.09.001.

Todorov, A.A. and Rao, D.C. (1997) Trade-off between false positives and false negatives in the linkage analysis of complex traits. *Genet. Epidemiol.*, **14**, 453–464.

Tuo, S. (2018) FDHE-IW: A fast approach for detecting high-order epistasis in genome-wide case-control studies. *Genes (Basel).*, **9**.

Wakefield, J. (2009) Bayes factors for Genome-wide association studies: Comparison with P-values. *Genet. Epidemiol.*, **33**, 79–86.

Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N.L.S. and Yu, W. (2010) BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am. J. Hum. Genet.*, **87**, 325–340. Available at: http://dx.doi.org/10.1016/j.ajhg.2010.07.021.

Wang, D., Salah El-Basyoni, I., Stephen Baenziger, P., Crossa, J., Eskridge, K.M. and Dweikat, I. (2012) Prediction of genetic values of quantitative traits with epistatic effects in plant breeding populations. *Heredity (Edinb).*, **109**, 313–319. Available at:

http://dx.doi.org/10.1038/hdy.2012.44.

Wei, W.H., Hemani, G. and Haley, C.S. (2014) Detecting epistasis in human complex traits. *Nat. Rev. Genet.*, **15**, 722–733. Available at: http://dx.doi.org/10.1038/nrg3747.

Xu, Y. and Crouch, J.H. (2008) Marker-assisted selection in plant breeding: From publications to practice. *Crop Sci.*, **48**, 391–407.

Yu, J., Pressoir, G., Briggs, W.H., et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.*, **38**, 203–208.

Yu, Y., Tang, T., Qian, Q., et al. (2008) Independent Losses of Function in a Polyphenol Oxidase in Rice: Differentiation in Grain Discoloration between Subspecies and the Role of Positive Selection under Domestication. *Plant Cell*, **20**, 2946–2959. Available at: http://www.plantcell.org/lookup/doi/10.1105/tpc.108.060426.

Zhou, Z., Jiang, Y., Wang, Z., et al. (2015) Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.*, **33**, 408–414.

Zhu, B.F., Si, L., Wang, Z., et al. (2011) Genetic control of a transition from black to straw-white seed hull in rice domestication. *Plant Physiol.*, **155**, 1301–1311.

Supporting Information

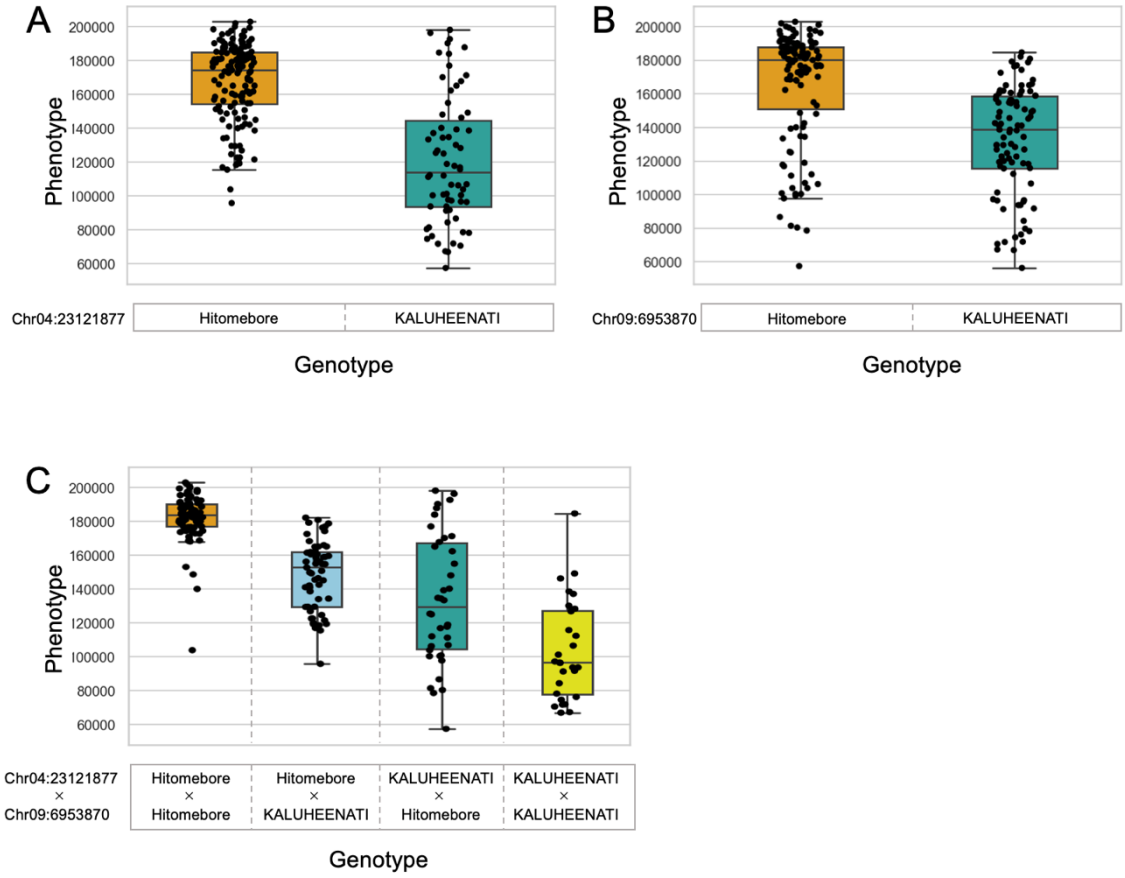

Figure S1. Relationships between rice seed hull color phenotypes and the genotypes of the two major QTLs. Boxplots showing the phenotypic values of RILs (Y-axis) in relation to

their genotypes at the two major QTLs (Y-axis). (A) The phenotypic values of RILs separately shown for Hitomebore or Kaluheenati genotypes at the SNP of chr04:23121877. (B) The phenotypic values of RILs separately shown for Hitomebore or Kaluheenati genotypes at the SNP of chr09:6953870. (C) The phenotypic values of RILs shown separately for the four combinations of genotypes at the SNPs of chr04:23121877 and chr09:6953870.

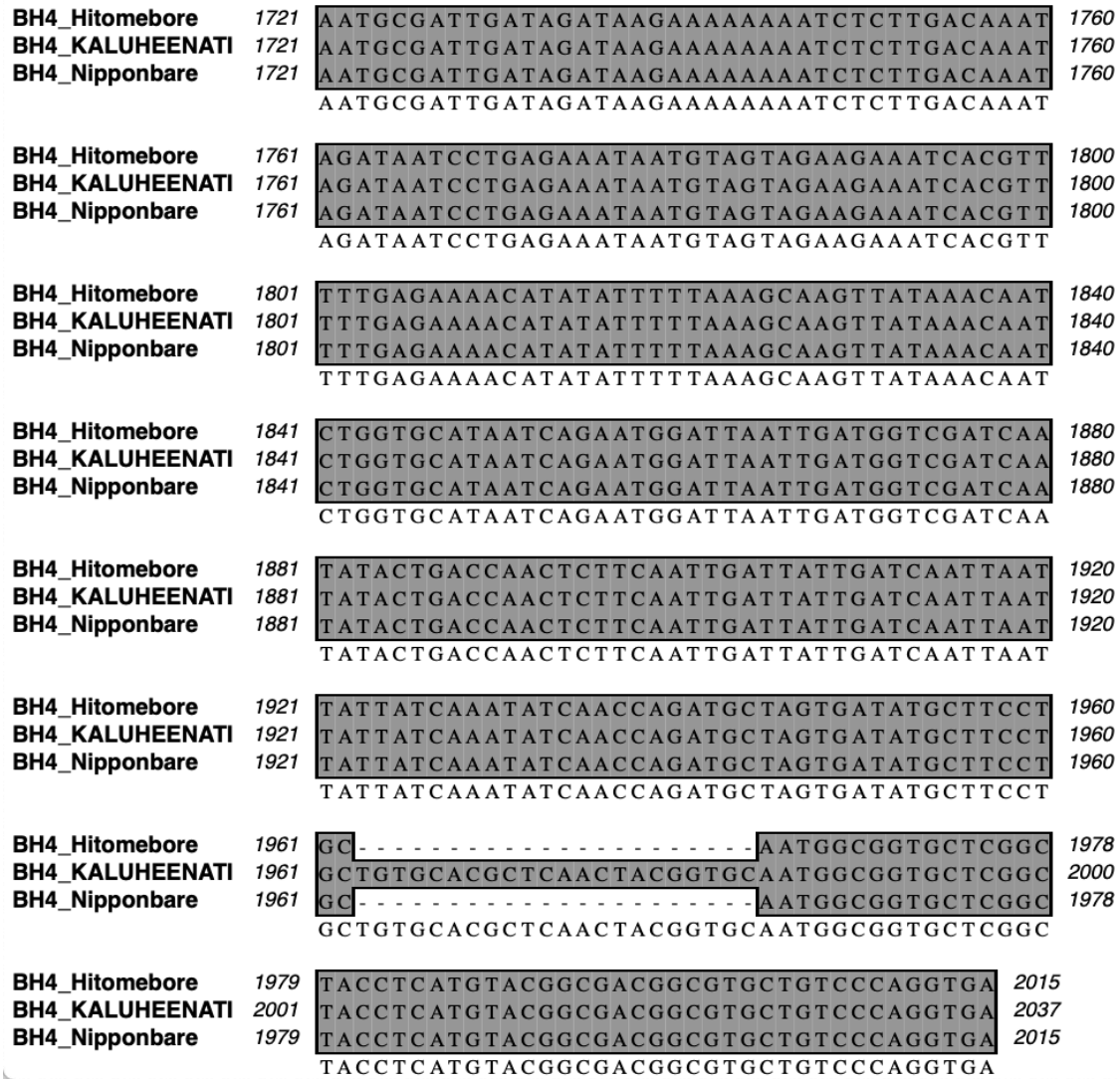

Figure S2A. DNA sequence alignment of the *BH4* gene. Genome sequences of the *BH4* gene for Hitomebore, KALUHEENATI, and Nipponbare are aligned and shown. Hitomebore and Nipponbare sequences have a 22bp deletion.

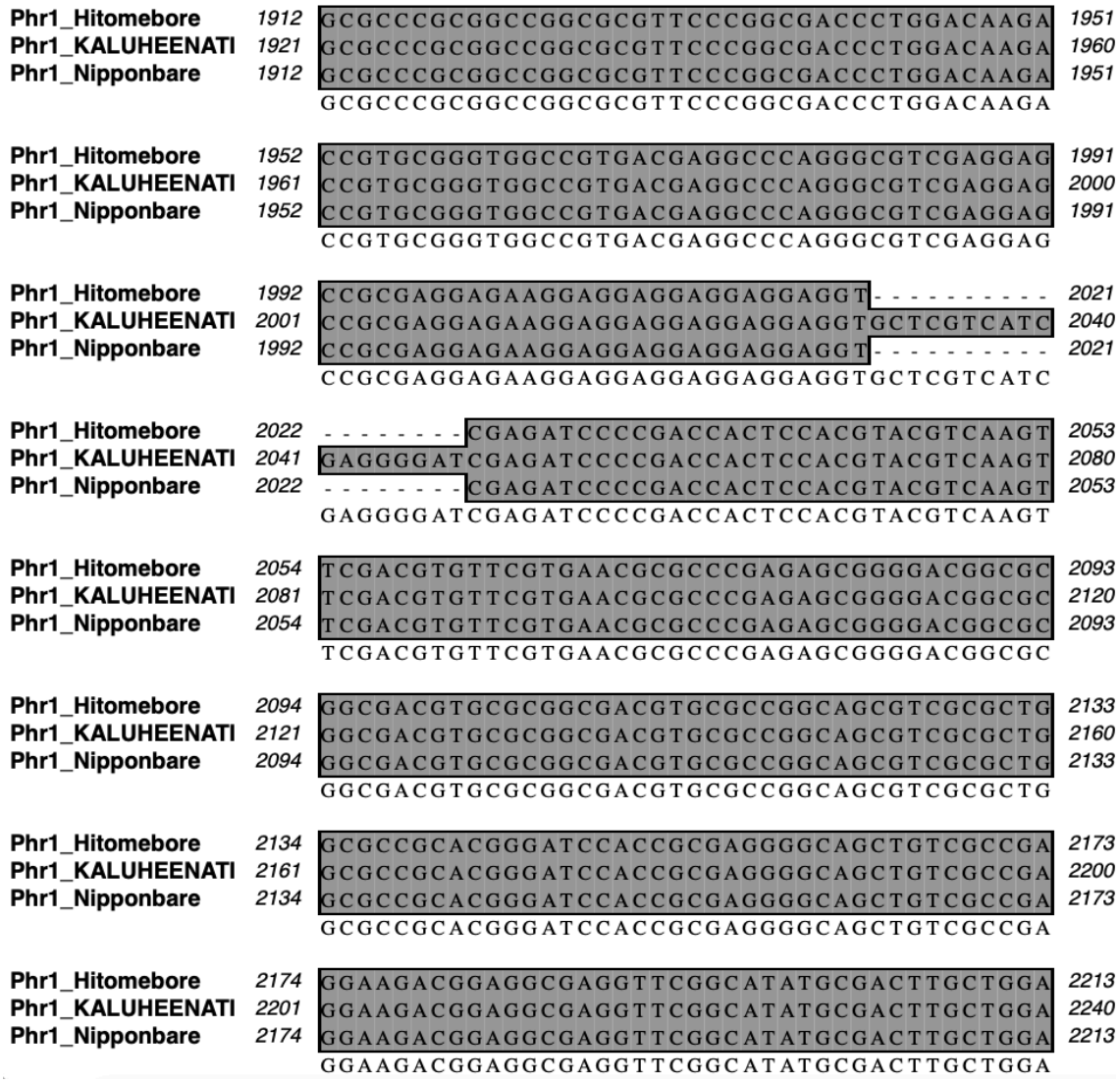

Figure S2B The sequence alignments of *Phr1.* Genome sequences of Hitomebore, KALUHEENATI, and Nipponbare in the region of *Phr1* gene. Hitomebore and Nipponbare have 18bp deletion.

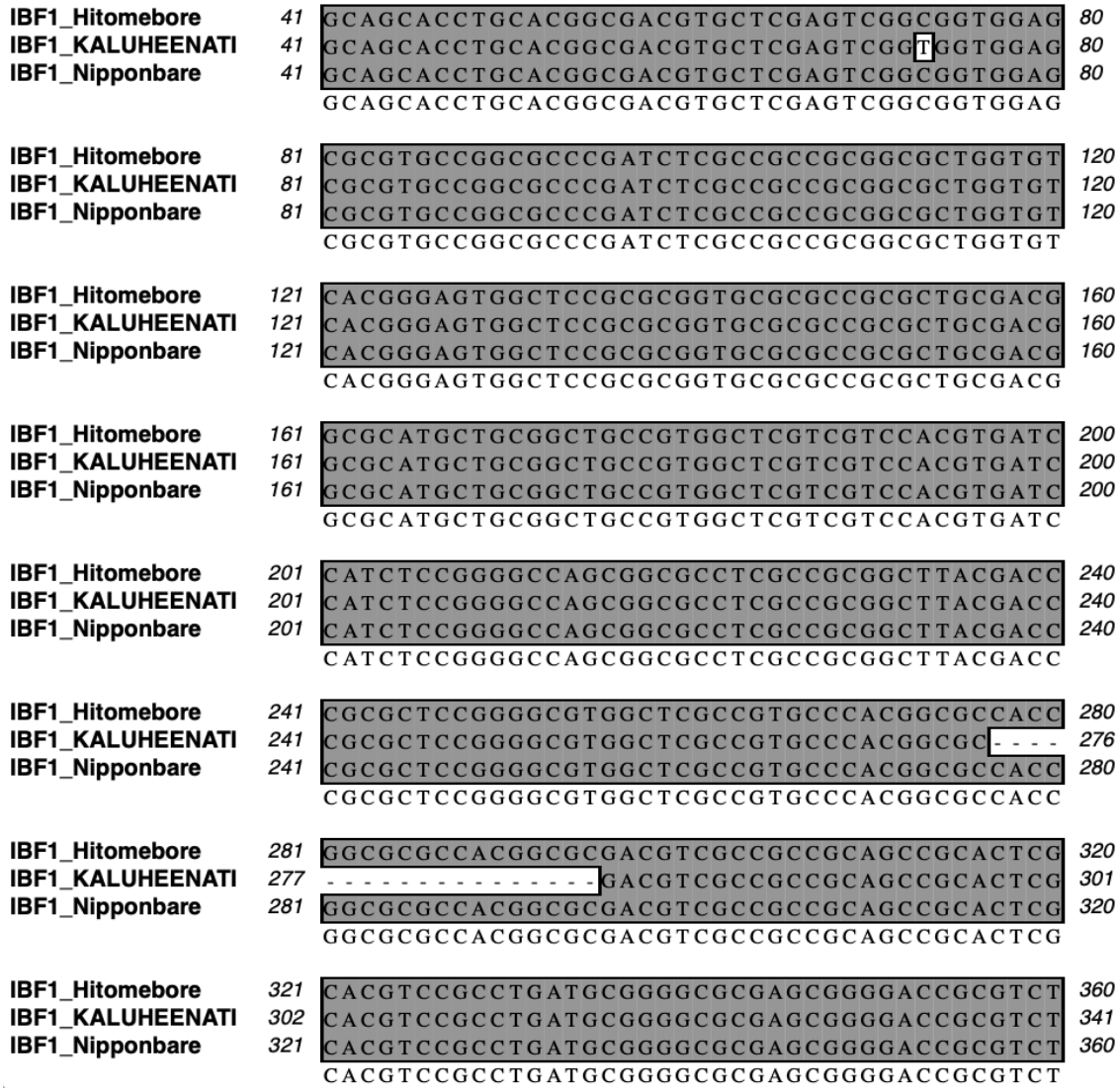

Figure S2C The sequence alignments of *IBF1*. Genome sequences of Hitomebore, KALUHEENATI, and Nipponbare in the region of *IBF1* gene. KALUHEENATI has 19bp deletion and 1 genetic variant.

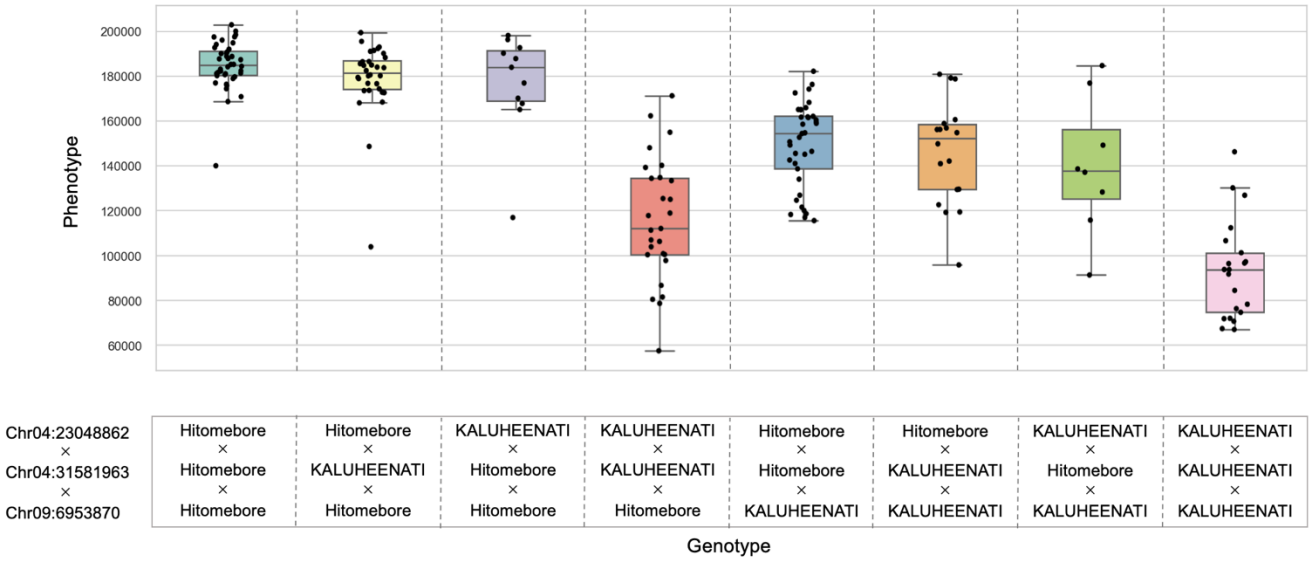

Figure S3. Relationships between rice seed hull color phenotypes and genotypes of the three loci. Boxplots showing the phenotypic values of RILs (Y-axis) separately for the different combinations of genotypes in the three loci (X-axis). When the genotype of chr09:6953870 is KALUHEENATI genotype, phenotypic values are consistently low independent of the genotypes of the other two loci. When the genotypes of chr04:23048862 and chr04:31581963 are both KALUHEENATI types, the phenotypic values tend to be low independent of chr09:6953870.

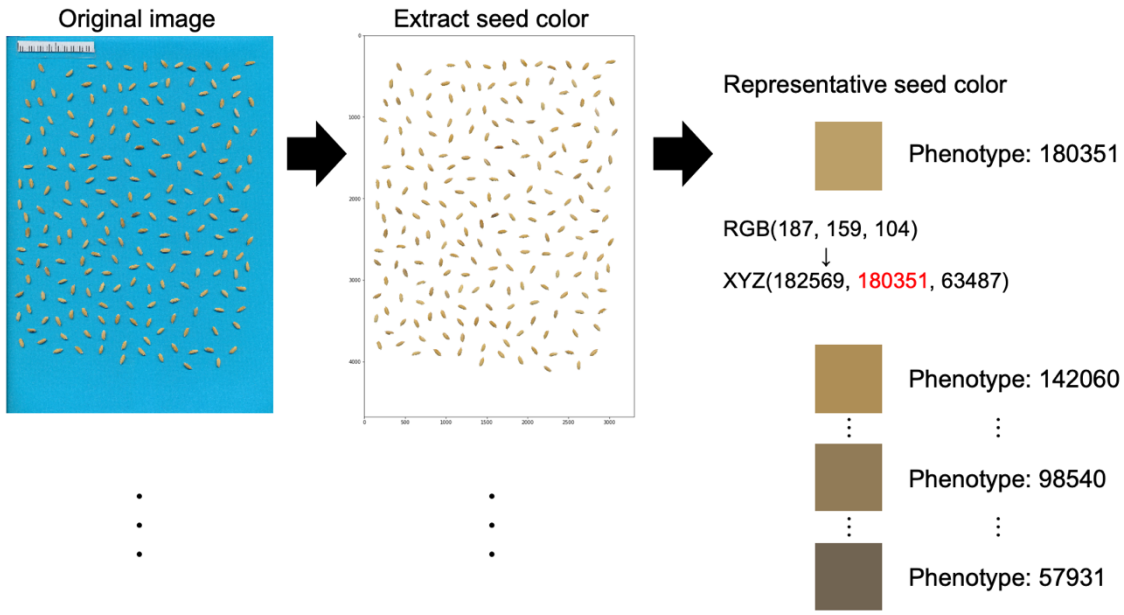

Figure S4. Phenotyping of rice seed hull color. We first trimmed only the image of seeds. Next, we extracted the representative color of the seeds using Principal Component Analysis(PCA). Finally, we converted RGB values of representative seed hull color to Y-axis values of CIE XYZ color space.