Abstract
Background: Difficulty in detecting rare variants is one of the problems in conventional genome wide association studies (GWAS). The problem is closely related to the complex gene compositions comprising multiple alleles, such as haplotypes. Several single nucleotide polymorphism (SNP) set approaches have been proposed to solve this problem. These methods, however, have been rarely discussed in connection with haplotypes. In this study, we developed a novel SNP-set GWAS method named “RAINBOW” and applied the method to haplotype-based GWAS by regarding a haplotype block as a SNP-set. Combining haplotype block estimation and SNP-set GWAS, haplotype-based GWAS can be conducted without prior information of haplotypes.

Results: We prepared 100 datasets of simulated phenotypic data and real marker genotype data of *Oryza sativa* subsp. *indica*, and performed GWAS of the datasets. We compared the power of our method, the conventional single-SNP GWAS, the conventional haplotype-based GWAS, and the conventional SNP-set GWAS. The results of the comparison indicated that the proposed method was able to better control false positives than the others. The proposed method was also excellent at detecting causal variants without relying on the linkage disequilibrium if causal variants were genotyped in the dataset. Moreover, the proposed method showed greater power than the other methods, i.e., it was able to detect causal variants that were not detected by the others, especially when the causal variants were located very close to each other and the directions of their effects were opposite.

Conclusion: The proposed method, RAINBOW, is especially superior in controlling false positives, detecting causal variants, and detecting nearby causal variants with opposite effects. By using the SNP-set approach as the proposed method, we expect that detecting not only rare variants but also genes with complex mechanisms, such as genes with multiple causal variants, can be realized. RAINBOW was implemented as the R package and is available at https://github.com/KosukeHamazaki/RAINBOW.

Keywords: GWAS; SNP-set; haplotype; mixed effects model; family relatedness

Background
With the decreasing cost and increasing throughput of next-generation sequencing, the number of accessions that can be used for genome wide association study (GWAS) is increasing [1, 2, 3]. Using such large sequencing data, GWAS is now widely used not only in human but also in plant and animal genetics and breeding, and has identified novel genes related to important agronomic traits [4, 5, 6]. One example of large next-generation sequencing data is that of “the 3,000 rice genomes
project” as used in this study [7, 8], data from which are available in the “Rice SNP-Seek Database” [9, 10, 11]. GWAS results using these data have already been reported [12].

Despite the enhancement of such public data, however, the conventional GWAS method still faces obstacles in the detection of unknown candidate genes. One common example is its difficulty in detecting rare alleles or rare variants. One problem caused by rare variants is that the non-causal markers that have a strong linkage disequilibrium (LD) with one causal rare variant indicate a higher detection power than the true causal rare variant, which may interfere with the detection of the true causal variant. This phenomenon is known as “synthetic association”, and often happens when the minor allele frequency (MAF) of the non-causal marker is higher than that of the true rare variant [13]. This problem is closely related to the complex gene compositions comprising multiple alleles such as haplotypes because genes related to important agronomic traits often consist of multiple rare alleles, and this is why haplotypes are hard to detect using GWAS [14].

Several methods have been proposed to solve this problem. The sequence kernel association test (SKAT) is one of the methods used to detect rare variants, and has been used mainly in human genomics [15]. The SKAT employs a single nucleotide polymorphism (SNP) set approach, which tests multiple SNPs in each SNP-set at the same time. The SKAT evaluates the significance of the variance explained by a SNP-set of interest as a random effect using a mixed effect model approach [16, 17]. The fatal drawback of the original SKAT is that the model does not take the effects of family relatedness into account as a random effect, which results in it generating false positives for GWAS in materials with a strong population structure or family relatedness, such as a world collection of rice germplasm as used in this study. In addition, several methods were also proposed to overcome another SKAT drawback, that a weighting scheme of the SKAT for rare and common variants can lead to loss of power of common variants, but their models also do not include the term for correcting the confounding effects of family relatedness [18, 19].

To solve the fatal drawbacks of the original SKAT, several methods whose models include the term of family relatedness as random effects to control false positives have been previously proposed [20, 21, 22]. From a statistical point of view, these methods usually perform the score test, which is the computationally efficient method because it requires variance component estimation only for the null model. In terms of the detection power, however, the score test is not necessarily the best method for testing the random effects in the mixed effects model. The likelihood-ratio (LR) test is another candidate used to test the variance of a SNP-set of interest, and several methods have been proposed that use the LR test for SNP-set GWAS in family samples [23, 24]. In particular, Lippert et al. implemented a computationally efficient SNP-set GWAS method using the LR test, and reported that the LR test showed greater power than the score test [24]. Despite being such an efficient method, however, Lippert et al. mainly used the linear kernel for the covariance matrix of each SNP-set, and therefore there were some restrictions on its use.

However, only a few methods for haplotype-based GWAS have so far been proposed. In plant genomics, Yano et al. performed a haplotype-based GWAS by testing
the effects of haplotypes while regarding dummy variables of haplotype groups as fixed effects, and found new candidate genes related to heading date for rice [25]. Other approaches have been proposed in animal genomics, which estimated ancestral haplotype effects by regarding them as random effects whose covariance matrix is a $k \times k$ identity matrix where $k$ is the number of ancestral haplotypes [26, 27]. However, these conventional haplotype-based GWAS methods require haplotype information a priori, and it is not so easy to apply these methods at the genome-wide level.

In this study, we extended the multi-kernel mixed effects model more generally to take family relatedness into account, while enabling computational speed-up for some limited cases, and developed a novel SNP-set GWAS approach named RAINBOW (Reliable Association INference By Optimizing Weights). We also estimated haplotype blocks from genome-wide marker genotype data, and used them as SNP-sets for analysis with RAINBOW to enable haplotype-based GWAS without prior haplotype information.

**Materials and Methods**

**Methods for RAINBOW**

In this subsection, we describe the basic idea of RAINBOW.

**RAINBOW model**

The RAINBOW model can be written as

$$y = X\beta + Z_c u_c + Z_r_i u_r_i + \epsilon,$$

where $y$ is a $n \times 1$ vector of phenotypic values, $X\beta$ is a $n \times 1$ vector of fixed effects including an intercept, a term to correct the population structure and other covariates, $Z_c u_c$ and $Z_r_i u_r_i$ are $n \times 1$ vectors of random effects, and $\epsilon$ is a $n \times 1$ vector of residual errors. Here $\beta$ is a $p \times 1$ vector of fixed effects, where $p$ is the number of fixed effects. $u_c$ and $u_r_i$ are $m_c \times 1$ and $m_r_i \times 1$ vector of genotypic values respectively, where $m_c$ is the number of genotypes for additive polygenetic effects and $m_r_i$ is the number of genotypes for $i$th SNP-set of interest. $X$, $Z_c$ and $Z_r_i$ are $n \times p$, $n \times m_c$ and $n \times m_r_i$ design matrices that correspond to $\beta$, $u_c$ and $u_r_i$ respectively. As the following formula Eq. (2), we assume that the polygenetic effect $u_c$ follows the multivariate normal distribution whose variance-covariance matrix is proportional to the additive numerator relationship matrix $K_c$.

$$u_c \sim \text{MVN} (0, K_c \sigma_c^2),$$

where $\sigma_c^2$ is the additive genetic variance to be estimated by solving the mixed effect model Eq. (1), and here $m_c \times m_c$ matrix $K_c = A$, where $A$ is the known additive genetic relationship matrix estimated from marker genotype data $W_c$ [28].

We also assume that the random effects from $i$th SNP-set of interest $u_r_i$ follows the multivariate normal distribution whose variance-covariance matrix is proportional to the Gram matrix $K_{r_i}$. 

Hamazaki and Iwata
\[ u_i \sim \text{MVN}(0, K_i, \sigma_i^2) , \] (3)

where \( \sigma_i^2 \) is the genetic variance for the \( i \)th SNP-set to be estimated by solving the mixed effect model Eq. (1) and \( K_i \) is the known \( m_r \times m_r \) Gram matrix estimated from marker genotype data \( W_r \) belonging to the \( i \)th SNP-set. We offer a linear, an exponential and a gaussian kernel for the Gram matrix \( K_i \), and faster computation can be realized for the linear kernel case (Supplementary Note in Additional file 1) [24].

Finally, the residual term is assumed to identically and independently follow a normal distribution as shown in the following equation.

\[ \epsilon \sim \text{MVN}(0, I_n, \sigma_e^2) , \] (4)

where \( I_n \) is a \( n \times n \) identity matrix and \( \sigma_e^2 \) is estimated by solving the mixed model.

**Estimation of variance components**

The variance components were estimated by maximum-likelihood (ML) and restricted maximum-likelihood (REML). Here we explain how to obtain ML and REML estimates of Eq. (1) for the general \( K_i \).

First we estimated the weights (we define \( w_c \) and \( w_r \)) between the genetic variances (\( \sigma_c^2 \) and \( \sigma_r^2 \)) by the following algorithm.

1 Setting initial parameters for \( w_c \) and \( w_r \):

\[ w_c = w_r = \frac{1}{2} \] (5)

2 Computing the following \( n \times n \) matrix \( K_s \):

\[ K_s = Z_c K_c Z_c^T w_c + Z_r K_r Z_r^T w_r \] (6)

3 Solving the following single-kernel linear mixed model (LMM) by using EMMA (efficient mixed model association) or GEMMA (genome wide efficient mixed model association) [29, 30].

\[ y = X \beta + u_s + \epsilon, \] (7)

where

\[ u_s \sim \text{MVN}(0, K_s, \sigma_s^2) \] (8)
4 Computing the full log likelihood \((l_F)\) or the restricted log likelihood \((l_R)\) of Eq. (7) by using estimated parameters; \(\hat{\beta}, \hat{\sigma}_s^2\) and \(\hat{\sigma}_e^2\):

\[
l_F(y; \hat{\beta}, \hat{\sigma}_s, \hat{\sigma}_e) = \frac{1}{2} \left[ -n \log(2\pi \hat{\sigma}_s^2) - \log|\hat{H}| \right. \\
\left. - \frac{1}{\hat{\sigma}_s^2} (y - X\hat{\beta})^T \hat{H}^{-1} (y - X\hat{\beta}) \right],
\]

\( (9) \)

\[
l_R(y; \hat{\sigma}_s, \hat{\sigma}_e) = l_F(y; \hat{\beta}, \hat{\sigma}_s, \hat{\sigma}_e) \\
+ \frac{1}{2} \left[ p \log(2\pi \hat{\sigma}_e^2) + \log|X^TX| \right. \\
\left. - \log|X^T \hat{H}^{-1} X| \right]
\]

\( (10) \)

Here \(\hat{H}\) is

\[
\hat{H} = \frac{\hat{\Sigma}}{\hat{\sigma}_s^2} = K_s + \hat{\delta}I_n
\]

where \(\hat{\Sigma}\) is a phenotypic variance-covariance matrix and \(\hat{\delta} = \frac{\hat{\sigma}_e^2}{\hat{\sigma}_s^2}\).

5 Optimizing \(w_c\) and \(w_r\) over maximization of the full/restricted log likelihood by using L-BFGS optimization method through repeating step 2-4 [31].

After estimating the weights \(w_c\) and \(w_r\), we estimated the variance components \((\sigma^2_s\) and \(\sigma^2_e)\) of the model Eq. (7), (8) by EMMA / GEMMA using \(\hat{w}_c\) and \(\hat{w}_r\). Then we obtained \(\hat{\sigma}_c^2 = \hat{w}_c \hat{\sigma}_s^2\) and \(\hat{\sigma}_r^2 = \hat{w}_r \hat{\sigma}_e^2\).

**Likelihood ratio test for GWAS**

To test the significance of each SNP-set, we performed the LR test of whether \(\sigma^2_r = 0\) or not. As a null hypothesis, the following model, which does not include the term of SNP-set effects was assumed.

\[
y = X\beta + Z_u u_c + \epsilon
\]

\( (12) \)

In contrast, as an alternative hypothesis model, the multi-kernel linear mixed model (MKLMM) of Eq. (1) was assumed. Therefore, we computed the following deviance after the estimation of variance components for each SNP-set.

\[
D = 2 \times (\hat{l}_R,_{\text{model}} - \hat{l}_R,_{\text{null}}),
\]

\( (13) \)

where \(\hat{l}_R,_{\text{model}}\) is the maximum of the restricted log likelihood for the model of Eq. (1) and \(\hat{l}_R,_{\text{null}}\) is the maximum of the restricted log likelihood for the model of Eq. (12).

Finally, we tested the significance of \(\sigma^2_r\) and calculate the \(p\)-value by assuming
that the deviance in Eq. (13) followed the mixture of two chi-square distributions with different degrees of freedom.

\[ D \sim \pi_0 \chi_0^2 + (1 - \pi_0) \chi_1^2, \]  

where \( \pi_0 \) is the mixture parameter and here we used \( \pi_0 = 1/2 \).

Materials and simulations

Genotype data

In this study, 414 accessions of *Oryza sativa* subsp. *indica* were collected from “the 3,000 rice genomes project” (Table S1 in Additional file 2) [7]. We used a marker genotype consisting of core SNPs defined by the Rice SNP-Seek Database as “404k CoreSNP Dataset”. Imputations were imputed using Beagle version 5.0 [32, 33]. We analyzed only bi-allelic sites over all accessions with a MAF \( \leq 0.025 \) by using VCFtools version 0.1.15 [34]. In the following analysis, genotypes are represented as -1 (homozygous of the reference allele), 1 (homozygous of the alternative allele) or 0 (heterozygous of the reference and alternative alleles). As a result of this data processing, marker genotypes with 112,630 SNPs were used for the following simulation study.

Estimation of haplotype block

To perform haplotype-based GWAS by regarding each haplotype block as a SNP-set, haplotype blocks were estimated from marker genotype data by using PLINK 1.9 [35, 36, 37]. As a result of estimation, we obtained 15,275 haplotype blocks consisting of 78,237 SNPs.

Simulation of phenotype data

We considered two scenarios to validate our novel haplotype-based GWAS approach. In both models, phenotypic values were simulated as follows.

\[ y = X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + Zu + e, \]  

where \( y \) is the vector of simulated phenotypic values of 414 accessions, \( X_1, X_2 \) and \( X_3 \) correspond to three quantitative trait nucleotides (QTNs) scored as -1, 0 or 1 (hereinafter, referred to as “QTN1”, “QTN2” and “QTN3” respectively), \( \beta_1, \beta_2 \) and \( \beta_3 \) are scalars representing the effects of the three QTNs, \( u \) is the vector of polygenetic effects, and \( e \) is the vector of the residuals.

Here, QTN1 and QTN2 were randomly selected from all genome-wide SNPs to satisfy that they belonged to the same haplotype block that harbored more than 4 SNPs. QTN3 was randomly selected from all the SNPs. We assumed that the effects of QTN1 and QTN2 had a variance 4 times greater than that of the effects of QTN3 to mainly check the detection power for the haplotype block.

The polygenetic effects \( u \) in Eq. (16) were sampled from the multivariate normal distribution whose variance-covariance matrix was the additive genetic relationship...
matrix \( A \), and were normalized so that their variance was equal to that of three QTN effects.

\[
\mathbf{u} \sim \text{MVN}(\mathbf{0}, G) \quad (16)
\]

\( G = A\sigma^2_A \) is the genetic covariance matrix, and the additive genetic variance \( \sigma^2_A \) is automatically determined from the relationship with the heritability. In this study, the additive numerator relationship matrix was estimated based on the marker genotype of 112,630 SNPs using the “A.mat” function of the R package “rrBLUP” version 4.6 [38].

The residual \( e \) in Eq. (17) was sampled identically and independently from a normal distribution, and then was normalized so that the narrow-sense heritability was equal to 0.6.

\[
\mathbf{e} \sim \text{MVN}(\mathbf{0}, I\sigma^2_e), \quad (17)
\]

where \( I \) is an identity matrix, and \( \sigma^2_e \) is the residual variance determined so that the heritability is equal to 0.6.

The difference between two scenarios is based on the directions of the two QTN effects \( \beta_1 \) and \( \beta_2 \). Scenario 1 assumed that the directions of two effects were identical. That is,

\[
\beta_1 = \begin{cases} 
\beta_2 & (\rho_{12} \geq 0) \\
-\beta_2 & (\rho_{12} < 0)
\end{cases}, \quad (18)
\]

where \( \rho_{12} \) is the correlation between \( X_1 \) and \( X_2 \). We call this model as “coupling”.

Conversely, scenario 2 assumed that the directions of the two effects were opposite. That is,

\[
\beta_1 = \begin{cases} 
-\beta_2 & (\rho_{12} \geq 0) \\
\beta_2 & (\rho_{12} < 0)
\end{cases}, \quad (19)
\]

We call this scenario 2 as “repulsion”.

**Evaluation of RAINBOW**

*Comparison of four methods*

To validate our novel approach, we compared the following four methods: Single-SNP GWAS [39], haplotype-based GWAS introduced by Yano et al. (hereinafter, referred to as “HGF”) [25], the SKAT [15] and our novel approach, RAINBOW. For all methods, to account for the population structure, the two eigen vectors (which correspond to the top two eigen values) of the additive genetic relationship matrix were included in the model as fixed effects.
Single-SNP GWAS  The single-SNP GWAS model was similar to Eq. (12), but this method regarded each SNP as a fixed effect and tested the significance of each SNP effect one by one [39]. In this study, we used the method called “P3D”, which first estimated $\sigma_c^2$ and $\sigma_e^2$ by solving the mixed model without SNP effects (Eq. (12)) by REML, and then tested each marker using the information of the estimated variance $\tilde{\sigma}_c^2$ and $\tilde{\sigma}_e^2$ [29, 40, 41]. We used the “GWAS” function of the R package “rrBLUP” version 4.6 [38].

HGF method  HGF tests the significance of each haplotype block effect. In this method, all accessions were divided into $k_r$ groups based on the haplotypes. The haplotype groups were then represented by dummy variables. The significance of the haplotype effect was tested by regarding a matrix of dummy variables as fixed effects [25]. In this study, the k-medoids method (the “pam” function of the R package “cluster” version 2.0.6) [42] or UPGMA (unweighted pair group method with arithmetic mean, “hclust” and “cutree” function of the R package “stats”) [43] were employed to group haplotypes. We set the number of groups $k_r$ as 2, 3 or 4. Therefore, we applied 6 different HGF models in total. To apply HGF models, we modified the R package “rrBLUP” version 4.6 to enable the use of a matrix of dummy variables as fixed effects [38].

SKAT  The SKAT is a SNP-set GWAS method, and is widely used in human genetics [15]. This method was invented to detect rare alleles, but the SKAT model does not include the random effects term that accounts for family relatedness, which may cause false positives for GWAS in materials with a strong population structure or family relatedness. In this study, we used the “SKAT” function of the R package “SKAT” version 1.3.2.1 and performed SNP-set GWAS by regarding each haplotype block as each SNP-set.

RAINBOW  The methods for RAINBOW were described above. In this study, we performed haplotype-based GWAS for RAINBOW by regarding each haplotype block as each SNP-set. We used the linear kernel of each haplotype block for the Gram matrix $K_r$.

Evaluation of the simulation results
The value of $-\log_{10}(p)$ of each marker or haplotype block was calculated by the four GWAS methods 100 times for the two simulated scenarios, coupling and repulsion. In this study, the following summary statistics were used to evaluate the simulation results.

$-\log_{10}(p)$ and adjusted $-\log_{10}(p)$  The first summary statistic is $-\log_{10}(p)$ of each causal SNP or haplotype block itself. For haplotype-based GWAS methods, HGF, SKAT and RAINBOW, the significance of $\beta_1$ and $\beta_2$ was represented by $-\log_{10}(p)$ of the causal haplotype block to which $X_1$ and $X_2$ belong. In the single-SNP GWAS method, the $-\log_{10}(p)$ of $\beta_1$ and $\beta_2$ were calculated separately, even though these SNPs were in the same haplotype. To compare the single-SNP GWAS method with the haplotype-based GWAS methods, the $-\log_{10}(p)$ values were averaged over $\beta_1$.
and $\beta_2$.

As some of these methods showed the results of inflated $-\log_{10}(p)$, we defined the following summary statistic to evaluate the degree of inflation.

$$inflator = \frac{1}{L} \sum_{l=1}^{L} (-\log_{10}(p_{false,l})),$$

(20)

where $p_{false,l}$ is the $l^{th}$ of $p$-values for false positives arranged in decreasing order. In this study, $L$ was set as 10. Then we adjusted $-\log_{10}(p)$ of the causal by using the inflator (Eq. (20)) as follows.

$$adjusted \ -\log_{10}(p) = -\log_{10}(p) - inflator$$

(21)

Here, we calculated each summary statistic in two ways. The first method is to calculate each summary statistic by directly using $-\log_{10}(p)$ of each causal SNP / haplotype block. The other method is to calculate the summary statistics by regarding multiple SNPs or haplotype blocks within the extent of the LD as one set. In this study, we defined SNPs or haplotype blocks that satisfy the condition that they are within 300 kb from the focused SNP or haplotype block and the condition that their square of the correlation coefficients with the focused SNP or haplotype block are 0.35 or more as one set considering the LD. The highest value of $-\log_{10}(p)$ in the LD region was assumed to represent the values of the SNPs or haplotype blocks within the extent of the LD.

Recall, Precision and F-measure We calculated the recall, precision and F-measure means as other summary statistics to evaluate the GWAS results. These summary statistics were calculated from the numbers of SNPs or haplotype blocks that were true positives, false positives, false negatives and true negatives. Here, we regarded a SNP or haplotype block as “positive” when that SNP or haplotype block exceeded the threshold. In this study, the value of $-\log_{10}(p)$ so that the FDR (false discovery rate) was 0.01 was set as the threshold by using the Benjamini-Hochberg method [44, 45]. In addition, these three summary statistics, recall, precision and F-measure, were calculated by assuming that the highest value of $-\log_{10}(p)$ in the LD region represented the values of the SNPs or haplotype blocks within the extent of the LD.

Therefore, recall represents the proportion of causals detected by GWAS. In contrast, precision represents the ratio of the detected SNPs or haplotype blocks that were causals. Finally, F-measure was calculated as the harmonic mean of the recall and the precision, which evaluates the GWAS results comprehensively. The greater these three summary statistics, the better the results of GWAS are. Here we simply took the average of each summary statistic from all the 100 simulation results.

AUC for regions around causals We calculated the mean of the AUC (area under the curve) for regions around the causals as a summary statistic. AUC refers to the area under the ROC (receiver operating characteristic) curve obtained by plotting the false positive rate on the horizontal axis and the true positive rate on the
vertical axis when the threshold is varied. In this study, the AUC was calculated for the SNPs or haplotype blocks near the causal SNP / haplotype block (QTN1 and QTN2). In other words, the non-causal markers that had a strong LD with the causal SNP / haplotype block were regarded as false positives under this summary statistic. Therefore, this summary statistic indicates the extent to which the causal itself can be detected by GWAS without relying on the LD. Here, when taking the average of the AUCs obtained from the simulation results, two methods were used, either using all the 100 results or only using the results whose QTN1 and QTN2 were “detected”. Here, QTN1 and QTN2 were regarded as “detected” if adjusted $-\log_{10}(p) \geq 1.5$ for each method.

Results

The detection power of four methods

The detection power of the four methods was evaluated by the value of $-\log_{10}(p)$ and adjusted $-\log_{10}(p)$ of QTN1 and QTN2 for the two models, coupling and repulsion (Fig. 1). RAINBOW outperformed the other methods when the significance was evaluated by the causal itself (Fig. 1a,c,e,g). However, when the significance was evaluated by the highest values of SNPs or haplotypes within the extent of the LD, other methods, e.g., HGF (k = 2, k-medoids method), showed a greater detection power than RAINBOW (Fig. 1b). When the detection power was evaluated by taking the extent of inflation into account, RAINBOW showed as great a power as HGF (k = 2, 3) even if the significance was evaluated by the unit of the LD block (Fig. 1d). Moreover, although the detection power of all the GWAS methods for the repulsion scenario was less than that for the coupling scenario, the tendency for RAINBOW to outperform the other methods was clearer for the repulsion scenario than the coupling scenario (Fig. 1). Finally, as compared with the other haplotype-based GWAS methods, RAINBOW showed smaller variation among iterations, indicating that the causal variants can be stably detected (Fig. 1).

The detection power for QTN3 was also evaluated. The single-SNP GWAS method showed a greater power than RAINBOW when evaluated by $-\log_{10}(p)$ (Fig. S1a,b,e,f in Additional file 3). However, if the detection power was evaluated by adjusted $-\log_{10}(p)$, RAINBOW showed as great a power as single-SNP GWAS (Fig. S1c,d,g,h). Contrary to the results for QTN1 and QTN2, the detection power of all the GWAS methods for the repulsion scenario was greater than for the coupling scenario (Fig. S1).

Recall, Precision and F-measure

The characteristics of each GWAS method were evaluated by the recall, precision and F-measure means (Fig. 2). For the mean of recall, the HGF methods and SKAT showed higher values than RAINBOW and single-SNP GWAS for both scenarios (Fig. 2). However, the haplotype-based GWAS methods other than RAINBOW showed low precision. That is, these methods may cause too many false positives. In contrast, RAINBOW and single-SNP GWAS showed higher precision than the remainders, and RAINBOW showed the highest precision among all the scenarios. From the results for the three summary statistics, RAINBOW also showed the
highest value for F-measure among the methods. In particular, for the repulsion scenario, the recall of RAINBOW was also higher than that of single-SNP GWAS, which resulted in the large difference of F-measure between these two methods (Fig. 2b). A similar tendency was also confirmed when changing the criterion of how to determine the threshold for the Bonferroni’s correction for the significance level $\alpha = 0.01$ (Fig. S2 in Additional file 4).

To compare the three summary statistics of the two scenarios in more detail, these values for each QTN were also calculated (Fig. S3 in Additional file 5). For both scenarios, RAINBOW showed the highest recall for QTN1 and QTN2 among the methods (Fig. S3ab). In particular, RAINBOW outperformed the other methods in all summary statistics for QTN1 and QTN2 for the repulsion scenario. However, it showed lower recall for QTN3 than the other methods (Fig. S3cd). In particular, the recall of RAINBOW for QTN3 was 0 for the coupling scenario (Fig. S3c). In addition, the three summary statistics of QTN3 for the repulsion scenario were greater than those for the coupling scenario in almost all the methods (Fig. S3cd). Regarding these results, a similar trend was confirmed even when changing the criterion of how to determine the threshold for the Bonferroni’s correction for the significance level $\alpha = 0.01$ (Fig. S4 in Additional file 6).

**AUC for regions around causals**

To evaluate how the causal itself can be detected by GWAS without relying on the LD, the AUC means for regions around the causals (QTN1 and QTN2) were compared (Fig. 3). The mean of AUC was almost the same when using all simulation results or using only the cases in which QTN1 and QTN2 were detected, although the value of the latter was slightly larger than that of the former in some methods. The results show that RAINBOW outperformed the other methods in both models (Fig. 3). Especially, the AUC mean of the single-SNP GWAS method in the repulsion scenario was much smaller than that in the coupling scenario, while RAINBOW was able to maintain a high AUC even in the repulsion scenario (Fig. 3b).

**Examples in the repulsion scenario**

Of the 100 simulations for the repulsion scenario, there were 7 cases in which QTN1 and QTN2 were detected only by RAINBOW. These cases were selected to satisfy three conditions that $\text{adjusted} - \log_{10}(p_R) \geq 1.5$, $\text{adjusted} - \log_{10}(p_O) \leq 1.2$ and the recall for QTN1 and QTN2 equals to 1. Here, $p_R$ represents the $p$-value of RAINBOW and $p_O$ represents the $p$-value of all the other methods. Although the same analysis was done for the other methods, no method satisfied the three conditions described above. One example of these cases (iteration 48) was shown by comparing the four GWAS methods; RAINBOW, single-SNP GWAS, HGF (the number of groups is 2, the grouping method is UPGMA) and SKAT (Fig. 4). The Manhattan plot shows that RAINBOW succeeded in detecting the causal haplotype block (of QTN1 and QTN2) that was not detected by the other methods. Although both QTN1 and QTN2 were also detected by the single-SNP method in one case (iteration 85), the same trend as the results for iteration 48 was seen for the remaining five results (Fig. S5 in Additional file 7).
Discussion

As shown in Results section, when \(- \log_{10}(p)\) was evaluated by the LD block unit for the coupling scenario, RAINBOW did not necessarily outperform the other methods. However, if we considered the inflation level of each result and evaluated the results with the adjusted \(- \log_{10}(p)\), RAINBOW showed as great a detection power as other methods (Fig. 1d), which means RAINBOW succeeded in controlling false positives compared with other haplotype-based GWAS methods. This can also be seen from the fact that the precision of RAINBOW was much higher than the other GWAS methods including single-SNP GWAS (Fig. 2).

Moreover, \(- \log_{10}(p)\) of RAINBOW was the highest when evaluated by that of the causal SNP/haplotype block itself, which implies that RAINBOW can detect causal haplotype blocks themselves without relying on the LD beyond the scope of the haplotype block. This can also be confirmed by the results that showed that the AUC for the regions around the causal was larger in RAINBOW than in any other methods (Fig. 3).

In addition, for the repulsion scenario, RAINBOW succeeded in detecting causal haplotype blocks that were not able to be detected by any other methods including single-SNP GWAS (Fig. 4). This result affected other results that RAINBOW outperformed the other methods especially when evaluated by the detection power, recall, precision and F-measure in the repulsion scenario. This fact suggests that RAINBOW is good for detecting the causal haplotype block with multiple causal variants. For example, RAINBOW can be applied to the detection of genes that have more than one variant. Therefore, for future analysis, RAINBOW can be used for gene-set GWAS (which regards one gene as one SNP-set) by using gene annotation information.

The only drawback of RAINBOW is that the detection power for the causal with small effects (QTN3) was not so high (Fig. S3cd). The drawback may be related to the fact that RAINBOW succeeded in detecting QTN1 and QTN2 well. In other words, RAINBOW cannot account for the loci of large effects well when testing other loci, and the loci of relatively small effects may be concealed by these loci of large effects. This drawback, however, can be easily resolved by using methods that condition the loci of large effects, such as composite interval mapping for QTL analysis [46, 47] or a multi-locus mixed model for GWAS [48]. For future analysis, we will implement this function to condition the loci of large effects when testing other loci of small effects.

Conclusion

We proposed a new method named “RAINBOW”, which is a haplotype-based GWAS method using a novel SNP-set approach, and showed that it outperformed the conventional single-SNP GWAS method, the conventional SNP-set method, and the conventional haplotype-based method. Our method is especially superior in controlling false positives, detecting the causal itself, and detecting causal variants under complex conditions, e.g., allelic effects that are in a repulsion phase (i.e., the opposite direction of effects of closely linked causal variants). By using the SNP-set approach as our method, we expect that detecting not only rare alleles but also genes with complex mechanisms, such as genes with multiple causal variants, can be realized.
List of abbreviations

AUC: Area under the curve  
EMMA: Efficient mixed model association  
GEMMA: Genome wide efficient mixed model association  
GWAS: Genome wide association study  
IRRI: International rice research institute  
LD: Linkage disequilibrium  
LMM: Linear mixed model  
LRT: Likelihood-ratio test  
MAF: Minor allele frequency  
MKLMM: Multi-kernel linear mixed model  
ML: Maximum likelihood  
P3D: Population parameters previously determined  
QTL: Quantitative trait loci  
QTN: Quantitative trait nucleotide  
RAINBOW: Reliable association inference by optimizing weights  
REML: Restricted maximum likelihood  
ROC: Receiver operating characteristic  
SKAT: Sequence kernel association test  
SNP: Single nucleotide polymorphism  
UPGMA: Unweighted pair group method with arithmetic mean

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and material
The datasets and scripts generated and/or analysed during the current study are available in the "KosukeHamazaki/HGRAINBOW" repository in the GitHub, https://github.com/KosukeHamazaki/HGRAINBOW.

Competing interests
The authors declare that they have no competing interests.

Funding
This work was supported by JST CREST Grant Number JPMJCR1602, Japan.

Author's contributions
KH developed the method of the novel haplotype-based GWAS with the SNP-set approach, RAINBOW, conducted all statistical analyses, and drafted the manuscript. HI was involved in the conception and design of the study, provided administrative support, and supervised the study. Both authors have read and approved the final manuscript.

Acknowledgements
We are grateful to Ryokei Tanaka for fruitful discussions, Motoyuki Ishimori and Goshi Sasaki for debugging the package, and Ryusuke Hamazaki for naming the package, RAINBOW.

References
1. Metzker ML. Sequencing technologies the next generation. Nat Rev Genet. 2010; 11(1): 31–46.
2. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. The next-generation sequencing revolution and its impact on genomics. Cell. 2013; 155(1): 27–38.
3. Ott J, Wang J, Leal SM. Genetic linkage analysis in the age of whole-genome sequencing. Nat Rev Genet. 2015; 16(5): 275–284.
4. Atwell S, Huang YS, Vilhjálmsdóttir BJ, Willems G, Horton M, Li Y, et al. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature. 2010; 465(7298): 627–631.
5. Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, et al. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet. 2010; 42(11): 961–967.
6. Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. Plant Methods. 2013; 9(1): 29.
7. Li JY, Wang J, Zeigler RS. The 3,000 rice genomes project: New opportunities and challenges for future rice research. GigaScience. 2014; 3(1): 1–3.
8. Zheng T, Fuentes RR, Zhang F, Mansuetu L, Copetti D, Sanciangco M, et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature. 2018; 557: 43–49.
9. Alexandrov N, Tai S, Wang W, Mansuetu L, Palis K, Fuentes RR, et al. SNP-Seek database of SNPs derived from 3,000 rice genomes. Nucleic Acids Res. 2015; 43(D1): 1023–1027.
10. Mansuetu L, Fuentes RR, Chebotarov D, Borja FN, Detras J, Abriol-Santos JM, et al. SNP-Seek II: A resource for allele mining and analysis of big genomic data in Oryza sativa. Curr Plant Biol. 2016; 7:8: 16–25.
11. Mansuetu L, Fuentes RR, Borja FN, Detras J, Abriol-Santos JM, Chebotarov D, et al. Rice SNP-Seek database update: New SNPs, indels, and queries. Nucleic Acids Res. 2017; 45(D1): 1075–1081.
12. Misra G, Badoni S, Anaclerio R, Graner A, Alexandrov N, Sreenivasulu N. Whole genome sequencing-based association study to unravel genetic architecture of cooked grain width and length traits in rice. Nat Sci Reports. 2017; 7(1): 12478.
13. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare Variants Create Synthetic Genome-Wide Associations. PLoS Biol. 2010; 8(1).

14. Stram D. Design, Analysis, and Interpretation of Genome-Wide Association Scans. Springer, Heidelberg, New York 2014;

15. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011; 89(4): 82–93.

16. Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. Biometrics. 2007; 63(4): 1079–1088.

17. Liu D, Ghosh D, Lin X. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. BMC Bioinformatics. 2008; 9: 1–11.

18. Sha Q, Wang X, Wang X, Zhang S. Detecting Association of Rare and Common Variants by Testing an Optimally Weighted Combination of Variants. Genet Epidemiol. 2012; 36(6): 561–571.

19. Ionita-Laza I, Lee S, Makarov V, Burbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. Am J Hum Genet. 2013; 92(6): 841–853.

20. Schifano ED, Epstein MP, Bielak LF, Jhun MA, Kardia SLR, Peyser PA, et al. SNP Set Association Analysis for Familial Data. Genet Epidemiol. 2012; 36(8): 797–810.

21. Chen H, Meigs JB, Dupuis J. Sequence Kernel Association Test for Quantitative Traits in Family Samples. Genet Epidemiol.. 2013; 37(2): 196–204.

22. Oualkacha K, Dastani Z, Li R, Cingolani PE, Spector TD, Hammond CJ, et al. Adjusted Sequence Kernel Association Mapping that accounts for sample structure in genome-wide association studies. Nat Genet. 2010; 42(4): 355–360.

23. Zhang Z, Guillaume F, Sartelet A, Charlier C, Georges M, Farnir F, et al. Ancestral haplotype-based association mapping with generalized linear mixed models accounting for stratification. Bioinformatics. 2012; 28(19): 2467–2473.

24. Lippert C, Xiang J, Horta D, Widmer C, Kadie C, Heckerman D, et al. Greater power and computational efficiency for kernel-based association testing of sets of genetic variants. Bioinformatics. 2014; 30(22): 3206–3214.

25. Yano K, Yamamoto E, Aya K, Takeuchi H, Lo PC, Hu L, et al. Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. Nat Genet. 2016; 48(8): 927–934.

26. Druet T, Georges M. A hidden Markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. Genetics. 2010; 184(3): 789–798.

27. Zhang Z, Guillaume F, Sartelet A, Charlier C, Georges M, Farnir F, et al. Ancestral haplotype-based association mapping with generalized linear mixed models accounting for stratification. Bioinformatics. 2012; 28(19): 2467–2473.

28. Endelman JB, Jannink JL. Shrinkage Estimation of the Realized Relationship Matrix. G3 (Bethesda). 2012; 2(11): 1405–1413.

29. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, et al. Efficient Control of Population Structure in Model Organism Association Mapping. Genetics. 2008; 178(3): 1709–1723.

30. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012; 44(7): 821–824.

31. Byrd R, Lu P, Nocedal J, Zhu C. A limited memory algorithm for bound constrained optimization. SIAM Journal on Scientific Computing. 1995; 16(5): 1190–1208.

32. Browning SR, Browning BL. 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(5): 1084–1097.

33. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation Reference Panels. Am J Hum Genet. 2018; 103(3): 338–348.

34. Danek P, Auvot A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27(15): 2156–2158.

35. Purcell S, Neale B, Todd-Brown KE, Thomas L, Ferreira MA, Bender D, et al. PLINK 1.9. 2018;.

36. Gaunt TR, Rodríguez S, Day INM. Cubic exact solutions for the estimation of pairwise haplotype frequencies: Implications for linkage disequilibrium analyses and a web tool ‘CubeX’. BMC Bioinformatics. 2007; 8: 1–9.

37. Taliun D, Gamper J, Pattaro C. Efficient haplotype block recognition of very long and dense genetic sequences. BMC Bioinformatics. 2014; 15(1): 1–18.

38. Endelman JB. Ridge Regression and Other Kernels for Genomic Selection with R Package rBLUP. Plant Genome. 2011; 4(3): 250.

39. Yu J, Pressoir G, Briggs WH, Voh B, Yamazaki M, Doebley JF, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 2006; 38(2): 203–8.

40. Kang HM, Sui JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010; 42(4): 348–354.

41. Zhang Z, Eeck S, Liu CQ, Todhunter RJ, Tiwari HK, Gore MA, et al. Mixed linear model approach adapted for genome-wide association studies. Nat Genet. 2010; 42(4): 355–360.

42. Maechler M, Rousseau P, Struyf A, Hubert M, Hornik K. Cluster: Cluster Analysis Basics and Extensions. 2018; R package version 2.0.7-1 | For new features, see the ‘Changelog’ file (in the package source).

43. Stram D. Design, Analysis, and Interpretation of Genome-Wide Association Scans. Springer, Heidelberg, New York 2014;

44. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing Author (s) Yoav Benjamini and Yosef Hochberg Source: Journal of the Royal Statistical Society. Series B ( Methodological ), Vol. 57 , No. 1 Published by : Wi. J R Stat Soc Ser B. 1995; 57(1): 289–300.

45. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci. 2003; 100(16): 9440–9445.
46. Zeng ZB. Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. Proc Natl Acad Sci. 1993; 90(23): 10972–10976.
47. Zeng ZB. Precision Mapping of Quantitative Trait Loci. Genetics. 1994; 136: 1457–1468.
48. Segura V, Vilhjalmsson BJ, Platt A, Korte A, Seren U, Long Q, et al. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet. 2012; 44(7): 825–830.

Figures

**Figure 1** The detection power of each GWAS method. Boxplot of the detection power evaluated by \(- \log_{10}(p)\) and adjusted \(- \log_{10}(p)\). a-d: The results for the “coupling” scenario. e-h: The results for the “repulsion” scenario. a,b,e,f: The results evaluated by \(- \log_{10}(p)\) with the scale on the vertical axis aligned in these four figures. c,d,g,h: The results evaluated by adjusted \(- \log_{10}(p)\) with the scale on the vertical axis aligned in these four figures. a,c,e,g: The results evaluated by the unit of the causal SNP or haplotype block itself. b,d,f,h: The results evaluated by the unit of the regions within the extent of LD. The abbreviation of each method is as follows. R: RAINBOW. SS: Single-SNP GWAS. H2k-H4p: HGF methods. The numbers in the method names correspond to the numbers of the groups they assume. The last letters of the methods are “k” or “p”. “k” corresponds to the k-medoids method and “p” corresponds to UPGMA method for the grouping method. SK: SKAT.

**Figure 2** Recall, Precision and F-measure of each GWAS method. Bar plot of the mean of each summary statistic for 100 simulation results. The red bars show the results for recall, the green bars show the results for precision, and the blue bars show the results for F-measure. a: Results for the coupling scenario. b: Results for the repulsion scenario. The abbreviations of each method are the same as those of Fig. 1.

Additional Files

Additional file 1 — Supplementary Note
Supplementary Note for additional RAINBOW methods. A faster computational method for the linear kernel and effective testing method for dominance and epistatic effects are mainly described. (PDF 88kb)
Figure 3 AUC for regions around causals. Bar plot of the mean of AUC for regions around causals. This summary statistic indicates the extent to which the causal itself can be detected by GWAS without relying on the LD. The red bars show the results for the means of 100 simulation results and the blue bars show the results for the means of the simulation results whose QTN1 and QTN2 were detected. a: Results for the coupling scenario. b: Results for the repulsion scenario. The abbreviations of each method are the same as those of Fig. 1.

Figure 4 An example of GWAS results for the repulsion scenario. Manhattan plots of 4 GWAS methods (RAINBOW, Single-SNP GWAS, HGF (the number of groups is 2, the grouping method is UPGMA), and SKAT) for one simulation result of the Repulsion model. The black horizontal dashed lines represent the thresholds determined by the Benjamini-Hochberg method (FDR = 0.01) for each result of the Repulsion model. The red vertical dashed lines show the positions of QTN1, QTN2, and the purple ones show the position of QTN3. The red points show $\log_{10}(p)$ of causal SNPs or haplotypes including QTN1 and QTN2, and the purple ones show $\log_{10}(p)$ of QTN3 or haplotypes including QTN3.

Additional file 2 — Table S1
Supplementary table for accession information used in this study. (CSV 19kb)

Additional file 3 — Fig. S1
Supplementary figure for $-\log_{10}(p)$ and adjusted $-\log_{10}(p)$ of QTN3 for each method. How to view this figure (including legends and abbreviations) is the same as that of Fig. 1. (PNG 1275kb)
Additional file 4 — Fig. S2
Supplementary figure for recall, precision and F-measure determined by the threshold criterion of Bonferroni correction whose significance level equals to 0.01. How to view this figure (including legends and abbreviations) is the same as that of Fig. 2. (PNG 578kb)

Additional file 5 — Fig. S3
Supplementary figure for recall, precision and F-measure of each QTN. How to view this figure (including legends and abbreviations) is the same as that of Fig. 2. (PNG 1169kb)

Additional file 6 — Fig. S4
Supplementary figure for recall, precision and F-measure of each QTN determined by the threshold criterion of Bonferroni correction whose significance level equals to 0.01. How to view this figure (including legends and abbreviations) is the same as that of Fig. 2. (PNG 1180kb)

Additional file 7 — Fig. S5
Supplementary figures (6 pages) for the examples of the cases where only RAINBOW succeeded in detecting causals, for the repulsion scenario. How to view this figure (including legends and abbreviations) is the same as that of Fig. 4. (PDF 7881kb)