OXFORD

Genome analysis

A rapid epistatic mixed-model association analysis by linear retransformations of genomic estimated values

Chao Ning[1], Dan Wang[1], Huimin Kang[1], Raphael Mrode[2], Lei Zhou[1], Shizhong Xu[3] and Jian-Feng Liu[1],*

[1]National Engineering Laboratory for Animal Breeding, Key Laboratory of Animal Genetics, Breeding and Reproduction, Ministry of Agriculture, College of Animal Science and Technology, China Agricultural University, Beijing 100193, China, [2]Animal Biosciences, International Livestock Institute, Nairobi 00100, Kenya and [3]Department of Botany and Plant Science, University of California, Riverside, CA 20705, USA

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Abstract

Motivation: Epistasis provides a feasible way for probing potential genetic mechanism of complex traits. However, time-consuming computation challenges successful detection of interaction in practice, especially when linear mixed model (LMM) is used to control type I error in the presence of population structure and cryptic relatedness.

Results: A rapid epistatic mixed-model association analysis (REMMA) method was developed to overcome computational limitation. This method first estimates individuals' epistatic effects by an extended genomic best linear unbiased prediction (EG-BLUP) model with additive and epistatic kinship matrix, then pairwise interaction effects are obtained by linear retransformations of individuals' epistatic effects. Simulation studies showed that REMMA could control type I error and increase statistical power in detecting epistatic QTNs in comparison with existing LMM-based FaST-LMM. We applied REMMA to two real datasets, a mouse dataset and the Wellcome Trust Case Control Consortium (WTCCC) data. Application to the mouse data further confirmed the performance of REMMA in controlling type I error. For the WTCCC data, we found most epistatic QTNs for type 1 diabetes (T1D) located in a major histocompatibility complex (MHC) region, from which a large interacting network with 12 hub genes (interacting with ten or more genes) was established.

Availability and implementation: Our REMMA method can be freely accessed at https://github.com/chaoning/REMMA.

Contact: liujf@cau.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genome-wide association studies (GWAS) have located many genetic variants associated with various traits. However, these variants totally explain a small proportion of phenotypic variance for most complex traits.

Epistasis, which is generally defined as interaction between different genes, is considered as one of potential explanations of the 'missing heritability' (Mackay and Moore, 2014; Upton *et al.*, 2016). In statistics, the term epistasis for two loci is defined as the deviation of the two-locus joint genotypic value from the sum of the

contributing single-locus genotypic values (Fisher, 1918). The effect of a single gene may be null or too weak to test, but it may largely contribute to the phenotypic variation jointly with other genes (Zhang et al., 2014).

For some traits, epistasis may contribute a considerable proportion of phenotypic variance, and even surpass the additive variance. For example, the estimated additive variance and epistatic variance in proportion to phenotypic variance were 35.7 and 9.7% for daily gain in pigs according to the study of Su et al. (2012); Xu (2013) reported 17 and 22% of phenotypic variance explained by additive and epistatic effect for rice tiller number. Epistasis also plays a crucial role in human complex disease. Both Wan et al. (2010) and Lippert et al. (2013) located thousands of epistatic signals for type 1 diabetes (T1D) in the major histocompatibility complex (MHC) region. Investigating epistasis will help to uncover the complex genetic system.

However, computational inefficiency hinders the wide application of the genome-wide epistatic association analysis. For example, in order to examine the pairwise interaction of 500 000 SNPs, a total of 125 billion statistical tests should be conducted, which is 250 thousand times that of the single locus analysis. This problem has attracted wide attentions and a growing number of methods have emerged to deal with the computational challenge.

The most commonly used methods are regression based methods, implemented by software packages such as PLINK (Chang et al., 2015), BOOST (Wan et al., 2010) and FastEpistasis (Schupbach et al., 2010). The advantage of these methods is the high computational efficiency due to low time complexity. Zhang et al. (2014) proposed a functional regression (FRG) model to detect gene-gene interactions with the next-generation sequencing (NGS) data. The FRG model reduced the dimension of collective information in gene by orthogonal polynomials and collectively tested the interaction between all possible pairs of SNPs within two genes. The number of interactions is therefore substantially reduced. However, in these analyses, individuals with cryptic relationship should be removed to guarantee individuals within the population to be unrelated, which may result in loss of detection power. In order to deal with population with structure and cryptic relatedness, additive kinship matrix was included in the model to control the type I error by software FaST-LMM (Lippert et al., 2011).

Linear mixed models (LMM) can correct systematic environmental factors, control population stratification and account for relatedness between individuals, and have been widely applied to GWAS. Numerous efficient methods based on LMM have emerged, such as EMMAX (Kang et al., 2010), TASSEL (Zhang et al., 2010), FaST-LMM (Lippert et al., 2011) and GEMMA (Zhou and Stephens, 2012). However, the computational efficiency can be extremely low when performing exhaustive epistatic association with the increases of sample size and marker density. Lippert et al. (2013) analyzed an expanded Wellcome Trust dataset (Wellcome Trust Case Control, 2007) deployed across 28 000 cores using FaST-LMM. It would take 950 computer years, with a wall-clock time of 13 days, across all seven phenotypes for 14 925 individuals genotyped with 356 441 SNPs.

To substantially reduce computational burden of LMM-based epistasis analysis, we proposed a rapid epistatic mixed-model association analysis method (REMMA) to detect pairwise interactions across the genome. The basic idea of our method is to estimate epistatic effects by linear retransformations of genomic estimated values. The idea is similar to the equivalence between genomic best linear unbiased prediction (G-BLUP) and SNP-BLUP which is

derived for additive effects analyses (Shen et al., 2013; Stranden and Garrick, 2009). We expand this equivalence aforementioned to the field of GWAS specially focusing on epistatic effects detection herein. Basically, each individual's epistatic value is the accumulation of pairwise interaction effects. We first estimate individuals' epistatic effects by an extended genomic best linear unbiased prediction (EG-BLUP) model (Jiang and Reif, 2015) with epistatic kinship matrix, and then estimate pairwise interaction effects by linear retransformations of the individuals' epistatic effects. The Wald Chi-squared tests based on transformed interaction effects are used to examine significant association with phenotypes. The time complexity has been substantially reduced with REMMA (detail shown in RESULT). Moreover, additive and epistatic polygenic effects are simultaneously included in our model to control multiple polygenic background effects. To further lower the computational burden, we developed a fast-speed version of REMMA (REMMA-scan). REMMA-scan first scans the genome-wide epistatic effects with running time that is linear in the cohort size, and then select top interactions to calculate their P-values. We performed a series of simulation studies to investigate the properties of the proposed method. We then further validated our method with two real datasets, an intercross mouse dataset (Jarvis and Cheverud, 2011) and a human dataset from the Wellcome Trust Case Control Consortium (WTCCC).

2 Materials and methods

2.1 From individuals' genomic estimated values to SNP effects

The whole additive and epistatic SNP effect model is shown below,

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu_a} + \mathbf{Qu}_i + \mathbf{e}. \tag{1}$$

Here, we assume that there are n individuals and m SNPs. Thus, $\mathbf{y}$ is a $n \times 1$ vector of phenotypic values; $\mathbf{b}$ is a vector of fixed effects; $\mathbf{u}_a$ is a $m \times 1$ vector of additive SNP effects; $\mathbf{u}_i$ is an $m(m-1)/2 \times 1$ vector of exhaustive epistatic SNP effects; $\mathbf{e}$ is a vector of residual errors; $\mathbf{X}$ is the design matrix for the fixed effects; $\mathbf{Z}$ and $\mathbf{Q}$ are standardized design matrices for additive and epistatic SNP effects, respectively. Matrix $\mathbf{Z}$ is constructed as follows,

$$\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_j, \ldots, \mathbf{z}_m) / \sqrt{2 \sum p_j(1 - p_j)}, \tag{2}$$

where p_j is the allele frequency of allele A for the jth SNP, and $\mathbf{z}_j$ is the jth SNP vector with elements defined as

$$\mathbf{z}_j = \begin{cases} 2 - 2p_j & AA \\ 1 - 2p_j & Aa \\ 0 - 2p_j & aa \end{cases}. \tag{3}$$

The $\mathbf{Q}$ matrix is constructed from $\mathbf{Z}$,

$$\mathbf{Q} = \{\mathbf{Z}_j \# \mathbf{Z}_k\}_{j \neq k, j < k}^{j=1..m, k=2..m}. \tag{4}$$

The symbol # represents the Hadamard matrix product (element wise multiplication). Matrix $\mathbf{Q}$ has n rows and $m(m-1)/2$ columns, and each column represents the interaction SNP vector of the jth and kth SNPs.

For Equation (1), we define the following variance matrices of all random effects,

$$\mathbf{u}_a \sim (0, \mathbf{I}\sigma_a^2), \ \mathbf{u}_i \sim (0, 2\mathbf{I}\sigma_i^2), \ \mathbf{e} \sim (0, \mathbf{I}\sigma_e^2).$$

where **I** is identity matrix. Following Henderson (1949), the mixed model equation (MME) for (1) is (Here, $\mathbf{R} = \mathbf{I}\sigma_e^2$)

$$
\begin{bmatrix}
\mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Q} \\
\mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + (I\sigma_a^2)^{-1} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Q} \\
\mathbf{Q}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Q}'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{Q}'\mathbf{R}^{-1}\mathbf{Q} + (2I\sigma_i^2)^{-1}
\end{bmatrix}
\begin{bmatrix}
\hat{\mathbf{b}} \\
\hat{\mathbf{u}}_a \\
\hat{\mathbf{u}}_i
\end{bmatrix}
$$

$$
= \begin{bmatrix}
\mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\
\mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \\
\mathbf{Q}'\mathbf{R}^{-1}\mathbf{y}
\end{bmatrix}
\tag{5}
$$

However, this MME is computationally unsolvable due to too many random effects when m is large (m for additive effects and $m(m-1)/2$ for epistatic effects).

If we set $\mathbf{a} = \mathbf{Z}\mathbf{u}_a$, $\mathbf{i} = \mathbf{Q}\mathbf{u}_i$, and **a** and **i** are defined as individuals' additive and epistatic effect vector. Then, Equation (1) can be rewritten as:

$$
\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{a} + \mathbf{i} + \mathbf{e} \tag{6}
$$

The variance matrices for **a** and **i** are:

$$
\mathrm{var}(\mathbf{a}) = \mathrm{var}(\mathbf{Z}\mathbf{u}_a) = \mathbf{Z}\mathrm{var}(\mathbf{u}_a)\mathbf{Z}' = \mathbf{Z}\mathbf{Z}'\sigma_a^2,
$$

$$
\mathrm{var}(\mathbf{i}) = \mathrm{var}(\mathbf{Q}\mathbf{u}_i) = \mathbf{Q}\mathrm{var}(\mathbf{u}_i)\mathbf{Q}' = 2\mathbf{Q}\mathbf{Q}'\sigma_i^2.
$$

If we define $\mathbf{K}_a = \mathbf{Z}\mathbf{Z}'$ and $\mathbf{K}_i = 2\mathbf{Q}\mathbf{Q}'$, then $\mathrm{var}(\mathbf{a}) = \mathbf{K}_a\sigma_a^2$, $\mathrm{var}(\mathbf{i}) = \mathbf{K}_i\sigma_i^2$. $\mathbf{K}_a$ is additive kinship matrix as defined by VanRaden (2008) and $\mathbf{K}_i$ is the epistatic kinship matrix as defined by Xu (2013). However, the epistatic kinship matrix is computationally inefficient using the formula of 2QQ'. In genomic selection of animal breeding, Su *et al.* (2012) utilized equation of $\mathbf{K}_a \# \mathbf{K}_a$ referred to traditional pedigree-based BLUP (Henderson, 1985). However, the equation included interactions of loci with themselves. An accurate and efficient form proved by Jiang and Reif (2015) is

$$
\mathbf{K}_i = \mathbf{K}_a\#\mathbf{K}_a - (\mathbf{Z}\#\mathbf{Z})(\mathbf{Z}\#\mathbf{Z})'. \tag{7}
$$

where # represents the Hadamard matrix product.

The MME for Equation (6) is:

$$
\begin{bmatrix}
\mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{I} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{I} \\
\mathbf{I}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{I}'\mathbf{R}^{-1}\mathbf{I} + (\mathbf{K}_a\sigma_a^2)^{-1} & \mathbf{I}'\mathbf{R}^{-1}\mathbf{I} \\
\mathbf{I}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{I}'\mathbf{R}^{-1}\mathbf{I} & \mathbf{I}'\mathbf{R}^{-1}\mathbf{I} + (\mathbf{K}_i\sigma_i^2)^{-1}
\end{bmatrix}
\begin{bmatrix}
\hat{\mathbf{b}} \\
\hat{\mathbf{a}} \\
\hat{\mathbf{i}}
\end{bmatrix}
$$

$$
= \begin{bmatrix}
\mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\
\mathbf{I}'\mathbf{R}^{-1}\mathbf{y} \\
\mathbf{I}'\mathbf{R}^{-1}\mathbf{y}
\end{bmatrix}. \tag{8}
$$

To demonstrate the derivation process, we solve (5) and get

$$
\hat{\mathbf{u}}_i = (Q'\mathbf{R}^{-1}Q + I/\sigma_i^2)^{-1}Q'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \mathbf{Z}\hat{\mathbf{u}}_a) \tag{9}
$$

where hat vectors means corresponding estimated values. Giving that matrices M and S are nonsingular, the Woodbury Matrix identity states that

$$
(\mathbf{M} + \mathbf{B}\mathbf{S}\mathbf{C}')^{-1} = \mathbf{M}^{-1} - \mathbf{M}^{-1}\mathbf{B}(\mathbf{S}^{-1} + \mathbf{C}'\mathbf{M}^{-1}\mathbf{B})^{-1}\mathbf{C}'\mathbf{M}^{-1} \tag{10}
$$

Applying (10) to $(Q'\mathbf{R}^{-1}Q + I/\sigma_i^2)^{-1}$ in (9) assuming $\mathbf{M} = \mathbf{I}/\sigma_i^2$, $\mathbf{Q}' = \mathbf{B} = \mathbf{C}$ and $\mathbf{S} = \mathbf{R}^{-1}$, we obtain

$$
\hat{\mathbf{u}}_i = (I\sigma_i^2 - I\sigma_i^2 Q'(R + QI\sigma_i^2 Q')^{-1}QI\sigma_i^2)Q'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \mathbf{Z}\hat{\mathbf{u}}_a).
$$

This can be simplified to

$$
\hat{\mathbf{u}}_i = (IQ'\sigma_i^2 - IQ'(R/\sigma_i^2 + K_i)^{-1}K_i\sigma_i^2)\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \mathbf{Z}\hat{\mathbf{u}}_a). \tag{11}
$$

Applying (10) again ($\mathbf{S} = \mathbf{R}/\sigma_i^2$, $\mathbf{B} = \mathbf{C} = \mathbf{I}$ and $\mathbf{M} = \mathbf{K}_i$), gives

$$
(\mathbf{R}/\sigma_i^2 + \mathbf{K}_i)^{-1} = \mathbf{K}_i^{-1} - \mathbf{K}_i^{-1}(\mathbf{R}^{-1}\sigma_i^2 + \mathbf{K}_i^{-1})^{-1}\mathbf{K}_i^{-1}.
$$

Applying $\hat{\mathbf{a}} = \mathbf{Z}\hat{\mathbf{u}}_a$ and further simplifying Equation (11) leads to

$$
\hat{\mathbf{u}}_i = \mathbf{Q}'\mathbf{K}_i^{-1}(\mathbf{R}^{-1} + \mathbf{K}_i^{-1}/\sigma_i^2)^{-1}\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \hat{\mathbf{a}}). \tag{12}
$$

Solving (8), we get the estimated values of individuals' epistatic effects, i.e.

$$
\hat{\mathbf{i}} = (\mathbf{R}^{-1} + \mathbf{K}_i^{-1}/\sigma_i^2)^{-1}\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \hat{\mathbf{a}}). \tag{13}
$$

Combining (13) and (12) leads to

$$
\hat{\mathbf{u}}_i = \mathbf{Q}'\mathbf{K}_i^{-1}\hat{\mathbf{i}}. \tag{14}
$$

Defining $\mathbf{O} = \mathbf{K}_i^{-1}\hat{\mathbf{i}}$, which can be computed prior to the epistatic scan, we get $\hat{\mathbf{u}}_i = \mathbf{Q}'\mathbf{O}$.

We need to calculate the error matrix of $\hat{\mathbf{u}}_i$, which is $\mathrm{var}(\hat{\mathbf{u}}_i) = \mathbf{Q}'\mathbf{K}_i^{-1}\mathrm{var}(\hat{\mathbf{i}})\mathbf{K}_i^{-1}\mathbf{Q}$. Following Henderson (1975), we have

$$
\mathrm{var}(\hat{\mathbf{i}}) = \mathbf{K}_i\sigma_i^2 - \mathbf{C}_{33}, \tag{15}
$$

where $\mathbf{C}_{33}$ is the block in the lower-right corner of the inverse of the coefficient matrix of (8) as shown below,

$$
\begin{bmatrix}
\mathbf{C}_{11} & \mathbf{C}_{12} & \mathbf{C}_{13} \\
\mathbf{C}_{12} & \mathbf{C}_{22} & \mathbf{C}_{23} \\
\mathbf{C}_{13} & \mathbf{C}_{23} & \mathbf{C}_{33}
\end{bmatrix}.
$$

Assume $\mathbf{P} = \mathbf{K}_i^{-1}\mathrm{var}(\hat{\mathbf{i}})\mathbf{K}_i^{-1}$ and the Cholesky decomposition of **P** gives $\mathbf{P} = \mathbf{L}\mathbf{L}'$. Now

$$
\mathrm{var}(\hat{\mathbf{u}}_i) = \mathbf{Q}'\mathbf{L}\mathbf{L}'\mathbf{Q} = (\mathbf{Q}'\mathbf{L})(\mathbf{Q}'\mathbf{L})' \tag{16}
$$

The diagonal elements of $\mathrm{var}(\hat{\mathbf{u}}_i)$ are the estimated variance for the estimated pairwise interaction effects.

2.2 Test statistics

For the interaction effect between the jth SNP and the kth SNP, the estimate and its variance are

$$
u_{j,k} = (\mathbf{z}_j\#\mathbf{z}_k)'\mathbf{O}, \quad \mathrm{var}(\hat{u}_{j,k}) = (\mathbf{z}_j\#\mathbf{z}_k)'\mathbf{L}\mathbf{L}'(\mathbf{z}_j\#\mathbf{z}_k)' \tag{17}
$$

The Wald Chi-squared test is defined as,

$$
\frac{\hat{u}_{j,k}^2}{\mathrm{var}(\hat{u}_{j,k})} \sim \chi^2(1) \tag{18}
$$

We can see that interaction effects are calculated by vector-vector multiplication whose computational complexity is $O(n)$, and the variance of estimated effects are calculated by matrix-vector multiplication whose computational complexity is $O(n^2)$.

Similarly, for the additive SNP effects, we have

$$
\hat{\mathbf{u}}_a = \mathbf{Z}'\mathbf{K}_a^{-1}\hat{\mathbf{a}}, \quad \mathrm{var}(\hat{\mathbf{u}}_a) = \mathbf{Z}'\mathbf{K}_a^{-1}\mathrm{var}(\hat{\mathbf{a}})\mathbf{K}_a^{-1}\mathbf{Z}, \text{ and } \mathrm{var}(\hat{\mathbf{a}}) = \mathbf{K}_a\sigma_a^2 - \mathbf{C}_{22} \tag{19}
$$

The Wald Chi-squared test for the kth additive SNP effect is

$$
u_k = \mathbf{z}_k'\mathbf{K}_a^{-1}\hat{\mathbf{a}}, \quad \mathrm{var}(\hat{u}_k) = \mathbf{z}_i'\mathbf{L}\mathbf{L}'\mathbf{z}_i' \tag{20}
$$

$$
\frac{\hat{u}_k^2}{\mathrm{var}(\hat{u}_k)} \sim \chi^2(1) \tag{21}
$$

If we only focus on the additive SNP effects, the epistatic effects can be excluded from Equations (1) and (6) and similar test statistics for additive SNP effects can also be obtained.

2.3 REMMA-scan

To further relax the computational burden of full REMMA method, we develop a fast-speed version of REMMA (REMMA-scan). REMMA-scan firstly scans the genome-wide epistatic effects with running time that is linear in the cohort size, then selects top interactions to calculate their P-values. The detailed workflow is specifically described below:

Step 1: Randomly select S (e.g. 10^6) SNP pairs and calculate their epistatic effects.
Step 2: Normal distribution is fitted with sample mean and standard deviation of the S epistatic effects, and different quantiles are calculated.
Step 3: Scan the genome-wide epistatic effects and select the top SNP–SNP pairs passing different quantiles.
Step 4: Examine these top SNP pairs with full REMMA.
Step 5: If no or less than the expected additional significant interactions are observed between continuous quantiles, take the result from Step 4 as final; otherwise, add additional quantiles and go to Step 3.

2.4 Simulation studies

In our simulation studies, the real genotypes of 959 individuals from the Human Genetic Analysis Workshop 18 (GAW18) dataset (Bickeboller *et al.*, 2014) were used for phenotypic simulation. We randomly sampled 1K, 10K, 100K and 300K SNPs from the whole genome to achieve different marker densities. The phenotypes were generated with equation: phenotype = population mean + additive + epistatic + residual. The population mean was set to 4.0. The additive effects were allocated to 100 SNPs, and the epistatic effects were allocated to 5, 10, 100 or 500 SNP–SNP pairs. We showed the detailed information of additive and epistatic QTNs in Supplementary Table S5. Following the previous studies (Liu *et al.*, 2016; Zhang *et al.*, 2010), the additive and epistatic effects followed a geometric distribution, i.e. the ith QTN effect was set as a^i, where a = 0.9. The residuals were drawn from a normal distribution. The phenotypic variance explained by the additive effects (h_a^2) was 30%, and the phenotypic variance contributed by the epistatic effects (h_i^2) was set at the following levels: 10, 30 and 60%. The epistatic variance was fixed, and the additive variance and residual variance were scaled to achieve the required simulation experiments. For each experiment, the simulation was repeated 100 or 1000 times. A positive interaction was considered to have been identified if a SNP–SNP pair which showed high linkage disequilibrium (LD) ($r^2 > 0.5$) (Gabriel *et al.*, 2002; Wu *et al.*, 2014) with the respective true QTN pairs passed the thresholds of different Type I error. All the other pairwise SNPs were unlinked to the true interaction QTNs, and the cumulative distribution of the P-values should show a uniform distribution with no inflated false positive (Kang *et al.*, 2008; Yu *et al.*, 2006). In our study, we also explored the influence on testing additive SNP effects with different levels of h_i^2.

2.5 Real data

The mouse dataset contained 1304 samples from the F10 generation of an intercross line, and each individual was genotyped with 1470 SNPs. The phenotype used in the study was reproductive fat pad weight. The human dataset was obtained from the Wellcome Trust

Case Control Consortium (WTCCC). The data contained genotypes of 13 999 individuals for seven common diseases: bipolar disorder (BD), coronary artery disease (CAD), crohn disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D) and type 2 diabetes (T2D). The 1500 controls from the UK Blood Service Control Group (NBS) were also included in our analysis. To achieve high quality data, SNPs with minor allele frequency less than 5% or more than 1% missing data were removed for the control group and also from each case group. We also performed the Hardy-Weinberg Equilibrium (HWE) test for each SNP in the control group and SNPs with a P-value < 0.001 were excluded from the data. After filtering, there were 15 490 individuals with 368 895 SNPs for the subsequent analysis. The Bonferroni correction was used to control false-positive rates for the two real datasets.

3 Results

3.1 Analytical models

We employed different analytical methods with respective model fittings for performance comparison in the analyses as below:

PLINK: Simple linear model is utilized to examine the additive and epistatic effect (PLINK-add and PLINK-epi).
FaST-LMM: The method is based on linear mixed model and additive kinship matrix is utilized to control the type I error in testing additive and epistatic effect (FaST-LMM-add and FaST-LMM-epi).
REMMA: The method is based on the linear mixed model. Besides the additive, epistatic kinship matrix is also included in the model in testing additive and epistatic effect (REMMA-epi-add and REMMA-epi). We also include REMMA-add model to test additive effect in the study, which only incorporates additive kinship matrix in the model.
REMMA-scan: A fast-speed version of REMMA in testing genome-wide epistatic effect and its detailed workflow is shown in 'Materials and methods' section.

3.2 Simulation

Extensive simulation studies were performed to systematically compare the performance of REMMA with two alternative methods (PLINK and FaST-LMM).

The quantile–quantile (Q–Q) plots of different methods in testing epistatic SNP effects at different h_i^2s (phenotypic variance contributed by epistatic effects; i.e. epistatic heritability) and marker densities are shown in Figure 1 and Supplementary Figure S1. For all simulation experiments, REMMA controls type I error better than PLINK and FaST-LMM. The simple linear model implemented by PLINK has the inflated type I error with the increases of h_i^2 level. However, marker density does not influence the performance of PLINK. FaST-LMM includes additive kinship matrix to control polygenic effects. It performs well at relatively low epistatic heritability, but still leads to the inflated type I error as the epistatic heritability increases. Nevertheless, we observed that type I error by FaST-LMM reduced when marker density increases.

The receiver operating characteristic (ROC) curve is a plot of statistical power against type I error, which is a useful way to compare different methods in detection of significant signals. Figure 2 and Supplementary Figure S2 show the ROC curves at different simulated epistatic heritabilities and marker densities. In general, REMMA has higher or similar performance compared with FaST-LMM in most cases. FaST-LMM slightly outperforms REMMA at
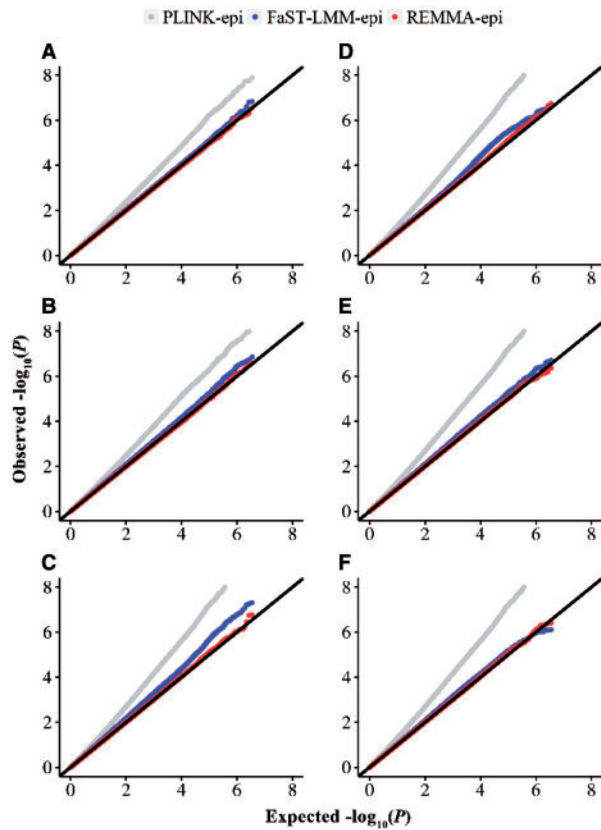
Fig. 1. The quantile–quantile plots of different methods in testing epistatic SNP effects. The epistatic effects were allocated to 100 SNP–SNP pairs. We randomly selected 122 500 pairs of SNPs from regions not in linkage disequilibrium with epistatic QTNs for each simulation and the simulation was repeated 100 times. Their observed *P*-values are expected to show a uniform distribution. Upward departures from the expectation indicate that the methods cause spurious associations. The first three (left panel) settings are simulated with 1K SNPs at epistatic heritability of 10, 30 and 60%, respectively. The last three (right panel) settings are simulated at epistatic heritability of 60% with 10K, 100K and 300K SNPs, respectively
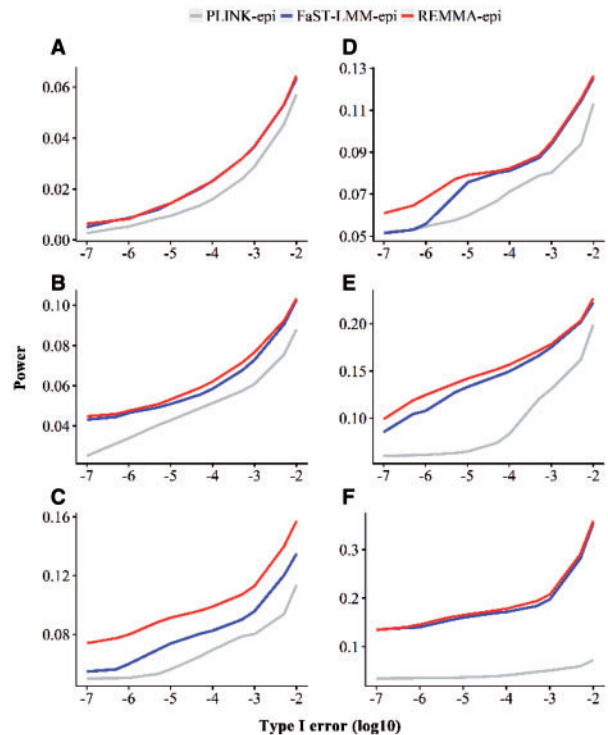

Fig. 2. Statistical power plotted against Type I error (in a $\log_{10}$ scale) of different methods in testing epistatic effects. The epistatic effects were allocated to 100 SNP–SNP pairs. The first three (left panel) settings are simulated with 1K SNPs at epistatic heritability of 10, 30 and 60%, respectively. The last three (right panel) settings are simulated at epistatic heritability of 60% with 10K, 100K and 300K SNPs, respectively

the simulation of high maker density with $h_i^2 = 10\%$. Simple linear model of PLINK has the lowest adjusted power for all simulation studies.

To explore the impact of the number of epistatic QTN pairs on the performance of different methods, we conducted simulations with varied number of interaction pairs (5, 10, 100 and 500). The results in Supplementary Figure S3 show REMMA still surpasses other methods in all situations. It should be noted that, in these simulations, we set a certain number of QTNs with both epistatic and additive effects. To widely investigate the performance with different scenarios, we alternatively mimicked the situation where all additive QTNs being outside the epistatic QTN regions, and the results (Supplementary Fig. S4) showed that REMMA still outperformed other methods.

It is notable that there exists a relatively high correlation (above 0.8) between the test statistics and the absolute values of epistatic effects in our simulations (Fig. 3A). The results suggest that we can implement REMMA-scan strategy to our simulated data, which scans the pairwise epistatic effects first and then selects the top interactions to calculate their *P*-values (see detailed workflow in 'Materials and methods' section). Considering the power of full REMMA after Bonferroni correction as standard (i.e. $0.05*2/[m*(m-1)]$, where *m* is

the number of SNPs), we found that detection power of REMMA-scan improved as the increases of quantiles of normal distribution fitted by random selected epistatic effects, and gradually reached a plateau at 1/10 000 quantile (Fig. 3B–D). The time complexity for calculating each epistatic effect is $O(n)$ (*n* is the number of individuals), while the time complexity for calculating its test statistic is $O(n^2)$. If we discard massive interaction with low epistatic effects (about one in ten thousand interaction pairs remain), the actual computing time will be significantly reduced (shown in 'Computational efficiency' section).

We also investigated the influence of the epistatic heritability on testing additive SNP effects. The quantile–quantile (Q–Q) plots of four methods in testing additive SNP effects are shown in Supplementary Figure S5A–C. It was shown that all methods except the simple linear model control the type I error well. Even at very high epistatic heritability, the methods only including additive kinship (FaST-LMM-add and REMMA-add) will not inflate the false positives. The ROC curves in Supplementary Figure S5D–F show that FaST-LMM-add, REMMA-add and REMMA-epi-add have very similar power after controlling the type I error. The results indicated that additive effects could be reliably tested even when epistatic variance was not incorporated, which was in agreement with the study of Kruijer *et al.* (2015).

3.3 Application to real data

To further evaluate our method, we applied REMMA to two public real datasets from mouse and human. The mouse dataset (Jarvis and Cheverud, 2011) contained 1304 samples from the F10 generation of an intercross line, and each individual was genotyped with 1470
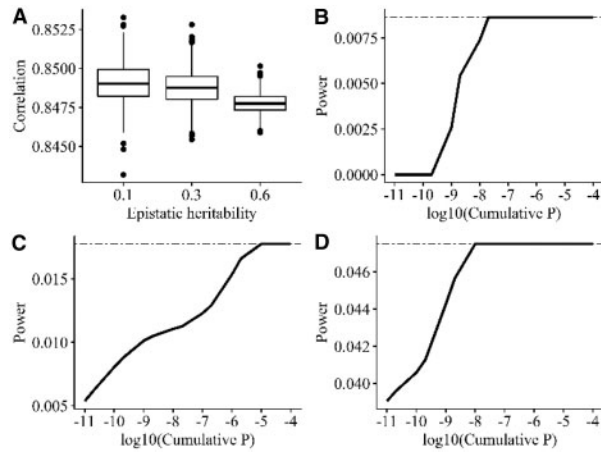
Fig. 3. The performance of REMMA-scan in the simulation study. The simulation is based on 100 pairs of epistatic QTNs at marker density of 100K. (**A**) Boxplots of Pearson correlation coefficient between epistatic effects (absolute values) and corresponding *P*-values (-log$_{10}$*P*). (**B–D**) The change of detection power with increasing quantile of normal distribution fitted by random selected epistatic effects at the epistatic heritability of 0.1 (B), 0.3 (C) and 0.6 (D). The dashed line means the detection power of the full REMMA after Bonferroni correction
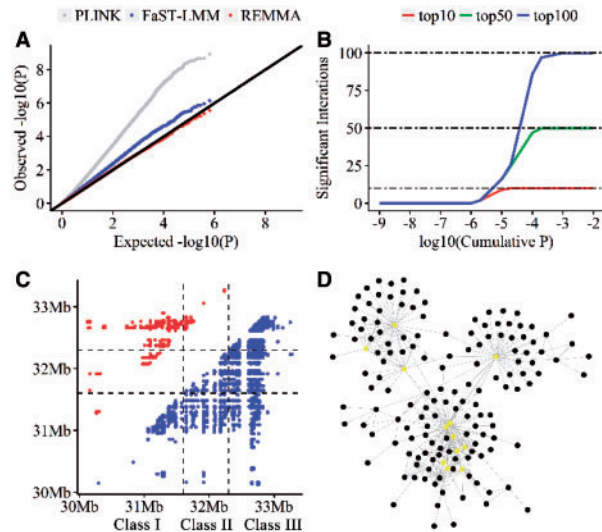

Fig. 4. Genome-wide epistatic association analysis for reproductive fat pad weight in mouse data (up panel) and T1D in WTCCC data (bottom panel). (**A**) Quantile–quantile (Q–Q) plots of genome-wide epistatic associations by three different method (PLINK, FaST-LMM and REMMA). (**B**) The change of detected significant interactions with increasing quantile of normal distribution fitted by random selected epistatic effects. (**C**) Significant interactions in the MHC region. The upper left points (red) corresponds to significant and well separated (>1 Mb) interactions in the MHC region, and the lower right points (blue) corresponds to all significant interactions in the MHC region. (**D**) Interacting networks of 310 pairs of genes with 12 hub genes. Hub genes involved in ten or more interactions are highlighted with yellow

SNPs. For reproductive fat pad weight in mouse, the phenotypic variance explained by additive variance, epistatic variance and residual variance are 0.285 ± 0.047, 0.347 ± 0.063 and 0.369 ± 0.057, respectively. The quantile–quantile (Q–Q) plots of genome-wide epistatic associations (Fig. 4A) indicate that both FaST-LMM and PLINK lead to the inflated type I error in the mouse data. REMMA did not locate any significant epistatic signals after Bonferroni correction ($0.05*2/(1470*(1470-1) = 4.63e-8$) in this study, and it indicated that there were no major epistatic genomic regions affecting reproductive fat pad weight in mouse. We defined top 10, 50 or 100 pair SNPs (based on their *P*-values) as significant interactions and applied REMMA-scan to detect the interactions. The number of detected interactions reached a plateau at 1% quantile (Fig. 4B).

The human dataset was obtained from the Wellcome Trust Case Control Consortium (WTCCC) (Wellcome Trust Case Control, 2007). The data contain genotypes and phenotypes of 13 990 individuals for seven common diseases: bipolar disorder (BD), coronary artery disease (CAD), crohn disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D) and type 2 diabetes (T2D). In addition, 1500 controls from the UK Blood Service Control Group (NBS) were also included in the analysis. The quality control procedure is presented in 'Materials and methods' section. T1D has been located abundant significant interactions in previous study (Lippert *et al.*, 2013; Wan *et al.*, 2010), therefore, we focused on T1D to validate our methods. To further achieve a sufficient high statistical power, we expanded the control set by including all other disease cohorts and closely related individuals. The tactic has been applied to the WTCCC data by Lippert *et al.* (2013) with another LMM-based methods (FaST-LMM). We first analyzed the data with the REMMA-scan strategy. We randomly examined 10^6 epistatic effects, and found a high correlation between absolute values of epistatic effects and the corresponding *P*-values (-log$_{10}$(*P*)) (Pearson's $r = 0.85$, *P*-value $< 2.2e-16$). The Q–Q plots of these *P*-values appear to be uniform, indicating that REMMA does not inflate false positive rate (Supplementary Fig. S6). The number of significant interactions passing the Bonferroni correction (7.35e−13) reached a

plateau at 1/10 000 quantile (Supplementary Fig. S7). We observed 5610 significant interactions with REMMA-scan. To estimate the performance of REMMA-scan, we ran the full REMMA and observed 5623 significant SNP–SNP pairs that had passed the Bonferroni corrected threshold (Supplementary Table S1). The REMMA-scan method only missed 13 interactions. Expanding the controls could lead to spurious associations arising from any of the other diseases in the controls. We examined the 5623 significant interactions in other six traits and removed the more significant epistatic signals in any of the other diseases from the expanded analysis for T1D, which was suggested by Lippert *et al.* (2013). Then, 5571 significant interactions were remained.

We observed several remarkable features from these significant interactions for T1D, which are listed below. (i) A mass of pairwise interactions have physical distance within 1 Mb. Epistatic interactions for SNP pairs within 1 Mb distance may be false positives due to linkage (Lippert *et al.*, 2013; Wan *et al.*, 2010). Therefore, we removed such potential spurious associations, leaving 1056 well separated SNP pairs (18.96%) for subsequent analysis. (ii) Most of significant interactions for T1D have weak marginal effects. We calculated the *P*-values of marginal effects for these significant interactions, and observed that most of them (944/1056) had weak marginal effects (*P*-value < 1e-6), which is consistent with the findings of Wan *et al.* (2010) (698 out of 789). (iii) Most of the detected epistatic interactions are in the MHC region. Both Wan *et al.* (2010) and Lippert *et al.* (2013) reported that the MHC region (29.8– 33.4 Mb on chromosome 6) could be a candidate interaction region for T1D. Among the 5571 significant epistatic interactions in our study, 5287 interactions (94.90%) are in the MHC region, where 859 interactions are well separated (Fig. 4C). (iv) A considerable number of epistatic QTNs interact with five or more loci. Forsberg *et al.* (2017) found that most epistatic QTNs interacted with one or

a few loci in yeast. However, we observed 385 epistatic QTNs for 1056 well separated interactions, and 110 epistatic QTNs (28.57%) interacted with five or more loci (Supplementary Fig. S8A, Supplementary Table S2). We searched the protein-coding genes within or nearest to the epistatic QTNs, and 310 unique paired interacting genes (including 159 epistatic genes) were located (Supplementary Table S3). Among the 159 epistatic genes, 39 genes interacted with five or more loci (Supplementary Fig. S8B, Supplementary Table S4). (v) We drew the epistatic network regulating T1D with Cytoscape (Doerks, 2002), and a large interacting network with 12 hub genes (interacting with ten or more genes) was established (Fig. 4D). (vi) Most of the genes interacting with five or more other genes locate in MHC region. We observed 39 genes interacting with five or more other genes, and 24 of them were in MHC region.

3.4 Computational efficiency

Theoretical complexity and actual performances were utilized to evaluate the computational efficiency of REMMA (including REMMA-scan and full REMMA), FaST-LMM and PLINK in testing epistatic SNP effects (Table 1). For FaST-LMM, REMMA or REMMA-scan, the computational complexity for building kinship matrix and estimating variance components is $O(mn^2)$ and $O(n^3)$, respectively. However, the association step takes up much more computing time. Therefore, we compared the computational complexity for association step in detail. For each epistatic test of full REMMA, computing time of preparation for interaction vector ($z_i \# z_j$), predicting epistatic effect and estimating corresponding estimated error are $O(n)$, $O(n)$ and $O(0.5n^2 + 1.5n)$, respectively. Therefore, the overall time complexity for the step of epistatic association test is $O(0.5Tn^2 + 3.5Tn)$, where n is the sample size and T is the total number of pairwise tests. REMMA-scan firstly random selects S SNP–SNP pairs to fit normal distribution and then scans genome-wide epistatic effects, whose time complexity is linear on the sample size. The SNP–SNP pairs passing the optimized quantile of normal distribution (assuming K) are for subsequent analysis. The overall time complexity is $O(2Sn + 2Tn + 0.5Kn^2 + 3.5Kn)$. FaST-LMM is an exact method and estimates variance parameters for each test. The method ingeniously takes advantage of spectral decomposition of the kinship matrix and the computational

complexity of the optimization step is reduced from quadratic of n to linear of n. However, the computational complexity of rotating the LMM to the simple linear model is square of n. The overall time complexity is shown $O(3Tn^2 + Tt_1c_1^2n + Tt_2 c_1^2n)$. As we showed earlier, just including the additive kinship matrix would lead to the inflated type I error when the epistatic heritability is high. However, the time complexity of the optimization step would be $O(n^3)$ when the epistatic kinship matrix is included in the model, which makes the exact epistatic test impossible. The simple linear model by PLINK is the most efficient and its computational complexity is linear of sample size for each test.

Actual performance for the mouse and the WTCCC datasets also proved that REMMA was more efficient than LMM-based FaST-LMM. REMMA was 182 times faster than FaST-LMM for the mouse data analysis. For the WTCCC dataset with tens of thousands of individuals and hundreds of thousands of SNPs, it took 5 days for REMMA using 2400 cores on Intel Xeon E5 2.2 GHz CPU. However, FaST-LMM failed due to the maximal core limit (2400 cores). When REMMA-scan was applied to the two datasets, the amount of real computational time was in the same order as PLINK.

4 Discussion

It has been widely accepted that additive genetic effects could explain most of genetic variance (Bloom *et al.*, 2015; Maki-Tanila and Hill, 2014), while epistasis also substantially contribute to genetic variance of some complex traits (Wan *et al.*, 2010; Xu, 2013). Nevertheless, exhaustive epistasis analysis is computationally expensive, especially when linear mixed model is utilized to explain relatedness among samples and to control population stratification. In this study, we proposed a rapid epistatic mixed-model association analysis method (REMMA) by linear retransformations of the genomic estimated values. REMMA included both additive and epistatic kinship matrices, and could theoretically control type I error efficiently in examining pairwise epistatic SNPs. Simulation studies using genotypes of the human population also demonstrated that REMMA performed better in controlling false positives than FaST-LMM (include additive kinship matrix only) and PLINK (simple linear kinship matrix). Application to the mouse data further confirmed the performance of REMMA in controlling type I error.

Both Wan *et al.* (2010) and Lippert *et al.* (2013) demonstrated that the MHC region was the major epistatic region regulating T1D. Almost all the T1D epistatic interactions located by REMMA were in the MHC region, which was consistent with previous studies (Lippert *et al.*, 2013; Wan *et al.*, 2010). The results proved that REMMA was efficient in detecting epistatic signals in real datasets. The results of Zhang *et al.* (2014) indicated that interacting genes formed small interacting networks for high-density lipoprotein (HDL), and Forsberg *et al.* (2017) showed similar findings of quantitative traits in yeast. However, we observed a large interacting network for T1D. Furthermore, a considerable number of epistatic QTNs or genes interacted with five or more the others. Among these hub genes, *HLA-DQB1* interacted with the most number of other genes (46 genes), and participated in immune response and antigen processing and presentation. These results indicated that epistasis played a nontrivial and complex role in the regulating process of T1D.

Linear mixed models can control population stratification and explain hidden relatedness to correct for the inflated false positives (Kang *et al.*, 2008, 2010; Yu *et al.*, 2006; Zhang *et al.*, 2010) in

Table 1. Theoretical complexity and actual computational time of different methods in testing epistatic SNP effects

Methods	Time Complexity for Association Test	Computing Time for Association step	
		Mouse	Human
PLINK	$O(Tn)$	1.2 s	10 min
FaST-LMM	$O(3Tn^2 + Tt_1c_1^2n + Tt_2 c_1^2n)$	79 min	NA
REMMA	$O(0.5Tn^2 + 3.5Tn)$	26 s	5 d
REMMA-scan	$O(2Sn + 2Tn + 0.5 Kn^2 + 3.5Kn)$	2.3 s	24 min

Note: All computing was performed on Intel Xeon E5 2.2 GHz CPU. A single core was used for mouse data, while 2400 cores was used for the WTCCC data. S is the number of random selected epistatic effects; T is the total number of pairwise test; n is the number of individuals; c_1 is the number of covariates except interaction and c_2 is the number of covariates; t_1 and t_2 are the number of optimization iterations for null model (without interaction) and full model (including interaction); K is the selected number of top interactions to calculate the P-values.

GWAS. Computational efficiency has improved dramatically for single locus analysis and can deal with groups with tens of thousands of samples genotyped millions of SNPs on standard desktop. However, the computational complexity of association statistics for each test is $O(n^2)$ (Yang *et al.*, 2014) compared with $O(n)$ by the simple linear model, where n is the sample size. REMMA reduces the coefficient of $O(n^2)$ to 0.5, but the computational cost for large sample size is still high compared to PLINK. We proposed REMMA-scan to overcome the problem. REMMA-scan firstly scans genome-wide epistatic effects and then selects top interactions for subsequent examination. Application to real datasets proved that REMMA-scan had similar performance with PLINK in computational efficiency.

In our study, the randomly selected 10^6 SNP–SNP pairs were merely used to develop an empirical normal distribution with their sample mean and standard deviation. However, for extremely high dense/imputed genotype data, 10^6 SNP–SNP pairs may be not enough to accurately estimate the empirical distribution thus maybe weakening the performance REMMA-scan. In this situation, a larger number of randomly selected SNP–SNP pairs could be considered, or alternatively even calculation of interaction effects for all pairs, since computational time is linear with population size. This is only the case in the framework of REMMA-scan. The choice of initial quantile for normal distribution fitted by random selected epistatic effects also will significantly affect performance of REMMA-scan. The relatively lower initial quantile will increase the repetition times, and higher initial quantile will increase the number of examination. In current study, 1/10 000 quantile is appropriate for hundreds of thousands of markers. For low-density SNP chip, the full REMMA method can work well.

Theoretically, each random SNP–SNP interaction effect should have respective variance, and it is not the best consideration to assume that the effects of all SNP–SNP interactions are sampled from normal distribution with the identical variance. Under the framework of Bayesian inference (Gianola *et al.*, 2009; Meuwissen *et al.*, 2001; Xu, 2003), we can address this issue allowing SNP–SNP interaction variance varying with its effect size. However, the extremely high computational demand based on Bayesian model makes it infeasible in practice when all the additive and epistatic SNP effects are considered. Taking this aspect aforementioned into consideration, we avoided estimating the SNP effects directly as well as assuming different variance for different pair of SNP–SNP interaction under the framework of GBLUP-based strategy, i.e. estimated SNP effects by linear retransformations of the individuals' genomic estimated values. The basic idea of REMMA enlightens us as to improving the accuracy of individuals' genomic estimated values to enhance the QTN detected power. In the field of improving individuals' genomic prediction, Speed and Balding (2014) propose MultiBLUP, which groups the SNPs into different classes based on SNP functional annotations and effect-size variances; Legarra *et al.* (2009) and Christensen and Lund (2010) propose single-step GBLUP in the meanwhile, which can include the not genotyped individuals in the model. These methods are promising to further improve the performance of REMMA.

References

Bickeboller,H. *et al.* (2014) Genetic Analysis Workshop 18: methods and strategies for analyzing human sequence and phenotype data in members of extended pedigrees. *BMC Proc.*, **8**, S1.

Bloom,J.S. *et al.* (2015) Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. *Nat. Commun.*, **6**, 8712.

Chang,C.C. *et al.* (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, **4**, 7.

Christensen,O.F. and Lund,M.S. (2010) Genomic prediction when some animals are not genotyped. *Genet. Select. Evol. GSE*, **42**, 2.

Doerks,T. (2002) Systematic identification of novel protein domain families associated with nuclear functions. *Genome Res.*, **12**, 47–56.

Fisher,R.A. (1918) The correlation between relatives on the supposition of Mendelian. *Philos. Trans. Royal Soc. Edinburgh*, **52**, 399–433.

Forsberg,S.K. *et al.* (2017) Accounting for genetic interactions improves modeling of individual quantitative trait phenotypes in yeast. *Nat. Genet.*, **49**, 497–503.

Gabriel,S.B. *et al.* (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.

Gianola,D. *et al.* (2009) Additive genetic variability and the Bayesian alphabet. *Genetics*, **183**, 347–363.

Henderson,C. (1949) Estimation of changes in herd environment. *J. Dairy Sci*, **32**, 706–715.

Henderson,C. (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics*, **31**, 423–447.

Henderson,C. (1985) Best linear unbiased prediction of nonadditive genetic merits. *J. Anim. Sci.*, **60**, 111–117.

Jarvis,J.P. and Cheverud,J.M. (2011) Mapping the epistatic network underlying murine reproductive fatpad variation. *Genetics*, **187**, 597–610.

Jiang,Y. and Reif,J.C. (2015) Modeling epistasis in genomic selection. *Genetics*, **201**, 759–768.

Kang,H.M. *et al.* (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.*, **42**, 348–354.

Kang,H.M. *et al.* (2008) Efficient control of population structure in model organism association mapping. *Genetics*, **178**, 1709–1723.

Kruijer,W. *et al.* (2015) Marker-based estimation of heritability in immortal populations. *Genetics*, **199**, 379–398.

Legarra,A. *et al.* (2009) A relationship matrix including full pedigree and genomic information. *J. Dairy Sci.*, **92**, 4656–4663.

Lippert,C. *et al.* (2013) An exhaustive epistatic SNP association analysis on expanded Wellcome Trust data. *Sci. Rep.*, **3**, 1099.

Lippert,C. *et al.* (2011) FaST linear mixed models for genome-wide association studies. *Nat. Methods*, **8**, 833–835.

Liu,X. *et al.* (2016) Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet.*, **12**, e1005767.

Mackay,T.F. and Moore,J.H. (2014) Why epistasis is important for tackling complex human disease genetics. *Genome Med.*, **6**, 124.

Maki-Tanila,A. and Hill,W.G. (2014) Influence of gene interaction on complex trait variation with multilocus models. *Genetics*, **198**, 355–367.

Meuwissen,T.H. *et al.* (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, **157**, 1819–1829.

Schupbach,T. *et al.* (2010) FastEpistasis: a high performance computing solution for quantitative trait epistasis. *Bioinformatics*, **26**, 1468–1469.

Shen,X. *et al*. (2013) A novel generalized ridge regression method for quantitative genetics. *Genetics*, **193**, 1255–1268.

Speed,D. and Balding,D.J. (2014) MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.*, **24**, 1550–1557.

Stranden,I. and Garrick,D.J. (2009) Technical note: derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *J. Dairy Sci.*, **92**, 2971–2975.

Su,G. *et al*. (2012) Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. *PloS One*, **7**, e45293.

Upton,A. *et al*. (2016) Review: high-performance computing to detect epistasis in genome scale data sets. *Brief. Bioinf.*, **17**, 368–379.

VanRaden,P.M. (2008) Efficient methods to compute genomic predictions. *J. Dairy Sci.*, **91**, 4414–4423.

Wan,X. *et al*. (2010) BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am. J. Hum. Genet.*, **87**, 325–340.

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.

Wu,L. *et al*. (2014) Variants associated with susceptibility to pancreatic cancer and melanoma do not reciprocally affect risk. *Cancer Epidemiol. Biomark. Prevent. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prevent. Oncol.*, **23**, 1121–1124.

Xu,S. (2003) Estimating polygenic effects using markers of the entire genome. *Genetics*, **163**, 789–801.

Xu,S. (2013) Mapping quantitative trait loci by controlling polygenic background effects. *Genetics*, **195**, 1209–1222.

Yang,J. *et al*. (2014) Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.*, **46**, 100–106.

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

Zhang,F. *et al*. (2014) Epistasis analysis for quantitative traits by functional regression model. *Genome Res*., **24**, 989–998.

Zhang,Z. *et al*. (2010) Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.*, **42**, 355–360.

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