Rapid Epistatic Mixed Model Association Studies by Controlling Multiple Polygenic Effects

Dan Wang¹, Hui Tang¹, Jian-Feng Liu², Shizhong Xu³, Qin Zhang^{1,*} and Chao Ning^{1,*}

¹Shandong Provincial Key Laboratory of Animal Biotechnology and Disease Control and Prevention, College of Animal Science and Technology, Shandong Agricultural University, Tai'an 271018, China

²National 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,

³Department of Botany and Plant Science, University of California, Riverside 20705, USA.

*To whom correspondence should be addressed: Qin Zhang (qzhang@sdau.edu.cn) and Chao Ning (ningchao@sdau.edu.cn).

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29

Summary: We have developed a rapid mixed model algorithm for exhaustive genome-wide epistatic association analysis by controlling multiple polygenic effects. Our model can simultaneously handle additive by additive epistasis, dominance by dominance epistasis and additive by dominance epistasis, and account for intrasubject fluctuations due to individuals with repeated records. Furthermore, we suggest a simple but efficient approximate algorithm, which allows examination of all pairwise interactions in a remarkably fast manner of linear with population size. Application to publicly available yeast and human data has showed that our mixed model-based method has similar performance with simple linear model-based Plink on computational efficiency. It took less than 40 hours for the pairwise analysis of 5,000 individuals genotyped with roughly 350,000 SNPs with five threads on Intel Xeon E5 2.6GHz CPU.

Availability and implementation: Source codes are freely available at <https://github.com/chaoning/GMAT>.

1

2 **1. Introduction**

3 Epistasis is defined as the deviation of the combined allele effect at two or more loci
4 from the sum of their individual effects (Fisher, 1918). It is considered a promising
5 approach to understanding genetic causes of complex traits (Mackay and Moore, 2014;
6 Phillips, 2008; Upton, et al., 2016).

7 Linear mixed models have been widely used in genome-wide association studies
8 (GWAS) due to its advantage in correcting environmental factors, controlling
9 population stratification and accounting for cryptic relatedness between individuals
10 (Kang, et al., 2008; Lippert, et al., 2011; Zhou and Stephens, 2012). However,
11 computational inefficiency hindered the application of linear mixed models in
12 performing exhaustive epistatic association studies with the increases of sample size
13 and marker density. For example, Lippert, et al. (2013) applied linear mixed models to
14 the exhaustive epistatic SNP association studies across seven traits for 14,925
15 individuals genotyped with 356,441 SNPs. It would take 950 computer years, with a
16 wall-clock time of 13 days using 28,000 cores.

17 To address the issues, we developed a rapid epistatic mixed model association
18 study (REMMA) in previous studies (Ning, et al., 2018). REMMA has an advantage
19 over existing linear mixed models in terms of high computational efficiency, reduced
20 type I error and high QTN detection power. However, the time complexity of
21 computation is still $O(n^2)$, where n is the population size. Furthermore, REMMA only
22 focuses on the additive by additive epistatic analysis and cannot deal with study
23 population consisting of individuals with repeated records. We report here a REMMA
24 eXpedited (REMMAX) version of the method. REMMAX can simultaneously deal
25 with the association studies for additive by additive epistasis, dominance by
26 dominance epistasis and additive by dominance epistasis by incorporating multiple
27 polygenic effects as background control. Furthermore, individual-specific residual
28 effect has been included in the model to control intrasubject fluctuations due to
29 individuals with repeated records. To improve the computational efficiency, we also

propose an approximate REMMAX algorithm, which first implements an approximate statistical test with computational complexity of $O(n)$ to filter out many nonsignificant interactions and then employs the exact Wald Chi-square test to test the interaction effects of selected SNP pairs. Experiments on yeast (Bloom, et al., 2015) and WTCCC (Wellcome Trust Case Control, 2007) data sets show that REMMAX is an efficient method to perform the exhaustive epistatic analysis.

2. Methods

2.1 REMMAX Model

A linear mixed model including additive and non-additive SNP effects can be written as

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{M}_a\mathbf{u}_a + \mathbf{Z}\mathbf{M}_d\mathbf{u}_d + \mathbf{Z}\mathbf{M}_{aa}\mathbf{u}_{aa} + \mathbf{Z}\mathbf{M}_{ad}\mathbf{u}_{ad} + \mathbf{Z}\mathbf{M}_{dd}\mathbf{u}_{dd} + \mathbf{Z}\mathbf{p} + \mathbf{e} \quad (1)$$

where $\mathbf{y}$ is a vector of N observations for n individuals (including repeated measurements). Below is the description of the model effects and their design matrices. Vector $\mathbf{b}$ is a collection of fixed effects. Vectors $\mathbf{u}_a$, $\mathbf{u}_d$, $\mathbf{u}_{aa}$, $\mathbf{u}_{ad}$ and $\mathbf{u}_{dd}$ are additive, dominance, additive by additive, additive by dominance and dominance by dominance SNP effects, respectively. Vector $\mathbf{p}$ is an array of individual-specific residual effects. Vector $\mathbf{e}$ is a collection of random residual errors. $\mathbf{X}$ is a design matrix for the fixed effects. $\mathbf{Z}$ is a design matrix for all the random model effects, serving as connection between the model effects and observed data points in the response variable. $\mathbf{M}_a$, $\mathbf{M}_d$, $\mathbf{M}_{aa}$, $\mathbf{M}_{ad}$ and $\mathbf{M}_{dd}$ are additive, dominance, additive by additive, additive by dominance and dominance by dominance SNP matrixes, respectively. They are formalized as

$$\begin{aligned} \mathbf{M}_a &= [\mathbf{m}_{a,i}]_{n \times m}, \mathbf{M}_d = [\mathbf{m}_{d,i}]_{n \times m}, \mathbf{M}_{aa} = [\mathbf{m}_{a,i} \# \mathbf{m}_{a,j}]_{n \times [m*(m-1)/2]}^{i \neq j, i < j}, \\ \mathbf{M}_{ad} &= [\mathbf{m}_{a,i} \# \mathbf{m}_{d,j}]_{n \times [m*(m-1)]}^{i \neq j}, \mathbf{M}_{dd} = [\mathbf{m}_{d,i} \# \mathbf{m}_{d,j}]_{n \times [m*(m-1)/2]}^{i \neq j, i < j} \end{aligned} \quad (2)$$

Here, m is the number of SNPs; $\#$ is Hadamard (element-wise) product. $\mathbf{m}_{a,i}$ and $\mathbf{m}_{d,i}$ are additive and dominance SNP vectors for the i th SNP with elements defined

1 as:

$$2 \quad \mathbf{m}_{a,i} = \begin{cases} 2-2p_i & AA \\ 1-2p_i & Aa \\ 0-2p_i & aa \end{cases}, \mathbf{m}_{d,i} = \begin{cases} 0-2p_i(1-p_i) & AA \\ 1-2p_i(1-p_i) & Aa \\ 0-2p_i(1-p_i) & aa \end{cases} \quad (3)$$

3 Where p_i is the allele frequency of allele A for the i th SNP. Below are the assumed
4 normal distributions for the random effects:

$$5 \quad \mathbf{u}_a \sim N(\mathbf{0}, \mathbf{I}_n \sigma_a^2), \mathbf{u}_d \sim N(\mathbf{0}, \mathbf{I}_n \sigma_d^2), \mathbf{u}_{aa} \sim N(\mathbf{0}, \mathbf{I}_{m^*(m-1)/2} \sigma_{aa}^2), \\ \mathbf{u}_{ad} \sim N(\mathbf{0}, \mathbf{I}_{m^*(m-1)} \sigma_{ad}^2), \mathbf{u}_{dd} \sim N(\mathbf{0}, \mathbf{I}_{m^*(m-1)/2} \sigma_{dd}^2), \mathbf{p} \sim N(\mathbf{0}, \mathbf{I}_n \sigma_p^2), \mathbf{e} \sim N(\mathbf{0}, \mathbf{I}_m \sigma_e^2) \quad (4)$$

6 Here, $\sigma_a^2, \sigma_d^2, \sigma_{aa}^2, \sigma_{ad}^2$ and σ_{dd}^2 are additive, dominance, additive by additive,
7 additive by dominance and dominance by dominance variances; $\mathbf{I}_*$ is identity matrix;
8 Finally, σ_p^2 and σ_e^2 are individual-specific error and random error variances,
9 respectively. The variances can be estimated by maximizing the following restricted
10 log-likelihood function with efficient combined expectation maximization (EM) and
11 average information (AI) algorithm (Ning, et al., 2019).

$$12 \quad L = -\frac{1}{2}(\log |\mathbf{V}| + \log |\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| + \mathbf{y}'\mathbf{P}\mathbf{y}) \quad (5)$$

13 Where

$$14 \quad \mathbf{V} = \text{var}(\mathbf{y}) = \mathbf{Z}\mathbf{M}_a\mathbf{M}_a' \mathbf{Z}' \sigma_a^2 + \mathbf{Z}\mathbf{M}_d\mathbf{M}_d' \mathbf{Z}' \sigma_d^2 + \mathbf{Z}\mathbf{M}_{aa}\mathbf{M}_{aa}' \mathbf{Z}' \sigma_{aa}^2 \\ + \mathbf{Z}\mathbf{M}_{ad}\mathbf{M}_{ad}' \mathbf{Z}' \sigma_{ad}^2 + \mathbf{Z}\mathbf{M}_{dd}\mathbf{M}_{dd}' \mathbf{Z}' \sigma_{dd}^2 + \mathbf{Z}\mathbf{Z}' \sigma_p^2 + \mathbf{I}_m \sigma_e^2 \quad (6)$$

$$15 \quad \mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1} \quad (7)$$

16 However, the calculation of $\mathbf{V}$ is inefficient due to high-dimensional $\mathbf{M}_{aa}$, $\mathbf{M}_{ad}$
17 and $\mathbf{M}_{dd}$. According to previous studies (Ning, et al., 2018; Su, et al., 2012;
18 VanRaden, 2008; Xu, 2013), $\mathbf{M}_a\mathbf{M}_a'$, $\mathbf{M}_d\mathbf{M}_d'$, $\mathbf{M}_{aa}\mathbf{M}_{aa}'$, $\mathbf{M}_{ad}\mathbf{M}_{ad}'$ and $\mathbf{M}_{dd}\mathbf{M}_{dd}'$
19 can be used to respectively reflect the additive, dominance, additive by additive,
20 additive by dominance and dominance by dominance genomic relationship among
21 individuals with the following equation.

$$22 \quad \mathbf{K}_a = \mathbf{M}_a\mathbf{M}_a' / \lambda_a \quad (8)$$

$$\mathbf{K}_d = \mathbf{M}_d \mathbf{M}_d' / \lambda_d \quad (9)$$

$$\begin{aligned} \mathbf{K}_{aa} &= \mathbf{M}_{aa} \mathbf{M}_{aa}' / (\lambda_a \lambda_a) = (\mathbf{M}_a \mathbf{M}_a' / \lambda_a) \# (\mathbf{M}_a \mathbf{M}_a' / \lambda_a) \\ &\quad - (\mathbf{M}_a \# \mathbf{M}_a') (\mathbf{M}_a \# \mathbf{M}_a')' / (\lambda_a \lambda_a) \approx \mathbf{K}_a \# \mathbf{K}_a \end{aligned} \quad (10)$$

$$\begin{aligned} \mathbf{K}_{ad} &= \mathbf{M}_{ad} \mathbf{M}_{ad}' / (\lambda_a \lambda_d) = (\mathbf{M}_a \mathbf{M}_a' / \lambda_a) \# (\mathbf{M}_d \mathbf{M}_d' / \lambda_d) \\ &\quad - (\mathbf{M}_a \# \mathbf{M}_d') (\mathbf{M}_a \# \mathbf{M}_d')' / (\lambda_a \lambda_d) \approx \mathbf{K}_a \# \mathbf{K}_d \end{aligned} \quad (11)$$

$$\begin{aligned} \mathbf{K}_{dd} &= \mathbf{M}_{dd} \mathbf{M}_{dd}' / (\lambda_d \lambda_d) = (\mathbf{M}_d \mathbf{M}_d' / \lambda_d) \# (\mathbf{M}_d \mathbf{M}_d' / \lambda_d) \\ &\quad - (\mathbf{M}_d \# \mathbf{M}_d') (\mathbf{M}_d \# \mathbf{M}_d')' / (\lambda_d \lambda_d) \approx \mathbf{K}_d \# \mathbf{K}_d \end{aligned} \quad (12)$$

Where, $\lambda_a = \sum 2p_i(1-p_i)$, $\lambda_d = \sum 2p_i(1-p_i)(1-2p_i(1-p_i))$. Apply equation (8-12)

to (6), $\mathbf{V}$ can be efficiently calculated with time complexity of $O(n^2m)$ in the

following equations

$$\begin{aligned} \mathbf{V} = \text{var}(\mathbf{y}) &= \mathbf{ZK}_a \mathbf{Z}' \lambda_a \sigma_a^2 + \mathbf{ZK}_d \mathbf{Z}' \lambda_d \sigma_d^2 + \mathbf{ZK}_{aa} \mathbf{Z}' \lambda_a \lambda_a \sigma_{aa}^2 \\ &\quad + \mathbf{ZK}_{ad} \mathbf{Z}' \lambda_a \lambda_d \sigma_{ad}^2 + \mathbf{ZK}_{dd} \mathbf{Z}' \lambda_d \lambda_d \sigma_{dd}^2 + \mathbf{ZZ}' \sigma_p^2 + \mathbf{I}_m \sigma_e^2 \end{aligned} \quad (13)$$

Following Henderson (1975), the random SNP effects can be estimated with the

following equations:

$$\begin{aligned} \hat{\mathbf{u}}_a &= (\mathbf{I}_n \sigma_a^2) (\mathbf{ZM}_a)' \mathbf{Py}, \hat{\mathbf{u}}_d = (\mathbf{I}_n \sigma_d^2) (\mathbf{ZM}_d)' \mathbf{Py}, \hat{\mathbf{u}}_{aa} = (\mathbf{I}_{m^*(m-1)/2} \sigma_{aa}^2) (\mathbf{ZM}_{aa})' \mathbf{Py}, \\ \hat{\mathbf{u}}_{ad} &= (\mathbf{I}_{m^*(m-1)} \sigma_{ad}^2) (\mathbf{ZM}_{ad})' \mathbf{Py}, \hat{\mathbf{u}}_{dd} = (\mathbf{I}_{m^*(m-1)/2} \sigma_{dd}^2) (\mathbf{ZM}_{dd})' \mathbf{Py} \end{aligned} \quad (14)$$

The corresponding variances are

$$\begin{aligned} \text{var}(\hat{\mathbf{u}}_a) &= (\mathbf{I}_n \sigma_a^2) (\mathbf{ZM}_a)' \mathbf{PVP} (\mathbf{ZM}_a) (\mathbf{I}_n \sigma_a^2), \\ \text{var}(\hat{\mathbf{u}}_d) &= (\mathbf{I}_n \sigma_d^2) (\mathbf{ZM}_d)' \mathbf{PVP} (\mathbf{ZM}_d) (\mathbf{I}_n \sigma_d^2), \\ \text{var}(\hat{\mathbf{u}}_{aa}) &= (\mathbf{I}_{m^*(m-1)/2} \sigma_{aa}^2) (\mathbf{ZM}_{aa})' \mathbf{PVP} (\mathbf{ZM}_{aa}) (\mathbf{I}_{m^*(m-1)/2} \sigma_{aa}^2), \\ \text{var}(\hat{\mathbf{u}}_{ad}) &= (\mathbf{I}_{m^*(m-1)} \sigma_{ad}^2) (\mathbf{ZM}_{ad})' \mathbf{PVP} (\mathbf{ZM}_{ad}) (\mathbf{I}_{m^*(m-1)} \sigma_{ad}^2), \\ \text{var}(\hat{\mathbf{u}}_{dd}) &= (\mathbf{I}_{m^*(m-1)/2} \sigma_{dd}^2) (\mathbf{ZM}_{dd})' \mathbf{PVP} (\mathbf{ZM}_{dd}) (\mathbf{I}_{m^*(m-1)/2} \sigma_{dd}^2) \end{aligned} \quad (15)$$

For the i th and j th SNPs, the estimated SNP effects and corresponding variances

are

$$\begin{aligned} \hat{\mathbf{u}}_{a,i} &= \mathbf{m}_{a,i}' \mathbf{Z}' \mathbf{Py} \sigma_a^2, \text{var}(\hat{\mathbf{u}}_{a,i}) = \mathbf{m}_{a,i}' \mathbf{Z}' \mathbf{PV} (\sigma_a^2)^2 \mathbf{PZ} \mathbf{m}_{a,i}, \\ \hat{\mathbf{u}}_{d,i} &= \mathbf{m}_{d,i}' \mathbf{Z}' \mathbf{Py} \sigma_d^2, \text{var}(\hat{\mathbf{u}}_{d,i}) = \mathbf{m}_{d,i}' \mathbf{Z}' \mathbf{PV} (\sigma_d^2)^2 \mathbf{PZ} \mathbf{m}_{d,i}, \\ \hat{\mathbf{u}}_{aa,ij} &= (\mathbf{m}_{a,i} \# \mathbf{m}_{a,j})' \mathbf{Z}' \mathbf{Py} \sigma_{aa}^2, \text{var}(\hat{\mathbf{u}}_{aa,ij}) = (\mathbf{m}_{a,i} \# \mathbf{m}_{a,j})' \mathbf{Z}' \mathbf{PV} (\sigma_{aa}^2)^2 \mathbf{PZ} (\mathbf{m}_{a,i} \# \mathbf{m}_{a,j}), \\ \hat{\mathbf{u}}_{ad,ij} &= (\mathbf{m}_{a,i} \# \mathbf{m}_{d,j})' \mathbf{Z}' \mathbf{Py} \sigma_{ad}^2, \text{var}(\hat{\mathbf{u}}_{ad,ij}) = (\mathbf{m}_{a,i} \# \mathbf{m}_{d,j})' \mathbf{Z}' \mathbf{PV} (\sigma_{ad}^2)^2 \mathbf{PZ} (\mathbf{m}_{a,i} \# \mathbf{m}_{d,j}), \\ \hat{\mathbf{u}}_{dd,ij} &= (\mathbf{m}_{d,i} \# \mathbf{m}_{d,j})' \mathbf{Z}' \mathbf{Py} \sigma_{dd}^2, \text{var}(\hat{\mathbf{u}}_{dd,ij}) = (\mathbf{m}_{d,i} \# \mathbf{m}_{d,j})' \mathbf{Z}' \mathbf{PV} (\sigma_{dd}^2)^2 \mathbf{PZ} (\mathbf{m}_{d,i} \# \mathbf{m}_{d,j}) \end{aligned}$$

1 (16)

2 The Wald Chi-squared test is defined as:

$$\begin{aligned}
 & \hat{\mathbf{u}}_{a,i}^2 / \text{var}(\hat{\mathbf{u}}_{a,i}) \sim \chi^2(1), \\
 & \hat{\mathbf{u}}_{d,i}^2 / \text{var}(\hat{\mathbf{u}}_{d,i}) \sim \chi^2(1), \\
 & \hat{\mathbf{u}}_{aa,ij}^2 / \text{var}(\hat{\mathbf{u}}_{aa,ij}) \sim \chi^2(1), \\
 & \hat{\mathbf{u}}_{ad,ij}^2 / \text{var}(\hat{\mathbf{u}}_{ad,ij}) \sim \chi^2(1), \\
 & \hat{\mathbf{u}}_{dd,ij}^2 / \text{var}(\hat{\mathbf{u}}_{dd,ij}) \sim \chi^2(1).
 \end{aligned}
 \tag{17}$$

4 Note that $\mathbf{Z}'\mathbf{P}\mathbf{y}$ and $\mathbf{Z}'\mathbf{P}\text{var}(\mathbf{y})\mathbf{P}\mathbf{Z}$ can be precalculated. The SNP effects are
5 calculated by vector-vector multiplication whose computational complexity is $O(N)$,
6 while the variances of estimated effects are calculated by vector-matrix-vector
7 multiplication whose computational complexity is $O(N^2)$.

8

9 2.2 Approximate algorithm

10 To reduce the time complexity, an efficient algorithm with the approximate Wald
11 Chi-squared statistical test is implemented as following
12 Step 1: Randomly select k (e.g. 100,000) SNP pairs and calculate their epistatic effects
13 and the corresponding variances. The median of variances is defined as $\text{var}(\hat{\mathbf{u}})_{\text{median}}$.
14 This step has time complexity of $O(kN^2)$
15 Step 2: Scan the genome-wide epistatic effects and calculate the approximate Wald
16 Chi-squared statistical test with $\hat{\mathbf{u}}_{ij}^2 / \text{var}(\hat{\mathbf{u}})_{\text{median}} \sim \chi^2(1)$. This step has time
17 complexity of $O(Tm)$, where T is the total number of pairwise test. T is $m(m-1)/2$ for
18 additive by additive and dominance by dominance epistatic test, and is $m(m-1)$ for
19 additive by dominance epistatic test, where m is the number of SNPs.
20 Step 3: Select top significant SNP pairs based on the approximate Wald Chi-squared
21 statistical test using a p -value threshold (e.g. $p\text{-cut} = 10^{-5}$). This step has a time
22 complexity of $O(rN^2)$, where r is the number of top significant SNP pairs.

23 For population with over hundred thousand SNPs, kN and rN have similar size to
24 T . Therefore, for each SNP pair epistatic test, the time complexity can be

1 approximated by $[O(km^2)+O(Tm)+O(rm^2)]/T \approx O(m)$.

2

3 **2.3 Data Analysis**

4 We applied the REMMAX method to the yeast data (Bloom, et al., 2015) and the
5 WTCCC data (Wellcome Trust Case Control, 2007). The yeast data contained 4,390
6 BY×BM segregants genotyped with 28,220 markers (no heterozygous genotype).
7 Twenty quantitative traits with repeated measurements were analyzed with the
8 following model

$$9 \quad \mathbf{y} = \mathbf{Xb} + \mathbf{ZM}_a \mathbf{u}_a + \mathbf{ZM}_{aa} \mathbf{u}_{aa} + \mathbf{Zp} + \mathbf{e} \quad (18)$$

10 The rheumatoid arthritis (RA) and type-I diabetes (T1D) traits were analyzed for
11 the WTCCC data. After quality control of genotypes following Ning, et al. (2018),
12 4,961 individuals genotyped by 353,859 SNPs for RA and 4,962 individuals
13 genotyped by 353,751 for T1D were remained for subsequent analysis. As no
14 individuals has repeated records, $\mathbf{p}$ is removed from model and $\mathbf{Z}$ is the identity
15 matrix. Then, the data was analyzed with the following model

$$16 \quad \mathbf{y} = \mathbf{Xb} + \mathbf{M}_a \mathbf{u}_a + \mathbf{M}_d \mathbf{u}_d + \mathbf{M}_{aa} \mathbf{u}_{aa} + \mathbf{M}_{ad} \mathbf{u}_{ad} + \mathbf{M}_{dd} \mathbf{u}_{dd} + \mathbf{e} \quad (19)$$

17 To observe whether different levels of polygenic effects will affect the additive by
18 additive epistatic test, the following models were also included to analyze WTCCC
19 data.

$$20 \quad \text{REMMAX-A_AA:} \quad \mathbf{y} = \mathbf{Xb} + \mathbf{M}_a \mathbf{u}_a + \mathbf{M}_{aa} \mathbf{u}_{aa} + \mathbf{e} \quad (20)$$

$$21 \quad \text{REMMAX-A_D_AA:} \quad \mathbf{y} = \mathbf{Xb} + \mathbf{M}_a \mathbf{u}_a + \mathbf{M}_d \mathbf{u}_d + \mathbf{M}_{aa} \mathbf{u}_{aa} + \mathbf{e} \quad (21)$$

22 Here, the REMMAX-A_AA model with only additive and additive by additive SNP
23 effects is the same to REMMA method of our previous studies (Ning, et al., 2018).

24

Figure 1. Comparisons of additive by additive epistatic P values from models with different levels of polygenic effects in the analysis of RA and T1D traits of WTCCC data. REMMAX-A_AA, REMMAX-A_D_AA and REMMAX-ALL refers to equation (20), (21) and (19), respectively. Random 100,000 SNP pairs were selected for comparison.

Figure 2. The quantile-quantile plots of exact (red) and approximate (blue) Wald Chi-squared statistical tests in testing additive by additive (left), additive by dominance (middle) and dominance by dominance (right) epistatic SNP effects. We random selected 100,000 SNP pairs to estimate the null distribution.

1

2

3 **Figure 3.** Comparisons of P values obtained by exact (X axis) and approximate (Y
4 axis) REAMMX statistical tests for additive by additive (left), additive by dominance
5 (middle) and dominance by dominance epistasis (right) in the WTCCC data. The blue
6 lines mean lines of best fit, while black lines mean $y=x$.

7

8 **3. Results**

9 We observed high Pearson correlation coefficients of additive by additive epistatic P
10 values from models with different levels of polygenic effects (**Figure 1**). It indicates
11 that if we focus on additive by additive epistatic test, previous REMMA method can
12 also work well.

13 In order to assess whether REMMAX can control type I error well, we randomly
14 shuffled analysis SNP across individuals at each locus, which can purposely destroy
15 the association of the phenotypes with the scanned SNP. Under the expectation that
16 random SNP pairs are unlinked to polymorphisms controlling these traits, the
17 cumulative P -value distribution follows a uniform distribution of $U(0, 1)$. The **Figure**
18 **2 and Supplementary Figure 1** show that the type I errors are well controlled by the

exact Wald Chi-squared statistical tests, but are not controlled by approximate Wald Chi-squared statistical tests in some cases.

Then, we compared the *P*-values obtained by the approximate and exact Wald Chi-squared statistical tests in the real data analysis of the yeast and WTCCC data. A high correlation coefficient was observed in the WTCCC data for the RA and T1D traits (**Figure 3**). The Pearson correlation coefficients between the approximate and exact test are all very closed to unity for the yeast data (**Supplementary Figure 2**). Integrating the results of type I error and high correlation coefficients of approximate and exact Wald Chi-squared statistical tests, the approximate REMMAX method, which first implements an approximate statistical test to filter out many nonsignificant interactions and then employs the exact Wald Chi-square test to confirm the true significant SNP pairs, is theoretically feasible to genome-wide epistatic association studies.

In the analysis of 20 traits in yeast data, we located 665,195 significant SNP pairs by approximate REMMAX after Bonferroni correction. Only nine SNP pairs were missed compared to exact REMMAX (**Supplementary Table 1**). As it is very time consuming to apply the exact REMMAX to the genome-wide pairwise analysis for the WTCCC data, we only applied the approximate REMMAX method to the WTCCC data. It took no more than 40 hours for each the pairwise analysis with five threads on Intel Xeon E5 2.6GHz CPU. The predicted time is one year for exact REMMAX method under the same condition. For the RA trait in the WTCCC data, the approximate REMMAX method located 978 significant SNP pairs (198 for additive by additive, 606 for additive by dominance and 174 for dominance by dominance), and 664 of them (67.9%) were in the MHC region (chr6:28,510,120-33,480,577). For T1D trait, we also observed most of the significant SNP pairs (2808/2836 for additive by additive, 1404/2303 for additive by dominance and 3555/3619 for dominance by dominance) were in the MHC region. These results are consistent with previous studies (Ning, et al., 2018; Wan, et al., 2010). For comparison, we applied the exact REMMAX method to pairwise analysis of the

chromosome 6. Taking the results of exact REMMAX as the gold standard, approximate REMMAX located most of the significant SNP pairs (94.6% for RA and 99.8% for T1D).

We further compared the time complexities and the actual running times of approximate REMMA and exact REMMAX compared to simple linear model implemented by Plink (Purcell, et al., 2007) (**Table 1**). Simple linear model is the most efficient with time complexity of $O(N)$. As k and r are much smaller than T , the time complexity for REMMAX is nearly linear with N . For both the yeast and human data, approximate REMMAX was able to finish the exhaustive analysis within a tolerable time frame over a hundred times faster than exact REMMAX.

Table 1. Time complexities and computing times of simple linear model, exact REMMAX and approximate REMMAX for exhaustive genome-wide epistatic studies.

Method	Time complexity	Computing time	
		Yeast	WTCCC
Simple linear model	$O(N)$	5.4min	4.7min
exact REMMAX	$O(N^2)$	32.79h	38h
approximate REMMAX	$O(kN^2) + O(TN) + O(rN^2)$	6.5min	7.2min

Note: All computing was performed with five threads on Intel Xeon E5 2.6GHz CPU. N is the number of observations for all individuals; k is the number of randomly selected SNP pairs; T is the total number of pair-wise tests; r is the remaining number of SNP pairs filtered by the approximate Wald Chi-squared statistical test. The simple linear model is implemented by Plink. As Plink cannot deal with study population of individuals with repeated records, we used the average values as the response variable for Plink in the yeast data. We only applied the additive by additive epistatic analysis to the chromosome 6 in the WTCCC data, as Plink can only deal with additive by additive epistatic analysis and it is time consuming for exact REMMAX to perform

the exhaustive analysis.

4. Discussion

In this study, we proposed a rapid epistatic mixed model association studies method, REMMAX. REMMAX can handle the additive by additive epistasis, additive by dominance epistasis and dominance by dominance epistasis, and individual-specific residual effect has been included in the model to analyze data with repeated measurements. Furthermore, we build an approximate Wald Chi-squared statistical test for REMMAX to expedite the exhaustive genome-wide epistatic association studies. The time complexity is nearly linear with the population size. Experiments on yeast and WTCCC data sets show that REMMAX is an efficient method to help understand genetic constitutions of complex traits.

Our previous REMMA (Ning, et al., 2018) method is the special case of REMMAX, where only genome-wide additive and additive by additive SNP effects are included. When we analyze the cross-sectional data (one record per individual), the individual-specific residual effect is same to random residual and should be removed from the model. In the study, we have learned that different levels of polygenic effects have little influence on the additive by additive epistatic association studies. Therefore, the REMMA method is also preferred when we focus on the additive by additive analysis of cross-sectional data. Additionally, in the estimation of variances of estimated SNP effects, REMMA has to calculate the inverse of huge mixed model equation which is affected by the number of random effects included in linear mixed model. For REMMAX, the computational bottleneck is to calculate the inverse of phenotypic covariance matrix ($\mathbf{V}$), whose dimension is always of N (N is equal to n for cross-sectional data) regardless of the number of random effects. For example, the dimension of mixed model equation and phenotypic covariance matrix is $6n$ and N , respectively. Therefore, the amount of computation and memory is reduced to about $1/27$, thus the computation time and memory consumption is shortened significantly.

1

2 **Acknowledgments**

3 This study makes use of data generated by the Wellcome Trust Case-Control
4 Consortium. A full list of the investigators who contributed to the generation of the
5 data is available from www.wtccc.org.uk. Funding for the project was provided by the
6 Wellcome Trust under award 076113 and 085475.

7

8 **References**

- 9 Bloom, J.S., *et al.* Genetic interactions contribute less than additive effects to quantitative trait
10 variation in yeast. *Nature communications* 2015;6:8712.
- 11 Fisher, R.A. The correlation between relatives on the supposition of Mendelian. 1918.
- 12 Henderson, C.R. Best linear unbiased estimation and prediction under a selection model.
13 *Biometrics* 1975:423-447.
- 14 Kang, H.M., *et al.* Efficient control of population structure in model organism association
15 mapping. In, *Genetics*. 2008. p. 1709-1723.
- 16 Lippert, C., *et al.* An exhaustive epistatic SNP association analysis on expanded Wellcome
17 Trust data. *Scientific reports* 2013;3:1099.
- 18 Lippert, C., *et al.* FaST linear mixed models for genome-wide association studies. *Nature*
19 *methods* 2011;8(10):833-835.
- 20 Mackay, T.F. and Moore, J.H. Why epistasis is important for tackling complex human disease
21 genetics. *Genome medicine* 2014;6(6):124.
- 22 Ning, C., *et al.* A rapid epistatic mixed-model association analysis by linear retransformations
23 of genomic estimated values. *Bioinformatics* 2018;34(11):1817-1825.
- 24 Ning, C., *et al.* Efficient multivariate analysis algorithms for longitudinal genome-wide

1 association studies. *Bioinformatics* 2019;35(23):4879-4885.

2 Phillips, P.C. Epistasis--the essential role of gene interactions in the structure and evolution of

3 genetic systems. *Nature reviews. Genetics* 2008;9(11):855-867.

4 Purcell, S., *et al.* PLINK: a tool set for whole-genome association and population-based

5 linkage analyses. *The American Journal of Human Genetics* 2007;81(3):559-575.

6 Su, G., *et al.* Estimating additive and non-additive genetic variances and predicting genetic

7 merits using genome-wide dense single nucleotide polymorphism markers. *PloS one*

8 2012;7(9):e45293.

9 Upton, A., *et al.* Review: High-performance computing to detect epistasis in genome scale

10 data sets. *Briefings in bioinformatics* 2016;17(3):368-379.

11 VanRaden, P.M. Efficient methods to compute genomic predictions. *Journal of dairy science*

12 2008;91(11):4414-4423.

13 Wan, X., *et al.* BOOST: A fast approach to detecting gene-gene interactions in genome-wide

14 case-control studies. *Am J Hum Genet* 2010;87(3):325-340.

15 Wellcome Trust Case Control, C. Genome-wide association study of 14,000 cases of seven

16 common diseases and 3,000 shared controls. *Nature* 2007;447(7145):661-678.

17 Xu, S. Mapping quantitative trait loci by controlling polygenic background effects. *Genetics*

18 2013;195(4):1209-1222.

19 Zhou, X. and Stephens, M. Genome-wide efficient mixed-model analysis for association

20 studies. *Nature genetics* 2012;44(7):821-824.

21