Genome-wide association studies using binned genotypes

Bingxing An¹ · Xue Gao¹ · Tianpeng Chang¹ · Jiangwei Xia² · Xiaoqiao Wang¹ · Jian Miao¹ · Lingyang Xu¹ · Lupei Zhang¹ · Yan Chen¹ · Junya Li¹ · Shizhong Xu³ · Huijiang Gao¹

Received: 26 February 2019 / Revised: 25 September 2019 / Accepted: 26 September 2019 / Published online: 22 October 2019
© The Author(s), under exclusive licence to The Genetics Society 2019

Abstract

Linear mixed models (LMM) that tests trait association one marker at a time have been the most popular methods for genome-wide association studies. However, this approach has potential pitfalls: over conservativeness after Bonferroni correction, ignorance of linkage disequilibrium (LD) between neighboring markers, and power reduction due to overfitting SNP effects. So, multiple locus models that can simultaneously estimate and test all markers in the genome are more appropriate. Based on the multiple locus models, we proposed a bin model that combines markers into bins based on their LD relationships. A bin is treated as a new synthetic marker and we detect the associations between bins and traits. Since the number of bins can be substantially smaller than the number of markers, a penalized multiple regression method can be adopted by fitting all bins to a single model. We developed an innovative method to bin the neighboring markers and used the least absolute shrinkage and selection operator (LASSO) method. We compared BIN-Lasso with SNP-Lasso and Q + K-LMM in a simulation experiment, and showed that the new method is more powerful with less Type I error than the other two methods. We also applied the bin model to a Chinese Simmental beef cattle population for bone weight association study. The new method identified more significant associations than the classical LMM. The bin model is a new dimension reduction technique that takes advantage of biological information (i.e., LD). The new method will be a significant breakthrough in associative genomics in the big data era.

Introduction

With the advent of high-throughput genotyping and sequencing technologies, genome-wide association studies (GWAS) have become the most popular methods for investigating the genetic architecture of quantitative traits in random populations. However, the ever-growing number of markers makes it difficult to automatically fit all markers

simultaneously in a single model. The classical mixed model GWAS estimates and tests trait association one marker at a time (genome scanning approach) to avoid the high dimensionality problem, but this has turned the problem into a multiple test problem instead. The linear mixed model (LMM) association study (Yu et al. 2006) has demonstrated its effectiveness in correcting family structure and cryptic relatedness among samples (Kang et al. 2010). However, there are three caveats with the above one-dimensional scanning GWAS. The first caveat is the problem of multiple tests. Single-locus GWAS requires a very conservative Bonferroni correction to determine the critical value of the *p* value used to declare statistical significance. As a result, when the number of markers is extremely large, true biologically important markers may be missed due to the over stringent test criterion (decreased power) (Wang et al. 2016). Secondly, if a trait is controlled by more than one gene, the single locus GWAS model is never true and the estimated effects of markers will be seriously biased due to linkage disequilibrium between neighboring markers (Wen et al. 2018). Finally, the LMM implemented GWAS doubly fits candidate markers because the fitted markers are also included in the genetic

These authors contributed equally: Bingxing An, Xue Gao

Supplementary information The online version of this article (<https://doi.org/10.1038/s41437-019-0279-y>) contains supplementary material, which is available to authorized users.

✉ Huijiang Gao
gaohuijiang@caas.cn

¹ Institute of Animal Sciences, Chinese Academy of Agricultural Sciences, Beijing, China

² Institute of Basic Medical Science, Westlake Institute for Advanced Study, Hangzhou, China

³ Department of Botany and Plant Sciences, University of California, Riverside, CA, USA

relationship matrix. Such a double fitting leads to a competition between the fixed effect tested and the random effect in the polygene and thus causes reduced power as a consequence (Yang et al. 2014).

The three problems associated with the genome scanning GWAS can be avoided by using a multiple regression model where all markers are simultaneously included in the model. The polygenic nature of many quantitative traits makes such a multiple marker association study more appealing than the single marker scanning approach (Segura et al. 2012; Wang et al. 2016; Wen et al. 2018). However, the ordinary least squares regression cannot handle the situations where the number of predictors (p) is larger than the sample size (n), which is often the case in GWAS. If the number of variables is indeed larger than the sample size, there is no unique solution for the estimated marker effects and the estimation errors will be arbitrarily large (Kyung et al. 2010). In many GWAS populations, the number of markers can easily reach millions whereas the sample size is often in the order of thousands. It is commonly accepted that only a small subset of markers is truly associated with any traits. Therefore, some types of variable selection, penalized regression, or Bayesian shrinkage models must be adopted in multiple marker GWAS (de Los Campos et al. 2013). Bayesian methods require Markov chain Monte Carlo sampling, which can be very intensive in terms of computational time (Wang et al. 2005; Xu 2003; Yi 2004). Several typical penalized regression models have been proposed to address the high dimensionality problem, including ridge regression (RR) (Hoerl and Kennard 1970; Shen et al. 2013), least absolute shrinkage selection operator (LASSO) (Tibshirani 1996), and elastic net (Zou and Hastie 2005) that combines the L_1 penalty of LASSO and the L_2 penalty of RR.

LASSO is the most tested method for variable selection and prediction. It uses a penalized function to simultaneously reduce dimensionality of markers (variable selection) and to penalize large estimated effects of regression coefficients (shrinkage). The cyclic coordinate descent algorithm adopted in the GLMNET/R software package (Friedman et al. 2010) is surprisingly efficient for finding the LASSO estimates of regression coefficients. While demonstrating the promising performance of LASSO for many problems, the method does have some limitations or shortcomings. For example, when the number of sample size does not match the growth of the number of variables, the model often fails (Hu et al. 2012). Therefore, we must borrow some biological information to reduce the number of variables prior to calling the LASSO method for association studies.

One piece of the biological information is the physical position of each marker on the genome. If two markers are located side by side physically, they may co-segregate or

segregate with a similar pattern. In statistics, this is called linkage disequilibrium (LD), but LD is primarily caused by physical linkage in most populations. Therefore, in most GWAS populations, LD is the biological information we can borrow for using it to reduce model dimension. We can combine multiple SNPs with high LD into a block. Markers of the entire genome may be divided into some manageable number of blocks, called bins (Hu et al. 2012; Xu 2013). Similar techniques by dividing markers into groups have been proposed by other research teams (Liu et al. 2008) with different names in different contexts, e.g., haplotypes (Xia et al. 2017), gene-based SNP groups, pathway-based SNP groups, and collapsing SNP groups and so on. These authors have shown that the group-based SNP association studies can be more powerful than the single SNP analysis. Motivated by the above studies, we proposed a bin model to reduce the data dimension by combining neighboring SNPs into bins. Once the number of bins is reduced to a manageable level, we then used LASSO to shrink and estimate the effects of bins on the phenotype of a trait under investigation. The number of bins is determined by the LD level of the genome. High LD requires a small number of bins to capture the whole genomic information while low LD needs a large number of bins to capture the same amount of information. The thumb rule is to convert SNPs into a desirable number of bins so that LASSO can comfortably handle all bins in the sample model.

It is commonly believed that multiple locus models (e.g., LASSO) are better than single marker scanning models (e.g., GEMMA). However, the multiple locus models can only handle a limited number of markers, say a few hundred thousand markers. When the number of markers reaches a million, all multiple locus models fail. When markers are binned and the bins are used in place of markers, the model dimension can be substantially reduced so that all bins can be fitted into a single model. This study is to develop an efficient method to construct bins and demonstrate that the bin model can be as efficient as multiple marker models in terms of power and Type 1 error, but the bin model can handle virtually unlimited number of markers.

Material and methods

Binning SNP genotypes

Bins defined here are different from the classical definitions of bins in a line crossing experiment involving two or a few parents (Wang et al. 2016; Xu 2013). In the classical definition, a bin is defined as a genomic segment that contains no recombination breakpoints across the entire population under study. That criterion of defining bins is too stringent for a random population that involves many (unknown

number of) parents. In this study, we define a bin as a block of physically linked SNPs with high LD. There are a few steps in creating a bin, as described below.

The first step is to code the minor allele by 0 and the major allele by 1, or the other way around, for all markers consistently. This is a very important step because if the coding system is not consistent across markers, we will not be able to capture the genomic LD information. Let A_1 be the minor allele and A_2 be the major allele. The numerical codes for A_1A_1 , A_1A_2 , and A_2A_2 should be 0, 1, and 2, respectively. Again, A_1 is the minor allele (not the reference allele) for all markers in the genome.

The second step is to standardize the genotype code variable. Let Z_{jk} be the genotypic code for all n individuals (an $n \times 1$ vector) at the j th marker within the k th bin, for $j = 1, \dots, m_k$ and $k = 1, \dots, m$, where m_k is the number of markers within the k th bin and m is the number of bins. The total number of markers in the entire genome is $M = \sum_{k=1}^m m_k$. Each Z_{jk} variable must be standardized to have a mean 0 and a standard deviation 1.

The third step is to create bins by combining neighboring SNPs. Let Z_k be an $n \times 1$ vector of bin genotypes and it is the average value of all markers in the k th bin,

$$Z_k = \frac{1}{m_k} \sum_{j=1}^{m_k} Z_{jk}. \quad (1)$$

The number of markers in the bin (m_k) is determined in such a way that $\text{var}(Z_k) > c$ but adding one more marker will make $\text{var}(Z_k) < c$, where c is a constant number between 0 and 1 set by the investigator. One can choose the c criterion so that, with this c , the number of bins is a desired number that can be comfortably handled by LASSO.

Recall that each Z_{jk} is a standardized variable with variance 1. If all the m_k markers within this bin are independent, the variance would be $\text{var}(Z_k) = 1/m_k$. If all the m_k markers within the bin are in perfect LD, the variance would be

$$\text{var}(Z_k) = \frac{1}{m_k^2} [m_k + 2m_k(m_k - 1)/2] = 1. \quad (2)$$

Let d_{ij} be the LD parameter between markers i and j within the k th bin (correlation coefficient between Z_{jk} and Z_{ik}). The actual variance of bin k is

$$\text{var}(Z_k) = \frac{1}{m_k^2} \left[m_k + 2 \sum_{i=1}^{m_k-1} \sum_{j=i+1}^{m_k} d_{ij} \right]. \quad (3)$$

Therefore, the bin variance is a value between $1/m_k$ and 1. If we set $\text{var}(Z_k) = 0.5$, then m_k will vary across bins depending on the LD of consecutive markers within bins.

Least absolute shrinkage selection operator

Assume that based on a certain c criterion we create m bins with the bin code denoted by Z_k (an $n \times 1$ vector) for the k th bin. We now use LASSO to estimate all bin effects. The LASSO method requires a centered response variable, i.e., the phenotypic values need to be adjusted by the fixed effects (including population structure effects if any) prior to the LASSO analysis. The centered y is described by the following multiple regression model:

$$y' = \sum_{k=1}^m Z_k \gamma_k + e, \quad (4)$$

where γ_k is the effect of bin k on the trait of interest. Let Z be an $n \times m$ matrix for all bin genotypes and γ be an $m \times 1$ vector for all bin effects. The LASSO estimates of the bin effects are

$$\gamma^{\text{Lasso}} = \arg \min_{\gamma \in \Omega} \{ (y' - Z\gamma)^T (y' - Z\gamma) + \lambda \|\gamma\|_1 \}, \quad (5)$$

where λ is a LASSO shrinkage factor and $\|\gamma\|_1 = \sum_{k=1}^m |\gamma_k|$. The GLMNET/R package (Friedman et al. 2010) can be used to estimate the LASSO regression coefficients. However, the package does not provide a way to perform significance tests for the regression coefficients. We adopted the method to calculate the variance of the LASSO estimated effect for a bin and used that variance to calculate the Wald test and the p value for each bin (Xu 2013). The Wald test for bin k is

$$W_k = \frac{\hat{\gamma}_k^2}{\text{var}(\hat{\gamma}_k)},$$

and the p value is

$$p_k = 1 - \Pr(\chi_1^2 < W_k),$$

where χ_1^2 is a Chi-square variable with one degree of freedom (Kao et al. 1999). The beauty of the LASSO method is that no multiple test correction is needed because all markers are included in a single model. Therefore, the nominal critical value for the $-\log_{10}(p_k)$ test statistic can be used, which is $-\log_{10}(0.05) = 1.30103$.

LMM—the gold standard for comparison

Using LASSO to analyze the bin data is our new contribution. However, we need to compare the new method with alternative methods to demonstrate the advantages of

our method over existing methods. The LMM is the routine method for GWAS and has been the gold standard for comparison. It automatically controls population structure effects (Q) and polygenic effect via a genomic relationship (kinship) matrix (K). Therefore, it is called the Q + K model (Yu et al. 2006). Let y be an $n \times 1$ vector of phenotypic values for a trait of interest, the LMM describing y is

$$y = X\beta + W_j\mu_j + \xi + e, \quad (6)$$

where X is a design matrix capturing all non-genetic (fixed) effects, e.g., systematic environmental effects, population structural effects, and so on, β are the fixed effects, W_j is a vector of numerical codes for the j th marker for $j = 1, \dots, M$ (M is the total number of markers, not the number of bins), μ_j is the effect of marker j on the trait (treated as a fixed effect), ξ is a vector of polygenic (random) effects with an assumed $N(0, K\phi^2)$ distribution where ϕ^2 is the polygenic variance and K is a marker inferred kinship matrix (defined later), e is a vector of residual errors with an assumed $N(0, I\sigma^2)$ distribution, where σ^2 is the residual error variance. The marker inferred kinship matrix is defined as

$$K = \frac{1}{d} \sum_{j=1}^M Z_j Z_j^T, \quad (7)$$

where d (a normalization factor) is the average value of the diagonal elements of the unnormalized matrix K . The restricted maximum likelihood (REML) method was used to estimate parameters $\{\phi^2, \sigma^2\}$ and the best linear unbiased estimates (BLUE) of the fixed effects, β and μ_j , were obtained via mixed model equations. The Wald test for $H_0: \mu_j = 0$ is

$$W_j = \frac{\hat{\mu}_j^2}{\text{var}(\hat{\mu}_j)}. \quad (8)$$

Under the null hypothesis, W_j follows approximately a Chi-square distribution with one degree of freedom. The corresponding p value was calculated using

$$p_j = 1 - \Pr(\chi_1^2 \leq W_j). \quad (9)$$

Genome-wide association studies

The bin model analyzed with LASSO and further refined by the second stage of tuning is the new method, denoted by BIN-Ave-Lasso, where Ave in the middle means average. An intuitive method derived from it is to use one SNP marker per bin (the middle one in the bin) to form the independent variable matrix and use LASSO for multiple

Table 1 Annotation of methods used in this study

Method	Annotation	Status
BIN-Ave-Lasso	Bin data analyzed by Lasso	New
BIN-One-Lasso	One marker per bin analyzed by Lasso	New
SNP-Lasso	SNP data analyzed by Lasso	Current
Q + K-LMM	SNP data analyzed by Q + K and LMM	Current

marker analysis. Once the middle marker in a bin is detected, a fine tuning second stage follows to evaluate all markers in the significant bins. This method is also new and is denoted by BIN-One-Lasso. The original SNPs analyzed by LASSO are an existing method, denoted by SNP-Lasso, which is used as an alternative method for comparison. The classical Q + K model implemented in LMM is another existing method for comparison and is denoted by Q + K-LMM. The four methods are summarized in Table 1. The Bin-Ave-Lasso and Bin-One-Lasso methods also include a series of bin methods using different c criteria.

Simulated data from beef genotypes

We used the genotypes of 608,696 SNPs of 1217 cattle of our own population to simulate a quantitative trait. We used the mean squared error (MSE) to measure the accuracy of an estimated quantitative trait nucleotides (QTN) effect. Of the 608,696 SNPs, we assigned 20 loci as QTN. We evaluated the accuracies for all the 20 simulated QTNs. The QTNs with positions, effects, and MSE's are demonstrated in Table 2. We also assigned each of the 608,696 SNPs, a small effect randomly sampled from a normal distribution with mean 0 and a very small variance to mimic a polygenic effect. The 20 loci collectively contributed 50% of the phenotypic variance, and the polygenic effect contributed 25% of the phenotypic variance. The remaining 25% phenotypic variance was contributed by the residual errors. All the 608,696 SNPs were used to construct bins of the cattle genome under various thresholds. A series of threshold value c for constructing bins were examined, ranging from 0.1 to 1 incremented by 0.1, where $c = 1$ is equivalent to SNP-Lasso, i.e., using the original SNPs by LASSO.

To evaluate the performance of the four methods, we generated 100 simulated samples from the above set up. The statistical power for a method was defined as the total number of QTNs detected per sample divided by 20 and then averaged over the 100 replicated samples. Type 1 error of a method was defined as the total number of SNPs above the significant thresholds in regions beyond the ± 10 kb reserved interval of a simulated true effect divided by the total number of markers above the significant threshold, and then averaged across the 100 replicated samples. All four methods were used for analysis of this simulated data.

Table 2 Information about the 20 simulated QTNs

QTN	Chr	Position (bp)	Effect	Proportion ^a (%)	MSE
1	2	25128989	0.066859	0.7665	0.1376
2	3	18849505	0.645265	7.3985	0.4144
3	3	47643448	-0.01178	0.135	0.3855
4	5	88624357	0.083782	0.9605	0.2591
5	5	98951148	-0.1257	1.4415	0.4605
6	6	11888618	-0.10072	1.155	0.3903
7	6	28262330	0.22537	2.584	0.3721
8	7	39373442	0.261224	2.995	0.2649
9	7	103925447	-0.45702	5.24	0.4617
10	8	44609878	-0.34945	4.007	0.2344
11	9	7918734	0.031867	0.3655	0.3695
12	9	78871984	-0.11516	1.3205	0.5181
13	9	102649062	0.249944	2.866	0.3014
14	13	21139804	0.283293	3.2485	0.1847
15	14	8615666	0.271053	3.108	0.2635
16	15	37134429	0.123874	1.4205	0.3743
17	16	44300159	0.428834	4.917	0.2741
18	22	4643997	0.138882	1.5925	0.3742
19	24	53584633	-0.09543	1.094	0.4344
20	29	6770481	-0.29518	3.3845	0.2813

^aProportion of the phenotypic variance contributed by each individual QTN

Actual beef data

The resource population consisted of 1225 Chinese Simmental cattle that were born in Ulagai, Inner Mongolia of China from 2008 to 2014. Each individual was observed and measured for growth and developmental traits until slaughtered at ages from 16 to 18 months old. All the 1225 cattle were genotyped by the Illumina Bovine HD Bead-Chip which contained 774,660 SNPs. Quality control procedures were carried out using the PLINK 1.9 (Purcell et al. 2007) to filter out SNPs with call rate <95%, minor allele frequency <0.05, and significant deviation from Hardy-Weinberg equilibrium ($p < 10^{-5}$). Moreover, cattle with >10% missing genotypes were removed. After filtering with the above criteria, a total of 608,696 SNPs and 1217 individuals remained in the sample for analysis. Imputation for sporadic missing alleles were performed using the Beagle 3.3.2 (Browning and Browning 2009). The binning criterion c ranged from 0.1 to 1.0 incremented by 0.1, as done in the simulated data analysis.

In this real data analysis, bone weight was the trait of interest (response variables). Since the LASSO method cannot adjust for covariates, phenotypic values were adjusted prior to the GWAS, i.e., the response variables were centered prior to the analysis. The covariates subject to adjustment included the first three principal components

(population structure) and the following fixed effects: (1) year effect, (2) fattening days effect, and (3) the effect of initial body weight (body weight when cattle entered the camp). The data centering process was conducted by fitting the original phenotypic values to the above three fixed effects and PCs using a general linear model. The residuals of the general linear model were the centered phenotypic values of the traits, which were eventually subject to the proposed GWAS analyses. All four methods listed in Table 1 were used for the real beef data.

Results

Analysis of simulated data

To attain an overall understanding of the bin model, we constructed bins for the 608,696 SNPs of the 1217 individuals using criterion c from 0.1 to 1.0 incremented by 0.1. Table 3 lists the total number of bins and the size per bin under different c values. Supplementary Tables S1–S9 show the number of bins per chromosome under different c values. There is a trend that the longer the chromosome, the more the number of bins. The numbers of bins also indicate the LD levels for chromosomes with similar length. For example, chromosomes 26 and 29 are similar in size, but when $c < 0.8$, the number of bins for chromosome 26 is smaller than the number of bins for chromosome 29. This phenomenon indicates that chromosome 29 has more markers within high LD blocks than chromosome 26.

Under each c criterion and using 0.05 as the p value criterion of significance test, we used the first 20 simulated samples to perform GWAS with BIN-Ave-Lasso and BIN-One-Lasso to evaluate the relationship between c and the power (Fig. 1a) and the relationship between c and the Type 1 error (Fig. 1b). Figure 1a shows that both methods have increased powers when c is increased, which is expected because binning neighboring SNPs will reduce the variance of a binned genotype compared with an individual SNP. In most cases, BIN-One-Lasso had a higher power than the BIN-Ave-Lasso except when $c = 0.1$ and $c = 0.3$ where the latter showed a higher power. Figure 1b shows that the two methods have decreased Type 1 errors when c is increased. Apparently, BIN-Ave-Lasso is superior over BIN-One-Lasso due to its better control of the Type 1 error. In subsequent analyses, we chose $c = 0.9$ as the binning criterion to compare the performances of the methods.

We now compare the powers and Type 1 errors of the new methods under $c = 0.9$ with existing methods from 100 replicated simulated samples. The results are given in Fig. S1, which show that BIN-One-Lasso has a much higher power than all existing methods and BIN-Ave-Lasso is much more efficient in controlling the Type 1 error than all other

methods. We used the power to Type 1 error ratio to evaluate the overall performance of these four methods. As shown in Fig. 2, BIN-Ave-Lasso outperformed the other methods and the performance of BIN-One-Lasso was roughly equal to the SNP-Lasso method. The Q + K-LMM method, the most commonly used GWAS approach, was the worst one in terms of power to Type 1 error ratio. The computational speeds (AMD Opteron 6344 CPU 2.59 GHz, Memory 272 G) of the four methods are compared in Fig. 3, showing that our new methods are most efficient in terms of computational time, only taking up to one-tenth of the computing time taken by the Q + K-LMM method.

Table 3 Bin information using different criteria

Threshold (<i>c</i>)	No. of bins	Average bin size (kb)	Standard deviation	<i>m/n</i> ratio ^a
1.0	608,696	—	—	500.16
0.9	360,084	2.02	6.33	295.88
0.8	293,252	3.44	9.59	240.96
0.7	243,510	5.13	13.41	200.09
0.6	206,034	7.07	18.20	169.30
0.5	169,597	9.85	25.49	139.36
0.4	122,541	15.71	36.65	100.69
0.3	75,589	28.49	56.77	62.11
0.2	39,566	58.83	100.71	32.51
0.1	13,489	181.55	251.13	11.08

^a*m* is the number of bins and *n* is the sample size

Since different methods used different threshold values for the test statistics to declare significance, we compared the ROC curves of the methods. An ROC curve is a plot of power against Type 1 error, the higher the curve, the better the method. We also compared the methods for a different type of ROC curve, where the power is plotted against the false discovery rate (FDR). The FDR is defined as the number of falsely detected markers or bins divided by the total number of detected markers or bins. The ROC curve plots are shown in Fig. 4, indicating that the new methods are always more powerful than the current methods compared. There is a slight difference between the two new methods: when the false positive rate is low, BIN-Ave-Lasso is more powerful than BIN-One-Lasso, and vice versa when the false positive rate is high. We also recorded the statistical powers for each QTN of the four different methods and ranked the 20 QTNs by their absolute values of effects in Table S10. The power of each QTN was defined as the fraction of the samples in which the QTN was detected. These QTN specific power comparisons are demonstrated in Fig. 5, showing that Q + K-LMM was too conservative to detect any QTN with absolute effect < 0.23. When QTNs with absolute effects > 0.35, all methods had 100% power except SNP-Lasso. In general, the power of BIN-One-Lasso is higher than other methods in detecting either large-effect QTNs or small-effect QTNs. Furthermore, because of the stringent significant threshold, Q + K-LMM has very low powers compared to other methods.

In summary, our new methods have similar performance compared to SNP-Lasso, but much model efficient to the

Fig. 1 Statistical powers and Type 1 errors under various *c* criteria of binning. **a** Statistical powers averaged across 20 replicated samples. **b** Type 1 errors averaged across 20 replicated samples

Fig. 2 Comparison of performance for BIN-Ave-Lasso, BIN-One-Lasso, Q + K-LMM, and SNP-Lasso averaged across 100 replicated samples

Fig. 3 Computing time (min) of BIN-Ave-Lasso, BIN-One-Lasso, Q + K-LMM, and SNP-Lasso across 100 replicated samples

Fig. 4 Statistical power plotted against Type 1 error and false discovery rate. **a** Statistical powers at different levels of Type 1 error. **b** Statistical powers at different levels of false discovery rate

Q + K-LMM method. In addition, the computational efficiency of the new methods is much higher than SNP-Lasso. Under the conventional significant threshold (p value = 0.05), BIN-One-Lasso is more powerful than BIN-Ave-Lasso while BIN-Ave-Lasso has a better control of Type 1 error.

Analysis of beef cattle data

We also used the four methods to analyze the real beef data for a trait called bone weight, which is defined as the total weight of all bones including the forequarter and

hindquarter joints. Significant SNPs (bins) for all methods are listed in Table 4 for the two new methods and Table 5 for the two current methods. Manhattan plots for the new methods (BIN-Ave-Lasso and BIN-One-Lasso) are also shown in Fig. S2 along with the plots from Q + K-LMM and SNP-Lasso. In the analysis of bone weight, we found that BIN-Ave-Lasso is relatively conservative and BIN-One-Lasso is more powerful than all other methods, consistent to the observation from the simulation studies. Moreover, regions identified by BIN-One-Lasso has covered all SNPs identified by Q + K-LMM and SNP-Lasso.

Fig. 5 Statistical powers of all simulated QTNs for BIN-Ave-Lasso, BIN-One-Lasso, SNP-Lasso, and Q + K-LMM

Table 4 Description of significant SNPs (bins) associated with bone weight identified by BIN-Ave-Lasso and BIN-One-Lasso

Method	Numbers of SNPs ^a	Chr	Position ^b (bp)	Nearest gene	Distance to the gene (bp)	p value
BIN-Ave-Lasso	1	14	62,769,117	RIMS2	Within	8.58E-04
	6	6	39,988,294-39,997,880	LCORL	996,182	9.37E-03
BIN-One-Lasso	1	14	62,769,117	RIMS2	Within	3.52E-53
	6	6	39,988,294-39,997,880	LCORL	996,182	6.86E-09
	1	21	5,309,649	SELENOS	75,449	1.81E-08
	2	15	9,229,362-9,236,537	Novel gene	306,521	9.61E-05
	2	6	38,466,628-38,468,901	LAP3	105,689	1.43E-04
	1	25	21,737,543	CHP2	76,232	8.23E-04
	1	10	38,729,744	EPB42	190,928	9.92E-04
	1	1	42,230,085	CLDND1	221,660	2.56E-03
	1	12	87,865,066	MYO16	53,674	4.62E-03
	1	6	39,932,557	LCORL	940,445	8.98E-03
	1	6	41,587,207	bta-mir-218-1	42,069	2.82E-02
	1	1	63,037,677	Novel gene	644,796	2.94E-02
	3	6	41,186,810-41,194,565	SLIT2	42,024	3.04E-02
	1	4	62,537,378	NPSR1	Within	3.05E-02

^aNumber of SNPs within the significant bin

^bWhen more than one SNP within this bin, the position of this bin is defined as the region from the position of the first SNP to the last SNP

Discussion

The next-generation sequencing technologies have produced large volumes of genomic data. Such big data face two major bottlenecks in GWAS: high dimensionality and

heavy computational burden. Combining SNPs into bins is an efficient way to solve both problems simultaneously. The SNP block type of methods can effectively reduce the number of markers for analysis, e.g., only about 20,000 genes to fit a model when using a gene-based method to

Table 5 Description of significant SNPs (bins) associated with bone weight identified by Q + K-LMM and SNP-Lasso

Method	SNP name	Chr	Position (bp)	Nearest gene	Distance to the gene (bp)	<i>p</i> value
Q + K-LMM	BovineHD1400017455	14	62,769,117	RIMS2	Within	6.87E-17
	BovineHD0600010952	6	39,990,876	LCORL	998,764	7.42E-10
	BovineHD0600010956	6	39,997,880	LCORL	1,005,768	9.88E-10
	BovineHD0600010953	6	39,991,818	LCORL	999,706	2.14E-09
	BovineHD0600010950	6	39,988,294	LCORL	996,182	5.16E-09
	BovineHD0600010951	6	39,989,868	LCORL	997,756	5.97E-09
	BovineHD0600010955	6	39,994,844	LCORL	1,002,732	7.48E-09
	BovineHD0600010646	6	38,468,901	LAP3	105,689	6.27E-08
SNP-Lasso	BovineHD1400017455	14	62,769,117	RIMS2	Within	1.57E-54
	BovineHD2100000949	21	5,308,649	SELENOS	75449	2.87E-07
	BovineHD0600010952	6	39,990,876	LCORL	998,764	1.20E-06
	BovineHD0600010646	6	38,468,901	LCORL	523,211	4.03E-04
	BovineHD1000011977	10	38,729,744	EPB42	190,928	1.21E-03
	BovineHD1500002403	15	9,236,537	Novel gene	306,521	1.22E-03
	BovineHD4100004740	6	41,194,565	SLIT2	42,024	4.64E-03
	BovineHD1200025651	12	87,865,066	MYO16	53,674	1.54E-02
	BovineHD2500006031	25	21,737,543	CHP2	76,232	1.77E-02
	BovineHD0100012067	1	42,230,085	CLDND1	221,660	2.04E-02
	BovineHD0400017134	4	62,537,378	NPSR1	Within	3.80E-02

perform GWAS in cattle (Capomaccio et al. 2015). This is far less than the number of genome-wide markers (770K) in the Illumina Bovine HD BeadChip. Many previous studies also elucidated that SNP block-like methods are more powerful than the single-marker analysis counterpart.

In this study, the whole genome is divided into a manageable number of bins, and we can easily control the number of bins through controlling the *c* threshold of variance across individuals in the population. Our bin model is based on the LD level between consecutive markers within bins and thus takes advantage of the biological information (physical relationships of SNPs). This is quite different from dimension reduction statistics such as the principal component analysis. Our results showed that the bin model can effectively compress markers. We only compare the performance of the bin model with *c* = 0.9 to other methods based on the prior knowledge from Fig. 1. When we set *c* = 0.5, our bin model also has comparable performance to Q + K-LMM. However, if the *c* threshold is too small, the bin sizes will be relatively large and the LD levels within bins will become relatively low. In that case, the bin model cannot capture QTNs with small effects and the effects will be diluted by so many markers within bins. The bin model seems to contradict with the purpose of improving power in GWAS. With the bin model, we actually emphasize the ability of handling almost unlimited number of markers while still using a multiple marker model. The gain comes from the multiple marker model, not from the bins themselves. Without the bin model, we cannot use the multiple

marker approach to performing GWAS when the number of markers is extremely large.

In practice, the LD level has a positive correlation with the marker density. For this reason, the efficiency of dimensionality reduction will increase when the density of markers is increased. Therefore, the proposed bin model will be more advantageous with high density sequence data than with low density marker data. However, the proposed bin model assumes that there is at most one true QTN in a bin. This means that the bin size cannot be too large. If the bin size is indeed too large, we need another assumption to compensate for it, that is all QTNs in the same bin must have effects in the same direction. An obvious consequence when the above assumptions are violated is the compromise of statistical power. Adding a weight to each marker is a feasible solution to this pitfall. In rare variant analysis, Han and Pan (2010) proposed to estimate the direction of effect for each marker in a marginal model and then assigned a weight to represent the direction of that marker effect. Lin and Tang (2011) used the estimated regression coefficient of each marker as the weight, where the regression was done one marker at a time. Moreover, the principal component analysis can also determine the weight. When the bin sizes are not too large, the chance of having more than one causal mutation within a single bin is minimum under the assumption that only a finite number of genes are causative for most quantitative traits.

LASSO is a penalized least-squares method with constraints on the absolute values of the regression

coefficients. Due to this special penalty, LASSO leads to sparse selection of independent variables by shrinking most of the regression coefficients to zero. As a consequence, it is the ideal model to fit most quantitative traits, which are assumed to be controlled by a few genes with large effects and many genes with undetectably small effects (Hayes and Goddard 2001). Xu (2007) compared LASSO with several other methods (empirical Bayes, penalized likelihood (Zhang and Xu 2005), and stochastic search variable selection (Yi et al. 2003) for multiple QTL mapping (Xu 2007) and found that LASSO was as efficient as the empirical Bayes method but much more efficient in terms of computational speed. However, LASSO has an intrinsic limit in the ratio of model size to sample size—the sample size must be sufficiently large to match the growth of model size. Based on the above considerations, we proposed the new bin model to improve GWAS performance by taking advantage of the high efficiency of LASSO. We used the GlmNet/R package (Friedman et al. 2010) to perform the LASSO analysis and the program is extremely fast for a median sized data. In GlmNet/R, the optimal value of tuning parameter λ , which controls the degree of shrinkage, can be obtained through a k-fold cross-validation. The λ value from the cross-validation is used to calculate the MSE. In this study, the λ value producing the minimum MSE + 1 standard error is deemed the optimal tuning parameter in order to control the Type 1 error rate (Waldmann et al. 2013). The relaxed λ value producing the minimum MSE also can be used to obtain more significant markers. The procedure of cross-validation may result in unstable estimation of the tuning parameter partly because of refitting the model many times (Li et al. 2010). An alternative method is the Bayesian LASSO (Yi and Xu 2008), which treats the tuning parameter as an unknown parameter to be estimated from the data. This method provides an automatic way to determine the tuning parameter. However, the method involves time consuming MCMC samplings and cannot generate a sparse model for variable selection (Hans 2010).

The bin model cannot pinpoint causative variants when bin sizes are relative large. However, this issue can be solved by the two-stage approach proposed here, where we detected significant bins first and then fit a model included all markers within the significant bins. Many two-stage methods have been proposed for GWAS including, e.g., LMML (Shi et al. 2011), mrMLM (Wang et al. 2016), and ISIS EM-BLASSO (Tamba et al. 2017), which appear to be advantageous over a single stage genome scanning approach. In the first stage, these methods select a promising subset of markers through a correlation test between markers and traits of interest. The aim of this step is to reduce the number of markers. After

this stage, the number of preliminarily associated markers is relatively small and can be easily handled by various penalized regression models in the second stage. The second stage model fitting can further reduce false positives, while a relaxed threshold in the first step can keep many potentially interested SNPs in the model. In our two-stage bin model, we used a second step of fine tuning to pick up specific markers within the associated bins. The model in the second stage LASSO analysis contains all markers in the significant bins. For example, if we identify 20 significant bins in the first stage and each significant bin contains 100 markers, then the total number of markers included in the second stage LASSO model would be just $20 \times 100 = 2000$, a small number easily handled by LASSO.

Conclusions

Multi-locus GWAS methods are more suitable for identifying causative variants than LMM that separates the connection between markers. However, the high dimensional problem has become a bottleneck for genome-wide multi-locus GWAS. In this study, we proposed a bin model to compress the scale of genome data by changing the c criterion of binning. The BIN-Ave-Lasso and BIN-One-Lasso methods can be used for genome-wide multi-locus GWAS. Our new approaches are faster and more powerful than Q + K-LMM and SNP-Lasso as demonstrated in the simulated studies. Moreover, the BIN-Ave-Lasso method has a much lower FPR while BIN-One-Lasso shows higher statistical power under the conventional threshold.

Acknowledgements This work was supported by the National Natural Science Foundations of China (31872975, 31472079, and 31702084) and Cattle Breeding Innovative Research Team (cxgc-ias-03) Grants to HG and JL.

Data availability Datasets are available from the Dryad Digital Repository <https://datadryad.org/stash/dataset/doi:10.5061/dryad.4qc06>

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* 84(2):210–223
- Capomaccio S, Milanese M, Bomba L, Cappelli K, Nicolazzi EL, Williams JL et al. (2015) Searching new signals for production

- traits through gene-based association analysis in three Italian cattle breeds. *Anim Genet* 46(4):361–370
- Friedman J, Hastie T, Tibshirani R (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* 33(1):1–22
- Han F, Pan W (2010) A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Heredity* 70(1):42–54
- Hans C (2010) Model uncertainty and variable selection in Bayesian lasso regression. *Stat Comput* 20(2):221–229
- Hayes B, Goddard ME (2001) The distribution of the effects of genes affecting quantitative traits in livestock. *Genet, selection, evolution: GSE* 33(3):209–229
- Hoerl AE, Kennard RW (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 12(1):55–67
- Hu Z, Wang Z, Xu S (2012) An infinitesimal model for quantitative trait genomic value prediction. *PLOS ONE* 7(7):e41336
- Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-Y, Freimer NB et al. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* 42:348
- Kao C-H, Zeng Z-B, Teasdale RD (1999) Multiple Interval Mapping for Quantitative Trait Loci. *Genetics* 152(3):1203
- Kyung M, Gill J, Ghosh M, Casella G (2010) Penalized regression, standard errors, and Bayesian lassos. *Bayesian Anal* 5(2):369–411
- Li J, Das K, Fu G, Li R, Wu R (2010) The Bayesian lasso for genome-wide association studies. *Bioinformatics* 27(4):516–523
- Lin DY, Tang ZZ (2011) A general framework for detecting disease associations with rare variants in sequencing studies. *Am J Hum Genet* 89(3):354–367
- Liu N, Zhang K, Zhao H (2008) Haplotype-association analysis. *Adv Genet* 60:335–405
- de Los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MP (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193(2):327–345
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3):559–575
- Segura V, Vilhjalmsdottir BJ, Platt A, Korte A, Seren U, Long Q et al. (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat Genet* 44(7):825–830
- Shen X, Alam M, Fikse F, Ronnegard L (2013) A novel generalized ridge regression method for quantitative genetics. *Genetics* 193(4):1255–1268
- Shi G, Boerwinkle E, Morrison AC, Gu CC, Chakravarti A, Rao DC (2011) Mining gold dust under the genome wide significance level: a two-stage approach to analysis of GWAS. *Genet Epidemiol* 35(2):111–118
- Tamba CL, Ni YL, Zhang YM (2017) Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. *PLoS computational Biol* 13(1):e1005357
- Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc: Ser B (Methodol)* 58(1):267–288
- Waldmann P, Meszaros G, Gredler B, Fuerst C, Solkner J (2013) Evaluation of the lasso and the elastic net in genome-wide association studies. *Front Genet* 4:270
- Wang H, Zhang YM, Li X, Masinde GL, Mohan S, Baylink DJ et al. (2005) Bayesian shrinkage estimation of quantitative trait loci parameters. *Genetics* 170(1):465–480
- Wang SB, Feng JY, Ren WL, Huang B, Zhou L, Wen YJ et al. (2016) Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci Rep* 6:19444
- Wen YJ, Zhang H, Ni YL, Huang B, Zhang J, Feng JY et al. (2018) Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief Bioinforma* 19(4):700–712
- Xia J, Fan H, Chang T, Xu L, Zhang W, Song Y et al. (2017) Searching for new loci and candidate genes for economically important traits through gene-based association analysis of Simmental cattle. *Sci Rep* 7:42048
- Xu S (2003) Estimating polygenic effects using markers of the entire genome. *Genetics* 163(2):789
- Xu S (2007) An empirical Bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics* 63(2):513–521
- Xu S (2013) Genetic mapping and genomic selection using recombination breakpoint data. *Genetics* 195(3):1103–1115
- Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL (2014) Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet* 46(2):100–106
- Yi N (2004) A unified Markov chain Monte Carlo framework for mapping multiple quantitative trait loci. *Genetics* 167(2):967–975
- Yi N, George V, Allison DB (2003) Stochastic search variable selection for identifying multiple quantitative trait loci. *Genetics* 164(3):1129–1138
- Yi N, Xu S (2008) Bayesian LASSO for quantitative trait loci mapping. *Genetics* 179(2):1045–1055
- Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 38(2):203–208
- Zhang YM, Xu S (2005) A penalized maximum likelihood method for estimating epistatic effects of QTL. *Heredity* 95(1):96–104
- Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J R Stat Soc: Ser B (Stat Methodol)* 67(2):301–320