Send Orders for Reprints to reprints@benthamscience.net

The Open Bioinformatics Journal, 2013, *7*, (Suppl-1, M2) 27-33 27

Open Access

Genetic Studies: The Linear Mixed Models in Genome-wide Association Studies

Gengxin Li[1,*] and Hongjiang Zhu[2]

[1]*Department of Mathematics and Statistics, Wright State University, 201 MM, 3640 Colonel Glenn Highway, Dayton, OH 45435-0001*

[2]*Division of Biostatistics, Coordinating Center for Clinical Trials, The University of Texas School of Public Health at Houston*

Abstract: With the availability of high-density genomic data containing millions of single nucleotide polymorphisms and tens or hundreds of thousands of individuals, genetic association study is likely to identify the variants contributing to complex traits in a genome-wide scale. However, genome-wide association studies are confounded by some spurious associations due to not properly interpreting sample structure (containing population structure, family structure and cryptic relatedness). The absence of complete genealogy of population in the genome-wide association studies model greatly motivates the development of new methods to correct the inflation of false positive. In this process, linear mixed model based approaches with the advantage of capturing multilevel relatedness have gained large ground. We summarize current literatures dealing with sample structure, and our review focuses on the following four areas: (i) The approaches handling population structure in genome-wide association studies; (ii) The linear mixed model based approaches in genome-wide association studies; (iii) The performance of linear mixed model based approaches in genome-wide association studies and (iv) The unsolved issues and future work of linear mixed model based approaches.

Keywords: Genetic similarity matrix, genome-wide association study (GWAS), linear mixed model (LMM), population stratification, sample structure, single nucleotide polymorphisms (SNPs).

1. INTRODUCTION

The recent breakthrough in genotyping technology induces the high density genome-wide collection, allowing researchers to access to an extraordinarily large number of single nucleotide polymorphisms (SNPs), even those newly identified markers in a fast and cost efficient way. The genome-wide association studies (GWAS), which traditionally tested the disease-causing genetic variants within some particular genes and regions, can be applied on a genomic scale. The success of the International HapMap Project [1] in cataloguing the genetic variation has brought the identification of over millions of SNPs for the large scale genome-wide association studies [2]. These studies have discovered numerous genetic variants contributing to major human diseases successfully [1, 3-5]. It is well known that GWAS may be confronted by the inflated false positive rates if the population structure, which is derived from individuals from different populations within one study, is not properly corrected in the model [6, 7]. The presence of related individual within either a case-control cohort or a population cohort is principally termed as the sample structure that includes family structure and cryptic relatedness as well as population structure [8]. In particular, the unknown genetic relationship of any two individuals results in the family structure and cryptic relatedness. Specifying these three structures simultaneously in GWAS is more challenging, and a number of new methods have been developed due to this limitation.

The once dominant methods controlling the inflation rate in GWAS are Genomic Control (GC), which measures and adjusts the inflation of the test statistics due to the population structure, Structured Association (SA), Principal Components Analysis (PCA) and Multidimensional Scaling (MDS) which describe the population structure in the GWAS model to correct the population stratification properly. These four methods are phenomenal at the control of population stratification, but fail to account for the complete genealogy of all the individuals. In particular, justifying family structure in GWAS with family-based data is essential, and the cryptic relatedness has been observed to occur frequently in the wide range of GWAS data. Somewhat differently, family-based association tests exploit both within and between family structures to improve the statistical power in GWAS. More recently, new approaches have been prevalently developed based on the linear mixed models (LMM), and the principle of this strategy makes the interpretation of sample structure possible in GWAS. It estimates the genetic similarity between a pair of individuals to account for the genealogy of population.

In this paper, we briefly review each of the above approaches to account for the sample structure in genome-

*Address correspondence to this author at the Department of Mathematics and Statistics, Wright State University, 259 MM, 3640 Colonel Glenn Highway, Dayton, OH 45435-0001; Tel: 937-775-4211:
E-mail: gengxin.li@wright.edu

1875-0362/13

wide association studies. These methods include once dominant approaches interpreting partial sample structure and new approaches using linear mixed models to capture the genealogy of population in GWAS. More importantly, primary challenges in spurious association tests due to partially or not correctly interpreting the familial relatedness of samples are discussed. For new approaches using the linear mixed model, the performances in applicability, computational speed and significance of calls are evaluated. Besides, a number of methods for estimating genetic similarity matrix are explored. A brief guideline about the efficacy of each LMM-based method is provided. Finally, the unsolved questions and future works of new LMM-based approaches are discussed.

2. THE APPROACHES HANDLING POPULATION STRUCTURE IN GWAS

Genomic Control has widely been applied in GWAS to adjust the extendibility of confounding risk due to population stratification. It defines the confounding factor, λ, which is calculated by the ratio of the observed median of test statistics to its theoretical median, to measure the inflation rate [9-11]. Ideally, there is no population stratification when λ is equal to 1. Once λ is above 1, some confounders (stratification, or family structure, or cryptic relatedness) may occur in GWAS. The practical experiments reveal $\lambda <$ 1.03-1.05 to be the sound inflation rate that interprets the individuals' relatedness sufficiently. Regarding the subpopulation differences (stratification) due to recent genetic divergence, the confounding factor can correct the stratification adequately. However, the genetic divergence starting from ancestral population [12] may result in the unexpected current SNPs and markers with unusual allele frequency, which makes the stratification more severe, and then the uniform confounding factor is not enough to adjust this inflation. Moreover, as for other confounders, such as family or cryptic relatedness, GC has limited application.

Structured Association and Principal Component Analysis are other prevalent approaches, which account for genetic ancestry explicitly in GWAS model, to correct the inflation due to stratification. Before computing association statistics, SA applies a clustering program [13, 14] to assign all individuals to different population groups to estimate the effect of stratification. In fact, incorporating the fractional cluster information brings large pressure on computation. A faster cluster program (ADMIXTURE [15]) has recently been proposed in SA which makes the genomic scale association test with inferring genetic ancestry practically. Compared with SA, PCA is implemented simply in GWAS [16-19], where PCA selects top components to capture the broad relatedness across individuals, and these principal components are fitted as covariates correcting the inflation due to stratification [20, 21]. In particular, EIGENSTRAT [20] is an improved PCA method to explicitly explore ancestry differences and treatment difference in laboratory. Additionally, both approaches provide a great correction to markers having large allele frequency differences across ancestral population. However, these two methods assume that there are only a small number of ancestral populations and admixture, and this assumption indicates that multiple levels of sample relatedness are partially captured [22-24]. Eventually, both approaches are limited to control family structure or cryptic relatedness in GWAS data. More recently, people tempt to combine these strategies, which removes closely related samples using SA, corrects the broad sample structure using PCA and adjusts the residual inflation using GC [1, 25, 26].

In real experiment, Principal Component Analysis is very sensitive to outliers that may result in the bias of detection and reduce power. An alternative approach, Multid-imensional Scaling, is applied to visualize substructure and to explicitly explore the ancestry of samples [27, 28]. The classical MDS has the data reduction technique *via* measuring substructure by a k-dimensional representation. When MDS is built on an Euclidean distance metric, it becomes identical to PCA and can be widely applied in other methods [20]. Spectral-GEM [28] has been proved to outperform PCA in separating the effect of outliers from population stratification. This method connects MDS and spectral graph theory to efficiently capture the stratification. But, MDS is limited to account for the multilevel relatedness of samples, especially when family structure and cryptic relatedness cannot be ignored in the data.

3. THE LINEAR MIXED MODEL BASED APPRO-ACHES IN GWAS

3.1. The Linear Mixed Model

The linear mixed model has widely been applied in human linkage analysis [29, 30]. Variance components model partitions genetic effects as additive and polygene effects whereby each one is treated as random. Each marker is tested to see whether the variance of a genetic effect at this locus is significantly deviated from zero. This variance component model can be expressed as:

$$y = \mu + a + g + e \quad (1)$$

Here, μ as the fixed effect denotes the overall mean; *a* measures the additive genetic effect; g denotes the polygene effect, and e denotes the random residual effect. In case of some more complex data, the fixed effects may include covariates in addition to overall mean (μ), such as: gender and age, then the fixed effect could be replaced by Xβ where β is a coefficient vector of fixed effects and X is an incidence matrix. Compared to simple regression models [31-34], variance component model shows notable merits in mapping significant loci related to phenotypic traits. The superiority of this strategy benefits from the mapping principle that sibpairs with similar phenotypes should have higher expected sharing of genetic material near genes. Thus sample structure resulting from the sharing genealogy could be captured by variance component models which could explicitly account for genetic relationships between two individuals. Even when the relatedness information is unknown, people could empirically estimate sample structure across individuals using high density marker genotypes. However, variance components models cannot be applied to the genomic data with millions of SNPs and thousands of individuals owing to the heavy burden on estimating random parameters. Motivated by this limitation of variance components model, some efficient LMM-based approaches in GWAS have been proposed (Yu's model [23], compressed MLM with P3D [35], EMMAX [36], FaST-

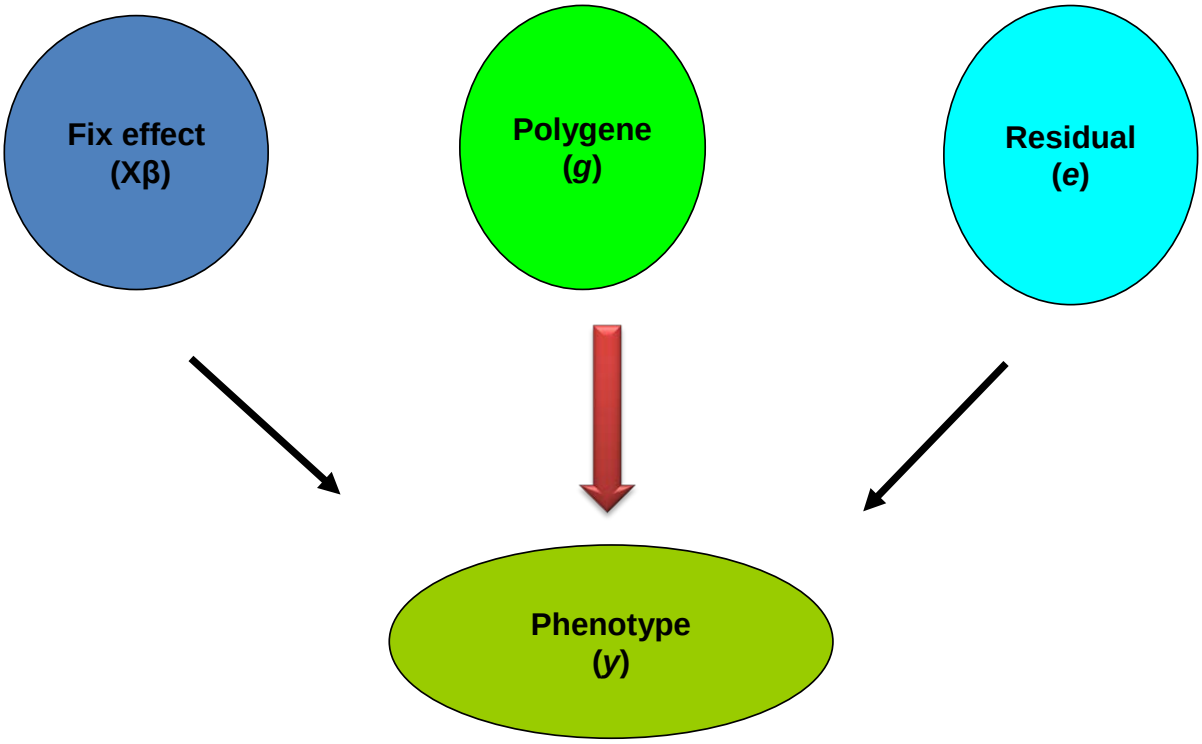

Fig. (1). The model setting of the linear mixed model based approaches in GWAs.

LMM [37] and GRAMMAR-Gamma [38]) recently. The model of linear mixed model based methods is set as,

$$y = X\beta + g + e \quad (2)$$

Here, X is the matrix of fixed effects including overall mean, covariates and the testing SNP; the vector β denotes the coefficients of fixed effects; g is a random effect reflecting polygene background, and its variance is dependent on the kinship matrix (Var(g) = $K\sigma_g^2$ where K is the kinship matrix that measures the genetic similarity across individuals). Indeed, the structure of this kinship matrix reflects population structure, family structure and cryptic relatedness. In particular, Fig. (**1**) displays the model setting of linear mixed model based approaches in GWAS. Consequently, this new linear mixed model is widely applied to GWAS to correct the inflation of false positive.

3.2. The Development of LMM-Based Approaches

The classical linear mixed models capturing family structure and cryptic relatedness in addition to stratification were initially developed from the animal models [23, 24, 39]. A unified mixed-model method [23] in GWAS has been proposed recently based on this strategy. Because of complex characteristics of GWAS data, this new method builds Q+K model, K model and Q model to adjust the inflation due to multilevel relatedness including population stratification and familial relatedness. In this process, Q as the fixed effect measures population structure that is calculated by package Structure [11], K as the random effect detects the relative kinship matrix that empirically measures the genetical similarity across individuals from markers or SNPs and computed by SPAGeDi [40]. It is known that K is superior to the co-ancestry matrix (G) in absence of pedigree information or under biases from genetic drift [41, 42]. Additionally, this unified linear mixed-model is flexible to be adopted in population data including both stratification and familial relatedness and family based data containing family structure alone [23]. Compared with previous strategies (GC, QTDT and simple regression model), this unified linear mixed-model method is good at estimating genetic effect and the control of errors simultaneously. This method has been implemented in Package TASSEL [43].

In the case of more complex traits, the small effect of genetic variants is important to be detected in GWAS, so a large GWAS data with millions of SNPs and tens of thousands individuals is available. But the previous Yu's model is computationally intractable when estimating random parameters for each of high density SNPs. Consequently, the compressed mixed linear model (compressed MLM) and population parameters previously determined (P3D) approach [35] are developed to overcome this limitation, and these two methods can be used separately or jointly. Specifically, the compressed MLM extends the animal breeding sir model [44-47] to reduce the size of random effects from the individual level (n^3) to the group level (u^3) where ($u \leq n$), and the compression level, the average number of individuals within one group, is optimized by clustering individuals into groups. P3D is a two-step approach where the first step focuses on estimating population parameters (controlling sample structure) once using compressed MLM without testing SNP effect, and the second step continues to estimate the testing SNP effect and the random genetic effect with priors of population parameters determined in the first step. When jointly using these two methods, the compressed MLM can be applied in

the first step of P3D to further reduce the computing time. Additionally, the gradual improvement in statistical power is achieved on optimizing the compression level in genome-wide association tests. The compressed MLM with P3D has also been built in package TASSEL [43].

Because of the heavy computational burden from the previous Yu's model [23] in GWAS, another efficient mixed-model association eXpedited (EMMAX [36]) which is an extension of previous linear mixed model EMMA [48], has been proposed. It is also a two-step approach where the first step estimates population parameters measuring sample structure and tests the significance of these parameters to the phenotypic variance once, and the second step uses an F-test (generalized least square (GLS)) [49] or a score test [50] for each SNP with population parameters as dependent variable. The practicability of this approach is based on one assumption that the effect of each marker on the trait is small for a large GWAS data, and then it is not necessary to estimate random variances for each marker in the second step which greatly reduces the computing time from years to hours. In the comparison of P value distribution in GWAS, this approach performs better than GC and PCA in both the population cohort [25, 51] and case-control cohort [1].

More recently, the cost of high density SNPs for each subject is gradually acceptable, making the super data with hundreds of thousands of individuals available for genome-wide association study. But, the cohort size of this super GWAS data is beyond the upper bound of the sample size allowed by EMMAX and compressed MLM with P3D [37]. Besides, the applicability of these two methods is based on the assumption that random variances are same across SNPs [35, 36]. A more flexible method, factored spectrally transformed linear mixed model (FaST-LMM [37]), without needing this assumption is developed for the super GWAS data. The dramatic improvement in computing time and memory is achieved through the dimension reduction on both SNPs and kinship matrix. FaST-LMM is implemented in two steps. The first step is to estimate the realized relationship matrix (RRM) [52-54], which measures the genetic similarity between a pair of individuals, using partial SNPs uniformly sampled from the whole SNPs pool. It is shown that different SNP sets have almost the identical effect on association tests (Fig. (**2**) in paper [37]). In the second step, maximizing the complex log likelihood function for each SNP is mainly reduced to optimize one dimensional parameter δ ($\delta = \frac{\sigma_g^2}{\sigma_e^2}$) as well as the fixed effect (β) and σ_g^2 after transforming phenotype, SNP and covariates using the spectral decomposition of RRM. At that time, FaST-LMM is named as FaST – LMM_{full}. In particular, k-spectral decomposition of RRM is applied, and then the procedure goes with further reducing the size of kinship matrix to optimize the log likelihood function. Now FaST-LMM is defined as FaST – LMM_{low}. If δ under null is fixed, running time is greatly improved.

All previous LMM-based methods (Yu's model, compressed MLM with P3D, EMMAX and FaST-LMM) are likelihood ratio test (LRT) based approaches that are computationally demanding for large GWAS data. The pressure on computation leads to the development of the score test based variance components methods (Fast association score test based analysis (FASTA [50]) and genome-wide rapid association mixed model and regression with GRAMMAR Gamma factor (GRAMMAR-Gamma [38])). FASTA is a two-stage approach where the population parameters and genetic similarity matrix are estimated once in the first step, and the score test with the previously computed population parameters and kinship matrix are applied to each marker to detect its effect, moreover, the likelihood ratio test is applied again to few candidate markers from previous score test to achieve more accurate significance in the second step. GRAMMAR-Gamma is a more advanced approach built on FASTA and GRAMMAR [55], jointly borrowing these two methods' merits. In fact, it is not a typical LMM-based approach, and two steps are also involved in GWAS analysis. In the first step, the null model of GRAMMAR-Gamma is similar as those of previous LRT-based methods under null. The population parameters measuring sample relatedness and GRAMMAR-Gamma factor that is a function of kinship matrix to correct the inflation are estimated, and transformed phenotypic traits by the kinship matrix are achieved. In the second step, a new score test adjusted by the previous GRAMMAR-Gamma factor is applied to each marker, and its computation complexity is close to the theoretical minimum [38] compared with LRT based approaches.

4. THE PERFORMANCE OF LMM-BASED APPROACHES IN GWAS

The notable advantage of LMM-based methods in GWAS is that this strategy empirically estimates the genetic similarity across individuals from SNP genotype information to efficiently control the inflation of false positive. In fact, the genetic similarity matrix has the capability of accounting for a wide range of sample relatedness which is sure to win out over the non-mixed model based methods (GC, SA, PCA and MDS) only capturing partial genealogy of population [23, 35-38]. Another superiority of LMM-based approaches is that this strategy can be flexibly applied to population-based GWAS data and family-based GWAS data as well as both population cohort and case-control cohort GWAS.

Additionally, we summarize and compare the performance of primary LMM-based approaches in testing, applicability, dimension reduction and time complexity in GWAS study (Table **1**). It is clearly seen that Yu's model [23], which is one step procedure, is applicable to a small data set because it needs to estimate population and non-population parameters for each marker, and is computationally demanding. Other mixed model based approaches (compressed MLM with P3D, EMMAX, FaST-LMM and GRAMMAR-Gamma) are two-step procedures which estimate random variances accounting for familial relatedness once in the first step, and then optimize the log-likelihood function to test genetic effect of each marker incorporating random variances as dependent variables. Besides, the compressed MLM with P3D further reduces the running time by clustering individuals into few groups to estimate population parameters in the first step. FaST-LMM randomly samples a subset of SNPs to estimate sample structure and lowers the rank of genetic similarity matrix to additionally reduce computing time. Overall, the

Table 1. Comparisons of each LMM-based Approaches

Method	Testing	Applicability	Reduction	Time Complexity	
				Step I	Step II
TASSEL	LRT	Pop & CC		$O(pn^3c)$	NA
TASSEL+P3D	LRT	Pop & CC	Sample	$O(u^3 + uc)$	$O(pn^2)$
EMMAX	LRT	Pop		$O(n^3 + nc)$	$O(pn^2)$
$FaST - LMM_{full}$	LRT	Pop & CC		$O(n^3 + nc)$	$O(pn^2)$
$FaST - LMM_{low}$	LRT	Pop & CC	SNP & kinship matrix	$O(ns^3 + nc)$	$O(pnk)$
GRAMMAR-Gamma	ST	Pop		$O(n^3 + nc)$	$O(pn)$

(TASSEL: is Yu's model (yu's model); TASSEL+P3D: is compressed MLM with P3D (p3d); LRT: is the likelihood ratio test taken in GWAS; ST: is the score test in GWAS; Pop: population cohort; CC: case-control cohort; n: the number of individual within one study; u: the number of group; s: the number of randomly selected SNPs; c: the average number of iteration; p: the number of testing SNPs; k: the number of eigenvectors deviating from zero.)

effectiveness of the compressed MLM with P3D and EMMAX is similar and can be applied to a relative large GWAS with tens of thousands of individuals. In the case of the super GWAS data, FaST-LMM is more appropriate approach which greatly improves both time and space complexity in computation. Finally, GRAMMAR-Gamma is the score test based approach which deeply reduces the time complexity in the second step, but the significance of calls is gently weaker than that of likelihood ratio test based methods.

5. THE UNSOLVED ISSUES AND FUTURE WORK OF LMM-BASED APPROACHES

How to account for population structure in LMM-based approaches is still an open question. In most cases, population structure collecting with family structure and cryptic relatedness is modeled as a random effect in the linear mixed models. Using the random effect to capture the complete genealogy of population will greatly simplify the model structure. A large number of real data analysis have shown that population stratification could be sufficiently corrected by the random effect. Alternatively, population structure could be separated from the familial relatedness and fitted as a fixed effect, and this fixed effect is estimated by PCA. In fundamental respects, population structure measuring the differences of samples from different subpopulations has identical effect for all individuals. Once the allele frequency of some markers is abnormally beyond the range of allele frequency of ancestral populations, the marker based kinship matrix may fail to capture the stratification in GWAS. On the other hand, fitting population structure in the fixed effect is not an appropriate way to control the stratification when samples possess more complex relationship including family structure as well as population structure, a few number of spurious principal components are fitted in fixed effects [18]. To avoid the spurious components, the components from the unrelated samples [56] are calculated, but it may lead to a new problem that some biased principal components may be produced due to some noise in SNP selection from unrelated samples [12]. This motivates the further improvement of the marker based random effect and efforts on PCA calculated from related samples.

One advantage of the linear mixed model based methods is to use the marker based genetic similarity matrix to account for a wide range of relatedness among individuals. In different populations, this genetic sharing matrix plays an important role in quantitative inheritance studies, but accurately estimating kinship matrix is challenging. One primary estimation method for kinship matrix is to compute identity by descent (IBD) between two individuals by adjusting with the identity by state (IBS) between random individuals. Regarding interpreting as much information of sample structure as possible, this marker based kinship matrix is superior to the pedigree based co-ancestry matrix especially when pedigree information is unknown [23]. The IBS or Balding Nichols matrix [57] is an alternative option that more accurately captures the long distance relationship due to stratification compared with IBD matrix which is phenomenal to account for short distance relatedness. In the case of GWAS data with admixture population, SNP bias in estimating sample structure may occur due to the probe error. We could weight each SNP to estimate IBS genetic similarity matrix [48]. To further improve the computational efficacy, FaST-LMM uses the realized relationship matrix (RRM) [52-54] to lower the rank of genetic sharing matrix instead of IBS or IBD matrix. However, there is no unified estimation method for genetic similarity matrix.

Extending the above linear mixed model based approaches to next generation sequencing studies requires a lot of research. These mixed model based approaches are developed for typical GWAS where common SNPs with medium and high minor allele frequency are widespread across whole genome. Since a large proportion of genetic heritability [58] is unexplained, we have to investigate an amount of variants with low minor allele frequency (MAF, MAF < 5%). As for whole exome or genome sequencing studies, variants with low minor allele frequency account for a large proportion across the whole genome. The use of these variants to estimate the genetic similarity matrix may lose some information and the greater precision of estimation for complete genealogy of population is difficult to maintain. Finding a more efficient estimation method is urgently demanding.

6. DISCUSSIONS

Recently, the linear mixed model based approaches are newly developed models that efficiently account for the

complete genealogy of population (including population stratification, family structure and cryptic relatedness) in GWAS. Compared with once dominant methods (GC, SA, PCA and MDS), the linear mixed model based methods are shown to be a comprehensive approach to correct the inflation of false positive due to not completely interpreting complex sample structure. Additionally, LMM-based approaches perform favorably in both population and case-control cohort.

The applicability of the linear mixed models is highly related to the relatedness property of GWAS data. As for GWAS data containing only population stratification, the once dominant methods (GC, SA, PCA and MDS) are adequate. As for GWAS data containing strong familial relatedness and weak population stratification, the linear mixed model based methods are sufficient to capture sample structure via the genetic similarity matrix in the random effect. As for GWAS data having strong population stratification and familial relatedness, a linear mixed model based approach plus PCA separately accounting for stratification in the fixed effect is sufficient. As for GWAS data with normal sample size, EMMAX and compressed MLM with P3D are appropriate methods. As for GWAS data with extremely large sample size, FaST – LMM_{low} is the best method at present.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

This project was supported in part by startup fund from Wright State University and University of Texas School of Public Health at Houston.

REFERENCES

[1] WTCCC, "Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls," *Nature*, vol. 447, pp. 661-678, 2007.

[2] D. E. Reich, S. B. Gabriel, D. Altshuler, "Quality and completeness of snp databases," *Nat. Gene.*, vol. 33, pp. 457-458, 2003.

[3] R. Sachidanandam, "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms," *Nature.*, vol. 409, pp. 928-933, 2001.

[4] R. J. Klein, "Complement factor h polymorphism in age-related macular degeneration," *Nature.*, vol. 308, pp. 385-389, 2005.

[5] R. Sladek, G. Rocheleau, J. Rung, C. Dina, L. Shen, S. David, B. Philippe, V. Daniel, B. Alexandre, H. Samy, B. Beverley, H. Barbara, C. Guillaume, H. Thomas, M. Alexandre, P. Alexey V, P. Marc, P. Barry I, B. David J, M. David, P. Constantin, F. Philippe ,"A genomewide association study identifies novel risk loci for type 2 diabetes," *Nature*, vol. 445, pp. 881-885, 2007.

[6] C. D. Campbell, E. L. Ogburn, K. L. Lunetta, H. N. Lyon, M. L. Freedman, L. C. Groop, D. Altshuler, K. G. Ardlie, and J. N. Hirschhorn, "Demonstrating stratification in a european american population," *Nat. Genet.*, vol. 37, pp. 868-872, 2005.

[7] C. Tian, R. M. Plenge, M. Ransom, A. Lee, P. Villoslada, C. Selmi, L. Klareskog, A. E. Pulver, L. Qi, P. K. Gregersen, and M. F. Seldin, "Analysis and application of european genetic substructure using 300 k snp information," *PLoS Genet.*, vol. 4, 2008.

[8] A. L. Price, N. A. Zaitlen, D. Reich, and N. Patterson, "New approaches to population stratification in genome-wide association studies," *Nat. Rev. Genet.*, vol. 11, pp. 459-463, 2010.

[9] B. Evlin and K. Roeder, "Genomic control for association studies," *Biometrics*, vol. 55, pp. 997-1004, 1999.

[10] J. K. Pritchard and N. A. Rosenberg, "Use of unlinked genetic markers to detect population stratification in association studies," *Am. J. Hum. Genet.,* vol. 65, pp. 220-228, 1999.

[11] D. E. Reich and D. B. Goldstein, "Detecting association in a casecontrol study while correcting for population stratification," *Genet. Epidemiol.,* vol. 20, pp. 4-16, 2001.

[12] A. L. Price, A. Helgason, S. Palsson, H. Stefansson, D. S. Clair, O. A. Andreassen, D. Reich, A. Kong, and K. Stefansson, "The impact of divergence time on the nature of population structure: an example from iceland," *PLoS Genet.*, vol. 5, p. e1000505, 2009.

[13] J. K. Pritchard, M. Stephens, and P. Donnelly, "Inference of population structure using multilocus genotype data," *Genetics*, vol. 155, pp. 945-959, 2000.

[14] N. A. Rosenberg, J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. W. Feldman, "Genetic structure of human populations," *Genetics*, vol. 298, pp. 2381-2385, 2002.

[15] D. H. Alexander, J. Novembre, and K. Lange, "Fast model-based estimation of ancestry in unrelated individuals," *Genome. Res.*, vol. 19, pp. 1655-1664, 2009.

[16] P. Menozzi, A. Piazza, and L. Cavalli-Sforza, "Synthetic maps of human gene frequencies in europeans," *Science.*, vol. 201, pp. 786-792, 1978.

[17] L. L. Cavalli-Sforza, P. Menozzi, and A. Piazza, "*The history and geography of human genes*," Princeton: Princeton University Press, 1994.

[18] N. Patterson, A. L. Price, and D. Reich, "Population structure and eigenanalysis," *PLoS Genet.*, vol. 2, p. e190, 2006.

[19] J. Novembre and M. Stephens, "Interpreting principal component analyses of spatial population genetic variation," *Nat. Genet.*, vol. 40, pp. 646-649, 2008.

[20] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nat Genet.*, vol. 38, pp. 904-909, 2006.

[21] X. Zhu, S. Zhang, H. Zhao, and R. S. Cooper, "Association mapping, using a mixture model for complex traits," *Genet. Epidemiol.,* vol. 23, pp. 181-196, 2002.

[22] M. J. Aranzana, S. Kim, K. Zhao, E. Bakker, M. Horton, K. Jakob, C. Lister, J. Molitor, C. Shindo, C. Tang, C. Toomajian, B. Traw, H. Zheng, J. Bergelson, C. Dean, P. Marjoram, and M. Nordborg, "Genome-wide association mapping in arabidopsis identifies previously known flowering time and pathogen resistance genes." *PLoS Genet.*, vol. 1, p. e60, 2005.

[23] J. Yu, G. Pressoir, W. H. Briggs, B. I. Vroh, M. Yamasaki, J. F. Doebley, M. D. McMullen, B. S. Gaut, D. M. Nielsen, J. B. Holland, S. Kresovich, and E. S. Buckler, "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness," *Nat. Genet.*, vol. 38, pp. 203-208, 2006.

[24] K. Zhao, M. J. Aranzana, S. Kim, C. Lister, C. Shindo, C. Tang, C. Toomajian, H. Zheng, C. Dean, P. Marjoram, and M. Nordborg, "An arabidopsis example of association mapping in structured samples," *PLoS Genet.*, vol. 3, p. e4, 2007.

[25] C. Sabatti, S. K. Service, A. L. Hartikainen, A. Pouta, S. Ripatti, J. Brodsky, C. G. Jones, N. A. Zaitlen, T. Varilo, M. Kaakinen, U. Sovio, A. Ruokonen, J. Laitinen, E. Jakkula, L. Coin, C. Hoggart, A. Collins, H. Turunen, S. Gabriel, P. Elliot, M. I. McCarthy, M. J. Daly, M. R. J′arvelin, N. B. Freimer, and L. Peltonen, "Genome-wide association analysis of metabolic traits in a birth cohort from a founder population." *Nat. Genet.*, vol. 41, pp. 35-46, 2009.

[26] Y. S. Cho, M. J. Go, Y. J. Kim, J. Y. Heo, J. H. Oh, H. J. Ban, D. Yoon, M. H. Lee, D. J. Kim, M. Park, S. H. Cha, J. W. Kim, B. G. Han, H. Min, Y. Ahn, M. S. Park, H. R. Han, H. Y. Jang, E. Y. Cho, J. E. Lee, N. H. Cho, C. Shin, T. Park, J. W. Park, J. K. Lee, L. Cardon, G. Clarke, M. I. McCarthy, J. Y. Lee, J. K. Lee, B. Oh, and H. L. Kim, "A large-scale genome-wide association study of asian populations uncovers genetic factors influencing eight quantitativetraits," *Nat. Genet.*, vol. 41, pp. 527-534, 2009.

[27] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham, "Plink: A tool set for whole-genome association and population-based linkage analyses," *Am. J. Hum. Genet.*, vol. 81, pp. 559-575, 2007.

[28] A. B. Lee, D. Luca, L. Klei, B. Devlin, and K. Roeder, "Discovering genetic ancestry using spectral graph theory," *Genet. Epidemiol.*, vol. 34, pp. 51-59, 2010.

[29] D. E. Goldgar, "Multipoint analysis of human quantitative genetic variation," *Am. J. Hum. Genet.*, vol. 47, pp. 957-967, 1990.

[30] N. J. Schork, "Extended multipoint identity-by-descent analysis of human quantitative traits: efficiency, power, and modeling considerations." *Am. J. Hum. Genet.*, vol. 53, pp. 1306-1319, 1993.

[31] C. I. Amos and R. C. Elston, "Robust methods for the detection of genetic linkage for quantitative data from pedigrees," *Genet. Epidemiol.*, vol. 6, pp. 349-360, 1989.

[32] C. I. Amos, R. C. Elston, A. F. Wilson, and J. E. Bailey-Wilson, "A more powerful robust sib-pair test of linkage for quantitative traits," *Genet. Epidemiol.*, vol. 6, pp. 435-449, 1989.

[33] C. I. Amos, D. V. Dawson, and R. C. Elston, "The probabilistic determination of identity-by-descent sharing for pairs of relatives from pedigrees," *Am. J. Hum. Genet.*, vol. 47, pp. 842-853, 1990.

[34] L. Andersson, C. S. Haley, H. Ellegren, S. A. Knott, M. Johansson, K. Andersson, L. Andersson-Eklund, I. Edfors-Lilja, M. Fredholm, I. Hansson, "Genetic mapping of quantitative trait loci for growth and fatness in pigs," *Science*, vol. 263, pp. 1771-1774, 1994.

[35] Z. Zhang, E. Ersoz, C.-Q. Lai, R. J. Todhunter, H. K. Tiwari, M. A. Gore, P. J. Bradbury, J. Yu, D. K. Arnett, J. M. Ordovas, and E. S. Buckler, "Mixed linear model approach adapted for genome-wide association studies," *Nat. Genet.*, vol. 42, pp. 355-360, 2010.

[36] H. M. Kang, J. H. Sul, S. K. Service, N. A. Zaitlen, S. yee Kong, N. B. Freimer, C. Sabatti, and E. Eskin, "Variance component model to account for sample structure in genome-wide association studies," *Nat. Genet.*, vol. 42, pp. 348-354, 2010.

[37] C. Lippert, J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson, and D. Heckerman, "Fast linear mixed models for genome-wide association studies," *Nat. Methods*, vol. 8, pp. 833-835, 2011.

[38] G. R. Svishcheva, T. I. Axenovich, N. M. Belonogova, C. M. van Duijn, and Y. S. Aulchenko, "Rapid variance components-based method for whole-genome association analysis," *Nat. Genet.*, vol. 44, pp. 1166-1170, 2012.

[39] C. Ober, M. Abney, and M. S. McPeek, "The genetic dissection of complex traits in a founder population," *Am. J. Hum. Genet.*, vol. 69, pp. 1068-1079, 2001.

[40] G. R. Abecasis, L. R. Cardon, and W. O. Cookson, "A general test of association for quantitative traits in nuclear families," *Am. J. Hum. Genet.*, vol. 66, pp. 279-292, 2000.

[41] S. Mylesa, J. Peiffera, P. J. Browna, E. S. Ersoza, Z. Zhanga, D. E. Costicha, and E. S. Buckler, "Association mapping: critical considerations shift from genotyping to experimental design," *Plant. Cell.*, vol. 21, pp. 2194-2202, 2009.

[42] L. Zhu, Z. Zhang, S. Friedenberg, S.-W. Jung, J. Phavaphutanon, M. Vernier-Singer, E. Corey, R. Mateescu, N. Dykes, J. Sandler, G. Acland, G. Lust, and R. Todhunter, "The long (and winding) road to gene discovery for canine hip dysplasia," *Vet. J.*, vol. 181, pp. 97-110, 2009.

[43] P. J. Bradbury1, Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss, and E. S. Buckler, "Tassel: software for association mapping of complex traits in diverse samples." *Bioinformatics*, vol. 23, pp. 2633-2635, 2007.

[44] C. R. Henderson, "Comparison of alternative sire evaluation methods." *J. Anim. Sci.*, vol. 41, pp. 760-770, 1975.

[45] E. J. Pollak and R. L. Quaas, "Definition of group effects in sire evaluation models." *J. Dairy. Sci.*, vol. 66, pp. 1503-1509, 1983.

[46] R. Thompson, "Sire evaluation." *Biometrics*, vol. 35, pp. 339-353, 1979.

[47] R. L. Quass and E. J. Pollak, "Mixed model methodology for farm and ranch beef cattle testing programs." *J. Anim. Sci.*, vol. 51, pp. 1277-1287, 1980.

[48] H. M. Kang, N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin, "Efficient control of population structure in model organism association mapping," *Genetics*, vol. 178, pp. 1709-1723, 2008.

[49] T. Kariya and H. Kurata, "*Generalized least squares*." Newyork: John Wiley & Sons, 2004.

[50] W. M. Chen and G. R. Abecasis., "Family-based association tests for genomewide association scans." *Am. J. Hum. Genet.*, vol. 81, pp. 913-926, 2007.

[51] P. Rantakallio, "Groups at risk in low birth weight infants and perinatal mortality." *Acta. Paediatr. Scand.*, vol. 193, pp. 1-71, 1969.

[52] E. Michael, N. R. Goddard, K. V.Wray, and M. V. Peter, "Estimating effects and making predictions from genome-wide marker data." *Stat. Sci.*, vol. 24, pp. 517-529, 2009.

[53] B. J. Hayes, P. M. Visscher, and M. E. Goddard, "Increased accuracy of artificial selection by using the realized relationship matrix." *Genet. Res. (Camb)*, vol. 91, pp. 47-60, 2009.

[54] J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher, "Common snps explain a large proportion of heritability for human height," *Nat. Genet.*, vol. 42, pp. 565-569, 2010.

[55] Y. S. Aulchenko, D. J. de Koning and C. Haley, "Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigreebased quantitative trait loci association analysis," *Genetics.*, vol. 177, pp. 577-585, 2007.

[56] X. Zhu, S. Li, R. S. Cooper, and R. C. Elston., "A unified association analysis approach for family and unrelated samples correcting for stratification," *Am. J. Hum. Genet.,* vol. 82, pp. 352-365, 2008.

[57] D. J. Balding and R. A. Nichols, "A method for quantifying differentiation between populations at multiallelic loci and its implications for investigating identity and paternity," *Genetica.*, vol. 96, pp. 3-12, 1995.

[58] P. M. Visscher, W. G. Hill, and N. R. Wray, "Heritability in the genomics era concepts and misconceptions," *Nat. Rev. Genet.*, vol. 9, pp. 255-266, 2008.

Received: August 06, 2013 Revised: September 06, 2013 Accepted: September 15, 2013