Efficient Multivariate Analysis Algorithms for Longitudinal Genome-wide Association Studies

Chao Ning[1,3], Dan Wang[1,3], Lei Zhou[1], Huimin Kang[1], Shengli Zhang[1], Jian Yang[2] and Jian-Feng Liu[1]

[1]National Engineering Laboratory for Animal Breeding; Key Laboratory of Animal Genetics, Breeding and Reproduction, Ministry of Agriculture; College of Animal Science and Technology, China Agricultural University, Beijing 100193, China.
[2]Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland 4072, Australia.
[3]These authors contributed equally to this work.
Correspondence should be addressed to J.F.L. (liujf@cau.edu.cn).

Current dynamic phenotyping system introduces time as an extra dimension to genome-wide association studies (GWAS). We develop efficient genome-wide multivariate analysis algorithms (GMA) to deal with longitudinal data. In contrast with existing univariate mixed linear model analyses, our algorithms have improved statistic power and computation speed, and can deal with unbalanced longitudinal data of thousands individuals with more than ten thousand records in a few hours and balanced longitudinal data in a few minutes. We package the analysis algorithms in a friendly and efficient software GMA (https://github.com/chaoning/GMA) for facilitating investigators in relevant fields.

Longitudinal traits, also known as dynamic traits or function-valued traits, are dynamically changing over a period of time controlled by the genetic developmental processes. Advanced dynamic phenotyping system in animal and plant genetic filed[1, 2]

makes it feasible to acquire high-throughput time-varied datasets. It is beneficial with such repeated measurements under varying environmental conditions to improve the quantitative trait nucleotide (QTN) detection power and explore dynamical genetic control mechanism of complex longitudinal traits.[3, 4] Analyzing such types of datasets also promotes early prediction of longitudinal traits or diseases.[5, 6]

However, the existing commonly employed analytical methods, such as varying-coefficient regression[7] and estimation equation[8], is computationally intensive compared to univariate analyses. One alternative way to reduce computing time is to analyze each single time point separately and then integrate test statistics across time points to determine the overall significance.[9] However, the method is less efficient in QTN detection because it ignores the dependency among the time course observations.

Random regression model (RRM), which is a kind of multivariate mixed linear model (mvMLM), has been widely applied to longitudinal data analysis in animal breeding[10]. Our previous studies have proved its advantage in dealing with longitudinal GWAS[11]. The method regards SNP effects as fix regression and use the sparse matrix technique in ASReml[12] and population parameters previously determined (P3D) algorithm[13] to reduce computing time. However, it is computationally challenging when dense marker-based kinship matrix instead of sparse pedigree-derived relationship matrix is used to reflect individual genetic relationships (computational complexity of $O(m^3)$, where m is the total number of phenotypic records).

In the study, we present two novel efficient algorithms, *i.e.*, the fix regression with Eigen decomposition strategy[14, 15, 16] (GMA-fix) and linear transformation of genomic estimation values[17, 18] (GMA-lt), for performing both unbalanced (every individual can be recorded at different time points) and balanced (every individual is recorded at the same time points) longitudinal GWAS. Novelty of the proposed algorithms in GMA-fix for longitudinal GWAS, we treat SNP effects as fix regressions under the

framework of RRM and conduct Eigen decomposition of phenotypic (co)variance matrix to rotate the RRM, which transforms the model into weighted least squares analysis. The computational complexity of longitudinal GWAS step is reduced to $O(m^2)$ hence. Parallel in GMA-lt, we improve the computing speed by linear transformation of genomic estimated values algorithm. GMA-lt takes advantage of the intermediate matrixes in estimating the variances of null model and avoid calculating the phenotypic (co)variance matrix and its Eigen decomposition. It reduces the computational complexity from $O(m^2)$ to $O([(df + 1)\ n]^2)$, where n is the number of individuals and df is the order of Legendre polynomials fitting the SNP effect. In order to guarantee the convergence in the process of variance component estimation, $(df + 1)*n$ is usually less than m.

Additionally, we further enhance the GMA performance for balanced longitudinal data through conducting Eigen decomposition of the genomic relatedness matrix (complexity of $O(n^3)$) to rotate the RRM (complexity of $O(n^2)$). The time complexity of variance component estimation for rotated RRM is $O(n)$ compared with $O(n^3)$ of unbalanced longitudinal data. Taking this advantage, we improve the QTN detection power of GMA-fix through re-estimating the variance components for each SNP test. In order to guarantee the variance components converged within the legal parameter space rapidly, we incorporate expectation maximization (EM) information into average information (AI) matrix to build a weighted information matrix[19].

We conduct empirical analyses using two datasets, an intercross F2 mouse dataset[20] with body weight trait and a dairy cow dataset[11] with milk yield trait, to validate the benefits of our novel algorithms in longitudinal GWAS. The mouse dataset is a small balanced data with 1,212 mice measured from age 1 week to age 16 weeks; and the dairy cow dataset is a large unbalanced data with 5,982 cows of 52,732 records across the first lactation (from day 5 to day 305).

Ahead of GWAS, we firstly compare our efficient algorithm in variance components

estimation with Wombat[21] and MTG2[16] (Table 1). In general, GMA takes fewer iterations and much less time to convergence. In balanced longitudinal mouse data, our algorithm takes only 2 seconds compared with 5 seconds for MTG2 and 40 minutes for Wombat. Even for unbalanced longitudinal dairy cow data, GMA is substantially faster than Wombat.

In the application of longitudinal GWAS, GMA-lt algorithm has similar performance with uvMLM for unbalanced data and is much faster for balanced data. Furthermore, the current GMA-lt algorithm in unbalanced data is several times faster than GMA-fix algorithm. We compared the *P*-values from GMA-fix and GMA-lt and find that they entail the exact the same values (Supplementary Fig. 1C). For balanced data, GMA-fix will optimize the variance components for per-SNP and is much slower than GMA-lt. The correlation coefficient of their *P*-values is very high (Pearson's r = 0.995) while GMA-fix *P*-values are much more significant than GMA-lt *P*-values (Supplementary Fig. 1A). Considering the strengths of GMA-lt in computational speed and GMA-fix with re-estimated variance components in statistical power, we can pre-select SNPs less than a certain *P*-values (*e.g.*, 0.01) through running GMA-lt, and recalculate their GMA-fix *P*-values, thus the loss power will be rescued and computation time will be reduced to the same level with GMA-lt (about 7 min).

In the meantime, we use simulation data to further validate the performance of our algorithm. We calculate kinship matrix from original SNPs and randomly shuffle the genome-wide SNPs. These shuffled SNPs are not related to the phenotypic values and their *P*-values are expected to follow uniform distribution of $U(0, 1)$. The results in Fig. 1 show that our longitudinal GWAS algorithm and uvMLM algorithm can control type I error well, while uvMLM-min has high type I error. The uvMLM-min algorithm origins from Kwak *et al*.[9]. Here, in balanced data, we analyse each single time point with MLM and use the minimum *P*-values to determine the significance. In the unbalanced data, we random select five records and include the time as covariant in the MLM. The power evaluation is based on adding time-varied QTN effects to

original phenotypic values. Nine different QTN curves are simulated for balanced and unbalanced data (Supplementary Fig. 2 and Supplementary Fig. 3). Our analysis algorithms gains more power compared to other two methods. The exact GMA-fix algorithm in the balanced data is the most powerful, which verifies the results of real data analysis. In uvMLM algorithm, the power estimation is obtained by averaging across different time points. We use the algorithm to simulate the expectation power of ignoring time variable and gathering phenotypes at a random time point. The algorithm has the lowest statistic power, which demonstrates the benefit of longitudinal data,

In summary, we develop two analysis algorithms for balanced and unbalanced longitudinal GWAS. For balanced data, fix regression algorithm, which optimizes variance components per-SNP, is more powerful, while linear transformation of genomic estimation values algorithm is much faster. The GMA-lt following GMA-fix is recommended. For unbalanced data, it is time consuming to optimize variance components for each SNP. GMA-fix and GMA-lt have the same power, but GMA-lt is faster. Thus, GMA-lt is recommended for unbalanced longitudinal GWAS.

GMA software is available at https://github.com/chaoning/GMA.

Online Methods

Data. We analyse two datasets in our study: a mouse data[20] and a dairy cow data[11].
The mouse data 1,212 F2 intercross mice between Gough Island mice and WSB/EiJ. The body weight are measured from age 1 week to age 16 weeks (16 phenotypes) through age for all mice. There are 11,833 markers to be used in the analysis. The dairy cow data includes 5,982 individuals. The milk yield trait of the first parity were analysed in this study. The cows with less than six records were filtered out, which resulted a total of 52,732 records. The SNPs with a minor allele frequency (MAF) lower than 0.03 and those that failed the Hardy–Weinberg equilibrium (HWE) test

(P-value < 10^{-6}) were removed, resulting in 71,527 SNPs for the subsequent longitudinal GWAS analyses.

Simulation. In order to assess the null distribution of different models, we randomly shuffle the genome-wide SNPs. Under the expectation that random SNPs are unlinked to polymorphisms controlling these traits, the cumulative P-value distribution follows a uniform distribution of $U(0, 1)$. We estimate the statistic power of different methods by adding maker effects back to the original phenotypes.[22] We random select 100 SNPs from the genome and allocate them with nine different maker effect curve. The time-varied SNP effects are then adjusted to different explain different phenotypic variances (average across the time points, 0.02-2% at MAF of 0.5). The genetic effect was assigned to the 100 random SNPs, one at a time, and each model was run to determine whether the effect could be detected with the empirical threshold (the 5th percentile of a null distribution).

GMA. We detail the GMA in the Supplementary Information.

Reference

1. Porto SMC, Arcidiacono C, Anguzza U, Cascone G. The automatic detection of dairy cow feeding and standing behaviours in free-stall barns by a computer vision-based system. *Biosystems Engineering* **133**, 46-55 (2015).

2. Fahlgren N, Gehan MA, Baxter I. Lights, camera, action: high-throughput plant phenotyping is ready for a close-up. *Current opinion in plant biology* **24**, 93-99 (2015).

3. Wu R, Lin M. Functional mapping—how to map and study the genetic architecture of dynamic complex traits. *Nature Reviews Genetics* **7**, 229-237 (2006).

4. Li Z, Sillanpaa MJ. Dynamic Quantitative Trait Locus Analysis of Plant Phenomic Data. *Trends in plant science* **20**, 822-833 (2015).

5. Kellogg EC, Thrasher A, Yoshinaga-Itano C. Early predictors of autism in young children who are deaf or hard of hearing: three longitudinal case studies. *Seminars in speech and language* **35**, 276-287 (2014).

6. McSweeney J, *et al.* Predicting coronary heart disease events in women: a longitudinal cohort study. *The Journal of cardiovascular nursing* **29**, 482-492 (2014).

7. Gong Y, Zou F. Varying coefficient models for mapping quantitative trait loci using recombinant inbred intercrosses. *Genetics* **190**, 475-486 (2012).

8. Xiong H, Goulding EH, Carlson EJ, Tecott LH, McCulloch CE, Sen S. A flexible estimating equations approach for mapping function-valued traits. *Genetics* **189**, 305-316 (2011).

9. Kwak IY, Moore CR, Spalding EP, Broman KW. A simple regression-based method to map quantitative trait loci underlying function-valued phenotypes. *Genetics* **197**, 1409-1416 (2014).

10. Schaeffer LR. Application of random regression models in animal breeding. *Livest Prod Sci* **86**, 35-45 (2004).

11. Ning C, *et al.* Performance Gains in Genome-Wide Association Studies for Longitudinal Traits via Modeling Time-varied effects. *Scientific reports* **7**, (2017).

12. Gilmour A, *et al.* ASReml user guide. Release 4.1 structural specification. *VSN International Ltd, Hemel Hempstead, HP1 1ES, UK www vsni co uk,* (2014).

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

14. Kang HM, *et al.* Efficient control of population structure in model organism association mapping. In: *Genetics* (ed^(eds) (2008).

15. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature methods* **11**, 407-409 (2014).

16. Lee SH, van der Werf JH. MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information. *Bioinformatics* **32**, 1420-1422 (2016).

17. Gualdron Duarte JL, Cantet RJ, Bates RO, Ernst CW, Raney NE, Steibel JP. Rapid screening for phenotype-genotype associations by linear transformations of genomic evaluations. *BMC bioinformatics* **15**, 246 (2014).

18. Ning C, *et al.* A rapid epistatic mixed-model association analysis by linear retransformations of genomic estimated values. *Bioinformatics* **34**, 1817-1825 (2018).

19. Jensen J. Residual maximum likelihood estimation of (co) variance components in multivariate mixed linear models using average information. *Journal of the Indian Society of Agricultural Statistics* **49**, 215-236 (1997).

20. Gray MM, *et al.* Genetics of Rapid and Extreme Size Evolution in Island Mice. *Genetics* **201**, 213-228 (2015).

21. Meyer K. WOMBAT: a tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). *Journal of Zhejiang University Science B* **8**, 815-821 (2007).

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

23. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics* **44**, 821-824 (2012).

24. Liu C, Rubin DB, Wu YN. Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika* **85**, 755-770 (1998).

Table 1 Computation time of different method in variances estimation (including iteration times) and GWAS.

	Computation time	
Method	**Mouse dataset**	**Dairy cow dataset**
Variances estimation		
Wombat	40 min (11)	105 (12)
MTG2	5 s (15)	-
GMA	2 s (9)	5.3 h (7)
GWAS		
uvMLM	14.4 min	3.7 h
GMA-fix	5.1 h	16.5 h
GMA-lt	1.7 min	3.8 h
GMA-lt + GMA-fix	7 min	-

All computing was performed on Intel Xeon E5 2.2 GHz CPU. We use three orders of Legendre polynomials for mouse dataset and four orders for dairy cow dataset as higher order encounters convergence problem. The same convergence criteria is used for different methods in variances estimation, *i.e.*, increase in log likelihood values between subsequent iterates is 0.001. The uvMLM is implemented with GEMMA[23]. In variance components estimation, Wombat uses a hybrid algorithm consisting of a few initial rounds of PX-EM[24], followed by AI algorithm; MTG2 uses the pure AI algorithm and moderates the magnitude of updates when the parameters go outside the parameter space; GMA incorporates EM information into AI matrix to build a weighted information matrix.

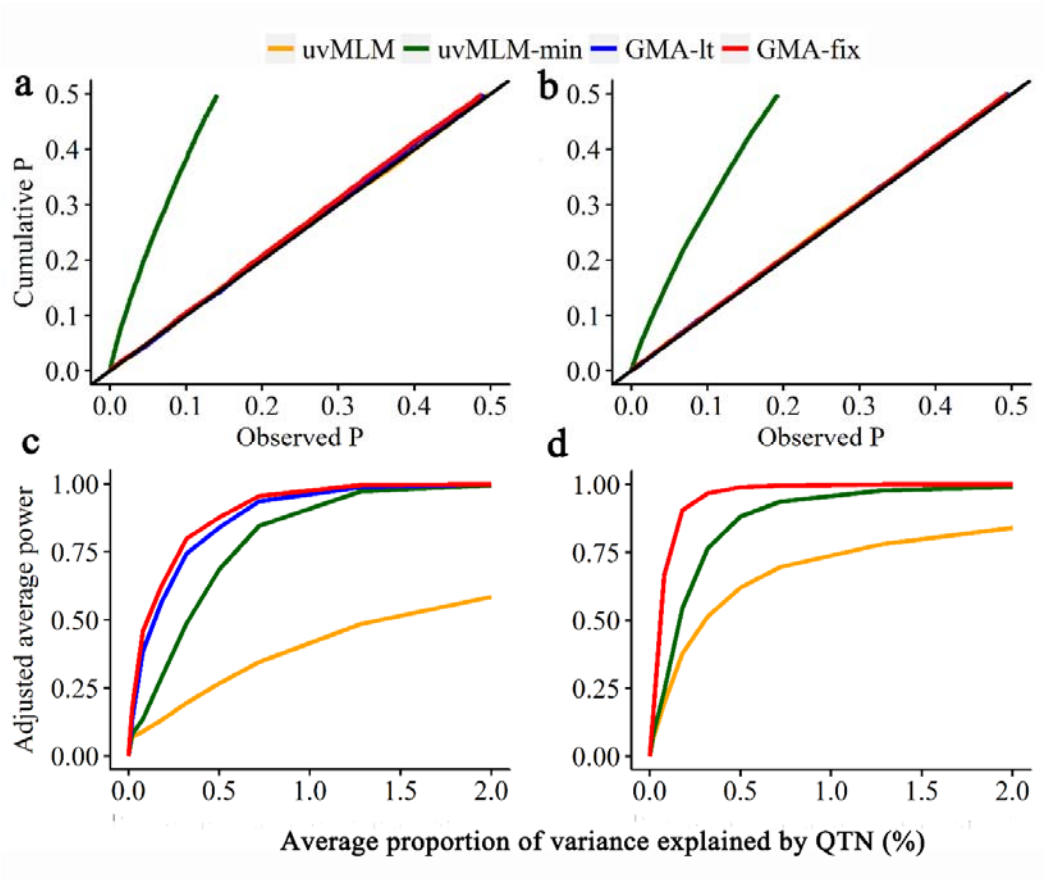

Fig. 1 Cumulative *P*-value distributions and adjusted statistical powers of different methods in the simulation study. The left panel is based on the balanced mouse data, and the right panel is based on the unbalanced dairy cow data. (a-b) The cumulative *P*-value distributions of random shuffled SNPs. Under the expectation that random SNPs are unlinked to polymorphisms controlling these traits, the cumulative *P*-value distribution should follow a uniform distribution of *U*(0, 1). Upward departures from the expectation indicate that the methods cause spurious associations. (c-d) The adjusted average power at different QTN explained phenotypic variances. The phenotypic variances are average across the time points for QTN with allele frequency 0.5. The average adjusted power are calculated from 100 QTNs with nine different genetic curve effects.