Bioinformatics, 35(23), 2019, 4879–4885
doi: 10.1093/bioinformatics/btz304
Advance Access Publication Date: 9 May 2019
Original Paper

OXFORD

Genome analysis

Efficient multivariate analysis algorithms for longitudinal genome-wide association studies

Chao Ning [1], Dan Wang[1], Lei Zhou[1], Julong Wei[2], Yuanxin Liu[3], Huimin Kang[1], Shengli Zhang[1], Xiang Zhou[2], Shizhong Xu[4] 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]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA, [3]School of English, Beijing International Studies University, Beijing 100024, China and [4]Department of Botany and Plant Science, University of California, Riverside, CA 20705, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on October 8, 2018; revised on April 16, 2019; editorial decision on April 18, 2019; accepted on April 25, 2019

Abstract

Motivation: Current dynamic phenotyping system introduces time as an extra dimension to genome-wide association studies (GWAS), which helps to explore the mechanism of dynamical genetic control for complex longitudinal traits. However, existing methods for longitudinal GWAS either ignore the covariance among observations of different time points or encounter computational efficiency issues.

Results: We herein developed efficient genome-wide multivariate association algorithms for longitudinal data. In contrast to existing univariate linear mixed model analyses, the proposed method has improved statistic power for association detection and computational speed. In addition, the new method can analyze unbalanced longitudinal data with thousands of individuals and more than ten thousand records within a few hours. The corresponding time for balanced longitudinal data is just a few minutes.

Availability and implementation: A software package to implement the efficient algorithm named GMA (https://github.com/chaoning/GMA) is available freely for interested users in relevant fields.

Contact: liujf@cau.edu.cn

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

1 Introduction

Genome-wide association studies (GWAS) have been used to detect many genetic variants associated with various quantitative traits and complex diseases. Linear mixed models (LMM) adopted to GWAS (Kang *et al.*, 2008; Lippert *et al.*, 2011; Yu *et al.*, 2006; Zhou and Stephens, 2012) are able to capture genetic correlation among individuals, correct confounding environmental factors and control population stratification. However, most LMM-based GWAS analytical tools, such as EMMA/EMMAX (Kang *et al.*, 2008, 2010), FaST-LMM (Lippert *et al.*, 2011), GEMMA (Zhou and Stephens, 2012) and GCTA (Yang *et al.*, 2011), focus on cross-sectional traits that are measured only once. There are few methods available for GWAS dealing with longitudinal traits that are repeatedly measured during the life span of individual development.

Longitudinal traits, also known as dynamic traits or functional traits, are dynamically changing over a period of time controlled by both genetic effects and environmental factors. Multiple measurements at various time points during a life cycle are usually collected as longitudinal traits. Recently, advanced dynamic phenotyping system in animal and plant genetic experiments (Fahlgren *et al.*, 2015; Porto *et al.*, 2015) makes it feasible to acquire high-throughput time-varied datasets. Such repeated measurements under varying environmental conditions can improve statistical power of quantitative trait nucleotide (QTN) detection and help to

further explore the mechanism of dynamical genetic control for complex longitudinal traits (Li and Sillanpää, 2015; Wu and Lin, 2006). Analyzing such types of datasets also promotes early prediction of longitudinal traits and diseases (Kellogg *et al.*, 2014; McSweeney *et al.*, 2014).

However, currently employed analytical methods, such as varying-coefficient regression (Gong and Zou, 2012) and estimation equation (Xiong *et al.*, 2011), are computationally intensive compared to the univariate counterpart. An alternative way to improve computational efficiency is to analyze each single time point separately and then integrate test statistics across time points to determine the overall significance (Kwak *et al.*, 2014). However, the single time point analysis is inefficient in QTN detection because it ignores the covariance among observations of different time points.

Random regression models (RRM) are multivariate linear mixed models (mvLMM) and have been widely applied to longitudinal data analysis in animal breeding (Schaeffer, 2004). Our previous studies demonstrated the advantages of longitudinal GWAS over single-trait GWAS (Ning *et al.*, 2017). In our previous methods, we treat SNP effects as fixed regression coefficients and use a sparse matrix technique in ASReml (Gilmour *et al.*, 2014) along with the population parameters previously determined (P3D) algorithm (Zhang *et al.*, 2010) to reduce computing time. However, it is still computationally challenging when a marker inferred dense kinship matrix (rather than a sparse pedigree-derived numerator relationship matrix) is used to capture individual genetic relationships. With the marker-inferred kinship matrix, the computational complexity is $O(m^3)$, where m is the total number of phenotypic records. To address the low computational efficiency problem, we previously proposed to use eigen decomposition to rotate the RRM and transform the model into weighted least squares analysis (Ning *et al.*, 2018), which reduced the computational complexity to $O(m^2)$. However, the method depends on the existing software to estimate variance parameters.

To further address the computational efficiency limitations, we developed two efficient multivariate algorithms for longitudinal GWAS (GMA): fixed regression strategy with eigenvalue decomposition (GMA-fixed) and linear transformation of genomic estimation values (GMA-trans) for unbalanced and balanced longitudinal traits, where unbalanced means that different individuals may be recorded at different time points and balanced means that all individuals are measured at the same time points. GMA-fixed for unbalanced data is similar to our previous study (Ning *et al.*, 2018). Here, we applied the weighted information matrix method to variance parameters estimation, which is faster than existing software. We developed an efficient and user-friendly software for unbalanced and balanced longitudinal GWAS, which is independent from existing software.

In order to investigate the properties of our new methods, a series of simulation studies were conducted to compare the methods with the existing univariate linear mixed model (uvLMM) method. Furthermore, we validated our methods using an unbalanced dairy cow milk production dataset and a balanced mouse growth dataset.

2 Materials and methods

2.1 Multivariate analysis algorithms for longitudinal GWAS (GMA)

Details of GMA are presented in Supplementary Note. A typical RRM to model time-varied genetic and environment effects (Mrode, 2014) can be written as

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Qa} + \mathbf{Zp} + \mathbf{e} \tag{1}$$

Where $\mathbf{y}$ is the vector of phenotypic values; $\mathbf{b}$ is the vector of fixed effects and fixed regression polynomials coefficients; $\mathbf{a}$ and $\mathbf{p}$ are the vectors of random regression polynomials coefficients for additive genetic effects and subject-specific permanent environmental effects, respectively; $\mathbf{e}$ is the vector of random residuals. $\mathbf{X}$, $\mathbf{Q}$ and $\mathbf{Z}$ are the corresponding design matrices. It is assumed that

$$\mathbf{a} \sim N(0, \mathbf{K} \otimes \sum\nolimits_a), \quad \mathbf{p} \sim N(0, \mathbf{I} \otimes \sum\nolimits_p), \quad \mathbf{e} \sim N(0, \mathbf{R}) \tag{2}$$

Here, $\mathbf{K}$ is a marker-derived relationship matrix; $\mathbf{I}$ is the identity matrix; $\otimes$ is the Kronecker product; $\sum_a$ is the (co)variance matrix for random regression coefficients of additive polygenic effects; $\sum_p$ is the variance–covariance matrix of random regression coefficients for permanent environmental effects; $\mathbf{R}$ is a diagonal matrix with different values at different time periods. In the variance parameters estimation, we incorporated the expectation-maximization (EM) algorithm into the average information (AI) matrix to build a weighted information matrix (Jensen, 1997), which guarantees the variance parameters to converge rapidly within their legal domain.

$$\mathbf{I}_{AE} = \lambda \mathbf{I}_{EM} + (1-\lambda)\mathbf{I}_{AI}, \quad \theta^{(f+1)} = \theta^{(f)} + (\mathbf{I}_{AE}{}^{(f)})^{-1} \frac{\partial L}{\partial \theta} | \theta^{(f)} \tag{3}$$

Where $\mathbf{I}_{EM}$, $\mathbf{I}_{AI}$ and $\mathbf{I}_{AE}$ are the EM information matrix, AI matrix and weighted information matrix, respectively. $\boldsymbol{\theta}$ is a vector of variance components including the unique values in $\sum_a$, $\sum_p$ and $\mathbf{R}$; f is the iteration round; $\partial L/\partial \boldsymbol{\theta}$ is a vector of the first derivatives of the log-likelihood function with respect to each variance component.

In the longitudinal GWAS analysis, the GMA-fixed and GMA-trans algorithms are applied in unbalanced and balanced data, respectively. GMA-fixed algorithm for unbalanced data is similar to our previous published paper (Ning *et al.*, 2018). Detailed formula derivation is shown in the Supplementary Note. We treated each SNP effect as fixed regression coefficients and used the Legendre polynomials to model the time-dependent SNP effects. The matrix form is

$$\mathbf{y} = \mathbf{Xb} + \mathbf{X}_{SNP}\mathbf{b}_{SNP} + \mathbf{Qa} + \mathbf{Zp} + \mathbf{e} \tag{4}$$

Where, $\mathbf{b}_{SNP}$ is a vector of fixed regression polynomials coefficients for SNP effect and $\mathbf{X}_{SNP}$ is corresponding design matrix. We construct Wald χ^2 statistic to examine whether all elements in $\mathbf{b}_{SNP}$ are zeros. With the eigenvalue decomposition technique, computational complexity of such a longitudinal GWAS step is reduced from $O(m^3)$ to $O(m^2)$ per SNP, where m is the total number of phenotypic records.

In the GMA-trans for unbalanced data, we first estimate the individuals' genetic effects with the mixed-model equations (MME) as follows

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Q} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Q}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Q}'\mathbf{R}^{-1}\mathbf{Q} + \mathbf{K}^{-1} \otimes \sum\nolimits_a^{-1} & \mathbf{Q}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Q} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{I} \otimes \sum\nolimits_p^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \\ \hat{\mathbf{p}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Q}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \tag{5}$$

In the Supplementary Note, we show that the estimated values of random regression coefficients for jth SNP and corresponding (co)variances are

$$\hat{\mathbf{u}}_j = (\mathbf{s}_j{}'\mathbf{K}^{-1} \otimes \mathbf{I})\hat{\mathbf{a}} \tag{6}$$

$$\mathrm{var}(\hat{\mathbf{u}}_j) = (\mathbf{s}_j{}'\mathbf{K}^{-1} \otimes \mathbf{I})\mathrm{var}(\hat{\mathbf{a}})(\mathbf{s}_j{}'\mathbf{K}^{-1} \otimes \mathbf{I})' \tag{7}$$

The Wald χ^2 test for time-dependent SNP effect is

$$\hat{\mathbf{u}}_j'[\mathrm{var}(\hat{\mathbf{u}}_j)]^{-1}\hat{\mathbf{u}}_j \sim \chi^2(nr_1 + 1) \tag{8}$$

Where nr_1 is the order of the Legendre polynomials fitting time-dependent genetic effects. Compared with GMA-fixed, GMA-trans

takes advantage of some intermediate results of matrix calculation in the variance parameter estimation step and avoids calculation of the phenotypic (co)variance matrix and its eigenvalue decomposition. This has reduced the computational complexity from $O(m^2)$ to $O([n(df+1)]^2)$, where n is the number of individuals and df is the order of the Legendre polynomials fitting the SNP effect. To ensure convergence of the iterations in the process of variance component estimation, df is usually less than five and thus $n(df+1)$ is smaller than m for the usual condition of more than five measures per individual.

For balanced longitudinal data, if model the additive genetic effect and permanent environmental effect with the same order of Legendre polynomials, matrix **Q** is equal to **Z** and the submatrices for every individual in **Q** and **Z** are the same. Here we define the submatrix as **T**, then Equation (2) can be rewritten as

$$\mathbf{y} = \mathbf{Xb} + (\mathbf{I} \otimes \mathbf{T})\mathbf{a} + (\mathbf{I} \otimes \mathbf{T})\mathbf{p} + \mathbf{e} \tag{9}$$

Eigen decomposition of the genomic relationship matrix **K** gives $\mathbf{K} = \mathbf{UDU}'$. We rotate Equation (9) with $\mathbf{U}' \otimes \mathbf{I}$, and the phenotypic (co)variance matrix for the rotated model is

$$\mathbf{V}* = \mathrm{var}((\mathbf{U} \otimes \mathbf{I})\mathbf{y}) = \mathbf{D} \otimes (\mathbf{TGT}') + \mathbf{I} \otimes (\mathbf{TPT}') + \mathbf{I} \otimes \sum \tag{10}$$

V* is a block diagonal matrix and the size for each block is the number of records per individual (s). Therefore, we can obtain its inversion with time complexity of $O(ns^3)$ compared with $O(m^3)$ for the unbalanced longitudinal data. With the rotated RRM, we improved the QTN detection power of GMA-fixed through re-estimating the variance components for each tested SNP. The computational complexity for GMA-trans is also reduced to $O(ns^3)$ in the rotated RRM. The details of GMA-fixed and GMA-trans for balanced longitudinal data are given in the Supplementary Note.

2.2 Univariate linear mixed model

Parallel to GMA-fixed and GMA-trans methods, we also consider the following standard LMM

$$\begin{aligned} &\mathbf{y} = \mathbf{X\beta} + \mathbf{w}\alpha + \mathbf{u} + \mathbf{e} \\ &\mathbf{u} \sim N(\mathbf{0}, \mathbf{K}\sigma^2{}_a), \quad \mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma^2{}_e) \end{aligned} \tag{11}$$

Where **y** is the vector of cross-sectional phenotypic values at a single time point; **β** is a vector of fixed effects; **X** is the design matrix which relates records to fixed effects; **w** is a vector of marker genotypes; α is the effect size of the marker; **u** is the random polygenic effects; **e** is the residuals; **K** is the marker-derived relationship matrix; $\sigma^2{}_a$ and $\sigma^2{}_e$ are the variances for polygenic effects and random residuals, respectively. The Wald χ^2 statistic-implemented GEMMA is used to examine whether the marker effect size (α) is zero. We also applied the following two methods based on univariate LMM in our analysis.

1. uvLMM-mean: It represents uvLMM to obtain the empirical detection power of cross-sectional analysis method that utilizes only phenotypic data at a single time point. We only applied the method in the simulation study. For each repeated simulation, we analysed a random measurement of each analysis and repeated a certain number of times for unbalanced data (sampling without replacement) or analysed the measurement of each single time point separately for balanced data with the LMM method. The power estimation was obtained by taking the mean power across different analyses.
2. uvLMM-min: It represents uvLMM via the minimum P-value. The algorithm was originated from Kwak *et al.* (2014), where they developed a simple regression-based method to map QTL for longitudinal data. In our study, we analysed a random measurement each time and repeated a certain number of times for unbalanced data (sampling without replacement) or analysed one measurement for each time point separately for balanced data with the LMM method. The minimum P-value across different analyses per SNP was used to determine the significance for a SNP.

2.3 Data

Two datasets were analysed in the study: a mouse data (Gray *et al.*, 2015) and a dairy cow data (Ning *et al.*, 2017). The mouse data contain $1212F_2$ from the cross between the Gough Island mice and the WSB/EiJ strain. The body weight and growth rate (estimated as the first derivative of the fitted cubic splines) was obtained from week 1 to week 16 incremented by 1 week (16 measurements per mouse). There are 11 833 available SNP markers across the mouse genome after proper quality control. The dairy cow data include 5982 individual cows. The milk yield, fat percentage and protein percentage 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) less than 0.03 and a failed the Hardy–Weinberg equilibrium (HWE) test (P-value $< 10^{-6}$) were removed, resulting in 71 527 SNPs for the subsequent longitudinal GWAS analyses.

2.4 Simulation

In order to assess whether different models can control type I error well, we calculated the kinship matrix from the original SNPs and randomly shuffled SNP analysis across individuals at each analysis, which can purposely destroy the association of the phenotypes with the scanned SNP and the linkage disequilibrium (LD) among SNPs. The covariance structure of original phenotypes induced by the complex cryptic genetic relationship among the individuals will not be disorganized in this way. Under the expectation that random SNPs are unlinked to polymorphisms controlling these traits, the cumulative P-value distribution follows a uniform distribution of $U(0, 1)$. The empirical power was obtained from populations simulated from the genotypes of the current populations (the mouse and the cattle data) by assigning genetic effects to selected markers and adding maker effects back to the original phenotypes (Yu *et al.*, 2006), i.e. $y_{i,\mathrm{new}}(t) = y_i(t) + s_i\mathrm{SNP}(t)$. Where $y_i(t)$ is the observed phenotypic value of individual i at time t; s_i is a genotype indicator for individual i which is assigned 0, 1 and 2 for genotype *aa*, *Aa* and *AA*, respectively; SNP(t) represents the simulated time-varied effect for selected marker; $y_{i,\mathrm{new}}(t)$ is the newly generated phenotypic value of individual i at time t. We randomly selected 100 SNPs from the genome and assigned them with nine different maker effect curves. The time-varied SNP effects were then adjusted so that they contributed to some predetermined proportions of the phenotypic variance (average proportion across the time points, 0.02–2% at MAF of 0.5). The genetic effect curves were assigned to the 100 random selected SNPs, one at a time. The simulated data were analysed by the proposed new methods and existing methods. Permutation test with method of Churchill and Doerge (1994) was used to determine whether a marker is significant at pointwise and genome-wide level. In order to maintain the kinship matrix intact, we calculated the kinship matrix from the original SNPs. To determine the pointwise significance, we randomly shuffled analysed marker and analysed the maker with different methods. We repeated the process with 1000 times. Then, we ordered the P-values and the fifth percentile was

used as empirical threshold. To determine the genome-wide significance, we permuted complete marker records and analysed the genome-wide maker to find the minimum *P*-values. The process was also repeated with 1000 times. We ordered these *P*-values and the fifth percentile was used to determine the significance.

3 Results

3.1 Simulation

We first validated the performance of GMA with simulated data. To make the simulation as close as possible to reality, we perform simulations based on two real datasets, a dairy cow dataset (Ning *et al.*, 2017) with milk yield, fat percentage and protein percentage traits and an inter-cross F_2 mouse dataset (Gray *et al.*, 2015) with body weight and growth rate traits. Figure 1A and B and Supplementary Figures S1A and B and S2A show that the type I errors are well controlled by our longitudinal GWAS algorithms and the uvLMM-mean algorithm, but are not controlled by the uvLMM-min method.

We obtained empirical statistic powers of different methods by adding QTN effects back to the original phenotypes (Yu *et al.*, 2006). Nine different QTN effect functions (curves) were simulated for the unbalanced dairy cow data and the balanced mouse data (Supplementary Figs S3 and S4). The results are illustrated in Figure 1C and D and Supplementary Figures S1C and D, S2B and S5 showing that the new methods have higher power than two uvLMM methods. In particular, the approximate GMA-fixed algorithm for the unbalanced data has almost the same power as GMA-trans, while the exact GMA-fixed algorithm for the balanced data (optimize variance parameters for each SNP) has the highest power. The uvLMM-mean algorithm has the lowest statistic power, which demonstrates the benefit of using the new GWAS methods of longitudinal traits.

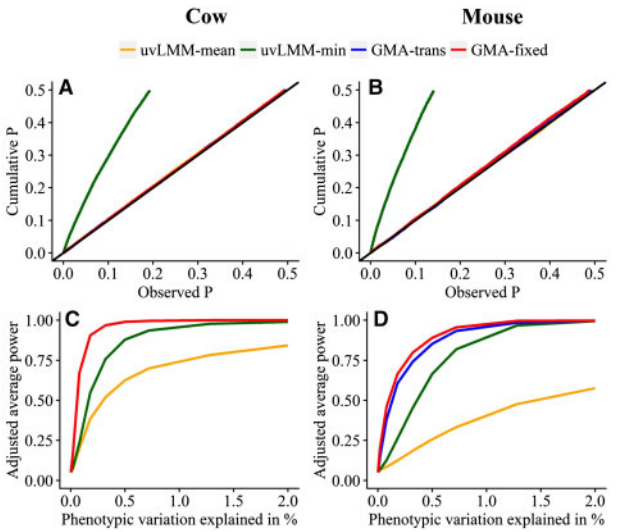

Fig. 1. Cumulative *P*-value distributions and pointwise powers of different methods in the simulation study. The left panels (**A** and **C**) represent the unbalanced dairy cow data with milk yield traits and the right panels (**B** and **D**) represent for the balanced mouse data with body weight traits. The upper panels (**A** and **B**) represent distributions of the randomly shuffled SNPs. Under the null model, the cumulative *P*-value distribution should follow a uniform distribution of *U*(0, 1) that overlaps with the diagonal line. Deviation from the diagonal line indicates spurious associations. The lower panels (**C** and **D**) represent the adjusted average power at different QTN contributions. The phenotypic variance is the average variance across different time points for QTN with allele frequency 0.5. The power estimation was based on whether the *P*-value was smaller the empirical threshold (the fifth percentile by permutation test), and we averaged the powers of different type of maker effect curve. The red line overlapping with the blue line in panel C indicates that GMA-fixed and GMA-trans have very similar power for the dairy cow data analysis

3.2 Application to real data

We applied the GMA to analyse milk yield of unbalanced dairy cow data and body weight of balanced mouse data. Prior to scanning markers in the GWAS, we first compared our efficient algorithms for variance component estimation to two existing methods, Wombat (Meyer, 2007) and MTG2 (Lee and van der Werf, 2016) (Table 1). In variance component estimation, the Wombat program uses a hybrid algorithm consisting of a few initial rounds of PX-EM (Liu *et al.*, 1998), followed by the AI algorithm, while MTG2 uses the pure AI algorithm with eigenvalue decomposition technique and moderates the magnitude of updates when the parameters go outside the legal domain of the parameter space. In general, the GMA methods converged faster with fewer iterations than the two methods. For the balanced longitudinal mouse data, our algorithm took only 2 s to complete the analysis while MTG2 took 5 s and Wombat took 40 min. Even for unbalanced longitudinal dairy cow data, the GMA method was substantially faster than Wombat.

We now compared results of the longitudinal GWAS obtained via the GMA-trans and uvLMM method. The two took about the same amount of time for the unbalanced data, but GMA-trans is much faster than uvLMM for the balanced data. Furthermore, the current GMA-trans algorithm for unbalanced data is several times faster than the GMA-fixed algorithm. We compared the *P*-values from GMA-fixed and GMA-trans and discovered that they are exactly the same (Supplementary Fig. S6A). For the balanced mouse data, GMA-fixed optimizes the variance components per SNP and is much slower than GMA-trans. However, the correlation coefficient of the *P*-values between the two methods is very high (Pearson's $r = 0.995$). The *P*-values of GMA-fixed are often smaller than the *P*-values of GMA-trans (Supplementary Fig. S6B), which means that GMA-fixed may detect more loci than GMA-trans. Taking into account the fast computational speed of GMA-trans and the high power of GMA-fixed (due to re-estimation of variance components; validated based on the above simulated datasets), we pre-selected SNPs based on a relaxed *P*-value criterion, say *P*-value < 0.01, from GMA-trans and then recalculated the *P*-values from GMA-fixed. As a result, the lost power by GMA-trans has been be rescued by GMA-fixed (Supplementary Fig. S6C), yet the reduced computational time remained at the same level (about 7 min) as the GMA-trans method.

All computations were performed on Intel Xeon E5 2.2 GHz CPU. We used the third order of Legendre polynomials for the mouse dataset and the fourth order for dairy cow dataset. The same convergence criterion was used for all methods in variance estimation, where the iteration stopped when the difference of the log-likelihood values between consecutive iterations is smaller than 0.001. The uvLMM method was implemented in the GEMMA (Zhou and Stephens, 2012) package. In variance component estimation, the Wombat program uses a hybrid algorithm consisting of a few initial rounds of PX-EM (Liu *et al.*, 1998), followed by the AI algorithm; MTG2 uses the pure AI algorithm and moderates the magnitude of updates when the parameters go outside the legal domain of the parameter space; GMA incorporates the EM algorithm into the AI matrix to build a weighted information matrix.

Table 1. Computational times of different methods for variance component estimation (including iteration number) and the subsequent step of GWAS

Category	Method	Computational time	
		Mouse data	Dairy cow data
Variance estimation	Wombat	40 min (11)	105 h (12)
	MTG2	5 s (15)	—
	GMA	2 s (9)	5.3 h (7)
GWAS	uvLMM	14.4 min	3.7 h
	GMA-fixed	5.1 h	16.5 h
	GMA-trans	1.7 min	3.8 h
	GMA-trans + GMA-fixed	7 min	—

For the unbalanced dairy cow data, both GMA-fixed and GMA-trans identified four significant SNPs (three at 1.65–1.81 Mb and one at about 4.36 Mb on chromosome 14) for milk yield without inflated false positives after multiple test correction using false discovery rate (FDR) with FDR < 5% (q-value < 0.05) (Supplementary Fig. S7). One of the SNPs (1 801 116 bp) is located within the *DGAT1* gene (1 795 351–1 804 562 bp) that is reported to be a major gene affecting milk production traits (Grisart *et al.*, 2004), and all significant SNPs are within the boundary of the reported QTL for milk yield (Hu *et al.*, 2015). We compared the additive effect curves of the four significant SNPs with milk yield trajectory in Supplementary Figure S8 and found very similar patterns between the curves, though the peak time of SNP effects (at about 200 days) is delayed compared to the peak time of the phenotypic trajectory (at about 80 days). The results indicate that *DGAT1* exhibits its main effects after the lactation peak and may contribute to the persistency of milk production (Strucken *et al.*, 2015).

For the balanced mouse data, GMA-fixed detected two candidate regions (112–128 Mb on chromosome 10 and 75–88 Mb on chromosome 13) (Fig. 2A and B) while GMA-trans only detected one of the two regions (119–125 Mb on chromosome 10) (Fig. 2C and D) after multiple testing correction with Benjamini–Hochberg (Benjamini and Hochberg, 1995) method at the q-value < 0.05. In this study, we also used the uvMLM-min method for comparison. The quantile–quantile (Q–Q) plot in Figure 2E shows that uvMLM-min appears to have higher type I errors than GMA, which is consistent with the simulation study. In order to control the type I error, we used the permutation test to determine the P-value threshold (genome-wide significance level of 0.05) for declaration of significance. This criterion led to the detection of one candidate region (118–125 Mb on chromosome 10) (Fig. 2F). Meanwhile, we compared the additive effect curves of the significant SNPs with the phenotypic trajectory (Fig. 3). The additive effect curves of significant SNPs on chromosome 10 have patterns similar to the phenotypic trajectory. The region has also been reported as a candidate QTL by Gray *et al.* (2015). However, the additive effect curves of the new candidate QTL on chromosome 13 are concave in shape and the QTL effect is inverse in the interim compared to the beginning and end (Fig. 3C).

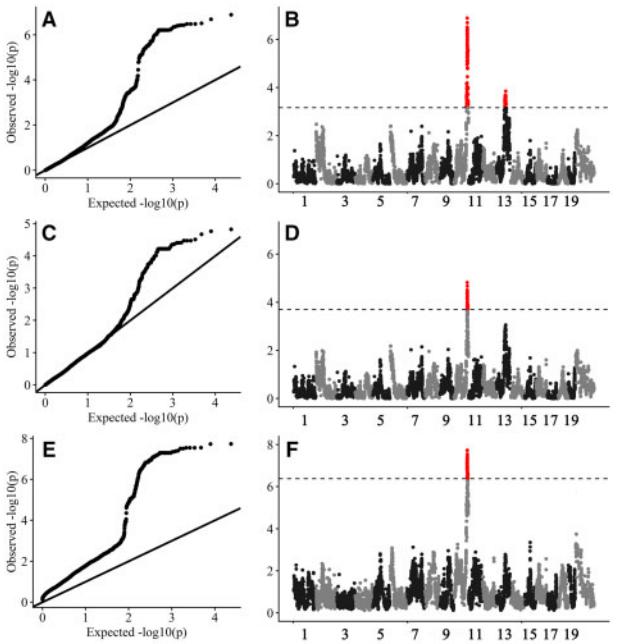

Fig. 2. Association studies of growth trajectory in the mouse population with the GMA-fixed method (panels at the top), the GMA-trans method (panels in the middle) and the uvMLM-min method (panels at the bottom)

4 Discussion

Longitudinal GWAS provides an appealing approach to probe the dynamic genetic mechanism of complex traits. However, successful application of the longitudinal GWAS is challenged by cryptic genetic relationship, dependency among the time course observations and time-consuming computation challenge. Here, we developed efficient analysis algorithms for longitudinal GWAS dealing with either balanced or unbalanced longitudinal data. Our algorithms are based on RRM, a mvMLM. The RRM includes a time-varied polygenic effect and a permanent environmental effect to explain the cryptic genetic relationship and dependency among observations. To improve the computational efficiency, we built a weighted information matrix from the EM algorithm and the AI information matrix, which guarantee the variance parameters to converge with fewer iterations. In the meantime, we proposed the fixed regression coefficient approach accompanied with eigenvalue decomposition strategy (GMA-fixed) and linear transformation of genomic estimation values (GMA-trans) algorithms. Simulations based on genotypes and phenotypes of actual populations show that our algorithms perform very well in terms of high statistical power and low false positive rate compared with the conventional uvLMM implemented GWAS. Application to the unbalanced dairy cow data and the balanced mouse data further validated the benefits of our longitudinal GMA.

There are various dynamic patterns of genetic controls represented by permanent QTLs, early QTLs, late QTLs and inverse QTLs (Wu and Lin, 2006). In this study, we used Legendre

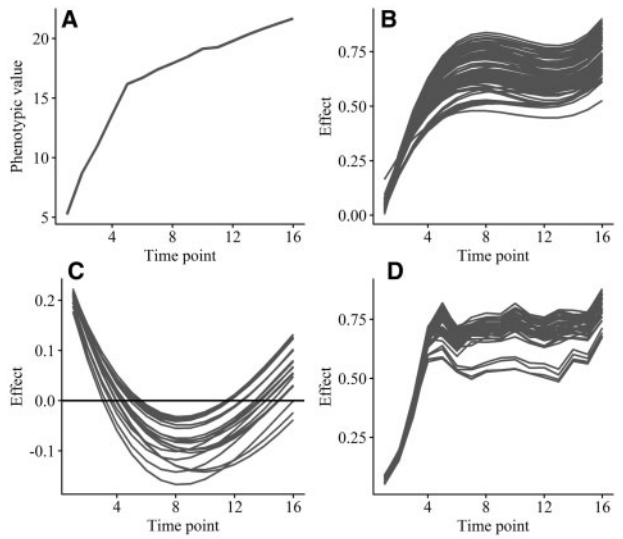

Fig. 3. The phenotypic and significant SNPs changing pattern for body weight in the mouse data. (**A**) The average phenotypic values plotted against age (from week 1 to week 16 incremented by 1). (**B**) The predicted growth trajectories of QTL effects for all significant SNPs between 112 and 128 Mb on chromosome 10 by the GMA-fixed method. (**C**) The predicted growth trajectories of QTL effects for all significant SNPs between 75 and 88 Mb on chromosome 13 by the GMA-fixed method. (**D**) The predicted growth trajectories of QTL effects for all significant SNPs between 118 and 125 Mb on chromosome 10 by uvMLM-min method

polynomials to model the dynamic changing process of QTL. This is a nonparametric approach because it makes no assumption about the shape of the curve. The method also reduces the correlations between the estimated random regression coefficients so that variance parameter estimation converges very rapidly. From the analyses of the two real data, we observed that the main QTLs tend to have similar changing patterns with the phenotypic curve, indicating that these QTLs determine the dynamic genetic mechanism of longitudinal traits. We also identified an inverse QTL (one genotype performs better than the other during early stage of growth, but the other genotype performs better during later stage of the growth) for the mouse data with GMA-fixed. These QTLs and others with minor effects can play a regulation role in shaping the final phenotypic trajectory.

For balanced data, GMA-fixed is more powerful than GMA-trans because it optimizes the variance parameters per SNP, but the latter is much faster. The GMA-trans step followed by the GMA-fixed step is recommended because it takes advantage of the high power of GMA-fixed and the high speed of GMA-trans. For unbalanced data, it is time consuming to optimize the variance components for each SNP. Since GMA-fixed and GMA-trans have similar power for unbalanced data, GMA-trans is recommended. In some cases, the contribution of genetic effects to the traits cannot be significantly greater than zeros, which can lead to overcorrection of polygenic effects. We provide an alternative based on the GMA-fixed method where the genetic effects are removed from the model.

In contrast to uvLMM with only two variance parameters (additive and residual variances), RRM has a complicated covariance structure with many variance parameters (depending on the orders of the Legendre polynomials). As a result, RRM may need more iterations to converge and, sometime, may encounter a convergence issue. If the iteration process stops early before convergence, the GMA algorithms may be subject to a higher Type I error. The orders of the Legendre polynomials can be determined by a model selection criteria, such as Akaike information criterion (AIC) (Akaike, 1974) and Bayesian information criterion (BIC) (Schwarz, 1978). To avoid any convergence issue, three or four orders of Legendre polynomials are recommended in practice. If the GMA algorithm encounters convergence issue even with low order of Legendre polynomials, the GMA-trans algorithm with an increased iteration number in variance parameter estimation step is recommended.

In our study, we focus on the traits changing over time. However, our developed GMA algorithm can be naturally applied to traits changing with other dynamic environmental covariates, such as solar radiation, solar radiation and temperature. Modern automatic information platforms can record abundant environmental data, while advanced genotyping technologies allow accessing to genomic information on a large scale. The GMA can utilize the two types of high dimensional information to tackle genome-wide genotypes and environments (G × E) interactions efficiently, which facilitates dissecting the complex genetic architecture of dynamic traits.

Author contributions

C.N. and J.F.L. conceived and designed the experiments. C.N. contributed analytic tools and analysed the data. D.W., L.Z., J.W., Y.L., H.K., S.Z, X.Z. and S.X. participated in the result interpretation and paper revision. C.N. and J.F.L. wrote the article with comments from X.Z. and S.X. All authors read and approved the final manuscript.

Acknowledgements

The project was supported by the National Natural Science Foundations of China (31661143013), Changjiang Scholars and Innovative Research Team in University (IRT_15R62) and Jinxinnong Animal Science Development Foundation. The authors are grateful to Jian Yang for his comments on an early version of the manuscript.

Conflict of Interest: none declared.

References

Akaike,H. (1974) A new look at the statistical model identification. *IEEE Trans. Automat. Control*, **19**, 716–723.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Series B Methodol.*, **57**, 289–300.

Churchill,G.A. and Doerge,R.W. (1994) Empirical threshold values for quantitative trait mapping. *Genetics*, **138**, 963–971.

Fahlgren,N. *et al.* (2015) Lights, camera, action: high-throughput plant phenotyping is ready for a close-up. *Curr. Opin. Plant Biol.*, **24**, 93–99.

Gilmour,A. *et al.* (2014) *ASReml User Guide. Release 4.1 Structural Specification*. VSN International, Hemel Hempstead, UK.

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

Gray,M.M. *et al.* (2015) Genetics of rapid and extreme size evolution in island mice. *Genetics*, **201**, 213–228.

Grisart,B. *et al.* (2004) Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. *Proc. Natl. Acad. Sci. USA*, **101**, 2398–2403.

Hu,Z.-L. *et al.* (2015) Developmental progress and current status of the animal QTLdb. *Nucleic Acids Res.*, **44**, D827-D833.

Jensen,J. (1997) Residual maximum likelihood estimation of (co) variance components in multivariate mixed linear models using average information. *J. Indian Soc. Agric. Statist.*, **49**, 215–236.

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

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

Kellogg,E.C. *et al*. (2014) Early predictors of autism in young children who are deaf or hard of hearing: three longitudinal case studies. *Semin. Speech Language*, **35**, 276–287.

Kwak,I.Y. *et al*. (2014) A simple regression-based method to map quantitative trait loci underlying function-valued phenotypes. *Genetics*, **197**, 1409–1416.

Lee,S.H. and van der Werf,J.H. (2016) MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information. *Bioinformatics*, **32**, 1420–1422.

Li,Z. and Sillanpää,M.J. (2015) Dynamic quantitative trait locus analysis of plant phenomic data. *Trends Plant Sci*., **20**, 822–833.

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

Liu,C. *et al*. (1998) Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika*, **85**, 755–770.

McSweeney,J. *et al*. (2014) Predicting coronary heart disease events in women. A longitudinal cohort study. *J. Cardiovasc. Nurs*., **29**, 482–492.

Meyer,K. (2007) WOMBAT: a tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). *J. Zhejiang Univ. Sci. B*, **8**, 815–821.

Mrode,R.A. (2014) *Linear Models for the Prediction of Animal Breeding Values*. 3rd edn. CABI Publishing, Wallinford, UK.

Ning,C. *et al*. (2017) Performance gains in genome-wide association studies for longitudinal traits via modeling time-varied effects. *Sci. Rep*., **7**, 590.

Ning,C. *et al*. (2018) Eigen decomposition expedites longitudinal genome-wide association studies for milk production traits in Chinese Holstein. *Genet. Select. Evol:*, **50**, 12.

Porto,S.M.C. *et al*. (2015) The automatic detection of dairy cow feeding and standing behaviours in free-stall barns by a computer vision-based system. *Biosyst. Eng*., **133**, 46–55.

Schaeffer,L.R. (2004) Application of random regression models in animal breeding. *Livest. Prod. Sci*., **86**, 35–45.

Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Statist*., **6**, 461–464.

Strucken,E.M. *et al*. (2015) Go with the flow-biology and genetics of the lactation cycle. *Front. Genet*., **6**, 118.

Wu,R. and Lin,M. (2006) Functional mapping—how to map and study the genetic architecture of dynamic complex traits. *Nat. Rev. Genet*., **7**, 229–237.

Xiong,H. *et al*. (2011) A flexible estimating equations approach for mapping function-valued traits. *Genetics*, **189**, 305–316.

Yang,J. *et al*. (2011) GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet*., **88**, 76–82.

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

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

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