Bioinformatics, 32(9), 2016, 1420–1422
doi: 10.1093/bioinformatics/btw012
Advance Access Publication Date: 10 January 2016
Applications Note

OXFORD

Genetics and population analysis

MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information

S. H. Lee[1,2,*] and J. H. J. van der Werf[1]

[1]School of Environmental and Rural Science, University of New England, Armidale, NSW 2351, Australia and [2]Queensland Brain Institute, The University of Queensland, Brisbane, QLD 4072, Australia

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on 17 September 2015; revised on 22 December 2015; accepted on 7 January 2016

Abstract

Summary: We have developed an algorithm for genetic analysis of complex traits using genome-wide SNPs in a linear mixed model framework. Compared to current standard REML software based on the mixed model equation, our method is substantially faster. The advantage is largest when there is only a single genetic covariance structure. The method is particularly useful for multivariate analysis, including multi-trait models and random regression models for studying reaction norms. We applied our proposed method to publicly available mice and human data and discuss the advantages and limitations.

Availability and implementation: MTG2 is available in https://sites.google.com/site/honglee0707/mtg2.

Contact: hong.lee@une.edu.au

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

1 Introduction

Previously, methods were developed to estimate genetic variance and genetic correlations between complex traits explained by genome-wide SNPs using linear mixed models (Lee *et al.*, 2012; Maier *et al.*, 2015; Yang *et al.*, 2011). As genetic relatedness among (conventionally) unrelated subjects could be estimated based on genomic information, which replaces family studies with population studies, the model allows estimation of the genetic effects to be much less confounded with family environmental effects. For this same reason, the approach has also been proposed as a more powerful tool to detect genotype–environment interaction (G × E) (Lee *et al.*, 2015). That is, in the presence of G × E, the genetic correlation between genetic effects in different environments is significantly lower than one (Falconer and Mackay, 1996). In order to capture G × E across a trajectory of multiple environments, random regression models have been proposed for evolutionary and livestock genetics (Kirkpatrick *et al.*, 1990; Meyer and Hill, 1997). The random regression model is also known as the reaction norm model (Kirkpatrick and Heckman, 1989).

In estimating genetic variance explained by genetic markers, Lee and Van der Werf (2006) introduced an efficient average information (AI) algorithm to obtain residual maximum likelihood (REML) estimates. As opposed to using Henderson's mixed model equation (MME) the algorithm was based on using the variance covariance matrix of phenotypic observations directly, hence the term 'direct AI algorithm'. The algorithm is particularly advantageous when using a dense covariance matrix, such as the genomic relationship matrix (GRM), and with a large number of multiple variance components. The direct AI algorithm has been implemented in GCTA-GREML (Lee *et al.*, 2012; Yang *et al.*, 2011, 2013) and MultiBLUP (Speed and Balding, 2014) that have been widely used in human, evolutionary and livestock genetics.

Here, we combine the direct AI algorithm with an eigen-decomposition of the GRM, as first proposed by Thompson and Shaw (1990). We apply the procedure to analysis of real data with univariate, multivariate and random regression linear mixed models with a single genetic covariance structure, and demonstrate that the computational

efficiency can increase by > 1000-fold compared with standard REML software based on MME.

2 Methods

2.1 Model

We used multivariate linear mixed models and random regression models to estimate genetic variances and covariances across multiple traits and among traits expressed in different environments. A linear mixed model can be written as

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b}_i + \mathbf{Z}_i\mathbf{g}_i + \mathbf{e}_i$$

where $\mathbf{y}_i$ is a vector of trait phenotypes, $\mathbf{b}_i$ is a vector of fixed effects, $\mathbf{g}_i$ is a vector of additive genetic value for individuals and $\mathbf{e}_i$ represents residuals for the trait or environment i. $\mathbf{X}$ and $\mathbf{Z}$ are incidence matrices. More details can be found in the Supplementary Notes. To model genotype-environment interactions, a random regression model attempts to fit effects as a function of a continuous variable (Kirkpatrick *et al.*, 1990; Meyer and Hill, 1997) as

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b}_i + \mathbf{Z}_i\mathbf{a}\mathbf{\Phi}'_i + \mathbf{e}_i$$

where $\mathbf{a}$ is a n (the number of records) by k matrix of genetic random regression coefficients, $\mathbf{\Phi}_i$ is the ith row in a p by k matrix of Legendre polynomials evaluated for p points on the trajectory, and k is the order of Legendre polynomials. This model is explicitly described in the Supplementary Notes. The genetic covariance structure was constructed based on genome-wide SNPs.

2.2 Algorithm

REML is often solved using the Newton–Raphson or Fisher's scoring method where variance components are updated based on observed (Hessian matrix) or expected second derivatives of the log likelihood (Fisher information matrix). In order to increase the computational efficiency of obtaining REML estimates, Gilmour *et al.* (1995) employed the average of the Hessian and Fisher information matrix that was estimated based on Henderson's MME. The MME-based AI algorithm is particularly efficient when the genetic covariance structure fit to the model is sparse. When using dense covariance structures such as GRM, the computational efficiency of the direct AI algorithm is substantially enhanced over the MME-based AI algorithm (Lee and Van der Werf, 2006). Here, we extend the direct AI algorithm by implementing an eigen-decomposition of the genetic covariance structure as proposed by Thompson and Shaw (1990).

In recent studies the eigen-decomposition technique has been made use of with the Newton–Raphson algorithm in univariate and multivariate linear mixed models (Zhou and Stephens, 2014). In the present work, we show that implementation in the direct AI algorithm is mathematically straightforward and is computationally more efficient, especially in multivariate linear mixed models (Supplementary Notes). Moreover, we demonstrate how our proposed algorithm can be efficiently applied to a random regression model (see Supplementary Notes).

2.3 Data

We used heterogeneous stock mice data (http://mus.well.ox.ac.uk/mo-use/HS/) to estimate genetic variances and covariances of complex traits explained by genome-wide SNPs. After a stringent QC of the genotypic data, we used 9258 autosomal SNPs from 1908 individuals. We used phenotypes of four glucose values (taken at 0, 15, 30 and 75 min after intraperitoneal glucose injection in a model of type 2 diabetes mellitus) as well as body mass index (BMI). We analyzed this data in a five-trait linear mixed model. We also applied a random regression model for the repeated glucose measures.

Second, we used human data from the Atherosclerosis Risk in Communities (ARIC) cohort (psh000280.v3.p1) (Sharrett, 1992). A similar stringent QC as above was applied to the available genotypes. In addition, we randomly removed one of each highly related pair of relatedness >0.05 to avoid bias because of population structure or family effects. After QC, 7263 individuals and 583 058 SNPs remained. We used BMI, triceps skinfold (TS), waist girth (WG), hip girth (HG), waist-to-hip ratio (WHR), systolic blood pressure (SP), diastolic blood pressure (DP) and hypertension (HP) that were fitted in an eight-trait linear mixed model.

Missing phenotypic values were less than 10% and 1% for each trait for the mice and the human data, respectively. They were imputed with their expected values from the univariate linear mixed model, each trait being fit separately.

2.4 Software

We implemented the direct AI algorithm and the eigen-decomposition technique with the MTG2 software. We compared MTG2 with GEMMA (Zhou and Stephens, 2014), ASReml (Gilmour *et al.*, 2006) and WOMBAT (Meyer, 2007). GEMMA uses the eigen-decomposition technique with the Newton-Raphson algorithm. ASReml and WOMBAT are well-known REML software that employed a MME-based AI algorithm.

3 Results

When using the heterogeneous mice data ($N = 1908$) for the multivariate linear mixed model with up to five traits, MTG2 only took a

Table 1. Computing time for each software run with a 2.7 GHz CPU when using the heterogeneous stock mice data ($N = 1908$)

	MTG2	GEMMA	ASReml	WOMBAT
# traits	Multivariate linear mixed model			
1	1 s	1 s	2 min	17 s
3	1 s	1 s	210 min	9 min
5	2 s	6 s	950 min	60 min
# order	Random regression model			
1	2 s	NA[a]	4 min	3 min
2	2 s	NA	82 min	30 min
3	2 s	NA	310 min	54 min

For MTG2 and GEMMA, it took ~4 s for the eigen-decomposition, which is only required to be done once per dataset then can then be reused for multiple analyses.

[a]GEMMA does not have a function for the random regression model.

few seconds, which was a few thousands times faster than ASReml and WOMBAT and few times faster than GEMMA (Table 1). Estimated SNP-heritability and genetic correlations between traits are shown in Supplementary Table S1. REML parameters after convergence were essentially the same between the different software suites, as shown in Supplementary Tables S8 and S9.

When employing a random regression model, the computing time for MTG2 was a few seconds, not changing with the higher-order models (Table 1). However, the computational efficiency of ASReml or WOMBAT was lower and the computing time increased substantially with the higher-order models (Table 1). GEMMA does not have a function for random regression models. The estimated results from the random regression model are described and depicted in Supplementary Data (Supplementary Table S2 and Figure S1).

When using the ARIC cohort human data (psh000280.v3.p1), the pattern of the computing time was similar to that for the heterogeneous mice in that MTG2 and GEMMA performed similarly although MTG2 became relatively faster when increasing the number of traits (Supplementary Table S4). ASReml and WOMBAT were too slow to run for this dataset. Supplementary Table S6 outlines the estimated SNP-heritability and genetic correlations between obesity and blood pressure traits.

4 Discussion

There are two main limitations to MTG2 as well as GEMMA. The eigen-decomposition technique cannot be used with more than one GRM as also noted by Zhou and Stephens (2014) unless a special condition is satisfied, i.e. one full-rank GRM and multiple low-rank GRMs are provided (Speed and Balding, 2014). In models with multiple GRMs, GEMMA cannot be used and MTG2 becomes slow although it is still considerably faster than ASReml and WOMBAT (Supplementary Table S5). Second, the eigen-decomposition technique requires a balanced design (i.e. no missing phenotypes across traits). Phenotypic imputation can be used for missing phenotypic values. In this work, we used imputed missing phenotypes for the mice data (<10% missing for each trait), although MTG2 without the eigen-decompostion could still be used for the data, including the missing values. We observed that the results from the data with and without the imputed missing phenotypes were not very different (Supplementary Table S2 and Figure S2). For the human data, missing phenotypes were less than 1%, therefore the results with and without the imputed missing phenotypes were almost identical (results not shown). Finally, both MTG2 and WOMBAT are able to facilitate a parallel computation that further raises efficiency.

5 Implication

There are three novel aspects in this application note. The first and foremost is estimating parameters for the random regression models with the direct AI algorithm. The second and third is to utilize the eigen-decomposition technique with the AI algorithm in the multivariate models and the random regression models, respectively. MTG2 can be used for a wider range of statistical models than GEMMA, including multivariate linear mixed models, random regression models and multiple variance components models. GEMMA can only be used for a single genetic variance component in multivariate linear mixed models (Supplementary Table S7). For random regression models or/and multiple variance components models, the computational efficiency for MTG2 (even without the eigen-decomposition) is considerably higher than that of ASReml or WOMBAT (Table 1, Supplementary Tables S5 and S7). Therefore, MTG2 can be a useful and efficient tool for complex traits analyses including estimating genetic variance and covariance and G × E.

Acknowledgements

This study makes use of publicly available data from Wellcome Trust Centre (http://mus.well.ox.ac.uk/mo-use/HS/) and from the database of Genotypes and Phenotypes (dbGaP) under accession psh000280.v3.p1 (see Supplementary Acknowledgements for the full statement).

Funding

This research is supported by the Australian National Health and Medical Research Council (APP1080157), the Australian Research Council (DE130100614 and DP160102126) and the Australian Sheep Industry Cooperative Research Centre.

Conflict of Interest: none declared.

References

Falconer,D.S. and Mackay,T.F.C. *Introduction to Quantitative Genetics*. Harlow, Essex, UK: Longman; 1996.

Gilmour,A.R. *et al*. *ASReml User Guide Release 2.0*. Hemel Hempstead, UK: VSN International; 2006.

Gilmour,A.R. *et al*. (1995) Average information REML: an efficient algorithm for variance parameters estimation in linear mixed models. *Biometrics*, **51**, 1440–1450.

Kirkpatrick,M. and Heckman,N. (1989) A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. *J. Math. Biol.*, **27**, 429–450.

Kirkpatrick,M. *et al*. (1990) Analysis of the inheritance, selection and evolution of growth trajectories. *Genetics*, **124**, 979–993.

Lee,S.H. *et al*. (2015) New data and an old puzzle: the negative association between schizophrenia and rheumatoid arthritis. *Int. J. Epidemiol*, **44**, 1706–1721.

Lee,S.H. and Van der Werf,J.H.J. (2006) An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree. *Genet. Sel. Evol.*, **38**, 25–43.

Lee,S.H. *et al*. (2012) Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*, **28**, 2540–2542.

Maier,R. *et al*. (2015) Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder and major depression disorder. *Am. J. Hum. Genet.*, **96**, 283–294.

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.

Meyer,K. and Hill,W. (1997) Estimation of genetic and phenotypic covariance functions for longitudinal or 'repeated' records by restricted maximum likelihood. *Livest. Prod. Sci.*, **47**, 185–200.

Sharrett,A.R. (1992) The Atherosclerosis Risk in Communities (ARIC) Study. Introduction and objectives of the hemostasis component. *Ann. Epidemiol.*, **2**, 467–469.

Speed,D. and Balding,D.J. (2014) MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.*, **24**, 1550–1557.

Thompson,E.A. and Shaw,R.G. (1990) Pedigree analysis for quantitative traits: variance components without matrix inversion. *Biometrics*, **46**, 399–413.

Yang,J. *et al*. Genome-Wide Complex Trait Analysis (GCTA): Methods, Data Analyses, and Interpretations. In: Gondro,C. *et al*. (eds.), *Genome-Wide Association Studies and Genomic Prediction*. New York, NY: Humana Press, 2013. Vol. **1019**. pp. 215–236.

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

Zhou,X. and Stephens,M. (2014) Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods*, **11**, 407–409.