Bioinformatics, 31(23), 2015, 3862–3864
doi: 10.1093/bioinformatics/btv448
Advance Access Publication Date: 4 August 2015
Applications Note

OXFORD

Genetics and population analysis

NAM: association studies in multiple populations

Alencar Xavier[1], Shizhong Xu[3], William M. Muir[2] and Katy Martin Rainey[1,*]

[1]Department of Agronomy and [2]Department of Animal Science, Purdue University, West Lafayette, IN 47907 and [3]Department of Plant Science, University of California, Riverside, CA 92521, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on May 4, 2015; revised on June 8, 2015; accepted on July 25, 2015

Abstract

Motivation: Mixed linear models provide important techniques for performing genome-wide association studies. However, current models have pitfalls associated with their strong assumptions. Here, we propose a new implementation designed to overcome some of these pitfalls using an empirical Bayes algorithm.

Results: Here we introduce NAM, an R package that allows user to take into account prior information regarding population stratification to relax the linkage phase assumption of current methods. It allows markers to be treated as a random effect to increase the resolution, and uses a sliding-window strategy to increase power and avoid double fitting markers into the model.

Availability and implementation: NAM is an R package available in the CRAN repository. It can be installed in R by typing install.packages ('NAM').

Contact: krainey@purdue.edu

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

1 Introduction

Since the advent of high-throughput genotyping technology, extensive efforts have focused on creating efficient mixed linear models (MLM) to address relatedness and computational issues in genome-wide association studies (Kang *et al.*, 2010; Zhou and Stephens, 2012). However, major pitfalls that still must be improved (Yang et al., 2014), including issues with resolution and detection power. Furthermore, MLM methods do not take into account the linkage phase associated with the multiple populations that comprise the association panel.

Association studies rely on persistent linkage disequilibrium (LD) between markers and quantitative trait loci (QTL). Such associations decay over time through recombination events, triggering LD that allows differentiation between populations (de Roos *et al.*, 2008). Therefore, association panels containing multiple populations are more likely to display diverging linkage phases, what makes QTL undetectable (Wientjes *et al.*, 2013).

Here we introduce NAM, a statistical package for association studies that aims to overcome some limitations of the mixed model framework and supports users to work with multiple populations when a stratification factor is known.

2 Structure and linkage phase

Structure, crypto-relatedness (Yu *et al.*, 2006) and unequal linkage phase across founders represent a major challenge for quantitative trait nucleotide (QTN) mapping (Lin *et al.*, 2003). Association methods deal with multiple levels of relatedness through genomic kinship, eigenvectors and model-based approaches (Kang *et al.*, 2010; Pritchard *et al.*, 2000; Zhang *et al.*, 2010) but are not able to handle linkage phase. Next-generation mapping populations such as NAM populations, it can address this issue by recoding the genotypic matrix to characterize haplotypes.

For example, in NAM populations alleles either come from the standard parent or from the founder. Thus, a given marker m can be represented as the number of alleles that come from each source: $m = [a_s, a_1, a_2, \ldots, a_f]$, where a_s represents the number of alleles inherited from the standard parent and a_1 to a_f represent alleles inherited from founder parents. The haplotype representation of genotypes works as follows: A given locus in an individual that belongs to family 2: if homozygous to the standard parent, it is coded as $m = [2,0,0,\ldots,f]$; if heterozygous, $m = [1,0,1,\ldots,f]$ and $m = [0,0,2,\ldots,f]$ if homozygous to the founder. Similar approaches can work for a random population if structural factors are known. This makes possible to relax assumptions regarding the linkage phase between the molecular marker and the QTN across populations, allowing different populations to pursue distinct coefficients for the marker under evaluation.

If the family term (stratification) is specified, the NAM package initiates the association study by recoding alleles and building the genomic relationship matrix (GRM). After solving the MLM through the EMMA algorithm (Kang *et al.*, 2008), NAM utilizes the P3D strategy (Zhang *et al.*, 2010) to avoid updating the polygenic term for every marker. Using the empirical Bayes approach, each molecular marker is treated as a random effect and the model is refitted using Eigen decomposition (Zhou and Stephens, 2012) and evaluated with the likelihood ratio test.

Datasets can still be analyzed by the empirical Bayes algorithm when no stratification factor is provided (Wang, 2015), applicable to multi-parent advanced generation inter-cross, random or bi-parental populations.

3 Major background effect

Most association algorithms attempt to control the diffuse background effect and are unable to control genes of major effect (Segura *et al.*, 2012) or use step-wise regression (Yu *et al.*, 2008). To address this issue, our package implements a sliding-window algorithm (Xu and Atchley, 1995). The approach consists of controlling the background by fitting a model with all markers outside a window, similar to whole-genome regression methods (Legarra *et al.*, 2015). The use of a sliding window prevents the double-fitting of the markers in the model, once the marker under evaluation is included in the GRM (Yang *et al.*, 2014). More details about the algorithm are available in the Supplementary file.

4 Methods comparison

To demonstrate the increase in power and resolution of the NAM package, we compared with three standard algorithms of MLM: the P3D/EMMAX algorithm with step-wise regression implemented in GAPIT (Lipka *et al.*, 2012), the GRAMMAR-Gamma algorithm implemented in GenABEL (Svishcheva *et al.*, 2012) and the GEMMA algorithm proposed and implemented by Zhou and Stephens (2012).

We used a simulated nested association panel with 840 individuals from six families, with 10 chromosomes of 100 cM and one marker by cM. A QTL was placed in the center of each chromosome (Fig. 1). The NAM package was able to capture most QTL with few false positives and little background noise, while other packages provided lower resolution QTL.

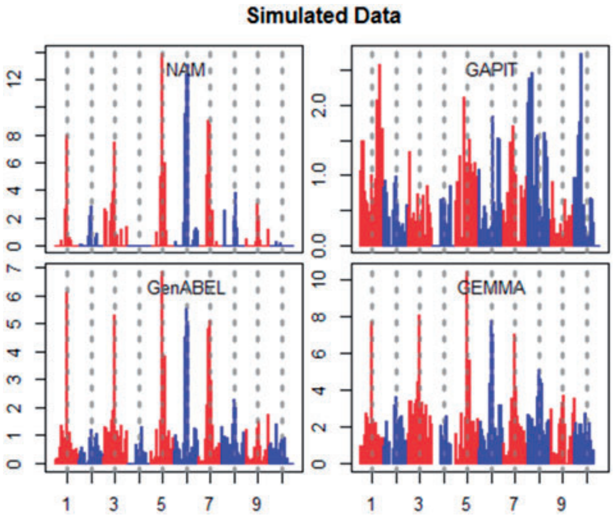

Fig. 1. Simulated data: genome-wide association mapping performed with four different implementations. Vertical lines represent the position of the QTL

5 Additional tools

The NAM package provides complimentary statistical tool, including the fixation indices (Weir and Cockerham, 1984), estimator of gene content (Forneris *et al.*, 2015), functions to deal with minor allele frequency and repeated markers, and the package performs imputation of missing loci through random forest (Stekhoven and Buhlmann, 2012). Best linear unbiased predictors (BLUP) are often used to replace raw phenotypes (Robinson, 1991) in association studies. Our package offers two algorithms to compute BLUP and variance components: REML (Kang *et al.*, 2008) and Bayesian Gibbs Sampling (Sorensen and Gianola, 2002). The latter allows users to perform Bayesian inferences.

6 Conclusions

The NAM package has implemented simple solutions to overcome pitfalls identified in association studies in mixed model frameworks, increasing the mapping power and resolution. The package includes an additional toolset for complimentary analysis of marker quality control, population stratification and to calculate BLUPs.

Acknowledgements

We acknowledge William Beavis for providing the simulated data, and Tiago Pimenta and Quishan Wang for reviewing algorithms and optimization of the source code.

Conflict of Interest: none declared.

References

de Roos,A.P.W. *et al.* (2008). Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. *Genetics*, **179**, 1503–1512.

Forneris,N.S. *et al.* (2015). Quality control of genotypes using heritability estimates of gene content at the marker. *Genetics*, **199**, 675–681.

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.

Legarra,A. *et al.* (2015). A comparison of methods for whole-genome QTL mapping using dense markers in four livestock species. *Gen. Sel. Evol.*, **47**, 6.

Lin,M. *et al.* (2003). A general statistical framework for mapping quantitative trait loci in nonmodel systems: issue for characterizing linkage phases. *Genetics*, **165**, 901–913.

Lipka,A.E. *et al.* (2012). GAPIT: genome association and prediction integrated tool. *Bioinformatics*, **28**, 2397–2399.

Pritchard,J.K. *et al.* (2000). Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Robinson,G.K. (1991). That BLUP is a good thing: the estimation of random effects. *Stat. Sci.*, **6**, 15–32.

Segura,V. *et al.* (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.*, **44**, 825–830.

Sorensen,D. and Gianola,D. (2002). *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. Springer Science & Business Media, New York.

Stekhoven,D.J. and Buhlmann,P. (2012). MissForest: non-parametric missing value imputation for mixed-TPE data. *Bioinformatics*, **28**, 112–118.

Svishcheva,G.R. *et al.* (2012). Rapid variance components-based method for whole-genome association analysis. *Nat. Genet.*, **44**, 1166–1170.

Wang,Q. (2015). An empirical Bayes method for genome-wide association studies. *In Plant and Animal Genome XXXII*. W799/Statistical Genomics.

Weir,B.S. and Cockerham,C.C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

Wientjes,Y.C. *et al.* (2013). The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics*, **193**, 621–631.

Xu,S. and Atchley,W.R. (1995). A random model approach to interval mapping of quantitative trait loci. *Genetics*, **141**, 1189–1197.

Yang,J. *et al.* (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.*, **46**, 100–106.

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.

Yu,J. *et al.* (2008). Genetic design and statistical power of nested association mapping in maize. *Genetics*, **178**, 539–551.

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.