Genetics and Population Analysis

NAM: Association Studies in Multiple Populations

Alencar Xavier\textsuperscript{1,*}, Shizhong Xu\textsuperscript{3}, William M. Muir\textsuperscript{2}, and Katy Martin Rainey\textsuperscript{1}

\textsuperscript{1}Department of Agronomy, Purdue University, 915 W. State St., Lilly Hall, West Lafayette, IN, 47907, USA.
\textsuperscript{2}Department of Animal Science, Purdue University, 915 W. State St., Lilly Hall, West Lafayette, IN, 47907, USA.
\textsuperscript{3}Department of Plant Science, University of California, 3134 Batchelor Hall, Riverside, CA, 92521, USA.

ABSTRACT

\textbf{Motivation:} Mixed linear models (MLM) provide important techniques for performing genome-wide associations studies (GWAS). 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.

\textbf{Results:} NAM is 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.

\textbf{Availability:} NAM is an R package available in the CRAN repository. It can be installed in R by typing \texttt{install.packages(“NAM”)}.

\textbf{Contact:} xaviera@purdue.edu

\textbf{Supplementary Information:} Supplementary information about the method and algorithms is 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 (GWAS) (Zhou and Stephens 2012, Kang et al. 2010). 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 (Pritchard et al. 2000, Kang et al. 2010, Zhang et al. 2010) but are not able to handle linkage phase. Next-generation mapping populations such as nested association mapping (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: \(m=\left[ a_1, a_1, a_2, \ldots, a_n \right] \), where \(a\) represents the number of alleles inherited from the standard parent and \(a_1\) to \(a_n\) represent alleles inherited from founder parents. The haplotype representation of genotypes works as follows. A given locus in an individual belongs to family 2: if homozygous to the standard parent, it is coded as \(m=\left[ 2,0,0,\ldots,f \right] \); if heterozygous, \(m=\left[ 1,0,1,\ldots,f \right] \); and \(m=\left[ 0,0,2,\ldots,f \right] \) 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 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 \texttt{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 (LRT).

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 (MAGIC), 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 (Se-
gura 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 to three standard algorithms of mixed linear models: 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 (Figure 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.

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.

REFERENCES