Bioinformatics Advance Access published July 13, 2012

GAPIT: Genome Association and Prediction Integrated Tool

Alexander E. Lipka[1], Feng Tian[2], Qishan Wang[3], Jason Peiffer[4], Meng Li[5], Peter J. Bradbury[1], Michael A. Gore[6], Edward S. Buckler[1,2,4], and Zhiwu Zhang[2*]

[1]Robert W. Holley Center, United States Department of Agriculture-Agricultural Research Service, Ithaca, New York, 14853, USA; [2]Institute for Genomic Diversity, Cornell University, Ithaca, New York, 14853 USA; [3]Department of Animal Science, Shanghai Jiao Tong University, Shanghai 200240, China; [4]Department of Plant Breeding and Genetics, Cornell University, Ithaca, New York, 14853, USA; [5]College of Agriculture, Nanjing Agricultural University, Nanjing 210095, China; [6]US Arid-Land Agricultural Research Center, United States Department of Agriculture-Agricultural Research Service, Maricopa, Arizona, 85138 USA;

Associate Editor: Dr. Jeffrey Barrett

ABSTRACT

Summary: Software programs that conduct genome-wide association studies and genomic prediction and selection need to use methodologies that maximize statistical power, provide high prediction accuracy, and run in a computationally efficient manner. We developed an R package called Genomic Association and Prediction Integrated Tool (GAPIT) that implements advanced statistical methods including the compressed mixed linear model (CMLM) and CMLM-based genomic prediction and selection. The GAPIT package can handle large data sets in excess of ten thousand individuals and one million single nucleotide polymorphisms with minimal computational time, while providing user-friendly access and concise tables and graphs to interpret results.

Availability: http://www.maizegenetics.net/GAPIT.

Contact: zhiwu.zhang@cornell.edu

Supplementary information: Available at *Bioinformatics* online.

1 INTRODUCTION

Advances in high-throughput single-nucleotide polymorphism (SNP) genotyping are enabling powerful genome-wide association studies (GWAS), thereby enhancing the ability to identify causal mutations that underlie human diseases and agriculturally important traits. The resulting SNPs are also valuable for genomic prediction and selection (GS), which provides criteria for disease risk management in humans and expedited selection in animal and plant breeding (Heffner, et al., 2009; Meuwissen, et al., 2001). Before the full potential of GWAS and GS are realized, inflated false positive rates, extensive computational requirements, and suboptimal prediction accuracies need to be addressed.

Newly developed GWAS statistical methods based on the mixed linear model (MLM) hold great promise to overcome these challenges. They are flexible because they incorporate fixed and random effects. To address the spurious associations that arise from population structure, covariates from either STRUCTURE (Pritchard, et al., 2000) or principal components (PCs) can be included as fixed effects. The cryptic relationships between individuals are accounted for through a kinship matrix in the unified MLM (Yu, et al., 2006). The more computationally efficient and powerful compressed MLM (CMLM) (Zhang, et al., 2010) uses a group kinship matrix calculated from clustered individuals.

Because the typical number of genotypic data points is exceeding hundreds of millions, solving MLMs using the traditional restricted maximum likelihood (REML) approach is computationally intensive. Therefore, the efficient mixed model association (EMMA) algorithm (Kang, et al., 2008) was developed to reduce this computational burden by reparameterizing the MLM likelihood functions. EMMA eXpedited (EMMAX) (Kang, et al., 2010) and population parameters previously determined (P3D) (Zhang, et al., 2010) were independently developed to further reduce computing time by eliminating the need to re-estimate variance components at each marker.

Most GS methods make predictions with the sum of the effects from all available SNPs or gBLUP (Genomic Best Linear Unbiased Prediction) based on a kinship matrix derived from these SNPs. The former approach offers higher prediction accuracies for simpler traits, while the latter approach is more accurate for complex traits (Daetwyler, et al., 2010). Our work implements an improved gBLUP method that increases accuracy, especially for simple traits.

Most software packages were developed for a particular GWAS or GS approach. For example, packages were written exclusively for the EMMA and EMMAX algorithms, respectively. Other software such as the Trait Analysis by aSSociation, Evolution, and Linkage (TASSEL) (Bradbury, et al., 2007) and PLINK (Purcell, et al., 2007) make multiple GWAS approaches available in one package. We continue these software development efforts by creating Genome Association and Prediction Integrated Tool (GAPIT), which integrates the most powerful, accurate, and computationally efficient GWAS and GS methods into a single R package.

2 IMPLEMENTATION

The GAPIT program accepts several combinations of genotypic data, phenotypic data, externally obtained kinship matrices, and covariates such as population structure and age. Multiple traits can be stored in a single phenotypic data set, which allows sequential analysis of each trait. The genotypic data may be stored in HapMap or numerical formats. If genotypic data are absent, then phenotypic data and a kinship matrix are required to perform GS.

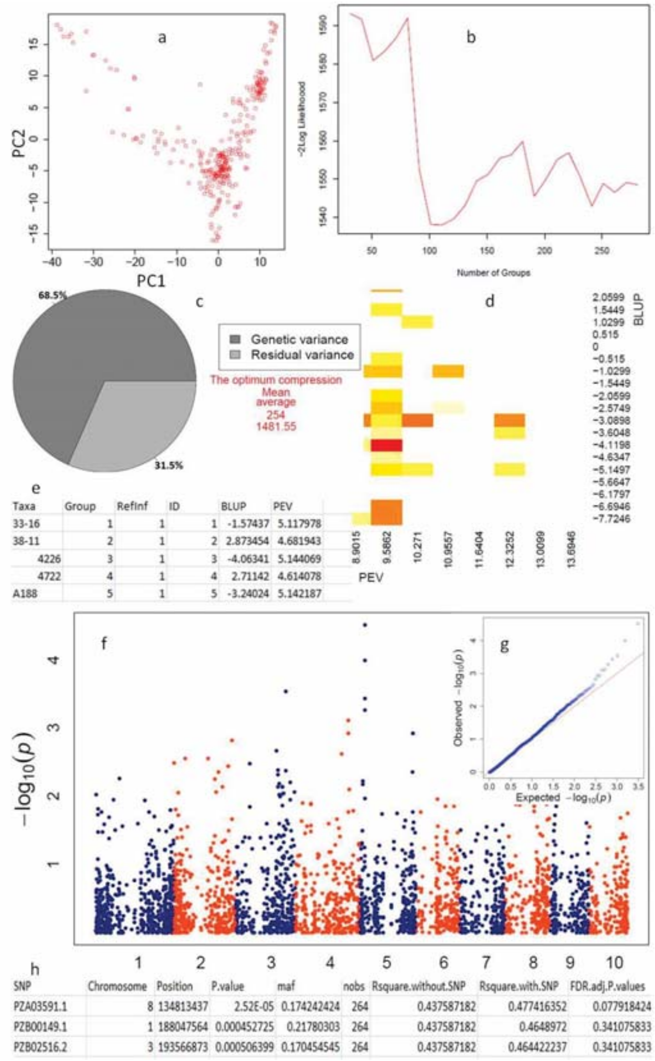
By default, GAPIT uses the CMLM approach with P3D/EMMAX for GWAS. GS is performed using the same optimization settings as GWAS (Supplementary Sections I&II and Figure S1). There is an option to perform GS only by specifying “SNP.test=FALSE”. Seven algorithms are available to cluster individuals into groups. GAPIT can also perform the MLM and GLM approaches by adjusting the “group.to” and “group.from” input parameters. When the kinship matrix is not provided, it will be calculated with the methods of VanRaden (VanRaden, 2008), Loiselle (Loiselle, et al., 1995), or EMMA (Kang, et al., 2008). GAPIT can also perform principal component analysis (PCA) of the genotypic data to control for population structure (Zhao, et al., 2007).

GAPIT has several strategies for analyzing large SNP data sets. One is to import genotypic data stored in multiple smaller files. If

[*]To whom correspondence should be addressed.

these files still exceed memory limits, the “file.fragment” parameter can be used to sequentially load fragments within each file. If there is not enough memory to use all SNPs to calculate the kinship matrix and PCs, then the “SNP.fraction” input parameter will select a random sample of the SNPs for these calculations (Yu, et al., 2009).

Results from GAPIT are accessed as both objects within the R workspace and as external files. The R objects, which include GWAS and GS results, may be used for follow-up analyses in R. The external files include publication-ready summaries of GWAS and GS results. GWAS results are summarized by Manhattan plots, Quantile-Quantile plots, and a table. Similarly, GS results are presented in a heat map and a table. Graphs of the heritability estimates and the likelihood function at various compression levels are included. A subset of the graphs and tables produced by GAPIT are presented in Figure 1.



Taxa	Group	RefInf	ID	BLUP	PEV
33-16	1	1	1	-1.57437	5.117978
38-11	2	1	2	2.873454	4.681943
4226	3	1	3	-4.06341	5.144069
4722	4	1	4	2.71142	4.614078
A188	5	1	5	-3.24024	5.142187

SNP	Chromosome	Position	P.value	maf	nobs	Rsquare.without.SNP	Rsquare.with.SNP	FDR.adj.P.values
PZA03591.1	8	134813437	2.52E-05	0.174242424	264	0.437587182	0.477416352	0.077918424
PZB00149.1	1	188047564	0.000452725	0.21780303	264	0.437587182	0.4648972	0.341075833
PZB02516.2	3	193566873	0.000506399	0.170454545	264	0.437587182	0.464422237	0.341075833

Figure 1. Gallery of GAPIT output. (a) Plot of the first two principal components (PC1 and PC2). (b) Plot of twice the negative log likelihood (-2LL, smaller is better) at various number of groups. (c) Graph showing the optimum cluster algorithm, method to calculate group kinship, group number, -2LL, and the proportion of genetic variance (group heritability) and residual variance. (d) Distribution of best linear unbiased predictors (BLUPs) and their prediction error variance (PEV) (e) Genomic prediction and selection output summary. The individual id (taxa), group, RefInf which indicates whether the individual is in the reference group (1) or not (2), the group ID number, the BLUP, and the PEV of the BLUP. (f) Manhattan plot.–log *P*-values are plotted against physical map position of SNPs. Chromosomes are alternatingly colored. (g) Quantile-quantile (QQ)-plot determines how GWAS results compare to the expected results under the null hypothesis of no association. (h) Output table of GWAS results. The SNP id, chromosome, bp position, *P*-value, minor allele frequency (maf), sample size (nobs), R^2 of the model without the SNP, R^2 of the model with the SNP, and adjusted *P*-value following a false discovery rate (FDR)-controlling procedure (Benjamini and Hochberg, 1995).

3 PERFORMANCE TESTS

EMMA and TASSEL were compared with GAPIT. These two packages were selected because both use the EMMA algorithm, while TASSEL also implements the CMLM approach with P3D. When the same approach was used, identical results were obtained (Figures S2 and S3). The computing time of all three packages increases linearly with the number of SNPs (Figure S4). However, the average computing time per SNP in GAPIT is 7-fold and 180-fold faster than TASSEL and EMMA, respectively (Figure S4). It took 69.5 hours to analyze a data set with 11,000 individuals and 500,000 SNPs, which extrapolates to 7,195 SNPs/CPU hours or less than 6 days to analyze one million SNPs.

4 CONCLUSIONS

This R package uses state-of-the-art mixed model methods to conduct GWAS and GS. GAPIT analyzes large data sets with minimum computational time and produces comprehensive results including R objects and high quality graphs.

ACKNOWLEDGEMENTS

J.C. Glaubitz is acknowledged for assistance with data analysis in UNIX.

Funding: Supported by NSF DBI-0321467, DBI-0820619, and DBI-0922493 and USDA-ARS.

REFERENCES

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to mutliple testing, *J.R. Statis. Soc. B.*, **57**, 289-300.

Bradbury, P.J., *et al.* (2007) TASSEL: software for association mapping of complex traits in diverse samples, *Bioinformatics*, **23**, 2633-2635.

Daetwyler, H.D., *et al.* (2010) The impact of genetic architecture on genome-wide evaluation methods, *Genetics*, **185**, 1021-1031.

Heffner, E.L., Sorrells, M.E. and Jannink, J.L. (2009) Genomic Selection for Crop Improvement, *Crop Science*, **49**, 1-12.

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

Kang, H.M., *et al.* (2008) Efficient Control of Population Structure in Model Organism Association Mapping, *Genetics*, **178**, 1709-1723.

Loiselle, B.A., *et al.* (1995) Spatial Genetic-Structure of a Tropical Understory Shrub, Psychotria Officinalis (Rubiaceae), *Am J Bot*, **82**, 1420-1425.

Meuwissen, T.H., Hayes, B.J. and Goddard, M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps, *Genetics*, **157**, 1819-1829.

Pritchard, J.K., *et al.* (2000) Association mapping in structured populations, *Am J Hum Genet*, **67**, 170-181.

Purcell, S., *et al.* (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses, *Am J Hum Genet*, **81**, 559-575. .

VanRaden, P.M. (2008) Efficient methods to compute genomic predictions, *J Dairy Sci*, **91**, 4414-4423.

Yu, J., *et al.* (2009) Simulation Appraisal of the Adequacy of Number of Background Markers for Relationship Estimation in Association Mapping, *Plant Genome*, **2**, 63-77

Yu, J.M., *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.

Zhao, K., *et al.* (2007) An Arabidopsis example of association mapping in structured samples, *PLoS Genet*, **3**, e4.