Published July 7, 2016

ORIGINAL RESEARCH

GAPIT Version 2: An Enhanced Integrated Tool for Genomic Association and Prediction

You Tang†, Xiaolei Liu†, Jiabo Wang, Meng Li, Qishan Wang, Feng Tian, Zhongbin Su, Yuchun Pan, Di Liu, Alexander E. Lipka, Edward S. Buckler, and Zhiwu Zhang*

Abstract

Most human diseases and agriculturally important traits are complex. Dissecting their genetic architecture requires continued development of innovative and powerful statistical methods. Corresponding advances in computing tools are critical to efficiently use these statistical innovations and to enhance and accelerate biomedical and agricultural research and applications. The genome association and prediction integrated tool (GAPIT) was first released in 2012 and became widely used for genome-wide association studies (GWAS) and genomic prediction. The GAPIT implemented computationally efficient statistical methods, including the compressed mixed linear model (CMLM) and genomic prediction by using genomic best linear unbiased prediction (gBLUP). New state-of-the-art statistical methods have now been implemented in a new, enhanced version of GAPIT. These methods include factored spectrally transformed linear mixed models (FaST-LMM), enriched CMLM (ECMLM), FaST-LMM-Select, and settlement of mixed linear models under progressively exclusive relationship (SUPER). The genomic prediction methods implemented in this new release of the GAPIT include gBLUP based on CMLM, ECMLM, and SUPER. Additionally, the GAPIT was updated to improve its existing output display features and to add new data display and evaluation functions, including new graphing options and capabilities, phenotype simulation, power analysis, and cross-validation. These enhancements make the GAPIT a valuable resource for determining appropriate experimental designs and performing GWAS and genomic prediction. The enhanced R-based GAPIT software package uses state-of-the-art methods to conduct GWAS and genomic prediction. The GAPIT also provides new functions for developing experimental designs and creating publication-ready tabular summaries and graphs to improve the efficiency and application of genomic research.

Published in Plant Genome
Volume 9. doi: 10.3835/plantgenome2015.11.0120

Core Ideas

- Genome-wide association study
- Genomic prediction
- Simulation and experimental design

Background

THE INCREASING volume of genomic data during the last 10 yr has out-paced Moore's Law, which describes the exponential growth of computer chip development. Biomedical and agricultural researchers have

Y. Tang, Z. Su, College of Electrical and Information, Northeast Agricultural Univ., Harbin, China; X. Liu, Key Laboratory of Agricultural Animal Genetics, Breeding and Reproduction, Ministry of Education & College of Animal Science and Technology, Huazhong Agricultural Univ., Wuhan, China; X. Liu, E. S. Buckler, Institute for Genomic Diversity, Cornell Univ., Ithaca, New York 48824; J. Wang and Z. Zhang, Dep. of Animal Science and Technology, North East Agricultural Univ., Harbin, China; J. Wang and D. Liu, Institute of Animal Husbandry, Heilongjiang Academy of Agricultural Science, Harbin, China; M. Li, College of Horticulture, Nanjing Agricultural Univ., Nanjing 210095, China; Q. Wang and Y. Pan, School of Agriculture and Biology, Shanghai Jiaotong Univ., Shanghai, China; F. Tian, National Maize Improvement Center of China, China Agricultural Univ., Beijing, China; A. E. Lipka, Dep. of Crop Sciences, Univ. of Illinois, Urbana, IL 61801; E. S. Buckler, USDA- ARS, Ithaca, NY 48824; Z. Zhang, Department of Crop and Soil Sciences, Washington State University, Pullman, WA 99164. † Y. Tang and X. Liu contributed equally to this work. Received 30 Nov. 2015. Accepted 7 Jan. 2016. *Corresponding author (Zhiwu.Zhang@WSU.edu).

Abbreviations: CMLM, compressed mixed linear model; ECMLM, enriched compressed mixed linear model; EMMA, efficient mixed-model association; FaST-LMM, factored spectrally transformed linear mixed models; FDR, false discovery rate; GAPIT, genome association and prediction integrated tool; gBLUP, genomic best linear unbiased prediction; GWAS, genome-wide association studies; LD, linkage disequilibrium; MAF, minor allele frequency; MLM, mixed linear model; PC, principal components; QTN, quantitative trait nucleotide; SNP, single-nucleotide polymorphism; SUPER, settlement of mixed linear models under progressively exclusive relationship

the unprecedented capacity to generate extremely large amounts of data (Shendure et al., 2004; Ober et al., 2012; Georges, 2014). These large datasets require increasingly greater computing capabilities for performing analyses such as GWAS (Cantor et al., 2010) and genomic prediction of phenotypes from genotypes (VanRaden et al., 2009; Endelman, 2011; Poland et al., 2012). Genome-wide association studies has proven to be one of the most efficient methods—relative to time, cost, and precision—for identifying candidate genes that control human diseases and agriculturally important traits. However, besides insufficient computing speed, low statistical power and false positives are also factors that influence GWAS performance and reliability (Atwell et al., 2010; Yang et al., 2014).

A typical GWAS can have an inflated false positive rate if the statistical model used includes only a tested genetic marker, such as a single-nucleotide polymorphism (SNP), as an explanatory variable. Indeed, associations between a genetic marker and a phenotype occur for many reasons, in addition to the genetic linkage between the tested genetic markers and functional causal polymorphisms. For example, population structure and relatedness among individuals are two common sources of false positives (Falush et al., 2007). Consequently, population structure and individuals' total genetic effects are often fitted as covariates in a mixed linear model (MLM) to reduce the false discovery rate (FDR) (Yu et al., 2006). Unfortunately, this reduction of false positives can also increase false negatives through confounding phenotypes with population structure and individuals' total genetic effects (Atwell et al., 2010). Therefore, new analysis methods with greater statistical power are critical for resolving these confounding issues and improving interpretive reliability (Yang et al., 2014).

Although several methods have been developed to improve the computing speed of MLMs, including efficient mixed-model association (EMMA) (Kang et al., 2008), EMMA expedited and population parameter previously determined (Kang et al., 2010; Zhang et al., 2010), genome-wide EMMA (Zhou and Stephens, 2012), FaST-LMM (Lippert et al., 2011), and GenABEL (Svishcheva et al., 2012), methods to improve statistical power were limited before 2012. Among them is the CMLM. The CMLM replaces the individuals' genetic effects with those of the group to which each individual belongs (Zhang et al., 2010). That is, individuals are clustered into groups on the basis of their relationships derived from all the available genetic markers. Simulations demonstrate that CMLM improves statistical power by 5 to 15% compared to regular MLM (Zhang et al., 2010). Additional benefits of CMLM include a dramatic reduction in computing time. The CMLM approach was implemented in the first release of the GAPIT in 2012 (Lipka et al., 2012).

Since 2012, several new powerful statistical approaches have been developed to improve statistical power for GWAS, including ECMLM (Li et al., 2014), FaST-LMM-Select (Listgarten et al., 2012), and SUPER (Wang et al., 2014). Next, we describe the implementation of these methods and new functions in the new release of the GAPIT.

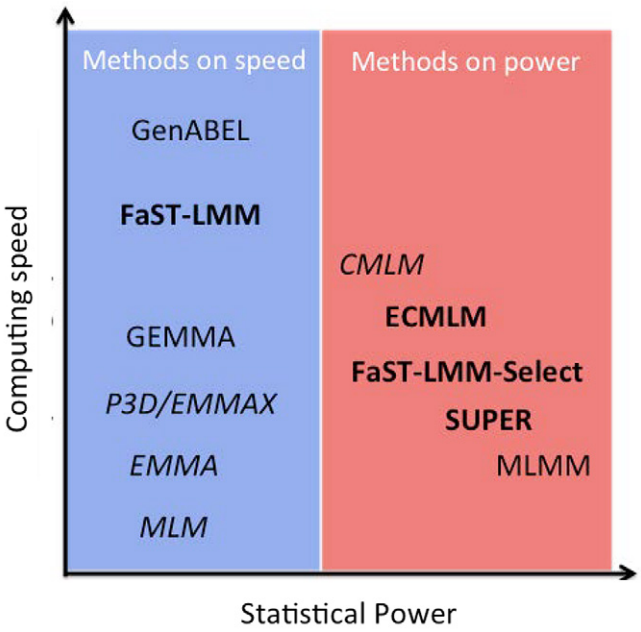

Fig. 1. Mixed linear model (MLM) methods and implementations in the genome association and prediction integrated tool (GAPIT). Since the introduction of MLM for genome-wide association studies (GWAS) in 2006, multiple methods have been developed to improve computing speed (blue area) and statistical power (red area). The methods implemented in GAPIT at first release are in italic text, including MLM, efficient mixed-model association (EMMA), population parameter previously determined (P3D), EMMA eXpedited, and compressed MLM (CMLM). Additional state-of-the-art methods (bold text) have been implemented in GAPIT's current release, including factored spectrally transformed linear mixed models (FaST-LMM), FaST-LMM-Select, enriched CMLM (ECMLM), and settlement of mixed linear models under progressively exclusive relationship (SUPER).

Implementation

The GAPIT implemented several new methods that improve statistical power and computing speed for data analyses (Fig. 1). Furthermore, the GAPIT includes an enhanced set of data interpretation and evaluation functions. These functions can help users interpret results or perform pre-experimental statistical power analyses by evaluating existing genotypes for phenotypes with specific genetic architecture similar to the traits of interest. These functions can also aid in experimental design by evaluating genotype data from similar populations.

Enriched Compressed Mixed Linear Models

Compressed mixed linear models cluster individuals into groups based on kinship among individuals so that the genetic effects of individuals in the regular MLM are replaced by the genetic effects of the corresponding groups. Consequently, the kinship matrix among individuals in the regular MLM is replaced by the kinship among groups. Compressed mixed linear models calculate group kinship as the average kinship among individuals. However, group kinship can also be derived with

alternative methods, including the maximum, median, or minimum kinship among corresponding individuals. Enriched CMLM provides users the ability to choose from multiple algorithms to cluster individuals into groups (e.g., the average and Ward methods) and multiple measures to derive group kinship from individual kinship (e.g., the average and minimum methods). Thus the implementation of ECMLM in GAPIT offers users the advantage of optimizing combinations of methods to improve statistical power with a negligible increase in computing time. Currently, GAPIT includes seven cluster algorithms ("ward", "single", "complete", "average", "mcquitty", "median", and "centroid")(Li et al., 2014) and four group kinship derivation methods (average, minimum, maximum, and median).

The FaST-LMM Algorithm

The computing time complexity of MLM is the cubic power of the sample size. To overcome this computational obstacle in the new version of the GAPIT, we implemented a rank-reduced kinship algorithm, called FaST-LMM (Lippert et al., 2011), which is a revolutionary improvement in computing speed for MLMs. The rank-reduced kinship depends on a subset of genetic markers that is less than the number of individuals. The subset of markers is used directly to define the relationship among individuals without building the kinship matrix first for optimizing the likelihood on genetic variance and the residual variance ratio. The subset of markers can be selected at random or with the methods described in the next paragraph. With FaST-LMM, computing time is linear to the size of the subset of markers and independent of sample size. Because FaST-LMM was implemented as a function, this function can be directly used by other R packages for further development.

The FaST-LMM-Select and SUPER Algorithms

Focusing on two objectives, algorithms from the FaST-LMM-Select and SUPER methods were selectively implemented in the enhanced version of the GAPIT. One objective was to avoid the potential reduction in statistical power that can occur when randomly selecting the subset of markers to define kinship among individuals. The other objective was to eliminate the confounding problem between kinship and testing markers (Listgarten et al., 2012). Whether kinship is derived from randomly selected markers or all markers, kinship has confounding problems with testing markers that leads to reduced statistical power. Thus we implemented the marker selection algorithm and the exclusion algorithm from FaST-LMM-Select and SUPER.

First, we replaced the associated marker approach from FaST-LMM-Select with the bin approach from SUPER to further reduce the number of markers. After the primary association test, especially with simple and fast methods (e.g., the General Linear Model), the genome is divided into small bins. Each bin is represented by the most significant marker. The size of bins and the number of bins selected are optimized iteratively via the maximum likelihood method. The converged SNPs, resulting from the iterations are named pseudo-quantitative trait nucleotides (QTNs) for descriptive convenience.

Second, the exclusion algorithm from FaST-LMM-Select and SUPER was implemented to derive the exclusive kinship for tested markers. Only the pseudo-QTNs that are not in linkage disequilibrium (LD) with the marker are used to define the relationship among individuals. Consequently, the kinship is trait-specific and complementary to the testing SNPs. The SUPER method boosts statistical power compared to the regular MLM, which derives kinship from all SNPs or a randomly selected subset.

The GAPIT provides multiple options for the primary association tests, including the General Linear Model, MLM, CMLM, ECMLM, and FaST-LMM. Although each option will produce a slightly different initial set of *P*-values, the final results are similar after the process runs through a couple of iterations, regardless of the options of the primary association tests.

Analyses of Power, Type I Error and FDR

From a GWAS perspective, statistical power is the probability of finding a gene of interest or a genetic marker that is physically and closely linked to the real gene. Ideally, a power analysis should be conducted before an experiment starts; unfortunately, this rarely happens for many reasons. One reason is that the analysis is complicated and imprecise and researchers are often forced to use a trial-and-error approach.

To fill this need for a practical and accurate pre-experimental power analysis, we implemented a function in the GAPIT. Before beginning an experiment, researchers can examine a population that is similar to their own. The GAPIT provides a function to simulate phenotypes from genotypes of the population. Researchers can also define the genetic architecture of the phenotype that is similar to the traits of interest according to heritability, the number of QTNs, and the roles of major genes.

The roles of major genes are defined by the distribution of QTN effects. Two distributions were implemented: a standard normal distribution and an approximated geometry distribution. When the geometry distribution is selected, the effect of the i^{th} QTN is assigned to a^i, where a is the effect of the first QTN, with a range from 0 to 1. When a is close to 1, the effects of all QTNs are nearly the same. When a is close to 0, variation in the QTN effect becomes larger. The first QTN has the most advantage over other QTNs (Yu et al., 2006).

Based on simulated phenotypes, the GAPIT performs the analyses of statistical power, Type I error, and FDR simultaneously. The whole genome is divided into QTN bins and non-QTN bins. A bin is defined as a QTN bin if it contains at least one QTN; otherwise, the bin is defined as a non-QTN bin. The strength of association of a bin is defined by the marker in the bin with the most significant p value. Users can specify the set of bin sizes or select a default setting (1 base pair; 10, 100, and 500 kb; and 1 Mb).

The non-QTN bins are used to derive a null distribution of *P*-values. For a specific threshold of Type I error, the proportion of QTNs detected is defined as the statistical power. The paired statistical power and the FDR are derived by sorting the bins, with the most significant one on the top. For each bin, the statistical power is defined as the proportion of QTNs included in that bin plus the QTNs in the bins above. The FDR for each bin is defined as the ratio of the number of non-QTN bins to the total number of bins located above the bin. The GAPIT provides a function to compare the differences in statistical power among the various statistical models implemented in GAPIT and also third party methods, such as PLINK (Purcell et al., 2007).

Genomic Prediction and Cross-validation

Although most have reported that the use of the GAPIT is primarily for GWAS, we expect a substantial proportion of researchers, especially plant and animal breeders, to use the GAPIT for genomic prediction. Genomic prediction in the GAPIT can be conducted with four models: regular MLM, CMLM, ECMLM, and SUPER. The regular MLM uses kinship among all individuals. The CMLM produces predictions at the group level. A group prediction can then be used for the associated individuals (i.e., individuals in the same group share the same prediction). Genomic prediction with ECMLM and SUPER is similar to the method used in CMLM, with the addition of optimization on the combination of cluster algorithms, group kinship methods, and pseudo-QTNs to define individual kinship. All these methods can be evaluated for specific genotype and phenotype datasets. Cross-validation can be performed for any model with a specific fold. Model fit in reference and prediction accuracy in inference are evaluated separately.

Enhanced Output

The GAPIT's original output functions were improved to help users interpret results. For example, the –log *P*-values are displayed on the basis of their magnitude. The insignificant values located at the bottom of the Manhattan plots are displayed as open circles. The most significant values, located at the top, are displayed as solid filled dots. The significance of the associations at values between these two extremes is indicated by all other dots with varying amounts of fill. Additionally, confidence intervals at 95% are displayed on the QQ plots, providing an objective criterion for differentiating the observed values from the expected values under the null hypothesis.

Beyond improvements to the original output functions, the enhanced version of GAPIT also contains new graphing options. For example, in addition to the plot of principal components (PCs) (Groth et al., 2013) in two-dimensional format, population structure is displayed in three-dimensional. Plots are also provided for pairs between minor allele frequency (MAF) and –log *P*-values, which serve as a flag, especially for the associated SNPs with small MAFs. Characteristics of genotype data can be revealed by the plots of frequency and accumulative frequency against marker density. This graph is accompanied by the decay of LD over distance. This comparison will help researchers determine if most of the markers are in strong LD with adjacent markers. Therefore, the hidden genes have the same chance of being in LD with existing markers.

User Manual and Forum

We created demonstration data and an associated demonstration script to help users quickly learn the essentials and begin using the GAPIT without delay. Details are included in the GAPIT User Manual. The source code, demonstration data, and demonstration script are available at http://zzlab.net/GAPIT (accessed 11 Feb. 2016). Frequently asked questions and answers are also included in the GAPIT User Manual. To ask additional questions or to report errors, we encourage users to post questions and comments to the GAPIT forum (https://groups.google.com/forum/#!forum/gapit-forum [accessed 11 Feb. 2016]).

Limitations

Although the GAPIT generates comprehensive tables and graphs to assist the interpretation of the results, the execution has to be conducted as command line in the R environment. The GAPIT takes longer to learn than other software packages with graphic user interfaces such as TASSEL (Bradbury et al., 2007). The software package also did not implement the functions to derive kinship from both pedigree and genetic markers (Misztal et al., 2009), models of multiple traits (Arthur et al., 2012), and multiple marker tests (Segura et al., 2012).

Results

The enhanced GAPIT was designed to implement state-of-the-art methods, produce a comprehensive set of high-quality and publication-ready graphs and tables, and to help users interpret the results through this variety of outputs. Importantly, these outputs aid in visualizing and understanding the results of each step in the GWAS and genomic prediction process, from diagnoses of phenotypes and genotypes to assessment of statistical power and genomic prediction accuracy. Next, we provide descriptions and illustrative examples of the GAPIT's enhanced capabilities.

Phenotype Diagnosis

The normality of distribution on the residual effects is required for all statistical models implemented in the GAPIT. Although approximation on non-normal distributed phenotypes can be conducted with the GAPIT, caution should be taken, as the statistical power may be reduced. Users of the GAPIT can visualize and validate phenotype distributions by checking the corresponding graphs such as histograms and box plots. Attention should be paid to the requirement of normality on

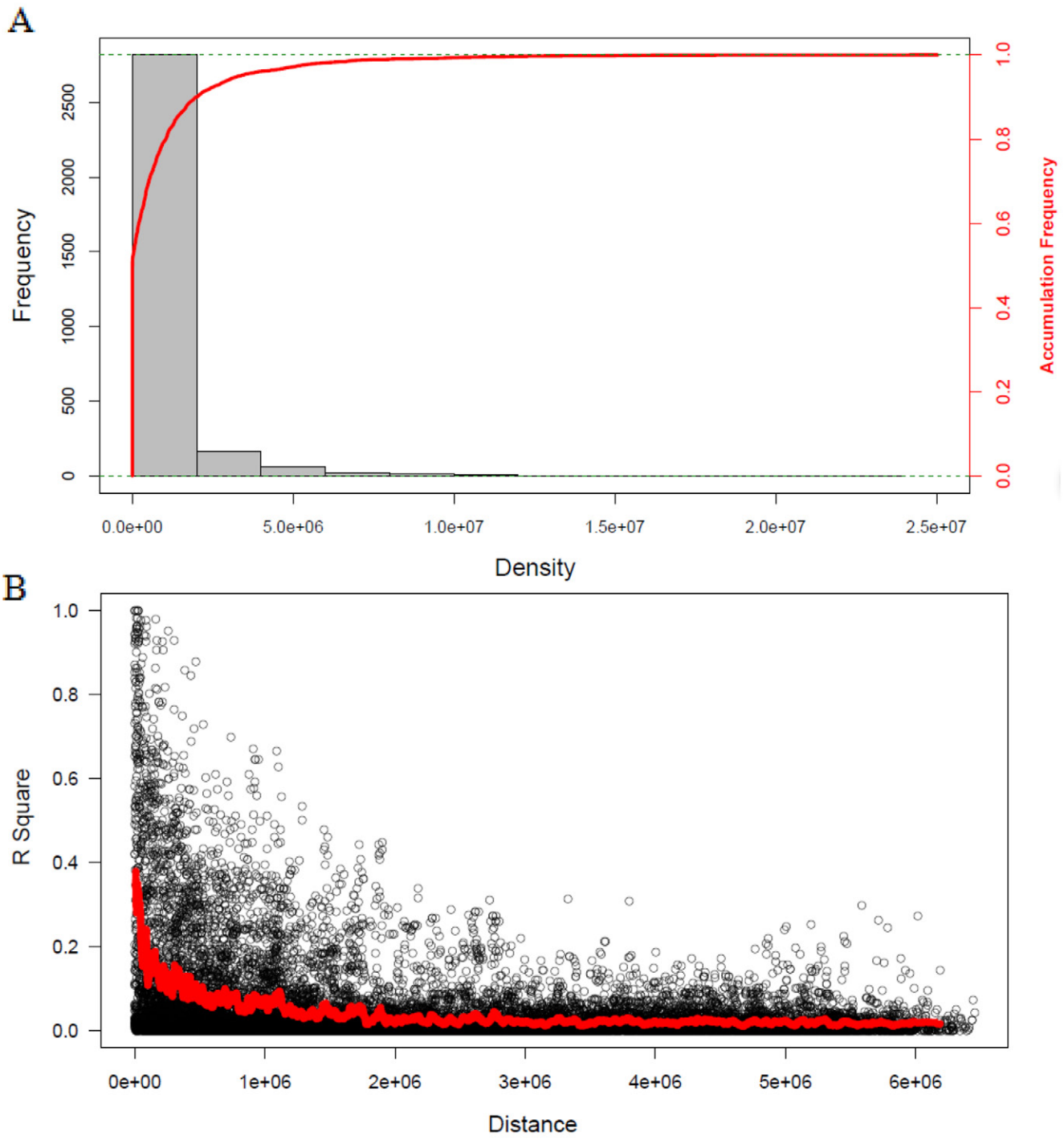

Fig. 2. Marker density and decay of linkage disequilibrium (LD) over distance. The distribution of marker density is displayed as a histogram and an accumulative distribution (A). Linkage disequilibrium was calculated on sliding windows with 100 adjacent genetic markers. Each dot represents a pair of distances between two markers on the window and their squared correlation coefficient (B). The red line is the moving average of the 10 adjacent markers.

residual effects. The raw phenotype distribution may appear to be bimodal for data from two treatments with distinct means. These plots also help in spotting outliers, which are important sources of error and may need correction to avoid false positives (Supplemental Fig. S1).

Genotype Diagnosis

Before analyses of GWAS and genomic prediction, researchers should validate and maintain genotype quality. The GAPIT provides a series of diagnostic tools to help users perform quality control on genotypes. These tools include histograms and accumulative distributions of marker density and decay plots of LD over distance (Fig. 2).

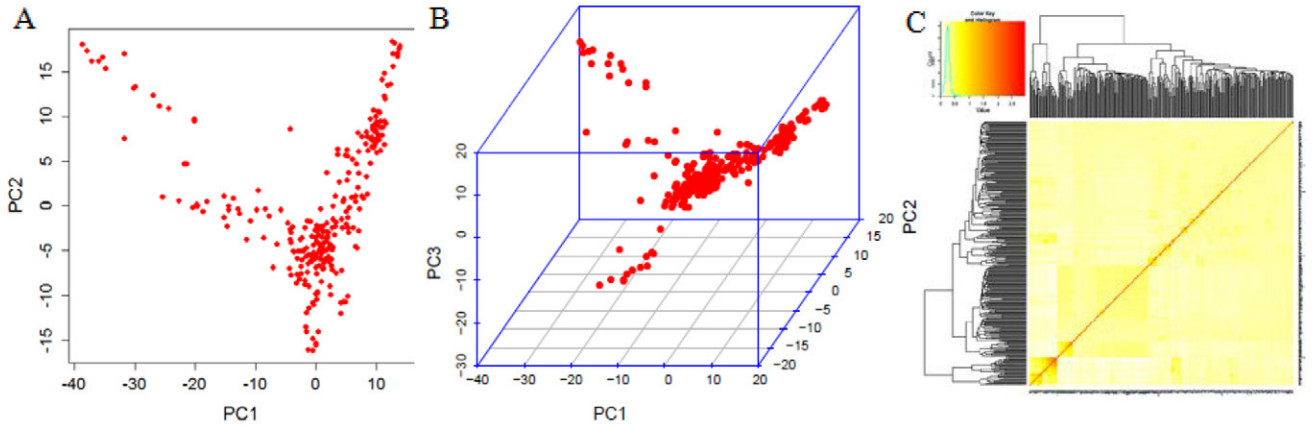

Fig. 3. Population structure and kinship. Principal components (PCs) and kinship are derived from all or a proportion of genetic markers to reveal the population structure. These PCs are virtually displayed in two (A) and three dimensions (B). Kinship is displayed virtually via a heat map and a tree (C).

Population Structure and Kinship

Analyses on population and kinship among individuals are always performed with or without phenotype data. When phenotype data are not specified (NULL), the GAPIT performs analyses on population structure and kinship only. Otherwise, GWAS and genomic prediction will continue. Principal components are outputted as tables with the number of PCs specified by users. The first three PCs are displayed in a three-dimensional plot (Fig. 3). The pairs of all PCs are displayed in regular two dimensional plots. The kinship matrix is displayed as a heat map, where red indicates the highest correlation between pairs of individuals and yellow indicates the lowest correlation. A hierarchy tree among individuals is displayed based on their kinship.

Associations between Phenotype and Population Structure

Scatter plots are used to reveal the relationship between phenotypes and each PC. Phenotype is plotted on the horizontal axis; PCs are plotted on the vertical axis, one PC at a time (Supplemental Fig. S2). Horizontally, non-sympatric distribution indicates the correlation (e.g., Supplemental Fig. S2b). The correlation, either positive or negative, indicates the impact of population structure on the phenotypes.

Optimizing Compression

When using CMLM in the GAPIT, several genetic parameters and statistics can be displayed to assess the optimization on the number of groups resulting from different clustering algorithms and kinship grouping methods (Supplemental Fig. S3). These parameters and statistics include −2log likelihood, genetic variance, residual variance, total variance, and heritability, which is defined as the proportion of genetic variance over total variance. When using ECMLM, multiple lines are added to each plot of parameters and statistics to assess the optimization resulting from various combinations of cluster algorithms and grouping methods (Li et al., 2014). Heritability at the optimum likelihood is outputted as a pie chart. This heritability is based on groups, not individuals. To produce an individual-based heritability, users can set the lower and upper bounds for the optimization so that the number of groups is equal to the number of individuals.

Vizualising Associations

Associations between phenotypes and genetic markers are outputted to tab-delimited text files (Supplemental Table S1) and displayed as Manhattan plots. Genetic markers are positioned by their chromosomes and ordered by their base-pair positions. Genetic markers on adjacent chromosomes are displayed with different colors. The strength of the association signal is displayed in two ways. One indicator of strength is the height on the vertical axis for –log *P*-values; the greater the height, the stronger the association (Fig. 4). The other indicator is the degree of filling in the dots; the greater the area filled within the dot, the stronger the association. Users can place vertical lines at specific positions along the *x*-axis to identify candidate genes or QTNs in simulation studies.

Interpreting Association Results

The GAPIT provides multiple graphs to interpret the results of GWAS (Fig. 5). First, QQ plots illustrate how well the majority of genetic markers fit the null hypothesis (i.e., the markers that are not associated with the phenotype). A red line indicates the expectation. The area of the 95% confidence interval is filled in gray. The dots above the confidence interval on the right indicate the genetic markers that are associated with the phenotype. Second, –log *P*-values are displayed for genetic markers against their MAFs. Users should use caution with the associated genetic markers that exhibit low MAFs. Third, the GAPIT provides a statistical power analysis for the data analyzed. Statistical power is defined as the proportion of markers detected after a genetic effect was assigned to them, one at a time. The effects on the original phenotype are used to derive the null distribution (Yu et al., 2006).

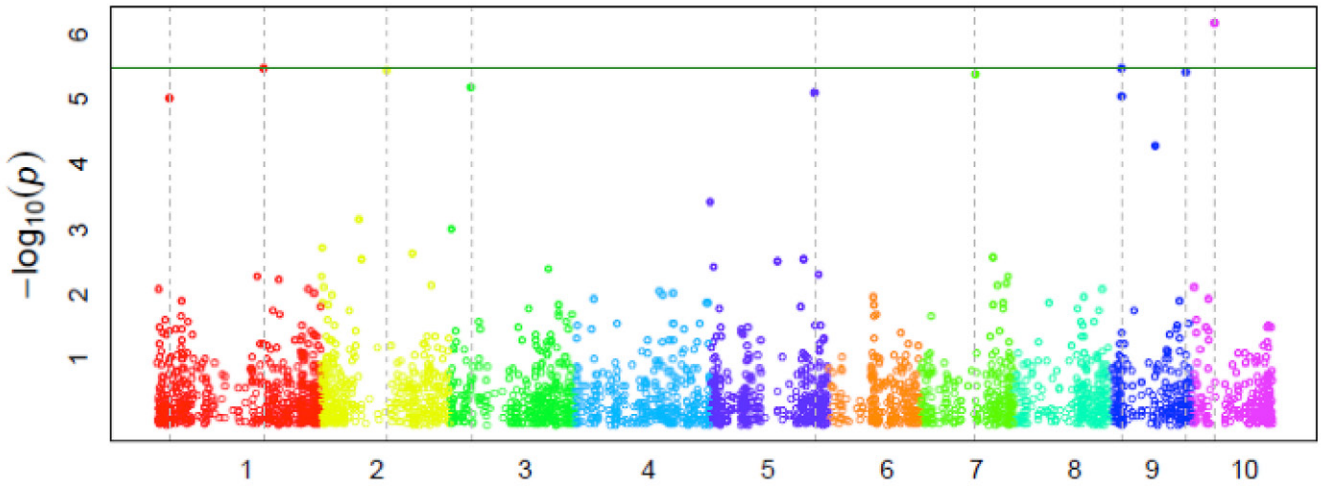

Fig. 4. Manhattan plot of a genome-wide association study. Probability values are displayed in negative log scale with base of 10 ($-\log_{10} P$) against the physical map positions of genetic markers. Chromosomes are designated with different colors. Candidate genes and quantitative trait nucleotides (QTNs) are marked with gray vertical dotted lines.

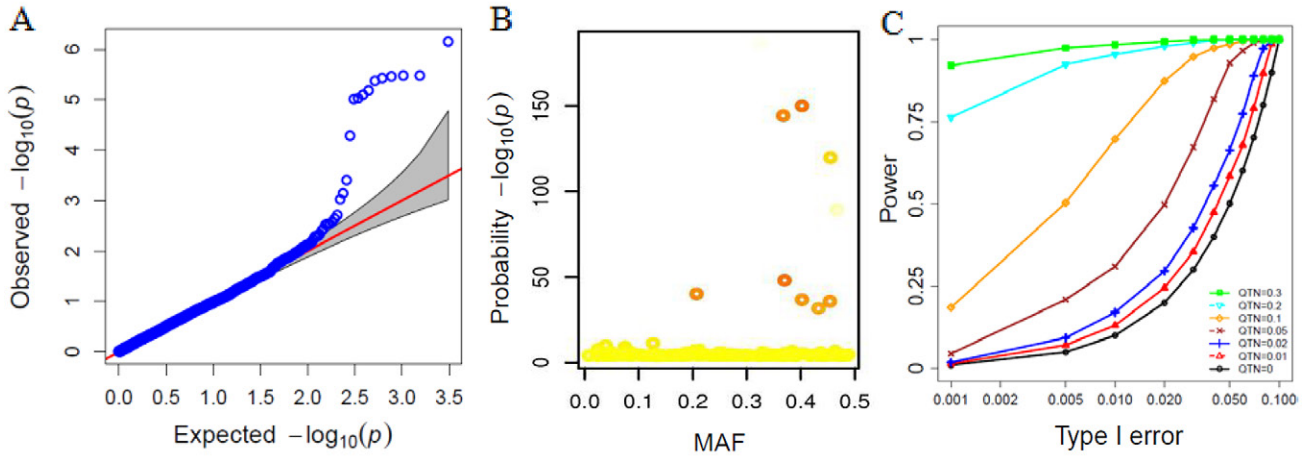

Fig. 5. Result diagnoses of the genome-wide association study. The observed distribution of *P*-values is displayed on a negative log scale with a base of 10 ($-\log_{10}$) and plotted against the null distribution (expected) (A) and minor allele frequency (MAF) (B). Statistical power is displayed against Type I error for the analyzed population and genetic markers. Statistical power is defined as the proportion of detected markers with an additional genetic effect added to the analyzed phenotype. The genetic effect was added to genetic markers one at time with magnitudes of 0, 0.01, 0.02, 0.05, 0.1, 0.2, and 0.3 units of phenotype SD.

Experimental Design

The GAPIT provides several functions to examine statistical power for experimental designs. The examination is based on a specific population for a trait with a specific genetic architecture. Power is presented in two tables, one against FDR (Supplemental Table S2) and the other against Type I error (Supplemental Table S3). A visualization function is available for comparing different statistical methods implemented in the GAPIT (Fig. 6). Users can also use the GAPIT to examine third-party software packages, such as PLINK (Purcell et al., 2007).

Genomic Prediction

Genomic predictions are outputted to tab-delimited text files (Supplemental Table S4) for all individuals with genotypes, whether an associated phenotype is specified or not. A prediction error variance is calculated for each prediction to serve as a confidence interval. The heat map remains from the first release of GAPIT and is used to illustrate the counts of number of individuals for combinations of genomic predictions and prediction error variance. Group assignments are also listed in the text file for each individual. Individuals within the same group assignment should have the same genomic prediction. The accuracy of genomic prediction can be evaluated with a GAPIT function through cross-validation at sets of different folds. The prediction accuracy of inference and model fit for reference are illustrated at different levels of fold (Fig. 7). The predicted phenotypes are displayed against the observed phenotypes for reference and inference separately.

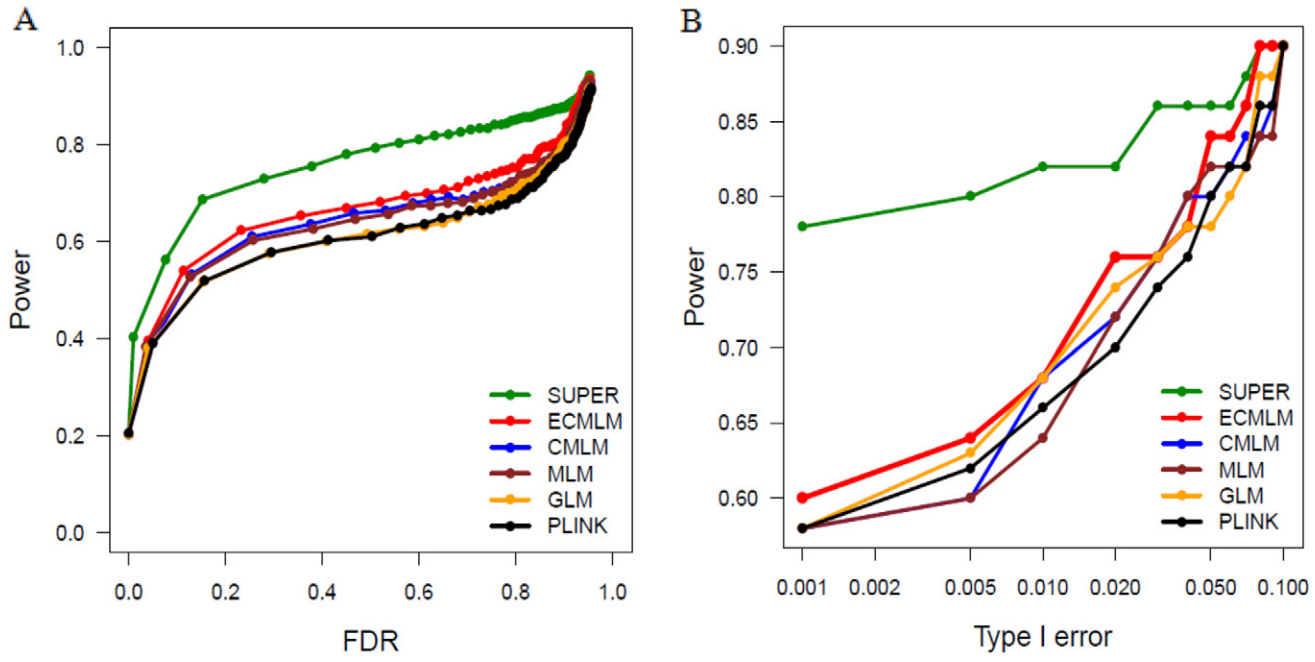

Fig. 6. Examinations of statistical power via different statistical methods. The examinations were performed with the genome association and prediction integrated tool's (GAPIT's) power comparison functions. The methods used can be those available in the GAPIT or in third-party software such as PLINK. The larger the area under the curve, the greater the power and the more desirable a method. Statistical power is displayed against false discovery rate (FDR) (A) and Type I error (B).

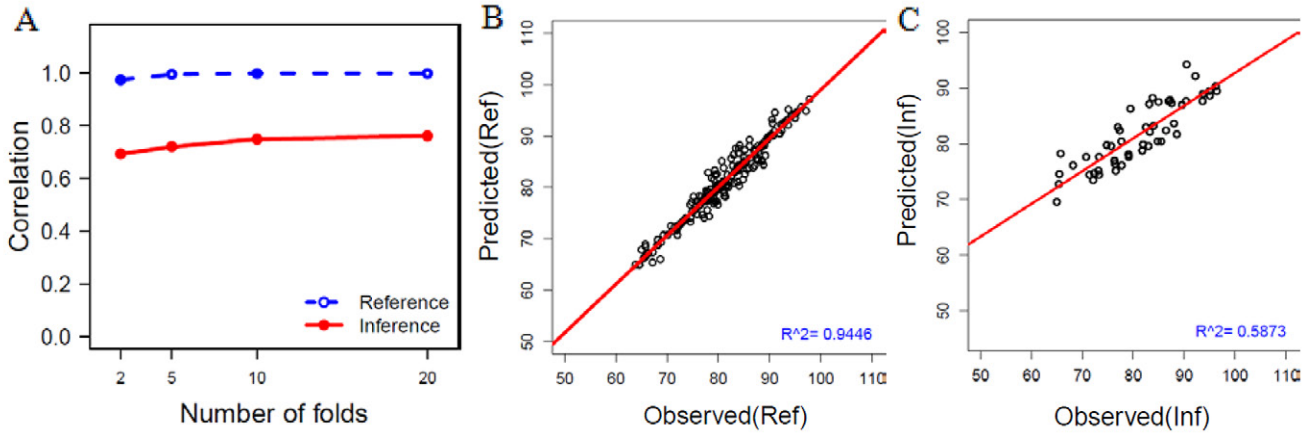

Fig. 7. Prediction accuracy and model fit under cross-validation. Pearson correlation coefficients between observed and predicted phenotypes are calculated as the model fit for reference and prediction accuracy for inference (A). Reference and inference are randomly assembled in the cross-validation at a user-specified fold (e.g. 2, 5, 10, or 20). Model fit and prediction accuracy for the fivefold level are displayed for reference (B) and inference (C).

Conclusions

We expect that the enhanced GAPIT will advance genomic research because it implements many state-of-the-art methods to improve the efficiency and accuracy of GWAS and genomic prediction. Similar to the original GAPIT package, the updated version is relatively easy to use and provides publication-ready tabular summaries and graphs. Our new, enhanced version of the GAPIT augments the popular original GAPIT by enabling free access to some of the most powerful and accurate GWAS and genomic prediction approaches available today.

Acknowledgments

This study was supported by NSF (0922493 and 1238014), USDA-ARS, an Emerging Research Issues Internal Competitive Grant from the Agricultural Research Center at Washington State University, College of Agricultural, Human, and Natural Resource Sciences, the Endowment and Research Project (No. 126593) from the Washington Grain Commission, the National Natural Science Foundation of China (Grant no. 31301748), and the China Postdoctoral Science Foundation (Grant no. 2014M551607). The authors thank Linda R. Klein for copyediting the manuscript.

References

Arthur, K., B.J. Vilhjálmsson, V. Segura, A. Platt, Q. Long, and M. Nordborg. 2012. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat. Genet. 44:1066–1071. doi:10.1038/ng.2376

Atwell, S., Y.S. Huang, B.J. Vilhjalmsson, G. Willems, M. Horton, Y. Li, et al. 2010. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 465:627–631. doi:10.1038/nature08800

Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, and E.S. Buckler. 2007. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23(19):2633–2635. doi:10.1093/bioinformatics/btm308

Cantor, R.M., K. Lange, and J.S. Sinsheimer. 2010. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. Am. J. Hum. Genet. 86(1):6–22. doi:10.1016/j.ajhg.2009.11.017

Endelman, J. 2011. Ridge regression and other kernels for genomic selection in the R package rrBLUP. Plant Gen. 4:250–255. doi:10.3835/plantgenome2011.08.0024

Falush, D., M. Stephens, and J.K. Pritchard . 2007. Inference of population structure using multilocus genotype data: Dominant markers and null alleles. Mol. Ecol. Notes 7(4):574–578. doi:10.1111/j.1471-8286.2007.01758.x

Georges, M. 2014. Towards sequence-based genomic selection of cattle. Nat. Genet. 46(8):807–809. doi:10.1038/ng.3048

Groth, D., S. Hartmann, S. Klie, and J. Selbig. 2013. Principal components analysis. Methods Mol. Biol. 930(May):527–547. doi:10.1007/978-1-62703-059-5_22

Kang, H.M., J.H. Sul, S.K. Service, N.A. Zaitlen, S.-Y. Kong, N.B. Freimer, et al. 2010. Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. 42:348–354. doi:10.1038/ng.548

Kang, H.M., N.A. Zaitlen, C.M. Wade, A. Kirby, D. Heckerman, M.J. Daly, et al. 2008. Efficient control of population structure in model organism association mapping. Genetics 178:1709–1723. doi:10.1534/genetics.107.080101

Li, M., X. Liu, P. Bradbury, J. Yu, Y.-M. Zhang, R.J. Todhunter, et al. 2014. Enrichment of statistical power for genome-wide association studies. BMC Biol. 12(1):73. doi:10.1186/s12915-014-0073-5

Lipka, A.E., F. Tian, Q. Wang, J. Peiffer, M. Li, P.J. Bradbury, et al. 2012. GAPIT: Genome association and prediction integrated tool. Bioinformatics 28(18):2397–2399. doi:10.1093/bioinformatics/bts444

Lippert, C., J. Listgarten, Y. Liu, C.M. Kadie, R.I. Davidson, and D. Heckerman. 2011. FaST linear mixed models for genome-wide association studies. Nat. Methods 8(10):833–835. doi:10.1038/nmeth.1681

Listgarten, J., C. Lippert, C.M. Kadie, R.I. Davidson, E. Eskin, and D. Heckerman. 2012. Improved linear mixed models for genome-wide association studies. Nat. Methods 9(6):525–526. doi:10.1038/nmeth.2037

Misztal, I., A Legarra, and I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J. Dairy Sci. 92(9):4648–4655. doi:10.3168/jds.2009-2064

Ober, U., J.F. Ayroles, E.A. Stone, S. Richards, D. Zhu, R.A. Gibbs, et al. 2012. Using whole-genome sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. PLoS Genet. 8(5):E1002685. doi:10.1371/journal.pgen.1002685

Poland, J., J. Endelman, J. Dawson, J. Rutkoski, S.Y. Wu, Y. Manes, et al. 2012. Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Gen. 5(3):103–113. doi:10.3835/plantgenome2012.06.0006

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, D. Bender, et al. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81(3):559–575. doi:10.1086/519795

Segura, V., B.J. Vilhjálmsson, A. Platt, A. Korte, Ü. Seren, Q. Long, et al. 2012. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat. Genet. 44(7):825–830. doi:10.1038/ng.2314

Shendure, J., R.D. Mitra, C. Varma, and G.M. Church. 2004. Advanced sequencing technologies: Methods and goals. Nat. Rev. Genet. 5(5):335–344. doi:10.1038/nrg1325

Svishcheva, G.R., T.I. Axenovich, N.M. Belonogova, C.M. van Duijn, and Y.S. Aulchenko. 2012. Rapid variance components-based method for whole-genome association analysis. Nat. Genet. 44(10):1166–1170. doi:10.1038/ng.2410

VanRaden, P.M., C.P. Van Tassell, G.R. Wiggans, T.S. Sonstegard, R.D. Schnabel, J.F. Taylor, et al. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92(1):16–24. doi:10.3168/jds.2008-1514

Wang, Q., F. Tian, Y. Pan, E.S. Buckler, and Z. Zhang. 2014. A SUPER powerful method for genome wide association study. PLoS ONE 9(9):E107684. doi:10.1371/journal.pone.0107684

Yang, J., N.A. Zaitlen, M.E. Goddard, P.M. Visscher, and A.L. Price. 2014. Advantages and pitfalls in the application of mixed-model association methods. Nat. Genet. 46:100–106. doi:10.1038/ng.2876

Yu, J., G. Pressoir, W.H. Briggs, I. Vroh Bi, M. Yamasaki, J.F. Doebley, et al. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38:203–208. doi:10.1038/ng1702

Zhang, Z., E. Ersoz, C.Q. Lai, R.J. Todhunter, H.K. Tiwari, M.A. Gore, et al. 2010. Mixed linear model approach adapted for genome-wide association studies. Nat. Genet. 42(4):355–360. doi:10.1038/ng.546

Zhou X., and M. Stephens. 2012. Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. 44(7):821–824. doi:10.1038/ng.2310.