GWAS and GS Are as Easy as Clicking and Dragging with iPat

Chunpeng James Chen[1], Zhiwu Zhang[1]*

[1]Department of Crop and Soil Sciences, Washington State University, Pullman, Washington, 99164, USA.

[*]Correspondences should be addressed to ZZ (Email: Zhiwu.Zhang@WSU.edu)

Introduction

One of the ultimate goals of genomic research is to predict phenotypes from genotypes, thereby improving the practices of medicine for humans and selective breeding for agricultural production. Dissecting the genetic architecture of complex traits, such as disease resistance and grain yield and quality, has been a major focus in efforts to identify genes underlying these traits. With increasing availability of dense genetic markers, such as single nucleotide polymorphisms (SNPs), insertions/deletions, and copy number variations, Genome-Wide Association Study (GWAS) has become the most efficient approach for mapping genes of interest. Compared with linkage analysis, which is based on recombination experiments, GWAS reveal statistical signals that associate phenotypes with genotypes. But, because the signals can occur for many reasons other than associated genetic markers locate near causal genes, false positives are a major concern in GWAS.

Many sophisticated statistical methods have been developed to reduce false positives and false negatives in Genome-Wide Association Studies (GWAS). The first method was introduced to reduce false positives due to population structure, the most common cause of false positives in GWAS. This method and its variations, such as Principal Component Analysis (PCA), are implemented in many software packages used to conduct GWAS[1], including PLINK[2]. PLINK is still the most common software for GWAS, especially in the human genetics research community.

In addition to population structure, the cryptic relationships among individuals also cause spurious associations. Consequently, the Mixed Linear Model (MLM) was introduced to incorporate kinship among individuals as random effects[3]. Kinship can be calculated either by using all available markers, associated markers not near the testing markers[4], or markers not in Linkage Disequilibrium (LD) with the testing markers[5]. Alternatively, the individuals in the regular MLM can be clustered into groups in the compressed MLM (CMLM), where the random effects are based on the groups[6]. Variations were introduced to derive the kinship among the groups, including the Enriched CMLM (ECMLM)[7]. Recently, to reduce the computing burden in MLM, GLM was jointly used with MLM in an iterative fashion[8]. Many software packages were developed to implement these methods, including TASSEL[9], GAPIT[10,11], and FarmCPU[8].

Because heritabilities are missing from genetic loci identified by GWAS, a different approach uses all available markers, whether they are significantly associated with the phenotypes or not. This approach is named Genomic Selection (GS). In fact, the first method developed for GS does not even estimate the effects of genetic markers. Instead, the method uses all markers in MLM to derive kinship among individuals and directly make the genomic Best Linear Unbiased Prediction (BLUP) for individuals' genetic merit. This method is now known as genomic BLUP, or gBLUP. gBLUP was implemented into MTDFREML[12], a widely used software package to estimate variance components and BLUP. An efficient algorithm was developed to derive kinship from markers. Inspired from GWAS, all available markers are simultaneously fitted as random effects and their effects are summed together to indirectly predict the individuals' genetic merit. Different distributions were applied to the random effects in the framework of Bayesian theory. Several software packages were developed to solve these Bayesian methods, including BLR and BGLR[13].

With such a large number of software packages, which are mostly executed through a Command Line Interface (CLI), genomic researchers are struggling with the steep learning curves for using them efficiently. These CLI packages have zero tolerance for grammar and syntax errors; users must input commands and specify parameters exactly as required by the packages. Furthermore, reductions in sequencing costs have made it possible to conduct genomic research on extremely large datasets, known as Big Data. But, researchers are unable to convert their genomic data into the specific formats, required by different software packages, with easy-to-use and familiar computing tools such as Excel. That is, computer programming skills are required to convert Big Data from one format to another. These restrictions limit genomic researchers from fully exploring the potential value of their data.

We developed a software package, named Intelligent Prediction and Association Tool (iPat) to achieve the following objectives: 1) incorporate common GWAS and GS CLI packages into a user-friendly Graphical User Interface (GUI) to easily navigate their use, 2) convert any typical data format into that required by a specific CLI package, and 3) provide standardized presentation graphics to help interpret input data and output results. With iPat, the computer programming requirement is completely eliminated to conduct GWAS and GS.

Method

To achieve the specific objectives of this study, iPat was designed to perform three roles: interpreter, translator, and presenter. As an interpreter, iPat understands the requirements from the users and transforms them into commands and parameters for a specific CLI package. As a translator, iPat takes any typical data format and converts it to the one specified by a particular CLI package. As a presenter, iPat provides a uniform set of tables and figures to help the user diagnose input data and explain the output results from the incorporated CLI packages. We used a mixture of programming in Java and R to develop iPat. Both interpreter and translator were written in Java and presenter was written in R. Because both Java and R can be used across operating systems, iPat functions independently from an operating system.

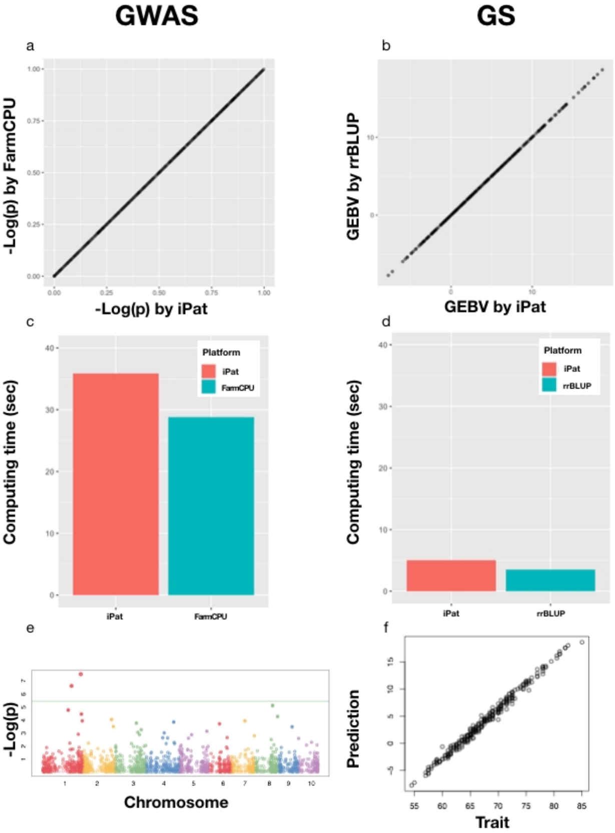
Interpreter: The initial frame of iPat is completely blank with the exception of the label "iPat". The blank frame prompts users to input data by dragging their files with any computer pointing device. Multiple data files can be dragged individually or simultaneously. A data file icon is assigned to each file and labeled with the original file name. The data icons can be repositioned or deleted. Deletion of a data icon on the iPat frame does not delete the file from its original folder. Right-clicking a file icon will display an option list, including opening the file, deleting the file, and changing the file type. Double-clicking a file icon will open the file with a default application, such as Excel or a text editor. Double-clicking on an empty space within the iPat frame will create another type of icon, named "project". This label is automatically created and can be renamed. Similar to file icons, project icons can be repositioned or deleted. Right-clicking a project icon will display an option list, including selecting the type of analysis (e.g., GWAS or GS) and running the analysis.

For the option of GWAS or GS, users can select a CLI package and specify parameters in the dialog window. When the Run option is selected, iPat will pass user-defined configurations (i.e. input arguments) to the specific CLI package using an R-script file. The

CLI package then runs behind the scenes in the CLI environment. After the package is launched, a log window pops up and displays messages returned from the package, allowing users to monitor the progress of the analysis. This window is visually the same as the message window in the CLI environment. When the analysis is finished, the project icon will either turn to green, indicating successful completion of the analysis, or red, indicating an error.

Translator: Currently, iPat accepts four types of genotype data formats: Hapmap, numerical, VCF, and PLINK. Any of these data formats can be used to conduct analyses with any of the incorporated CLI packages. The input file format is detected and automatic conversion is conducted if the chosen CLI package requires a different format. Format detection is based on four features of the input file: number of rows of genotype and map files, number of columns of genotype and map files, number of elements in one single row of a genotype file, and file extension names. For example, if a genotype file has 11 more columns than a map file's number of rows, this dataset would be detected as a Hapmap format. Or, if a single row of a genotype file contains '0', '1', and '2' and has less than 5 different elements (e.g., 3 genotypes and a missing value), it would be detected as a numerical file.

Presenter: Most of iPat's incorporated CLI packages, such as PLINK, rrBLUP[14], and BGLR, output results as text files. Graphical visualization of the results and input data are extremely helpful for data quality control, interpretation, and re-modeling. One of the incorporated packages, GAPIT, adds many graphs in PDF format to visualize the input data and output results. These graphs include phenotype distribution, marker density distribution, linkage disequilibrium decay, Manhattan plot, and QQ plot. We integrated most of the graphs from GAPIT into iPat. These graphs can be generated for any of the incorporated CLI packages. We also added extra graphs into iPat to further enhance the visualization of results, including plots of genomic-estimated breeding values against their observed phenotypes.



Results

Figure 1. Performances and outputs of iPat. iPat generates P values (a) and Genomic Estimated Breeding Values (b) that are identical to the ones by running the incorporated packages in Command Line Interface (CLI) environment for Genome-Wide Association (GWAS) and Genomic Selection (GS). The computing time is also similar (c and d) as iPat only add an extra time to load the CLI package. For any CLI packages incorporated, iPat provides uniform outputs, including Manhattan plot (e) and plot of prediction and observed trait(f).

The design of iPat not only allows users to easily interact with the incorporated CLI packages, but also retains the efficiency of the original packages. First, the outputs of iPat are identical to the outputs obtained by running the CLI packages under a CLI environment (**Figure 1a** and **b**). Second, computing times are very similar (**Figure 1c** and **d**). Once iPat interprets the user-defined commands and parameters for a chosen CLI package, the analysis program is executed under a CLI environment.

Moreover, iPat produces comprehensive graphs to help diagnose the quality and properties of phenotypic data and genotypic data, including distributions and marker densities. iPat provides standardized presentation graphics for the results generated from all of the incorporated CLI packages, including GAPIT, FarmCPU, rrBLUP, PLINK, and BGLR. When GWAS is conducted either through GAPIT, FARMCPU, or PLINK, both Manhattan plots (**Figure 1e**) and QQ plots are created by iPat. Similarly, when GS is conducted by either GAPIT, rrBLUP, or BGLR, estimated genomic breeding values are plotted against phenotypes for the assessment of model fit (**Figure 1f**).

With iPat as interpreter, translator, and presenter, users can employ any typical data format and access any incorporated CLI package without computer programming skills or memorizing commands and parameters. The iPat package, including executable files, user manual, and demonstration datasets are freely available at http://zzlab.net/iPat.

References

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

Endelman, J. Ridge regression and other kernels for genomic selection in the R package rrBLUP. *Plant Genome* **4,** 250–255 (2011).

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

Listgarten, J. *et al.* Improved linear mixed models for genome-wide association studies. *Nat. Methods* **9,** 525–526 (2012).

Li, M. *et al.* Enrichment of statistical power for genome-wide association studies. *BMC Biol.* **12,** 73 (2014).

Liu, X., Huang, M., Fan, B., Buckler, E. S. & Zhang, Z. Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. *PLoS Genet.* **12,** e1005767 (2016).

Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–959 (2000).

Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81,** 559–575 (2007).

Pérez, P. & De Los Campos, G. Genome-wide regression and prediction with the BGLR statistical package. *Genetics* **198,** 483–495 (2014).

Tang, Y. *et al.* GAPIT Version 2: An Enhanced Integrated Tool for Genomic Association and Prediction. *Plant J.* **9,** (2016).

Wang, Q., Tian, F., Pan, Y., Buckler, E. S. & Zhang, Z. A SUPER Powerful Method for Genome Wide Association Study. *PLoS One* **9,** e107684 (2014).

Yu, J. M. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38,** 203–208 (2006).

Zhang, Z., Todhunter, R. J., Buckler, E. S. & Van Vleck, L. D. Technical note: Use of marker-based relationships with multiple-trait derivative-free restricted maximal likelihood. *J. Anim. Sci.* **85,** 881–885 (2007).

Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42,** 355–360 (2010).