Bioinformatics, 34(11), 2018, 1925–1927
doi: 10.1093/bioinformatics/bty015
Advance Access Publication Date: 11 January 2018
Applications Note

OXFORD

Genome analysis

iPat: intelligent prediction and association tool for genomic research

Chunpeng James Chen and Zhiwu Zhang*

Department of Crop and Soil Sciences, Washington State University, Pullman, WA 99164, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on May 12, 2017; revised on September 1, 2017; editorial decision on January 5, 2018; accepted on January 9, 2018

Abstract

Summary: The ultimate goal of genomic research is to effectively predict phenotypes from genotypes so that medical management can improve human health and molecular breeding can increase agricultural production. Genomic prediction or selection (GS) plays a complementary role to genome-wide association studies (GWAS), which is the primary method to identify genes underlying phenotypes. Unfortunately, most computing tools cannot perform data analyses for both GWAS and GS. Furthermore, the majority of these tools are executed through a command-line interface (CLI), which requires programming skills. Non-programmers struggle to use them efficiently because of the steep learning curves and zero tolerance for data formats and mistakes when inputting keywords and parameters. To address these problems, this study developed a software package, named the Intelligent Prediction and Association Tool (iPat), with a user-friendly graphical user interface. With iPat, GWAS or GS can be performed using a pointing device to simply drag and/or click on graphical elements to specify input data files, choose input parameters and select analytical models. Models available to users include those implemented in third party CLI packages such as GAPIT, PLINK, FarmCPU, BLINK, rrBLUP and BGLR. Users can choose any data format and conduct analyses with any of these packages. File conversions are automatically conducted for specified input data and selected packages. A GWAS-assisted genomic prediction method was implemented to perform genomic prediction using any GWAS method such as FarmCPU. iPat was written in Java for adaptation to multiple operating systems including Windows, Mac and Linux.

Availability and implementation: The iPat executable file, user manual, tutorials and example datasets are freely available at http://zzlab.net/iPat.

Contact: zhiwu.zhang@wsu.edu

1 Introduction

Genome-wide association studies (GWAS) have become the primary method for dissecting complex traits. To incorporate population structure, a general linear model was implemented in PLINK (Purcell *et al.*, 2007) to reduce the spurious associations. Mixed linear models have been developed to incorporate cryptic relationships among individuals to further reduce the spurious associations. Software packages have been developed correspondingly to conduct the analyses, including TASSEL (Bradbury *et al.*, 2007), EMMA (Kang *et al.*, 2008), GAPIT (Lipka *et al.*, 2012; Tang *et al.*, 2016) and FarmCPU (Liu *et al.*, 2016).

Other recently developed analytical methods have also given genomic research a boost toward improving disease risk management in humans and molecular breeding of plants and animals—the ultimate goals of genomic prediction. These packages include rrBLUP (Endelman, 2011) and BGLR (Pérez and De Los Campos, 2014). rrBLUP implements ridge regression and genomic BLUP (gBLUP) and BGLR implements Bayesian methods such as Bayes A, B, CPi and LASSO. Some genomic prediction methods can be used for GWAS, for example, Bayes A, B and Cpi. In return, GWAS results can also enhance genomic prediction (Spindel *et al.*, 2016).

The multiple available software packages provide the potential to enhance data analyses, but also create challenges for users. Most packages only use a command-line interface (CLI), which has a very steep learning curve for non-programmers. Furthermore, users must spend great effort when shifting from one package to another due to inconsistent format requirements for input data. Users must take the time to reformat their data accordingly. As a result, a user-friendly graphical user interface (GUI)-based software package that can access multiple CLI packages, use any type of the input file format, and perform both GWAS and genomic prediction or selection is critically needed.

The objective of this study was to develop a software package with the following functions: (1) performs both GWAS and genomic prediction, including GWAS-assisted genomic prediction; (2) offers a friendly GUI to reduce user learning time and (3) requires only one input data format to conduct any analysis with any incorporated method.

2 GWAS-assisted genomic prediction

By default, Intelligent Prediction and Association Tool (iPat) conducts genomic prediction after GWAS with any implemented CLI package. Genomic prediction is conducted by gBLUP with associated loci fitted as fixed effects in the following model:

$$\boldsymbol{y} = \mathbf{W}\boldsymbol{\gamma} + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{u} + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{y}$ is a vector of phenotypes; $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ represent unknown fixed effects, with $\boldsymbol{\gamma}$ as inheritable factors (e.g. population structure and associated genetic loci) and $\boldsymbol{\beta}$ as uninheritable factors (e.g. environmental treatments); and $\boldsymbol{u}$ is a vector of genomic prediction with size n (number of individuals) for unknown random polygenic effects. These random effects follow a distribution with a mean of zero and a covariance matrix of $\boldsymbol{G} = 2\boldsymbol{K}\sigma_{\mathrm{a}}^2$, where $\boldsymbol{K}$ is the kinship with element k_{ij} (i, j = 1, 2, ..., n) representing the relationship between individuals i and j, and σ_{a}^2 is an unknown genetic variance. W, X and Z are the incidence matrices for $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$ and $\boldsymbol{u}$, respectively. $\boldsymbol{e}$ is a vector of random residual effects that are normally distributed with a mean of zero and a covariance of $\boldsymbol{R} = \boldsymbol{I}\sigma_{\mathrm{e}}^2$, where $\boldsymbol{I}$ is the identity matrix and σ_{e}^2 is the unknown residual variance. The predicted genetic merits (**GM**) of individuals are calculated by following equation:

$$\mathbf{GM} = \mathbf{W}\hat{\boldsymbol{\gamma}} + \mathbf{Z}\hat{\boldsymbol{u}} \tag{2}$$

where $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{u}}$ are the estimates and prediction of $\boldsymbol{\gamma}$ and $\boldsymbol{u}$, respectively.

The associated loci are defined as the genetic markers with P-values above the Bonferroni threshold. The associated loci are also filtered for markers that are in linkage disequilibrium (LD). Makers are sorted with the strongest associated marker on top. Any other marker with a LD of 50% (R^2) or above with the top marker is removed. Then, the second strongest associated marker is selected as the top marker and the same process is repeated until no markers can be removed. The sum of the associated markers and the other fixed effects must be less than the square root of the number of individuals. If not, the less significant markers are removed until this requirement is satisfied.

3 GUI, data and third party CLI packages

iPat's GUI is designed to drag and click input data and access third party CLI packages using a computer's pointing device (Fig. 1). Users can also use the keyboard to change parameters.

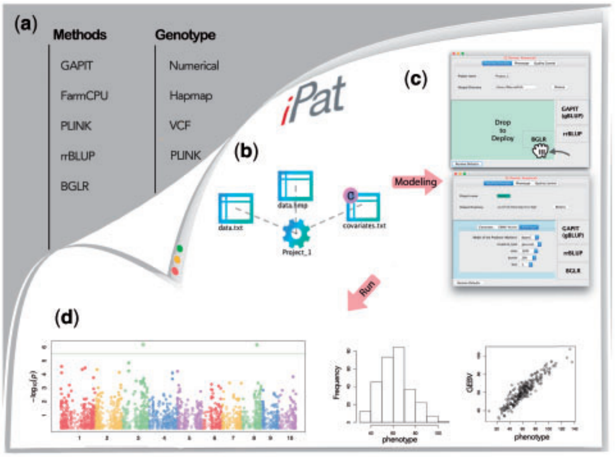

Fig. 1. Design of the iPat. iPat provides users the ability to access incorporated software packages and data inputs (**a**) by using a GUI. The GUI (**b**) allow users to control all the processes, including modeling (**c**) and displaying results (**d**). Currently, incorporated packages include GAPIT, PLINK, FarmCPU, BLINK, rrBLUP and BGLR. Genotype data can be input in any format, including numerical, hapmap, VCF and PLINK. The GUI allows users to drag any type of data file into the interface and create project icons to link data files, manage analyses and display results

After iPat is launched, the GUI appears as a blank frame labeled iPat. The frame is used to manage data files and project analyses. The frame behaves like a folder that users can drag any object into, including files and other folders. The graphical icons on the frame are links to the original files and folders. By double-clicking on these icons, the computer's operating system opens them with the appropriate default programs. For example, a folder is opened by file explore. A text file is opened by text editor.

A project icon can be created by double clicking anywhere on the iPat frame. Multiple project icons are acceptable. The project icons are used for linking the input files, defining parameters and initiating modeling analyses. Both project icons and file icons can be repositioned by dragging them with the pointing device. An icon can be deleted by dragging it to the bottom right-hand corner. When the icon is close to the corner, a trashcan will appear at the corner to indicate the deletion.

Overlapping a project icon and a data icon creates their connection and is indicated by a dashed line (Fig. 1). Clicking on the dashed line turns it into a solid line. Clicking again returns the solid line back to a dashed line. When a solid line, the connection can be dragged to the trashcan at the bottom right-hand corner for deletion. When a project icon is linked to required genotype and phenotype data files by the dashed line, the project icon can be opened as a dialog by right-clicking. In the dialogue box, the user can define parameters, select the desired model and execute the incorporated CLI packages to perform analyses. During the execution, the project icon will spin. The spinning will stop and display either a green or a red flag upon success or failure of the execution, respectively. Results of a successful run can be displayed by double-clicking the project icon.

4 Implementation

iPat's GUI was developed in Java. Input data and parameters are passed to specified CLI packages through the command-line interpreter. The interpreters are MS-DOS in a Windows operating system and Terminal in Mac OS or Linux systems. For an R-based package, the input parameters are translated into R script. The pre-requisite R packages are imported into a library before calling the R package for the analysis. iPat then opens a new thread and executes this R script file by calling the 'Rscript' function in the command-line interpreters. For instance, if a user would like to perform GWAS by FarmCPU, iPat will pass 'Rscript FarmCPU.r mydata.dat mydata.map mydata.txt . . .' to the command-line interpreter. The first argument of the function 'Rscript' signals which R script file should be compiled. The remaining arguments are used in FarmCPU.r, which defines the genotype data, genetic map and phenotype. For C-driven packages, iPat calls the command-line interpreter directly. For example, iPat will execute a command 'plink–bfile mydata–assoc –out mydata_out' if binary files are used to run GWAS in PLINK, where mydata and mydata_out specify the path and name of the input and output files, respectively.

Execution of the CLI packages are monitored using Java system functions. A new message panel is initiated to collect screen output for the CLI packages by calling java.lang.Process.getInputStream(). All information on the message panel is saved as a log file. A project can be terminated at any time by closing this message panel—an action that calls java.lang.Process.destroy(). iPat uses the commands java.io.IOException and java.lang.InterruptedException to catch exceptions in the executed command, allowing the program to detect whether or not the computation was completed successfully.

Input file formats are automatically converted to the formats corresponding to the specified CLI packages. iPat uses the first three lines of each input file to determine the formats. Acceptable genotype formats include hapmap, numerical, VCF, PLINK and BLINK. Phenotype formats are acceptable with or without individual identification. When input data formats match the chosen CLI package format requirements, analyses are conducted directly. Otherwise, format conversion is performed first.

Display results are presented uniformly with the same array of information and graphics, regardless of which CLI package is used. Most CLI packages produce a limited set of results, such *P*-values and genomic predictions. iPat uses the display functions in GAPIT as the universal set of result graphics, which include Manhattan plots, QQ plots and heat maps for prediction and accuracy distribution.

5 Conclusions

Because of its GUI, iPat allows users to perform genomic analyses without pre-requisite programming skills. Analyses include both mapping genes through GWAS and genomic prediction through understanding the relationships between genotypes and phenotypes. Additionally, iPat gives users the flexibility to combine different analysis methods (such as FarmCPU or rrBLUP) with different input formats (such as PLINK or hapmap genotype data) without requiring the tedious process of manual reformatting. These features should attract users of all levels. In turn, widespread use of iPat has the potential to spawn faster advances in genomic research.

Acknowledgement

The authors thank Linda R. Klein for helpful comments and editing the manuscript.

Funding

This work was partly supported by an Emerging Research Issues Internal Competitive Grant from the Agricultural Research Center at Washington State University, College of Agricultural, Human and Natural Resource Sciences; and the Endowment, Research Project (No. 126593) from the Washington Grain Commission, Department of Energy (awards of DE-SC0016366) and the National Institute of Food and Agriculture, U.S. Department of Agriculture (awards of 2015-05798 and 2016-68004-24770).

Conflict of Interest: none declared.

References

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

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

Kang,H.M. *et al.* (2008) Efficient control of population structure in model organism association mapping. *Genetics*, **178**, 1709–1723.

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

Liu,X. *et al.* (2016) Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet.*, **12**, e1005767.

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

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.

Spindel,J.E. *et al.* (2016) Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity (Edinb)*, **116**, 395–408.

Tang,Y. *et al.* (2016) GAPIT version 2: an enhanced integrated tool for genomic association and prediction. *Plant J.*, **9**, 1–9.