Bioinformatics, 37(9), 2021, 1324–1326
doi: 10.1093/bioinformatics/btaa824
Advance Access Publication Date: 25 November 2020
Applications Note

Genetics and population analysis

MMAP: a cloud computing platform for mining the maximum accuracy of predicting phenotypes from genotypes

Wei Huang[1], Ping Zheng[2], Zhenhai Cui[3], Zhuo Li[4], Yifeng Gao[4], Helong Yu[5], You Tang[4,5,*], Xiaohui Yuan[6,*] and Zhiwu Zhang[7,*]

[1]Economic and Management School, Jilin Agricultural Science and Technology University, Jilin, China, [2]Institute of Electrical and Information, Northeast Agricultural University, Harbin, China, [3]College of Life Sciences and Technology, Shenyang Agricultural University, Liaoning, China, [4]Electrical and Information Engineering College, JiLin Agricultural Science and Technology University, Jilin, China, [5]Information Technology Academy, Jilin Agricultural University, Changchun, China, [6]Department of Computer Sciences, Wuhan University of Technology, Wuhan, China and [7]Department of Crop and Soil Sciences, Washington State University, Pullman, WA, USA

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on June 3, 2020; revised on September 3, 2020; editorial decision on September 5, 2020; accepted on September 8, 2020

Abstract

Accurately predicting phenotypes from genotypes holds great promise to improve health management in humans and animals, and breeding efficiency in animals and plants. Although many prediction methods have been developed, the optimal method differs across datasets due to multiple factors, including species, environments, populations and traits of interest. Studies have demonstrated that the number of genes underlying a trait and its heritability are the two key factors that determine which method fits the trait the best. In many cases, however, these two factors are unknown for the traits of interest. We developed a cloud computing platform for Mining the Maximum Accuracy of Predicting phenotypes from genotypes (MMAP) using unsupervised learning on publicly available real data and simulated data. MMAP provides a user interface to upload input data, manage projects and analyses and download the output results. The platform is free for the public to conduct computations for predicting phenotypes and genetic merit using the best prediction method optimized from many available ones, including Ridge Regression, gBLUP, compressed BLUP, Bayesian LASSO, Bayes A, B, Cpi and many more. Users can also use the platform to conduct data analyses with any methods of their choice. It is expected that extensive usage of MMAP would enrich the training data, which in turn results in continual improvement of the identification of the best method for use with particular traits.

Availability and implementation: The MMAP user manual, tutorials and example datasets are available at http://zzlab.net/MMAP.

Contact: zhiwu.zhang@wsu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Accurate prediction of phenotypes from genotypes is one of the ultimate goals of genomic research, so that a medical treatment could be optimized to improve human and animal health, and breeding could be revamped to increase animal and plant production. Before a complete identification of genes underlying a particular trait of interest through techniques, such as genome-wide association study (GWAS), genomic prediction or genomic selection (GS), is a practical shortcut that plays a critical role in animal and plant breeding to predict phenotypes from genotypes without knowledge of where those genes are. Many statistical methods and computing tools have been developed to conduct GWAS and GS, including the common methods and tools for both GWAS and GS (Endelman, 2011; Kim *et al.*, 2019; Lipka *et al.*, 2012; Pérez and De Los Campos, 2004; Tang *et al.*, 2016). However, there is a fundamental difference between GWAS and GS. There are minimal interactions between GWAS methods and traits. For example, for some traits, all methods perform the same, either successfully detecting a major gene or failing to detect any association when either sample size or gene effects are too small. For other traits, these methods perform differently. Some methods detect more associations than others. The magnitude

of the statistical power varies from trait to trait. However, the orders of methods rarely change. The situation is different for GS. The order of GS methods varies from trait to trait depending on the genetic architecture of the traits (Wang *et al.*, 2018). For polygenic traits, genomic Best Linear Unbiased Prediction (gBLUP) performs better than SUPER BLUP. For Mendelian traits, the opposite is true. For traits with low heritability, compressed BLUP performs better than Bayesian LASSO, and the reverse applies for traits with high heritability (Wang *et al.*, 2018). It is challenging to choose a suitable method for a particular trait. Researchers have to examine a variety of methods before reaching a desirable prediction accuracy. Additional challenges, such as installation, steep learning curves and required computational resources intimidate many biological researchers. There is a critical need to develop a free computing platform that would automatically identify the best method and conduct analyses for users with minimal effort, such as uploading and downloading genotype and phenotype data. Herein, we present a cloud computing platform to solve the problem by Mining the Maximum Accuracy of Predicting phenotypes from genotypes (MMAP).

2 Method and implementation

MMAP is a knowledge-based cloud computing platform that continuously gains knowledge over time during application (Fig. 1a). It currently implements eight GS methods and a mining system to identify the best prediction method for a particular trait (Fig. 1b and Supplementary Fig. S1). The eight GS methods include gBLUP, compressed BLUP, SUPER BLUP, Bayes A, Bayes B, Bayes C, Bayes Cpi and Bayesian LASSO (Fig. 1c). The mining system consists of an existing database and an interactive and dynamic evaluation (IDE) across GS methods and datasets. The current database contains the essential characteristics of over a hundred datasets and their prediction accuracy using these GS methods (Supplementary Tables S1 and S2).

The essential characteristics include sample size, genome size, number of markers, linkage disequilibrium decade, heritability and parameters of principal component analysis. The IDE contains an initial evaluation of prediction accuracy using gBLUP, which was reported to have the highest prediction accuracy on substantial traits, especially the polygenic traits. The essential characteristics and the initial prediction accuracy using gBLUP are used as the input to predict the next methods with two objectives. First, the next method has a high probability of being the best among the implemented GS methods. The other objective is to provide relevant information to find the method that has the highest chance to be the next best GS method. We implemented single trait GS methods and the IDE using the efficient C/C++ programming language and incorporated several highly efficient open-source mathematical operation and optimization libraries. The computation is distributed across multiple nodes on our networked Linux High-Performance Computing cluster.

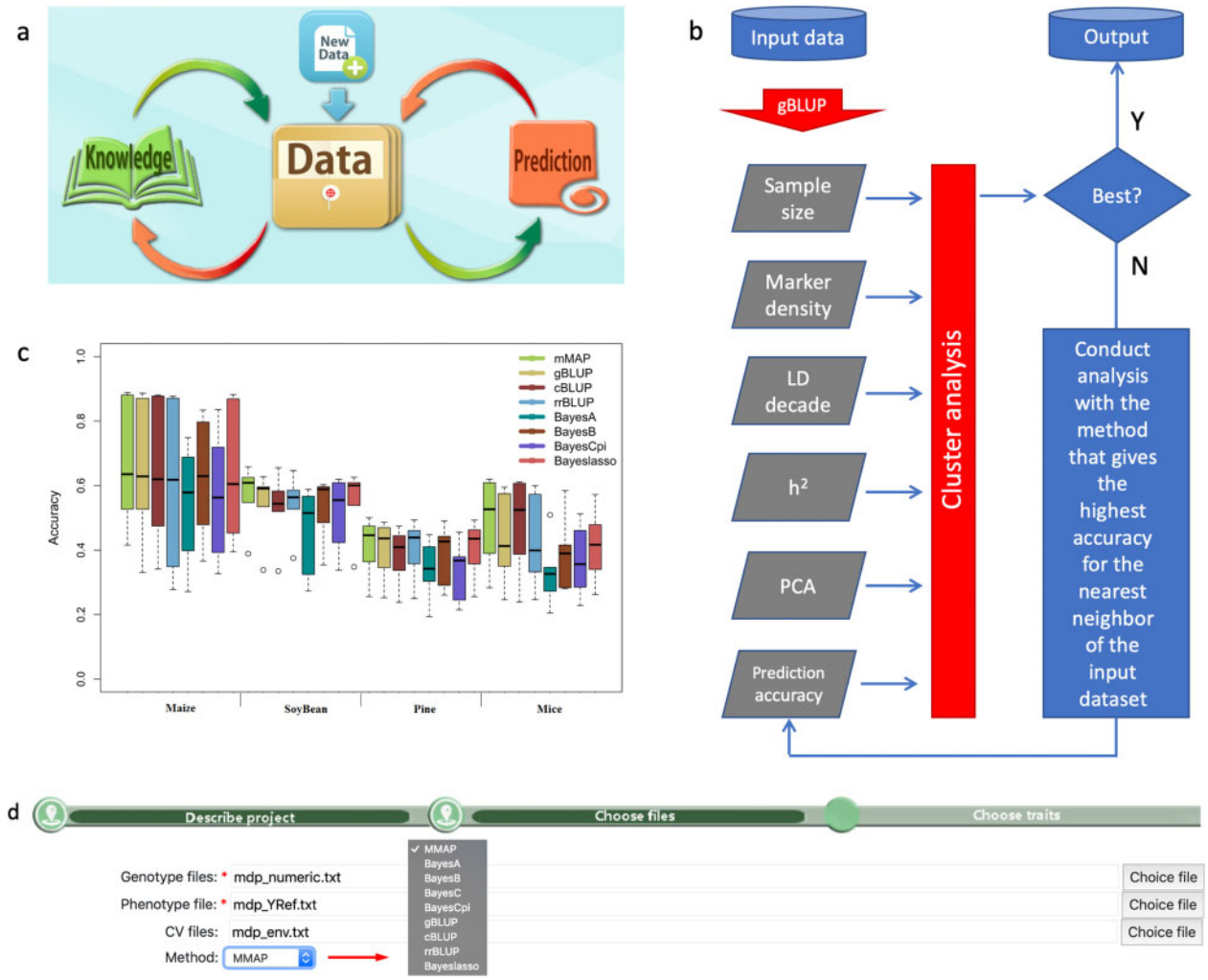

Fig. 1. The workflow and performance of MMAP. As a cloud computing platform, MMAP integrate existing knowledge and interactively search for the best GS method for a particular dataset (**a**). The search is based on the characteristics of the input data and IDE initiated with the gBLUP method (**b**). MMAP has the highest average prediction accuracy (**c**) with minimal effort required for uploading phenotypic data, genotypic data and covariable data (**d**)

3 Workflow and user interface

MMAP has four tabs to navigate in the platform operations, including User Account, File tab, Project tab and user manual. The File tab navigates to upload input data for phenotypes, genotypes and covariate variables. The Project tab specifies the input data and provides the link to download prediction results (Fig. 1d).

4 Results and discussion

The prediction methods implemented in MMAP can be selected specifically to generate identical or similar results depending on methods using random sampling or not (Supplementary Figs S2 and S3). Under automatic mode, MMAP took an average of 2.93 times to find the best method at 91% success, and 96% success at identifying at least one of the top three methods (Supplementary Fig. S4). Among the multiple traits across four species examined, MMAP had the highest average prediction accuracy compared to all implemented GS methods (Fig. 1c and Supplementary Figs S5 and S6).

5 Conclusion

MMAP is a cloud computing platform with a user-friendly interface that requires minimal effort to conduct GS without an explicit understanding of a variety of methods and computing tools. Researchers are entirely liberated from software installation, training with steep learning curves and allocating appropriate computing resources with this free platform.

Acknowledgements

The authors thank Cari Park for copyediting the manuscript.

Funding

This project was partially supported by the National Science Foundation [Award # DBI 1661348]; the USDA NIFA [Hatch project 1014919, Award #s 2016-68004-24770, 2018-70005-28792, 2019-67013-29171 and 2020-67021-32460]; the Washington Grain Commission [Endowment and Award #s 126593, 134574]; the PhD start-up Foundation Project of Jilin Agricultural Science and Technology University [20180337]; and the Digital Agriculture key discipline Foundation of Jilin Province.

Conflict of Interest: none declared.

Data availability

All required links to the program and data are provided in this manuscript.

References

1. Endelman,J. (2011) Ridge regression and other kernels for genomic selection in the R package rrBLUP. *Plant Genome*, **4**, 250–255.
2. Kim,B. *et al.* (2019) GWASpro: a high-performance genome-wide association analysis server. *Bioinformatics*, **35**, 2512–2514.
3. Lipka,A.E. *et al.* (2012) GAPIT: genome association and prediction integrated tool. *Bioinformatics*, **28**, 2397–2399.
4. Pérez,P. and De Los Campos,G. (2004) *BGLR: A Statistical Package for Whole Genome Regression and Prediction*. R package version 1. 0.2. 2013
5. Tang,Y. *et al.* (2016) GAPIT version 2: an enhanced integrated tool for genomic association and prediction. *Plant J.*, **9**(2), 1–9.
6. Wang,J. *et al.* (2018) Expanding the BLUP alphabet for genomic prediction adaptable to the genetic architectures of complex traits. *Heredity (Edinb)*, **121**, 648–662.