2018

# GWASpro: A High-Performance Genome-Wide Association Analysis Server

Bongsong KIm
*Noble Research Institute*

Xinbin Dai
*Noble Research Institute*

Wenchao Zhang
*Noble Research Institute*

Zhaohong Zhuang
*Noble Research Institute*

Darlene L. Sanchez
*Texas A&M AgriLife Research*

*See next page for additional authors*

# GWASpro: A High-Performance Genome-Wide Association Analysis Server

**Abstract**

We present GWASpro, a high-performance web server for the analyses of large-scale genome-wide association studies (GWAS). GWASpro was developed to provide data analyses for large-scale molecular genetic data, coupled with complex replicated experimental designs such as found in plant science investigations, and to overcome the steep learning curves of existing GWAS software tools. GWASpro supports building complex design matrices, by which complex experimental designs that may include replications, treatments, locations and times, can be accounted for in the linear mixed model (LMM). GWASpro is optimized to handle GWAS data that may consist of up to 10 million markers and 10,000 samples from replicable lines or hybrids. GWASpro provides an interface that significantly reduces the learning curve for new GWAS investigators.

**Disciplines**

Agriculture | Bioinformatics | Molecular Genetics | Plant Breeding and Genetics

**Comments**

This is a manuscript of an article published as Kim, Bongsong, Xinbin Dai, Wenchao Zhang, Zhaohong Zhuang, Darlene L. Sanchez, Thomas Lübberstedt, Yun Kang et al. "GWASpro: A High-Performance Genome-Wide Association Analysis Server." *Bioinformatics* (2018). doi: 10.1093/bioinformatics/bty989.

**Authors**

Bongsong KIm, Xinbin Dai, Wenchao Zhang, Zhaohong Zhuang, Darlene L. Sanchez, Thomas Lubberstedt, Yun Kang, Michael Udvardi, William D. Beavis, Shizhong Xu, and Patrick X. Zhao

*Bioinformatics*, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Applications Note

*Genetics and Population Analysis*

# GWASpro: A High-Performance Genome-Wide Association Analysis Server

Bongsong Kim[1], Xinbin Dai[1], Wenchao Zhang[1], Zhaohong Zhuang[1], Darlene L. Sanchez[2], Thomas Lübberstedt[3], Yun Kang[1], Michael Udvardi[1], William D. Beavis[3], Shizhong Xu[4,*] and Patrick X. Zhao[1,*]

[1]Noble Research Institute, Ardmore, OK 73401, USA, [2]Texas A&M AgriLife Research, Beaumont, TX 77713, USA, [3]Department of Agronomy, Iowa State University, Ames, IA 50011, USA, [4]Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on XXXXX; revised on XXXXX; accepted on XXXXX

## Abstract

**Summary:** We present GWASpro, a high-performance web server for the analyses of large-scale genome-wide association studies (GWAS). GWASpro was developed to provide data analyses for large-scale molecular genetic data, coupled with complex replicated experimental designs such as found in plant science investigations, and to overcome the steep learning curves of existing GWAS software tools. GWASpro supports building complex design matrices, by which complex experimental designs that may include replications, treatments, locations and times, can be accounted for in the linear mixed model (LMM). GWASpro is optimized to handle GWAS data that may consist of up to 10 million markers and 10,000 samples from replicable lines or hybrids. GWASpro provides an interface that significantly reduces the learning curve for new GWAS investigators.

**Availability and implementation:** GWASpro is freely available at https://bioinfo.noble.org/GWASPRO.

**Contact:** Shizhong Xu, shizhong.xu@ucr.edu; Patrick X. Zhao, pzhao@noble.org

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

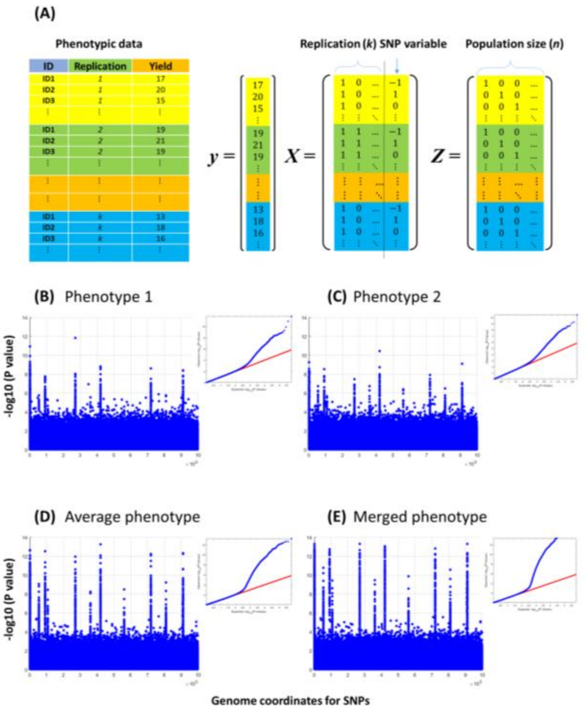
## 1 Introduction

Genome-wide association studies (GWAS) for crop improvements often confront significant challenges related to complex experimental designs and large data sets; there is a need for new GWAS analysis software that can address replicated phenotypic data related to complex experimental designs involving multiple environments along with a large-scale molecular marker data. Popular GWAS software tools (Bradbury, et al., 2007; Lipka, et al., 2012) are confined to a single population and using linear mixed models (LMMs), in particular the QK model, which incorporates both a population stratification structure (Q) matrix and a kinship (K) matrix (Yu, et al., 2006). Recently, several modified models, such as the compressed mixed linear model (CMLM) (Zhang, et al., 2010), multi-locus mixed model (MLMM) (Segura, et al., 2012), FarmCPU (Liu, et al., 2016), and the integration of Kruskal-Wallis test with empirical Bayes (pkWemEB) (Ren, et al., 2018), were proposed to achieve fast computation and high statistical power. However, all of the above software tools lack the capacity to account for the phenotypic variance across environments (Korte and Farlow, 2013). To solve this problem, we present GWASpro, a web-based platform that provides online GWAS data analysis services. GWASpro supports building complex design matrices to account for replicated phenotypic observations (years, treatments, locations, and/or replications), which advances the QK model toward better quantitative trait loci (QTL) mapping resolutions. GWASpro is capable of handling a large-scale data set consisting of up to 10 million markers and 10,000 samples representing the replicable genotypes.

## 2 Methods Features and Implementation

### 2.1. Design matrices

GWASpro supports flexible building design matrices for the LMM. **Figure 1A** shows how the design matrices for genotypic data consisting of *m* markers and *n* individuals with *k* replications are arranged.



| ID | Replication | Yield |
|---|---|---|
| ID1 | 1 | 17 |
| ID2 | 1 | 20 |
| ID3 | 1 | 15 |
| ⋮ | ⋮ | ⋮ |
| ID1 | 2 | 19 |
| ID2 | 2 | 21 |
| ID3 | 2 | 19 |
| ⋮ | ⋮ | ⋮ |
| ⋮ | ⋮ | ⋮ |
| ⋮ | ⋮ | ⋮ |
| ID1 | k | 13 |
| ID2 | k | 18 |
| ID3 | k | 16 |
| ⋮ | ⋮ | ⋮ |

$$y = \begin{pmatrix} 17 \\ 20 \\ 15 \\ \vdots \\ 19 \\ 21 \\ 19 \\ \vdots \\ \vdots \\ \vdots \\ 13 \\ 18 \\ 16 \\ \vdots \end{pmatrix} \quad X = \begin{pmatrix} 1 & 0 & \cdots & -1 \\ 1 & 0 & \cdots & 1 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & -1 \\ 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & -1 \\ 1 & 0 & \cdots & 1 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \end{pmatrix} \quad Z = \begin{pmatrix} 1 & 0 & 0 & \cdots \\ 0 & 1 & 0 & \cdots \\ 0 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \\ 1 & 0 & 0 & \cdots \\ 0 & 1 & 0 & \cdots \\ 0 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots \\ 0 & 1 & 0 & \cdots \\ 0 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Figure 1. (A) Example data and related design matrices for y, X, and Z, where y is the vector for phenotype, X is the design matrix for the fixed effect, and Z is the design matrix for the random genetic effect. (See Equations 1 and 2 in **Supplementary Material A**); (B) Manhattan plots and QQ plots, obtained using phenotype 1; (C) Manhattan plots and QQ plots, obtained using phenotype 2; (D) Manhattan plots and QQ plots, obtained using the average phenotype; (D) Manhattan plots and QQ plots, obtained using the merged phenotype.

### 2.2. Efficient computing for large-scale GWAS

In GWASpro, working procedures include building a kinship matrix, fitting the LMM, and performing Wald test for calculating *p*-values (see **Supplementary Material A**). GWASpro implements a distributed parallel-computing engine that can effectively utilize ~1,000 CPU cores and ~10TB RAM (**Figure S1**). We also implemented a multithreading and resumable data-uploading module, utilizing HTML5 protocol for robust and fast data transfer.

### 2.3. Genomic control for adjusting inflated *p*-values

We observed genomic (*p*-value) inflations given a replication factor in our simulation study (see **Section 3.1**) and the case study 3. To address this, GWASpro introduces a genomic correction function, by which the inflated *p*-values are adjusted using the genomic inflation factor ($\lambda_{GC}$) as demonstrated by (Devlin and Roeder, 1999; Devlin, et al., 2001; van Iterson, et al., 2017; Voorman, et al., 2011).

### 2.4. Input

GWASpro automatically establishes the LMM with required inputs including a genotypic file, a phenotypic file, and variable names with properties (categorical/numerical). Users are responsible for imputations of markers. The upload of kinship matrix is optional as it can be calculated using the genotypic matrix. Missing phenotypic records are automatically excluded. Users can either directly upload data files from a local computer or specify the URLs of user input data, including data sharing URLs of Google Drive and Dropbox for remote downloading using http/https/ftp protocols.

### 2.5. Output

The job queue management system in GWASpro assigns a user a unique session ID, which can be used to track the job progress and download final results. The GWASpro returns original *p*-values, adjusted *p*-values based on genomic control, QQ plot, and Manhattan plot.

## 3 Results and Discussions

### 3.1 Simulation study: Assessing QTL mapping resolution

Our simulated data set mimics a situation in which two identical plant populations (A and B) are grown in two environments (see **Supplementary Material C**). We prepared four phenotypic data sets: phenotype 1 (Figure 1B), phenotype 2 (Figure 1C), average phenotype (Figure 1D), and merged phenotype (Figure 1E). Heritability for each population was adjusted to 0.5. The principle of this simulation was introduced in (Kim, 2017). The resulting Manhattan plots reveal that Figure 1E produces the best QTL resolution with the highest QTL peaks and trivial background inflation, followed by Figure 1D. To compare the analysis performance between Figures 1D and 1E, the receiver operating characteristic (ROC) curves were drawn (**Figure S5**). The area under the curves (AUC) for Figures 1D and 1E were 0.9178 and 0.9276, respectively. This supports that Figure 1E shows better QTL resolution. This is a novel benefit of GWASpro, suggesting that accounting for the phenotypic variabilities can improve QTL mapping resolution by reducing the missing heritability (Korte and Farlow, 2013).

### 3.2 Case study 1: Comparing GWASpro, GAPIT, and PEPIS

We analyzed the thousand-grain weight (as phenotype) for the IMF2 rice population (Hua, et al., 2003; Hua, et al., 2002) using GAPIT (Lipka, et al., 2012), PEPIS (Zhang, et al., 2016), and GWASpro (**Figure S2)**. All significant peaks were consistent. In particular, GAPIT and GWASpro yielded similar plot outlines with different *p*-value scales, which indicates that different *p*-value thresholds must be applied to the GAPIT and GWASpro results. GAPIT, PEPIS, and GWASpro have different characteristics: GAPIT should be used for a single population in the additive QK model; PEPIS for a single population accounting for additive, epistasis, and dominant effects in the K model; and GWASpro for either a single or replicated genotypes in either the K or QK model.

### 3.3 Case study 2: *Medicago truncatula* data

Kang et al. (Kang, et al., 2015) published GWAS results for leaf size and shoot biomass weight traits with a *Medicago truncatula* HapMap population consisting of 220 accessions with 1,810,466 SNPs using TASSEL (see **Supplementary Material B**). We reanalyzed the same data set using GWASpro and TASSEL and compared their results. The resulting Manhattan plots and QQ plots are very similar to each other (**Figure S3**).

### 3.4 Case study 3: Maize data

Sanchez et al. (Sanchez, et al., 2018) published GWAS results with three replicated populations (302 maize accessions in each population) using GAPIT. We analyzed the same data using GWASpro. GAPIT and GWASpro produced different results because the GWASpro results were obtained directly using the replicated phenotypic data, whereas the GAPIT results were obtained using the breeding values (BVs) predicted from the replicated genotypes. The authors used the BVs for GWAS analyses because GAPIT is not capable of handling the replications. GAPIT required twice fitting the LMMs for BV prediction and GWAS, which might cause LMM overfitting. With GWASpro, this problem can be avoided.  The genomic inflation was observed in the GWASpro results, which is common given replicated genotypes (Ehret, 2010; van

Iterson, et al., 2017; Voorman, et al., 2011). To address this issue, the population stratification resulting from the principle component analysis (PCA) was first accounted for then, *p*-vaules were adjusted by the genomic control (see **Section 2.3**) in our analysis. **Figure S4** compares the results obtained by GWASpro and GAPIT.

### 3.5 Performance test

We performed benchmark tests of GWASpro by measuring runtimes (**Table S1**) given the various sizes of data (1 million, 3 million, 5 million, 10 million SNPs; 1k, 3k, 5k individuals). **Figure S6** summarizes that the runtime generally increases following $O(n^2m)$, where *n* is sample size and *m* is marker size.

## 4 Conclusion

GWASpro is an online platform for GWAS analysis that does not require the hassles of software installation and maintenance. The parallel computing engine allows GWASpro to quickly analyze a large-scale data set. In GWASpro, the QK model is implemented for unbiased QTL mapping by accounting for the kinship matrix (K) and population stratification (Q) (Yu, et al., 2006). GWASpro can address replicated phenotypic data, which are typically from self-pollinating plant species. Our simulation data sets demonstrate that GWASpro captures the amplified QTL signals when the gene-environment interactions in multiple replications are in similar patterns. Our Maize data sets demonstrate that GWASpro captures QTLs by accounting for the phenotypic variabilities across different environments. The environmental factors are crucial to identify robust environment-resistant QTL (Palomeque, et al., 2010; Xavier, et al., 2018). In addition, GWASpro supports breeding value estimation, which is introduced in **Supplementary material D**.

## Funding

This work was supported by the National Science Foundation collaborative research grant award DBI-1458597 to P.X.Z. and DBI-1458515 to S.X.; and by partial funding support from the Noble Research Institute to P.X.Z.; the North Central Soybean Research Program, Baker Center for Plant Breeding, USDA-NIFA project IOW04314 and the GF Sprague Endowment of the Agronomy Department at Iowa State University to W.D.B.

*Conflict of Interest:* none declared.

## References

Bradbury, P.J., *et al.* TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 2007;23(19):2633-2635.

Devlin, B. and Roeder, K. Genomic control for association studies. *Biometrics* 1999;55(4):997-1004.

Devlin, B., Roeder, K. and Wasserman, L. Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol* 2001;60(3):155-166.

Ehret, G.B. Genome-wide association studies: contribution of genomics to understanding blood pressure and essential hypertension. *Curr Hypertens Rep* 2010;12(1):17-25.

Hua, J., *et al.* Single-locus heterotic effects and dominance by dominance interactions can adequately explain the genetic basis of heterosis in an elite rice hybrid. *Proceedings of the National Academy of Sciences of the United States of America* 2003;100(5):2574-2579.

Hua, J.P., *et al.* Genetic dissection of an elite rice hybrid revealed that heterozygotes are not always advantageous for performance. *Genetics* 2002;162(4):1885-1895.

Kang, Y., *et al.* Genome-wide association of drought-related and biomass traits with HapMap SNPs in Medicago truncatula. *Plant, Cell & Environment* 2015;38(10):1997-2011.

Kim, B. Hierarchical Association Coefficient Algorithm: New Method for Genome-Wide Association Study. *Evolutionary bioinformatics online* 2017;13:1176934317713004.

Korte, A. and Farlow, A. The advantages and limitations of trait analysis with GWAS: a review. *Plant methods* 2013;9:29.

Lipka, A.E., *et al.* GAPIT: genome association and prediction integrated tool. *Bioinformatics* 2012;28(18):2397-2399.

Liu, X., *et al.* Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. *PLoS Genet* 2016;12(2):e1005767.

Palomeque, L., *et al.* Validation of mega-environment universal and specific QTL associated with seed yield and agronomic traits in soybeans. *Theor Appl Genet* 2010;120(5):997-1003.

Ren, W.L., *et al.* pKWmEB: integration of Kruskal-Wallis test with empirical Bayes under polygenic background control for multi-locus genome-wide association study. *Heredity (Edinb)* 2018;120(3):208-218.

Sanchez, D.L., *et al.* Genome-wide association studies of doubled haploid exotic introgression lines for root system architecture traits in maize (Zea mays L.). *Plant science : an international journal of experimental plant biology* 2018;268:30-38.

Segura, V., *et al.* An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat Genet* 2012;44(7):825-830.

van Iterson, M., *et al.* Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biology* 2017;18(1):19.

Voorman, A., *et al.* Behavior of QQ-plots and genomic control in studies of gene-environment interaction. *PLoS One* 2011;6(5):e19416.

Xavier, A., *et al.* Genome-Wide analysis of grain yield stability and environmental interactions in a multiparental soybean population. *G3: Genes, Genomes, Genetics* 2018;8(2):519-529.

Yu, J., *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 2006;38(2):203-208.

Zhang, W., *et al.* PEPIS: A pipeline for estimating epistatic effects in quantitative trait locus mapping and genome-wide association studies. *PLoS computational biology* 2016;12(5):e1004925.

Zhang, Z., *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* 2010;42(4):355-360.