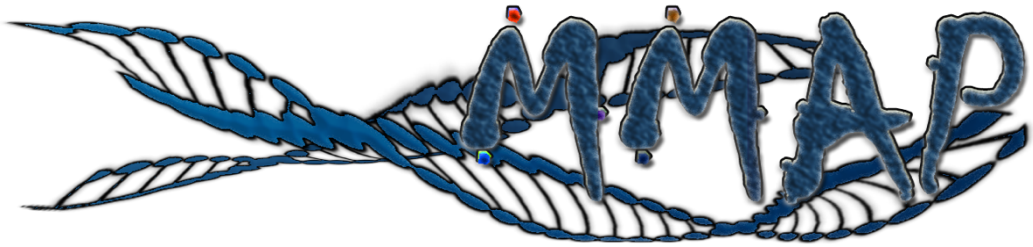
User Manual for



Mining the Maximum Accuracy of Prediction

(Version 1.0)

Last updated on June 2, 2020

Zhiwu Zhang Laboratory

WASHINGTON STATE UNIVERSITY

Disclaimer: While extensive testing has been performed by the Zhiwu Zhang Lab at Washington State University, results are, in general, reliable, correct or appropriate. However, results are not guaranteed for any specific set of data. We strongly recommend that users validate MMAP results with other software packages, such as GAPIT, rrBLUP, and BGLG.

Support documents: Extensive support documents, including this user manual, demo data, and results, are available at MMAP website hosted by Zhiwu Zhang Laboratory: http://zzlab.net/MMAP.

Questions and comments: Please email questions and comments to You Tang by email: 1098118439@qq.com.

MMAP Team: Wei Huang, Ping Zheng, Zhenhai Cui, Zhuo Li, Yifeng Gao, Helong Yu, You Tang, Xiaohui Yuan, and Zhiwu Zhang

Citation: Wei Huang, Ping Zheng, Zhenhai Cui, Zhuo Li, Yifeng Gao, Helong Yu, You Tang, Xiaohui Yuan, and **Zhiwu Zhang***. **MMAP: A Cloud Computing Platform for Mining the Maximum Accuracy of Predicting Phenotypes from Genotypes.** Bioinformatics, 2020, https://doi.org/10.1093/bioinformatics/btaa824.

Support: The MMAP project is partially supported by USDA, DOE, NSF, the Agricultural Research Center at Washington State University, Jilin Agricultural Science and Technology University and Washington Grain Commission.

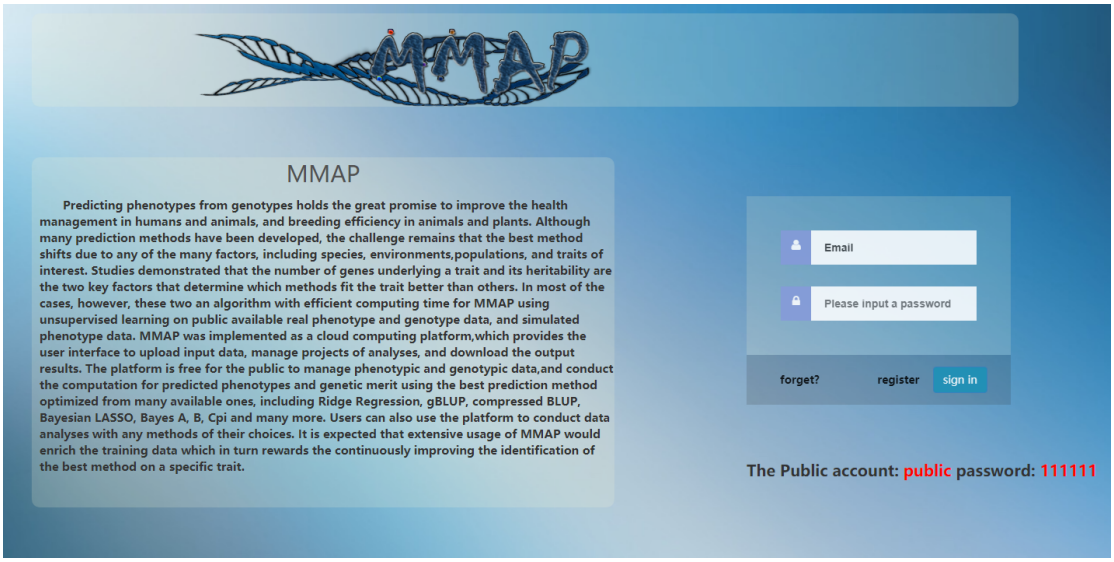
TABLE OF CONTENTS

INTRODUCTION

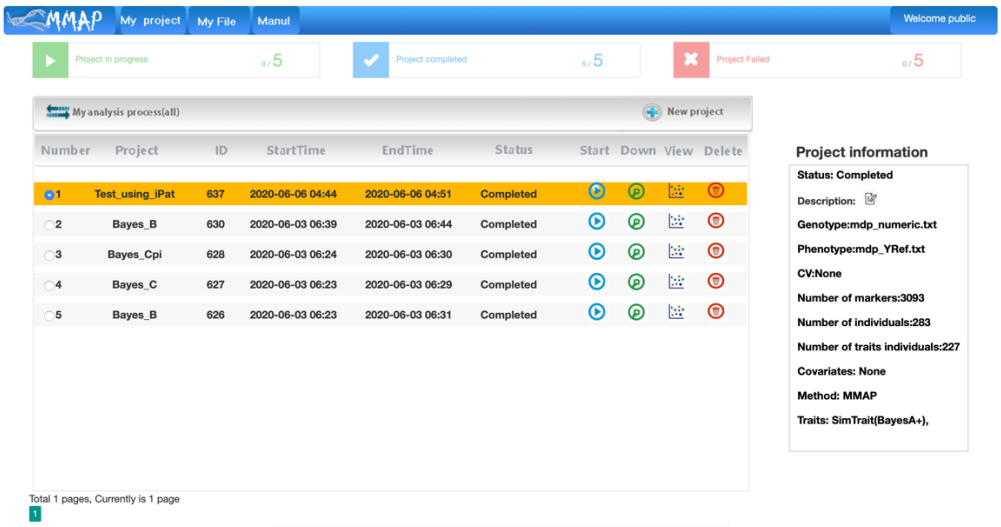
Accurate prediction of phenotypes from genotypes is the ultimate goal of genomic research, so that medical treatment could be optimized to improve human and animal health, and breeding could be revamped to increase animal and plant production. Before a complete identification of genes underlying a particular 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 field 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(Kim *et al.* 2018) and GS(Pérez and De Los Campos 2004; Endelman 2011; Lipka *et al.* 2012; Tang *et al.* 2016). However, there is a fundamental difference between GWAS and GS. There are 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, their orders barely 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. 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. 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 a user with minimal effort, such as uploading data for traits of interest. Herein, we present a cloud computing platform to solve the problem by mining the maximum accuracy of predicting phenotypes (MMAP) from genotypes.

1. GETTING START

Go to http://zzlab.net/MMAP and click MMAP icon on the top to start the login page. You can login via the public account, or your own account. A valid email is required for registration. The benefits to login individual accounts include privacy, storage of data and results for longer period, and email notification for computational status.



MMAP web interface contains four taps: 1) MMAP login page as demonstrated above, 2) Project; 3) File, and 4) User manual. The user manual tab link to the help document in PDF format. There are two ways reach the login page for using a different account. One is logout through the account tap on the top right. The other is to click the MMAP tab. The Project and File tabs will be described in next two sections. The project tab is displayed after login:

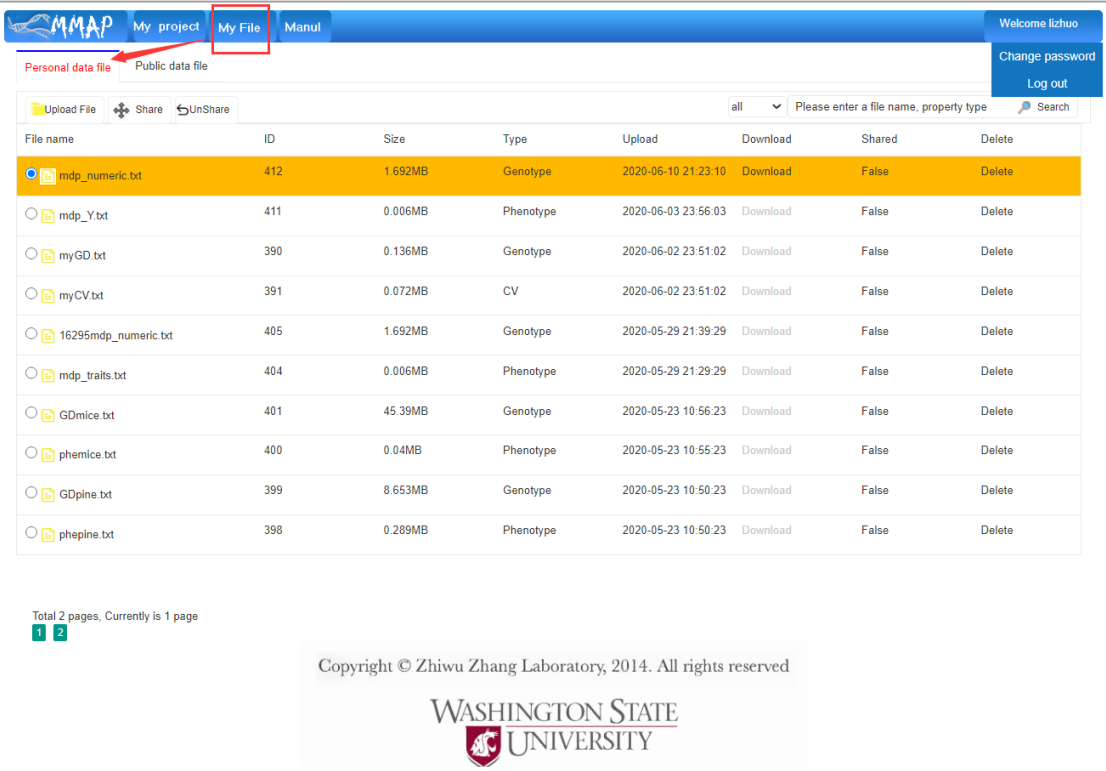


Copyright © Zhiwu Zhang Laboratory, 2014. All rights reserved
WASHINGTON STATE UNIVERSITY

2. FILE OPERATIONS

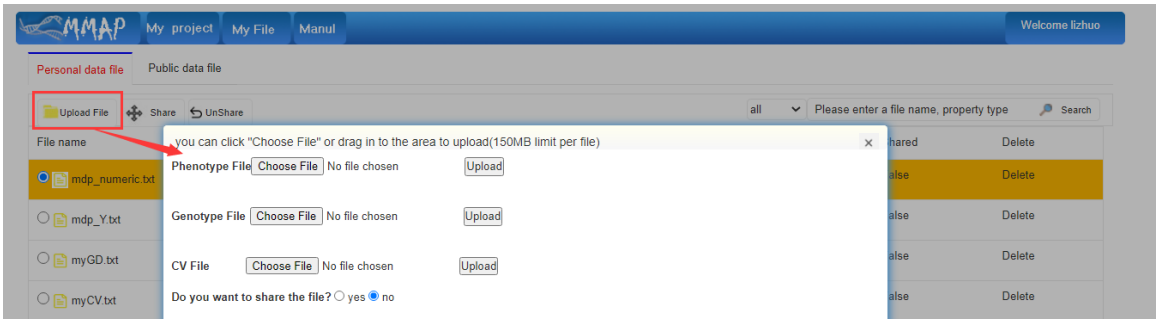
2.1 Upload Files

• First click on the "My File" button in the title bar to enter the file operation page.



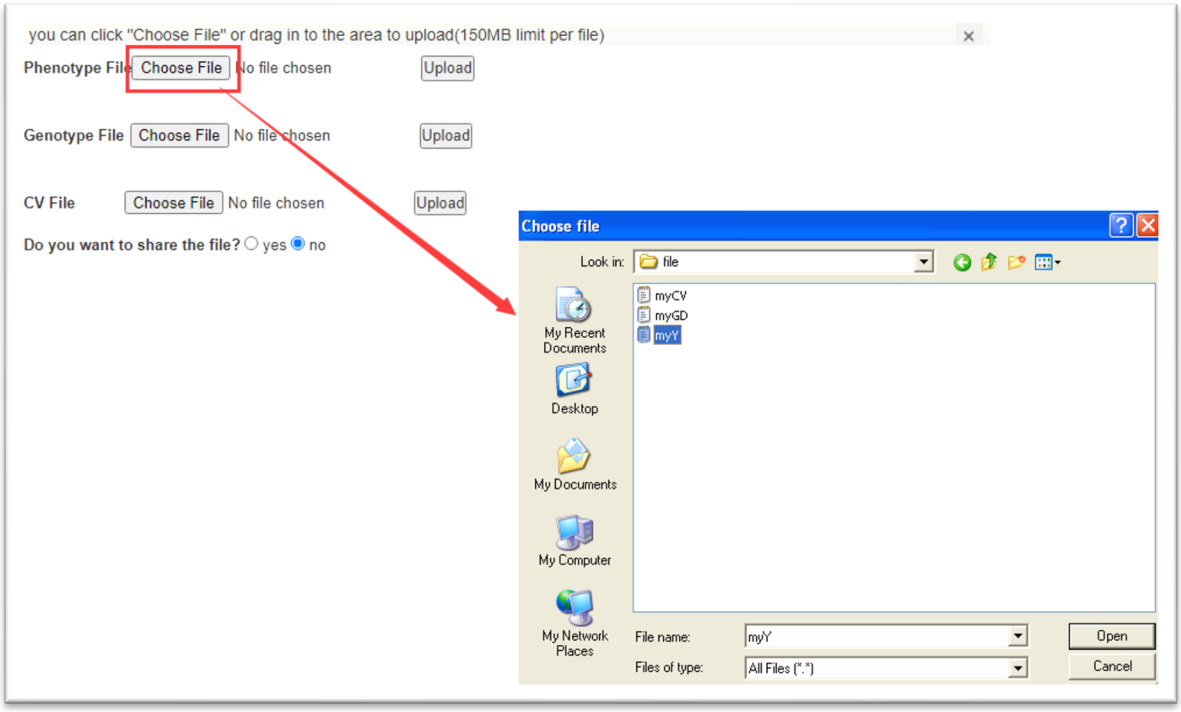
Copyright © Zhiwu Zhang Laboratory, 2014. All rights reserved
WASHINGTON STATE UNIVERSITY

• Click "Upload File".



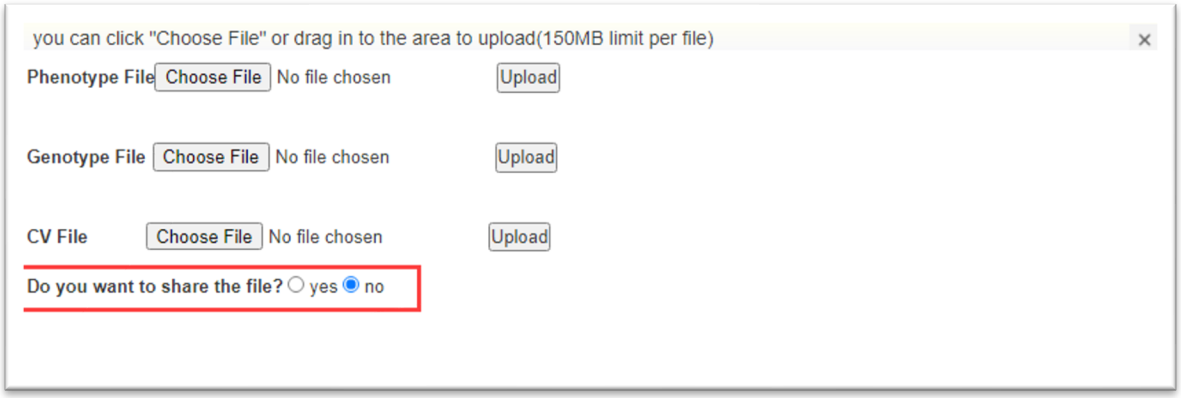
• Click on "choose file" after the type of file you want to upload to select the file you want to upload.

• Here is a phenotype file as an example.

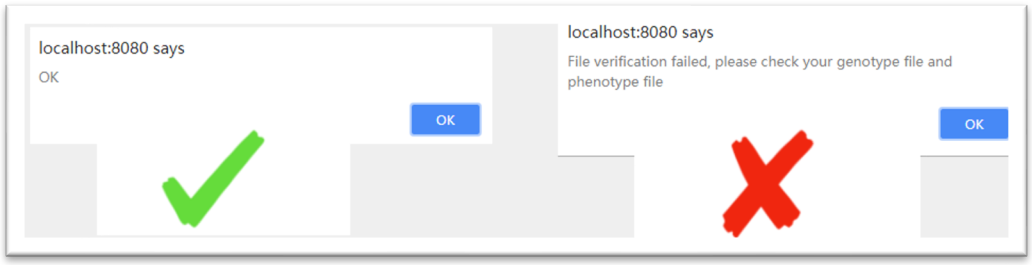


• After the file selection is completed, click "Upload" to upload the file you selected.

• Also, at the bottom of the page you can choose whether or not to share your uploaded files.

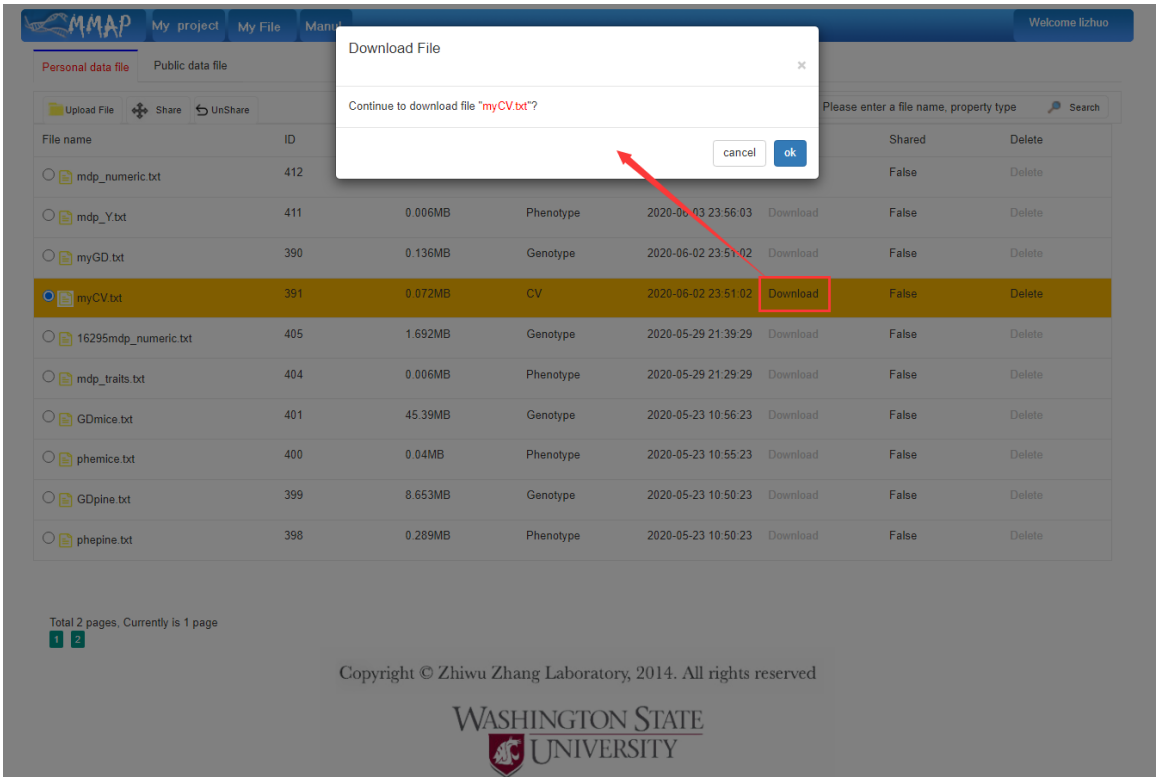


• After uploading the file, the system will automatically detect if your file is correct.



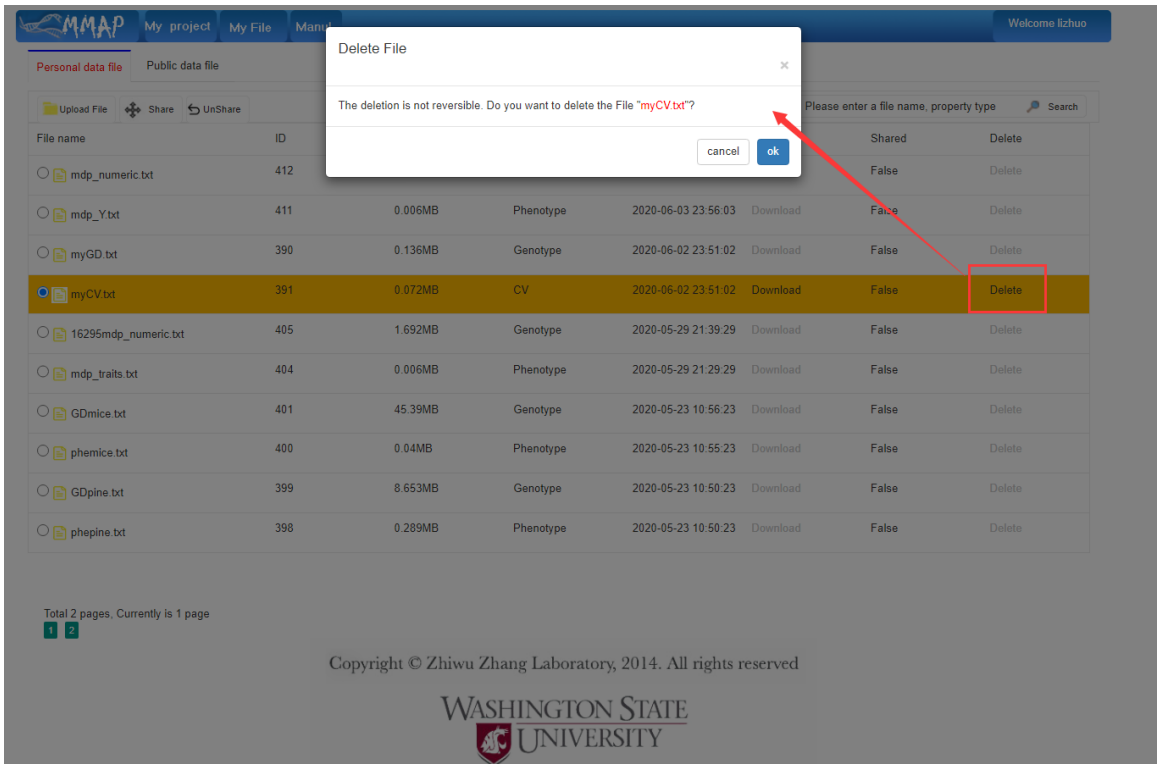
2.2 Download File

On the “My File” page, click the “Download” button behind the highlighted file to download the file you selected.



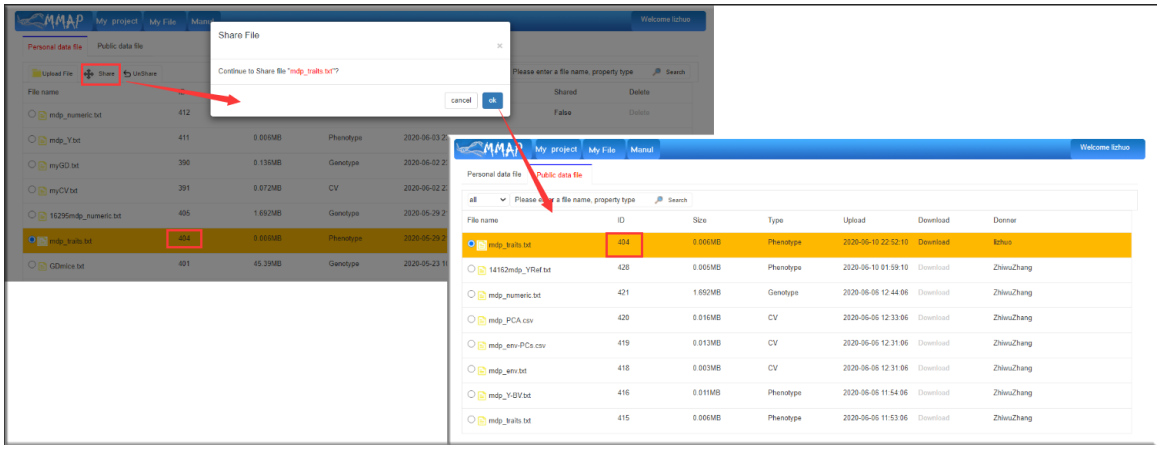
2.3 Delete Files

On the "My File" page, click on the "delete" button behind the highlighted file to delete the file you selected.



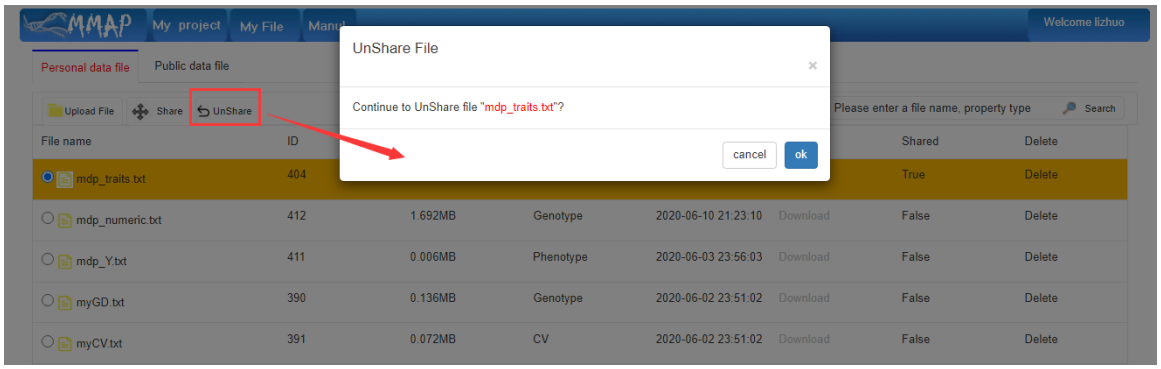
2.4 Share files

On the ”My File” page, click on the “share” button behind the highlighted file to share your selected files. Shared files can be viewed on the "public date file" page



2.5 Unshare files

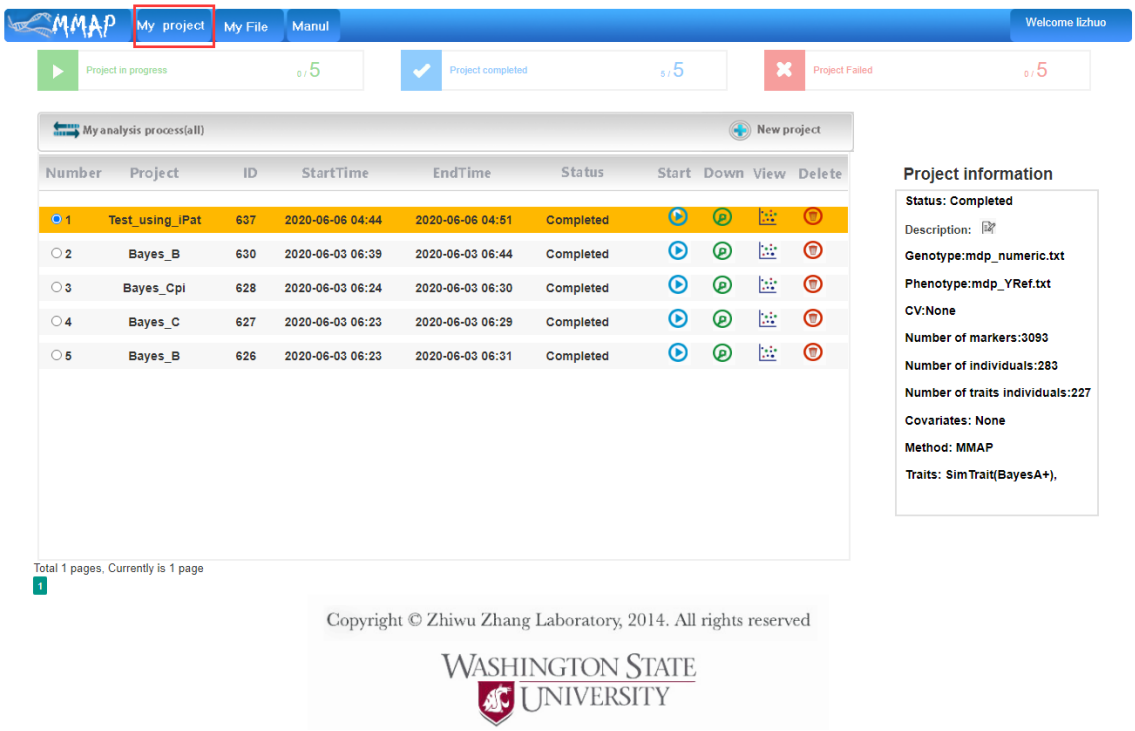
On the ”My File” page, click on the “Unshare” button behind the highlighted file to unshare your selected files


MMAP
My project
My File
Welcome lizhuo
Personal data file
Public data file
Upload File
Share
UnShare
Please enter a file name, property type
Search
File name
ID
Shared
Delete
mdp_traits.txt
404
True
Delete
mdp_numeric.txt
412
1.692MB
Genotype
2020-06-10 21:23:10
Download
False
Delete
mdp_Y.txt
411
0.006MB
Phenotype
2020-06-03 23:56:03
Download
False
Delete
myGD.txt
390
0.136MB
Genotype
2020-06-02 23:51:02
Download
False
Delete
myCV.txt
391
0.072MB
CV
2020-06-02 23:51:02
Download
False
Delete
UnShare File
Continue to UnShare file "mdp_traits.txt"?
cancel
ok

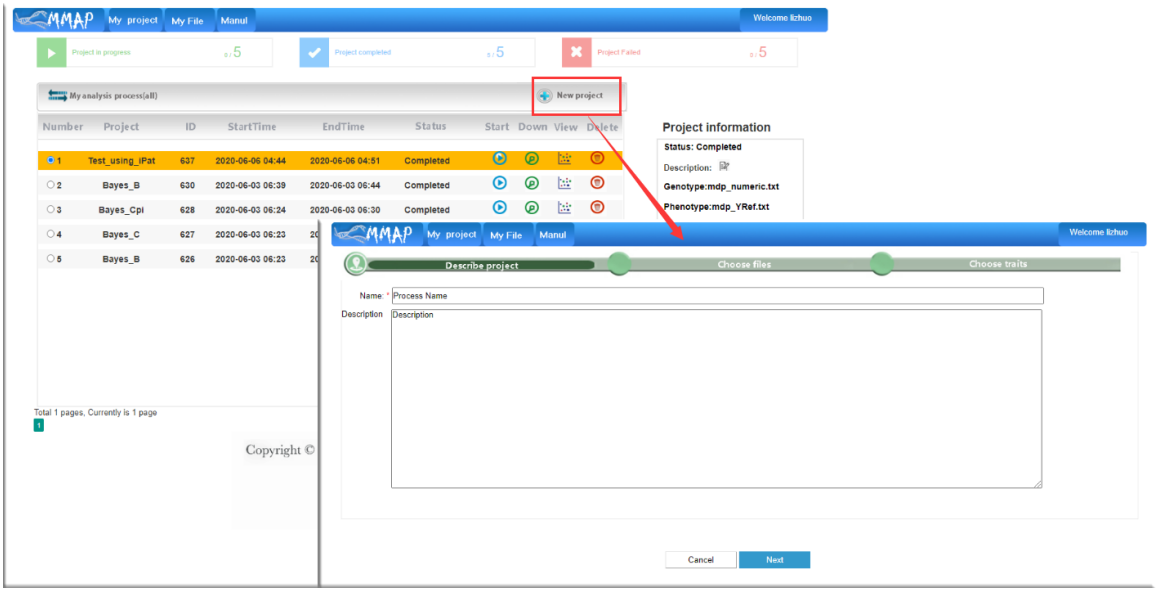
3. PROJECT OPERATIONS

3.1 How to Create A New Project

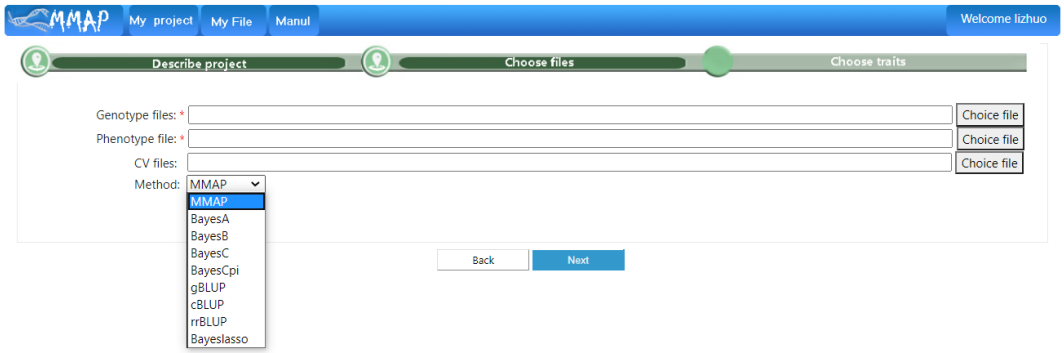
• First click on the "my project" button in the title bar to go to the home page.



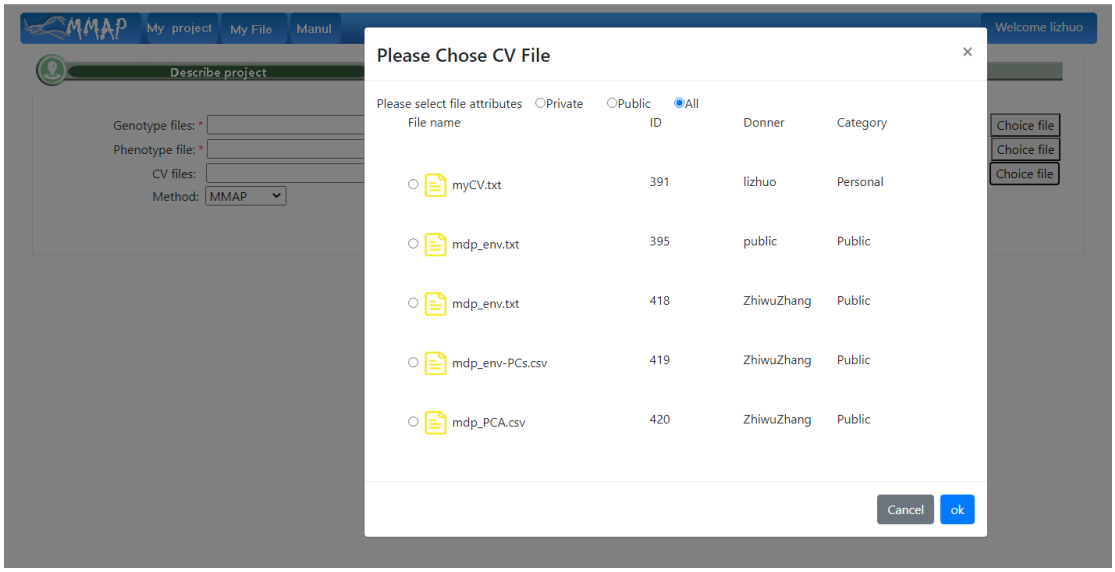
• Click on the "New Project" button on the homepage to start creating a new process.



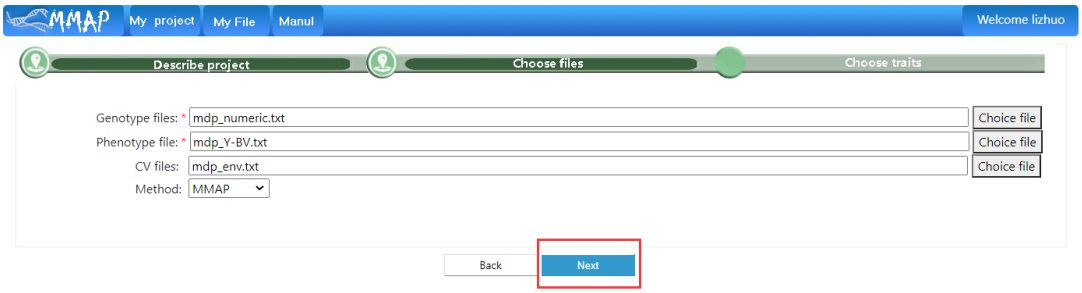
• Fill in the name and comments of the process in the corresponding location, click next to go to the next step.



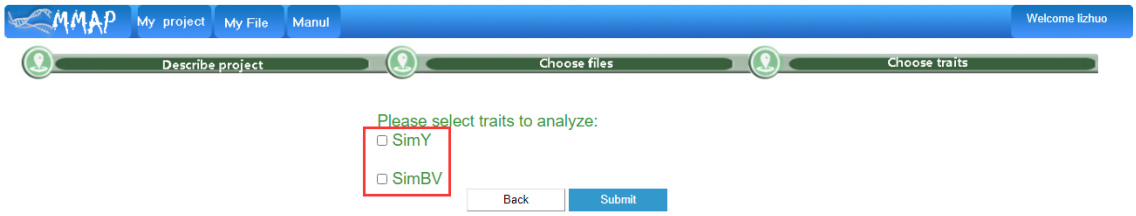
• On the Select File page, click on "choice file" to select the appropriate type of file. Here is a genotype file as an example.



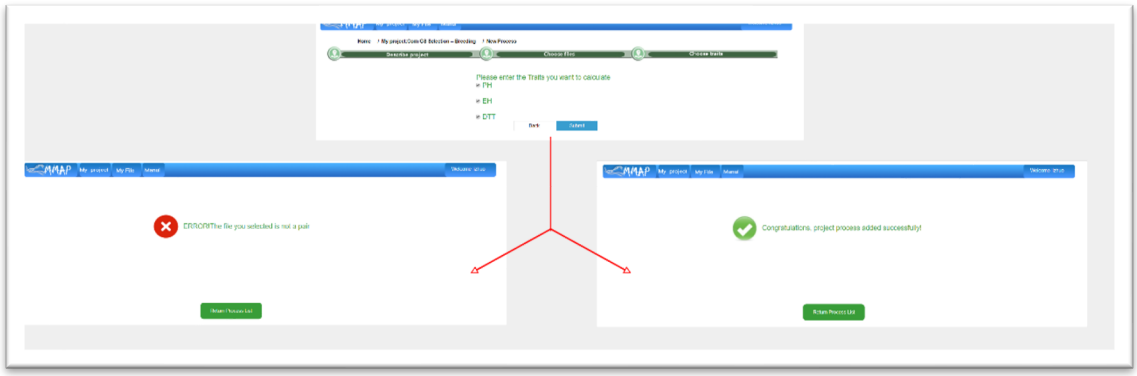
• After the Data name and file are selected, click next to go to the next step.



• Check the columns you want to calculate



• After the check is completed, click “next” to complete the process.

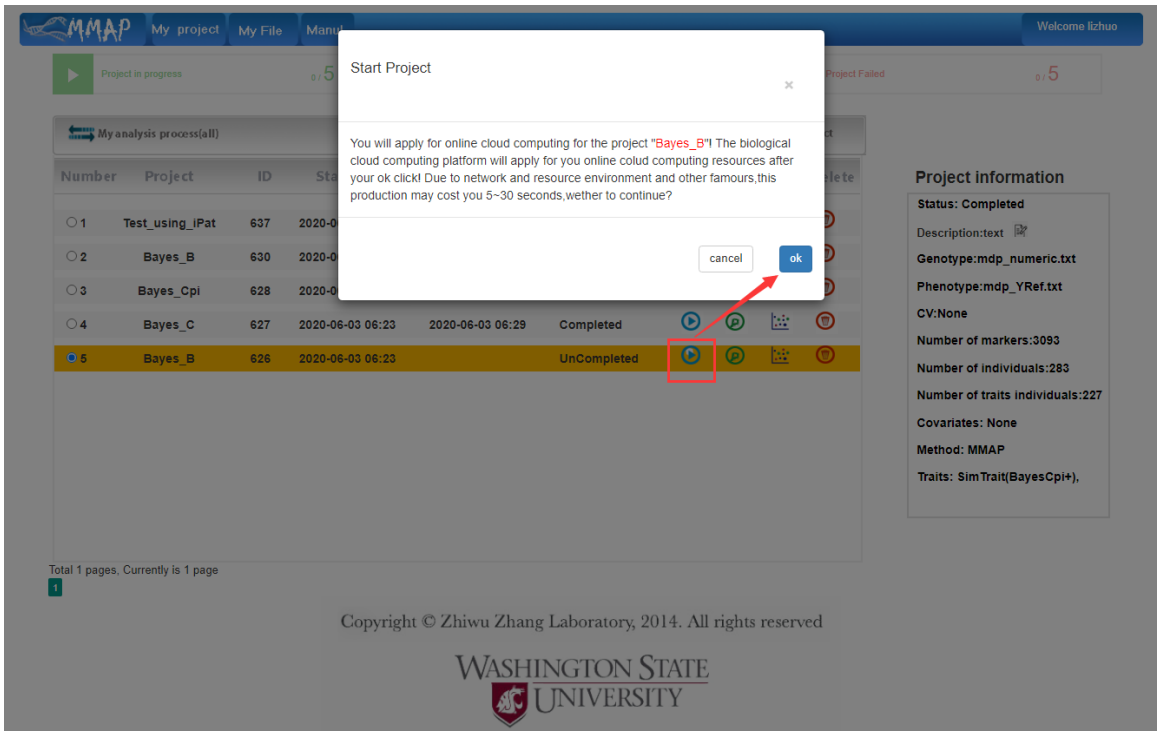


• At this point, the addition process is complete.

3.2 How to Start Project Calculation

On the home page, click on the process you want to calculate and the selected process will be highlighted. Click on the calculation icon of the highlighted process and a dialog will

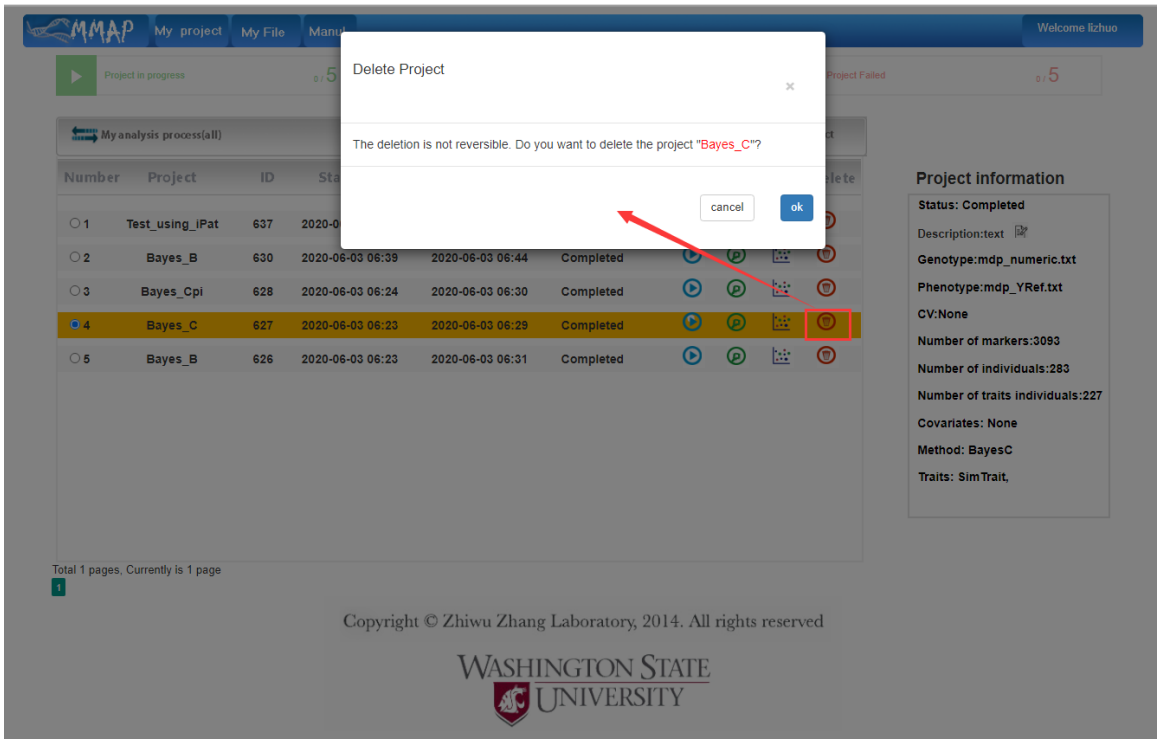
pop up asking if you want to enable the calculation.



Click "ok" to start cloud computing

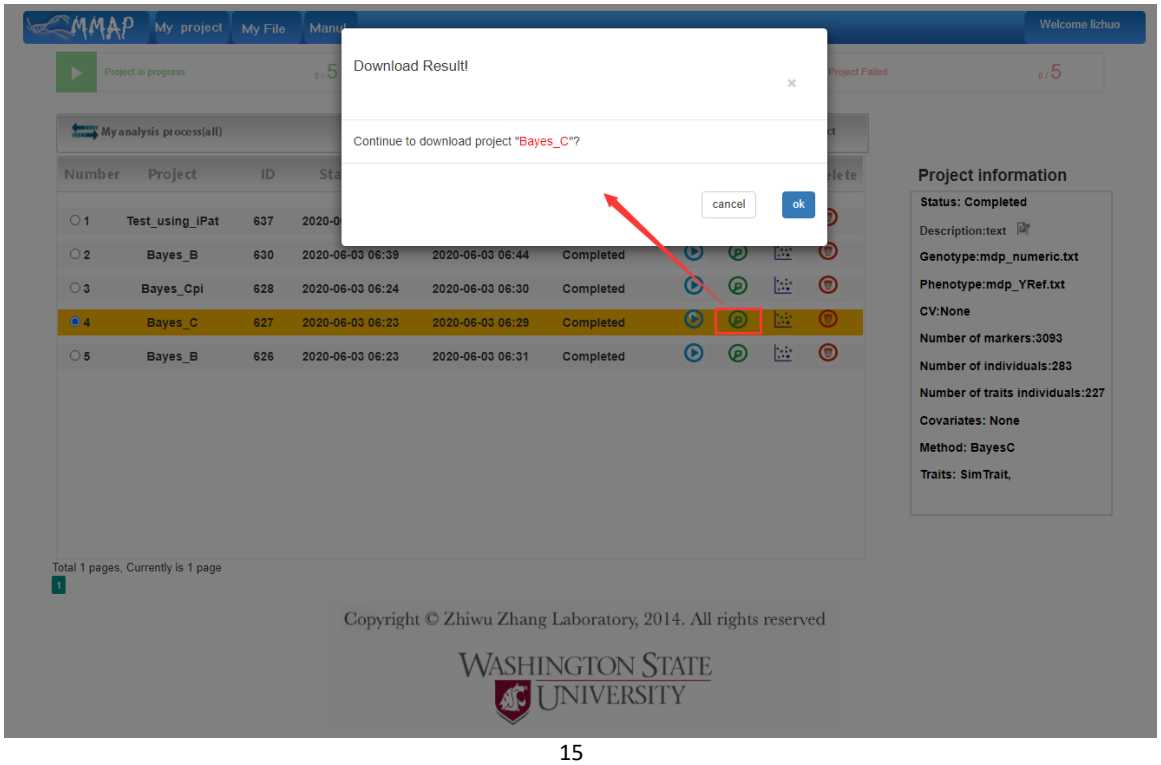
3.3 Delete A Project

On the home page, click on the process you want to delete and the selected process will be highlighted. Click the delete icon of the highlighted process, a dialog window will pop up asking if you want to delete the process.



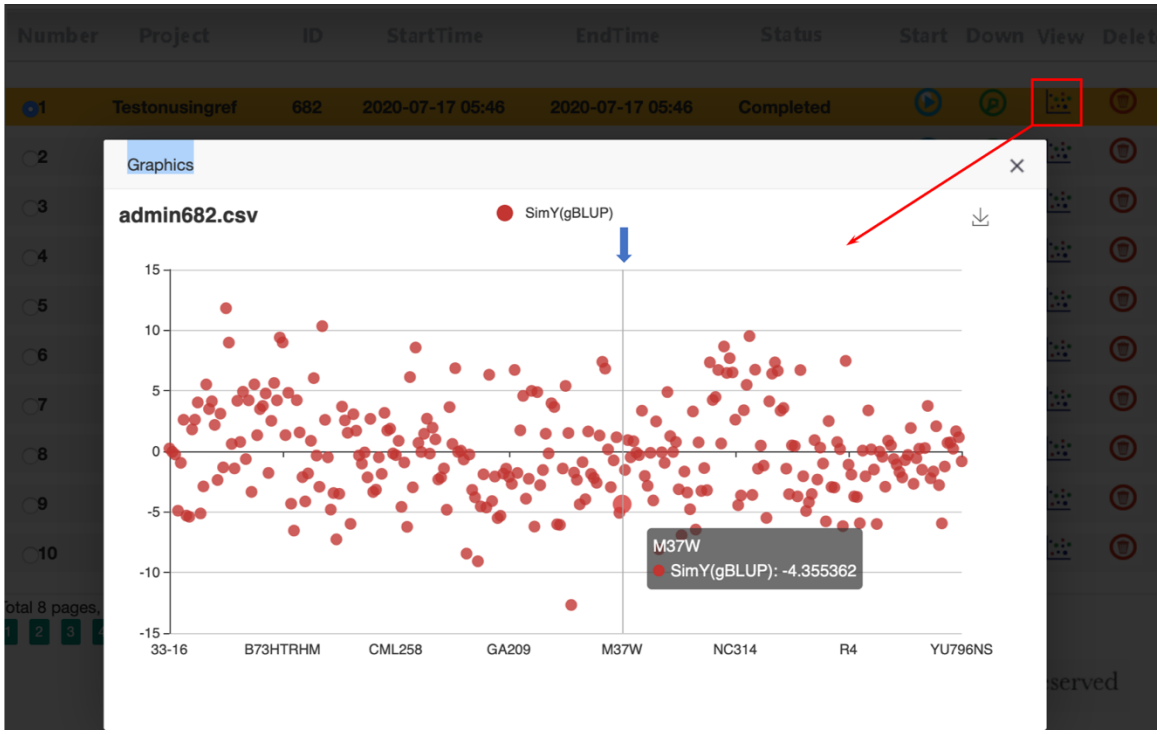
3.4 Download project Report

On the home page, click on the project you want to download the report. The selected project will be highlighted. Click on the download icon for the highlighted project. A dialog will pop up to confirm if you want to download the report.



3.5 How to View Project Report

On the homepage, click on the project you want to view the report. The selected project will be highlighted. Click on the view icon for the highlighted project. A dialog will pop up, asking if you want to view the report. The horizontal axis partially labels the taxa names. The vertical axis is the estimated breeding value. Each dot is an individual. When the cursor is on the plot, a vertical line appears. The information is displayed for the individual closest to the vertical line, including name and estimated breeding value. The dot for the closest individual is enlarged, as indicated by the blue arrow. The 2D plot helps to interpret prediction results, including identification of the individuals with the highest, or the lowest estimated breeding values.



4. INPUT FILES

4.1 Genotype Files

Numerical genotype format is accepted for MMAP. For examples, the two homozygous are coded as 0 and 2 and the heterozygous is coded as 1. Header is optional. The columns are delaminated by Tab, spaces, or comma. The first column is individual ID for linking to other files such as phenotype and covariate variable files. Missing genotypes are not allowed. Here is an example for the first 10 rows and columns displayed in Excel:

taxa	PZB00859.1	PZA01271.1	PZA03613.2	PZA03613.1	PZA03614.2	PZA03614.1	PZA00258.3	PZA02962.13	PZA02962.14	PZA00599.25
33-16	2	0	0	2	2	2	2	2	0	0
38-11	2	2	0	2	2	2	0	2	0	2
4226	2	0	0	2	2	2	0	2	0	0
4722	2	2	0	2	2	2	1	2	0	2
A188	0	0	0	2	2	2	0	2	0	2
A214N	2	0	2	0	2	0	0	2	0	2
A239	0	0	2	2	0	0	0	2	0	2
A272	0	0	2	2	0	0	2	2	0	2
A441-5	2	0	0	2	2	2	0	2	0	2
A554	2	2	2	2	0	2	0	2	0	2

Note: The first row (header) is optional.

4.2 Phenotype Files

The format of phenotype files is similar to genotype files except header is required. The first column is individual ID for linking to other files such as genotype and covariate variable files. The rest columns are for traits. Multiple traits are acceptable. At least one trait is required. The columns are delaminated by Tab, spaces, or comma. Missing data are indicated by “NA”, or “NaN”. The header is for selection of traits to analyze. Here is an example for the first five rows with three columns of traits displayed in Excel:

Taxa	EarHT	dpoll	EarDia
811	59.5	NaN	NaN
4226	65.5	59.5	32.21933
4722	81.13	71.5	32.421
33-16	64.75	64.5	NaN
38-11	92.25	68.5	37.897

A individual in the phenotype files must be included in the genotype file, otherwise, the individual will be excluded in the analysis. Conversely, the individuals in the genotype file do not have to be in the phenotype files, such as the individuals in testing population (inference) for prediction. Consequently, the individuals in the phenotype file do not have to be in the same order as genotype files.

4.3 Files for covariate variables

The format of covariate variable files is also similar to genotype and phenotype files except header and individual ID are both optional. If individual ID is included, it must be on the first column. The rest columns are for covariate variables. Multiple covariate variables are acceptable. At least one covariate variable is required. The columns are delaminated by Tab, spaces, or comma. Missing data are not allowed. Here is an example for the first ten rows with three columns of covariate variables (PCs) displayed in Excel:

taxa	PC1	PC2	PC3
33-16	1.6780775	-4.9373382	1.0029279
38-11	-1.6021749	-4.7322790	-0.6886187
4226	-0.8999517	-6.2186090	2.2737655
4722	2.1334477	-6.2879301	5.6837161
A188	0.6302372	-4.8947416	0.8189001
A214N	-13.6690754	0.8736302	-21.6324613
A239	-0.5680841	-5.9401064	0.3307201
A272	3.7670958	2.7504406	2.2384039
A441-5	5.4327297	-0.1516547	1.0950912
A554	-0.9801181	-5.4792220	1.0430492

The flexibility on the optional individual ID comes with the price of restriction. The rows of the covariate variable file must be corresponded to the genotype files, same size, same order.

5. OUTPUT FILES

5.1 Text File

A comma delaminated file with "csv" extension can be download for each project. The text file has multiple columns. The first column is the name of taxa. The rest columns are the predicted breeding values. Each trait name is followed by the method used. For example, when gBLUP is specified for the demo data (mdp_numeric.txt and mdp_traits.txt), the output should be as following for the first ten rows.

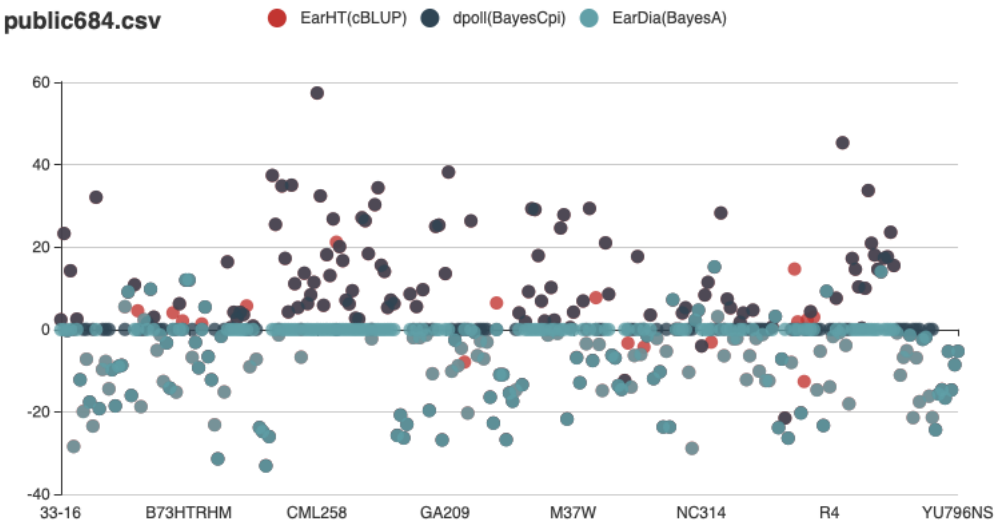
Name	EarHT(cBLUP)	dpoll(cBLUP)	EarDia(cBLUP)
33-16	2.342978898	-2.493965758	0.358728305
38-11	23.26918367	1.67717234	0.924505566
4226	-0.265996782	-7.421175365	-3.636198657
4722	14.24245672	3.157497099	-4.050049737
A188	-28.40417196	-5.376102708	-4.72233902
A214N	2.502748528	1.303327621	-3.282778994
A239	-12.20969125	-6.305223629	-0.906691382
A272	-19.89074779	2.183586922	-0.748929733
A441-5	-7.227049111	0.034001928	-1.602927767

When the default method (MMAP) is used for selecting the best method, the output will add a plus sign ("+") for the method selected by MMAP. For the same data above, the output should be as following when default method (MMAP) is selected.

Name	EarHT(cBLUP+)	dpoll(BayesCpi+)	EarDia(BayesA+)
33-16	2.342978898	0.00342905	-0.057326396
38-11	23.26918367	0.013262256	-0.033738287
4226	-0.265996782	-0.004771784	-0.034994867
4722	14.24245672	9.82E-04	-0.024248246
A188	-28.40417196	0.007761541	-0.035242291
A214N	2.502748528	0.004052183	-0.015484185
A239	-12.20969125	-0.004304375	-0.02754722
A272	-19.89074779	0.007729866	-0.056213018
A441-5	-7.227049111	0.017696258	-0.010969673

5.2 Graphic File

The interactive display of the predicted breeding values can be download in PNG format. Texas are displayed horizontally. Predicted breeding values are displayed vertically.



6. Frequently Asked Questions

1. How to cite MMAP?

A: Wei Huang, Ping Zheng, Zhenhai Cui, Zhuo Li, Yifeng Gao, Helong Yu, You Tang, Xiaohui Yuan, and Zhiwu Zhang, MMAP User manual (Mining the Maximum Accuracy of Prediction), version 1.0, http://zzlab.net/MMAP, accessed on MM/DD/YYYY.

2. What do I do if I get frustrated?

A: Try to go through this Q/A list first before asking help from MMAP team. If you need to contact MMAP team, email to Dr. You Tang (email: 1098118439@qq.com).

3. Why MMAP has different results from other software?

A: MMAP generates the identical results with other software packages such as GAPIT (Lipka *et al.* 2012; Tang *et al.* 2016) and rrBLUP for specific methods that do not involve random sampling. For methods with random sampling such as Bayesian methods (A, B, C, Cpi, and LASSO), the differences between MMAP and other software packages such BLR and BGLR(Pérez and De Los Campos 2004), are similar to the difference among multiple replicates within using the same packages.

4. What are the methods implemented in MMAP for genomic prediction?

A: Currently MMAP implemented eight methods for genomic selection, including genomic Best Linear Unbiased Prediction (gBLUP), compressed BLUP (cBLUP), Ridge Regression BLUP (rrBLUP), Bayes A, Bayes B, Bayes C, Bayes Cpi, Bayesian LASSO.

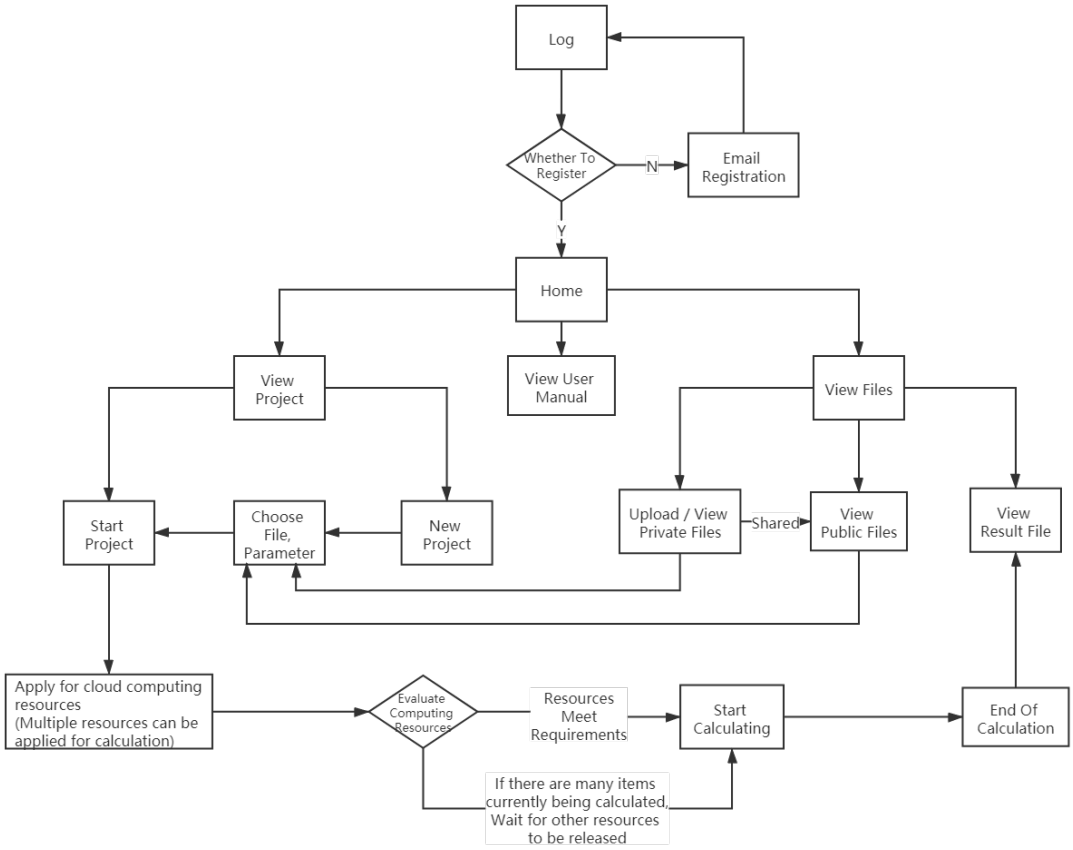
5. Which method is the best for genomic prediction?

A: The best method varies across datasets, depending on heritability and number of genes underlying the traits(Wang *et al.* 2018). For a new dataset, MMAP identifies the best method based on the similarity between the new dataset and the existing datasets.

6. Is it possible to analyze genotype files above the upload limitation?

A: Yes, please email with Dr. You Tang for options (1098118439@qq.com).

OPERATION FLOWCHART



REFERENCE

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

Kim B., Dai X., Zhang W., Zhuang Z., Sanchez D. L., *et al.*, 2018 GWASpro: a high-performance genome-wide association analysis server. Bioinformatics 35: 2512–2514.

Lipka A. E., Tian F., Wang Q., Peiffer J., Li M., *et al.*, 2012 GAPIT: genome association and prediction integrated tool. Bioinformatics 28: 2397–2399.

Pérez P., Los Campos G. De, 2004 *BGLR: A Statistical Package for Whole Genome Regression and Prediction*.

Tang Y., Liu X., Wang J., Li M., Wang Q., *et al.*, 2016 GAPIT Version 2: An Enhanced Integrated Tool for Genomic Association and Prediction. Plant J. 9.

Wang J., Zhou Z., Zhang Z., Li H., Liu D., *et al.*, 2018 Expanding the BLUP alphabet for genomic prediction adaptable to the genetic architectures of complex traits. Heredity (Edinb).