

**User manual for**

# **Predhy.GUI**

**Performs Genomic Prediction of Hybrid Performance**

**With**

**Graphical User Interface**

**(Vision 1.0)**

**Yang Xu, Guangning Yu, Yuxiang Zhang, Yanru Cui,**

**Shizhong Xu, Chenwu Xu**

**(xuyang\_89@126.com)**

**Last updated on December, 2022**

# contents

1. Getting started .....	3
1.1 installation .....	3
1.2 Run predhy.GUI .....	3
2. Dataset input .....	3
2.1 Genotype dataset .....	3
2.1.1 Input_genotype dataset .....	3
2.1.2 Inbred_gene dataset(*.csv format file) .....	4
2.2 Phenotype dataset(*.csv format file) .....	5
2.3 Parent names dataset(*.csv format file) .....	5
3. Operation process .....	6
3.1 cv .....	6
Dataset Input .....	6
Method select & Parameter setting .....	7
Run the software .....	8
3.2 predhy.predict .....	8
Dataset Input .....	9
Method select & Parameter setting .....	9
Run the software .....	10
3.3 predhy.predict_NCII .....	11
Dataset Input .....	11
Method select & Parameter setting .....	11
Run the software .....	12
3.4 convertgen .....	14
Dataset Input .....	14
Method select & Parameter setting .....	14
Run the software .....	15
3.5 crodesign .....	16
Dataset Input .....	16
Method selection & Parameter setting .....	16
Run the software .....	17

# 1. Getting started

The software package predhy.GUI runs only in the R software environment and can be freely downloaded from the R website (<https://cran.r-project.org>).

## 1.1 installation

Within R environment, the predhy.GUI software can be installed online using the below command:

```
install.packages("predhy.GUI")
```

## 1.2 Run predhy.GUI

Once the software predhy.GUI is installed, users may run the software using two commands:

```
library("predhy.GUI")  
predhy.GUI()
```

# 2. Dataset input

## 2.1 Genotype dataset

### 2.1.1 Input\_genotype dataset

**Numeric format for Genotypic dataset** (\*.csv or \*.txt format file)

The first column stands for marker ID. Among the remaining columns, each column lists all the genotypes for one individual while the first row shows the individual names. For each marker, homozygous genotypes are expressed by 1 and -1, respectively, and the heterozygous genotypes are indicated by zero, missing values are indicated by NA.

	R001	R002	R003	R004	R005	R006	R007	R008
SNP1	-1	1	1	1	-1	1	-1	-1
SNP2	-1	1	1	1	-1	1	-1	-1
SNP3	-1	1	1	1	-1	1	-1	-1
SNP4	-1	1	1	1	-1	1	-1	-1
SNP5	-1	1	1	1	-1	1	-1	-1
SNP6	-1	1	1	1	-1	1	-1	-1
SNP7	-1	1	1	NA	-1	1	-1	-1
SNP8	-1	1	1	1	-1	1	-1	-1
SNP9	-1	1	1	1	-1	1	-1	-1
SNP10	-1	1	1	1	-1	1	-1	-1

## Hapmap format for Genotypic dataset (\*.txt format file)

Please see the TASSEL software in details. Here we introduce simply. The first eleven columns describe the specific information of markers and individuals, and their column names must be "rs#", "alleles", "chrom", "pos", "strand", "assembly#", "center", "protLSID", "assayLSID", "panel" and "QCcode".

The values for marker genotypes should be character, such as AA, TT, CC, GG, NN, AC and AG, where the "NN" indicates missing or unknown genotypes. In the 2 and 5 to 11 columns, "NA" indicates **no information** available. All the individual genotypic information will be showed from the 12 to last columns. In each column, individual name is listed in the first row, i.e., "A002", and the others are the genotypes (character).

rs	alleles	chrom	pos	strand	assembly	center	protLSID	assayLSID	panel	QCcode	A002	A003	A004	A005	A006
SNP_1_14068	T/C	1	14068	NA	NA	NA	NA	NA	NA	NA	NA	TT	TT	NA	TT
SNP_1_338176	G/T	1	338176	NA	NA	NA	NA	NA	NA	NA	NA	NA	GG	NA	GG
SNP_1_703171	G/A	1	703171	NA	NA	NA	NA	NA	NA	NA	GG	GA	GG	GA	GA
SNP_1_1033512	C/T	1	1033512	NA	NA	NA	NA	NA	NA	NA	TT	TT	CC	NA	TT
SNP_1_1401306	A/C	1	1401306	NA	NA	NA	NA	NA	NA	NA	CC	CC	CC	NA	CC
SNP_1_1465404	C/T	1	1465404	NA	NA	NA	NA	NA	NA	NA	CC	CC	CC	CC	CT
SNP_1_1725463	C/T	1	1725463	NA	NA	NA	NA	NA	NA	NA	CT	CT	CC	CT	CT
SNP_1_1866006	C/T	1	1866006	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
SNP_1_2045326	G/A	1	2045326	NA	NA	NA	NA	NA	NA	NA	GG	AA	AA	GG	GG
SNP_1_2670571	A/G	1	2670571	NA	NA	NA	NA	NA	NA	NA	AA	AA	AA	AA	AA
SNP_1_2950255	G/C	1	2950255	NA	NA	NA	NA	NA	NA	NA	GG	GG	GG	GG	GG
SNP_1_3818861	A/T	1	3818861	NA	NA	NA	NA	NA	NA	NA	AA	AA	AA	AA	AA
SNP_1_4185501	C/G	1	4185501	NA	NA	NA	NA	NA	NA	NA	GG	CC	CC	CC	CC
SNP_1_4616639	G/T	1	4616639	NA	NA	NA	NA	NA	NA	NA	NA	GG	GG	GT	GT
SNP_1_5036129	G/A	1	5036129	NA	NA	NA	NA	NA	NA	NA	GG	GG	GG	GG	GG

## 2.1.2 Inbred\_gene dataset(\*.csv format file)

A matrix for genotypes of parental lines in numeric format, coded as 1, 0 and -1. The first columns indicates the names of inbred lines, which must be provided. Among the remaining columns, each column lists all the genotypes for a SNP while the first row shows the SNP names.

It can be obtained from the original genotype using convertgen function.

	SNP_1_14068	SNP_1_338176	SNP_1_703171	SNP_1_1033512	SNP_1_1401306	SNP_1_1465404	SNP_1_1725463	SNP_1_1866006
A002	0.521126761	0.800711744	1	-1	-1	1	0	0.580952381
A003	1	0.800711744	0	-1	-1	1	0	0.580952381
A004	1	1	1	1	-1	1	1	0.580952381
A005	0.521126761	0.800711744	0	-0.239875389	-0.865319865	1	0	0.580952381
A006	1	1	0	-1	-1	0	0	0.580952381
A007	0	-1	1	-1	-1	1	0	-1
A008	1	1	0	-1	-1	1	0	-1
A010	1	0	1	1	1	1	0	0.580952381
A011	1	1	0	-1	-1	1	0	-1
A012	1	1	1	-1	-1	1	0	1
A013	1	1	0	-1	-1	1	0	0.580952381
A014	1	1	1	-1	-0.865319865	0	0	-1
A015	-1	0.800711744	0	-0.239875389	-0.865319865	1	0	1
A016	0	0	1	-1	-1	1	0	-1
A017	-1	0	1	-1	-1	1	0	1
A018	1	0	1	-1	-1	1	0	1
A020	0.521126761	1	1	1	1	1	0	1
A021	-1	1	1	-1	-1	1	0	1
A022	1	0.800711744	1	-1	-1	1	0	-1
A023	1	1	1	1	-1	1	0	1

## 2.2 Phenotype dataset(\*.csv format file)

A data frame with three columns. The first column and the second column are the names of male and female parents of the corresponding hybrids, respectively; the third column is the phenotypic values of hybrids. The names of male and female parents must match the rownames of `inbred_gen`. Missing (NA) values are not allowed.

M	F	Trait1
A002	A017	1433.745
A003	A393	1451.795
A003	A256	952.38
A003	A187	522.58
A003	A071	1457.775
A003	A439	1320.1
A005	A429	1638.91
A005	A430	1592.485
A006	A017	2050.12
A006	A021	1948.125
A006	A304	1474.83
A006	A268	1499.175
A006	A010	1010.345
A006	A030	953.685
A007	A021	1541.34

### 2.3 Parent names dataset(\*.csv format file)

male\_name: a data frame with only one column, of the names of male parents, with “M” in the first row.

female\_name: a data frame with only one column, of the names of female parents, with “F” in the first row.

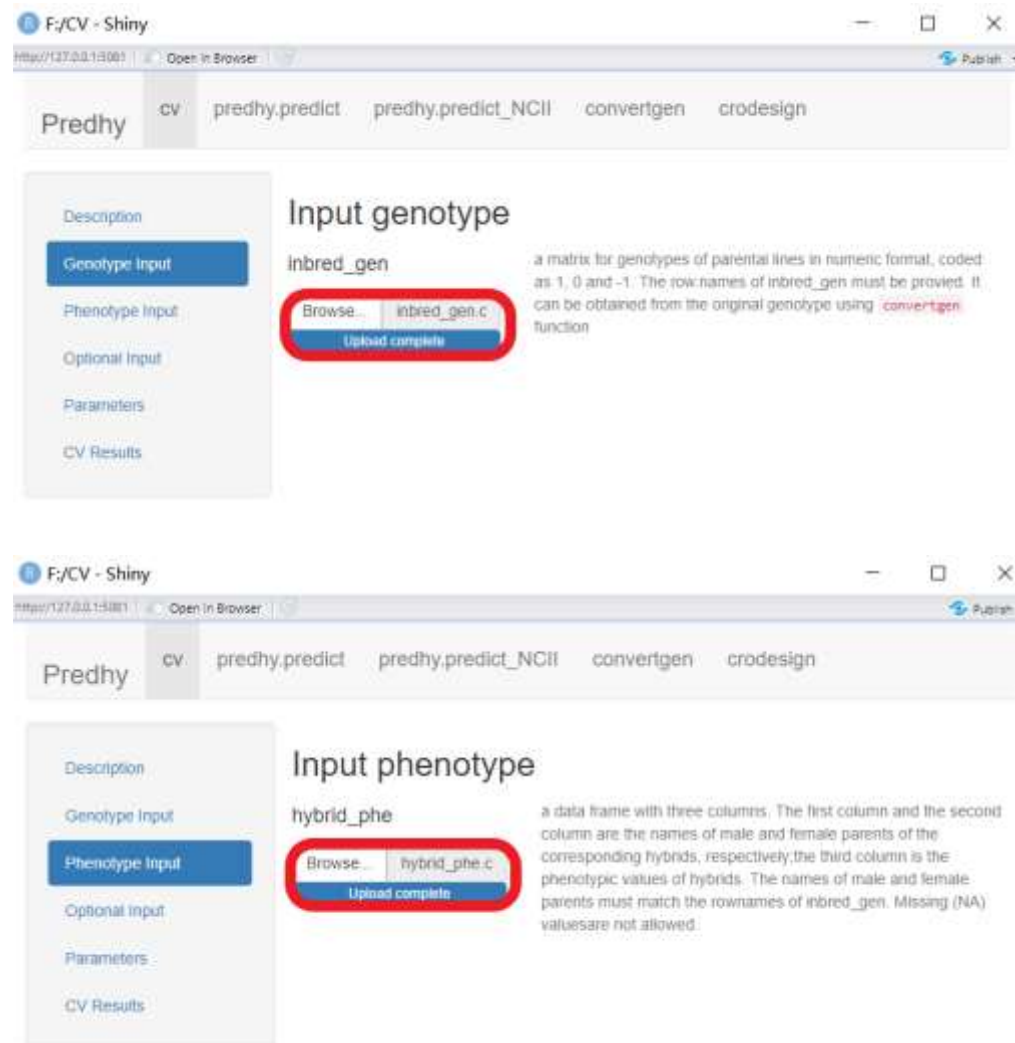
M	F
A002	A008
A003	A008
A003	A008
A003	A008
A003	A010
A003	A010
A005	A010
A005	A010
A006	A010
A006	A010
A006	A011
A006	A011
A006	A011
A006	A011
A007	A012

### 3. Operation process

#### 3.1 cv

##### Dataset Input

Users must upload the `inbred_gen` and phenotype files, while the design matrix are optional. In design matrix module, users should upload the design matrix if you select “**Input a design matrix**”; users don’t need to upload this file, which will be ignored, if you select “**Not included**”. The dominance genotypes is also optional, in dominance genotypes module, if you select “**Include dominance genotypes**”; users don’t need to upload this file, which will be calculated automatically, if you select “**Not included**”, it will be ignored.



Predhy CV predhy.predict predhy.predict\_NCII convertgen crodesign

Description  
Genotype Input  
Phenotype Input  
**Optional Input**  
Parameters  
CV Results

### Input design matrix of the fixed effects & dominance genotypes

design matrix of the fixed effects(Optional)

☐ Not included  
☒ Input a design matrix

fixed effects

Browse... No file selected

dominance genotypes(Optional)

☒ Not included  
☐ Include dominance genotypes

## Method select & Parameter setting

**Method:** There are eight GS methods in the predhy.GUI, including "GBLUP", "BayesB", "RKHS", "PLS", "LASSO", "EN", "XGBOOST", "RF". Users may select one of those methods or all of them simultaneously with "ALL".

**Number of folds:** The k for k-fold cross validation.

**Replicates:** Repeat number of independent replicates for the cross-validation.

**The random number:** The random number.

**CPU:** the number of CPU for parallel calculation.

Predhy CV predhy.predict predhy.predict\_NCII convertgen crodesign

Description  
Genotype Input  
Phenotype Input  
Optional Input  
**Parameters**  
CV Results

### Select models & other parameters

method, eight GS methods

GBLUP

the number of folds

5

replicates

1

the random number

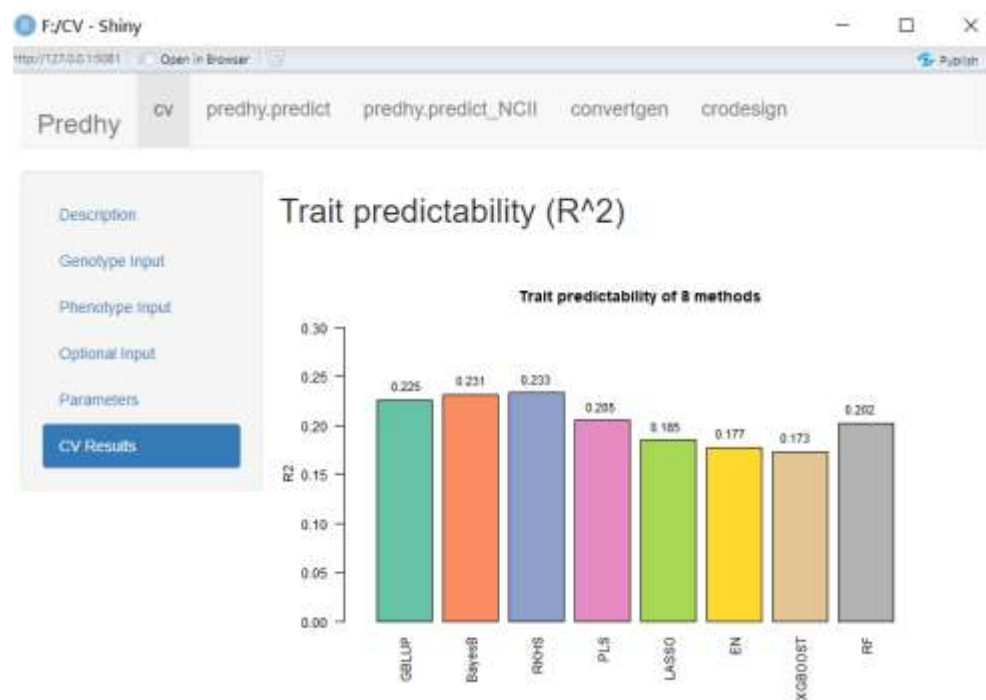
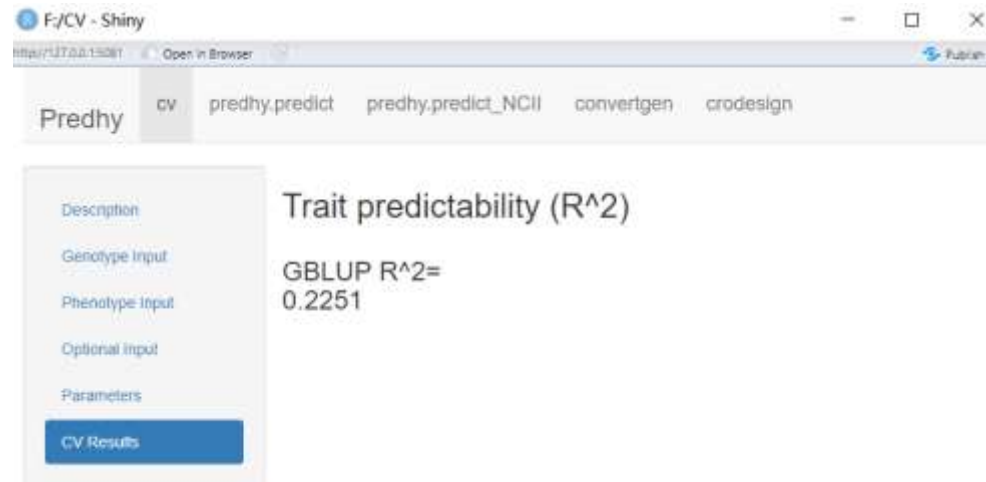
133

the number of CPU

1

## Run the software

After uploading all the needed files and setting the parameters, users can run the Software simply by clicking “CV Results”. The result will be print on the panel if a single method is selected. If you chose “ALL” in method, a plot of cross validation result for eight methods will be given.

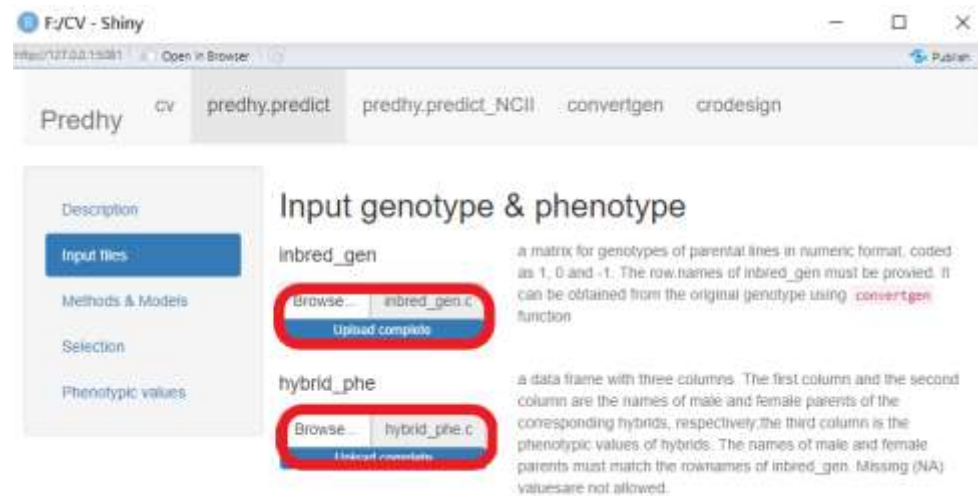


## 3.2 predhy.predict

This function was designed to predict all potential crosses of a given set of parents using a subset of crosses as the training sample.

## Dataset Input

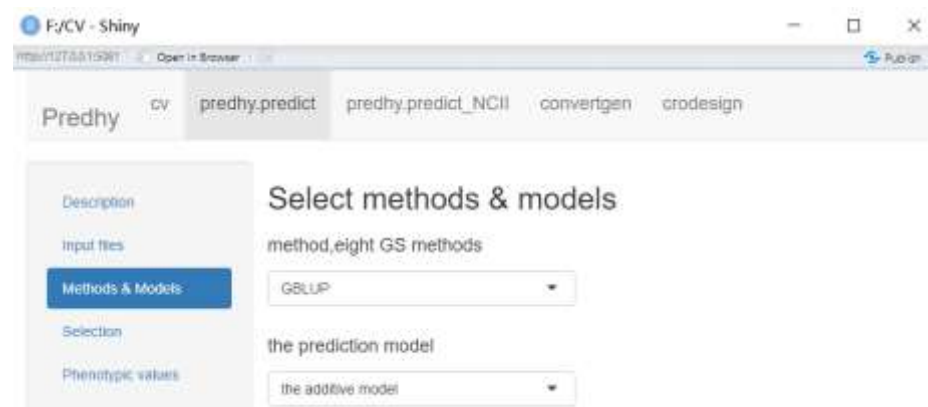
Users must upload the inbred\_gen and phenotype files.



## Method select & Parameter setting

**Method:** There are eight GS methods in the predhy.GUI for hybrid performance predicting, including "GBLUP", "BayesB", "RKHS", "PLS", "LASSO", "EN", "XGBOOST", "RF". Users should select one of those methods.

**Prediction model:** There are two options: the additive model, the additive-dominance model, user can choose one by select one of the choices.



**Select hybrids:** Selection of hybrids based on the prediction results. There are three options: select = "all", which selects all potential crosses. select = "top", which selects the top n crosses. select = "bottom", which selects the bottom n crosses. User can decide number hybrids to select when select = "top" or select = "bottom".

**Select hybrids**

the selection of hybrids based on the prediction results

all potential crosses

the number of selected top or bottom hybrids, only when select = "top" or select = "bottom".

100

## Run the software

After uploading all the needed files and setting the parameters, users can run the Software simply by clicking “Phenotypic values”. When calculation is down, the result will be given in the datatable below the panel, user may download the full data by clicking at “Predict & Download Results” bottom.

**Phenotypic values of the predicted hybrids.**

[Predict & Download Results](#)

Show 10 entries Search:

top\_100

A062/A291	1925.4469052247
A169/A291	1923.3362122002
A133/A291	1920.01656259097
A027/A291	1919.05920723499
A017/A291	1916.77925420535
A036/A291	1916.63947026575
A062/A169	1905.34090031377
A052/A291	1904.01580338883
A062/A133	1902.02325070454
A291/A396	1901.66370902396

Showing 1 to 10 of 100 entries Previous 1 2 3 4 5 ... 10 Next

### 3.3 predhy.predict\_NCII

This function was designed to predict all potential crosses of a given set of parents (usually between different heterotic groups) using a subset of crosses as the training sample, following the North Carolina mating design II.

#### Dataset Input

Users must upload the inbred\_gen and phenotype files, along with the Heterotic group dataset (two files, one contains male\_names, the other contains female\_names).

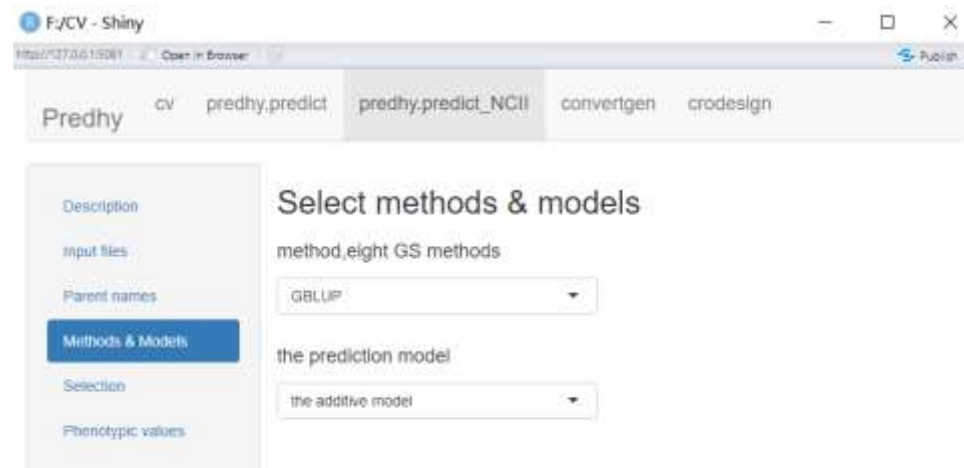
The first screenshot shows the 'predhy.predict\_NCII' tab in the Shiny app. The left sidebar has 'Input files' selected. The main area is titled 'Input genotype & phenotype'. It contains two input sections: 'inbred\_gen' and 'hybrid\_phe'. Each section has a 'Browse...' button, a text field with a file name, and an 'Upload complete' button. In the 'inbred\_gen' section, the 'Browse...' button and 'inbred\_gen.csv' are highlighted with a red box. The description for 'inbred\_gen' states it is a matrix for genotypes of parental lines in numeric format, coded as 1, 0 and -1. The 'hybrid\_phe' section also has its 'Browse...' button and 'hybrid\_phe.csv' highlighted with a red box. The description for 'hybrid\_phe' states it is a data frame with three columns: male parent names, female parent names, and phenotypic values.

The second screenshot shows the same app with the 'Parent names' section selected in the sidebar. The main area is titled 'Input names of parents'. It contains two input sections: 'male\_name' and 'female\_name'. Each section has a 'Browse...' button, a text field with a file name, and an 'Upload complete' button. In the 'male\_name' section, the 'Browse...' button and 'male\_name.csv' are highlighted with a red box. The description for 'male\_name' states it is a vector of the names of male parents. The 'female\_name' section also has its 'Browse...' button and 'female\_name.csv' highlighted with a red box. The description for 'female\_name' states it is a vector of the names of female parents.

#### Method select & Parameter setting

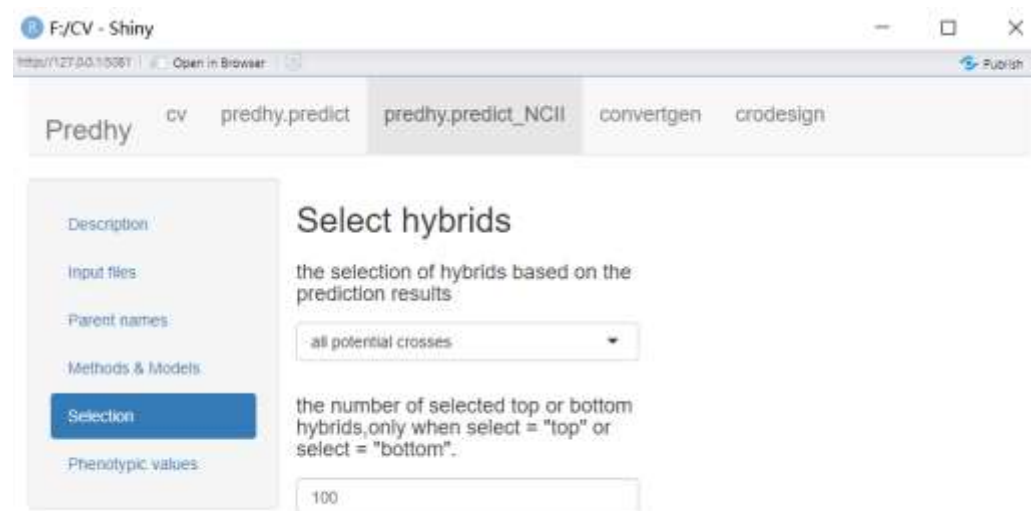
**Method:** There are eight GS methods in the predhy.GUI for hybrid performance predicting, including "GBLUP", "BayesB", "RKHS", "PLS", "LASSO", "EN", "XGBOOST", "RF". Users should select one of those methods.

Prediction model: There are two options: the additive model, the additive-dominance model, user can choose one by select one of the choices.



The screenshot shows a web browser window with the URL 'https://127.0.0.1:5081'. The browser tab is 'F2/CV - Shiny'. The application has a navigation bar with tabs: 'Predhy', 'cv', 'predhy.predict', 'predhy.predict\_NCII' (active), 'convertgen', and 'crodesign'. On the left is a sidebar with a vertical list of menu items: 'Description', 'input files', 'Parent names', 'Methods & Models' (highlighted in blue), 'Selection', and 'Phenotypic values'. The main content area is titled 'Select methods & models'. It contains three sections: 'method,eight GS methods' with a dropdown menu showing 'GBLUP'; 'the prediction model' with a dropdown menu showing 'the additive model'; and a 'Publish' button in the top right corner.

**Select hybrids:** Selection of hybrids based on the prediction results. There are three options: select = "all", which selects all potential crosses. select = "top", which selects the top n crosses. select = "bottom", which selects the bottom n crosses. User can decide number hybrids to select when select = "top" or select = "bottom".



The screenshot shows the same web browser window as the previous one, but the 'predhy.predict\_NCII' tab is active. The sidebar menu is the same, but the 'Selection' item is now highlighted in blue. The main content area is titled 'Select hybrids'. It contains two sections: 'the selection of hybrids based on the prediction results' with a dropdown menu showing 'all potential crosses'; and 'the number of selected top or bottom hybrids,only when select = "top" or select = "bottom".' with a text input field containing the number '100'. The 'Publish' button is still visible in the top right corner.

## Run the software

After uploading all the needed files and setting the parameters, users can run the Software simply by clicking “Phenotypic values”. When calculation is down, the result will be given in the datatable below the panel, user may download the full data by clicking at “Predict & Download Results” bottom.

Predhy

cv

predhy.predict

predhy.predict\_NCI

convertgen

crodesign

Description

Input files

Parent names

Methods & Models

Selection

Phenotypic values

## Phenotypic values of the predicted hybrids.

### Predict & Download Results

Show 10 entries

Search:

top\_100

A291/A291	1944.44221711114
A062/A291	1926.4469052247
A291/A062	1926.4469052247
A169/A291	1923.3362122002
A291/A169	1923.3362122002
A133/A291	1920.01856259097
A291/A133	1920.01856259097
A027/A291	1919.05920723499
A291/A027	1919.05920723499
A017/A291	1915.77925420535

Showing 1 to 10 of 100 entries

Previous

1

2

3

4

5

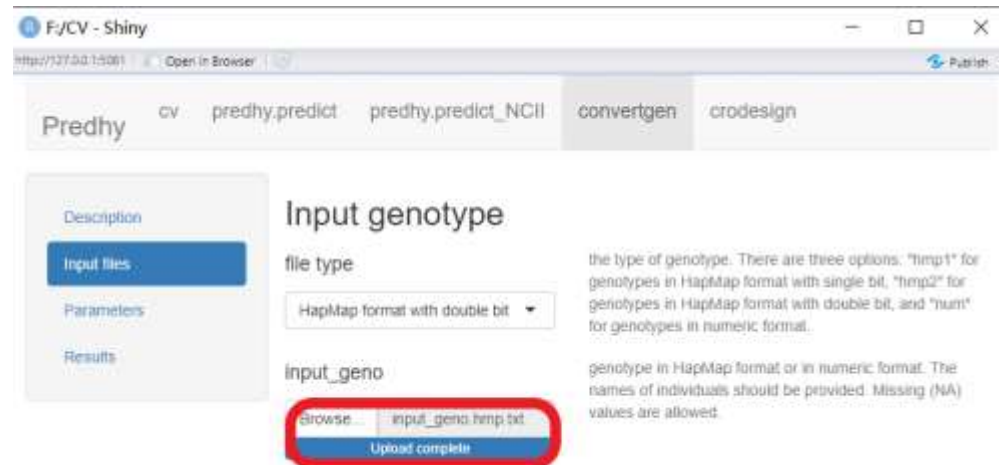
10

Next

### 3.4 convertgen

#### Dataset Input

Users must first click the drop-down menu to select the genotype file type, which includes “HapMap format with single bit”, “HapMap format with double bit”, “numeric format”. Then users can click the file input box to upload their data.



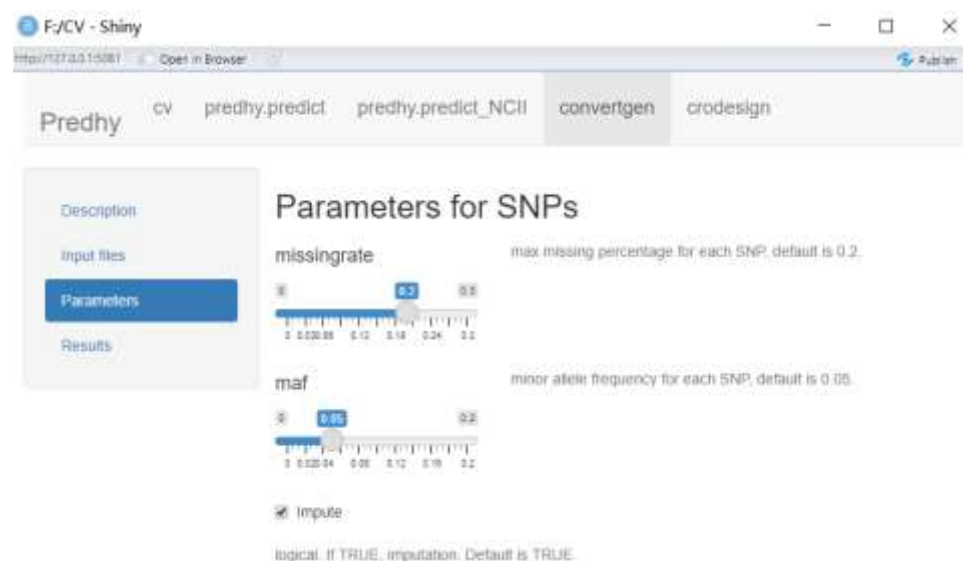
The screenshot shows the 'convertgen' tab of the 'F2/CV - Shiny' application. The 'Input genotype' section is active. It features a 'file type' dropdown menu set to 'HapMap format with double bit'. Below this is the 'input\_geno' section, which includes a 'browse...' button, a text input field containing 'input\_geno.hmp.txt', and an 'Upload complete' button. A red rectangle highlights the 'browse...' button and the 'Upload complete' button. To the right of the input fields, there is explanatory text: 'the type of genotype. There are three options: "hmp1" for genotypes in HapMap format with single bit, "hmp2" for genotypes in HapMap format with double bit, and "num" for genotypes in numeric format.' and 'genotype in HapMap format or in numeric format. The names of individuals should be provided. Missing (NA) values are allowed.'

#### Method select & Parameter setting

**missingrate:** max missing percentage for each SNP, users are allowed to choose one by sliding the bottom on the sliderInput.

**maf:** minor allele frequency for each SNP, users are allowed to choose one by sliding the bottom on the sliderInput.

**Impute:** users can click on the checkbox to decide whether to impute NA SNP or not.



The screenshot shows the 'Parameters for SNPs' section of the 'convertgen' tab. It features two slider inputs: 'missingrate' with a range from 0 to 0.2 and a default value of 0.2, and 'maf' with a range from 0 to 0.2 and a default value of 0.05. Below the sliders is a checkbox labeled 'Impute' which is checked. To the right of the sliders, there is explanatory text: 'max missing percentage for each SNP, default is 0.2.' and 'minor allele frequency for each SNP, default is 0.05.' At the bottom, there is a note: 'logical. If TRUE, imputation. Default is TRUE.'

## Run the software

After uploading all the needed files and setting the parameters, users can run the Software simply by clicking “Results”. When calculation is down, the result will be given in the datatable below the panel, user may download the full data by clicking at “Download Genotype” bottom.

The screenshot shows the Predhy web application interface. The browser address bar indicates the URL is <http://127.0.0.1:5081>. The application has a navigation bar with tabs: **Predhy**, **cv**, **predhy.predict**, **predhy.predict\_NCII**, **convertgen** (selected), and **crodesign**. On the left, a sidebar contains links for **Description**, **Input files**, **Parameters**, and a **Results** button. The main content area is titled **Converted genotype** and includes a **Show 10 entries** dropdown and a **Search:** input field. Below this is a table with 5 columns: an identifier column and four SNP columns: **SNP\_1\_14068**, **SNP\_1\_338176**, **SNP\_1\_703171**, and **SNP\_1\_1033512**. The table displays 12 rows of data (A002 to A012). At the bottom, it shows **Showing 1 to 10 of 348 entries** with pagination controls (Previous, 1, 2, 3, 4, 5, 35, Next) and a **Download Genotype** link.

	SNP_1_14068	SNP_1_338176	SNP_1_703171	SNP_1_1033512
A002	0.52112676056336	0.800711743772242	1	-1
A003	1	0.800711743772242	0	-1
A004	1	1	1	1
A005	0.52112676056336	0.800711743772242	0	-0.2396753894081
A006	1	1	0	-1
A007	0	-1	1	-1
A008	1	1	0	-1
A010	1	0	1	1
A011	1	1	0	-1
A012	1	1	1	-1

### 3.5 crodesign

This function was designed to generate a mating design for a subset of crosses based on a balanced random partial rectangle cross-design (BRPRCD) (Xu et al. 2016).

#### Dataset Input

Users need to upload the Parent names dataset(two files, one contains male\_names, the other contains female\_names).

The screenshot shows the 'crodesign' tab in the 'F2/VCV - Shiny' application. The sidebar on the left has 'Parent names' selected. The main content area is titled 'Input parent names'. It contains two sections: 'male parent name' and 'female parent name'. Each section has a 'Browse...' button and an 'Upload complete' button. The 'Browse...' buttons are highlighted with red circles. The top navigation bar shows tabs for 'cv', 'predhy.predict', 'predhy.predict\_NCII', 'convertgen', and 'crodesign'.

#### Method selection & Parameter setting

**percentage:** User can decide the percentage of all potential hybrids to be evaluated in the field by clicking the numericInput.

**seed:** The random number.

The screenshot shows the 'Parameters' section in the 'F2/VCV - Shiny' application. The sidebar on the left has 'Input parameters' selected. The main content area is titled 'Parameters'. It contains two sections: 'percentage' and 'seed'. The 'percentage' section has a numeric input field with the value '50'. The 'seed' section has a numeric input field with the value '123'. The top navigation bar shows tabs for 'cv', 'predhy.predict', 'predhy.predict\_NCII', 'convertgen', and 'crodesign'.

## Run the software

After uploading all the needed files and setting the parameters, users can run the Software simply by clicking “Results”. When calculation is down, the result will be given in the datatable below the panel, user may download the full data by clicking at “Download crodesign” bottom.

The screenshot shows a web browser window titled "F:/CV - Shiny" with the URL "http://127.0.0.1:5081". The browser has tabs for "Open in Browser" and "Publish". The application interface has a top navigation bar with tabs: "Predhy", "cv", "predhy.predict", "predhy.predict\_NCII", "convertgen", and "crodesign" (which is selected). On the left, there is a sidebar with links: "Description", "Parent names", "Input parameters", and "Results" (which is highlighted in blue). The main content area is titled "Results" and contains a link "Download crodesign". Below this, there is a "Show 10 entries" dropdown and a "Search:" input field. A table displays 10 rows of data with columns "crossID", "male\_Name", and "female\_Name". The table shows a list of IDs and names. At the bottom, it says "Showing 1 to 10 of 3,370 entries" and has pagination controls: "Previous", "1", "2", "3", "4", "5", "...", "337", and "Next".

	crossID	male_Name	female_Name
1	1	A008	A007
2	2	A054	A007
3	3	A156	A007
4	4	A005	A007
5	5	A335	A007
6	6	A426	A007
7	7	A025	A007
8	8	A011	A007
9	9	A166	A007
10	10	A092	A030