User manual for

Predhy.GUI

Performs Genomic Prediction of Hybrid Performance

With

Graphical User Interface

(Vision 2.0)

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

Shizhong Xu, Chenwu Xu

(xuyang_89@126.com)

Last updated on April, 2024

contents

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 datasets

2.1.1 Input_geno 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 datasets (*.csv format file)

2.2.1 Hybrid phenotype

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

2.2.2 Parent phenotype

A matrix of phenotypic values of parent. The names of the matrix must match the rownames of inbred_gen.

	parent_phe
A002	1
A003	1
A004	1
A005	1
A006	2
A007	1
A008	1
A010	1
A011	1
A012	1

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
A002
A003
A003
A003
A003
A003
A005
A005
A006
A006
A006
A006
A006
A006
A007

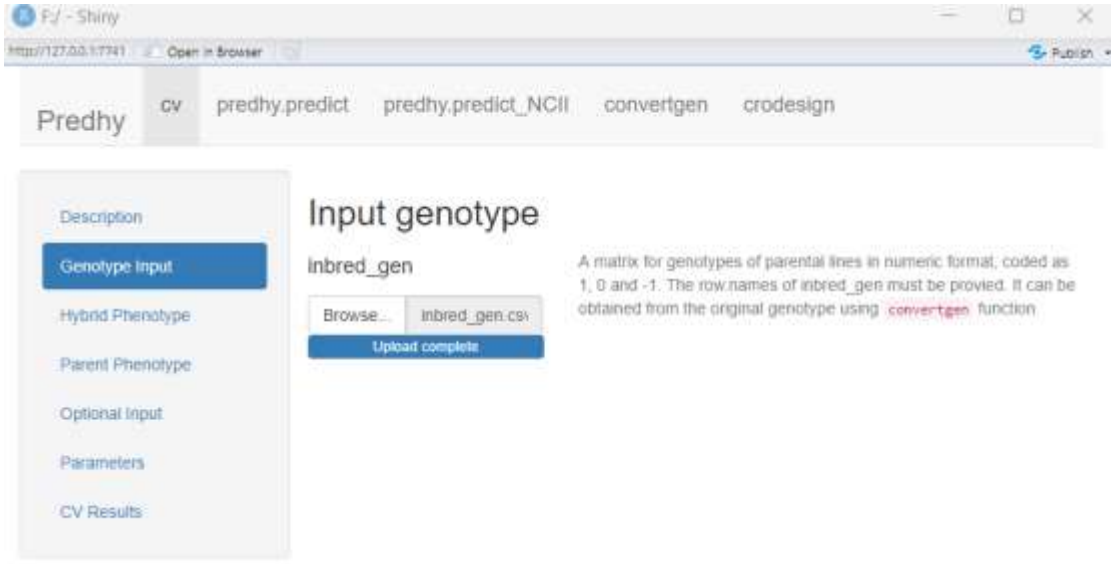
F
A008
A008
A008
A008
A010
A010
A010
A010
A010
A010
A011
A011
A011
A011
A012

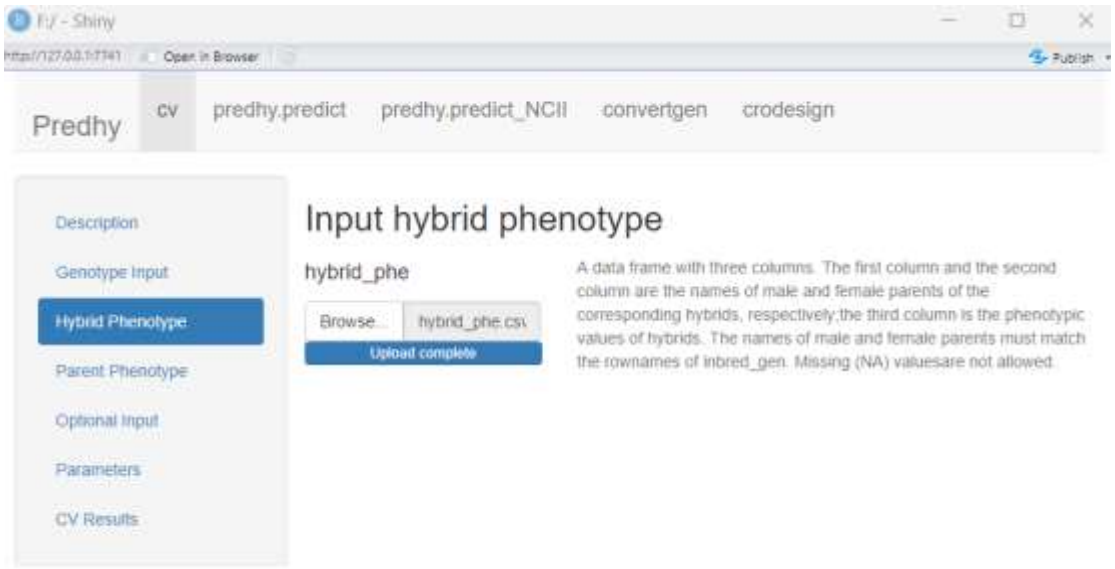
3. Operation process

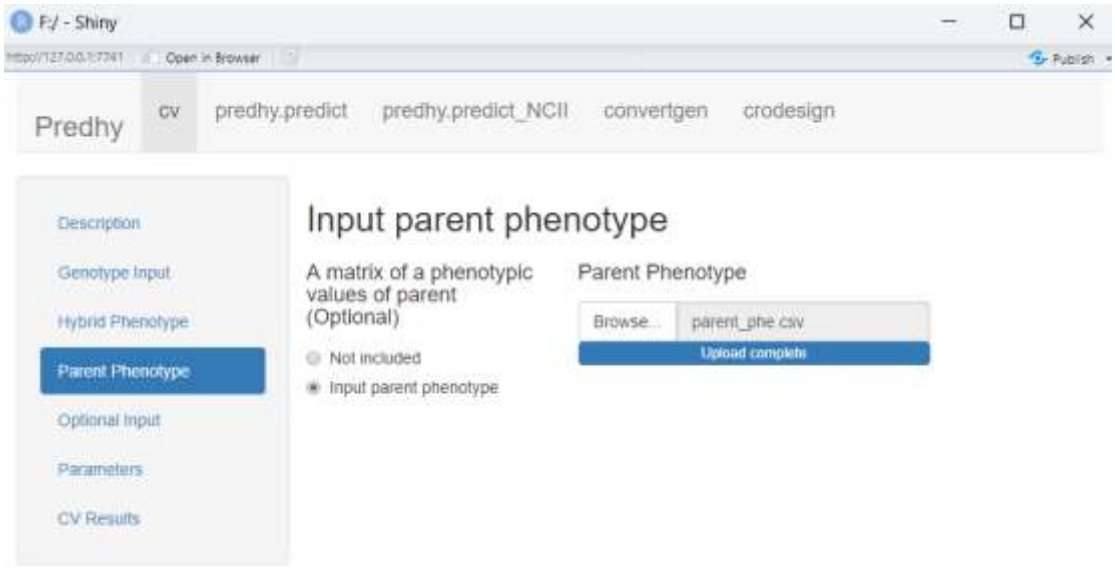
3.1 cv

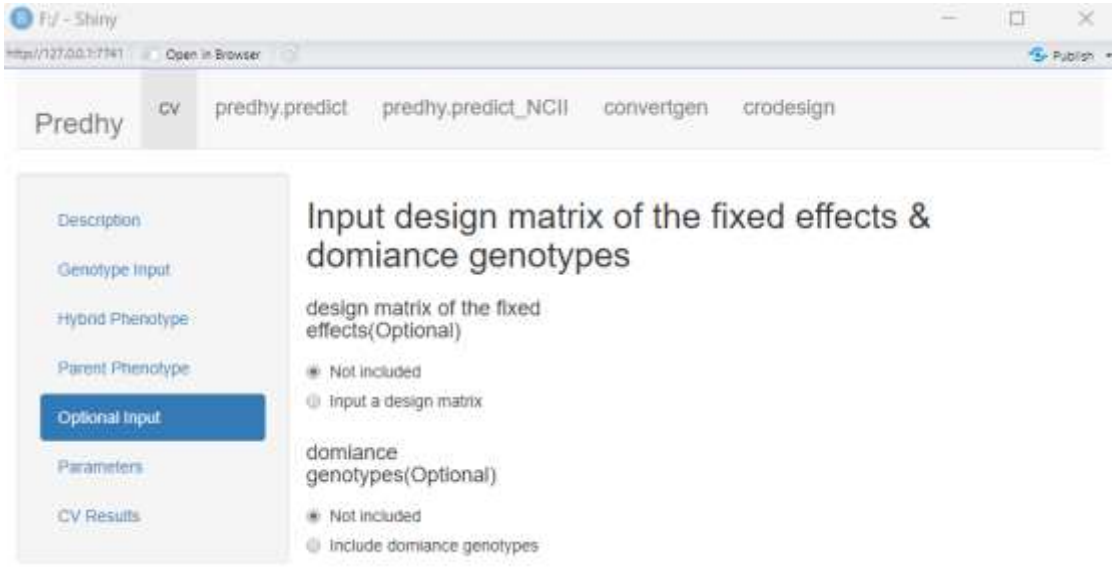
Dataset Input

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


F:/ - Shiny
Open in Browser
Publish
Predhy
cv
predhy.predict
predhy.predict_NCII
convertgen
crodesign
Description
Genotype Input
Hybrid Phenotype
Parent Phenotype
Optional Input
Parameters
CV Results
Input genotype
inbred_gen
Browse...
inbred_gen.csv
Upload complete
A matrix for genotypes of parental lines in numeric format, coded as 1, 0 and -1. The row.names of inbred_gen must be provied. It can be obtained from the original genotype using convertgen function

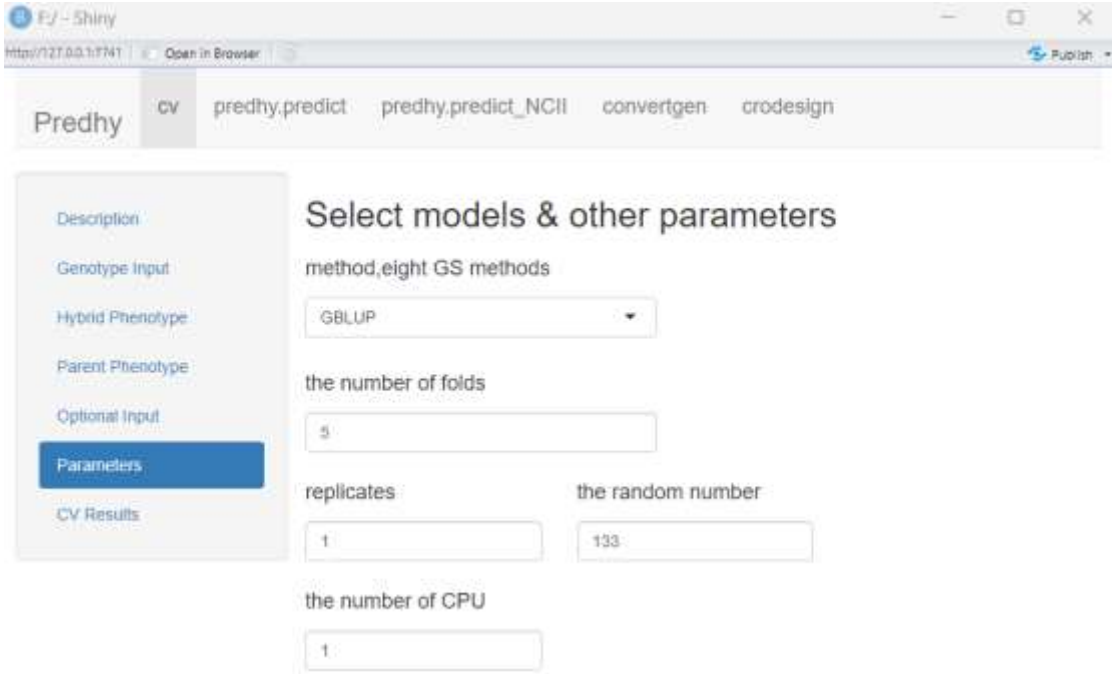

F:/ - Shiny
Open in Browser
Publish
Predhy
cv
predhy.predict
predhy.predict_NCII
convertgen
crodesign
Description
Genotype Input
Hybrid Phenotype
Parent Phenotype
Optional Input
Parameters
CV Results
Input hybrid phenotype
hybrid_phe
Browse...
hybrid_phe.csv
Upload complete
A data frame with three columns. 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) valuesare not allowed.


F:/ - Shiny
Open in Browser
Publish
Predhy
cv
predhy.predict
predhy.predict_NCII
convertgen
crodesign
Description
Genotype Input
Hybrid Phenotype
Parent Phenotype
Optional Input
Parameters
CV Results
Input parent phenotype
A matrix of a phenotypic values of parent (Optional)
Not included
Input parent phenotype
Parent Phenotype
Browse...
parent_phe.csv
Upload complete



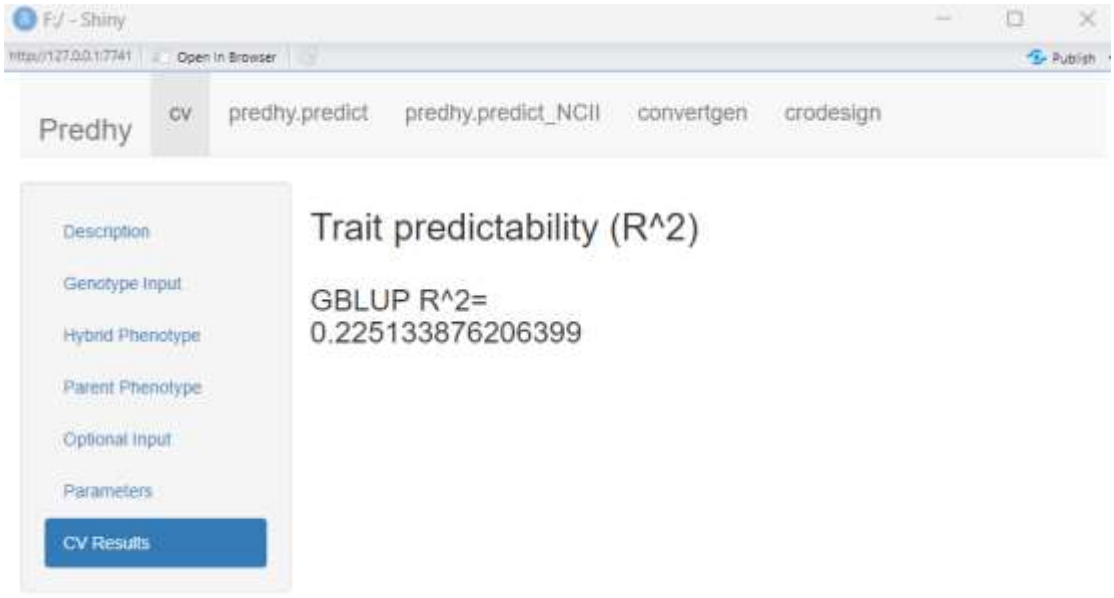
Method select & Parameter setting

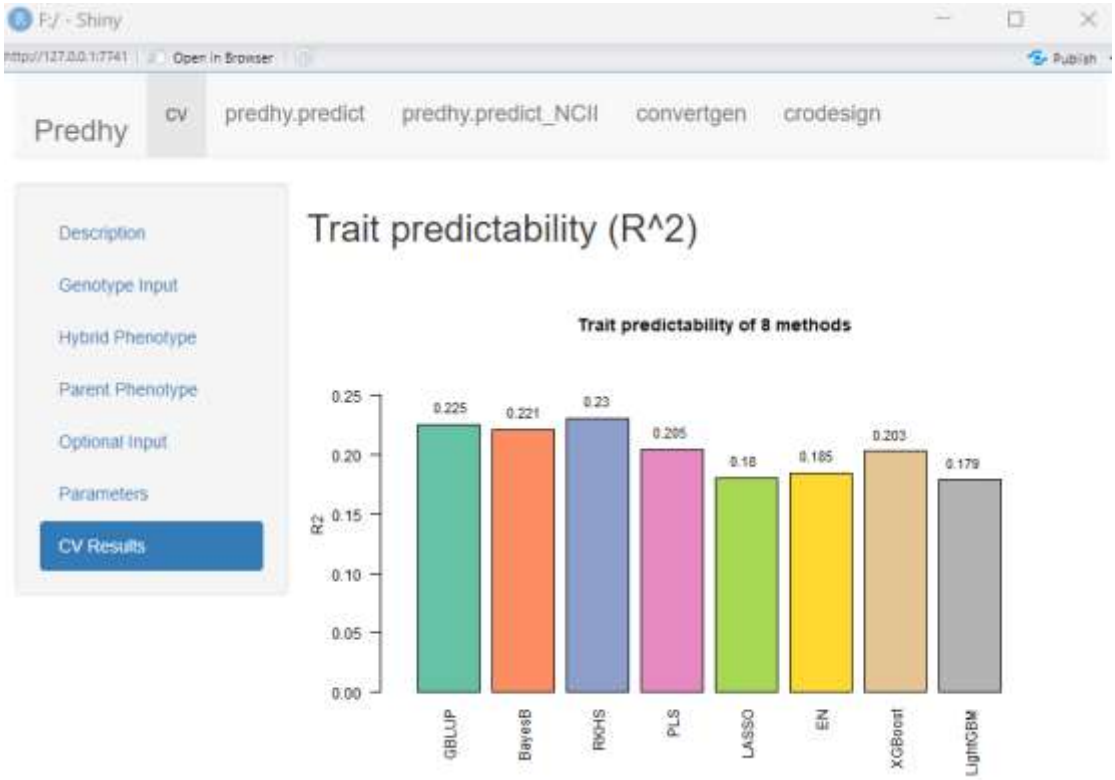
Method: There are eight GS methods in the predhy.GUI, including "GBLUP", "BayesB", "RKHS", "PLS", "LASSO", "EN", "XGBoost", "LightGBM". Users may select one of those methods or all of them simultaneously with "ALL".
Number of folds: The k for k-fold cross validation.
Replicates: Repeat number of independent replicates for the cross-validation.
The random number: The random number.
CPU: the number of CPU for parallel calculation.



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.



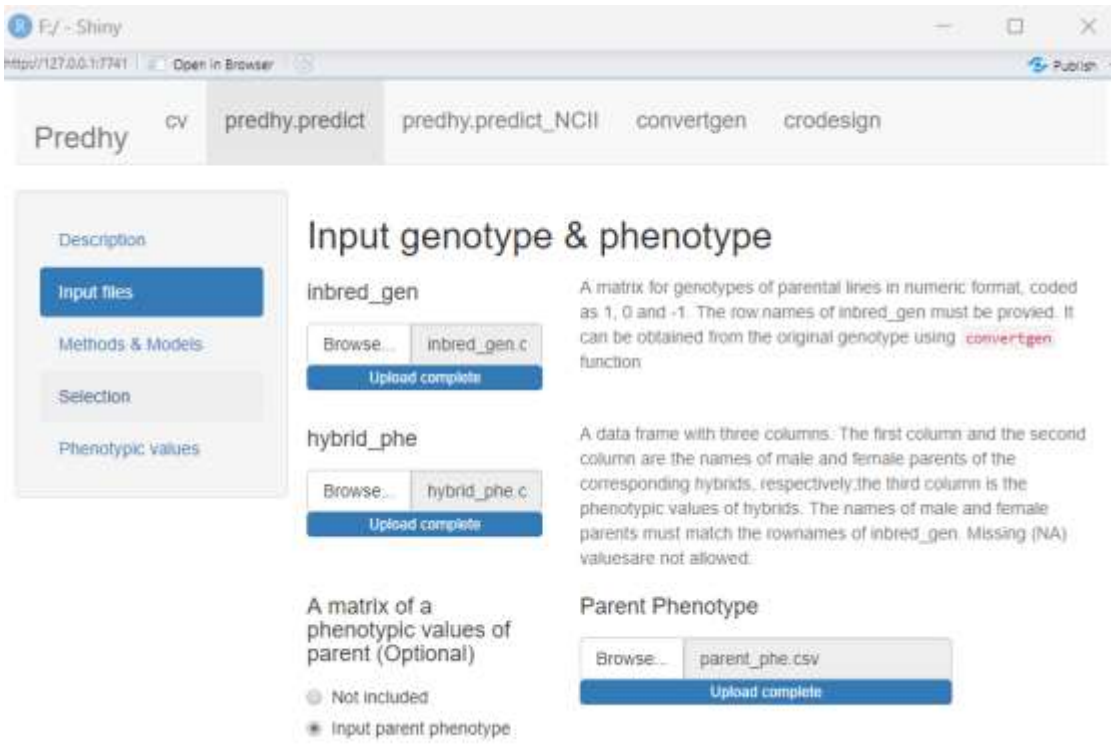


3.2 predhy.predict

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

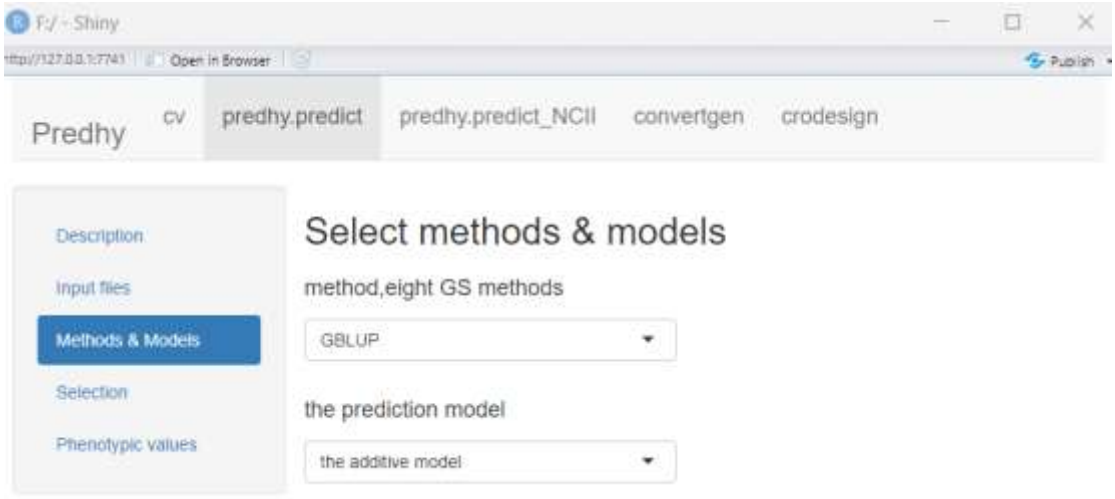
Dataset Input

Users must upload the inbred_gen and hybrid phenotype files, while the parent phenotype is optional.

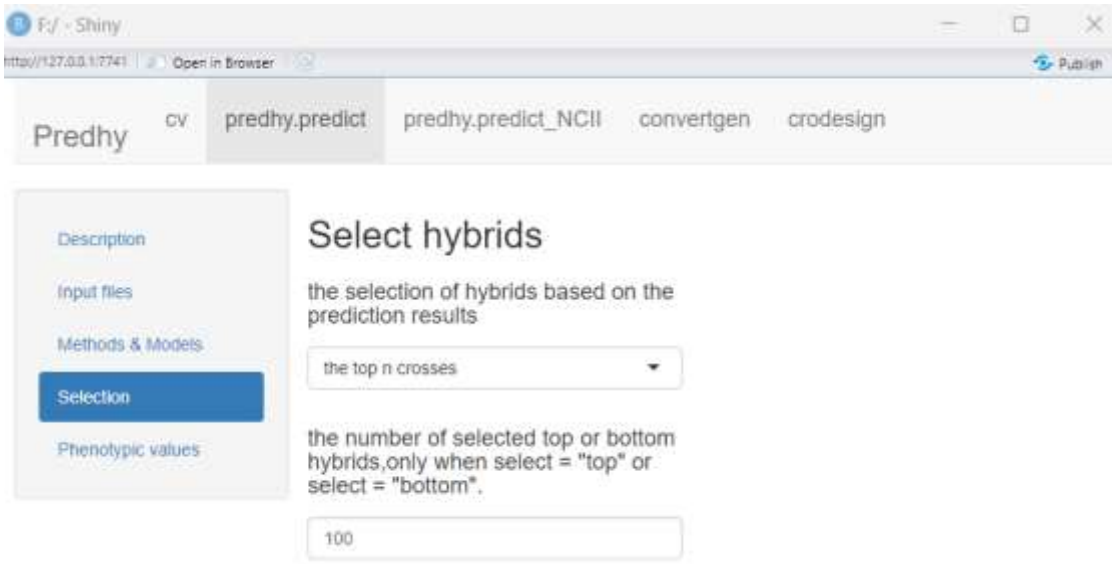


Method select & Parameter setting

Method: There are eight GS methods in the predhy.GUI for hybrid performance predicting, including "GBLUP", "BayesB", "RKHS", "PLS", "LASSO", "EN", "XGBoost", "LightGBM". Users should select one of those methods.
Prediction model: There are four options: the additive model, the additive-dominance model, the additive-phenotypic model, the additive-dominance-phenotypic model, user can choose one by select one of the choices.

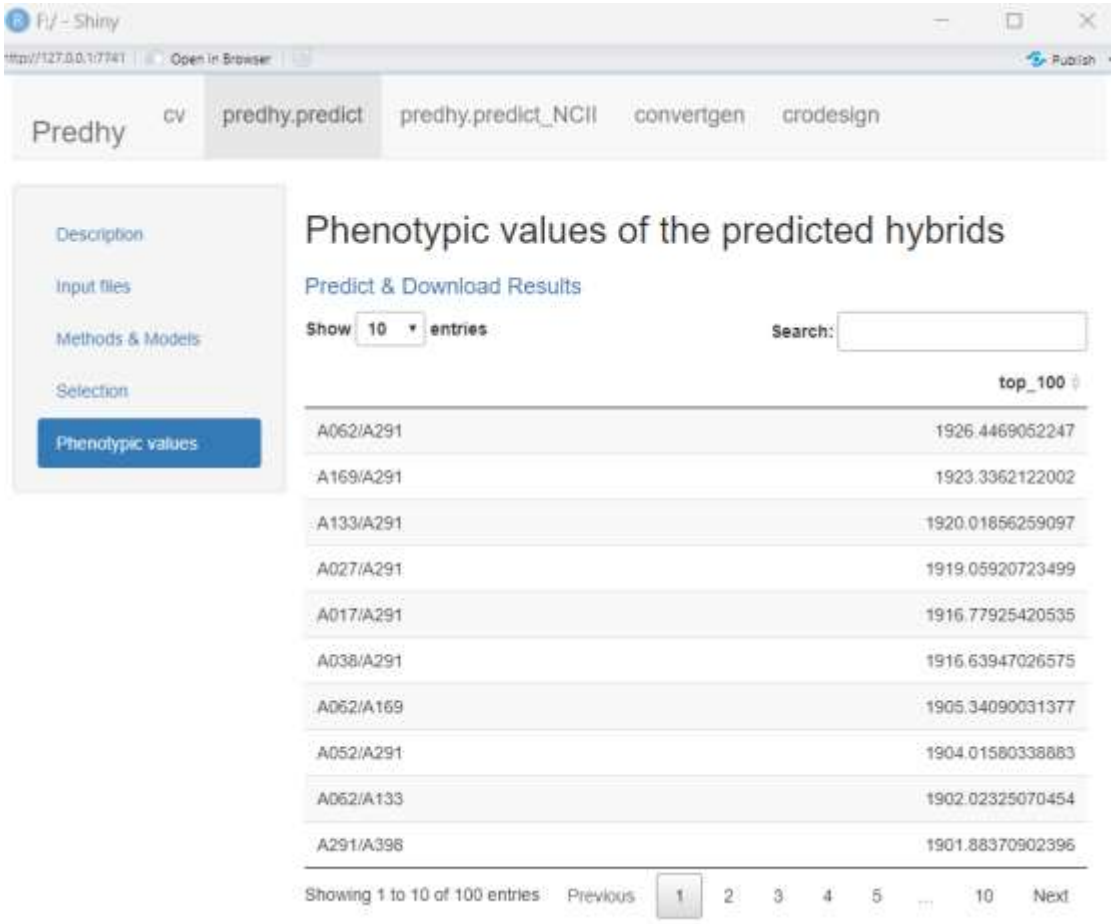


Select hybrids: Selection of hybrids based on the prediction results. There are three options: select = "all", which selects all potential crosses. select = "top", which selects the top n crosses. select = "bottom", which selects the bottom n crosses. User can decide number hybrids to select when select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



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.

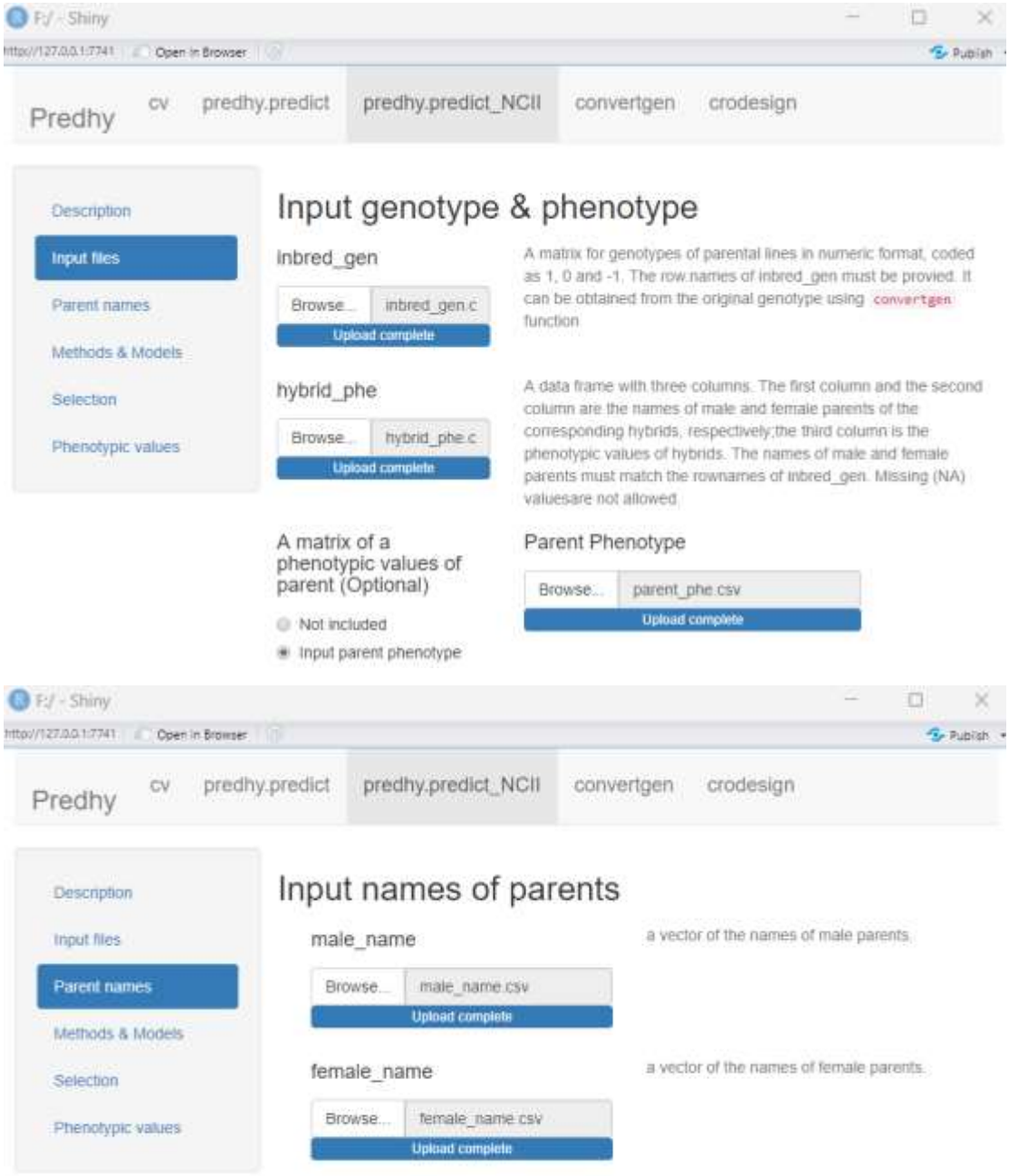


3.3 predhy.predict_NCII

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.

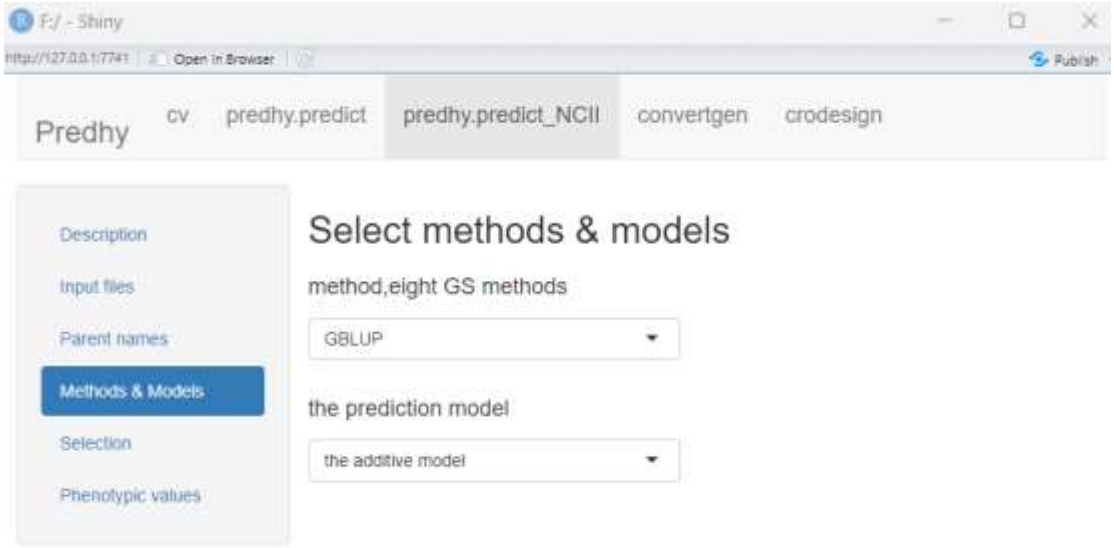
Dataset Input

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), while the parent phenotype is optional.

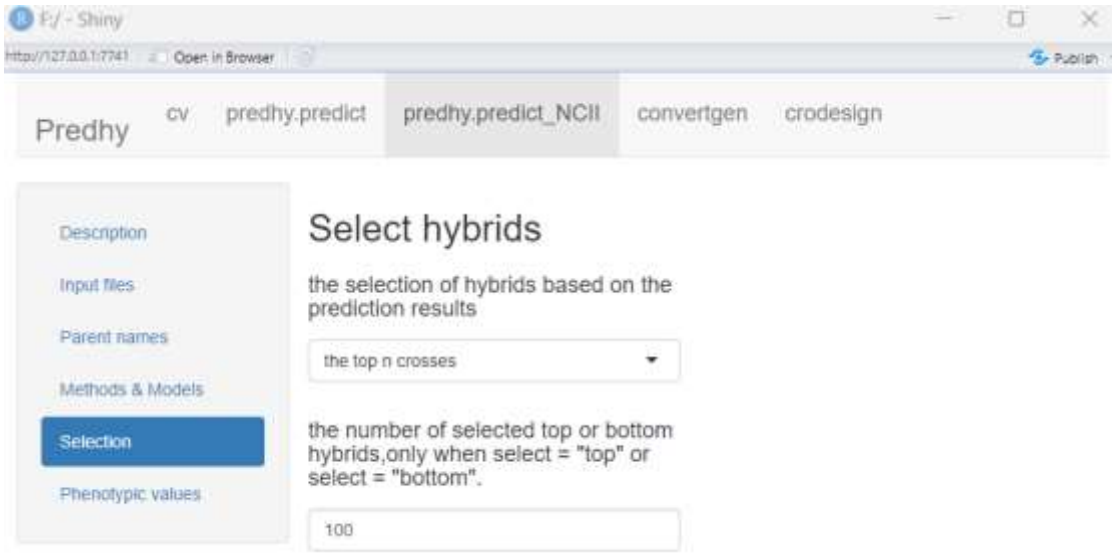


Method select & Parameter setting

Method: There are eight GS methods in the predhy.GUI for hybrid performance predicting, including "GBLUP", "BayesB", "RKHS", "PLS", "LASSO", "EN", "XGBoost", "LightGBM". Users should select one of those methods.
Prediction model: There are four options: the additive model, the additive-dominance model, the additive-phenotypic model, the additive-dominance-phenotypic model, user can choose one by select one of the choices.

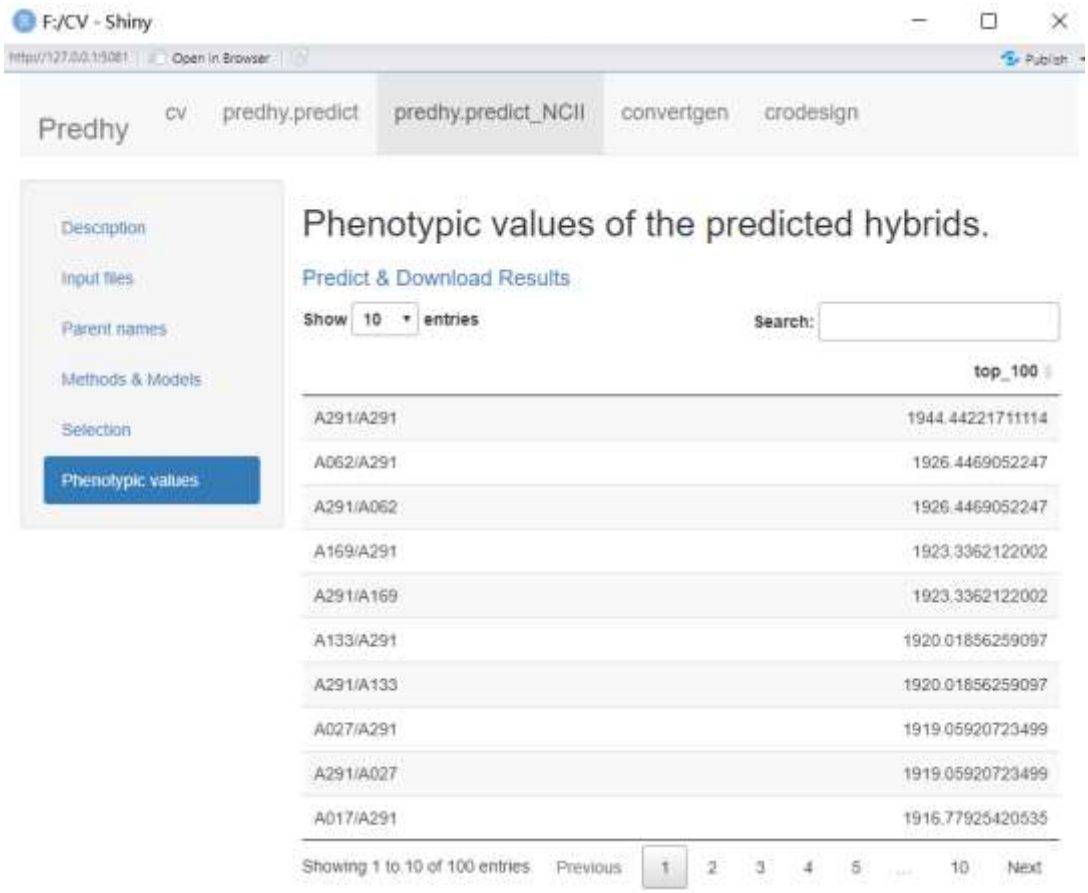


Select hybrids: Selection of hybrids based on the prediction results. There are three options: select = "all", which selects all potential crosses. select = "top", which selects the top n crosses. select = "bottom", which selects the bottom n crosses. User can decide number hybrids to select when select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



Run the software

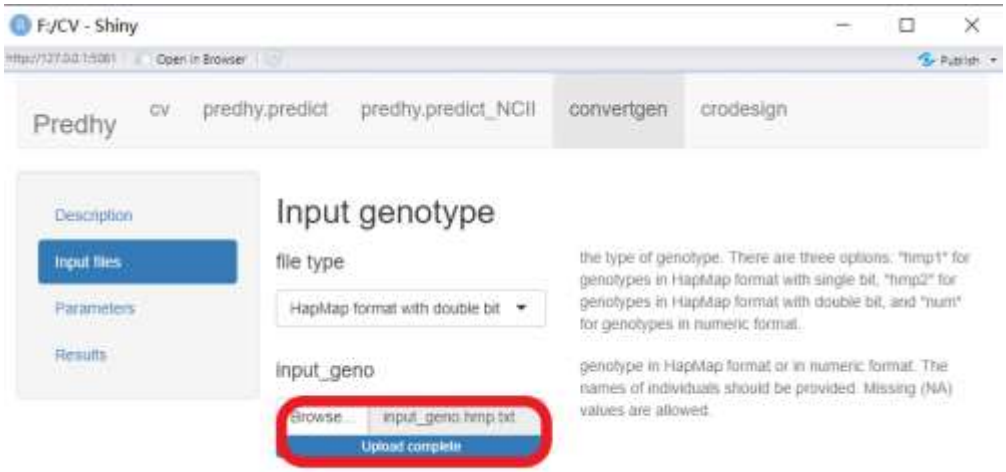
After uploading all the needed files and setting the parameters, users can run the Software simply by clicking “Phenotypic values”. When calculation is down, the result will be given in the datatable below the panel, user may download the full data by clicking at “Predict & Download Results” bottom.


F:/CV - Shiny
Open in Browser
Publish
Predhy
cv
predhy.predict
predhy.predict_NCII
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 1916.77925420535
Showing 1 to 10 of 100 entries
Previous
1
2
3
4
5
...
10
Next

3.4 convertgen

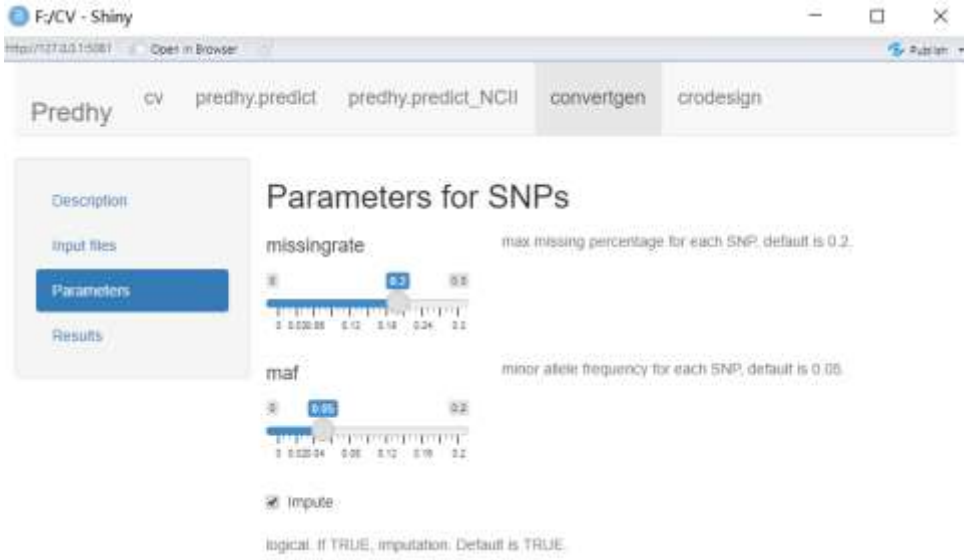
Dataset Input

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.



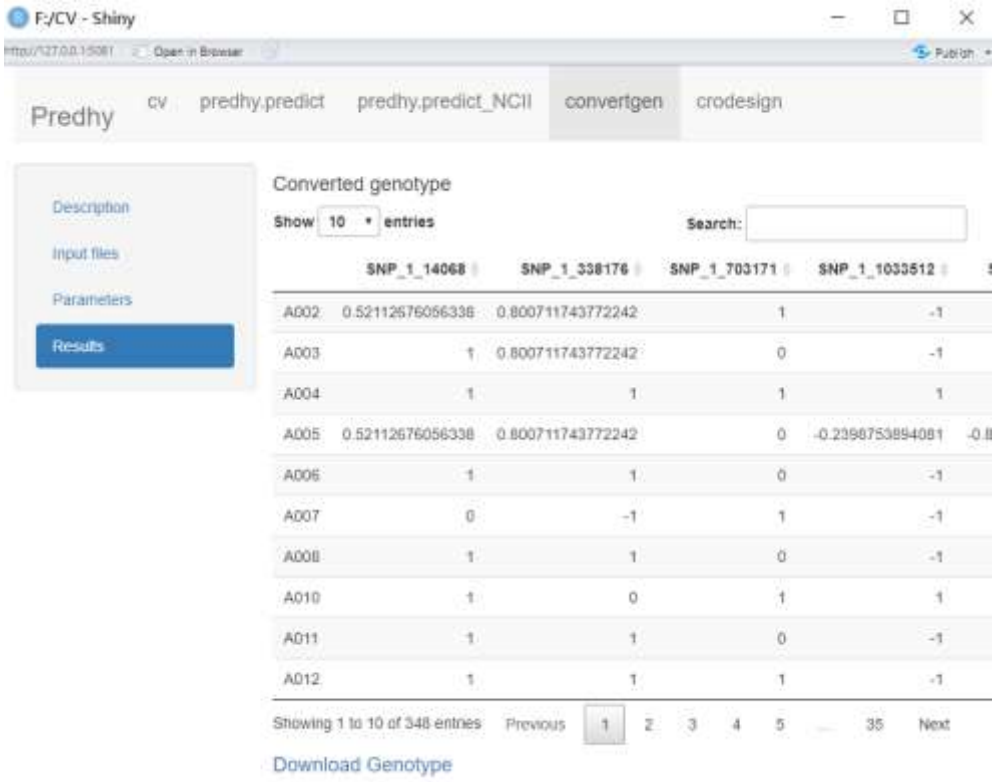
Method select & Parameter setting

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.



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.



3.5 crodesign

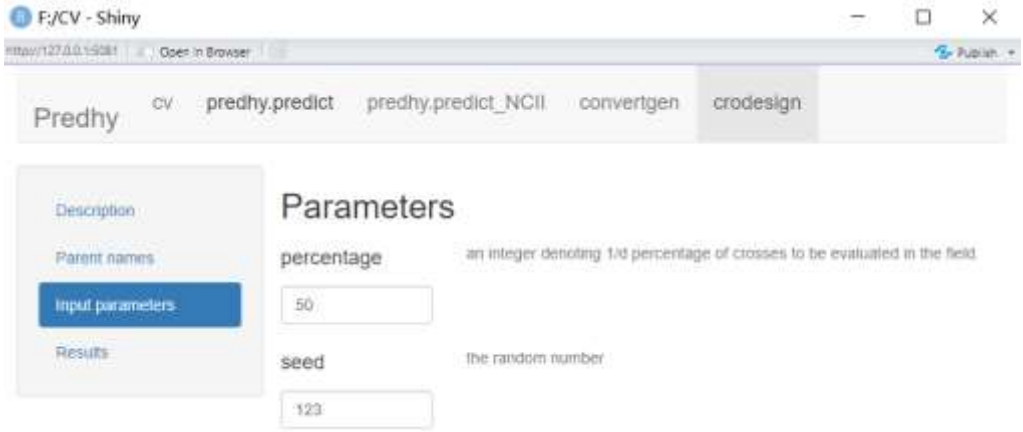
This function was designed to generate a mating design for a subset of crosses based on a balanced random partial rectangle cross-design (BRPRCD) (Xu et al. 2016).

Dataset Input

Users need to upload the Parent names dataset(two files, one contains male_names, the other contains female_names).

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.



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.