

# User Manual

*for*

## ISMU 2.0<sub>(beta)</sub>

*An Integrated SNP Mining & Utilization Pipeline*

### Genomic Selection

By

**International Crops Research Institute for the Semi-Arid Tropics (ICRISAT)**

*Abhishek Rathore, Sarwar Azam, Roma Rani Das, Manish Roorkiwal, Dadakhalar  
Doddamani, Mohan Telluri, David Marshall, Trushar Shah, A Bhanu Prakash, Dave Edwards,  
Alain Charcosset, Mark Sorrells, John M Hickey, Jean-Luc Jannink, Rajeev K Varshney*

**For further details, suggestion and reporting bugs please contact:**

Dr. Rajeev K. Varshney ([r.k.varshney@cgiar.org](mailto:r.k.varshney@cgiar.org))

Dr. Abhishek Rathore ([a.rathore@cgiar.org](mailto:a.rathore@cgiar.org))

**About:** ISMU 2.0 is developed to carry out Genomic Selection analysis from the available Genotypic and Phenotypic data. It can directly import data generated from the ISMU 1.0 together with available phenotypic information. Several data processing capabilities and genomic selection (GS) modules are being integrated in the existing pipeline. Most commonly used method for genomic selection methods like Ridge Regression Best Linear Unbiased Predictor (RR-BLUP), Kinship Gauss, Bayesian LASSO, BayesB, BayesCpi and Random Forest are being customized and integrated in to pipeline. As it is a beta version, we recommend validating of results by other means before using them.

### System Requirement:

Operating system: Windows XP, Windows 7, CentOS 6.2\* / Ubuntu 12\*

\* JRE 1.7 (or above) and R 3.0.1 (or above) with following packages (genetics, imputation, rrBLUP, BLR, randomForest, R2HTML and multicore)

## 1. Data Input

### File format:

ISMU 2.0 uses its own formats for data. For Genomic Selection analysis, two input files are mandatory.

**(a) Genotype / Marker Data:** Markers should be arranged in rows and genotypes in columns. It accepts both dominant and co-dominant markers (SNP). For co-dominant markers, alleles can be separated by a single character or with no separator.

**(b) Phenotype Data:** First column contains the genotype labels and phenotype values of single/multiple traits are arranged from second column.

Both genotype and phenotype files should be saved in comma separated values file format.

\* Please see demo files in **Sample** Folder

### Missing values

Missing values must be specified as "NA" in both the files.

**Other Files:** Relationship matrix, pedigree data & population structure (beta)

## 2. Installation of ISMU 2.0:

Installation of ISMU pipeline is fairly simple. One needs to copy folder "ISMU2.0" from CD to any user folder of workstation.

## 3. Executing ISMU 2.0 Pipeline:

### *For Windows:*

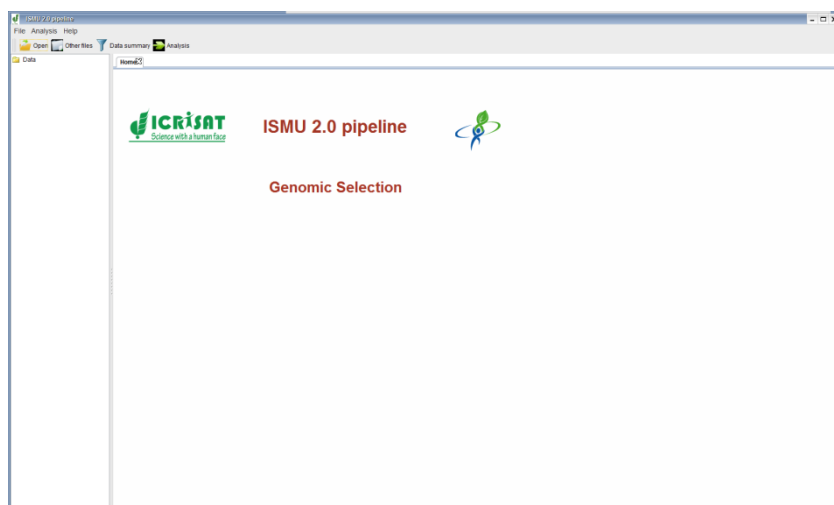
To run ISMU 2.0 double click on *ISMU2.0x64.bat* or *ISMU2.0x32.bat* file based on configuration of your workstation.

### *For CentOS / Ubuntu:*

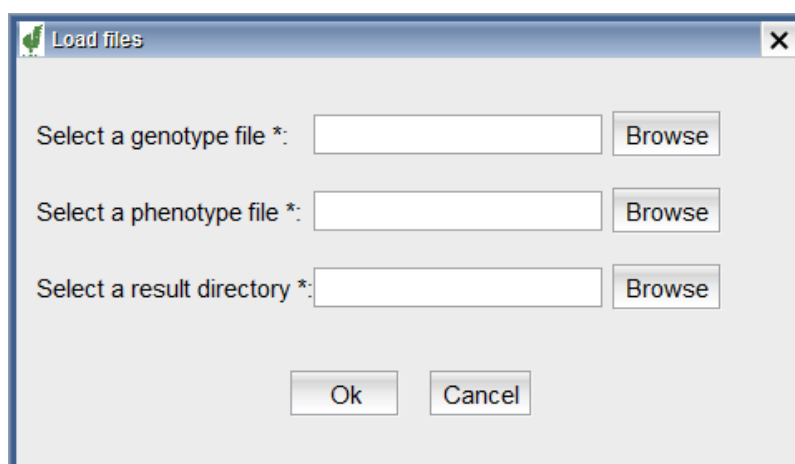
*To run ISMU 2.0 on command prompt type:*

***sh ISMU2.0.sh***

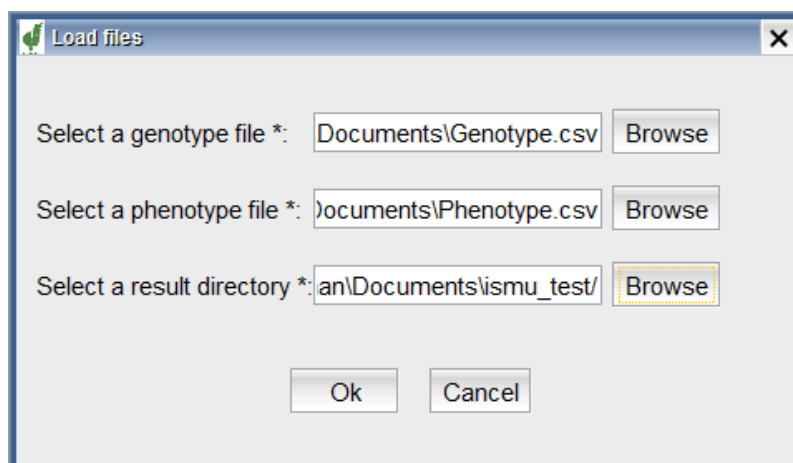
The GUI of ISMU 2.0 pipeline appears on user's screen as follows:



**(a) Loading of Input Files:** To start using the application, first go to menu File → Open. A GUI of open dialog box appears



Browse and load genotypic and phenotypic files. For the result directory create a folder and select it where all the results will be stored and this folder will also act as user's working directory.



On Click Ok button, the loaded files are displayed in home screen.

ISMU 2.0 pipeline

File Analysis Help

Open Other files Data summary Analysis

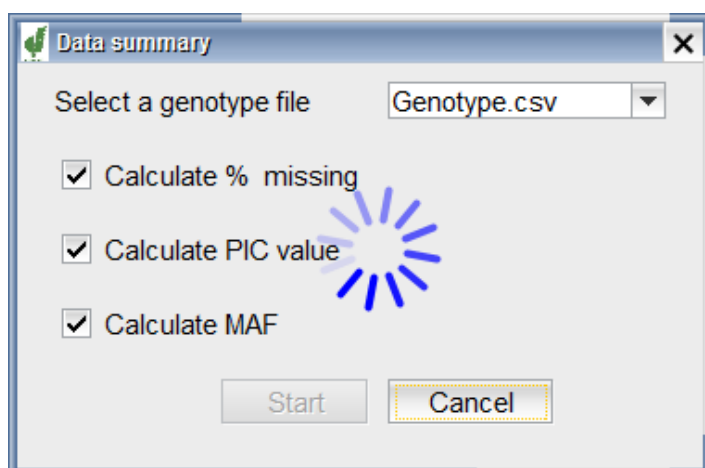
Data

- Genotype
- Genotype.csv
- Phenotype
- Phenotype.csv
- Relationship matrix
- Pedigree data
- Population structure
- Result Directory
- Log

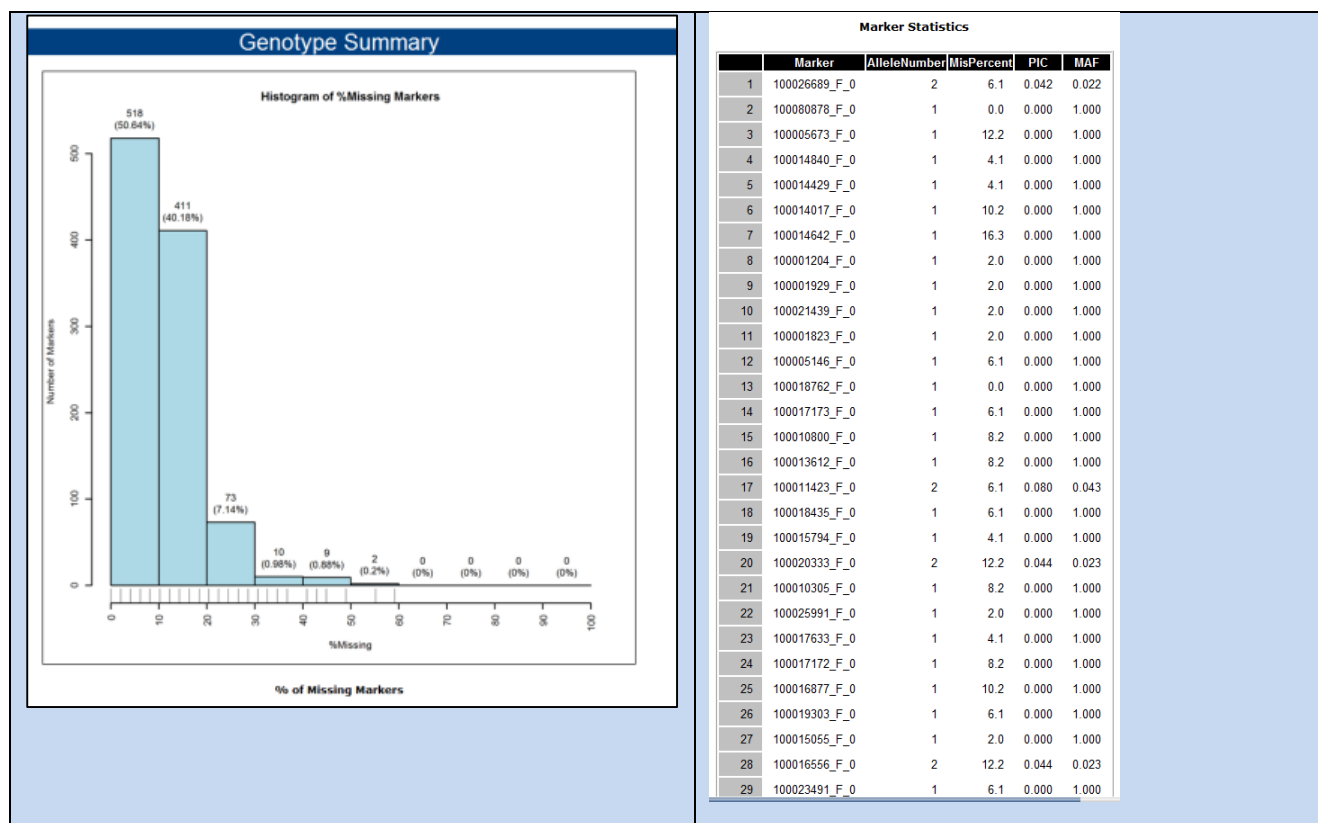
Marker	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14	G15	G16	G17	G18
100014840...	1	1	NA	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
100014429...	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
100014017...	1	1	1	NA	1	1	1	1	1	1	1	1	1	NA	1	1	1	1
100014642...	1	1	1	NA	1	1	1	1	1	1	1	1	1	1	1	1	1	NA
100001204...	1	1	1	1	1	1	1	1	1	1	1	NA	1	1	1	1	1	1
100001929...	1	1	1	1	1	1	1	1	1	1	1	NA	1	1	1	1	1	1
100021439...	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
100001823...	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
100005146...	1	1	NA	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
100018762...	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
100017173...	1	1	1	1	1	1	1	1	1	1	1	NA	1	1	1	1	1	1
100010800...	1	NA	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
100013612...	1	NA	1	1	1	1	1	1	1	1	1	NA	1	1	1	1	1	1
100011423...	1	1	NA	1	1	1	NA	1	1	1	1	1	NA	1	1	1	1	1
100018435...	1	1	1	1	1	1	NA	1	1	1	1	1	1	1	1	1	1	1
100015794...	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
100020333...	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	NA
100010305...	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	NA
100025991...	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
100017633...	1	1	1	NA	1	1	1	1	1	1	1	1	1	1	1	1	1	1
100017172...	1	NA	1	1	1	1	1	NA	1	1	1	1	1	1	1	1	1	1
100016877...	NA	1	1	1	1	NA	1	1	1	1	1	1	1	1	1	1	1	1
100019303...	1	1	1	1	1	NA	1	1	1	1	1	1	1	1	1	1	1	1
100015055...	1	1	1	1	1	1	1	1	1	1	1	NA	1	1	1	1	1	1
100016556...	1	1	1	1	1	1	1	1	1	1	1	1	NA	1	1	1	1	1
100023491...	1	1	1	1	NA	1	1	1	1	1	1	1	1	1	1	1	1	1
100018164...	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
100002825...	NA	1	1	1	1	1	1	NA	NA	1	1	1	1	1	1	1	1	1
100012208...	1	1	1	NA	1	1	1	1	1	1	1	1	1	NA	1	1	1	1
100030189...	1	1	1	1	NA	NA	1	1	1	1	1	1	1	1	1	1	NA	NA
100017145...	1	1	1	1	1	1	1	1	1	1	1	0	1	NA	1	1	1	1
100015746...	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
100004661...	1	1	NA	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
100017590...	NA	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	NA
100021128...	1	1	1	1	NA	1	1	1	1	1	1	1	1	1	1	1	1	1
100020360...	1	1	1	NA	NA	1	1	1	1	1	1	1	NA	NA	NA	NA	NA	NA
100036201...	1	1	1	1	NA	1	1	1	NA	1	NA	1	1	1	NA	1	1	NA
100019139...	NA	1	1	1	1	1	1	1	1	1	1	1	1	1	NA	1	1	1
100018479...	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
100025396...	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	NA	1
100017022...	NA	1	1	1	1	1	1	1	NA	NA	1	1	NA	1	1	1	1	1
100038481...	1	1	1	1	1	1	1	NA	1	1	NA	NA	1	1	1	1	NA	1

Genotype	Trait1	Trait2
G1	6.27	6.70545768
G2	4.478	4.6161323
G3	3.873	4.0557819
G4	6.871	7.6346435
G5	17.135	17.574085
G6	-3.358	-2.6491491
G7	8.491	8.8618438
G8	-14.047	-13.512185
G9	-13.282	-12.889361
G10	-7.6	-6.8687318
G11	-1.505	-0.74581737
G12	-7.636	-6.7533301
G13	-2.587	-1.8210518
G14	12.204	12.710490
G15	-0.518	0.42118718
G16	-6.531	-5.9538143
G17	-4.477	-4.10482738
G18	-14.712	-13.985555
G19	-1.34	-0.9350263
G20	3.128	3.7535075
G21	-7.071	-6.7463516
G22	-7.072	-6.7771447
G23	17.743	18.5774735
G24	-1.332	-1.0565223
G25	22.077	21.920236
G26	-1.12	-0.6706805
G27	12.034	12.751560
G28	10.606	10.636101
G29	-12.407	-11.992284
G30	-3.011	-2.7833136
G31	4.934	5.2866814
G32	-7.313	-6.8638112
G33	-2.646	-1.7420079
G34	17.136	18.091697
G35	12.345	12.458136
G36	12.789	13.708263
G37	6.105	6.5613664
G38	0.889	1.33365882
G39	1.376	2.1011192
G40	0.911	1.1888096
G41	7.753	8.5190906
G42	-5.578	-4.80939436
G43	5.48	6.4135364

**(b) Genotypic data Summary:** A basic summary statistics for genotypic data will be calculated using **Data summary** module. Click on "Data summary" tab and select the required loaded genotypic file.



This module has an option to calculate individual marker's percentage of missing values, Polymorphic Information Content (PIC) & Minor Allele Frequency (MAF). A summary table and histogram of the selected statistic is generated.



**(c) Running GS models:** Six different models (*Ridge Regression BLUP*, *Bayes Cpi*, *BayesB*, *Bayes LASSO*, *Random Forest* and *Kinship Gauss*) are implemented for fitting Genomic Selection. For GS analysis

- Select loaded genotype & phenotype files from the drop down menu
- Select the traits for which user want to fit GS model
- Select different methods for analysis
- As a data cleaning steps, remove few markers, if required, with specified percentage (%) of missing markers, PIC and Minor allele frequency (MAF) values

**Analysis**

**Select file names from combo box**

Genotype file name : Genotype.csv  
 Phenotype file name : Phenotype.csv  
 Relationship matrix file name : Select  
 Pedigree data file name : Select  
 Population structure file name : Select

**Select a method(s) to start analysis**

☒ Ridge Regression BLUP  
☒ Bayes Cpi  
☒ BayesB  
☒ Bayes LASSO  
☒ Random Forest  
☒ Kinship Gauss

**Parameters**

Percentage(%) of missing markers : 30  
 PIC value : 0  
 Minor allele frequency (MAF) : 0.1

**Additional Parameters**

☐ Cross validation  
 Replication : 2    Fold : 5  
 Bayes : 0    0    0  
 Random Forest : 0  
 No of processors (cpu's) : 1

**Select the trait(s) for analysis**

Trait2    Trait1

> >> << <

**Start** **Cancel**

Click on **Start**. The process will start and a screen will appear showing the progress status of each trait running for each selected method. Trait wise analysis is done for all different selected methods.

**Analysis**

Performing analysis of trait : Trait1  
 Present running method : Data Processing

☐ RidgeRegression  
☐ BayesCpi  
☐ BayesB  
☐ BayesLasso  
☐ RandomForest  
☐ KinshipGauss

**Cancel**

**Analysis**

Performing analysis of trait : Trait1  
 Present running method : BayesB...

☒ RidgeRegression  
☒ BayesCpi  
☐ BayesB  
☐ BayesLasso  
☐ RandomForest  
☐ KinshipGauss

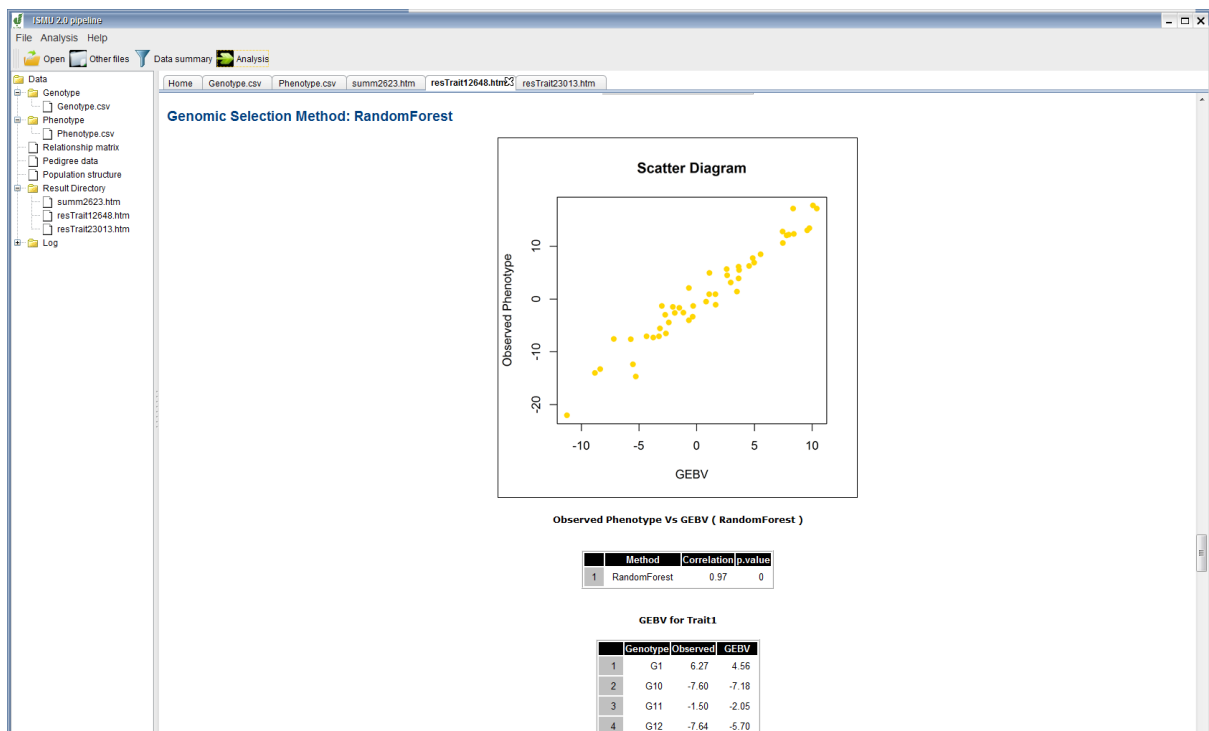
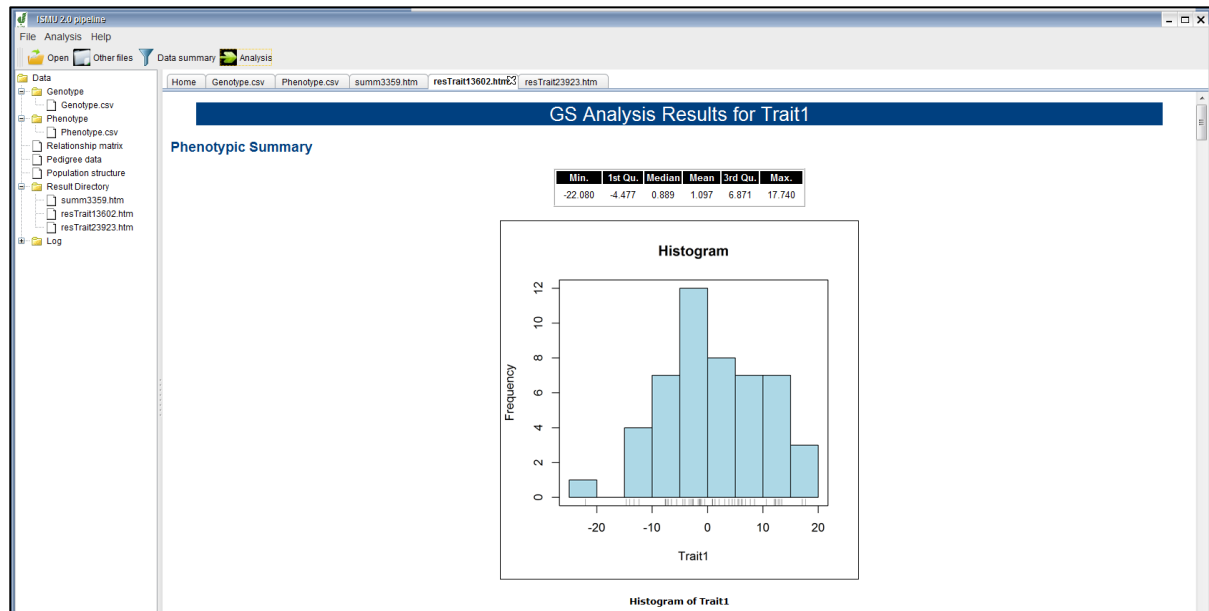
**Cancel**

**(d) Output:** A html file is generated for each trait with all selected GS method. For each selected trait, result contains

- Data summary, histogram and boxplot for phenotypic data
- Genomic Estimated Breeding Values (GEBV) for different selected methods



- Scatter plots between observed phenotype and GEBV's for different selected methods
- Summary table showing the correlation and p-value between observed phenotype and GEBV's for different selected methods

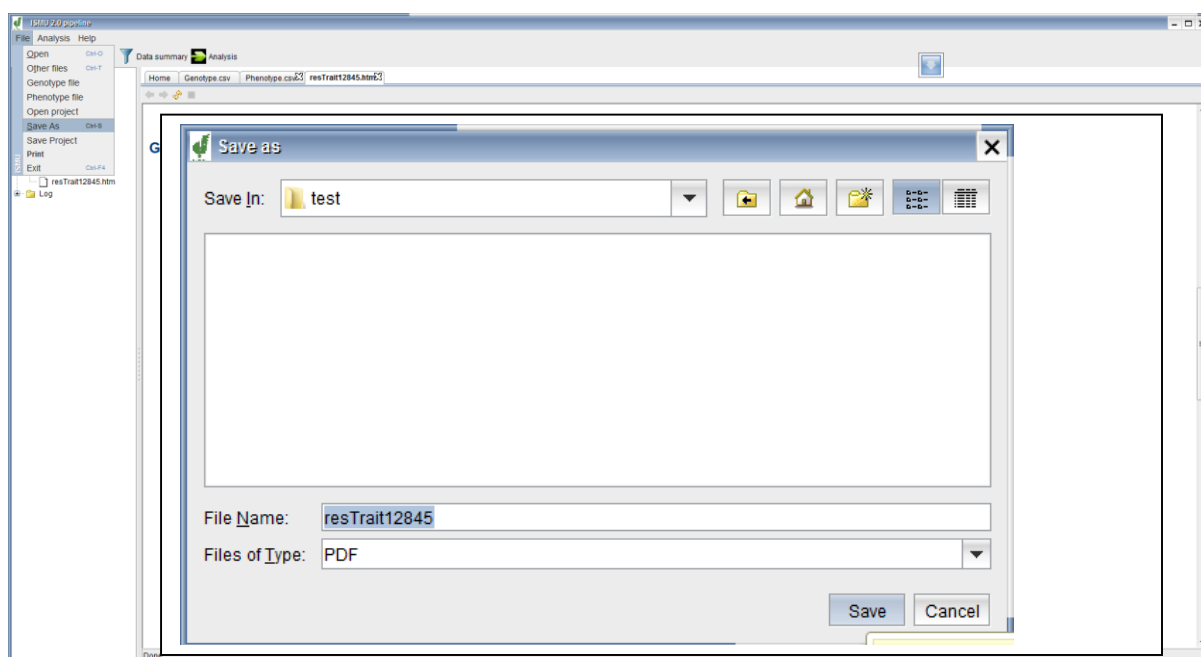


## Summary of Selected GS Methods

	Method	Correlation	Prob.t
1	RidgeRegression	0.551	0.000
2	BayesCpi	0.025	0.866
3	BayesB	0.257	0.074
4	BayesianLASSO	0.041	0.780
5	RandomForest	0.972	0.000
6	KinshipGauss	0.561	0.000

Results Generated by ISMU 2.0 : Tue Sep 24 2:30:13 PM 2013

The output file in html format can also be saved and imported in pdf format in desired folder.



----END----