User Manual

for

ISMU 2.0(beta)

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, Dadakhalandar 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)

Dr. Abhishek Rathore (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:

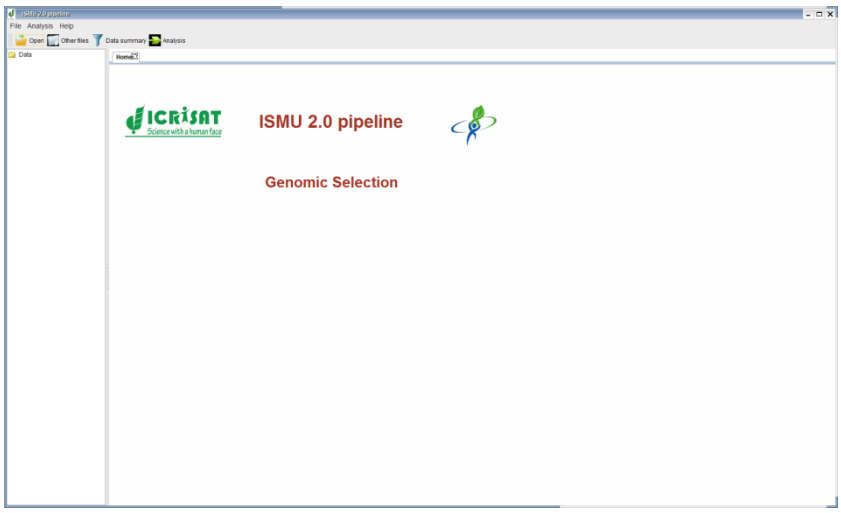
To run ISMU 2.0 double click on *ISMU2.0x64.bat* or *ISMU2.0x32.bat* file based on configuration of your workstation.

For CentOS / Ubuntu:

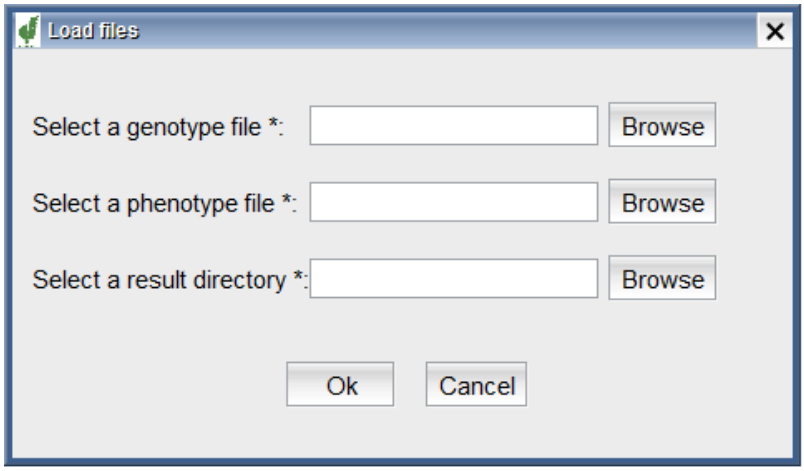
To run ISMU 2.0 on command prompt type:

sh ISMU2.0.sh

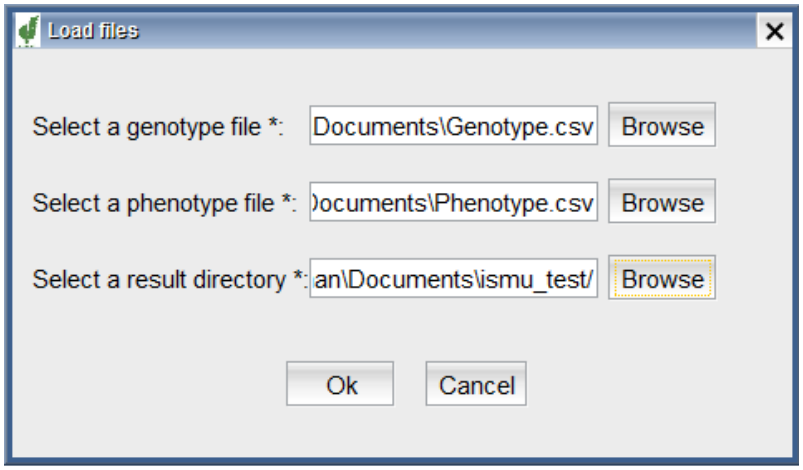
The GUI of ISMU 2.0 pipeline appears on user's screen as follows:



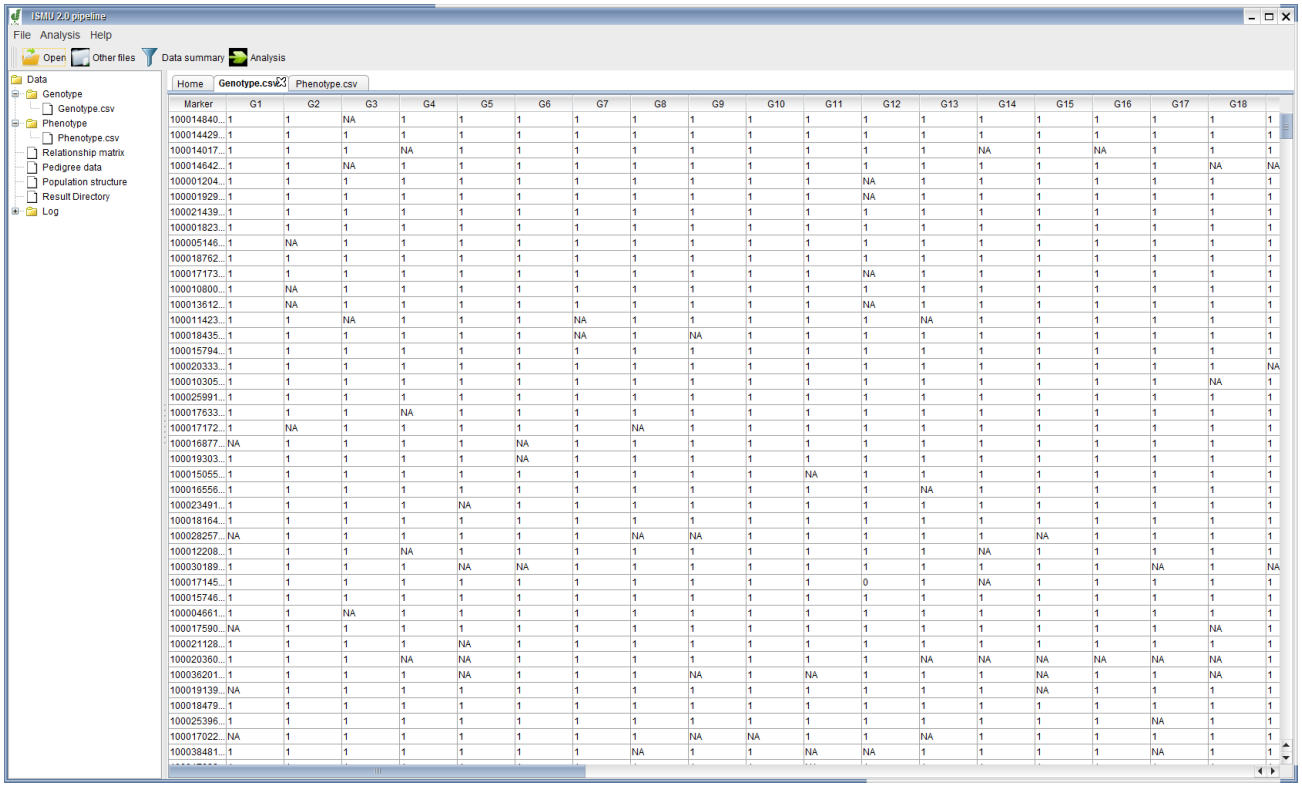
(a) Loading of Input Files: To start using the application, first go to menu File → Open. A GUI of open dialog box appears

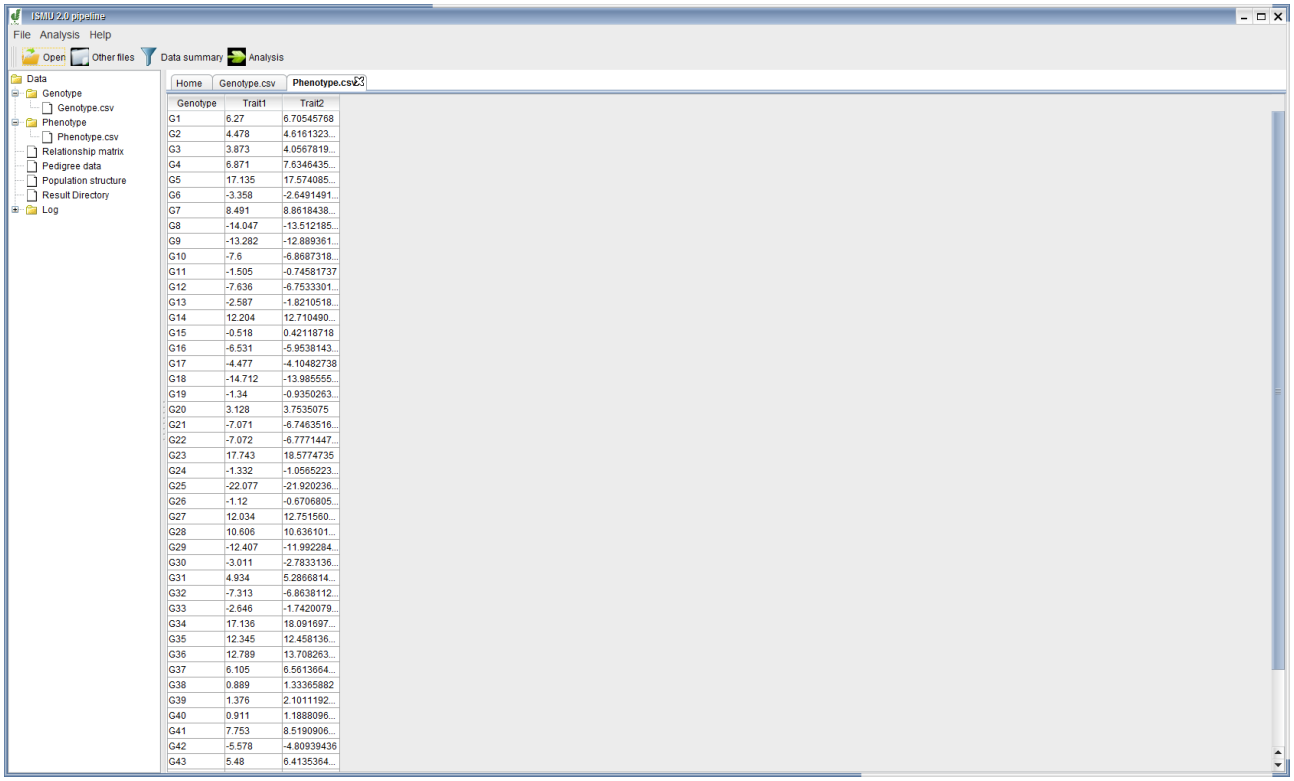


Browse and load genotypic and phenotypic files. For the result directory create a folder and select it where all the results will be stored and this folder will also act as user's working directory.

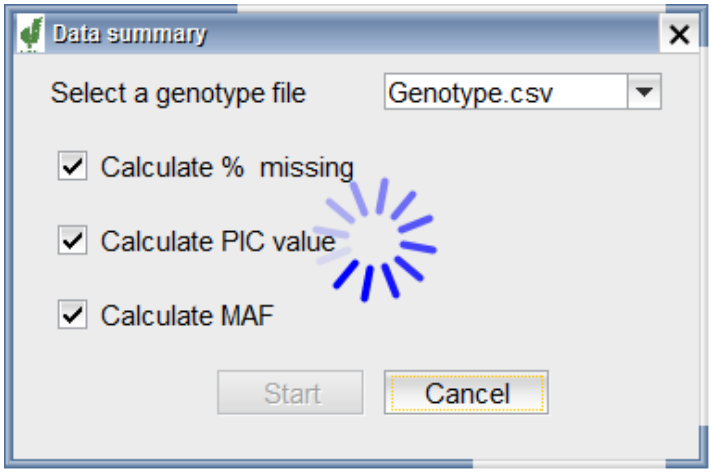


On Click **Ok** button, the loaded files are displayed in home screen.

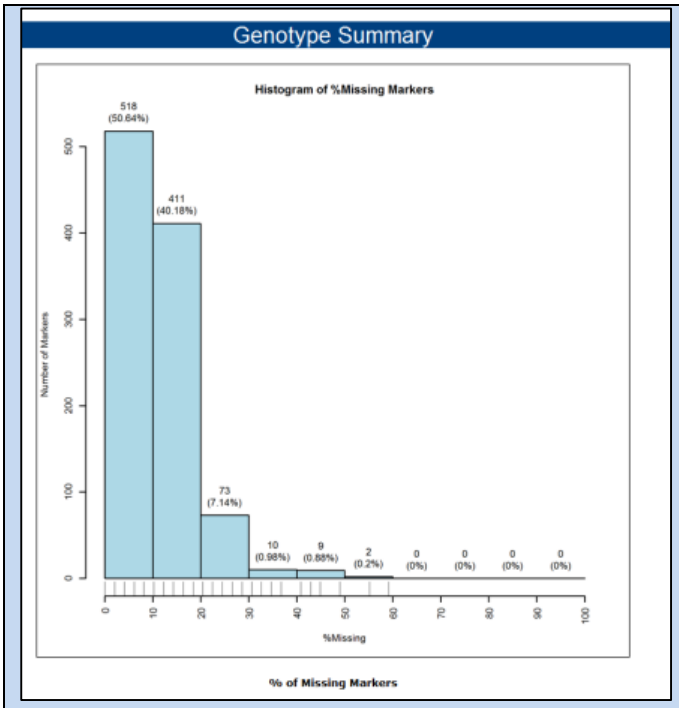




(**b) Genotypic data Summary:** A basic summary statistics for genotypic data will be calculated using **Data summary** module. Click on "Data summary" tab and select the required loaded genotypic file.



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.

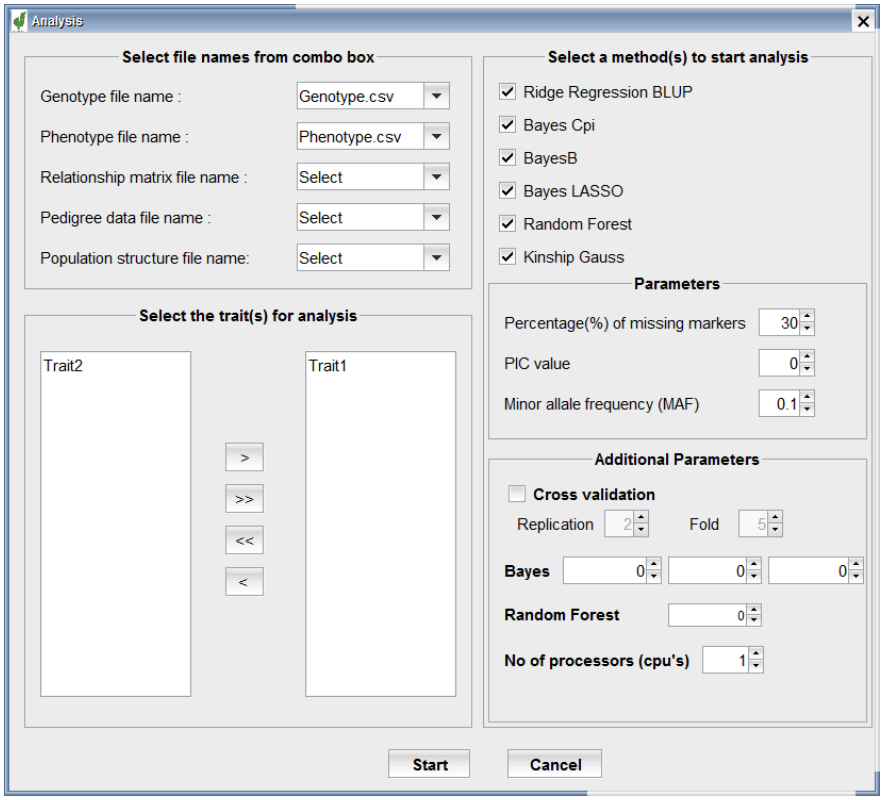


Marker Statistics

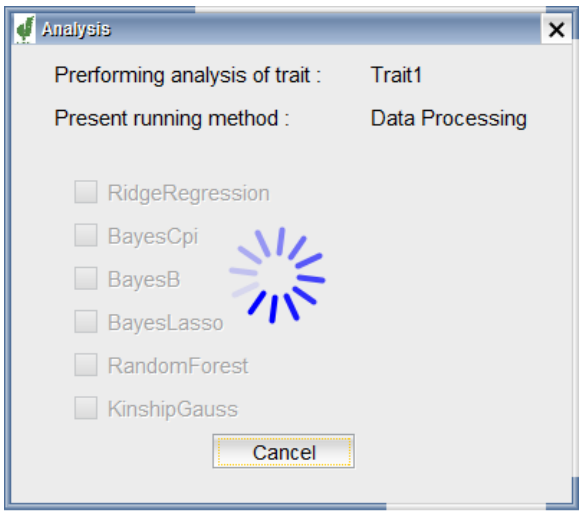
	Marker	AlleleNumber	MisPercent	PIC	MAF
1	100026689_F_0	2	6.1	0.042	0.022
2	100080878_F_0	1	0.0	0.000	1.000
3	100005673_F_0	1	12.2	0.000	1.000
4	100014840_F_0	1	4.1	0.000	1.000
5	100014429_F_0	1	4.1	0.000	1.000
6	100014017_F_0	1	10.2	0.000	1.000
7	100014642_F_0	1	16.3	0.000	1.000
8	100001204_F_0	1	2.0	0.000	1.000
9	100001929_F_0	1	2.0	0.000	1.000
10	100021439_F_0	1	2.0	0.000	1.000
11	100001823_F_0	1	2.0	0.000	1.000
12	100005146_F_0	1	6.1	0.000	1.000
13	100018762_F_0	1	0.0	0.000	1.000
14	100017173_F_0	1	6.1	0.000	1.000
15	100010800_F_0	1	8.2	0.000	1.000
16	100013612_F_0	1	8.2	0.000	1.000
17	100011423_F_0	2	6.1	0.080	0.043
18	100018435_F_0	1	6.1	0.000	1.000
19	100015794_F_0	1	4.1	0.000	1.000
20	100020333_F_0	2	12.2	0.044	0.023
21	100010305_F_0	1	8.2	0.000	1.000
22	100025991_F_0	1	2.0	0.000	1.000
23	100017633_F_0	1	4.1	0.000	1.000
24	100017172_F_0	1	8.2	0.000	1.000
25	100016877_F_0	1	10.2	0.000	1.000
26	100019303_F_0	1	6.1	0.000	1.000
27	100015055_F_0	1	2.0	0.000	1.000
28	100016556_F_0	2	12.2	0.044	0.023
29	100023491_F_0	1	6.1	0.000	1.000

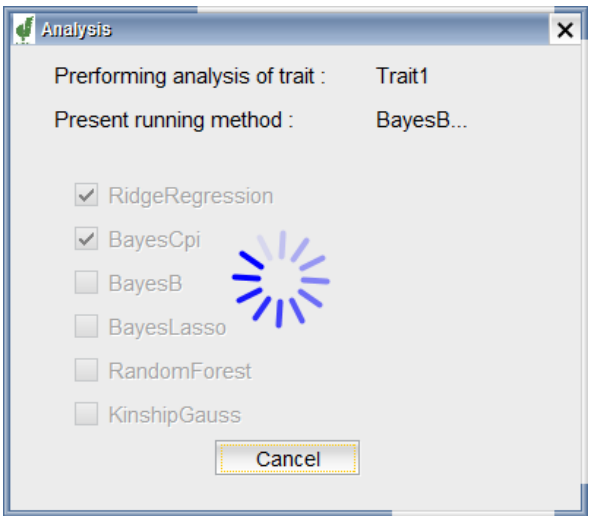
(c) Running GS models: Six different models (***Ridge Regression BLUP, Bayes Cpi, BayesB, Bayes LASSO, Random Forest and Kinship Gauss***) are implemented for fitting Genomic Selection. For GS analysis

- 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



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

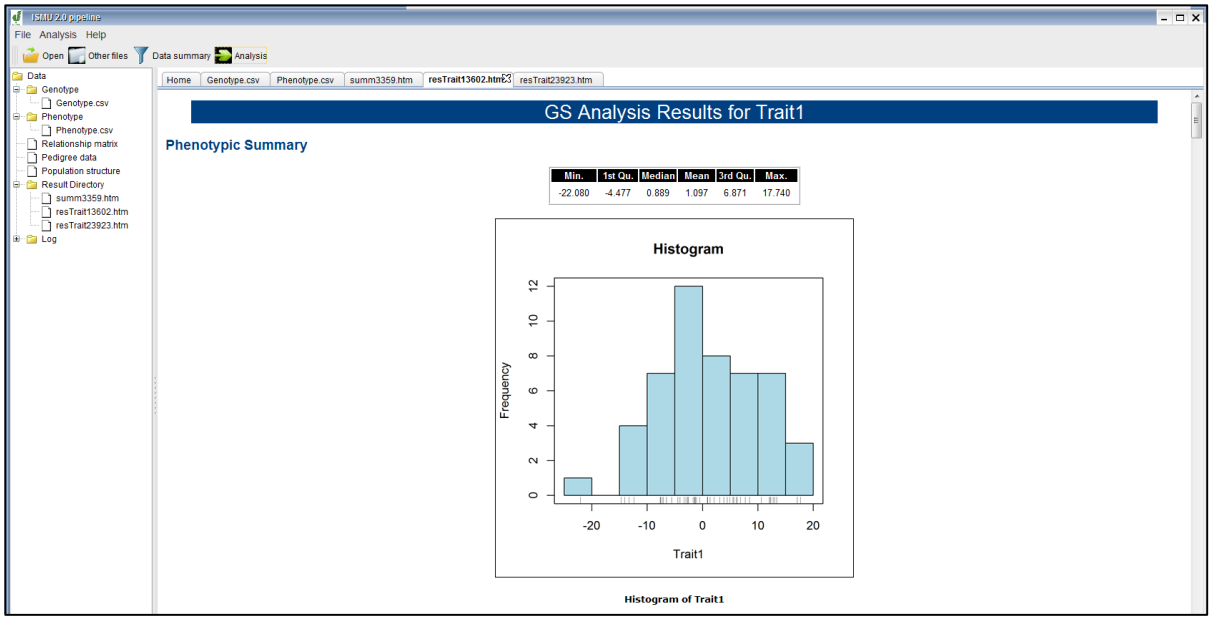


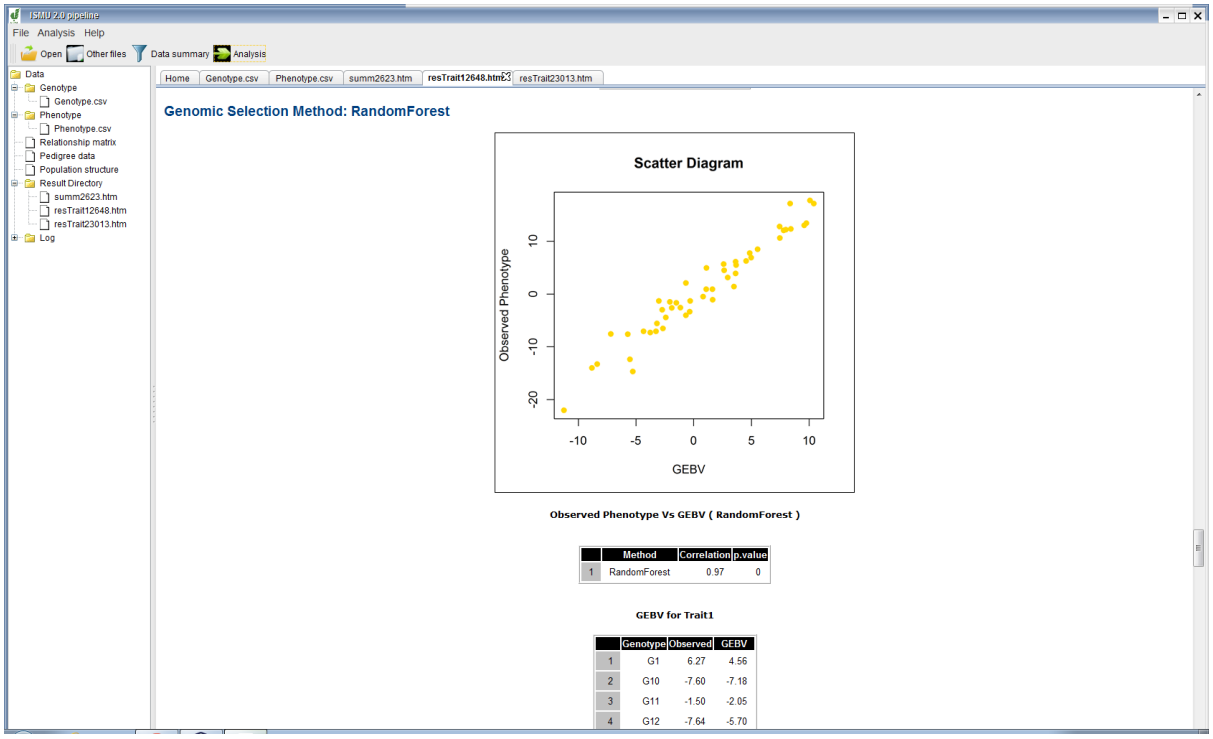


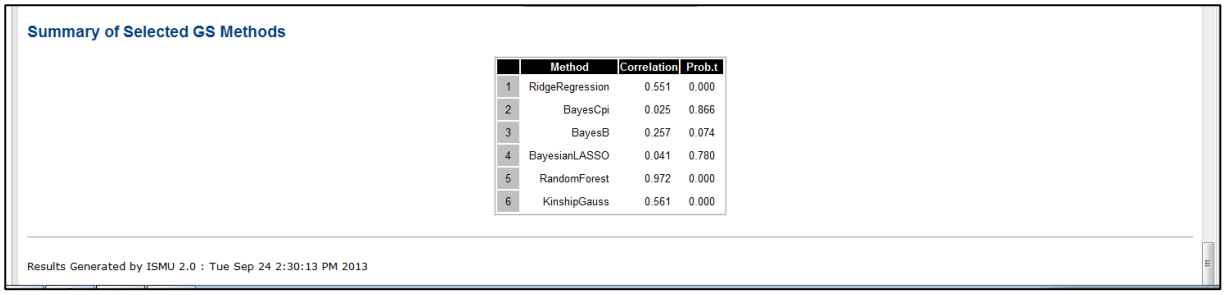
(d) Output: A html file is generated for each trait with all selected GS method. For each selected trait, result contains

- 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





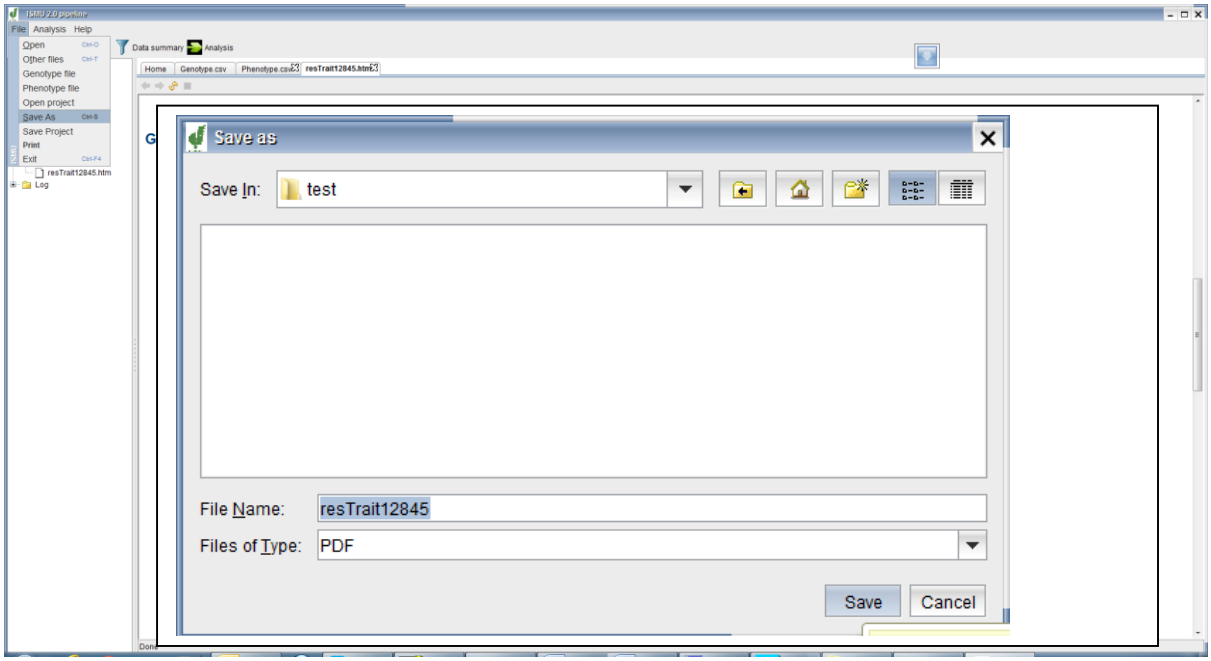


Summary of Selected GS Methods

	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---