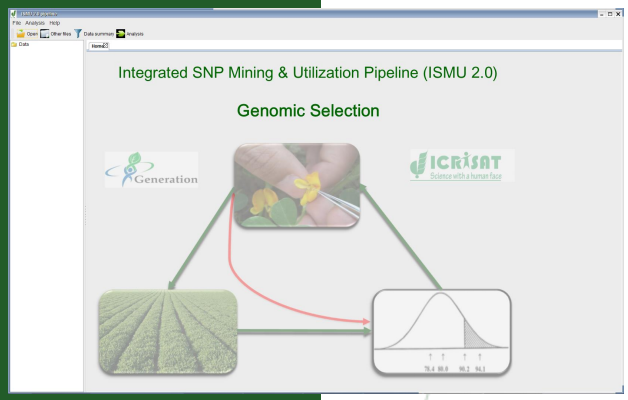
ISMU 2.0: A Multi-Algorithm Pipeline for Genomic Selection

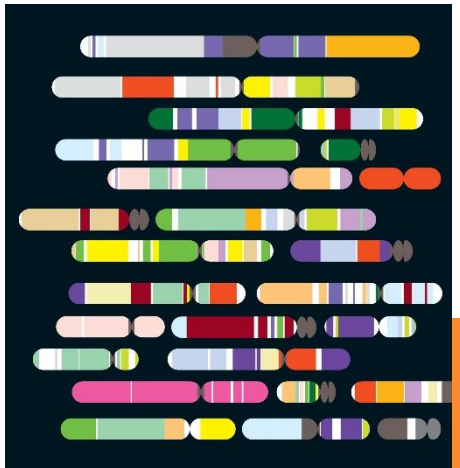
5th International Conference on Next Generation Genomics and Integrated Breeding for Crop Improvement

Wednesday, February 18, 2015

Abhishek Rathore[1], Roma R. Das[1], Manish Roorkiwal[1], Dadakhalandar Doddamani[1], Mohan Telluri[1], David Edwards[2], Mark E Sorrells[3], Janez Jenko[4], John Hickey[4], Jean-Luc Jannink[3] and Rajeev K. Varshney[1]

[1] ICRISAT, Hyderabad, India
[2] University of Queensland, Brisbane, Australia
[3] Cornell University, Ithaca, NY
[4] The University of Edinburgh, Scotland, United Kingdom

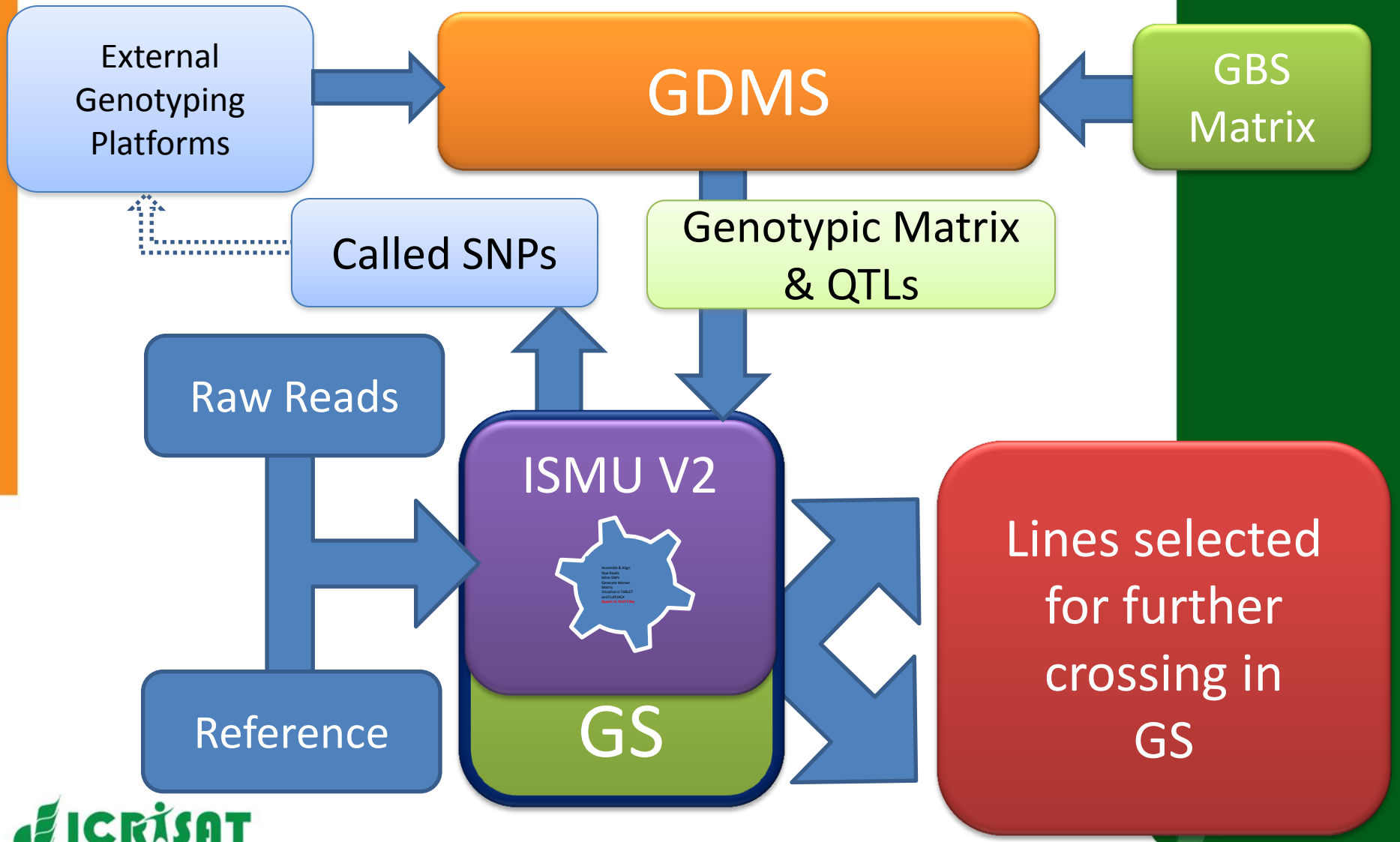




ICRISAT
Science with a human face

International Crops Research Institute for the Semi-Arid Tropics

ISMU V2.0



ICRISAT
Science with a human face

Genomic Selection (GS)

Genomic tool to accelerate breeding cycle

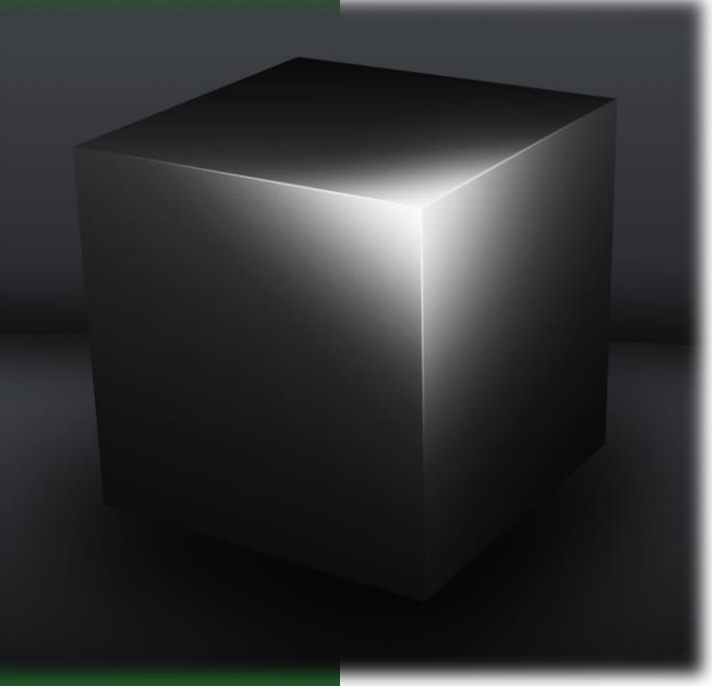
- Increases **genetic gain** per cycle through early selection
- Very useful for **complex traits** (Difficult/ expensive/takes long time to phenotype, etc.)
- Breeding values are predicted on the basis of genome wide markers, called Genomic Estimated Breeding Values (**GEBVs**)
- Several analytical approaches / GS models have been proposed for prediction of GEBVs

ICRISAT
Science with a human face

GS Approaches / Models?

- To meet the challenges, statistical methods that can handle high-dimensional data developed
- Respective properties are still not fully understood
- Causing considerable uncertainty about the choice of models for genomic prediction
- Factors affecting GS are also not very clear

ICRISAT
Science with a human face



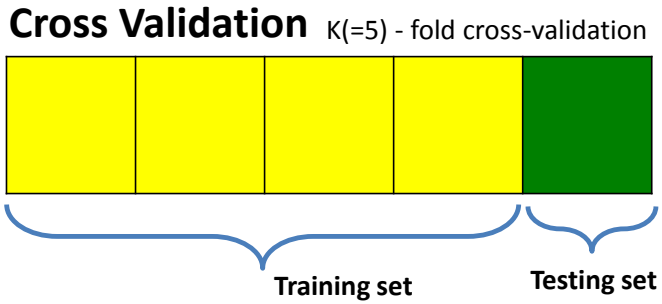
Factors Affecting GS-Models?

- Marker density, genome size and structure?
- Size of the training population?
- Historical effective population size?
- Trait heritability?
- Relationship between training population & selection candidates?
- Number of genes and distribution of their effects?
- Method used for the estimation of marker effects?
- GxE?

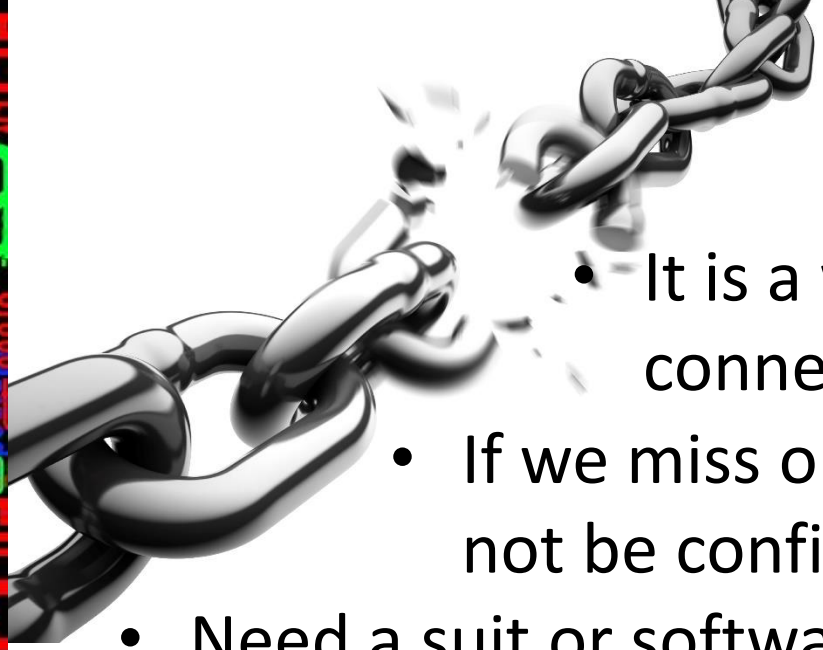
ICRISAT
Science with a human face

Many Steps in Genomic Selection…

- ✓ Get Training Population (Marker & Phenotype)
- ✓ Quality control / data filtering
- ✓ Model Population Structure / Covariates
- ✓ Fit available models
- ✓ Perform Cross Validation
- ✓ Prepare matrix of scores
- ✓ Select final method
- ✓ Get Testing Population, Predict GEBVs
- ✓ Make Selection based on GEBVs
- ✓ Add new data & rebuild model



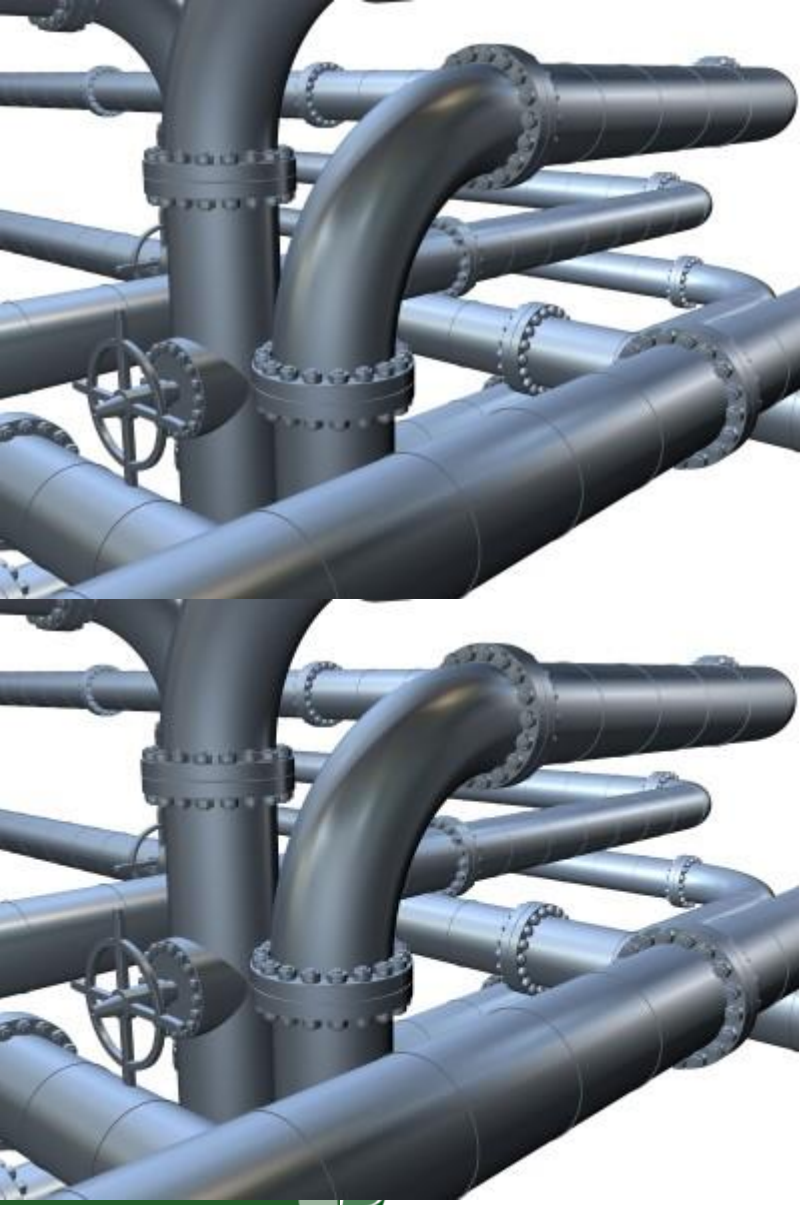
ICRISAT
Science with a human face



Difficulties in GS Application

- It is a whole chain of inter-connected tasks
- If we miss one link, predictions will not be confident
- Need a suit or software pipeline to deal with all steps with ease and confidence

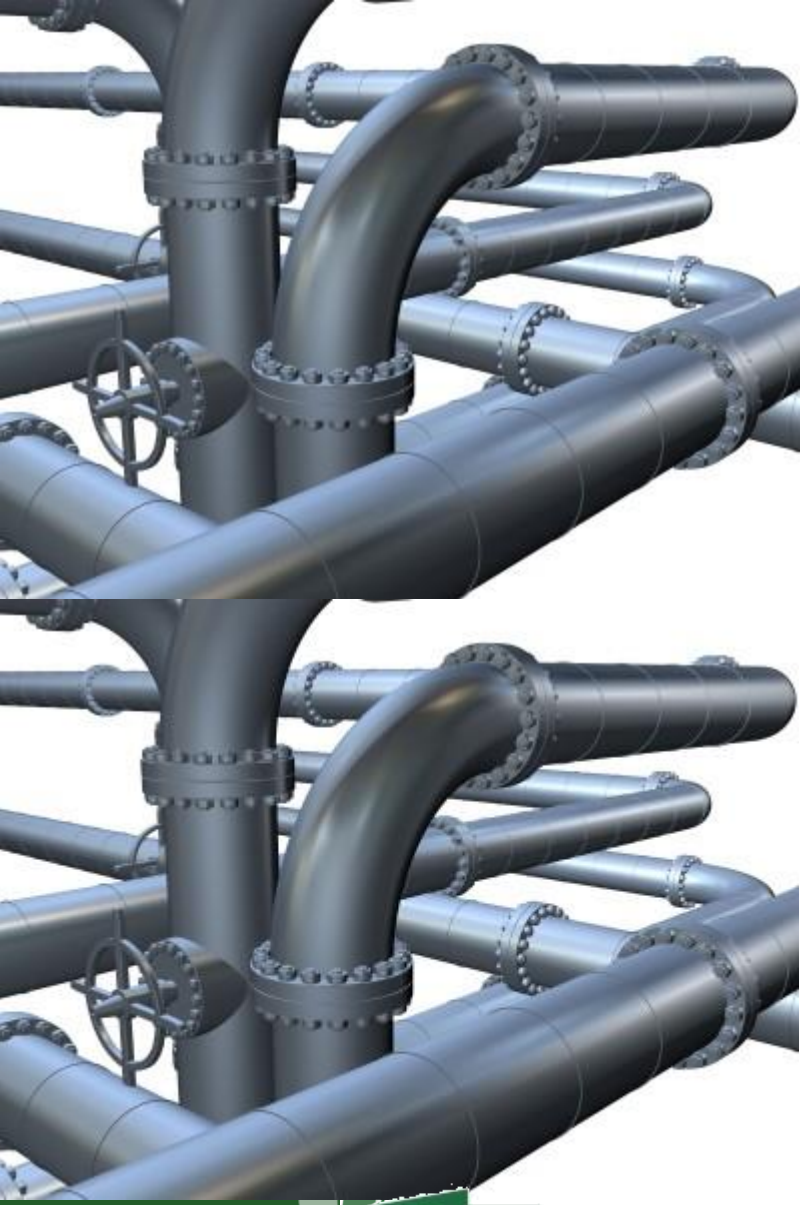
ISMU 2.0



ISMU 2.0 Pipeline

- GUI for Genomic Selection
- Multicore Support
- R and Fortran Libraries for GS
- Project Mode Development
- IDE Supports
- Multiple Method & Traits at once
- Platform Support
  - Windows x64
  - Windows x32
  - CentOS x64
  - Ubuntu x64

ICRISAT
Science with a human face

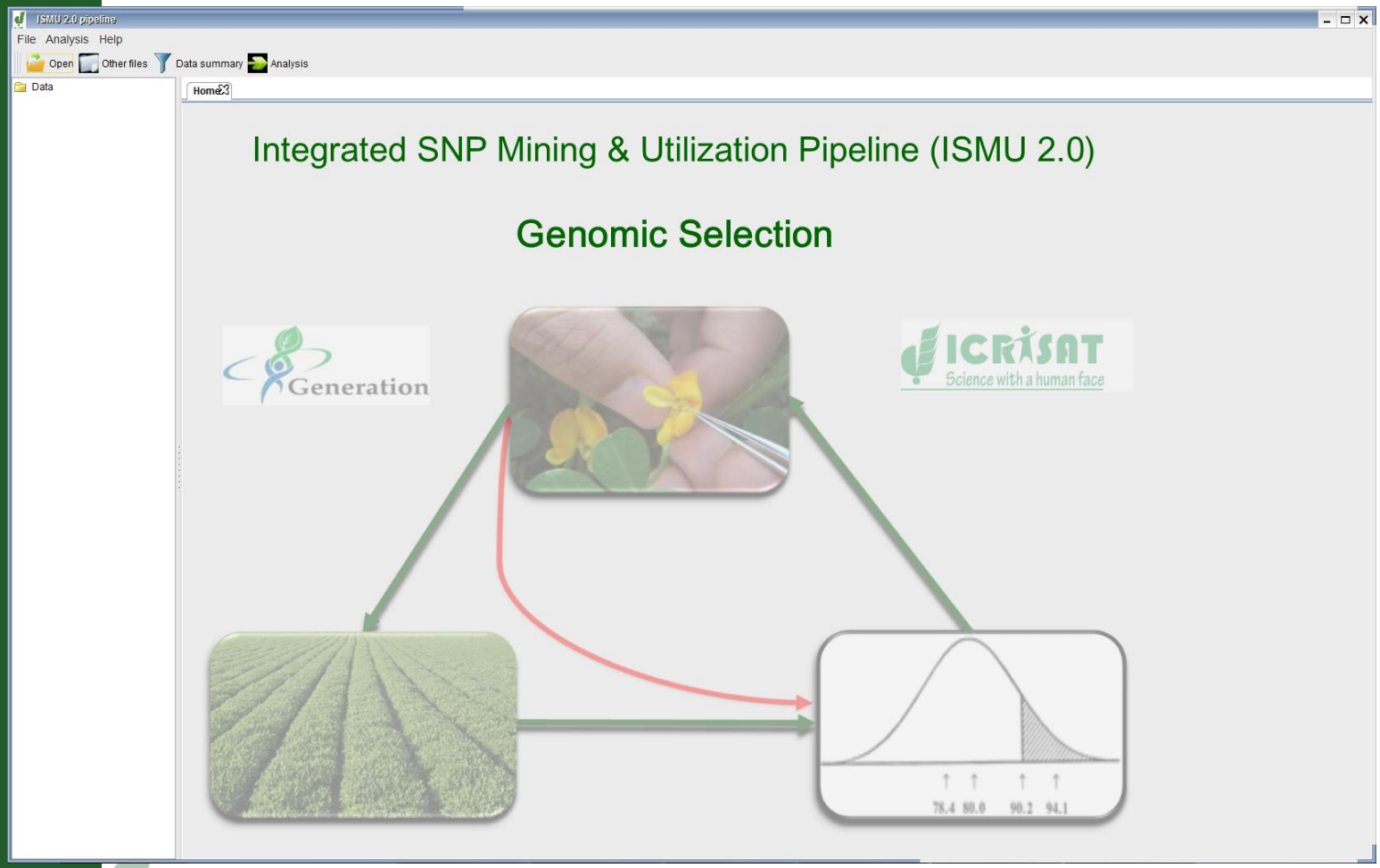


ISMU 2.0 Pipeline

- Data Diagnostics
  - Graphical Summary
  - Tabular Summary
- Subset Data
  - Missing %
  - MAF
  - PIC
- Genomic Selection
  - RR-BLUP
  - Kinship Gauss
  - Bayesian LASSO
  - BayesA, BayesB and BayesCπ
  - Random Forest Regression (RFR)
- Excel, HTML & PDF Output

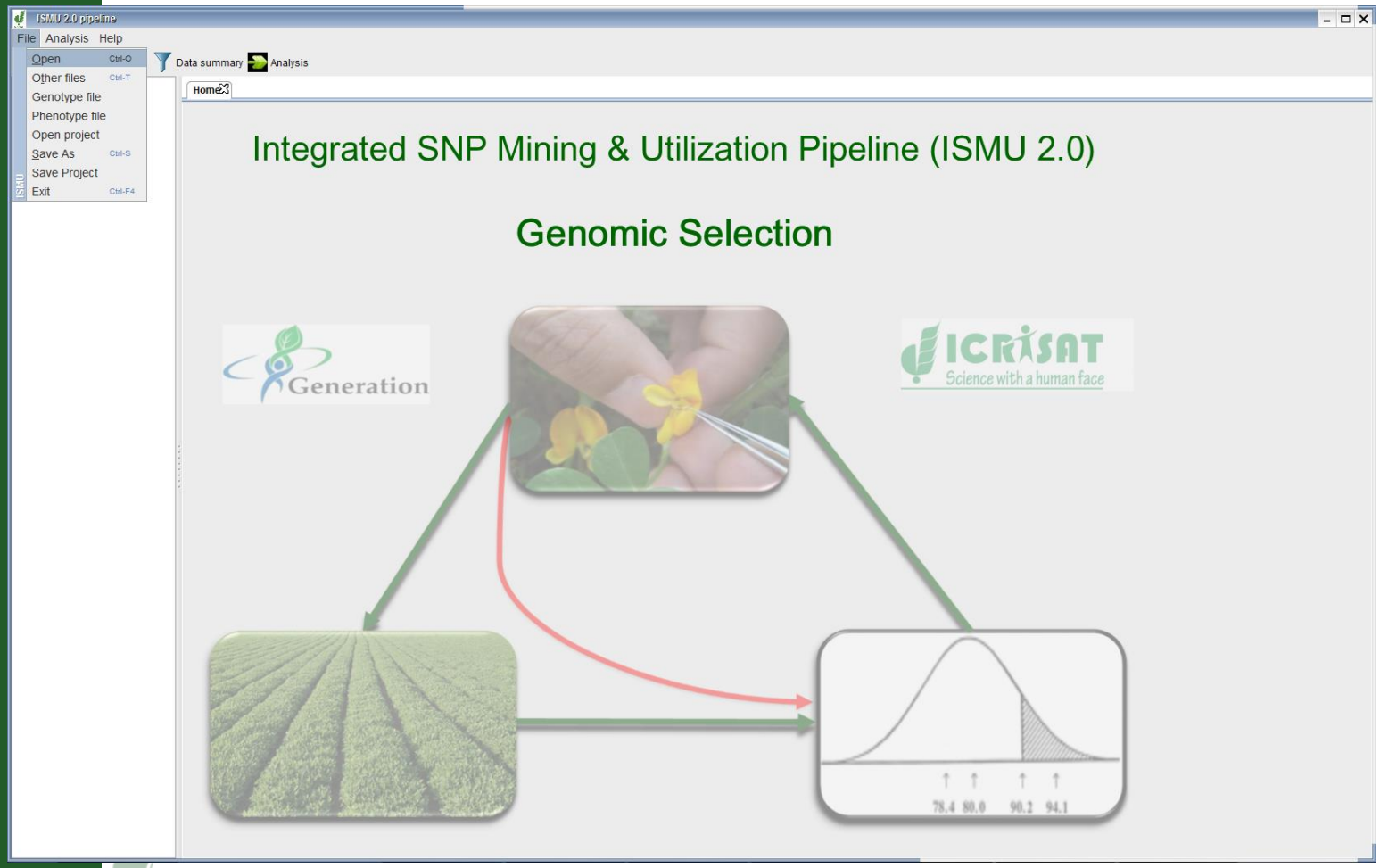
ICRISAT
Science with a human face

ISMU 2.0



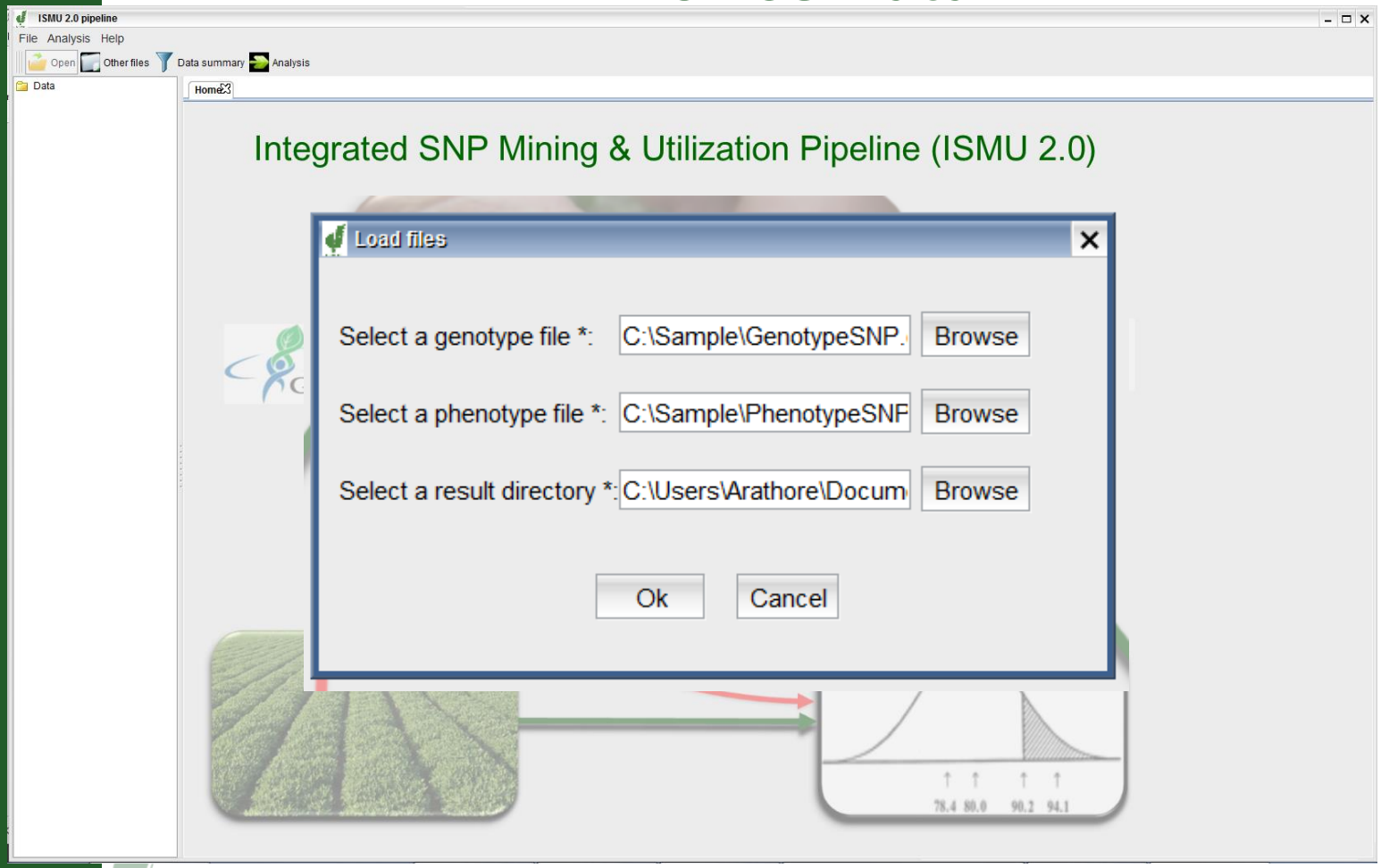
ICRISAT
Science with a human face

ISMU 2.0



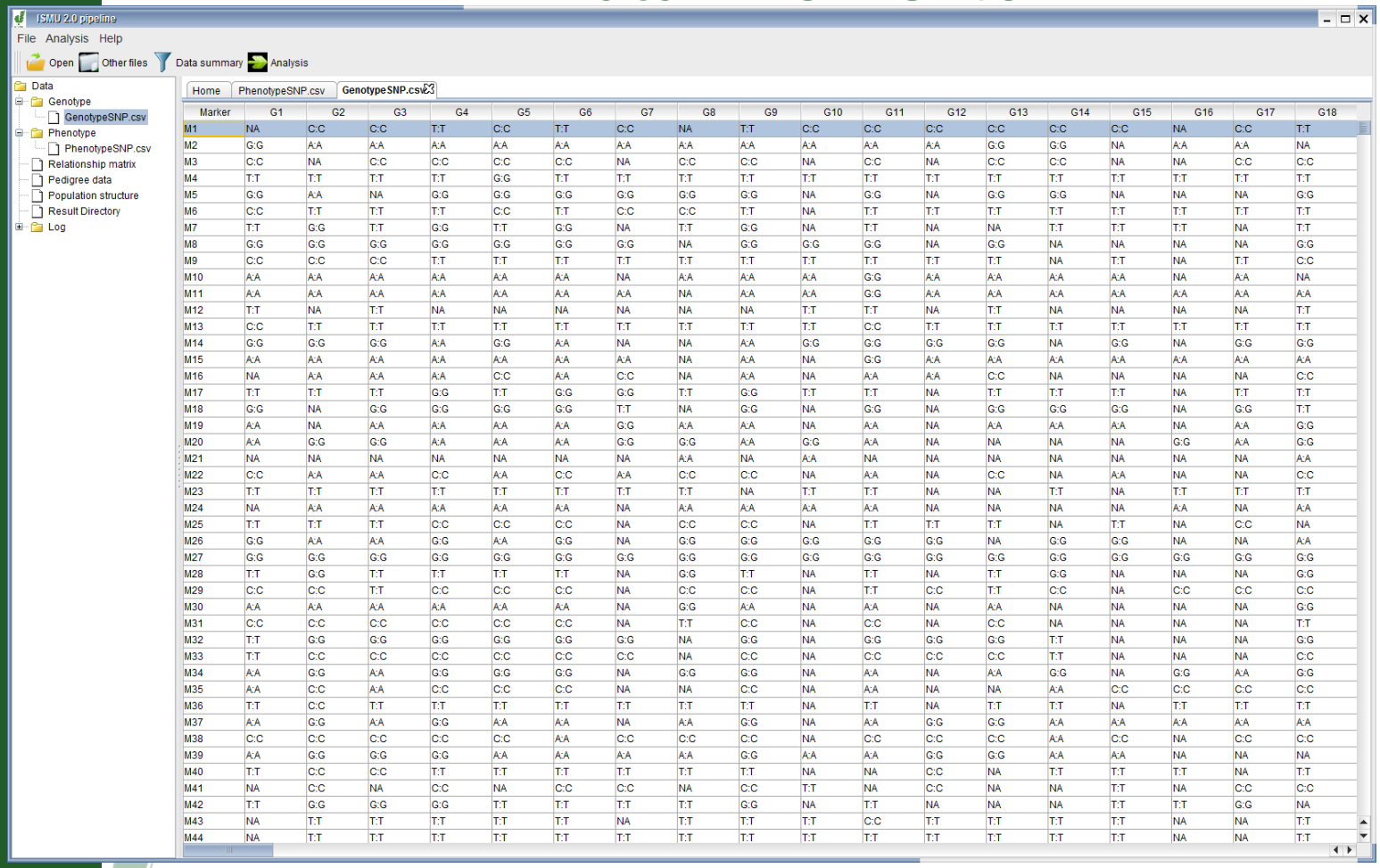
ICRISAT
Science with a human face

Browse Data



ICRISAT
Science with a human face

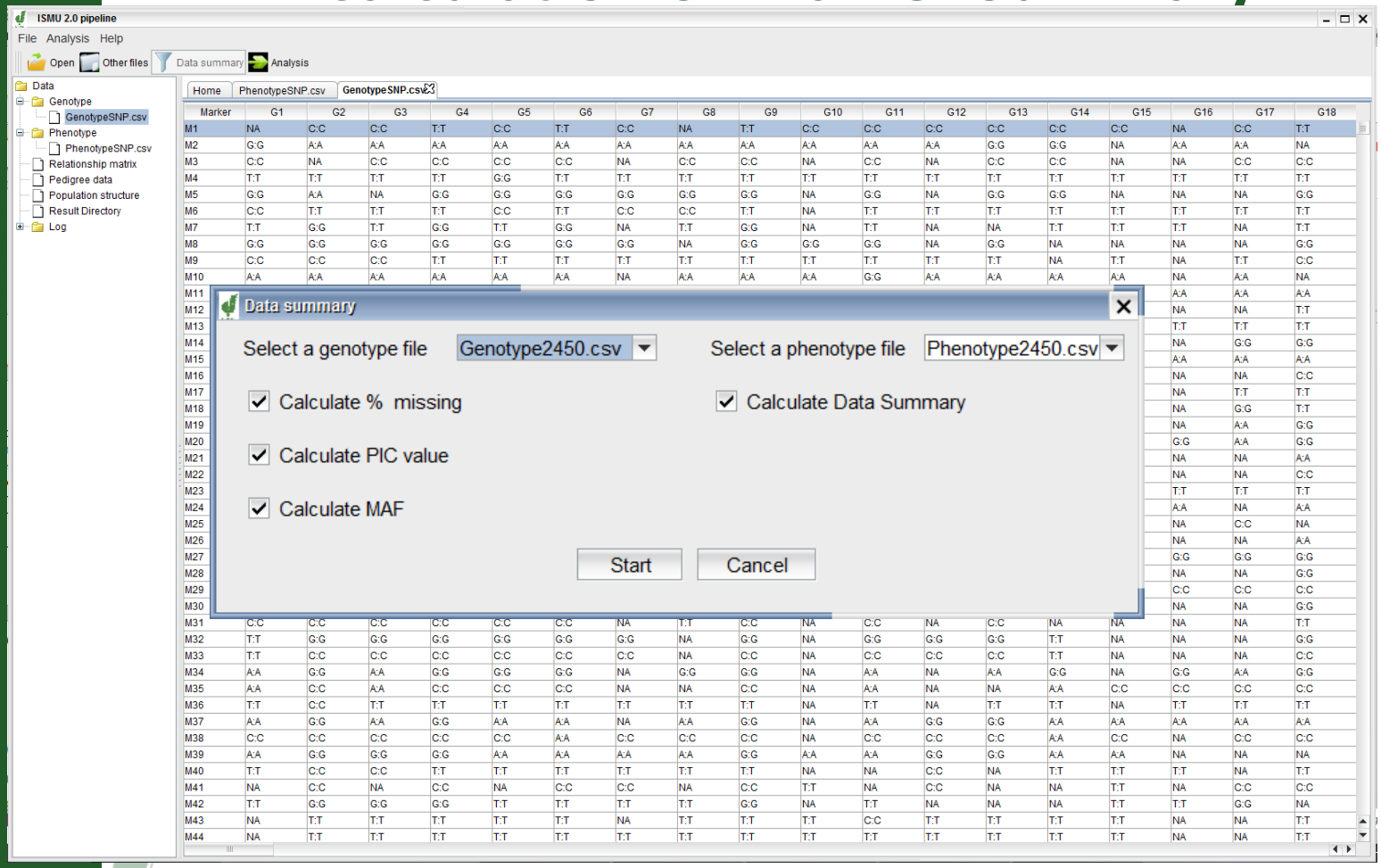
Data in ISMU2.0



Marker	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14	G15	G16	G17	G18
M1	NA	C:C	C:C	T:T	C:C	T:T	C:C	NA	T:T	C:C	C:C	C:C	C:C	C:C	C:C	NA	C:C	T:T
M2	G:G	A:A	A:A	A:A	A:A	A:A	A:A	A:A	A:A	A:A	A:A	A:A	G:G	G:G	NA	A:A	A:A	NA
M3	C:C	NA	C:C	C:C	C:C	C:C	NA	C:C	C:C	NA	C:C	NA	C:C	C:C	NA	NA	C:C	C:C
M4	T:T	T:T	T:T	T:T	G:G	T:T	T:T	T:T	T:T	T:T	T:T	T:T	T:T	T:T	T:T	T:T	T:T	T:T
M5	G:G	A:A	NA	G:G	G:G	G:G	G:G	G:G	G:G	NA	G:G	NA	G:G	G:G	NA	NA	NA	G:G
M6	C:C	T:T	T:T	T:T	C:C	T:T	C:C	C:C	T:T	NA	T:T	T:T	T:T	T:T	T:T	T:T	T:T	T:T
M7	T:T	G:G	T:T	G:G	T:T	G:G	NA	T:T	G:G	NA	T:T	NA	NA	T:T	T:T	T:T	NA	T:T
M8	G:G	G:G	G:G	G:G	G:G	G:G	G:G	NA	G:G	G:G	G:G	NA	G:G	NA	NA	NA	NA	G:G
M9	C:C	C:C	C:C	T:T	T:T	T:T	T:T	T:T	T:T	T:T	T:T	T:T	T:T	NA	T:T	NA	T:T	C:C
M10	A:A	A:A	A:A	A:A	A:A	A:A	NA	A:A	A:A	A:A	G:G	A:A	A:A	A:A	A:A	NA	A:A	NA
M11	A:A	A:A	A:A	A:A	A:A	A:A	A:A	NA	A:A	A:A	G:G	A:A	A:A	A:A	A:A	A:A	A:A	A:A
M12	T:T	NA	T:T	NA	NA	NA	NA	NA	NA	T:T	T:T	NA	T:T	NA	NA	NA	NA	T:T
M13	C:C	T:T	T:T	T:T	T:T	T:T	T:T	T:T	T:T	T:T	C:C	T:T	T:T	T:T	T:T	T:T	T:T	T:T
M14	G:G	G:G	G:G	A:A	G:G	A:A	NA	NA	A:A	G:G	G:G	G:G	G:G	NA	G:G	NA	G:G	G:G
M15	A:A	A:A	A:A	A:A	A:A	A:A	A:A	NA	A:A	NA	G:G	A:A	A:A	A:A	A:A	A:A	A:A	A:A
M16	NA	A:A	A:A	A:A	C:C	A:A	C:C	NA	A:A	NA	A:A	A:A	C:C	NA	NA	NA	NA	C:C
M17	T:T	T:T	T:T	G:G	T:T	G:G	G:G	T:T	G:G	T:T	T:T	NA	T:T	T:T	T:T	NA	T:T	T:T
M18	G:G	NA	G:G	G:G	G:G	G:G	T:T	NA	G:G	NA	G:G	NA	G:G	G:G	G:G	NA	G:G	T:T
M19	A:A	NA	A:A	A:A	A:A	A:A	G:G	A:A	A:A	NA	A:A	NA	A:A	A:A	A:A	NA	A:A	G:G
M20	A:A	G:G	G:G	A:A	A:A	A:A	G:G	G:G	A:A	G:G	A:A	NA	NA	NA	NA	G:G	A:A	G:G
M21	NA	NA	NA	NA	NA	NA	NA	A:A	NA	A:A	NA	NA	NA	NA	NA	NA	NA	A:A
M22	C:C	A:A	A:A	C:C	A:A	C:C	A:A	C:C	C:C	NA	A:A	NA	C:C	NA	A:A	NA	NA	C:C
M23	T:T	T:T	T:T	T:T	T:T	T:T	T:T	T:T	NA	T:T	T:T	NA	NA	T:T	NA	T:T	T:T	T:T
M24	NA	A:A	A:A	A:A	A:A	A:A	NA	A:A	A:A	A:A	A:A	NA	NA	NA	NA	A:A	NA	A:A
M25	T:T	T:T	T:T	C:C	C:C	C:C	NA	C:C	C:C	NA	T:T	T:T	T:T	NA	T:T	NA	C:C	NA
M26	G:G	A:A	A:A	G:G	A:A	G:G	NA	G:G	G:G	G:G	G:G	G:G	NA	G:G	G:G	NA	NA	A:A
M27	G:G	G:G	G:G	G:G	G:G	G:G	G:G	G:G	G:G	G:G	G:G	G:G	G:G	G:G	G:G	G:G	G:G	G:G
M28	T:T	G:G	T:T	T:T	T:T	T:T	NA	G:G	T:T	NA	T:T	NA	T:T	G:G	NA	NA	NA	G:G
M29	C:C	C:C	T:T	C:C	C:C	C:C	NA	C:C	C:C	NA	T:T	C:C	T:T	C:C	NA	C:C	C:C	C:C
M30	A:A	A:A	A:A	A:A	A:A	A:A	NA	G:G	A:A	NA	A:A	NA	A:A	NA	NA	NA	NA	G:G
M31	C:C	C:C	C:C	C:C	C:C	C:C	NA	T:T	C:C	NA	C:C	NA	C:C	NA	NA	NA	NA	T:T
M32	T:T	G:G	G:G	G:G	G:G	G:G	G:G	NA	G:G	NA	G:G	G:G	G:G	T:T	NA	NA	NA	G:G
M33	T:T	C:C	C:C	C:C	C:C	C:C	C:C	NA	C:C	NA	C:C	C:C	C:C	T:T	NA	NA	NA	C:C
M34	A:A	G:G	A:A	G:G	G:G	G:G	NA	G:G	G:G	NA	A:A	NA	A:A	G:G	NA	G:G	A:A	G:G
M35	A:A	C:C	A:A	C:C	C:C	C:C	NA	NA	C:C	NA	A:A	NA	NA	A:A	C:C	C:C	C:C	C:C
M36	T:T	C:C	T:T	T:T	T:T	T:T	T:T	T:T	T:T	NA	T:T	NA	T:T	T:T	NA	T:T	T:T	T:T
M37	A:A	G:G	A:A	G:G	A:A	A:A	NA	A:A	G:G	NA	A:A	G:G	G:G	A:A	A:A	A:A	A:A	A:A
M38	C:C	C:C	C:C	C:C	C:C	A:A	C:C	C:C	C:C	NA	C:C	C:C	C:C	A:A	C:C	NA	C:C	C:C
M39	A:A	G:G	G:G	G:G	A:A	A:A	A:A	A:A	G:G	A:A	A:A	G:G	G:G	A:A	A:A	NA	NA	NA
M40	T:T	C:C	C:C	T:T	T:T	T:T	T:T	T:T	T:T	NA	NA	C:C	NA	T:T	T:T	T:T	NA	T:T
M41	NA	C:C	NA	C:C	NA	C:C	C:C	NA	C:C	T:T	NA	C:C	NA	NA	T:T	NA	C:C	C:C
M42	T:T	G:G	G:G	G:G	T:T	T:T	T:T	T:T	G:G	NA	T:T	NA	NA	NA	T:T	T:T	G:G	NA
M43	NA	T:T	T:T	T:T	T:T	T:T	NA	T:T	T:T	T:T	C:C	T:T	T:T	T:T	T:T	NA	NA	T:T
M44	NA	T:T	T:T	T:T	T:T	T:T	T:T	T:T	T:T	T:T	T:T	T:T	T:T	T:T	T:T	NA	NA	T:T

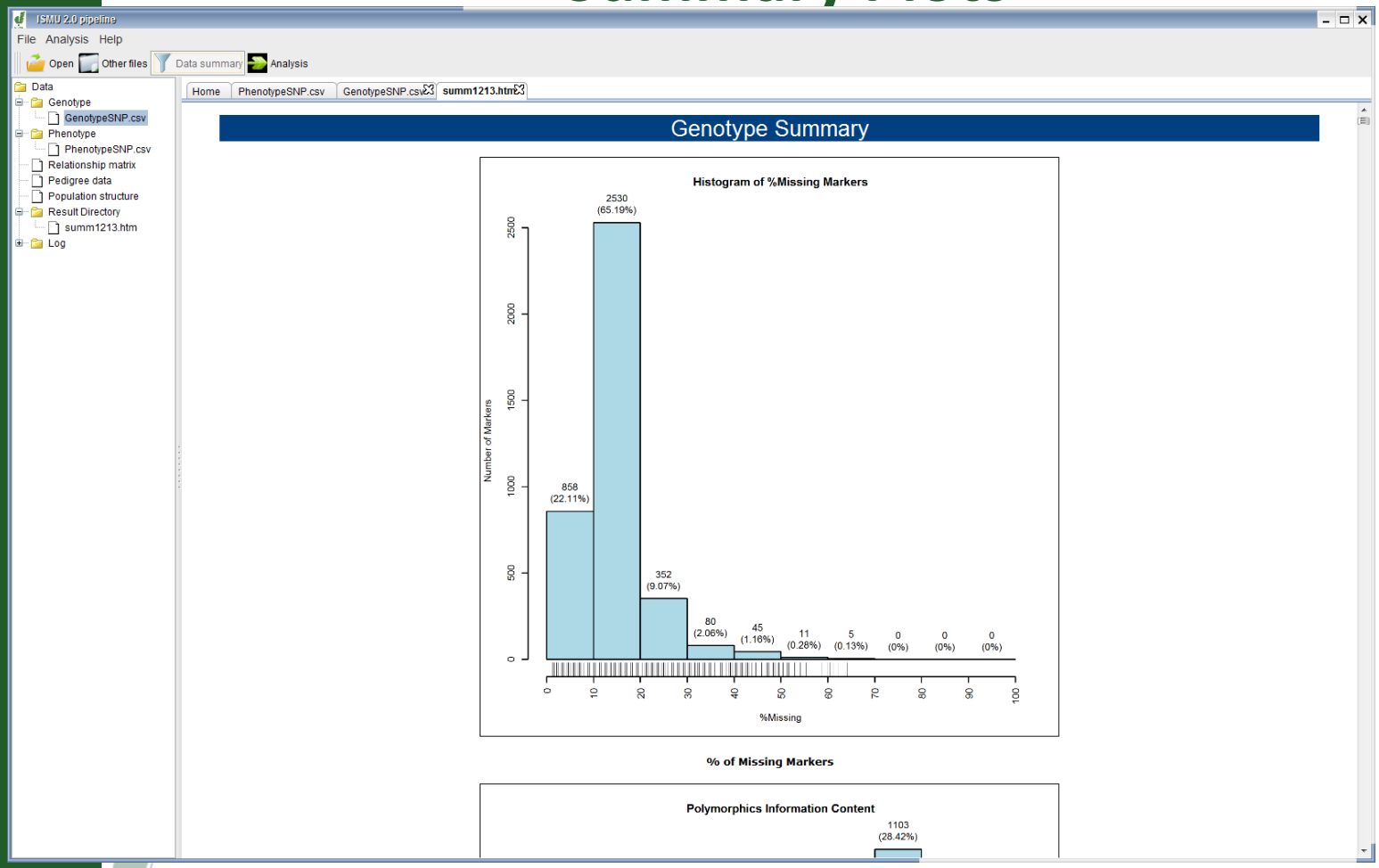
ICRISAT
Science with a human face

Calculation of Marker Summary



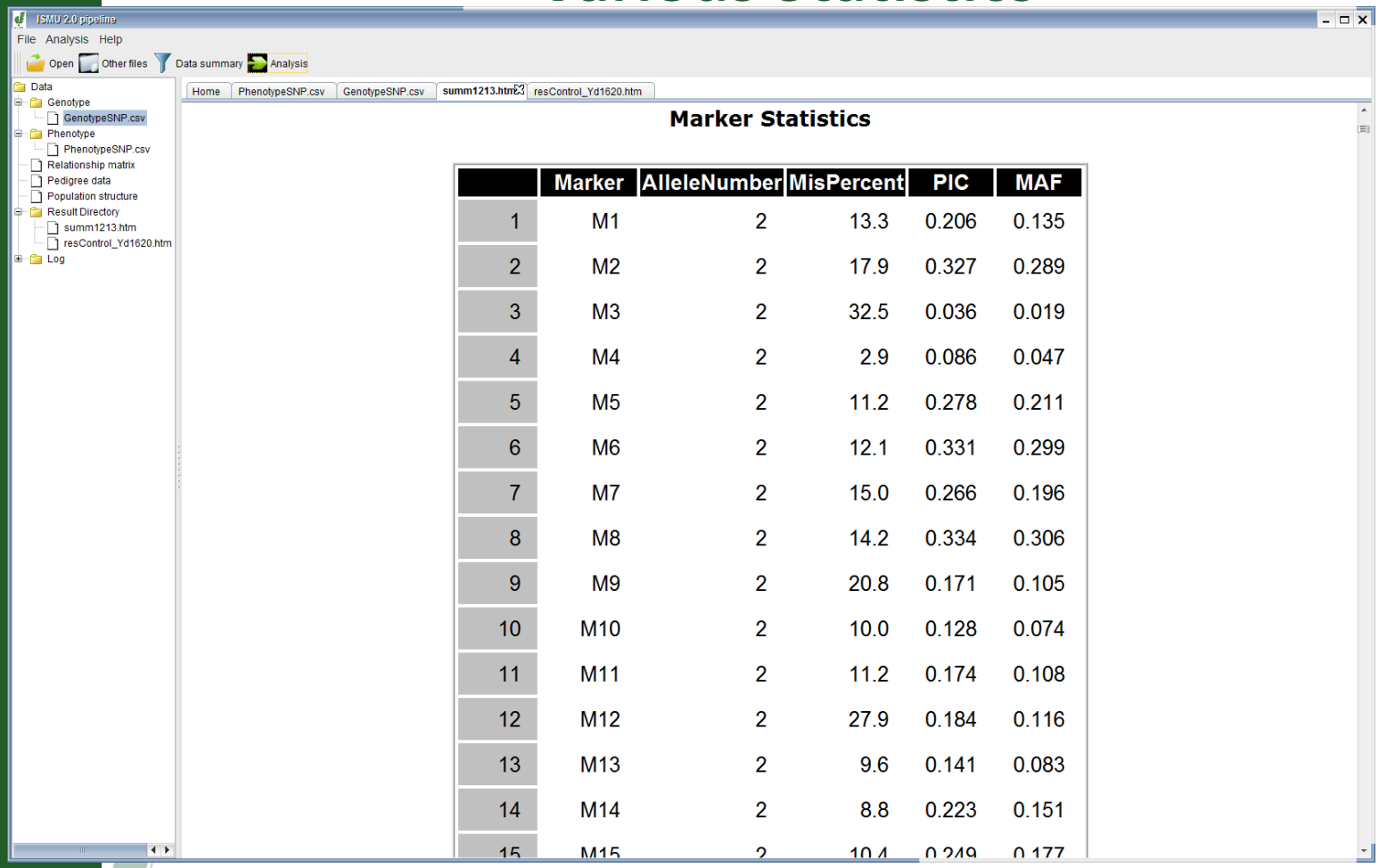
ICRISAT
Science with a human face

Summary Plots



ICRISAT
Science with a human face

Various Statistics

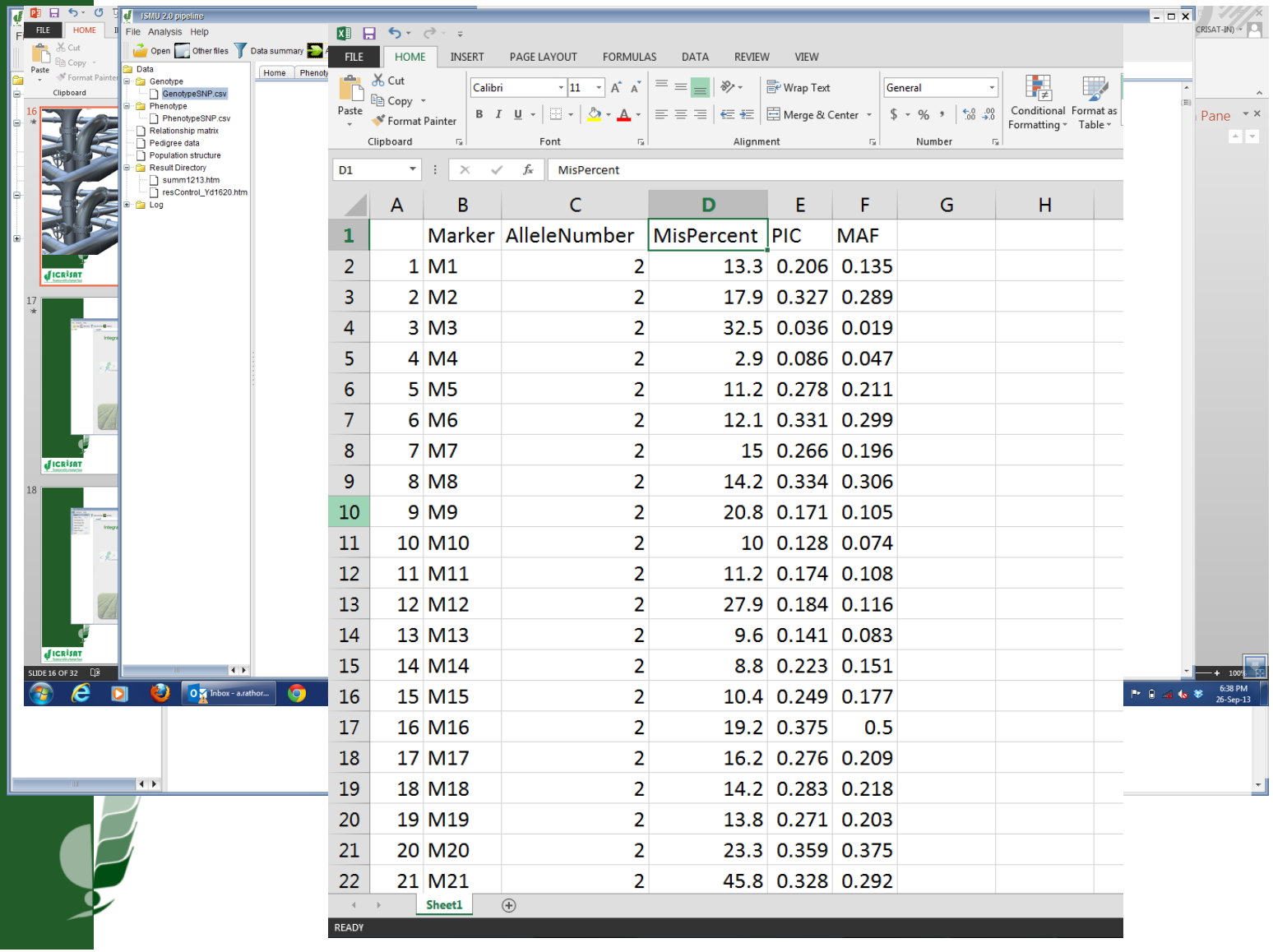


Marker Statistics

	Marker	AlleleNumber	MisPercent	PIC	MAF
1	M1	2	13.3	0.206	0.135
2	M2	2	17.9	0.327	0.289
3	M3	2	32.5	0.036	0.019
4	M4	2	2.9	0.086	0.047
5	M5	2	11.2	0.278	0.211
6	M6	2	12.1	0.331	0.299
7	M7	2	15.0	0.266	0.196
8	M8	2	14.2	0.334	0.306
9	M9	2	20.8	0.171	0.105
10	M10	2	10.0	0.128	0.074
11	M11	2	11.2	0.174	0.108
12	M12	2	27.9	0.184	0.116
13	M13	2	9.6	0.141	0.083
14	M14	2	8.8	0.223	0.151
15	M15	2	10.4	0.249	0.177

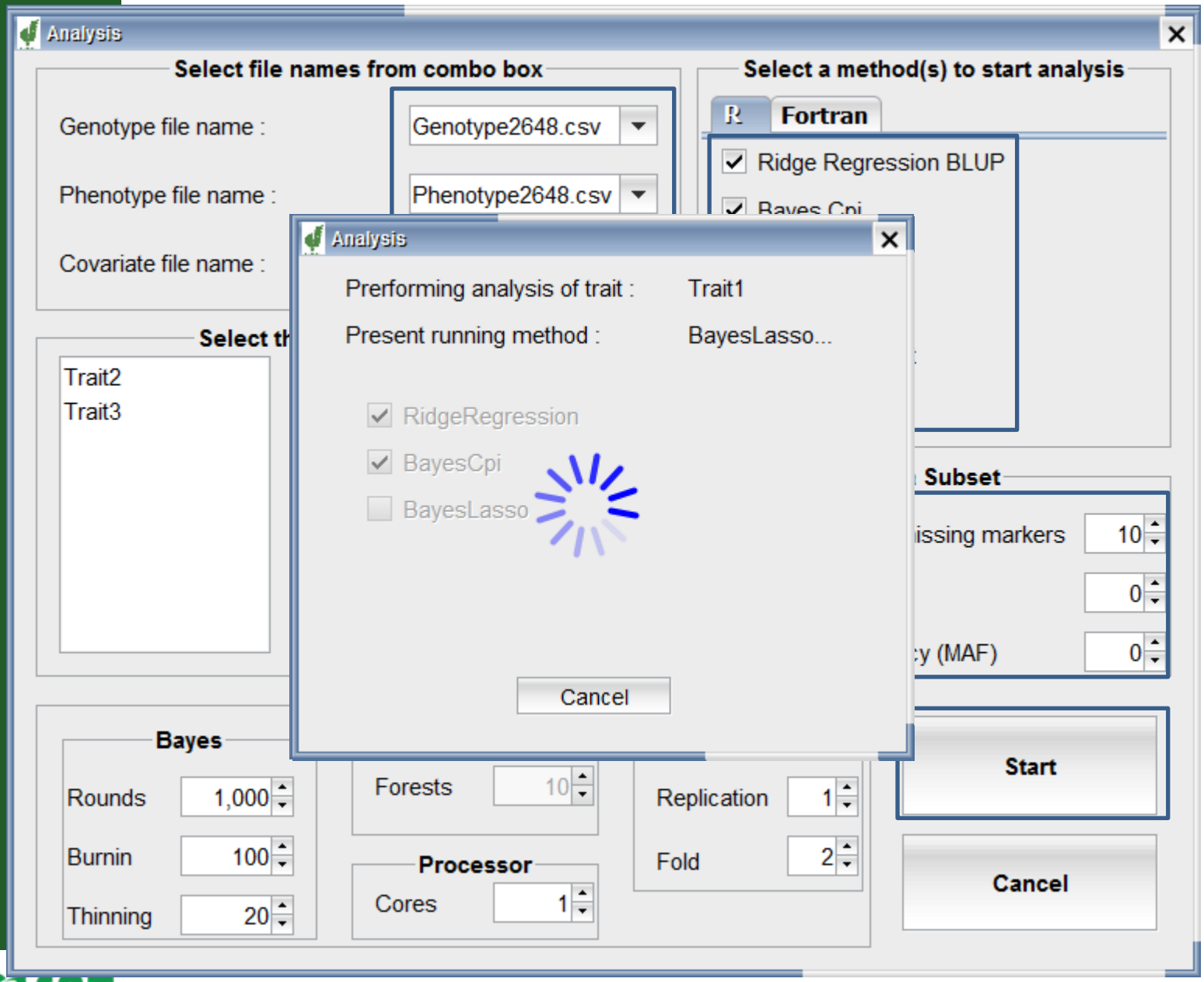
ICRISAT
Science with a human face

Export to MS-Excel (Windows)



	A	B	C	D	E	F
1		Marker	AlleleNumber	MisPercent	PIC	MAF
2	1	M1	2	13.3	0.206	0.135
3	2	M2	2	17.9	0.327	0.289
4	3	M3	2	32.5	0.036	0.019
5	4	M4	2	2.9	0.086	0.047
6	5	M5	2	11.2	0.278	0.211
7	6	M6	2	12.1	0.331	0.299
8	7	M7	2	15	0.266	0.196
9	8	M8	2	14.2	0.334	0.306
10	9	M9	2	20.8	0.171	0.105
11	10	M10	2	10	0.128	0.074
12	11	M11	2	11.2	0.174	0.108
13	12	M12	2	27.9	0.184	0.116
14	13	M13	2	9.6	0.141	0.083
15	14	M14	2	8.8	0.223	0.151
16	15	M15	2	10.4	0.249	0.177
17	16	M16	2	19.2	0.375	0.5
18	17	M17	2	16.2	0.276	0.209
19	18	M18	2	14.2	0.283	0.218
20	19	M19	2	13.8	0.271	0.203
21	20	M20	2	23.3	0.359	0.375
22	21	M21	2	45.8	0.328	0.292

ICRISAT
Science with a human face



ICRISAT
Science with a human face

GS Results

ISMU 2.0 pipeline

File Analysis Help

Open Import Data summary Analysis

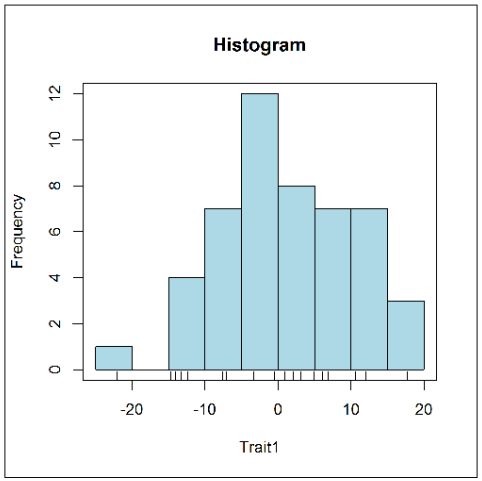
Data
- Genotype
  - Genotype2648.csv
- Phenotype
  - Phenotype2648.csv
- Covariate
- Result Directory
  - resTrait13315.htm
- Log

Home | Genotype2648.csv | Phenotype2648.csv | resTrait13315.htm

GS Analysis Results for Trait1

Phenotypic Summary

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-22.080	-4.477	0.889	1.097	6.871	17.740



Histogram of Trait1

GS Results

ISMU 2.0 pipeline

File Analysis Help

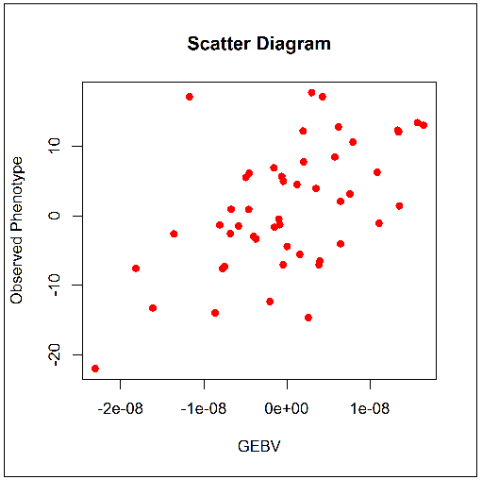
Open Import Data summary Analysis

Data
- Genotype
  - Genotype2648.csv
- Phenotype
  - Phenotype2648.csv
- Covariate
- Result Directory
  - resTrait13315.htm
- Log

Home | Genotype2648.csv | Phenotype2648.csv | resTrait13315.htm

Boxplot of Trait1

Genomic Selection Method: RidgeRegression (R)



Observed Phenotype Vs GEBV (RidgeRegression)

	Method	Correlation	p.value
1	RidgeRegression	0.55	4.2e-05

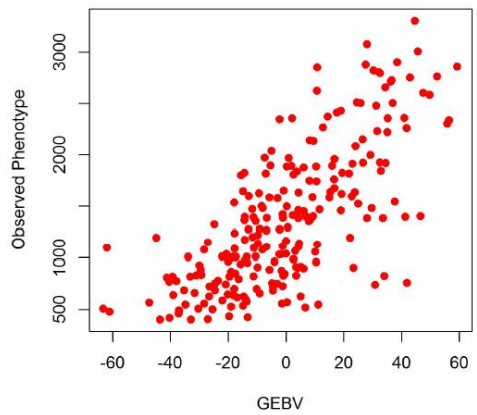
ICRISAT
Science with a human face

Export to PDF

Results.pdf - Adobe Reader

File Edit View Window Help

2 / 7 125% Tools Sign Comment



Observed Phenotype Vs GEBV (RidgeRegression)

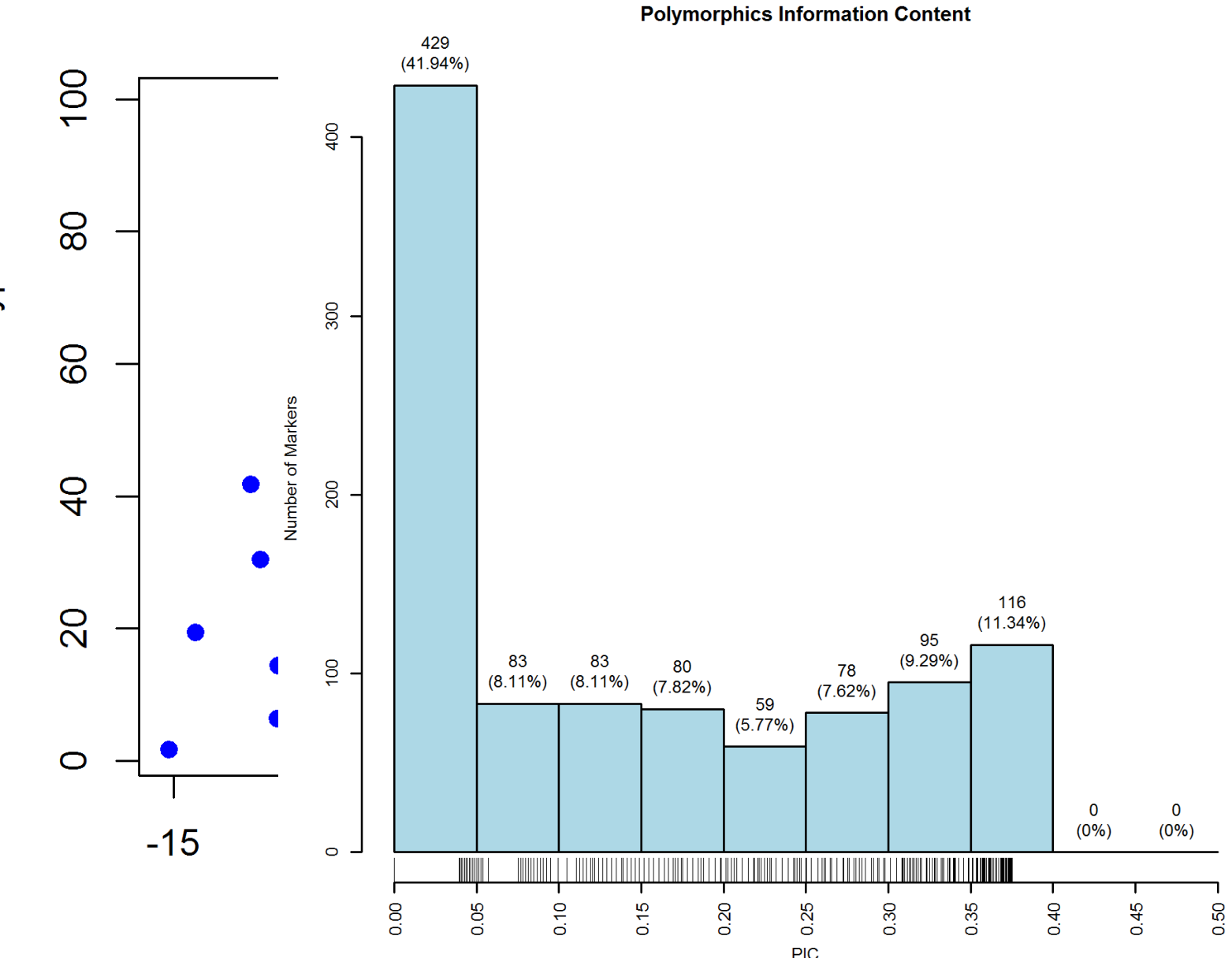
	Method	Correlation	p.value
1	RidgeRegression	0.72	0

GEBV for Control_Yd

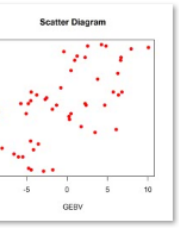
	Genotype	Observed	GEBV
1	G1	646	-20.0695430
2	G10	2822	30.2705130
3	G100	1886	10.4832539
4	G101	1884	0.3081230
5	G102	1874	6.0961027

ICRISAT
Science with a human face

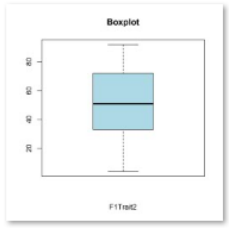
Export to High Quality Graphics 300DPI



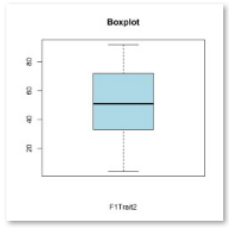
Graphics & HTML Reports saved



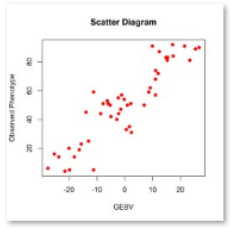
_Oct07_11063
90.303.png



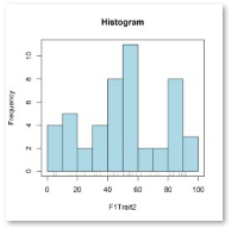
RES_Oct07_11063
91.5.png



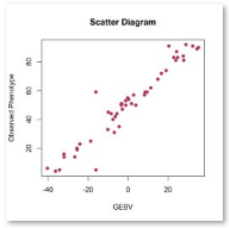
RES_Oct07_11072
30.745.png



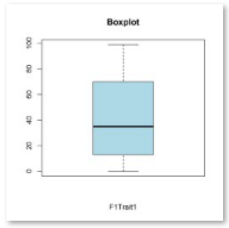
RES_Oct07_11072
31.358.png



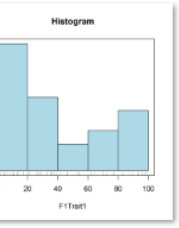
RES_Oct07_11072
31.484.png



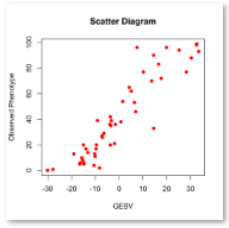
RES_Oct07_11072
41.608.png



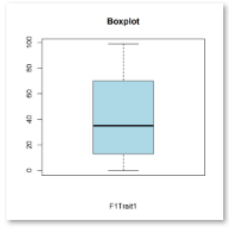
RES_Oct08_09272
90.496.png



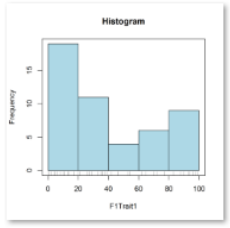
_Oct08_11252
81.011.png



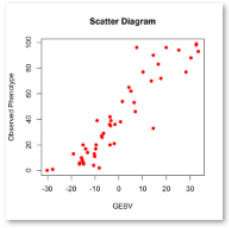
RES_Oct08_11252
91.224.png



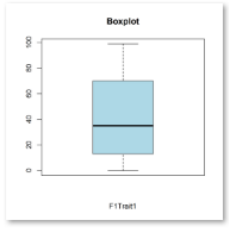
RES_Oct08_11301
00.254.png



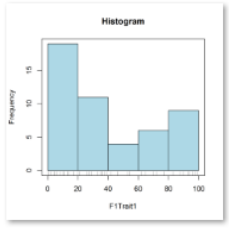
RES_Oct08_11301
00.312.png



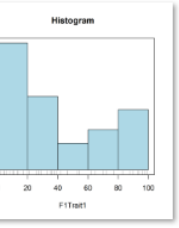
RES_Oct08_11301
10.107.png



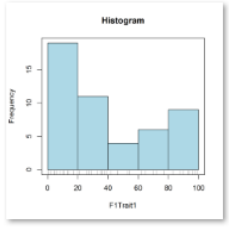
RES_Oct08_11362
40.314.png



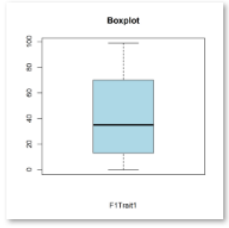
RES_Oct08_11362
40.529.png



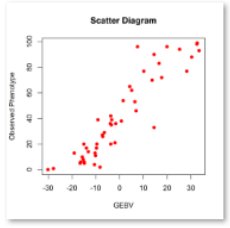
_Oct08_12301
20.488.png



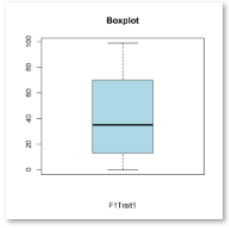
RES_Oct08_12332
30.528.png



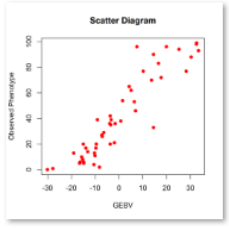
RES_Oct08_12332
31.728.png



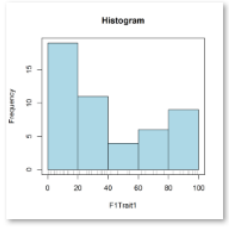
RES_Oct08_12332
40.127.png



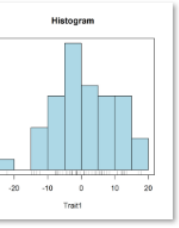
RES_Oct08_12354
50.024.png



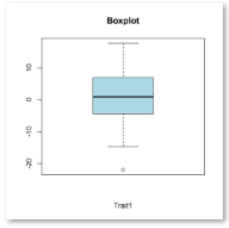
RES_Oct08_12354
50.88.png



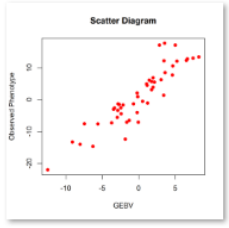
RES_Oct08_12354
50.902.png



_Oct27_16081
10.006.png



RES_Oct27_16081
11.071.png



RES_Oct27_16081
21.033.png

resF1Trait10537.h
tm

resF1Trait12304.h
tm

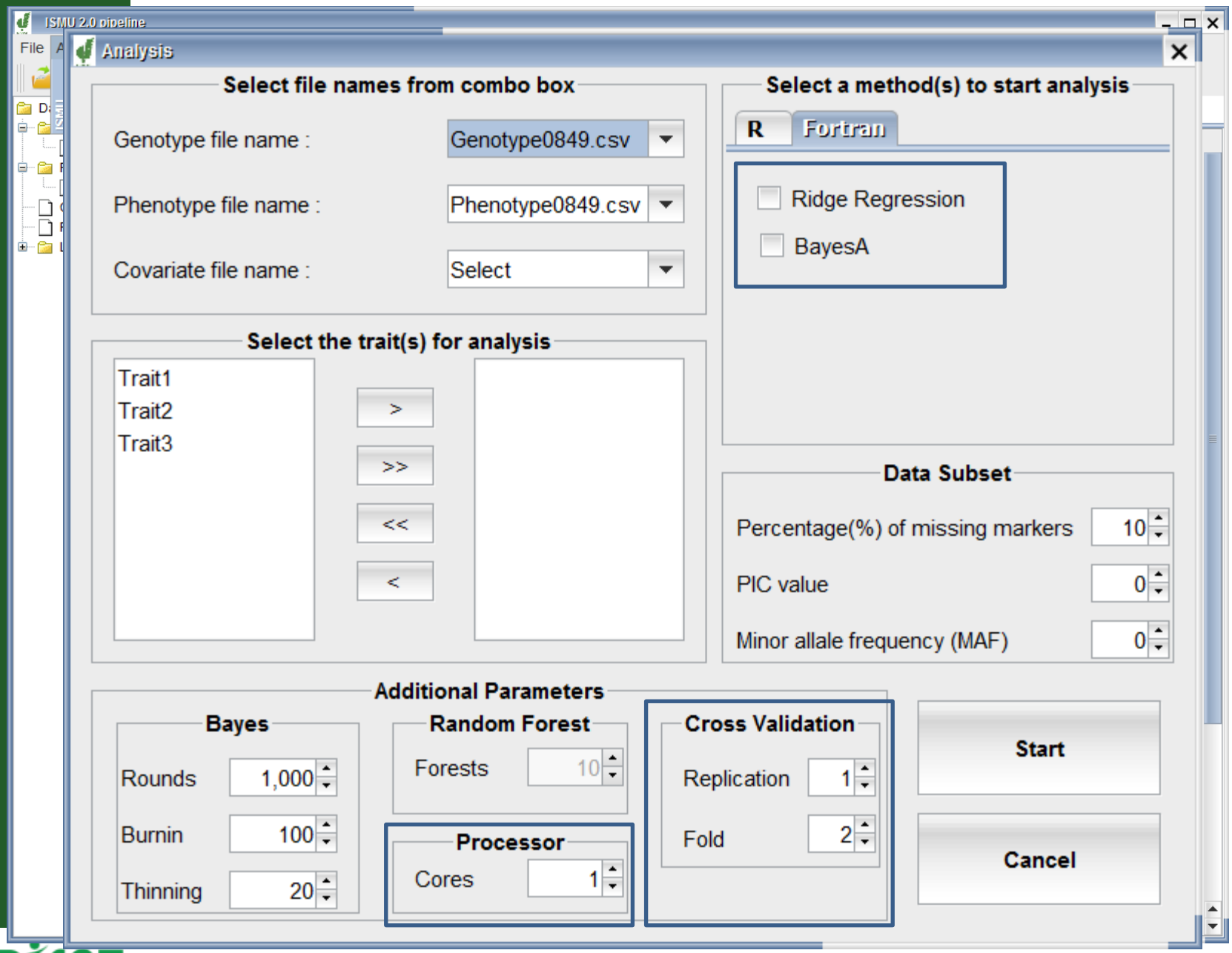
resF1Trait12527.h
tm

resF1Trait12727.h
tm

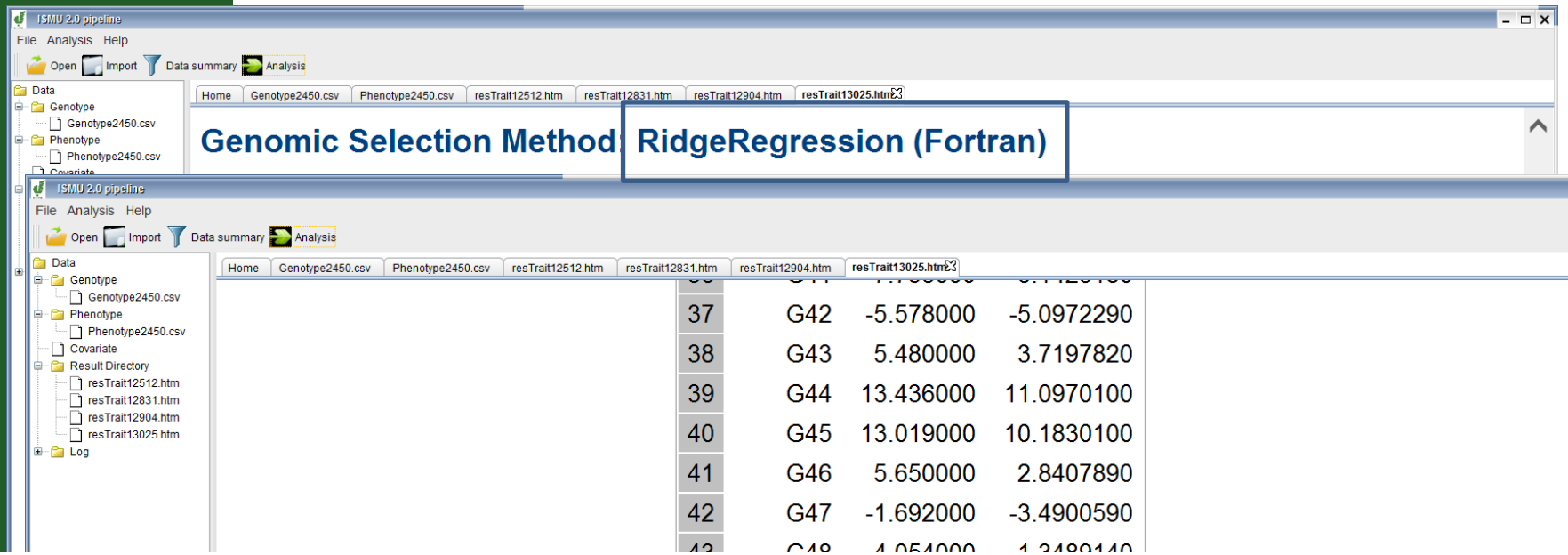
Support Large Data Sets : R & F Cocktail

- R is relatively slow when apply GS on large data sets
  - 1500 Individuals and 50 K Markers?
  - Or Even 5000 Individuals and 50 K Marker?
- A cocktail of Native FORTRAN binaries and R was used as a solution
  - 5-6 times faster
- FORTRAN was used for data processing and fitting GS Models
- R was used to compile generated results and produce high quality graphics and dynamic reports

ICRISAT
Science with a human face



ICRISAT
Science with a human face



Summary of Selected GS Methods

	Method	Correlation	Prob.t
1	RidgeRegression	0.627	0.000
2	CrossValidation(R=1,F=2)	0.513	0.0025
3	BayesA	0.57	0.000
4	CrossValidation(R=1,F=2)	0.45	0.002

Results Generated by ISMU 2.0 : Mon Jan 12 10:30:30 AM 2015 (Total Time: 00 Hr : 00 Min : 04 Sec)

Results Generated by ISMU 2.0 : Mon Jan 12 10:30:30 AM 2015 (Total Time: 00 Hr : 00 Min : 04 Sec)

Plans

- Make online version
- Support import of various popular formats
- VCF, PED, hapmap and etc
- Integration of newer methods
- Multiple trait GS
- GxE

ICRISAT
Science with a human face

Acknowledgements


Thanks…