

User Manual for



Intelligent Prediction and Association Tool

(Version 1.3)

Last updated on September 12, 2017

Zhiwu Zhang Laboratory
 *For Statistical Genomics*
ZZLab.Net

Disclaimer: While extensive testing has been performed by Zhiwu Zhang Lab at Washington State University, results are, in general, reliable, correct or appropriate. However, results are not guaranteed for any specific set of data. We strongly recommend that users validate iPat results with other original software packages, such as GAPIT, PLINK, rrBLUP and BGLR.

Support documents: Extensive support documents, including this user manual, source code, demonstration scripts, data, and results, are available at iPat website Zhiwu Zhang Laboratory: <http://zzlab.net/iPat>

Questions and comments: Users and developers are recommended to post questions and comments at iPat forum: <https://github.com/Poissonfish/iPat/issues>. Answers from other users and developers are appreciated. The iPat team members will periodically go through these questions and comments and address them accordingly.

The iPat project is partially under supports from USDA-ARS, DOE, NSF, the Agricultural Research Center at Washington State University, and Washington Grain Commission.

Citation: James Chen and Zhiwu Zhang, User manual for Intelligent Prediction and Association Tool, version 1.3, <http://zzlab.net/iPat>, accessed on MM/DD/YYYY.



TABLE OF CONTENTS

Why iPat?	5
1. Getting start.....	5
1.1 Operation environment	5
1.2 Set up R environment.....	5
1.3 Windows users	6
1.4 Mac OS users	6
2. Interface	7
2.1 Import files.....	7
2.2 Create a project	7
2.3 Covariates and kinship	8
2.4 Define Your Analysis	8
2.5 Run an analysis	10
2.6 Inspect the result.....	11
2.7 Remove files from iPat.....	11
3. File formats	12
3.1 Phenotype	12
3.2 Genotype.....	13
3.2.1 Hapmap	13
3.2.2 Numeric	13
3.2.3 VCF	13
3.2.4 PLINK (the header should be removed).....	13
3.2.5 Binary PLINK (the header should be removed)	14
3.3 Covariates	14
3.4 Kinship.....	14
4. Incorporated packages.....	15
4.1 GAPIT	15
4.2 FarmCPU.....	15
4.3 PLINK	15
4.4 rrBLUP.....	16
4.5 BGLR	16
5. Output files	17
5.1 Phenotype	17
5.1.1 Overview	17
5.2 Population structure	17
5.2.1 Heterozygosity	17
5.3 GWAS	18
5.3.1 Genomewise Manhattan plot.....	18
5.3.2 Q-Q plot.....	18
5.3.3 GWAS result	18
5.4 GS	19
5.4.1 Genomic estimated breeding value (GEBV)	19

5.4.2 Standard deviation of GEBV19

5.4.3 Histogram of GEBV20

5.4.4 Prediction result.....20

6. Tutorial.....21

6.1 Perform GWAS in iPat21

6.2 Perform GS and add user-define covariates in iPat.....22

6.3 Perform GWAS-assist GS in iPat23

7. References26

WHY IPAT?

Genome Wide Association Study (GWAS) and Genomic Prediction/Selection (GS) are two types of analyses in genomic research. Numerous software packages have been developed for the analyses with different models on different format of data. Most of these software packages were executed through Command Line Interface (CLI), including PLINK(Purcell *et al.* 2007), GAPIT(Lipka *et al.* 2012; Tang *et al.* 2016), FarmCPU(Liu *et al.* 2016), rrBLUP(Endelman 2011) and BGLR(Pérez and De Los Campos 2014). Researchers are hindered by factors such as the programming requirements, data format incompatibilities, and zero tolerance on typo of commands. Intelligent Prediction and Association Tool (iPat) is a software package with a user-friendly graphical user interface (GUI) to conduct GWAS and GS with multiple available CLI packages such as the ones listed above. Researchers can simply drag and/or click on graphical elements to specify input data files, select models, and choose define parameters. Multiple data formats are acceptable and converted automatically to the required format. Furthermore, a uniform and comprehensive presentation of results is provided to enhance interpretation of data analyses.

1. GETTING START

1.1 OPERATION ENVIRONMENT

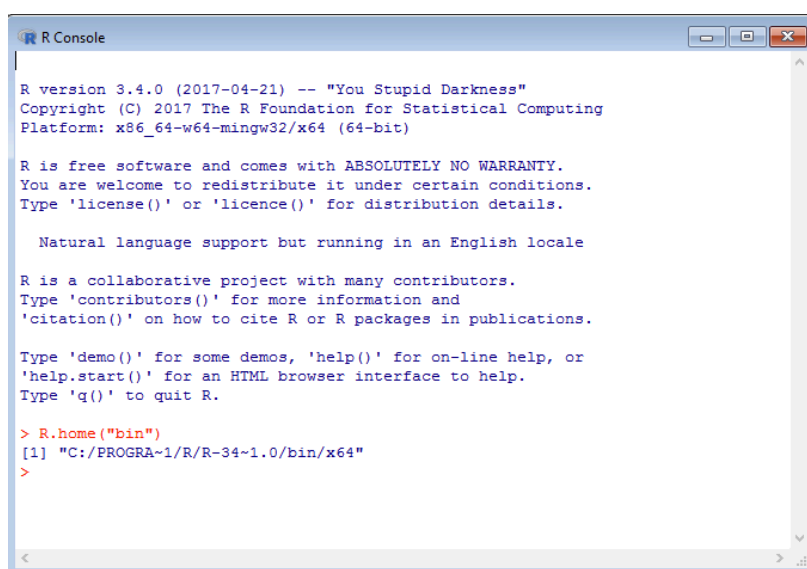
- The operation environment need to meet the following requirement:
- Operation System: Windows or Mac OS X.
- [Java Runtime Environment \(JRE\)](#): Version 8 or later.
- [R](#): Version 3.4.1 or later.

1.2 SET UP R ENVIRONMENT

Open R software and run,

```
source("http://zzlab.net/iPat/iPat_installation.r")
```

This command will install all the required r packages automatically



```
R Console

R version 3.4.0 (2017-04-21) -- "You Stupid Darkness"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

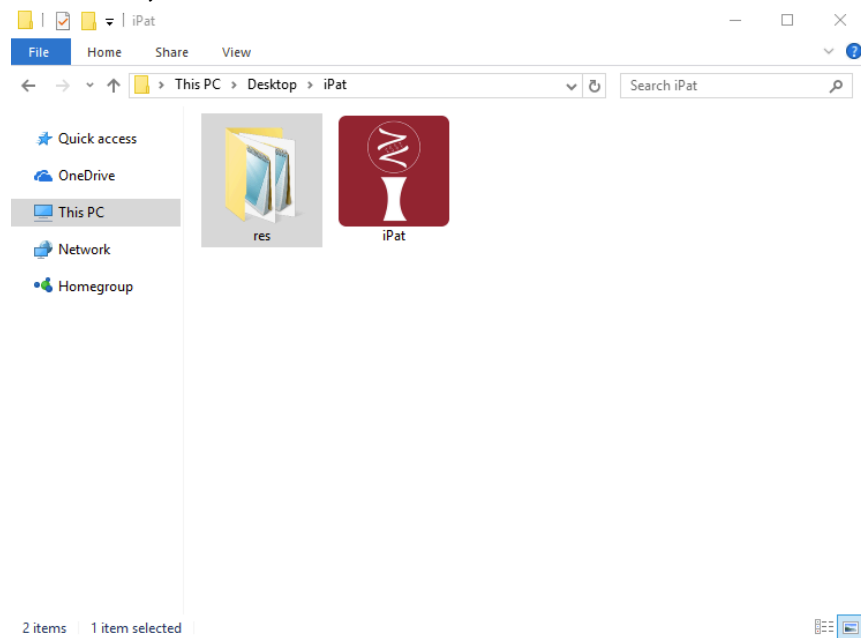
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> R.home("bin")
[1] "C:/PROGRA-1/R/R-34-1.0/bin/x64"
>
```

1.3 WINDOWS USERS

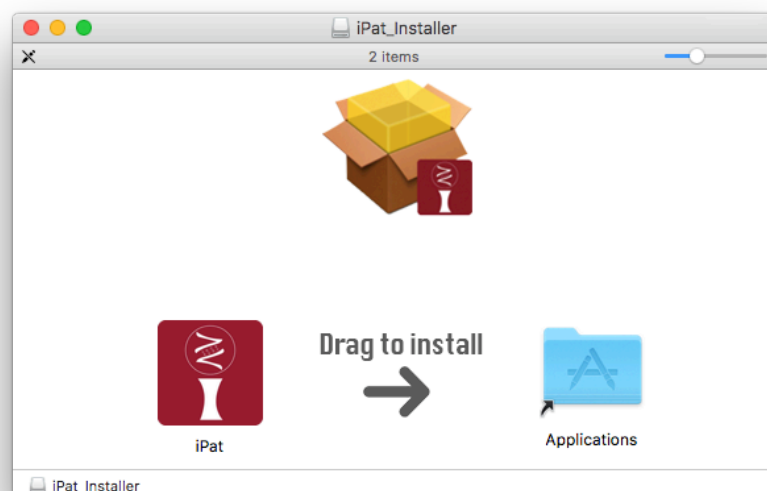
- Download [iPat.zip](#) and decompress it. You will then get a folder named "iPat", which contains a executable file "iPat.exe" and a folder "libs".
- It's noted that users are always required to place "iPat.exe" and the folder "res" in the same folder (directory) so that iPat can function normally.



- Double click 'iPat.exe' to launch iPat.

1.4 MAC OS USERS

- Download [iPat_Installer.dmg](#) and mount it on Mac.
- Follow the instruction to install iPat.

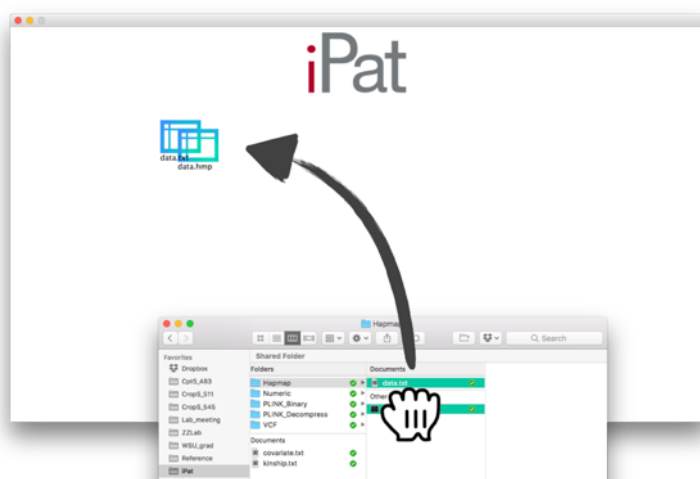


- Double click 'iPat.app' to launch iPat.

2. INTERFACE

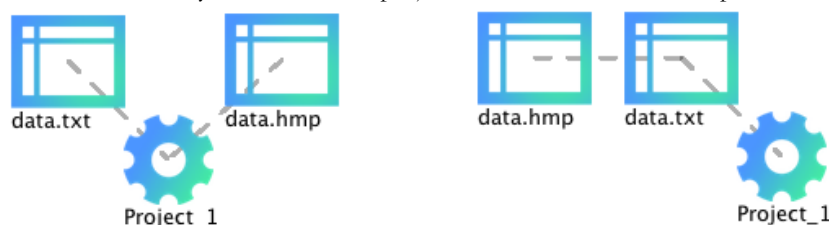
2.1 IMPORT FILES

- At beginning, iPat will show nothing but an icon "iPat" at the top of screen.
- Users can import files simply by dragging and dropping.



2.2 CREATE A PROJECT

- After importing the files, double clicking on anywhere in iPat to create a new project (a gear icon).
- Build a project by dragging a files over the project icon. A dash line will be shown between the file and project, which mean this file has been already included in this project. The below two are examples for a valid project.



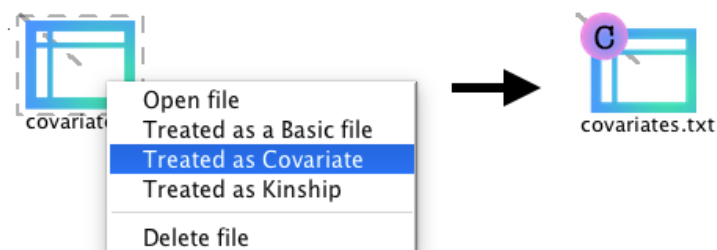
- [IMPORTANT]** A valid project must include a certain number of required files, **no less, no more**. Otherwise iPat won't work and will return an error message. Valid datasets for each format can be found from the table below (See section 3 for details):

Format	File 1 (required)	File 2 (required)	File 3 (required)	File 4 (required)
Hapmap	Genotype (.hmp)	Phenotype (.txt)	None	None
Numeric	Genotype (.dat)	Phenotype (.txt)	Map information (.map)	None
VCF	Genotype (.vcf)	Phenotype (.txt)	None	None
PLINK	Genotype (.ped)	Phenotype (.txt)	Map information (.map)	None
PLINK (binary)	Genotype (.bed)	Phenotype (.txt)	Map information (.bim)	Individual information (.fam)

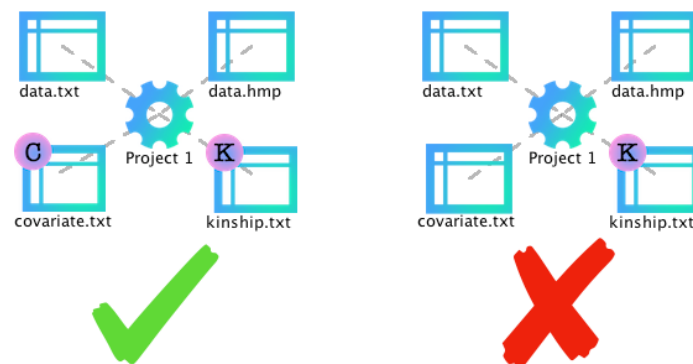
- Map information in numerical format is only required for GWAS
- In PLINK, phenotype file is only required for multiple traits analysis

2.3 COVARIATES AND KINSHIP

- Covariates provided by users will be treated as **fixed effect** in the selected model except in BGLR.
- It's **optional** that users can add **user-define** covariates or kinship into the project. Right clicking on the file to tell iPat what type of file it is. (i.e. covariates, kinship or a basic required file)



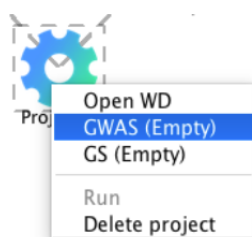
- Be aware that apart from the basic required file (i.e. phenotype and genotype), optional files must be properly labeled in a project.



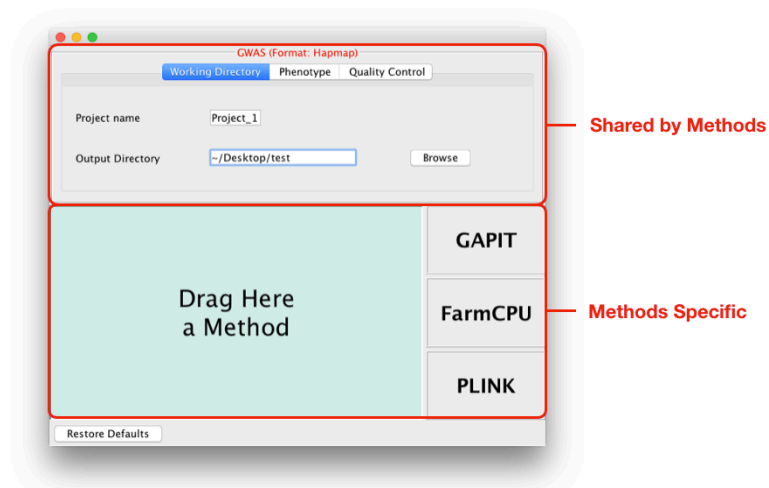
- The file labeled “C” stands for a covariate file, while files labeled “K” is identified as a kinship file by iPat. For the example of a valid project below (Left one), the file "covariate.txt" and "kinship.txt" are treated as covariates and a kinship in this project, respectively. Each project can contain **one** covariate files and **one** single kinship.

2.4 DEFINE YOUR ANALYSIS

- After linking every files needed in the project, right click on the project and choose either GWAS or GS to open a configuration panel.



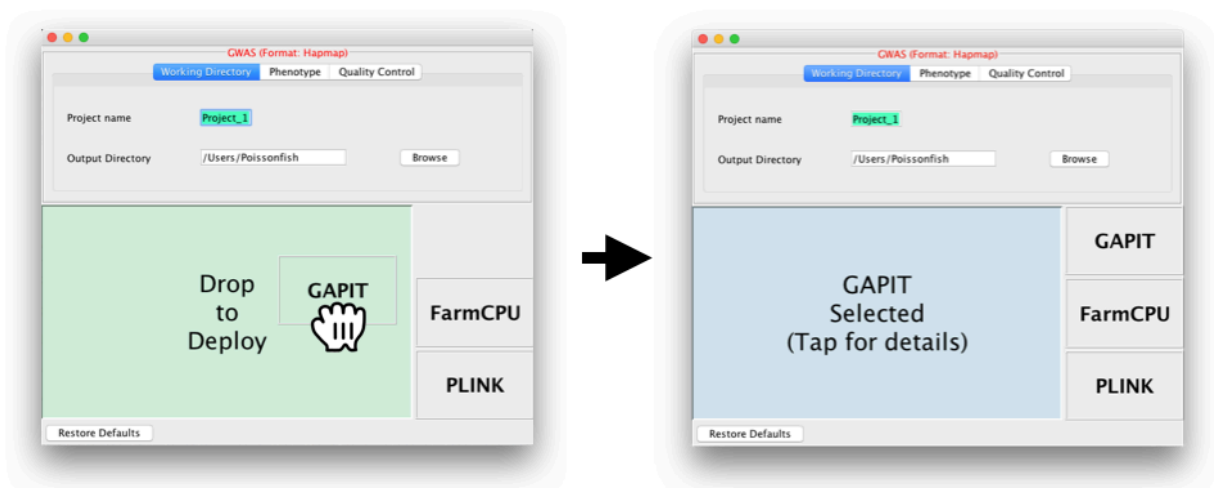
- The panel consist of two sections. The upper one presents a set of input arguments shared by all methods, while users can define method-specific arguments from the lower section.



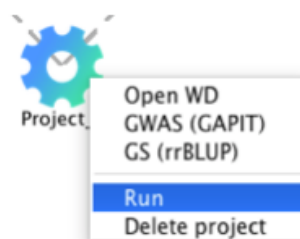
- Available parameters in the upper section:

Tab	Parameters	Definition	Default
Working Directory	Project name	Prefix for output files	Project_x (x is a number starts from 1)
Working Directory	Output Directory	A path where output files will be generated	Home directory
Phenotype	Trait names	Subsetting traits data	All traits are selected
Quality Control	By missing rate	Filtering out markers where certain rate of value is missing	No threshold
Quality Control	By MAF	Filtering out markers based on minor allele frequency	0.05

- To select a method, simply drag a "method block" to the left-side area. And tap on this area for further defining (see section 3 for details)

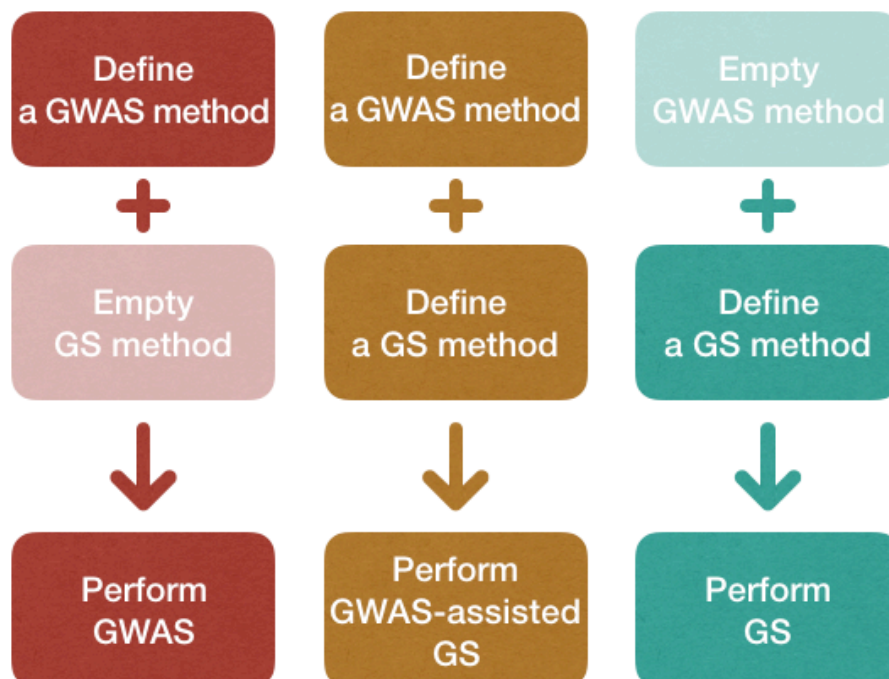


- After defining the analysis, user can start to run the procedure by clicking 'Run' at the pop-up menu of the project.

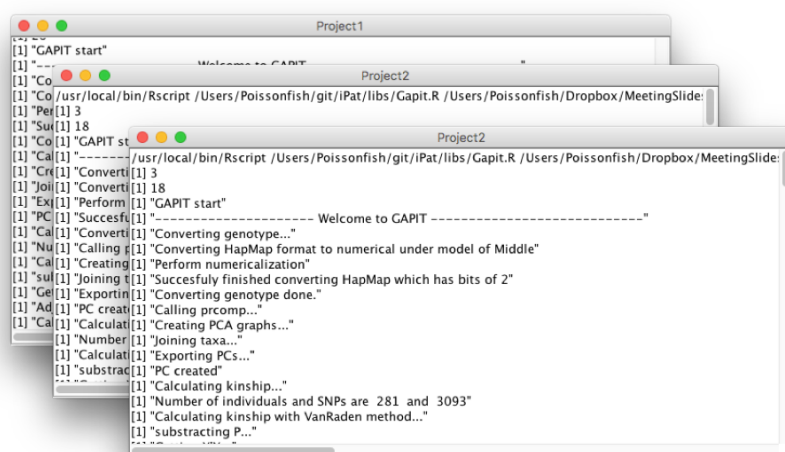


2.5 RUN AN ANALYSIS

- In iPat, users are allowed to do genomic studies such as GWAS, GS and GWAS-Assisted GS (Associated SNPs reported by GWAS will be treated as fixed effect in GS). iPat will detect the project configuration and decide which analysis should be implemented afterward.



- Each project will generate a console window while running the analysis. User can track the progress of the task from window messages.
- iPat also capable of multitasking. Users can arrange another project even when the previous one have not done yet.



2.6 INSPECT THE RESULT

- When iPat complete a project, the gear icon will show a green dot if the task run successfully without any error occurred. Otherwise it will show a red dot at its top-left to notify users that there're existing at least one error message during the analysis.



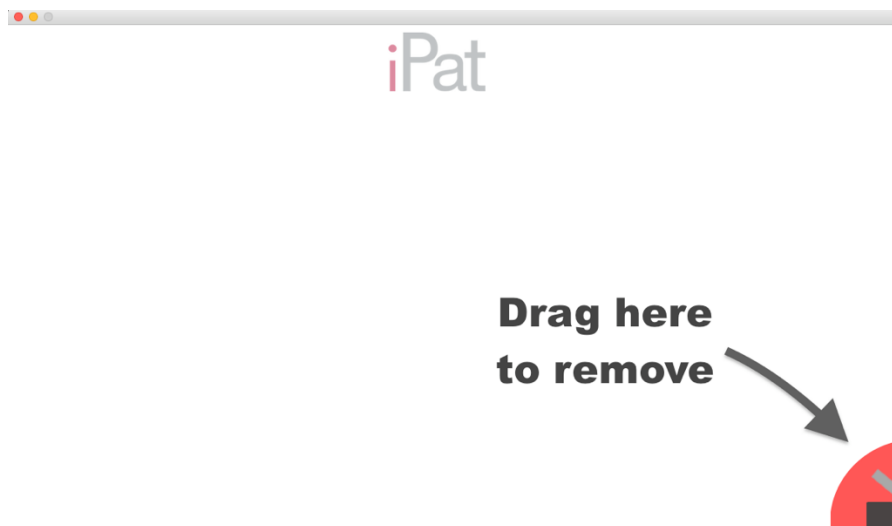
- Users can inspect the results by double clicking on the gear icon, which will direct users to the folder where output files generated (See section 4 for details of output files).

[illegible]

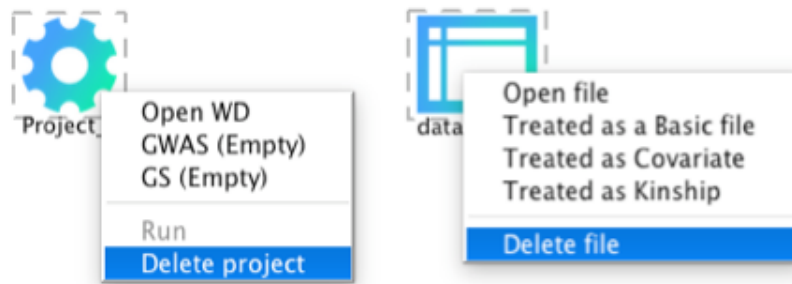
2.7 REMOVE FILES FROM IPAT

Users are allowed to remove projects, files and linkage from iPat in 3 ways:

- Drag any object to the bottom-left area and release to remove it.



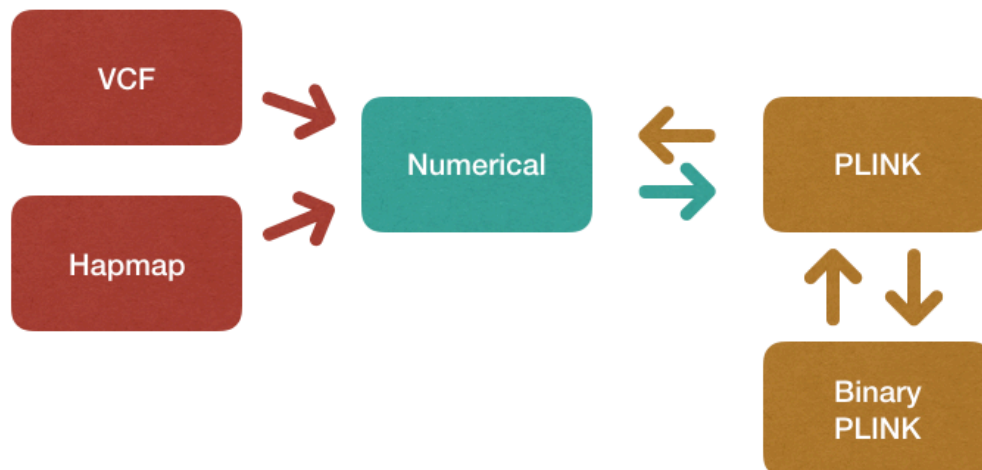
- Right click on an object to open a pop-up menu then delete it



- Or, simply press Backspace (press delete on Mac) after selecting an object.

3. FILE FORMATS

- iPat mainly work with files in numerical format, but it can also work fine with Hapmap, VCF and PLINK format. iPat will recognize the format of input files and do a format conversion automatically if needed.



3.1 PHENOTYPE

- Phenotype data for every formats except PLINK must contain **sample names** in the first column and **traits names** as the header:

taxa	trait 1	trait 2
sample1		
sample2		
sample3		

- Phenotype data for PLINK must contain **sample and family names** in the first 2 columns and **traits names** as the header:

FID	SID	trait 1	trait 2
family 1	sample1		
family 2	sample2		
family 3	sample3		

3.2 GENOTYPE

3.2.1 HAPMAP

- Genotype data, the header is **required** to be provided:

rs	alleles	Chr.	pos	strand	Assem	Cent.	protLS	assay	panel	QC	Sample 1	Sample 2
Marker 1	A/C	1	157104	+	AGPv1	Panzea	NA	NA	maize282	NA	CC	CC
Marker 2	C/G	1	194798	+	AGPv1	Panzea	NA	NA	maize282	NA	GG	GG

3.2.2 NUMERIC

- Genotype data, samples are recorded in rows. The header and sample names can be **omitted**:

taxa	marker 1	marker 2	marker 3
sample1	0	0	1
sample2	0	0	0
sample3	1	0	0

- Map information, the header is **required** to be provided:

SNP	Chromosome	Position
marker 1	1	157104
marker 2	1	1947984
marker 3	1	2914066

3.2.3 VCF

- Genotype data, the header is **required** to be provided:

Chr	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	sample 1	sample 2
1	157104	marker 1	A	C	.	PASS	.	GT	0/0	1/1
1	1947984	marker 2	C	G	.	PASS	.	GT	0/0	1/1

3.2.4 PLINK (THE HEADER **SHOULD** BE REMOVED)

- Genotype data (.ped). Missing value can be filled as "0":

Family ID	Sample ID	Paternal ID	Maternal ID	Sex	Affectio	marker	marker	marker
FAM1	NA0698	0	0	1	1	A A	T T	A A
FAM1	NA0699	0	0	1	1	A A	T T	A A
0	NA0699	0	0	1	1	C T	C C	T T

- Map information (.map):

Chromosome	Marker ID	Genetic distance	Physical Position
1	marker 1	0	157104
1	marker 2	0	1947984

3.2.5 BINARY PLINK (THE HEADER **SHOULD** BE REMOVED)

- Genotype data (.bed): Please follow the instruction from [here](#)
- FAM file:

Family ID	Sample ID	Paternal ID	Maternal ID	Sex	Affection
FAM1	NA06985	0	0	1	1
FAM1	NA06991	0	0	1	1

- BIM file:

Chromosome	Marker ID	Genetic distance	Physical Position	Allele 1	Allele 2
1	marker 1	0	157104	A	C
1	marker 2	0	1947984	A	T

3.3 COVARIATES

- Demo format for a covariate file. The header is **required** to be provided:

PC1	PC2	PC3
-1.8942149	-4.91532916	0.8674568
1.6858820	-5.08378277	-0.4069675
0.2579269	-6.29547725	2.6867939

3.4 KINSHIP

- Demo format for a kinship file. Taxa name is **required** while the header can be omitted:

taxa	sample 1	sample 2	sample 3
sample 1	2.00000000	0.22883683	0.22932180
sample 2	0.22883683	2.00000000	0.24496455
sample 3	0.22932180	0.24496455	2.00000000

- If there is no user-define kinship, a kinship will be generated by the selected package:

Package	Kinship algorithm
GAPIT	VanRaden (<i>VanRaden, 2008</i>), Loiselle (<i>Loiselle et al., 1995</i>) or EMMA (<i>Kang et al., 2008</i>)
FarmCPU	FARM-CPU (<i>Liu et al., 2016</i>)
PLINK	Not available
rrBLUP	VanRaden (<i>VanRaden, 2008</i>)
BGLR	User-provided

4. INCORPORATED PACKAGES

Tools implemented in iPat allow users to do genome-wide associate study (GWAS) and genomic selection (GS). Currently GWAS can be performed by GAPIT, FarmCPU and PLINK, and GS can be done by GAPIT, rrBLUP and BGLR in iPat. Tables below are the input arguments available in iPat:

4.1 GAPIT

Tab	Parameters	Definitions	Default
Covariates	Covariate names	Subsetting covariates data	All covariates are selected
GAPIT input	Model	Which linear model to be used in GWAS	GLM
GAPIT input	kinship.cluster	Clustering algorithm to group individuals based on their kinship	average
GAPIT input	kinship.group	Method to derive kinship among groups	Mean
Advance	SNP.fraction	Fraction of SNPs Sampled to Estimate Kinship and PCs	1
Advance	File.fragment	The Fragment Size to Read Each Time within a File	512
Advance	Model selection	Conduct Bayesian information criterion (BIC)-based model selection to find the optimal number of PCs for inclusion in the GWAS models	FALSE

4.2 FARMCPU

Category	Parameters	Definitions	Default
Covariates	Covariate names	Subsetting covariates data	All covariates are selected
FarmCPU input	method.bin	It uses fixed or optimized of possible QTN window size and number of possible QTNs selected into FarmCPU model.	static
FarmCPU input	maxLoop	Maximum number of iterations allowed	10

4.3 PLINK

Category	Parameters	Definitions	Default
Covariates	Covariate names	Subsetting covariates data	All covariates are selected
PLINK input	C.I.	The desired coverage for a confidence interval	0.95
PLINK input	Method	Regression methods of the study, available options are "GLM" and "Logistic regression"	GLM

4.4 rrBLUP

Category	Parameters	Definitions	Default
Covariates	Covariate names	Subsetting covariates data	All covariates are selected
rrBLUP input	Shrinkage estimation	Shrinkage estimation can improve the accuracy of genome-wide marker-assisted selection, particularly at low marker density (<i>Endelman and Jannink 2012</i>)	TRUE
rrBLUP input	impute.method	Imputation algorithm for missing values in markers data	mean

4.5 BGLR

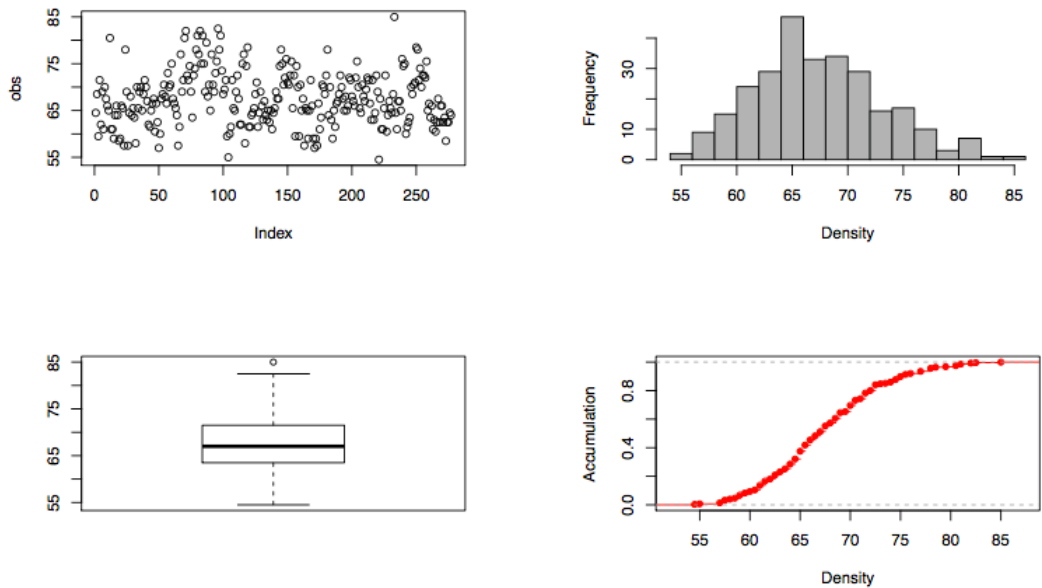
Category	Parameter	Definitions	Default
Subset	Subset of traits data	Users can select all or partial of traits to be analyzed	All traits
BGLR	Regression model for predictor (Markers)	The regression type for the markers data	BRR
BGLR	response_type	Data type of the response (y)	gaussian
BGLR	nIter	The number of iterations of the sampler	1200
BGLR	burnIn	The number of samples discarded	200
BGLR	thin	The number of thinning	5

5. Output files

5.1 PHENOTYPE

5.1.1 OVERVIEW

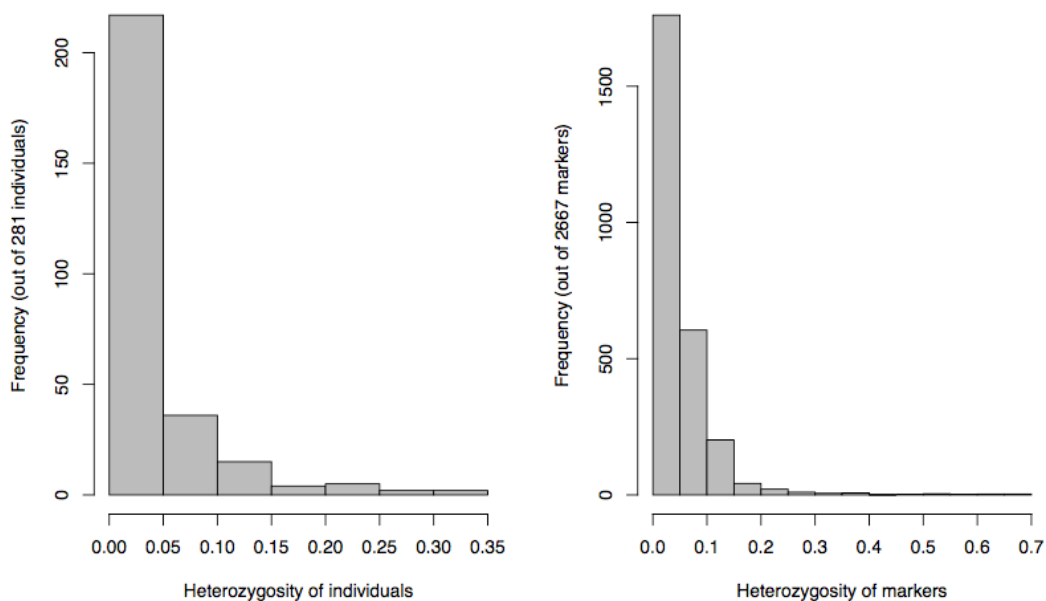
An overview of phenotype, including a scatter plot and histogram of the distribution.
(suffix: `_phenotype_view.pdf`)



5.2 Population structure

5.2.1 HETEROZYGOSITY

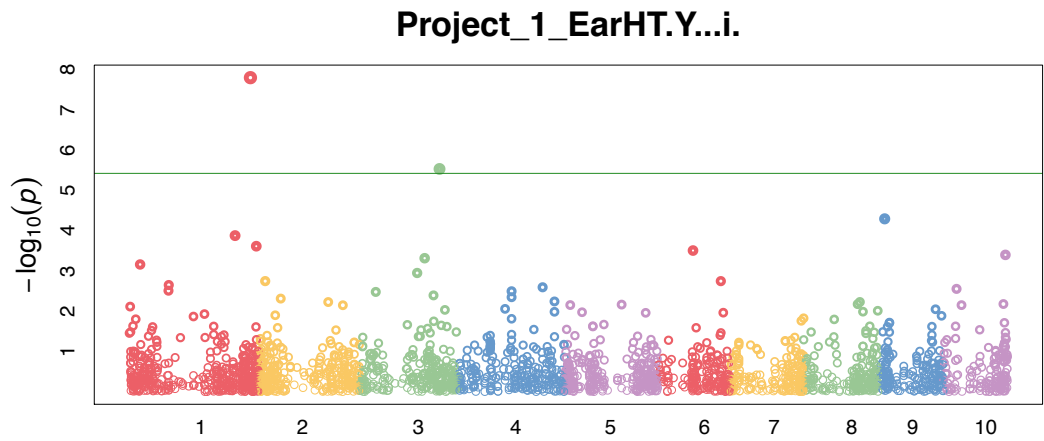
Histograms that show the heterozygosity by individuals and by markers. (suffix: `_heterozygosity.pdf`)



5.3 GWAS

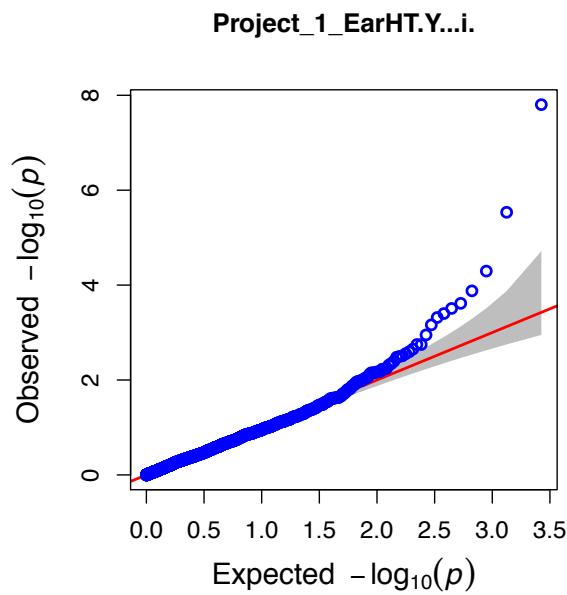
5.3.1 GENOMEWISE MANHATTAN PLOT

(suffix: Manhattan.Plot.Genomewise.pdf)



5.3.2 Q-Q PLOT

(suffix: QQ-Plot.pdf)



5.3.3 GWAS RESULT

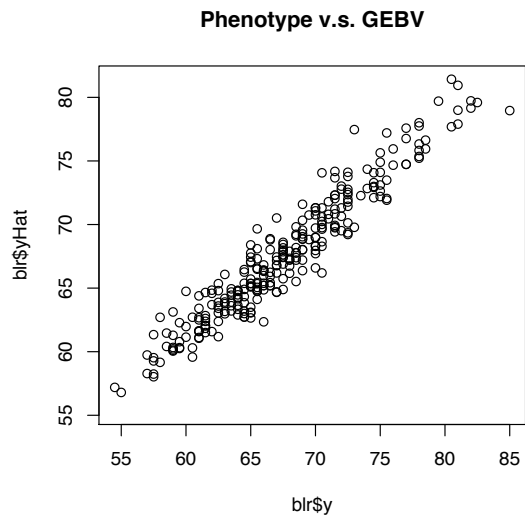
A table shows the marker information and its tested P-value. (suffix: _GWAS.txt)

SNP	Chromosome	Position	P.value	MAF
PZB00859.1	1	157104	0.4636	0.2402
PZA01271.1	1	1947984	0.1585	0.4893

5.4 GS

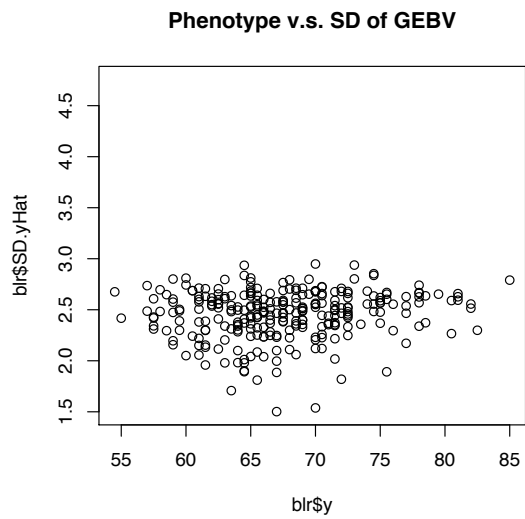
5.4.1 GENOMIC ESTIMATED BREEDING VALUE (GEBV)

A scatter plot shows the correlation between phenotype and genomic estimated breeding value (GEBV)
(suffix: _GEBV_value.pdf)



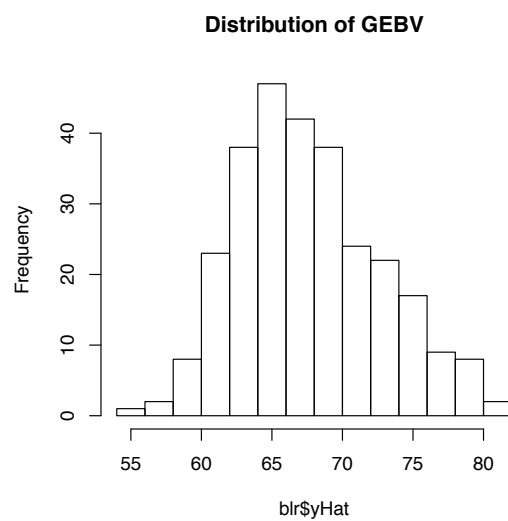
5.4.2 STANDARD DEVIATION OF GEBV

A scatter plot shows the correlation between phenotype and SD of GEBV
(suffix: _GEBV_var.pdf)



5.4.3 HISTOGRAM OF GEBV

(suffix: _GEBV_hist.pdf)



5.4.4 PREDICTION RESULT

(suffix: _EBV.dat)

Taxa	Prediction	SD of prediction
33-16	36.576341	4.0006
38-11	38.401256	0.7429

6. TUTORIAL

6.1 PERFORM GWAS IN IPAT

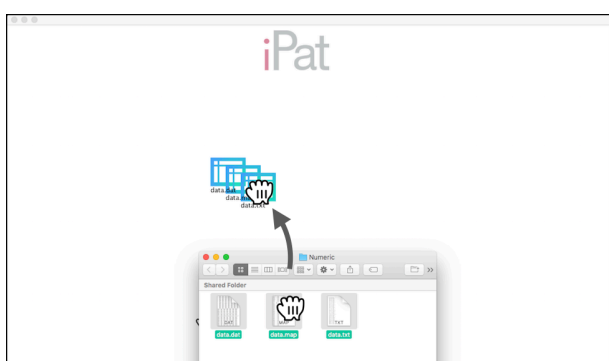
DATA FORMAT : NUMERICAL (COLUMNS AS MARKERS)

DATA REQUIRED :

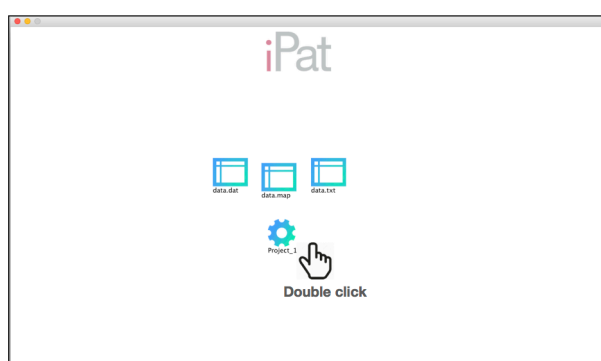
- Phenotype
- Genotype
- Map information

IMPLEMENTED PACKAGE: FARMCPU

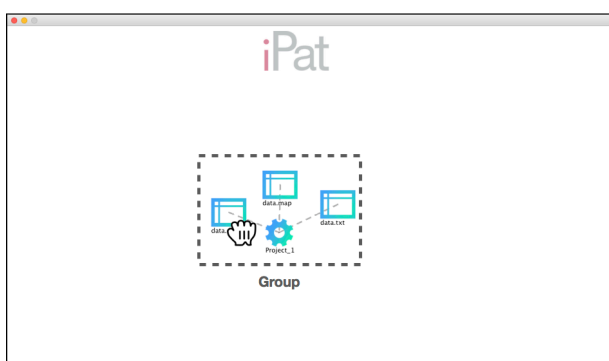
Step 1: Drag all the required files into iPat



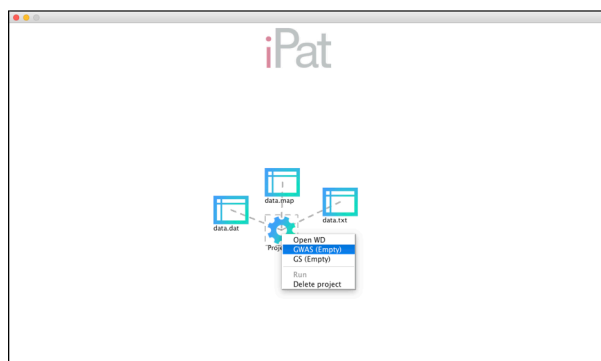
Step 2: Create a new project for files



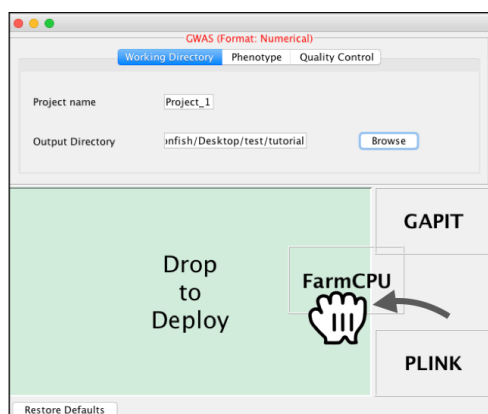
Step 3: Drag files and hover over the project to build a group



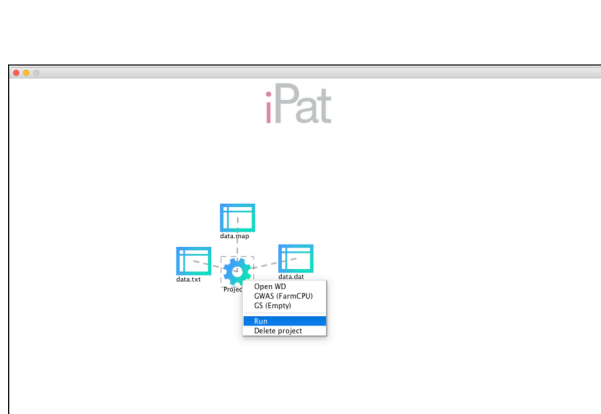
Step 4: Right click on the project and choose "GWAS"



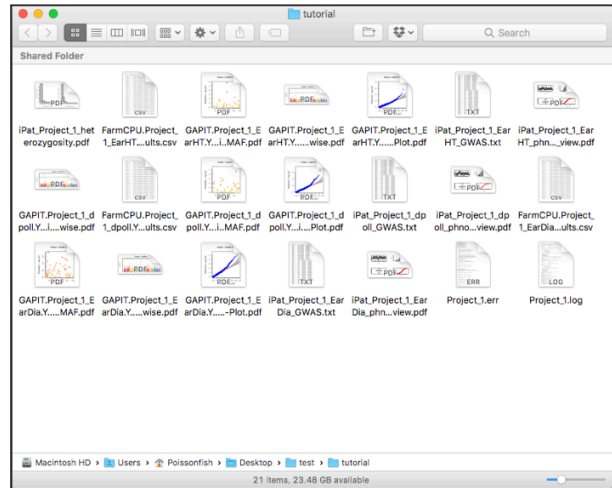
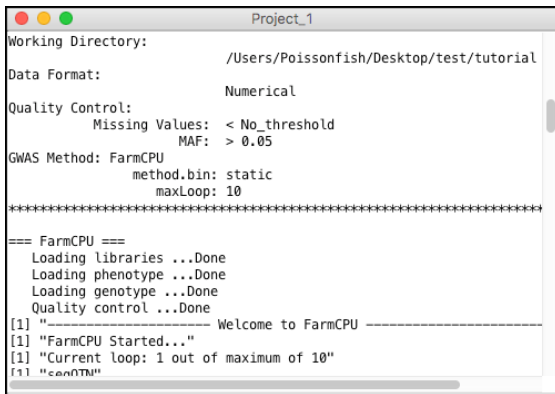
Step 5: Drag "FarmCPU" to the left, then close the window



Step 6: Right click on the project and choose "Run"



Step 7: Double click on the project to inspect results after a finished computing



6.2 PERFORM GS AND ADD USER-DEFINE COVARIATES IN IPAT

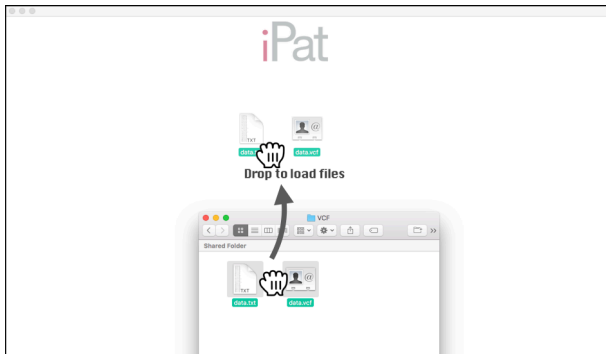
DATA FORMAT : VCF (COLUMNS AS MARKERS)

DATA REQUIRED :

- Phenotype
- Genotype (.vcf)
- Covariates

IMPLEMENTED PACKAGE: RRBLUP

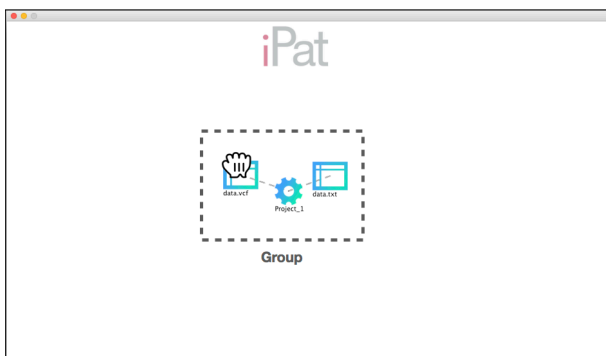
Step 1: Drag all the required files into iPat



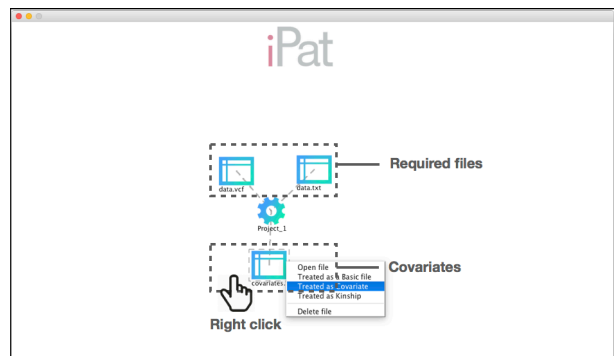
Step 2: Create a new project for files



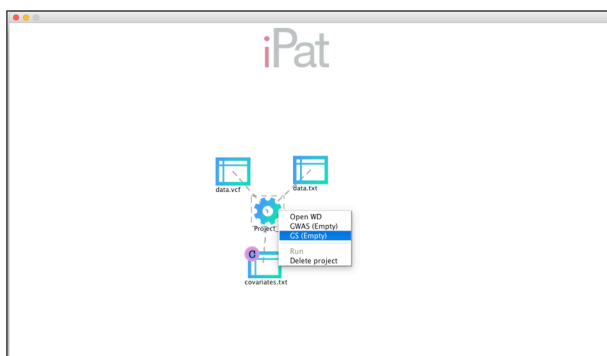
Step 3: Drag files and hover over the project to build a group



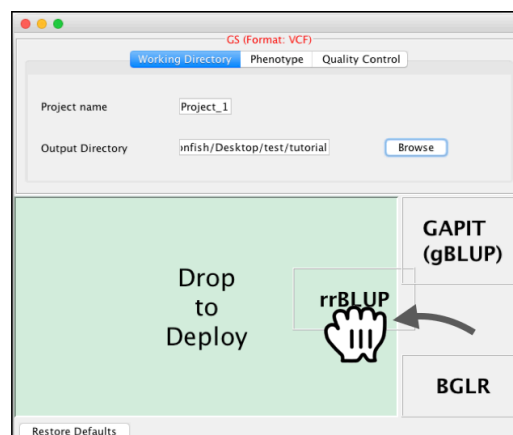
Step 4: Add a covariates file and assign it as covariates



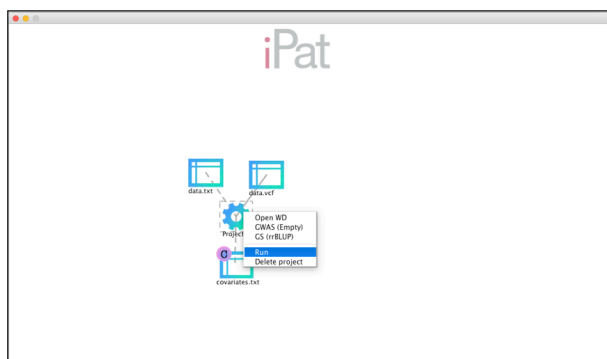
Step 5: Right click on the project and choose "GS"



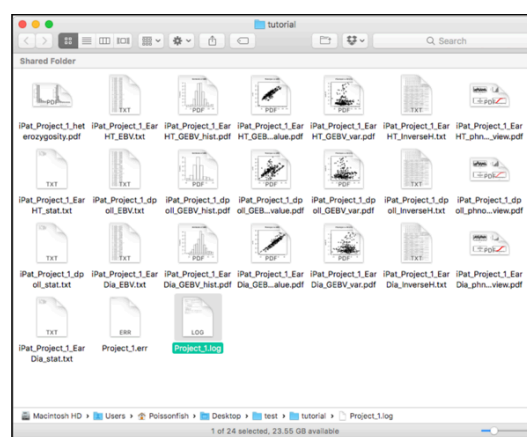
Step 6: Drag a label "GAPIT" to the left and close window



Step 7: Right click on the project and choose "Run"



Step 8: Double click on the project to inspect results



6.3 PERFORM GWAS-ASSIST GS IN IPAT

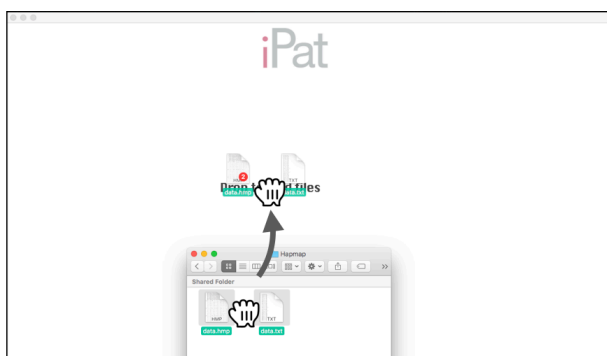
DATA FORMAT : HAPMAP

DATA REQUIRED :

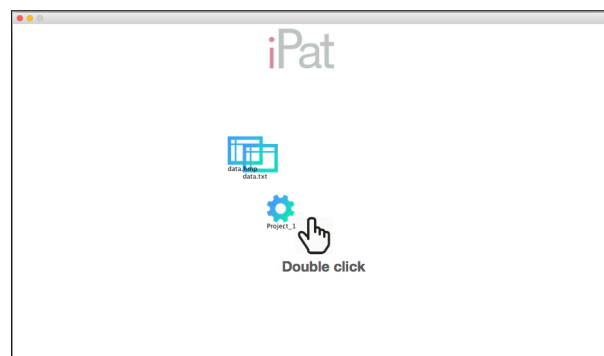
- Phenotype
- Genotype (.hmp)

IMPLEMENTED PACKAGE: BGLR

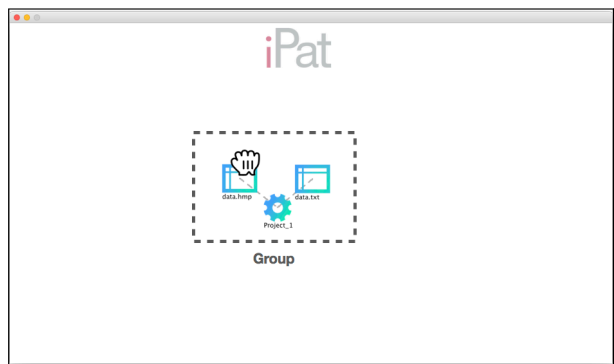
Step 1: Drag all the required files into iPat



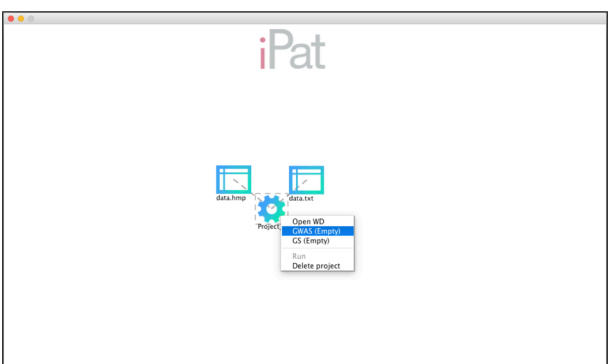
Step 2: Create a new project for files



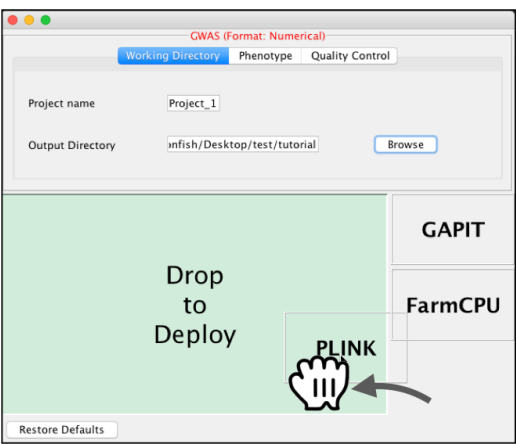
Step 3: Drag files and hover over the project to build a group



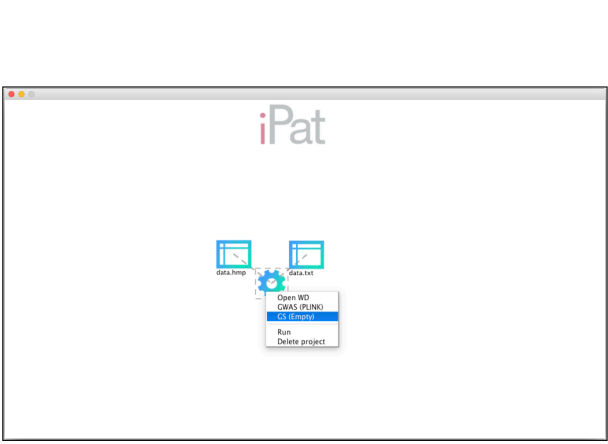
Step 4: Right click on the project and choose "GWAS"



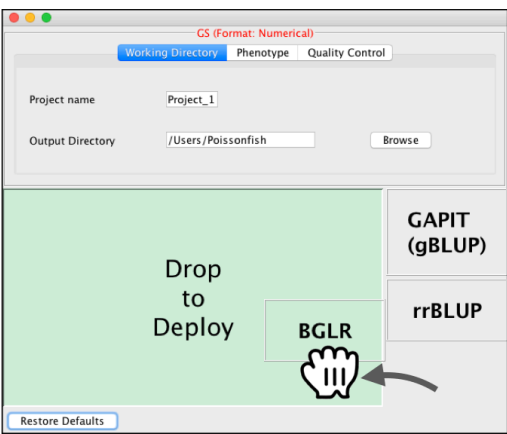
Step 5: Drag "PLINK" to the left, then close the window



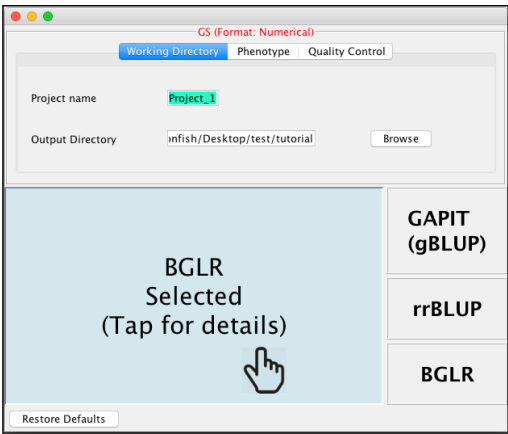
Step 6: Right click on the project and choose "GS"



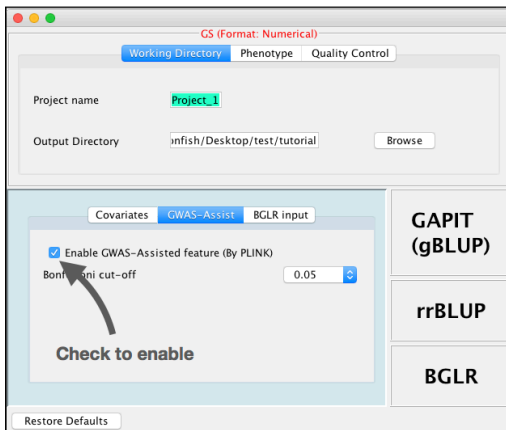
Step 7: Drag "BGLR" to the left, then close the window



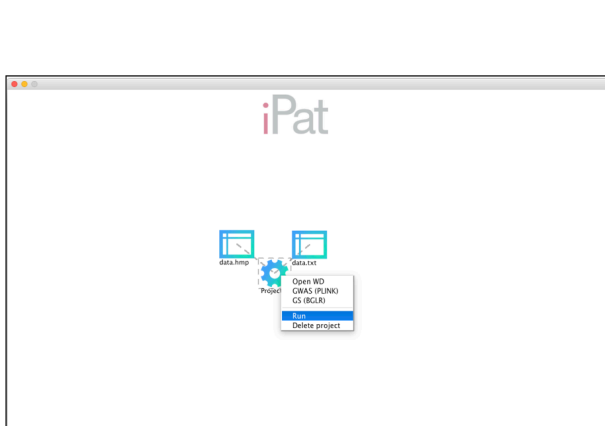
Step 8: Click on the left to further define GS



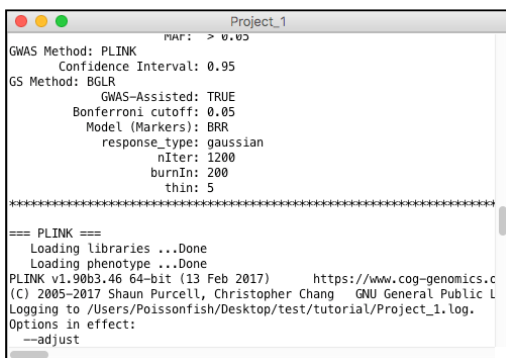
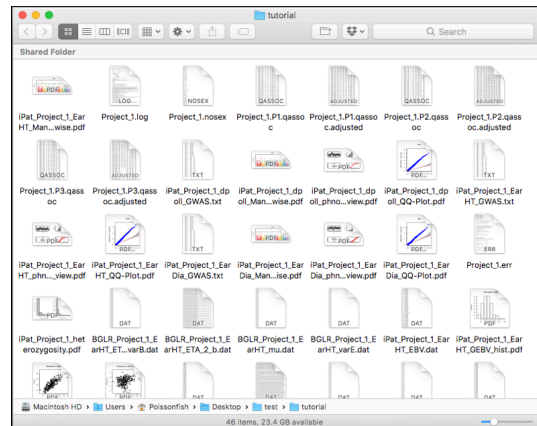
Step 9: Select “GWAS-Assist” tab and check “Enable GWAS-Assisted...” to enable GWAS-assisted GS feature



Step 10: Right click on the project and choose “Run”



Step 11: Double click on the project to inspect results



7. REFERENCES

- Endelman J., 2011 Ridge regression and other kernels for genomic selection in the R package rrBLUP. *Plant Genome* 4: 250–255.
- Lipka A. E., Tian F., Wang Q., Peiffer J., Li M., *et al.*, 2012 GAPIT: genome association and prediction integrated tool. *Bioinformatics* 28: 2397–2399.
- Liu X., Huang M., Fan B., Buckler E. S., Zhang Z., 2016 Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. *PLoS Genet.* 12: e1005767.
- Pérez P., Los Campos G. De, 2014 Genome-wide regression and prediction with the BGLR statistical package. *Genetics* 198: 483–495.
- Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M. A. R., *et al.*, 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81: 559–575.
- Tang Y., Liu X., Wang J., Li M., Wang Q., *et al.*, 2016 GAPIT Version 2: An Enhanced Integrated Tool for Genomic Association and Prediction. *Plant J.* 9.