Structured Association Mapping using STRUCTURE and TASSEL

K K Vinod

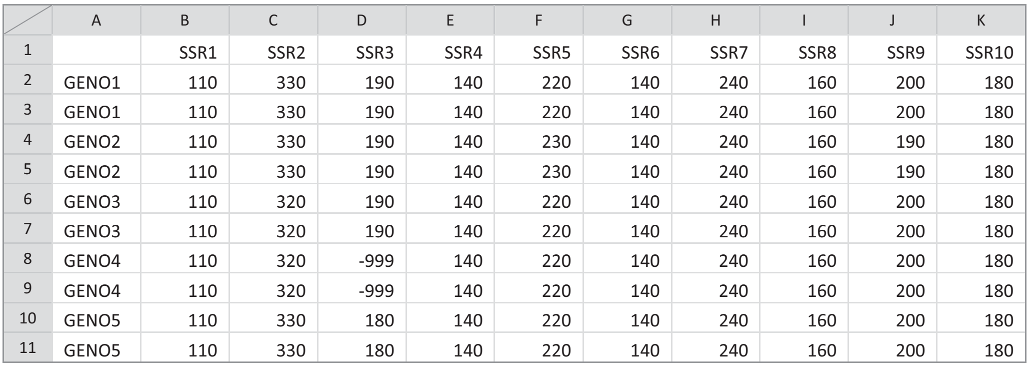
Indian Agricultural Research Institute

With advances in genotyping technology, including rapid increases in the number of genetic markers available for QTL studies, association analysis is now a viable approach for the dissection of complex genetic traits (Churchill et al. 2004). Association mapping involves assessment of population structure and using this population information and kinship information among individuals to assess marker – trait association. Two common software packages widely used today for association mapping are STRUCTURE (Pritchard et al. 2010) and TASSEL (Buckler et al. 2009). STRUCTURE implements a model-based clustering method for inferring population structure using genotype data consisting of unlinked markers. This program can demonstrate the presence of population structure, identify distinct genetic populations, assign individuals to populations, and identify migrants and admixed individuals. Trait Analysis by Association, Evolution and Linkage, or TASSEL, makes use of the most advanced statistical methods to maximize statistical power for finding QTL. Both a structured association approach (Pritchard *et al.* 2000; Thornsberry *et al.* 2001) and a unified mixed model method have been implemented to minimize the risk of false positives by integrating population structure and family relatedness within populations (Yu *et al.* 2006).

I. Determining population structure using Structure 3.2.2

A. Preparation of marker genotype data

Prepare a matrix of marker genotype data in Excel as given below, for microsatellite data:



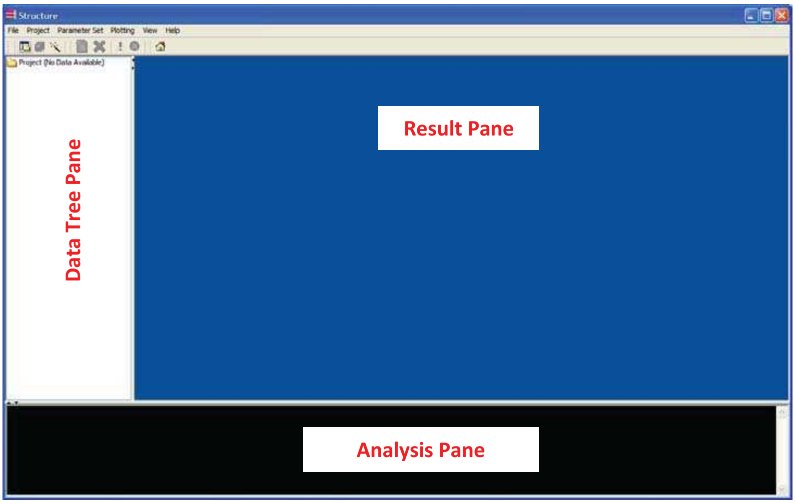
	A	B	C	D	E	F	G	H	I	J	K
1		SSR1	SSR2	SSR3	SSR4	SSR5	SSR6	SSR7	SSR8	SSR9	SSR10
2	GENO1	110	330	190	140	220	140	240	160	200	180
3	GENO1	110	330	190	140	220	140	240	160	200	180
4	GENO2	110	330	190	140	230	140	240	160	190	180
5	GENO2	110	330	190	140	230	140	240	160	190	180
6	GENO3	110	320	190	140	220	140	240	160	200	180
7	GENO3	110	320	190	140	220	140	240	160	200	180
8	GENO4	110	320	-999	140	220	140	240	160	200	180
9	GENO4	110	320	-999	140	220	140	240	160	200	180
10	GENO5	110	330	180	140	220	140	240	160	200	180
11	GENO5	110	330	180	140	220	140	240	160	200	180

SSR is the code for markers; GENO is for genotype; -999: missing data value

- Save the data file in Text (tab delimited) type with a suitable filename <genodata.txt>.

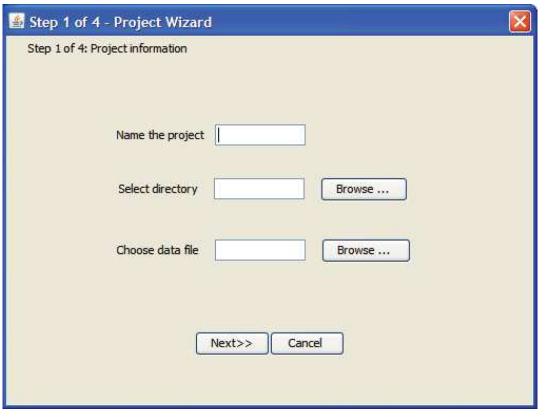
B. Download and install STRUCTURE. Latest version of STRUCTURE is available for download at http://pritch.bsd.uchicago.edu/structure.html.

- Run the Structure 3.2.2 software, by double clicking the icon at desktop.


Structure
File Project Parameter Set Plotting View Help
Project (No Data Available)
Result Pane
Data Tree Pane
Analysis Pane

(1) Building a project

- Click on **File > New Project**



Step 1 of 4 - Project Wizard

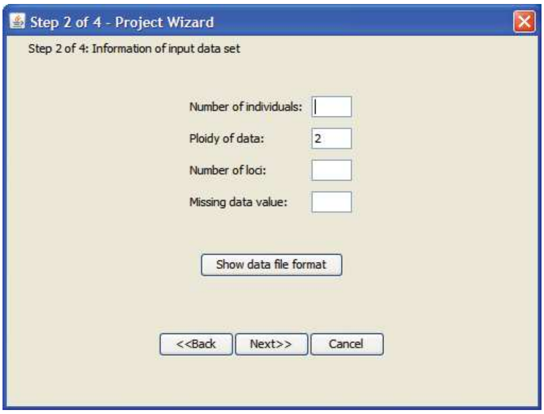
Step 1 of 4: Project information

Name the project

Select directory

Choose data file

- Fill in these boxes: Name of project, Select directory and Choose data file
- Select the file saved in step A and Click **Next**



Step 2 of 4 - Project Wizard

Step 2 of 4: Information of input data set

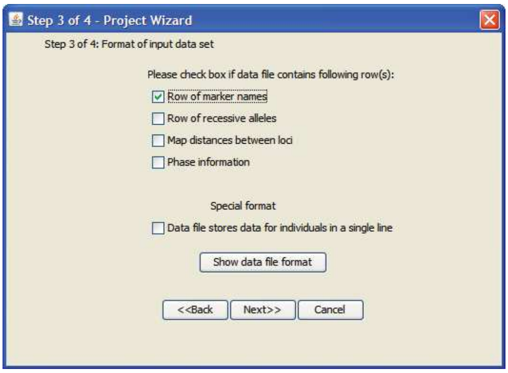
Number of individuals:

Ploidy of data: 2

Number of loci:

Missing data value:

- Fill in these boxes: number of individuals, ploidy of data ('2' for diploid), number of loci, and missing data value ('-999'). Click [**Next]**


Step 3 of 4 - Project Wizard

Step 3 of 4: Format of input data set

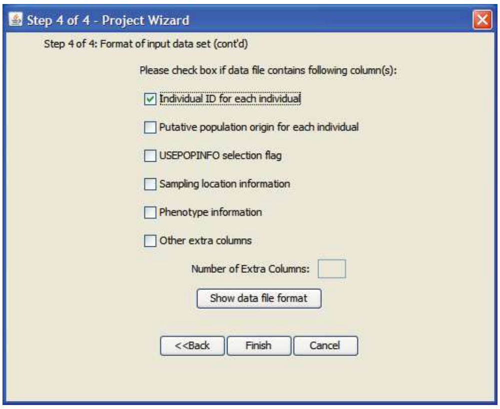
Please check box if data file contains following row(s):

☑ Row of marker names
☐ Row of recessive alleles
☐ Map distances between loci
☐ Phase information

Special format

☐ Data file stores data for individuals in a single line

- Since our data contains marker and genotype labels, check 'Row and marker names'. Click [**Next]**


Step 4 of 4 - Project Wizard

Step 4 of 4: Format of input data set (cont'd)

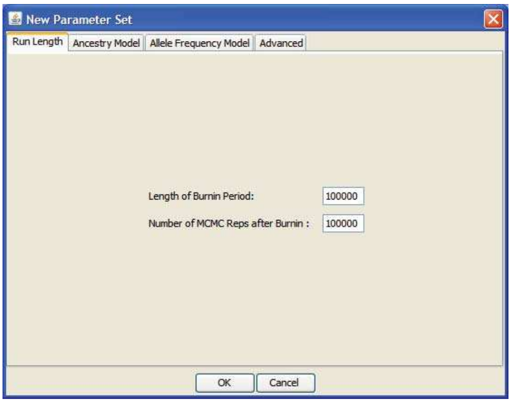
Please check box if data file contains following column(s):

☑ Individual ID for each individual
☐ Putative population origin for each individual
☐ USEPOPINFO selection flag
☐ Sampling location information
☐ Phenotype information
☐ Other extra columns

Number of Extra Columns:

- Since the data file contains genotypes labels, check **Individual ID for each Individual.** Click [**Finish]**

(2) When project is done, a parameter set needs to be configured. For this, in the STRUCTURE Main window, click on **Parameter Set > New**

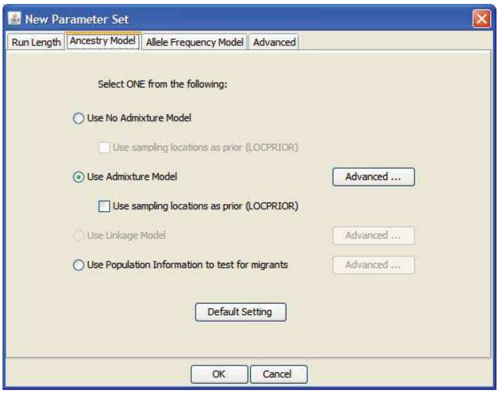

New Parameter Set

Length of Burnin Period: 100000

Number of MCMC Reps after Burnin : 100000

- Fill in these boxes: Length of Burnin Period: (100000), and Number of Markov chain Monte Carlo (MCMC) Reps (simulations) after Burnin: (100000).

  This number should be high, preferably more than 100000 to get reliable convergence.
- Click **[OK]** button
- In the **Ancestry Model** tab, select Use Admixture Model (This is the default). Click **[OK]** button.



Select ONE from the following:

○ Use No Admixture Model

□ Use sampling locations as prior (LOCPRIOR)

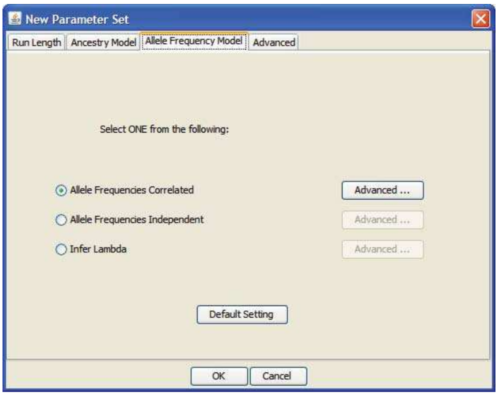
● Use Admixture Model

□ Use sampling locations as prior (LOCPRIOR)

○ Use Linkage Model

○ Use Population Information to test for migrants

- In the Allele Frequency Model tab, select **Alleles Frequencies Correlated**



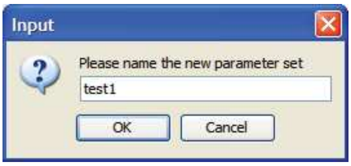
Select ONE from the following:

● Allele Frequencies Correlated

○ Allele Frequencies Independent

○ Infer Lambda

- Click **[OK]** button.



Please name the new parameter set

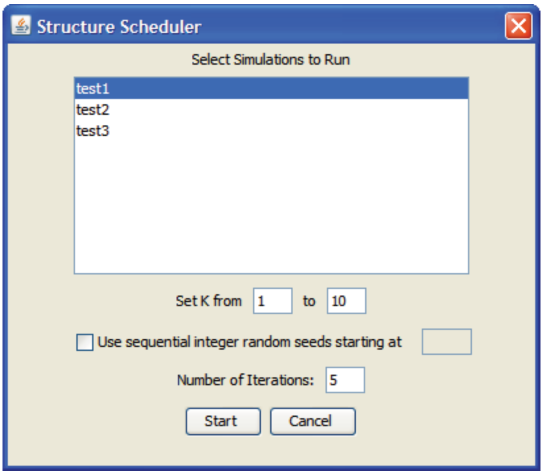
test1

- Name the newly created parameter set in the input dialogue (e.g. test1)
- Click **[OK]**.

(3) Running Simulations

[If optimum population structure (K) is already known by some other means, then skip this step and go to step (5)]

- In the STRUCTURE main window, click on Project > Start a job



Structure Scheduler

Select Simulations to Run

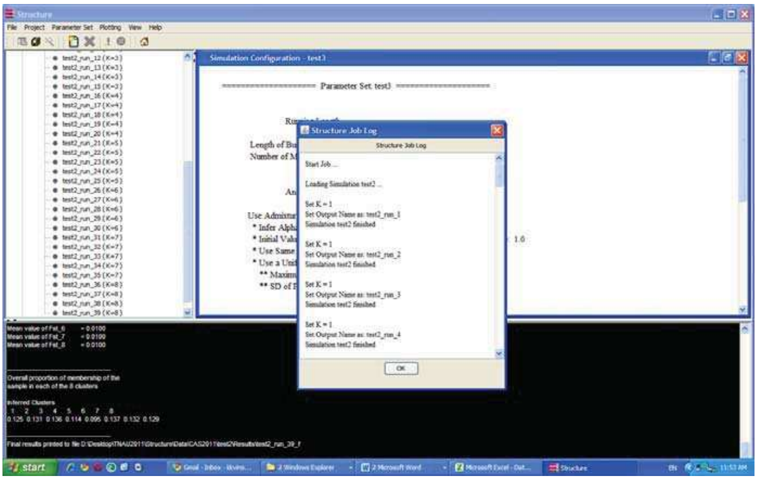
test1
test2
test3

Set K from 1 to 10

☐ Use sequential integer random seeds starting at

Number of Iterations: 5

- In the Scheduler dialogue, select the Parameter set you want to run (e.g. test1);
- Set K between a range say 1 to 10;
- Set the number of replications (Iterations) to run (e.g. 5)
- Click **[Start]**



```
test2_run_12 (K=3)
test2_run_13 (K=3)
test2_run_14 (K=3)
test2_run_15 (K=3)
test2_run_16 (K=4)
test2_run_17 (K=4)
test2_run_18 (K=4)
test2_run_19 (K=4)
test2_run_20 (K=4)
test2_run_21 (K=5)
test2_run_22 (K=5)
test2_run_23 (K=5)
test2_run_24 (K=5)
test2_run_25 (K=5)
test2_run_26 (K=6)
test2_run_27 (K=6)
test2_run_28 (K=6)
test2_run_29 (K=6)
test2_run_30 (K=6)
test2_run_31 (K=7)
test2_run_32 (K=7)
test2_run_33 (K=7)
test2_run_34 (K=7)
test2_run_35 (K=7)
test2_run_36 (K=8)
test2_run_37 (K=8)
test2_run_38 (K=8)
test2_run_39 (K=8)
```

Simulation Configuration - test3

```
==================== Parameter Set: test3 ====================
```

Structure Job Log

```
Start Job ...

Loading Simulation test2 ...

Set K = 1
Set Output Name as: test2_run_1
Simulation test2 finished

Set K = 1
Set Output Name as: test2_run_2
Simulation test2 finished

Set K = 1
Set Output Name as: test2_run_3
Simulation test2 finished

Set K = 1
Set Output Name as: test2_run_4
Simulation test2 finished
```

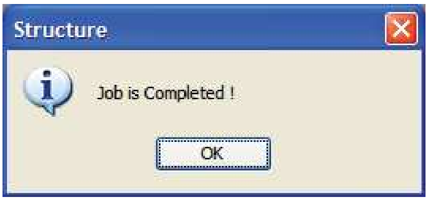
```
Mean value of Fst_6      = 0.0100
Mean value of Fst_7      = 0.0100
Mean value of Fst_8      = 0.0100

--------------------------------------------
Overall proportion of membership of the
sample in each of the 8 clusters

Inferred Clusters
  1     2     3     4     5     6     7     8
0.125 0.131 0.136 0.114 0.095 0.137 0.132 0.129

--------------------------------------------
Final results printed to file D:\Desktop\TNAU2011\Structure\Data\CAS2011\test2\Results\test2_run_39_f
```

- The program will run for a very long time (>72 hours) depending on the speed of the computer, size of the data, and number of iterations and the replications defined in the parameter set.



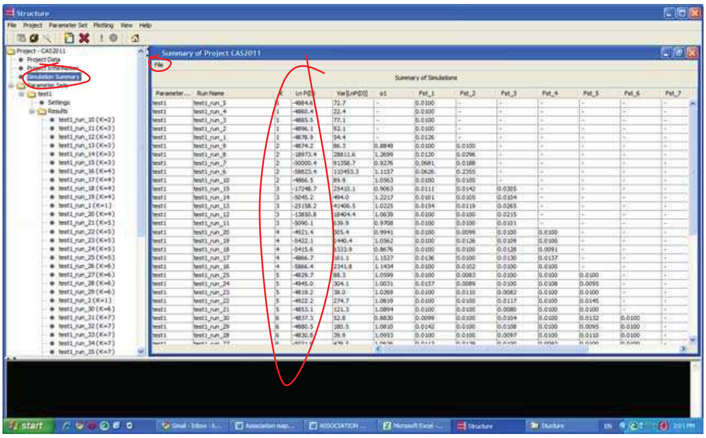
Structure

Job is Completed !

- STRUCTURE displays Job is Completed! dialogue after successful analysis.

(4) Determining the optimum population structure

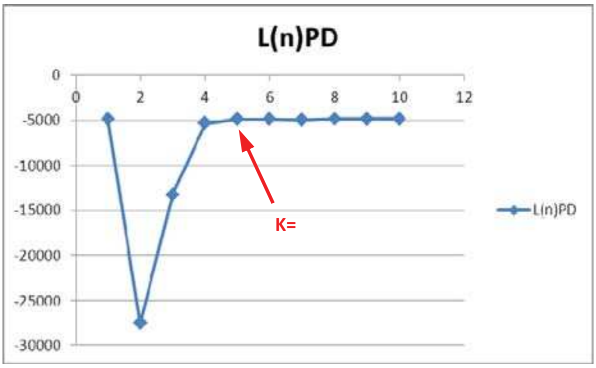
- To determine optimum value for K, Click on the Simulation Summary in the Tree Pane of STRUCTURE window



Summary of Simulations

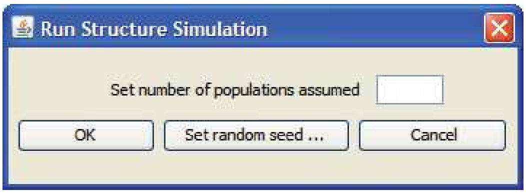
Parameter...	Run Name	K	Ln P(D)	Var[LnP(D)]	α1	Fst_1	Fst_2	Fst_3	Fst_4	Fst_5	Fst_6	Fst_7
test1	test1_run_5	1	-4884.6	72.7	-	0.0100	-	-	-	-	-	-
test1	test1_run_4	1	-4860.4	22.4	-	0.0100	-	-	-	-	-	-
test1	test1_run_3	1	-4885.5	77.1	-	0.0100	-	-	-	-	-	-
test1	test1_run_2	1	-4896.1	92.1	-	0.0100	-	-	-	-	-	-
test1	test1_run_1	1	-4878.9	54.4	-	0.0126	-	-	-	-	-	-
test1	test1_run_9	2	-4874.2	86.3	0.8849	0.0100	0.0100	-	-	-	-	-
test1	test1_run_8	2	-18973.4	28811.6	1.2699	0.0120	0.0296	-	-	-	-	-
test1	test1_run_7	2	-50000.4	91358.7	0.9276	0.0681	0.0188	-	-	-	-	-
test1	test1_run_6	2	-58825.4	110453.3	1.1157	0.0626	0.2355	-	-	-	-	-
test1	test1_run_10	2	-4866.5	89.9	1.0563	0.0100	0.0100	-	-	-	-	-
test1	test1_run_15	3	-17248.7	25410.1	0.9063	0.0111	0.0142	0.0365	-	-	-	-
test1	test1_run_14	3	-5045.2	494.0	1.2217	0.0101	0.0105	0.0104	-	-	-	-
test1	test1_run_13	3	-25158.2	41406.5	1.0225	0.0154	0.0119	0.0265	-	-	-	-
test1	test1_run_12	3	-13850.8	18404.4	1.0639	0.0100	0.0100	0.0215	-	-	-	-
test1	test1_run_11	3	-5090.1	639.9	0.9708	0.0100	0.0100	0.0101	-	-	-	-
test1	test1_run_20	4	-4921.4	305.4	0.9941	0.0100	0.0099	0.0100	0.0100	-	-	-
test1	test1_run_19	4	-5422.1	1440.4	1.0562	0.0100	0.0126	0.0109	0.0100	-	-	-
test1	test1_run_18	4	-5415.6	1333.9	0.8676	0.0100	0.0100	0.0128	0.0091	-	-	-
test1	test1_run_17	4	-4866.7	461.1	1.1527	0.0136	0.0100	0.0130	0.0137	-	-	-
test1	test1_run_16	4	-5866.4	2341.8	1.1434	0.0100	0.0102	0.0100	0.0100	-	-	-
test1	test1_run_25	5	-4829.7	88.3	1.0599	0.0100	0.0083	0.0100	0.0100	0.0100	-	-
test1	test1_run_24	5	-4945.0	304.1	1.0031	0.0157	0.0086	0.0100	0.0108	0.0095	-	-
test1	test1_run_23	5	-4819.2	38.0	1.0269	0.0100	0.0110	0.0082	0.0100	0.0100	-	-
test1	test1_run_22	5	-4922.2	274.7	1.0819	0.0100	0.0100	0.0117	0.0100	0.0145	-	-
test1	test1_run_21	5	-4853.1	121.3	1.0894	0.0100	0.0100	0.0080	0.0100	0.0100	-	-
test1	test1_run_30	6	-4837.3	52.8	0.8830	0.0099	0.0100	0.0104	0.0100	0.0132	0.0100	-
test1	test1_run_29	6	-4880.5	180.5	1.0810	0.0142	0.0100	0.0108	0.0100	0.0095	0.0100	-
test1	test1_run_28	6	-4830.8	39.9	1.0953	0.0100	0.0100	0.0097	0.0100	0.0110	0.0100	-

- In the Result pane, click File on the left hand top corner, to save the simulation summary in a text file.
- Copy the values of K and Ln P(D) into a convenient data editor (Excel is the best option) and calculate average of Ln P(D) against each K, across replications. The K at which Ln P(D) plateaus is to be taken as optimum K.



(5) Once optimum population structure (K) is known, estimate Inferred ancestry (Q matrix) of individuals,

- In the STRUCTURE main window, Select **Parameter Set >Run**



Set number of populations assumed

- Enter the value of K in the box and click **[OK].**

- When the job is complete, go to the result pane and select the latest run from the results folder
- From the results on the right pane, select inferred ancestry of individuals
- Copy and paste it in notepad.

Alternately,

- Go to the directory in where you save the project, open the folder with project name, and open the result folder. There will be several files with a "f" suffix.
- Open the file with later run number in Notepad. In this output file, copy the values of "Inferred ancestry of individuals"
- Inferred ancestry of individuals (Q matrix) is used as covariate in TASSEL.
- A typical Q matrix will look like as follows:

```
1   GENO1   0 :    0.212  0.020  0.131  0.043  0.593
2   GENO2   0 :    0.173  0.201  0.538  0.062  0.027
3   GENO3   0 :    0.200  0.270  0.131  0.189  0.211
4   GENO4   0 :    0.092  0.506  0.155  0.142  0.105
5   GENO5   0 :    0.124  0.329  0.046  0.427  0.074
6   GENO6   0 :    0.339  0.053  0.096  0.450  0.062
7   GENO7   0 :    0.343  0.039  0.246  0.120  0.251
8   GENO8   0 :    0.376  0.208  0.201  0.059  0.155
9   GENO9   0 :    0.172  0.044  0.590  0.137  0.058
10  GENO10  0 :    0.172  0.163  0.131  0.445  0.089
11  GENO11  0 :    0.093  0.470  0.101  0.165  0.171
12  GENO12  0 :    0.313  0.156  0.237  0.108  0.187
13  GENO13  0 :    0.184  0.371  0.299  0.030  0.117
14  GENO14  0 :    0.078  0.159  0.036  0.675  0.052
15  GENO15  0 :    0.705  0.076  0.065  0.078  0.077
```

- Save the Q matrix. This need to be formatted to be read in TASSEL

II. Association Analysis using TASSEL

Association mapping can produce spurious association between marker and phenotype; therefore, the population structure is an important component in estimating marker – trait associations. This is done by incorporating the Q matrix of inferred ancestry coefficients of the individuals across the sub-populations as covariate in the association mapping analysis.

To refine the results, the kinship coefficients are also used in association analysis. Kinship matrix (K matrix) can be estimated using software such as SPAGeDi (Hardy and Vekemans, 2002) or can be estimated within TASSEL itself.

Unlike that of STRUCTURE which is a complete program by itself, TASSEL stand-alone version runs only under Java runtime environment (JRE) version 1.5 and above. JRE is freely downloadable software from Sun Microsystems, http://java.sun.com/. Alternatively, online versions of TASSEL are also available.

Note: Latest version of TASSEL 3.0 does not support microsatellite data anymore. So SSR data analysis can be done only using TASSEL version 2.1.

Both TASSEL 2.1 and 3.0 are available for free download at the following website:

http://www.maizegenetics.net/index.php?option=com_content&task=view&id=89&Itemid=119.

- Once software platforms are ready, double clicking on the file sTASSEL.jar will run TASSEL 2.1.

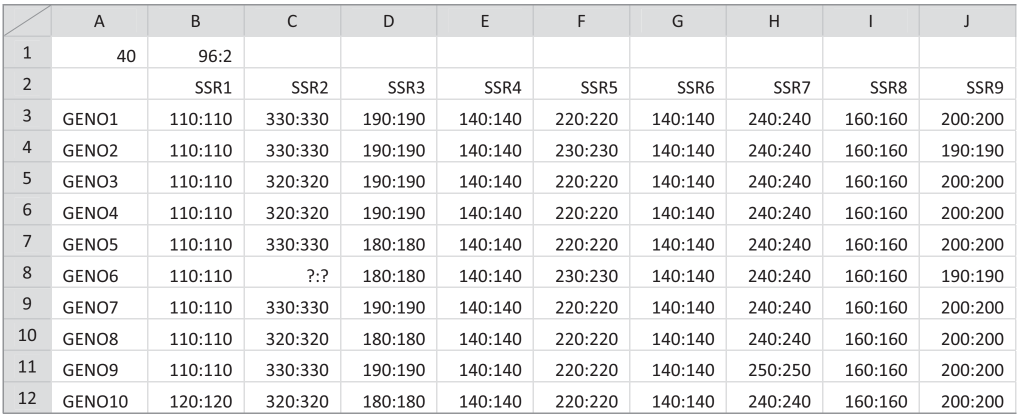
A. Preparation of data

TASSEL requires three types of data primarily for the analysis. (i) Marker segregation data (ii) Phenotype data and (iii) Ancestry coefficient data (Q matrix)

- Prepare these data in Excel, and save as Text Tab delimited (*.txt) files.

(a) Genotype data

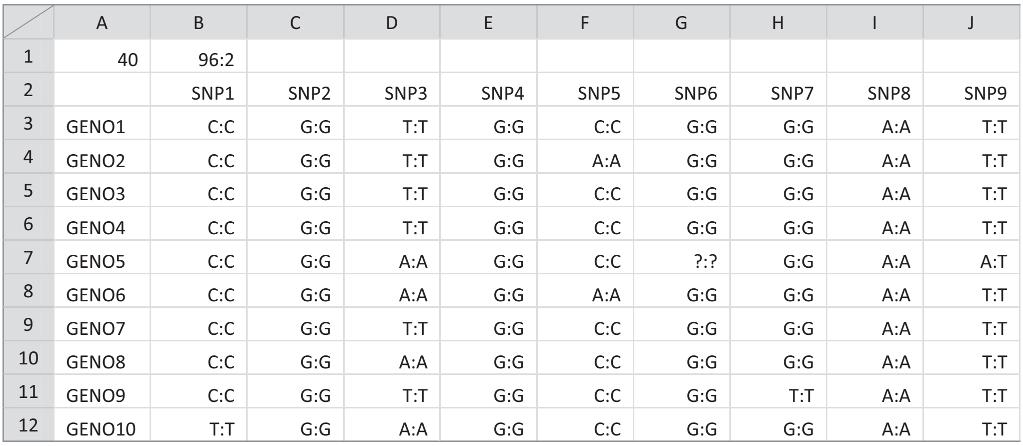
Genotype data using microsatellites uses the following format:



	A	B	C	D	E	F	G	H	I	J
1	40	96:2								
2		SSR1	SSR2	SSR3	SSR4	SSR5	SSR6	SSR7	SSR8	SSR9
3	GENO1	110:110	330:330	190:190	140:140	220:220	140:140	240:240	160:160	200:200
4	GENO2	110:110	330:330	190:190	140:140	230:230	140:140	240:240	160:160	190:190
5	GENO3	110:110	320:320	190:190	140:140	220:220	140:140	240:240	160:160	200:200
6	GENO4	110:110	320:320	190:190	140:140	220:220	140:140	240:240	160:160	200:200
7	GENO5	110:110	330:330	180:180	140:140	220:220	140:140	240:240	160:160	200:200
8	GENO6	110:110	?:?	180:180	140:140	230:230	140:140	240:240	160:160	190:190
9	GENO7	110:110	330:330	190:190	140:140	220:220	140:140	240:240	160:160	200:200
10	GENO8	110:110	320:320	180:180	140:140	220:220	140:140	240:240	160:160	200:200
11	GENO9	110:110	330:330	190:190	140:140	220:220	140:140	250:250	160:160	200:200
12	GENO10	120:120	320:320	180:180	140:140	220:220	140:140	240:240	160:160	200:200

Note: The number in the first row tell TASSEL, the number of individuals, followed by number of markers, and (:2) indicate diploid nature of the individuals. ? is commonly used for missing data. *Don't put individual with missing data in the first row. Instead, move it into another row.*

Genotype data using SNP uses the following format:



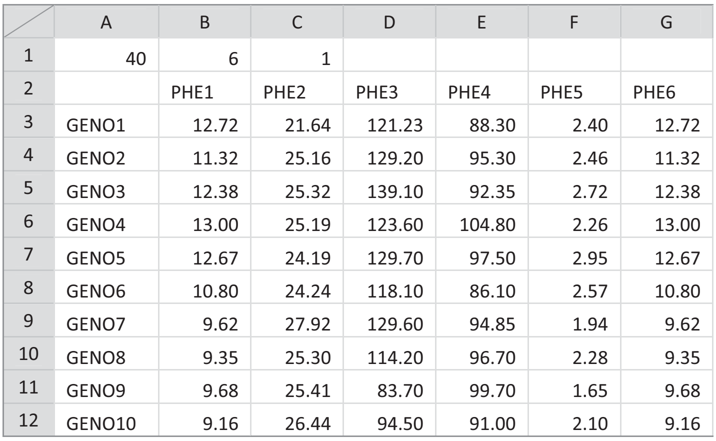
	A	B	C	D	E	F	G	H	I	J
1	40	96:2								
2		SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9
3	GENO1	C:C	G:G	T:T	G:G	C:C	G:G	G:G	A:A	T:T
4	GENO2	C:C	G:G	T:T	G:G	A:A	G:G	G:G	A:A	T:T
5	GENO3	C:C	G:G	T:T	G:G	C:C	G:G	G:G	A:A	T:T
6	GENO4	C:C	G:G	T:T	G:G	C:C	G:G	G:G	A:A	T:T
7	GENO5	C:C	G:G	A:A	G:G	C:C	?:?	G:G	A:A	A:T
8	GENO6	C:C	G:G	A:A	G:G	A:A	G:G	G:G	A:A	T:T
9	GENO7	C:C	G:G	T:T	G:G	C:C	G:G	G:G	A:A	T:T
10	GENO8	C:C	G:G	A:A	G:G	C:C	G:G	G:G	A:A	T:T
11	GENO9	C:C	G:G	T:T	G:G	C:C	G:G	T:T	A:A	T:T
12	GENO10	T:T	G:G	A:A	G:G	C:C	G:G	G:G	A:A	T:T

Note: The number in the first row tell TASSEL, the number of individuals, followed by number of markers, and (:2) indicate diploid nature of the individuals. ? is commonly used for missing data. *Don't put individual with missing data in the first row. Instead, move it into another row.*

- Save the data matrix in Text (Tab delimited) type with a suitable filename, <Markername.txt>.

(b) Phenotype data

Phenotype data uses following format:



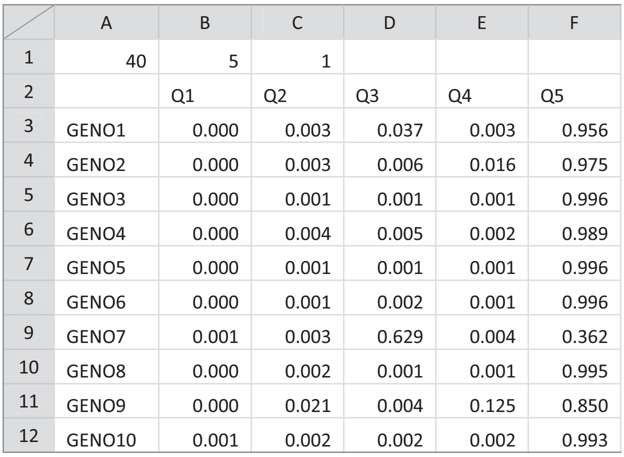
	A	B	C	D	E	F	G
1	40	6	1				
2		PHE1	PHE2	PHE3	PHE4	PHE5	PHE6
3	GENO1	12.72	21.64	121.23	88.30	2.40	12.72
4	GENO2	11.32	25.16	129.20	95.30	2.46	11.32
5	GENO3	12.38	25.32	139.10	92.35	2.72	12.38
6	GENO4	13.00	25.19	123.60	104.80	2.26	13.00
7	GENO5	12.67	24.19	129.70	97.50	2.95	12.67
8	GENO6	10.80	24.24	118.10	86.10	2.57	10.80
9	GENO7	9.62	27.92	129.60	94.85	1.94	9.62
10	GENO8	9.35	25.30	114.20	96.70	2.28	9.35
11	GENO9	9.68	25.41	83.70	99.70	1.65	9.68
12	GENO10	9.16	26.44	94.50	91.00	2.10	9.16

Note: The number in the first row tell TASSEL, the number of individuals, followed by number of traits, and 1 indicate number of header rows. -999 is commonly used for missing data.

- Save the phenotype data matrix in Text (Tab delimited) type with a suitable filename, <traitname.txt>.

(c) Population structure data

Population structure data (Q matrix) uses following format:



	A	B	C	D	E	F
1	40	5	1			
2		Q1	Q2	Q3	Q4	Q5
3	GENO1	0.000	0.003	0.037	0.003	0.956
4	GENO2	0.000	0.003	0.006	0.016	0.975
5	GENO3	0.000	0.001	0.001	0.001	0.996
6	GENO4	0.000	0.004	0.005	0.002	0.989
7	GENO5	0.000	0.001	0.001	0.001	0.996
8	GENO6	0.000	0.001	0.002	0.001	0.996
9	GENO7	0.001	0.003	0.629	0.004	0.362
10	GENO8	0.000	0.002	0.001	0.001	0.995
11	GENO9	0.000	0.021	0.004	0.125	0.850
12	GENO10	0.001	0.002	0.002	0.002	0.993

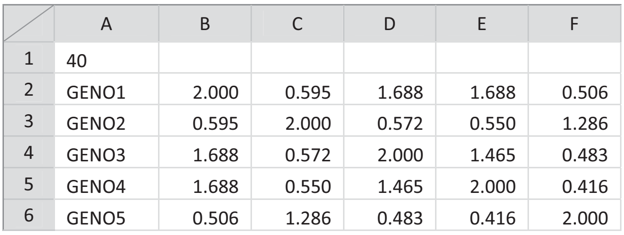
Note: The number in the first row tell TASSEL, the number of individuals, followed by number of sub-populations (K=5), and 1 indicate number of header rows.

- Save the Q matrix in Text (Tab delimited) type with a suitable filename, <Q_matrix name.txt>.

(d) Kinship data (K matrix)

Kinship data is an optional requirement for Association mapping. Structured association analysis is done using a general linear model (GLM) algorithm, which does not require K matrix. K matrix is however, essential for mixed linear model (MLM) analysis.

If the kinship output from SPAGeDi is used, it should be formatted to read in TASSEL as given below. For this, (i) add a value of "2" for relative kinship between same individuals and (ii) change the all negative values of relative kinship into "0".



	A	B	C	D	E	F
1	40					
2	GENO1	2.000	0.595	1.688	1.688	0.506
3	GENO2	0.595	2.000	0.572	0.550	1.286
4	GENO3	1.688	0.572	2.000	1.465	0.483
5	GENO4	1.688	0.550	1.465	2.000	0.416
6	GENO5	0.506	1.286	0.483	0.416	2.000

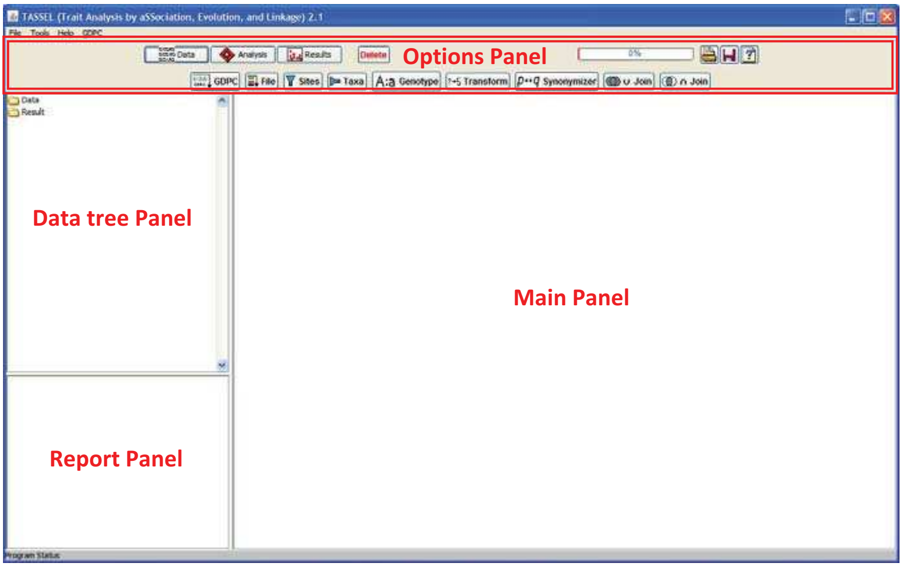
Note: The number in the first row tell TASSEL, the number of individuals. No missing data are permitted in K matrix.

- Save the K matrix in Text (Tab delimited) type with a suitable filename <kinship.txt>

B. Running Structured Association Mapping

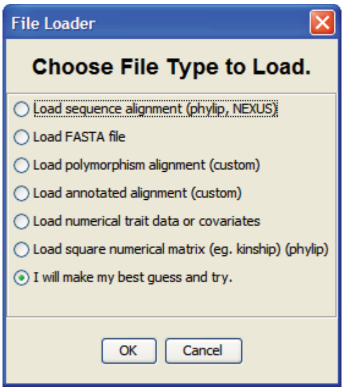
(i) Loading data

- Double click on the file sTASSEL.jar in the TASSEL 2-1 directory. Following window opens.



- Click on the **[Data]** button from Options panel.

- From the buttons below click on **[File]** button



File Loader

Choose File Type to Load.

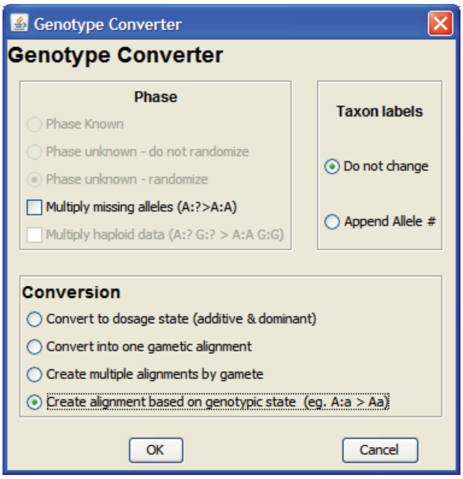
○ Load sequence alignment (phylip, NEXUS)
○ Load FASTA file
○ Load polymorphism alignment (custom)
○ Load annotated alignment (custom)
○ Load numerical trait data or covariates
○ Load square numerical matrix (eg. kinship) (phylip)
● I will make my best guess and try.

- Select the data type to load, if not sure about data, it is better to choose “I will make my best guess and try”, this will allow TASSEL to select the data type by itself.
- Click **[OK]** and select the file containing marker name <markername.txt>.
- Repeat the step, and load phenotype data <traitname.txt> and Q-matrix <Q_matrixname.txt> one after the other.
- Once the data are loaded, they appear in the Data tree panel.

(ii) Geneotype data processing

When diploid microsatellite data is used, convert the raw format to genetype state, to do this,

- Click on the button [**A:a Genotype]** to start Genetype Converter



Genotype Converter

Genotype Converter

Phase

○ Phase Known
○ Phase unknown - do not randomize
● Phase unknown - randomize
□ Multiply missing alleles (A:?>A:A)
□ Multiply haploid data (A:? G:? > A:A G:G)

Taxon labels

● Do not change
○ Append Allele #

Conversion

○ Convert to dosage state (additive & dominant)
○ Convert into one gametic alignment
○ Create multiple alignments by gamete
● Create alignment based on genotypic state (eg. A:a > Aa)

- Select **Create alignment based on genotypic state (eg. A:a > Aa)** and click **[OK]**
- This will add another dataset named “GenoStates” in the Data tree panel

(iii) Joining marker, phenotype and population structure data

- In the Data tree Panel, select data "GenoStates", "<Traitname>" and "<Q_Matrixname>" by clicking on them, while <Ctrl> key is pressed

- Click on the button **[U Join]** in the Options Panel
- A new Data set named “GenoStates+<Traitname>+<Q_matrixname>” appear on the data tree panel

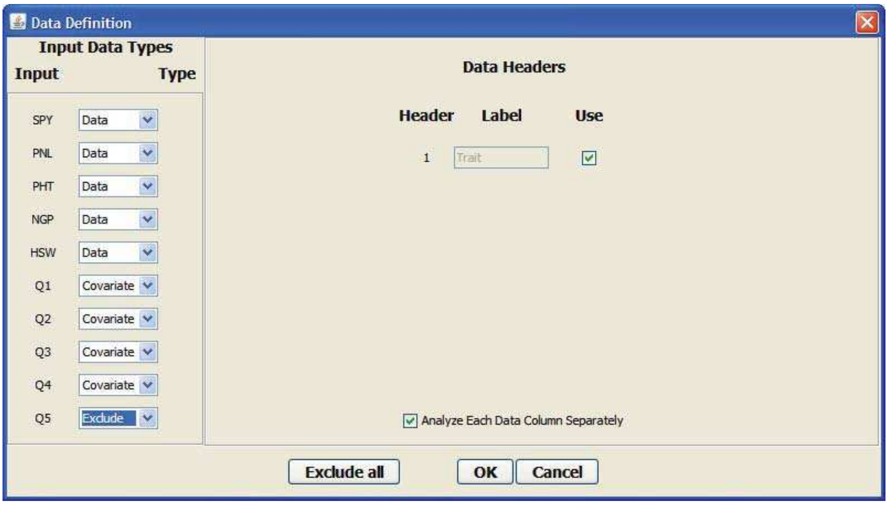
(iv) Loading relative kinship data (for MLM analysis only)

- If kinship information is available, load it by clicking **[File]** button on the Options Panel.
- Select either "Load square numerical matrix (eg. kinship) (phylip)" or “I will make my best guess and try” and click **[OK]**
- Select Kinship file and Open
- The kinship data appear under Matrix in the Data tree panel
- Alternately Kinship can be calculated within TASSEL, by selecting the “GenoStates” and clicking on **[Analysis]** and then **[Kinship]**.

  Note: This is a simple kinship matrix generated from the distance matrix. In order to use more robust Kinship estimates it is recommended to use SPAGeDi or SAS.

(v) Structured association analysis using least squares GLM

- Select “GenoStates+<Traitname>+<Q_matrixname>” from the Data tree panel, by clicking on it while holding the <Ctrl> key pressed
- Click on the **[Analysis]** button and then click on **[GLM]**



Input Data Types

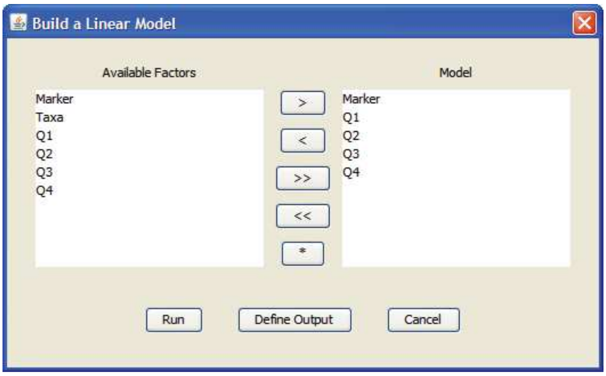
Input	Type
SPY	Data
PNL	Data
PHT	Data
NGP	Data
HSW	Data
Q1	Covariate
Q2	Covariate
Q3	Covariate
Q4	Covariate
Q5	Exclude

Data Headers

Header	Label	Use
1	Trait	☑

☑ Analyze Each Data Column Separately

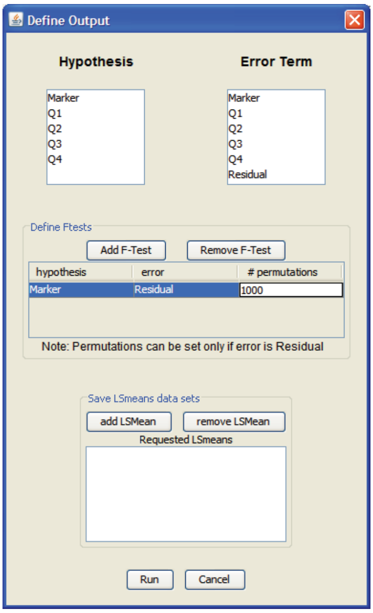
- Select all phenotype as data, and Sub-population data as covariate, Exclude the last sub-population
- Check “Analyse Each Data Column Separately”
- Click **[OK]**



Available Factors: Marker, Taxa, Q1, Q2, Q3, Q4

Model: Marker, Q1, Q2, Q3, Q4

- Click on **[Define Output]**



Hypothesis: Marker, Q1, Q2, Q3, Q4

Error Term: Marker, Q1, Q2, Q3, Q4, Residual

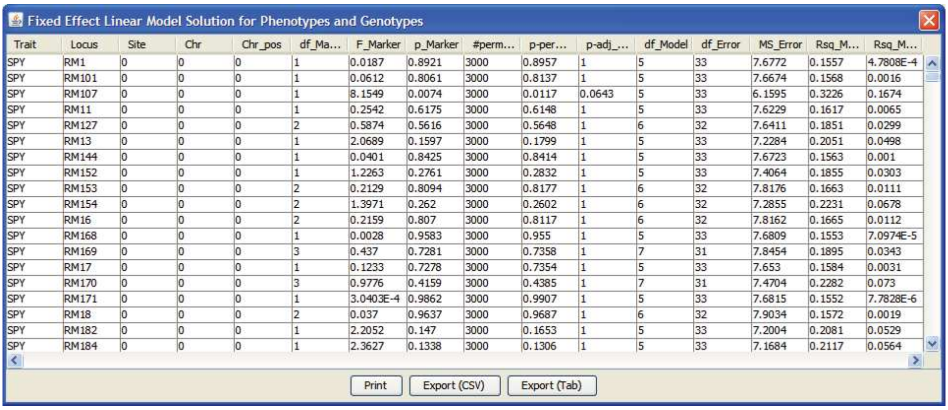
hypothesis	error	# permutations
Marker	Residual	1000

Note: Permutations can be set only if error is Residual

- In Define Ftests, we can set number of permutations, say 1000. Click **[Run]**
- A new data set appear under Result>Association in the Data tree Panel named “GLM_ GenoStates+<Traitname>+<Q_matrixname>”

(vi) Viewing and saving results

- Click on the button [Results] from Options Panel
- Select the result data from Data tree Panel, by holding the <Ctrl> key pressed and clicking on “GLM_ GenoStates+<Traitname>+<Q_matrixname>”
- Click [Table] button from the Options Panel



Fixed Effect Linear Model Solution for Phenotypes and Genotypes

Trait	Locus	Site	Chr	Chr_pos	df_Ma...	F_Marker	p_Marker	#perm...	p-per...	p-adj_...	df_Model	df_Error	MS_Error	Rsq_M...	Rsq_M...
SPY	RM1	0	0	0	1	0.0187	0.8921	3000	0.8957	1	5	33	7.6772	0.1557	4.7808E-4
SPY	RM101	0	0	0	1	0.0612	0.8061	3000	0.8137	1	5	33	7.6674	0.1568	0.0016
SPY	RM107	0	0	0	1	8.1549	0.0074	3000	0.0117	0.0643	5	33	6.1595	0.3226	0.1674
SPY	RM11	0	0	0	1	0.2542	0.6175	3000	0.6148	1	5	33	7.6229	0.1617	0.0065
SPY	RM127	0	0	0	2	0.5874	0.5616	3000	0.5648	1	6	32	7.6411	0.1851	0.0299
SPY	RM13	0	0	0	1	2.0689	0.1597	3000	0.1799	1	5	33	7.2284	0.2051	0.0498
SPY	RM144	0	0	0	1	0.0401	0.8425	3000	0.8414	1	5	33	7.6723	0.1563	0.001
SPY	RM152	0	0	0	1	1.2263	0.2761	3000	0.2832	1	5	33	7.4064	0.1855	0.0303
SPY	RM153	0	0	0	2	0.2129	0.8094	3000	0.8177	1	6	32	7.8176	0.1663	0.0111
SPY	RM154	0	0	0	2	1.3971	0.262	3000	0.2602	1	6	32	7.2855	0.2231	0.0678
SPY	RM16	0	0	0	2	0.2159	0.807	3000	0.8117	1	6	32	7.8162	0.1665	0.0112
SPY	RM168	0	0	0	1	0.0028	0.9583	3000	0.955	1	5	33	7.6809	0.1553	7.0974E-5
SPY	RM169	0	0	0	3	0.437	0.7281	3000	0.7358	1	7	31	7.8454	0.1895	0.0343
SPY	RM17	0	0	0	1	0.1233	0.7278	3000	0.7354	1	5	33	7.653	0.1584	0.0031
SPY	RM170	0	0	0	3	0.9776	0.4159	3000	0.4385	1	7	31	7.4704	0.2282	0.073
SPY	RM171	0	0	0	1	3.0403E-4	0.9862	3000	0.9907	1	5	33	7.6815	0.1552	7.7828E-6
SPY	RM18	0	0	0	2	0.037	0.9637	3000	0.9687	1	6	32	7.9034	0.1572	0.0019
SPY	RM182	0	0	0	1	2.2052	0.147	3000	0.1653	1	5	33	7.2004	0.2081	0.0529
SPY	RM184	0	0	0	1	2.3627	0.1338	3000	0.1306	1	5	33	7.1684	0.2117	0.0564

- By clicking on the **[Print]** results can now be printed, or exported to Tab delimited text file or Comma separated values (CSV) text file by clicking on buttons **[Export (CSV)]** and **[Export (Tab)]** respectively.

(vii) Understanding the result file

Trait	Locus	Site	Chr	Chr_ pos	df_ Marker	F_ Marker	p_ Marker	#perm_ Marker	p-perm_ Marker	p-adj_ Marker	df_ Model	df_ Error	MS_ Error	Rsq_ Model	Rsq_ Marker
SPY	RM1	0	0	0	1	0.0187	0.8921	3000	0.8957	1	5	33	7.6772	0.1557	4.78E-04
SPY	RM101	0	0	0	1	0.0612	0.8061	3000	0.8137	1	5	33	7.6674	0.1568	0.0016
SPY	RM107	0	0	0	1	8.1549	0.0074	3000	0.0117	0.0643	5	33	6.1595	0.3226	0.1674
SPY	RM11	0	0	0	1	0.2542	0.6175	3000	0.6148	1	5	33	7.6229	0.1617	0.0065
SPY	RM127	0	0	0	2	0.5874	0.5616	3000	0.5648	1	6	32	7.6411	0.1851	0.0299
SPY	RM13	0	0	0	1	2.0689	0.1597	3000	0.1799	1	5	33	7.2284	0.2051	0.0498
SPY	RM144	0	0	0	1	0.0401	0.8425	3000	0.8414	1	5	33	7.6723	0.1563	0.001
SPY	RM152	0	0	0	1	1.2263	0.2761	3000	0.2832	1	5	33	7.4064	0.1855	0.0303
SPY	RM153	0	0	0	2	0.2129	0.8094	3000	0.8177	1	6	32	7.8176	0.1663	0.0111
SPY	RM154	0	0	0	2	1.3971	0.262	3000	0.2602	1	6	32	7.2855	0.2231	0.0678
SPY	RM16	0	0	0	2	0.2159	0.807	3000	0.8117	1	6	32	7.8162	0.1665	0.0112

The result file, in addition to displaying the F-statistics and p-values for the requested F-tests, also contains information about degrees of freedom, the error mean square for the model, R-square of the model, and Rsquare for the marker. The model R-square is the portion of total variation explained by the full model. The marker R-square is the portion of total variation explained by the marker but not by the other terms in the model. When permutations are requested, #perm_Marker is the number of permutations run, pperm_ Marker is a test of individual markers, and p-adj_Marker is the marker p-value adjusted for multiple tests. The p-adj_Marker value is a permutation test derived using a step-down MinP procedure (Ge et al. 2003) and controls the family-wise error rate (FWER). For example, if only markers with p-adj values of .05 or less are accepted as significant, then the probability of rejecting a single true null hypothesis across the entire set of hypotheses is held to .05 or less. This test takes dependence between hypotheses into account and does not assume that hypotheses are independent as do other multiple test correction procedures.

Note:

Both STRUCTURE and TASSEL comes with well written tutorials. This document is no substitution for those. For any clarification and in depth information please read these tutorials carefully. Besides, there are online discussion forums available for these software packages, in which users post their doubts and suggestions. These discussions are watched by the developers of these software and they incorporate modifications/ fix bugs as and when required.

To join these forums visit following sites,

STRUCTURE: https://groups.google.com/d/forum/structure-software

TASSEL : http://groups.google.com/d/forum/tassel

Major References :

Buckler E, Casstevens T, Bradbury P, Zhang Z (2009) Trait Analysis by aSSociation, Evolution and Linkage (TASSEL): User Manual. Cornell University

http://www.maizegenetics.net/tassel/docs/TASSEL_help.pdf

Pritchard JK, Wena X, Falush D (2010) Documentation for structure software: Version 2.3. Department of Human Genetics, University of Chicago.

http://pritch.bsd.uchicago.edu/structure_software/release_versions/v2.3.3/structure_doc.pdf

Other references:

Bradbury PJ, Zhang Z , Kroon DE , Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23: 2633-2635

Churchill G, Airey DC, Allayee H, Angel JM, Attie AD et al. (2004) The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nature Genet 36: 1133-1137

Hardy OJ, Vekemans X (2002) SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol Ecol Notes 2: 618-620

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Human Genet 67: 170-181.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. Genetics, 155:945–959

Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D et al. (2001) Dwarf8 polymorphisms associate with variation in flowering time. Nature Genet 28: 286-289.

Yu JM, Pressoir G, Briggs WH, Bi IV, Yamasaki M et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genet 38:203-208