Genome-wide Association Study (GWAS) in TASSEL (GUI)


avikarn.com/2019-07-22-GWAS

TASSEL aslo known as `Trait Analysis by aSSociation, Evolution and Linkage` is a powerful statistical software to conduct `association mapping` such as `General Linear Model (GLM)` and `Mixed Linear Model (MLM)` . The GUI (graphical user interface) version of TASSEL is very well built for anyone who does not have a background or experience in working in `command line` . In this tutorial, I will show how to prepare `input` files and run assoication analysis in TASSEL. For detailed information on TASSEL, user's guide and further documentation please visit: `https://www.maizegenetics.net/tassel`

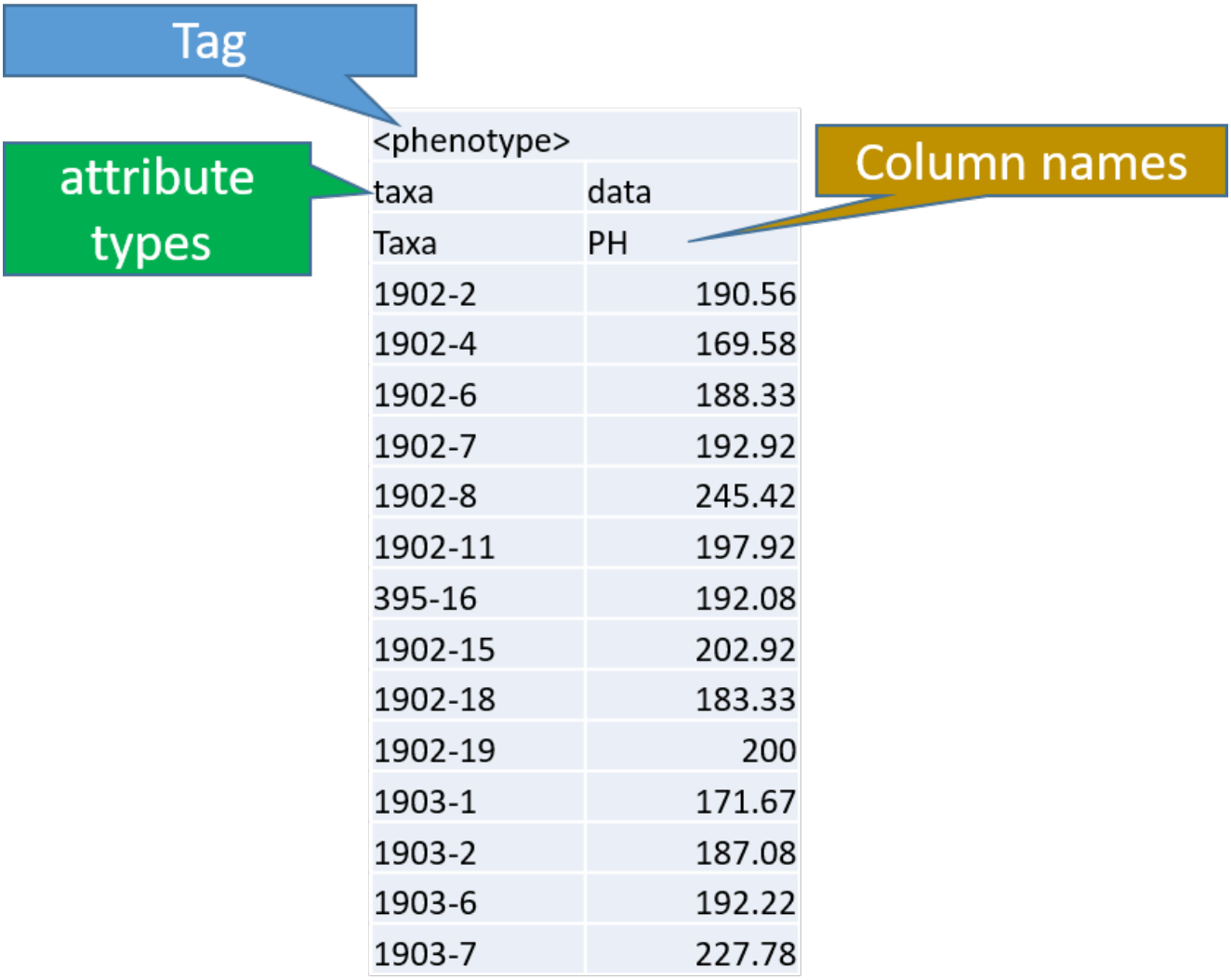
1.1 Download and install TASSEL software

Download and install the latest version of the **TASSEL software** at this link: `https://www.maizegenetics.net/tassel`

1.2: Preparing the Input files

Phenotype file

Prepare the phenotype file as shown below in the figure, and please remember if your data has `covariates` such as `sex` , `age` or `treatment` , then, please categories them with header name `factor` .



<phenotype>	
taxa	data
Taxa	PH
1902-2	190.56
1902-4	169.58
1902-6	188.33
1902-7	192.92
1902-8	245.42
1902-11	197.92
395-16	192.08
1902-15	202.92
1902-18	183.33
1902-19	200
1903-1	171.67
1903-2	187.08
1903-6	192.22
1903-7	227.78

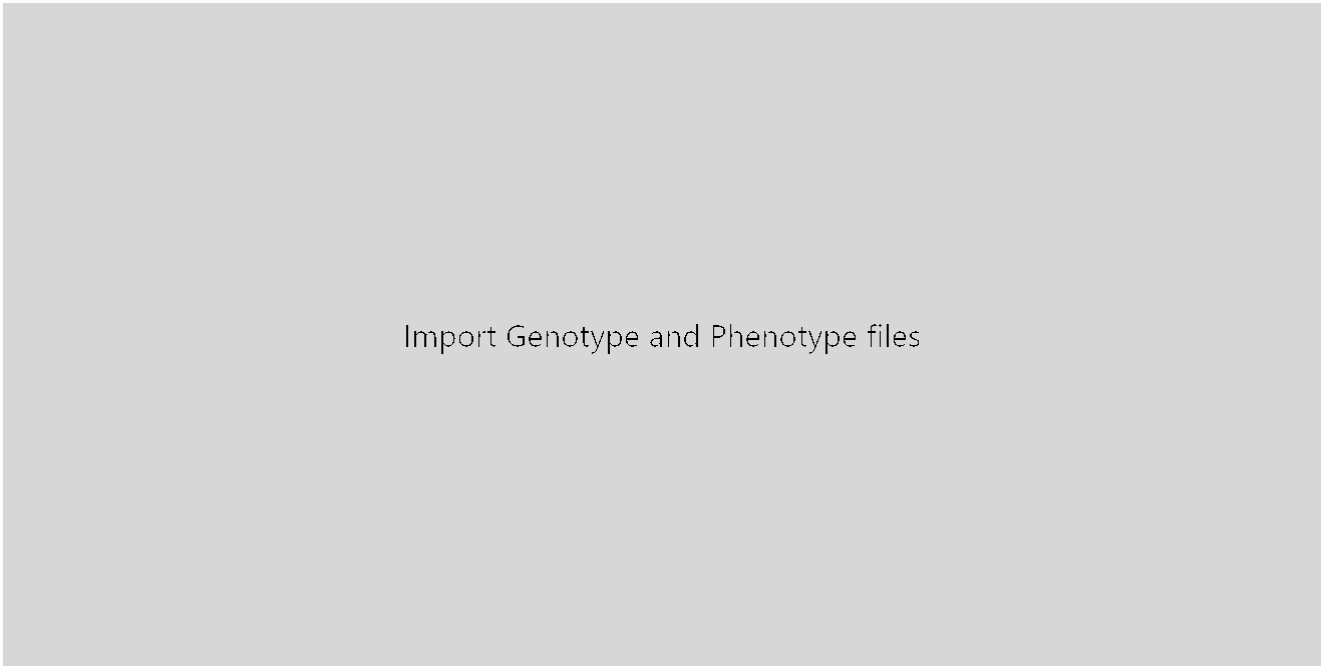
Genotype file

TASSEL allows various genotype file formats such as `VCF` (variant call format), `.hmp.txt` , and `plink` . In this tutorial, I am using the `hmp.txt` version of the genotype file. The below githe screenshot of the hmp.txt genotype file.

rs#	alleles	chrom	pos	strand	assembly#	center	protLSID	assayLSID	panelLSID	QCcode	374-19	374-18	374-15
S1_366383	A/T/G	1	366383	+	NA	NA	NA	NA	NA	NA	A	A	A	A	A	A	A	W	A	A	A	A	W	W
S1_366392	T/G	1	366392	+	NA	NA	NA	NA	NA	NA	T	T	T	T	T	T	T	T	T	T	T	T	T	T
S1_366424	T/C	1	366424	+	NA	NA	NA	NA	NA	NA	Y	Y	Y	Y	Y	Y	Y	T	Y	Y	Y	C	Y	T	Y	T
S1_366426	A/T	1	366426	+	NA	NA	NA	NA	NA	NA	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
S1_366438	C/T	1	366438	+	NA	NA	NA	NA	NA	NA	Y	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
S1_366440	T/C	1	366440	+	NA	NA	NA	NA	NA	NA	T	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	T	Y	Y	Y	Y
S1_374909	T/-	1	374909	+	NA	NA	NA	NA	NA	NA	N	T	T	T	T	N	T	T	T	T	0	N	T	T	T	0
S1_374910	A/T	1	374910	+	NA	NA	NA	NA	NA	NA	N	A	A	A	A	N	A	A	A	A	A	N	A	A	A	W
S1_374913	A/-/T	1	374913	+	NA	NA	NA	NA	NA	NA	N	A	A	A	A	A	N	A	A	A	A	A	N	W	A	A
S1_374916	G/T	1	374916	+	NA	NA	NA	NA	NA	NA	N	G	G	G	G	N	G	G	G	G	G	N	G	G	G	K
S1_374918	C/A	1	374918	+	NA	NA	NA	NA	NA	NA	N	C	C	C	C	N	C	C	C	C	C	N	C	C	C	C
S1_374919	T/C	1	374919	+	NA	NA	NA	NA	NA	NA	N	T	Y	T	T	N	T	T	T	T	T	N	T	T	T	T
S1_374927	C/-	1	374927	+	NA	NA	NA	NA	NA	NA	N	C	C	C	C	N	C	C	C	C	C	N	C	C	C	C
S1_374931	A/G	1	374931	+	NA	NA	NA	NA	NA	NA	N	R	A	R	A	N	A	A	R	A	A	N	R	A	A	R
S1_374933	A/G	1	374933	+	NA	NA	NA	NA	NA	NA	N	A	A	A	A	N	A	A	A	R	A	N	A	A	A	A
S1_374938	G/A/C	1	374938	+	NA	NA	NA	NA	NA	NA	N	G	G	G	G	N	G	G	G	G	G	N	G	G	G	G
S1_374942	A/T	1	374942	+	NA	NA	NA	NA	NA	NA	N	W	A	A	W	N	A	A	A	A	A	N	A	T	A	A
S1_374949	A/G	1	374949	+	NA	NA	NA	NA	NA	NA	N	A	A	A	A	N	R	G	A	A	A	N	A	A	A	A
S1_374953	T/C/A/G	1	374953	+	NA	NA	NA	NA	NA	NA	N	T	T	T	T	N	T	T	Y	T	T	N	T	T	C
S1_374954	T/C	1	374954	+	NA	NA	NA	NA	NA	NA	N	Y	Y	Y	T	N	T	T	T	T	T	N	Y	T	T	T
S1_374955	A/G/-/C	1	374955	+	NA	NA	NA	NA	NA	NA	N	A	A	A	A	N	A	A	A	A	A	N	A	A	A	A
S1_374956	C/G	1	374956	+	NA	NA	NA	NA	NA	NA	N	C	C	C	C	N	C	C	C	C	C	N	C	C	C	C
S1_374960	G/A	1	374960	+	NA	NA	NA	NA	NA	NA	N	G	G	G	G	N	G	G	G	G	G	N	G	G	G	G
S1_374962	A/C	1	374962	+	NA	NA	NA	NA	NA	NA	N	A	A	A	A	N	A	A	A	A	A	N	A	A	A	A
S1_1033975	A/G	1	1033975	+	NA	NA	NA	NA	NA	NA	N	A	A	A	A	A	A	N	A	A	A	N	A	A	A	A
S1_1034009	T/A	1	1034009	+	NA	NA	NA	NA	NA	NA	N	T	T	T	T	T	T	N	T	W	T	N	T	T	A	T
S1_1565068	T/G	1	1565068	+	NA	NA	NA	NA	NA	NA	T	T	T	T	T	N	T	T	T	T	T	T	T	T	T	N
S1_1762067	T/C/G	1	1762067	+	NA	NA	NA	NA	NA	NA	NA	T	T	T	C	T	T	T	N	T	Y	K	N	T	T	C

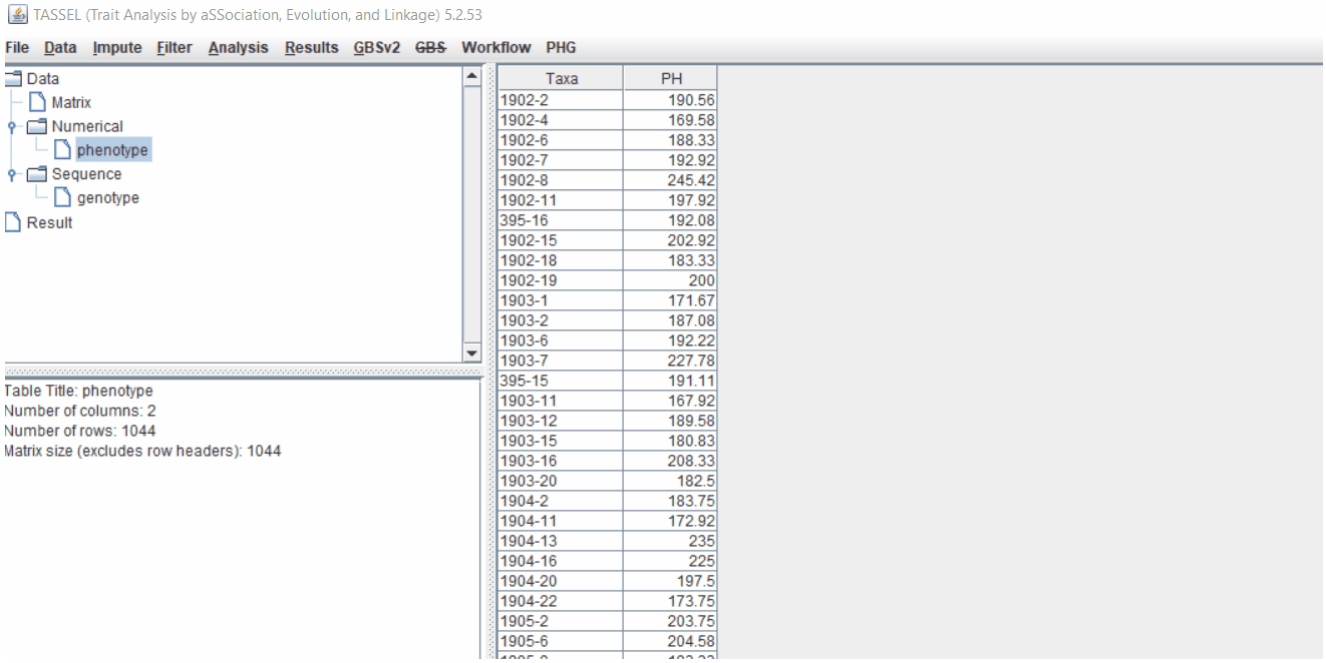
Step 1.2: Importing phenotype and genotype files

Import the files by following the steps shown below. **Tip!** Both files can be opened at same time holding `CTRL` and clicking the file names.



1.3 Phenotype distribution plot

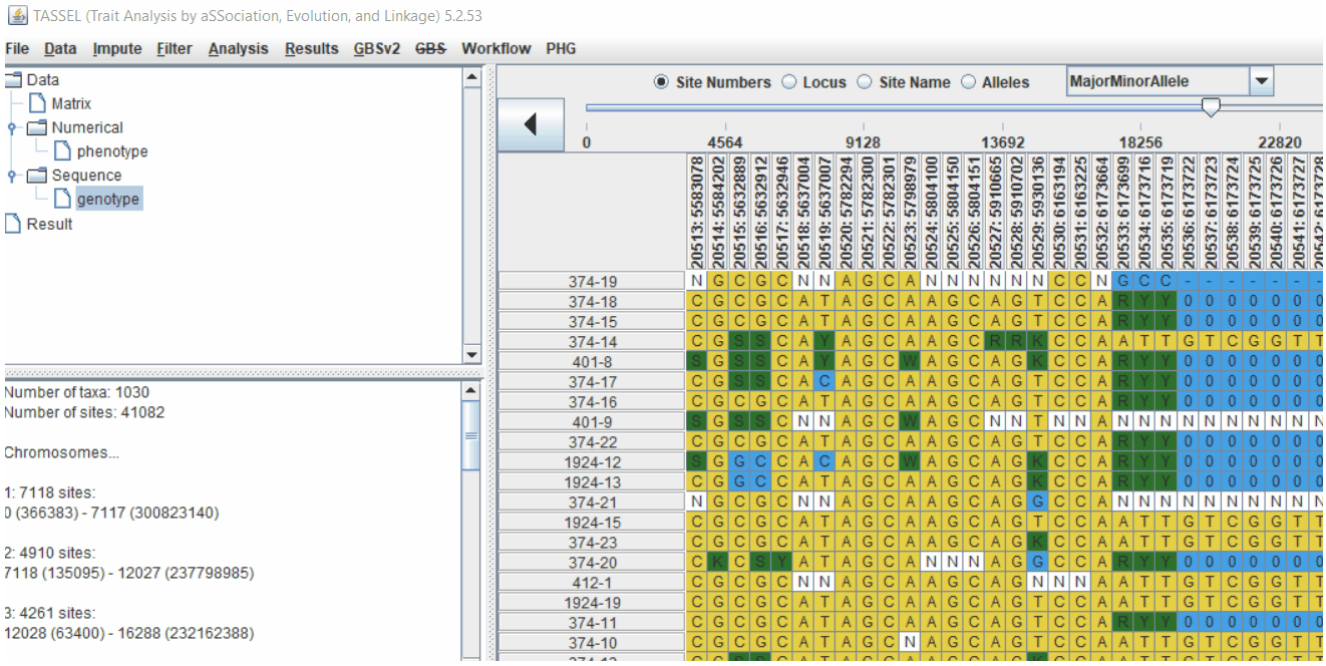
It is always a wise idea to look at the phenotype distribution by plotting to check for any outliers. Follow below steps to plot histogram of your phenotype data.



1.4 Genotype summary analysis

Next crucial step is to look at the genotype data by simply following the steps shown. Couple of keys things to look at are:

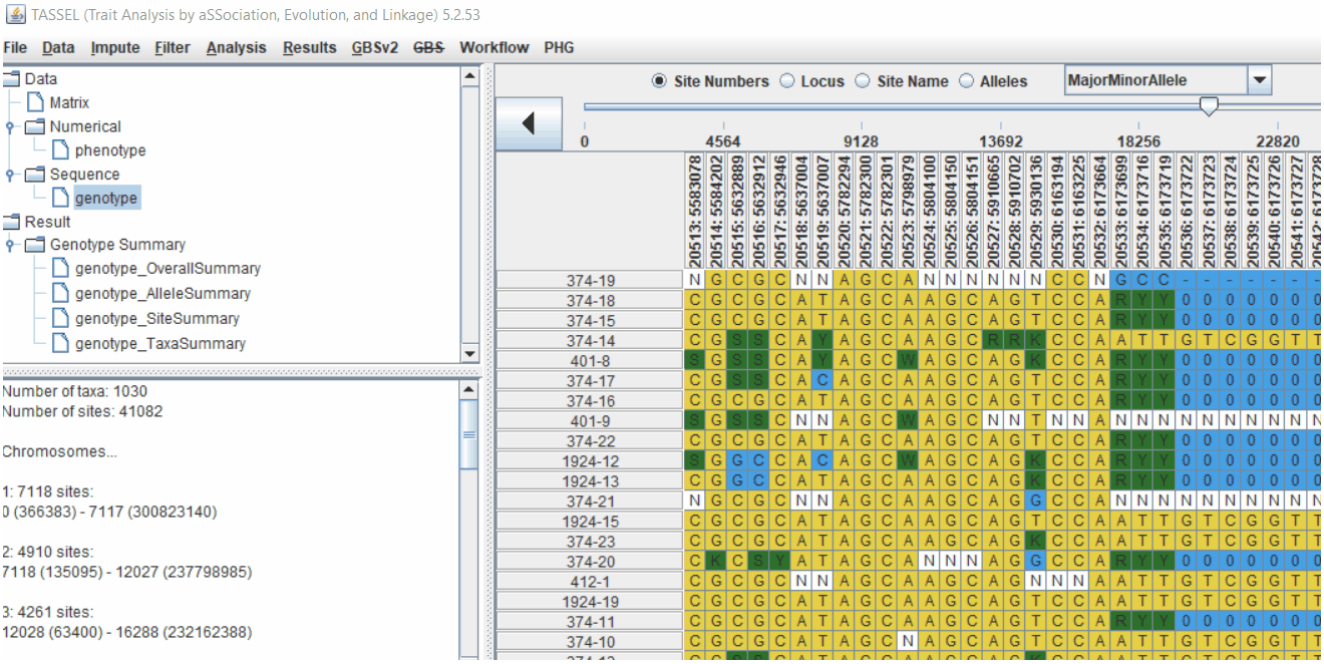
1. Minor allele frequency distribution
2. Missing genotypic data to see if it requires to be imputed
3. Proportion of heterozygous in the samples to check for self-ed samples



2.0 Conduct GWAS analysis

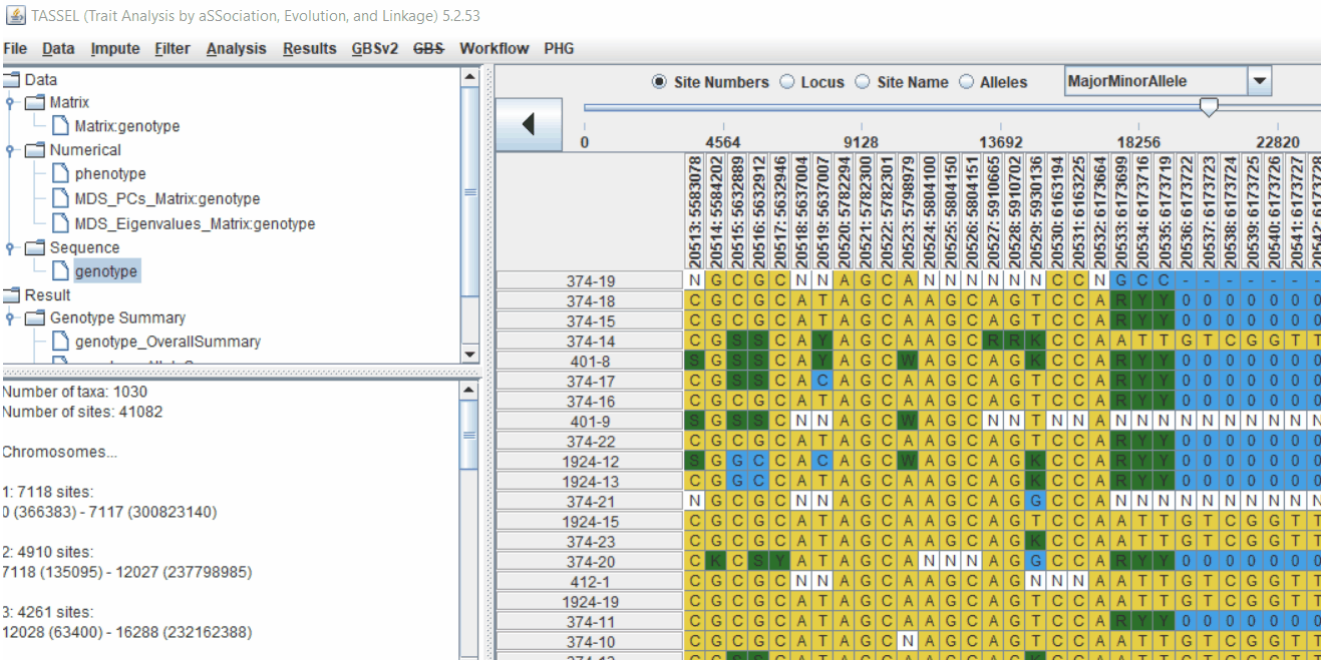
2.1 multidimensional scaling (MDS)

MDS output can be used as the covariate in the GLM or MLM to correct for population structure. Please follow the steps shown below:



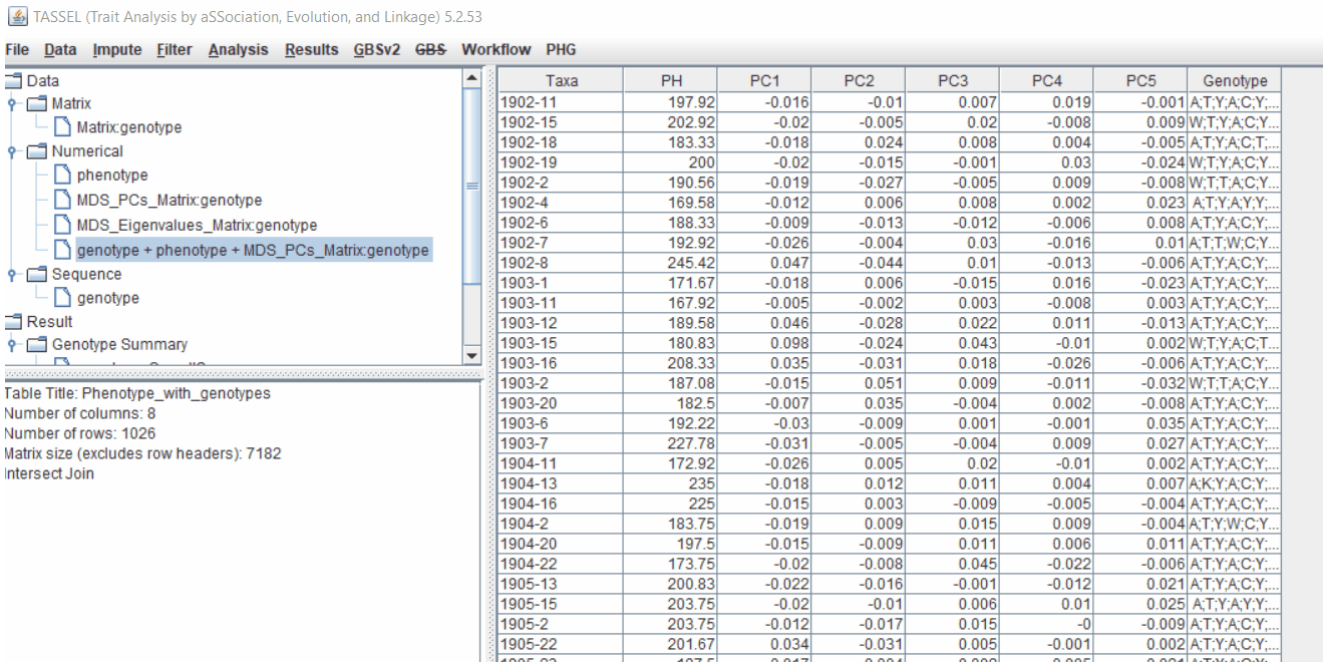
2.2 Intersecting the files

Intersect the `genotype`, `phenotype` and `MDS` files by following the steps below:



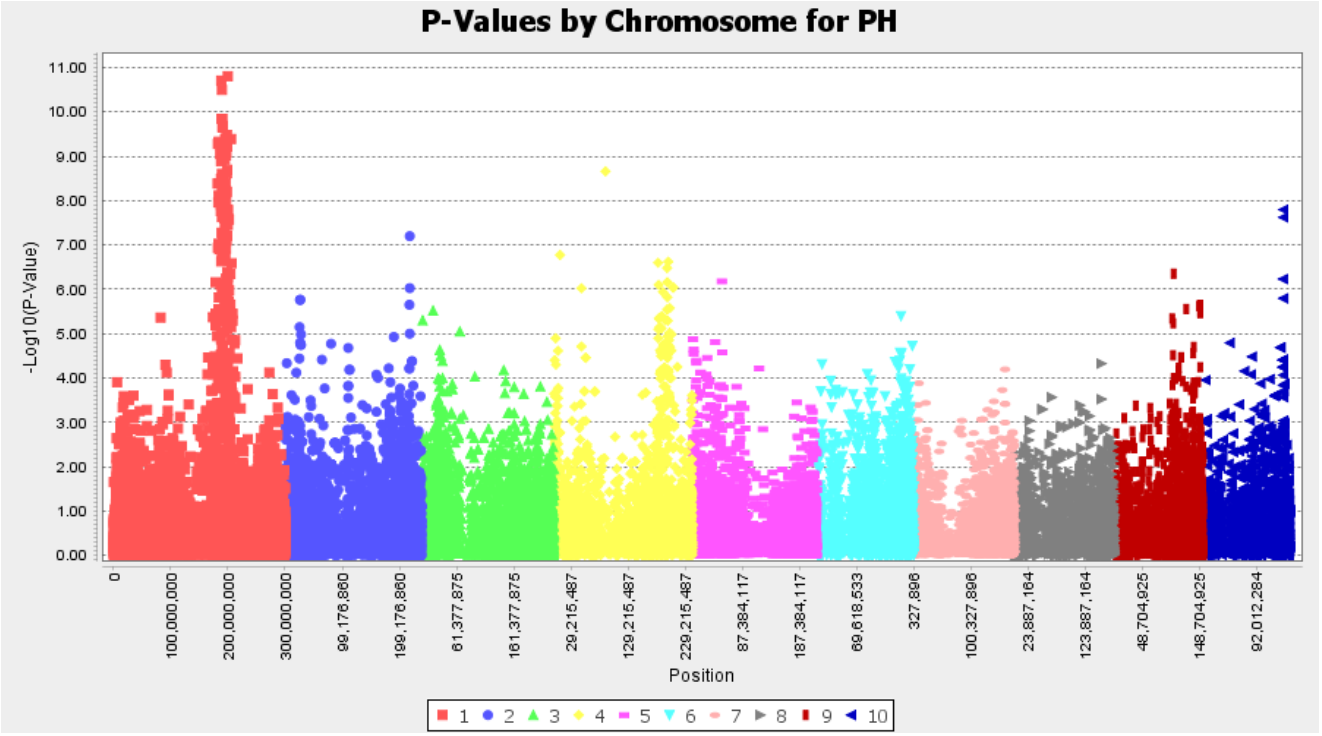
3.0 running General Linear Model (GLM)

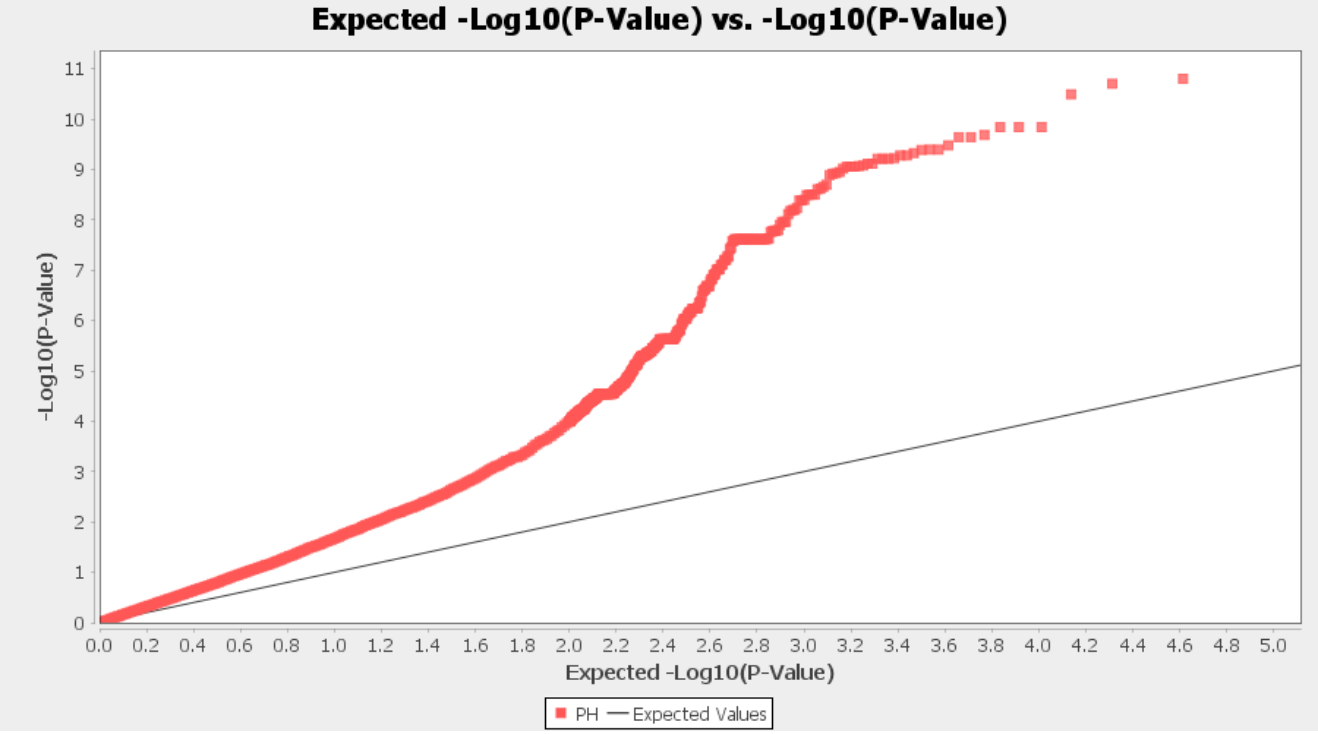
Run the GLM analysis by selecting the `intersected` files following the steps below:



Taxa	PH	PC1	PC2	PC3	PC4	PC5	Genotype
1902-11	197.92	-0.016	-0.01	0.007	0.019	-0.001	A;T;Y;A;C;Y;...
1902-15	202.92	-0.02	-0.005	0.02	-0.008	0.009	W;T;Y;A;C;Y...
1902-18	183.33	-0.018	0.024	0.008	0.004	-0.005	A;T;Y;A;C;T;...
1902-19	200	-0.02	-0.015	-0.001	0.03	-0.024	W;T;Y;A;C;Y...
1902-2	190.56	-0.019	-0.027	-0.005	0.009	-0.008	W;T;T;A;C;Y...
1902-4	169.58	-0.012	0.006	0.008	0.002	0.023	A;T;Y;A;Y;Y;...
1902-6	188.33	-0.009	-0.013	-0.012	-0.006	0.008	A;T;Y;A;C;Y;...
1902-7	192.92	-0.026	-0.004	0.03	-0.016	0.01	A;T;T;W;C;Y...
1902-8	245.42	0.047	-0.044	0.01	-0.013	-0.006	A;T;Y;A;C;Y;...
1903-1	171.67	-0.018	0.006	-0.015	0.016	-0.023	A;T;Y;A;C;Y;...
1903-11	167.92	-0.005	-0.002	0.003	-0.008	0.003	A;T;Y;A;C;Y;...
1903-12	189.58	0.046	-0.028	0.022	0.011	-0.013	A;T;Y;A;C;Y;...
1903-15	180.83	0.098	-0.024	0.043	-0.01	0.002	W;T;Y;A;C;T...
1903-16	208.33	0.035	-0.031	0.018	-0.026	-0.006	A;T;Y;A;C;Y;...
1903-2	187.08	-0.015	0.051	0.009	-0.011	-0.032	W;T;T;A;C;Y...
1903-20	182.5	-0.007	0.035	-0.004	0.002	-0.008	A;T;Y;A;C;Y;...
1903-6	192.22	-0.03	-0.009	0.001	-0.001	0.035	A;T;Y;A;C;Y;...
1903-7	227.78	-0.031	-0.005	-0.004	0.009	0.027	A;T;Y;A;C;Y;...
1904-11	172.92	-0.026	0.005	0.02	-0.01	0.002	A;T;Y;A;C;Y;...
1904-13	235	-0.018	0.012	0.011	0.004	0.007	A;K;Y;A;C;Y;...
1904-16	225	-0.015	0.003	-0.009	-0.005	-0.004	A;T;Y;A;C;Y;...
1904-2	183.75	-0.019	0.009	0.015	0.009	-0.004	A;T;Y;W;C;Y...
1904-20	197.5	-0.015	-0.009	0.011	0.006	0.011	A;T;Y;A;C;Y;...
1904-22	173.75	-0.02	-0.008	0.045	-0.022	-0.006	A;T;Y;A;C;Y;...
1905-13	200.83	-0.022	-0.016	-0.001	-0.012	0.021	A;T;Y;A;C;Y;...
1905-15	203.75	-0.02	-0.01	0.006	0.01	0.025	A;T;Y;A;Y;Y;...
1905-2	203.75	-0.012	-0.017	0.015	-0	-0.009	A;T;Y;A;C;Y;...
1905-22	201.67	0.034	-0.031	0.005	-0.001	0.002	A;T;Y;A;C;Y;...

The output of the GLM analyis is produced ubder the `Result` node. The GLM association test can be evaluated by plotting `Q-Q plot` and the `Manhattan` plot as shown below.

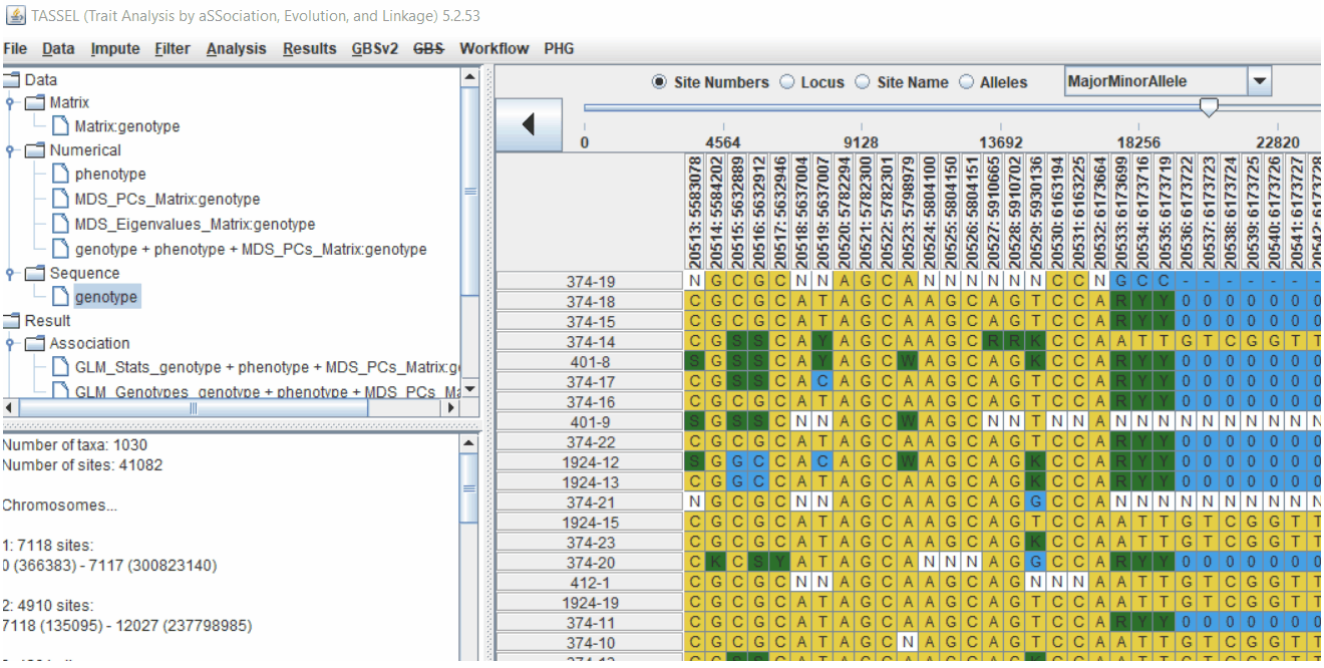




From the above Q-Q plot, we can see that are several markers that appear to be falsely associated with the trait, therefore, to control this confounding effect, use `Kinship` matrix

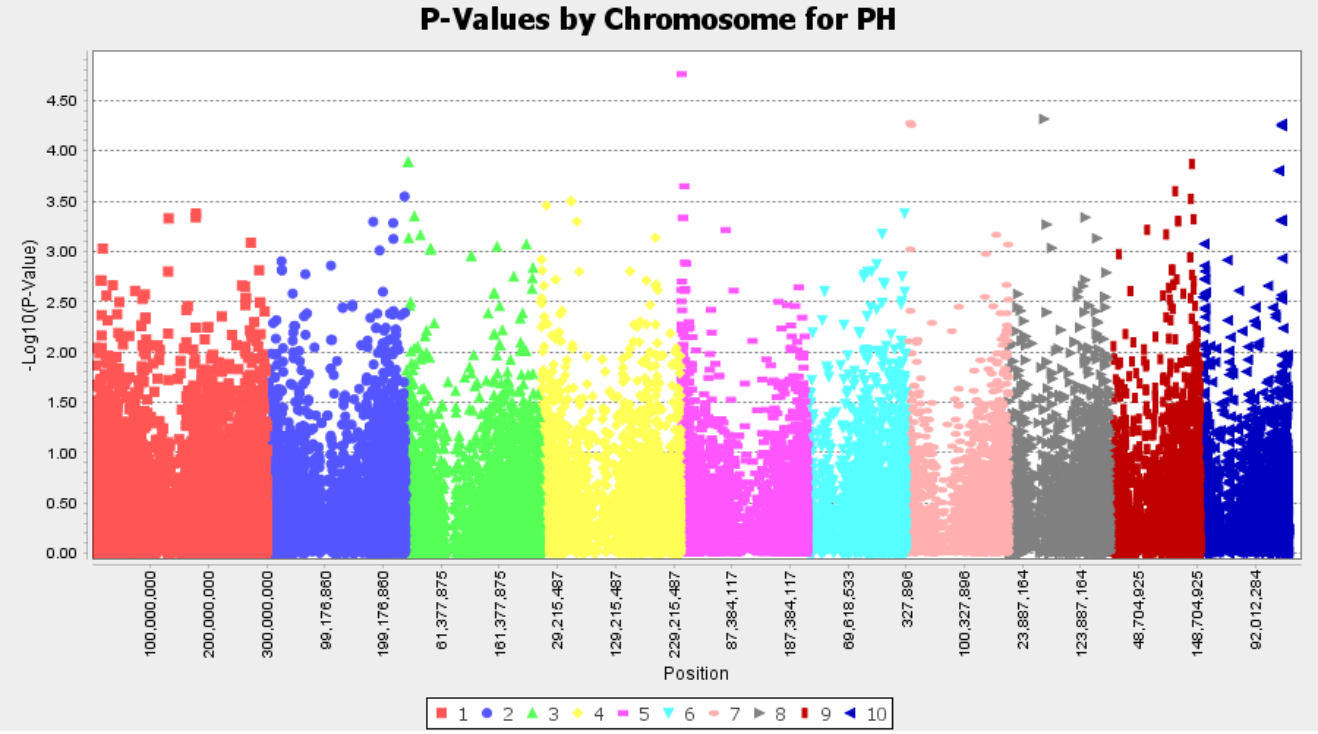
4.0 Calculating Kinship matrix

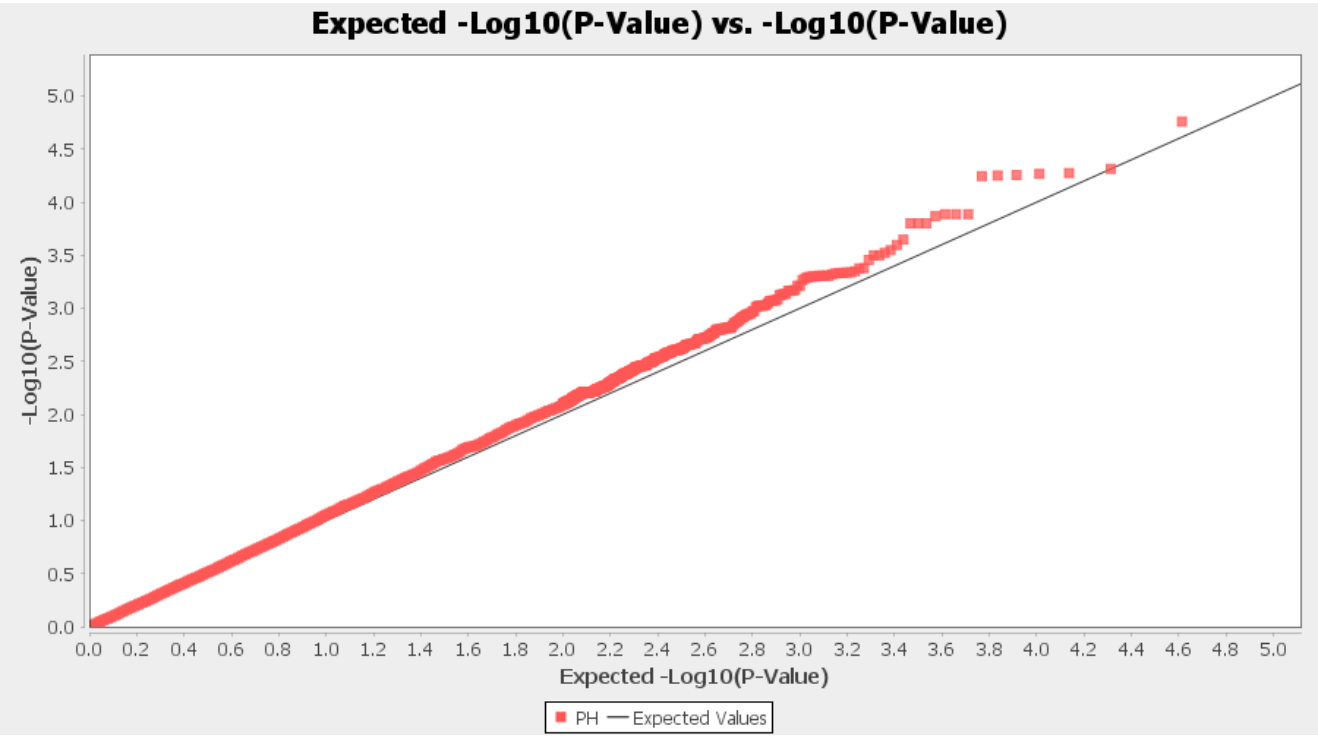
Follow the below steps to calcuate the kinship matrix.



4.1 running Mixed Linear Model (MLM)

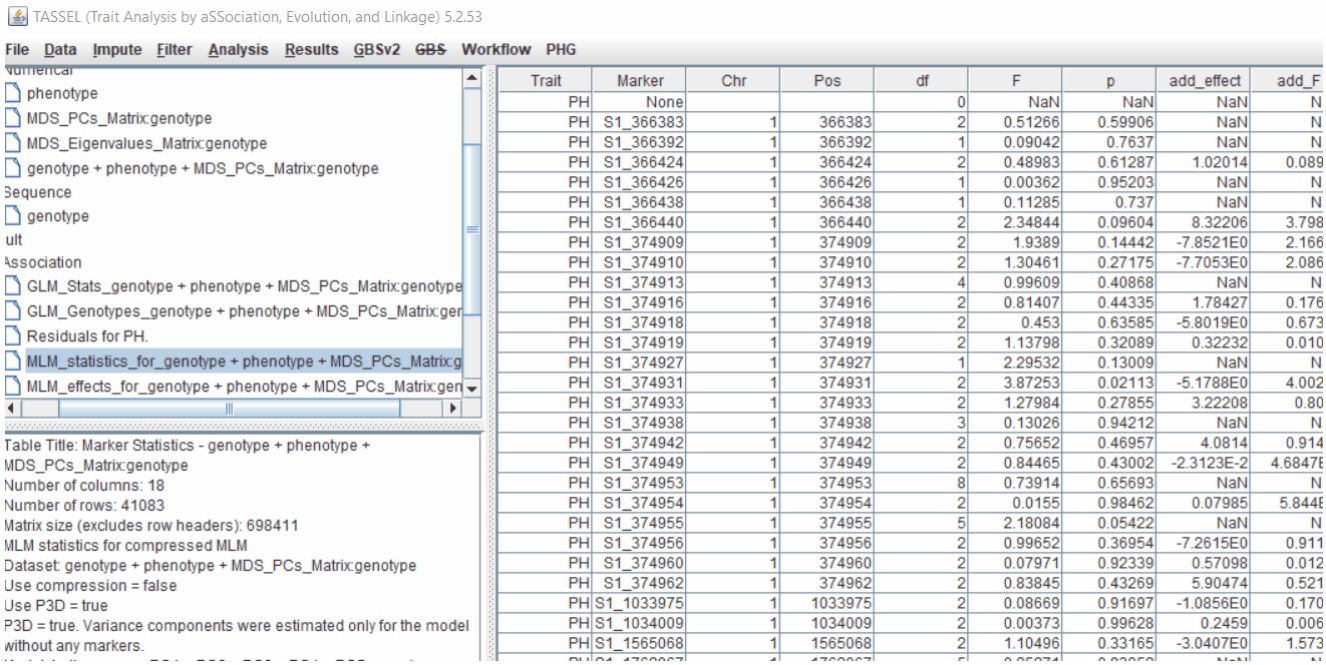
Once the Kinship matrix has been calculated, MLM can be now be conducted by below:





4.2 Exporting results

One may export the results in .txt format by the following the below steps:



Trait	Marker	Chr	Pos	df	F	p	add_effect	add_F
PH	None			0	NaN	NaN	NaN	N
PH	S1_366383	1	366383	2	0.51266	0.59906	NaN	N
PH	S1_366392	1	366392	1	0.09042	0.7637	NaN	N
PH	S1_366424	1	366424	2	0.48983	0.61287	1.02014	0.089
PH	S1_366426	1	366426	1	0.00362	0.95203	NaN	N
PH	S1_366438	1	366438	1	0.11285	0.737	NaN	N
PH	S1_366440	1	366440	2	2.34844	0.09604	8.32206	3.798
PH	S1_374909	1	374909	2	1.9389	0.14442	-7.8521E0	2.166
PH	S1_374910	1	374910	2	1.30461	0.27175	-7.7053E0	2.086
PH	S1_374913	1	374913	4	0.99609	0.40868	NaN	N
PH	S1_374916	1	374916	2	0.81407	0.44335	1.78427	0.176
PH	S1_374918	1	374918	2	0.453	0.63585	-5.8019E0	0.673
PH	S1_374919	1	374919	2	1.13798	0.32089	0.32232	0.010
PH	S1_374927	1	374927	1	2.29532	0.13009	NaN	N
PH	S1_374931	1	374931	2	3.87253	0.02113	-5.1788E0	4.002
PH	S1_374933	1	374933	2	1.27984	0.27855	3.22208	0.80
PH	S1_374938	1	374938	3	0.13026	0.94212	NaN	N
PH	S1_374942	1	374942	2	0.75652	0.46957	4.0814	0.914
PH	S1_374949	1	374949	2	0.84465	0.43002	-2.3123E-2	4.6847E
PH	S1_374953	1	374953	8	0.73914	0.65693	NaN	N
PH	S1_374954	1	374954	2	0.0155	0.98462	0.07985	5.844E
PH	S1_374955	1	374955	5	2.18084	0.05422	NaN	N
PH	S1_374956	1	374956	2	0.99652	0.36954	-7.2615E0	0.911
PH	S1_374960	1	374960	2	0.07971	0.92339	0.57098	0.012
PH	S1_374962	1	374962	2	0.83845	0.43269	5.90474	0.521
PH	S1_1033975	1	1033975	2	0.08669	0.91697	-1.0856E0	0.170
PH	S1_1034009	1	1034009	2	0.00373	0.99628	0.2459	0.006
PH	S1_1565068	1	1565068	2	1.10496	0.33165	-3.0407E0	1.573

4.3 Significance Threshold

Bonferroni threshold can be deterimined to identify significantly markers associated with the trait by using the below equation:

$P \leq 1/N \ (\alpha = 0.05)$

where, N is the total number of markers tested in association analysis) was used to identify the most significantly markers associated with the trait. Similarly, another way is to perform `FDR` (`False Discovey Rate`) correction method.

--- End of Tutorial ---

Thank you for reading this tutorial. If you have any questions or comments, please let me know in the comment section below or send me an email.

Bibliography

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.