

# TASSEL Pipeline: Analyzing Diversity Data

Terry Casstevens<sup>1</sup>, Peter Bradbury<sup>2,3</sup>, Dallas Kroon<sup>1</sup>, Zhiwu Zhang<sup>1</sup>, Edward Buckler<sup>1,2,4</sup>

<sup>1</sup> Institute for Genomic Diversity, Cornell University, Ithaca, NY <sup>2</sup> USDA-ARS <sup>3</sup> Cornell Theory Center, Cornell University <sup>4</sup> Dept. of Plant Breeding and Genetics, Cornell University

<http://www.maizegenetics.net/tassel>

## Introduction

TASSEL version 2.1 has been redesigned to allow command line batch analysis of diversity data. The various analysis functions are now built as individual modules (plugins). These plugins are used both in the graphical user interface and the new TASSEL pipeline. The plugins are used in the pipeline to create custom batch analysis jobs. Basically, data goes through the pipeline being acted upon by each plugin along the way. Also, more complicated logic can be designed into the pipeline if needed. Analysis plugins include functions, such as calculations of linkage disequilibrium, plots of linkage disequilibrium, association analysis using mixed linear model (MLM) and general linear model (GLM) algorithms, loading files, genotype transforms, and exporting results to files. Conditional plugins include functions, such as combining data sets, filtering, pass through, and setting analysis parameters. Once a pipeline has been designed and implemented, many data manipulations can be accomplished without user interaction. One pipeline that we developed can perform GLM and MLM analysis using flat file input (i.e. SNP data, Trait data, Population Structure, and Kinship Matrix) and producing tab delimited output of the results. TASSEL is an open source project.

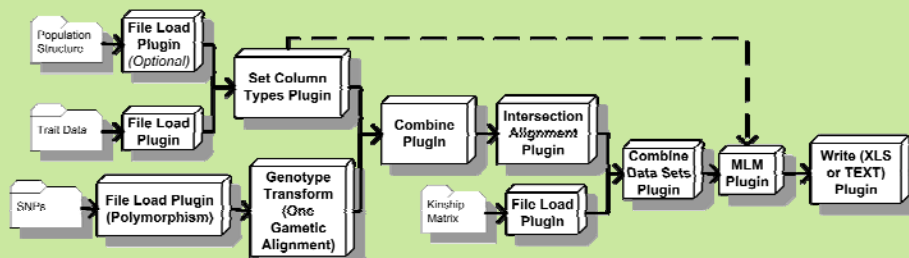
[www.maizegenetics.net/tassel](http://www.maizegenetics.net/tassel)  
[sourceforge.net/projects/tassel](http://sourceforge.net/projects/tassel)

## Available Plugins

[www.maizegenetics.net/tassel/docs/TasselPipeline.pdf](http://www.maizegenetics.net/tassel/docs/TasselPipeline.pdf)

- CombineDataSetsPlugin
- CombineSimpleTableReportsPlugin
- CreateTreePlugin
- FileLoadPlugin
- FilterAlignmentPlugin
- FilterDataSetPlugin
- FilterTaxaAlignmentPlugin
- GLMPlugin
- GenotypeTransformPlugin
- Grid2dDisplayPlugin
- IntersectionAlignmentPlugin
- KinshipPlugin
- LinkageDiseqDisplayPlugin
- LinkageDisequilibriumPlugin
- LogisticRegressionAssocPlugin
- MLMPlugin
- NumericalGenotypePlugin
- PassThroughPlugin
- SetColumnTypesPlugin
- SequenceDiversityPlugin
- StepwisePlugin
- StepwiseAnalysisPlugin
- SynonymizerPlugin
- TableDisplayPlugin
- TreeDisplayPlugin
- UnionAlignmentPlugin
- WriteXLSPlugin
- chart.ChartDisplayPlugin
- gdpc.ConvertGDPCToPALPlugin
- gdpc.GDPCPlugin
- numericaltransform.NumericalTransformPlugin
- pedigree.PedigreePlugin
- snpassay.ExtractSNPassaysPlugin

## Example Pipeline: MLM and GLM Analysis



Usage: `java net.maizegenetics.baseplugins.test.MLMGLMFileInputPipeline -t <trait file> [-s <SNP file> | -p <Poly file>] [-q <population file>] -o <output file> [-glm] [-mlm -k <kinship file>] [-xls | -txt] [-abc true | false] [-mni <iterations>] [-mim true | false] [-fim true | false]`

```
-t <trait file>      : trait data
-s <SNP file>       : SNP data (Loads as sequence alignment)
-p <Poly file>      : SNP data (Loads as polymorphism alignment)
                    Must specify either -s or -p
-q <population file>: optional population data
-k <kinship file>   : kinship data needed for mlm analysis

-o <output file>    : output file
-xls | -txt         : optionally specify output format
                    MS Excel or text (default: MS Excel)

-glm                : optionally specifies to run glm analysis
-mlm                : optionally specifies to run mlm analysis

-abc true | false  : optionally specifies whether to Analyze by Column
                    (default: false)
-mni <iterations> : optionally specifies Maximum Number of Iterations (only for MLM, default: 200)
-mim true | false  : optionally specifies whether Markers used in Model (only for GLM, default: true)
-fim true | false  : optionally specifies whether to use full model (reduced model if false)
                    (only for MLM, default: false)
```

## SNP Data

	91	7							
	s1297	s1411	s1666	s1786	s2245	s2276	s2356		
1	G	A	A	A	C	A	G		
2	38-11	G	A	A	C	A	G		
3	A272	G	A	A	C	A	G		
4	A441-5	G	A	A	G	C	A	G	
5	A554	G	A	A	G	C	A	G	
6	A6	G	A	A	C	A	G		
7	A619	G	G	A	A	G	C		
8	A632	G	G	G	C	A	G		
9	B103	G	A	A	G	C	A	G	
10	B104	G	G	G	C	A	G		
11	...	...	...	...	...	...	...		
12	...	...	...	...	...	...	...		

## Example Usages

```
java -classpath "%CP%" -Xms128m -Xmx512m
net.maizegenetics.baseplugins.test.MLMGLMFileInputPipeline -t
"three_traits_rn.txt" -p "d8coding_sites.txt" -q
"popStructure_taxa286_rn.txt" -o "output_poly.txt" -glm -txt
```

```
java -classpath "%CP%" -Xms128m -Xmx512m
net.maizegenetics.baseplugins.test.MLMGLMFileInputPipeline -t
"three_traits_rn.txt" -p "d8coding_sites.txt" -k
"kinship_277taxa_by_Spagedi_m_sq.txt" -q
"popStructure_taxa286_rn.txt" -o "output_poly.xls" -mlm -xls
```

## Kinship Matrix

	277								
	33-16	2	0.1816	0.0187	0	0.1536	...		
1	33-16	2	0.1816	0.0187	0	0.1536	...		
2	4226	0.01873	0	2	0	0	...		
3	4722	0	0.112	0	2	0	...		
4	A188	0.15357	0	0	0	2	...		
5	A214N	0	0	0	0	0.0231	...		
6	A239	0.01664	0	0	0	0.0076	...		
7	A272	0	0	0	0	0	...		
8	A441-5	0.18202	0	0	0	0	...		
9	A554	0.17447	0.0263	0.0618	0.0523	0	...		
10	...	...	...	...	...	...	...		
11	...	...	...	...	...	...	...		
12	...	...	...	...	...	...	...		

## Population Structure

	286	2	2						
		Q1	Q2						
		STIFF	STALK	NONSTIFF					
1	33-16	0.014	0.972						
2	38-11	0.003	0.993						
3	4226	0.071	0.917						
4	4722	0.035	0.854						
5	A188	0.013	0.982						
6	A214N	0.762	0.017						
7	A239	0.035	0.963						
8	A272	0.019	0.122						
9	...	...	...						
10	...	...	...						
11	...	...	...						
12	...	...	...						

## Trait Data

	301	3	1						
	81-1	EarHT	dpoll	EarDia					
1	81-1	59.5	-999	-999					
2	33-16	64.75	64.5	-999					
3	38-11	92.25	68.5	37.897					
4	4226	65.5	59.5	32.2193					
5	4722	81.13	71.5	32.421					
6	A188	27.5	62	31.419					
7	A214N	65	69	32.006					
8	A239	47.88	61	36.064					
9	A272	35.63	70	-999					
10	...	...	...	...					
11	...	...	...	...					
12	...	...	...	...					

## MLM Results

	Trait	Locus	Site	df	f	p	d_Model	d_Error	MS_Error	Rsq_model	Rsq_marker
1	dpoll	s2276	0	1	11.96	9.05E-04	3	74	11.5121	0.7105	0.0468
2	dpoll	s2356	0	1	11.96	9.05E-04	3	74	11.5121	0.7105	0.0468
3	EarHT	s2276	0	1	9.052	0.0036	3	75	251.9316	0.6067	0.0475
4	EarHT	s2356	0	1	9.052	0.0036	3	75	251.9316	0.6067	0.0475
5	dpoll	s2245	0	1	7.652	0.0072	3	74	12.1196	0.6952	0.0315
6	dpoll	s1411	0	1	7.406	0.0081	3	74	12.1562	0.6943	0.0306
7	EarHT	s2245	0	1	3.711	0.0578	3	75	269.0274	0.58	0.0208
8	dpoll	s1786	0	1	2.339	0.1305	3	74	12.9632	0.674	0.0103
9	EarDia	s2276	0	1	2.201	0.1429	3	64	9.0601	0.4815	0.0178

## GLM Results

	Trait	Locus	Site	Chr	Chr_pos	df	f	p	d_Model	d_Error	MS_Error	Rsq_Model	Rsq_Marker
1	dpoll	s2276	0	0	0	1	10	0.0019	3	78	23.1327	0.4503	0.0727
2	dpoll	s2356	0	0	0	1	10	0.0019	3	78	23.1327	0.4503	0.0727
3	EarHT	s2276	0	0	0	1	9.4	0.003	3	78	524.4422	0.2309	0.0926
4	EarHT	s2356	0	0	0	1	9.4	0.003	3	78	524.4422	0.2309	0.0926
5	dpoll	s2245	0	0	0	1	9.3	0.0031	3	78	23.3968	0.4441	0.0664
6	dpoll	s1411	0	0	0	1	7.5	0.0076	3	78	23.8924	0.4323	0.0547
7	EarHT	s2245	0	0	0	1	5.4	0.0222	3	78	549.2233	0.1946	0.0563
8	EarHT	s1411	0	0	0	1	2.2	0.1401	3	78	571.3066	0.1622	0.0239
9	EarDia	s1297	0	0	0	1	2.2	0.145	3	67	17.3664	0.0566	0.0305

Thanks to NSF and USDA-ARS for their support.

January 5, 2009