TASSEL 3.0 / 4.0 Pipeline Command Line Interface: *Guide to using Tassel Pipeline*

Terry Casstevens (*tmc46@cornell.edu*)

Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853-2703

October 11, 2012

Prerequisites

- Java SDK 6.0 or later (http://java.sun.com/javase/downloads/index.jsp).
- Tassel Standalone Build (http://www.maizegenetics.net/tassel/tassel3.0_standalone.zip)
- Tassel Standalone Build (http://www.maizegenetics.net/tassel/tassel4.0_standalone.zip)

Source Code

- https://tassel.svn.sourceforge.net/svnroot/tassel/maizegenetics

Install

Unzip the Tassel Standalone Build onto your file system. Change into the root directory: `tassel3.0_standalone or tassel4.0_standalone`.

Execute

On Windows, use `run_pipeline.bat` to execute the pipeline.

In UNIX, use `run_pipeline.pl` to execute the pipeline. If you are using a Bash Shell on Windows, you may need to change the following line to use a ; instead of a :.

```
my $CP = join(":", @fl);
```

To launch the Tassel GUI that automatically executes a pipeline, use `start_tassel.bat` or `start_tassel.pl` instead of `run_pipeline.bat` or `run_pipeline.pl` respectively.

These scripts have a $top variable that can be changed to the absolute path of your installation. That way, you can execute them any directory.

Increasing Heap Size

To modify the initial or maximum heap size available to the Tassel Pipeline, either edit `run_pipeline.pl` or specify values via the command line.

```
./run_pipeline.pl –Xms512m -Xmx10g -fork1…
```

Examples

```
./run_pipeline.pl -fork1 -h chr1_5000sites.txt -ld -ldd png -o
chr1_5000sites_ld.png -runfork1
```

```
./run_pipeline.pl -fork1 -h chr1_5000sites.txt -ld -ldd png -o
chr1_5000sites_ld.png -runfork1
```

```
./run_pipeline.pl -fork1 … -fork2 … -combine3 -input1 -input2 … -fork4 -
<flag> -input3 -runfork1 -runfork2
```

Examples *(XML Configuration Files)*

This command runs the Tassel Pipeline according to the specified configuration file... Configuration files are standard XML notation. The tags are the same as the below documented flags although no beginning dash is used. See the `example_pipelines` directory for some common XML configurations.

```
./run_pipeline.pl -configFile config.xml
```

This command creates the XML configuration file from the original command line flags. Simply insert the -createXML and filename at the beginning. Only the XML is created. It does not run the pipeline...

```
./run_pipeline.pl -createXML config.xml -fork1 ...
```

This command translates the specified XML configuration file back into the original command line flags... It does not run the pipeline...

```
./run_pipeline.pl -translateXML config.xml
```

Usage

Pipeline Controls	
-fork<id>	This flag identifies the start of a pipeline segment that should be executed sequentially. <id> can be numbers or characters (no spaces). No space between -fork and <id> either. Other flags can reference the <id>.
-runfork<id>	This flag identifies a pipeline segment to execute. This will usually be the last argument. This explicitly executes the identified pipeline segment. This should not be used to execute pipeline segments that receive input from other pipeline segments. Those will start automatically when it receives the input.
-input<id>	This specifies a pipeline segment as input to the plugin prior this flag. That plugin must be in the current pipeline segment. Multiple of these can be specified after plugins that accept multiple inputs. ./run_pipeline.pl -fork1 -h genotype.hmp.txt -fork2 -r phenotype.txt -combine3 -input1 -input2 -intersect -runfork1 -runfork2 ./run_pipeline.pl -fork1 -h genotype.hmp.txt -fork2 -includeTaxaInFile taxaList1.txt -input1 -export file1 -fork3 -includeTaxaInFile taxaList2.txt -input1 -export file2 -runfork1
-inputOnce<id>	This specifies a pipeline segment as a one-time input to a CombineDataSetsPlugin. As such, this flag should follow a CombineDataSetsPlugin in the current pipeline segment. After the CombineDataSetsPlugin has received data from this input, it will use it for every iteration. Whereas CombineDataSetsPlugin waits for data specified by -input each iteration. Multiple of these can be specified.
-combine<id>	This flag starts a new pipeline segment with a CombineDataSetsPlugin at the beginning. The CombineDataSetsPlugin is used to combine data sets from multiple pipeline segments. Follow this flag with -input<id> and/or -inputOnce<id> flags to specify which pipeline segments should be combined.
Data	
	If the filename to be imported begins with "http", it will be treated as an URL.
-t <trait file>	Loads trait file as numerical data.

-s <SNP file>	Loads SNP file as sequence alignment.
-p <Poly file>	Loads polymorphism file as polymorphism alignment.
-a <Anno file>	Loads annotated alignment file.
-r <phenotype file>	Loads file as Table Report. This is meant to replace numerical, polymorphism, and annotated alignment.
-k <kinship file>	Loads kinship file as square matrix.
-q <population structure file>	Loads population structure file as numerical data.
-h <hapmap file>	Loads hapmap file (.hmp.txt or .hmp.txt.gz)
-b <BLOB file>	Loads BLOB file (.zip). Only Tassel 3.
-g <BLOB file>	Loads BLOB file (.gz). Only Tassel 3.
-plink -ped <ped filename> -map <map filename>	Loads Plink format given ped and map files.
-flapjack -geno <geno file> -map <map file>	Loads Flapjack format given geno and map files.
-fasta <filename>	Loads FASTA file.
-geneticMap <filename>	Loads Genetic Map.
-maxAllelesToRetain <number>	Sets the preference maximum number of alleles to retain. Initial default is 2. Notice this is a preference setting and remains true on a given machine for all import operations that follow, even in subsequent executions of the Tassel Pipeline or Tassel GUI.
-retainRareAlleles true | false	Sets the preference whether to retain rare alleles. Notice this is a preference setting and remains true on a given machine for all import operations that follow, even in subsequent executions of the Tassel Pipeline or Tassel GUI.
-optimizeForTaxa	This should follow a file import, which instructs the loader to optimize the data for taxa based operations. By default, loaded data is optimized for site operations.
-convertToSiteOpt	Converts input Alignment to a Site Optimized Alignment. May return same Alignment if already optimized for Sites. This is mainly for testing purposes.
-convertToTaxaOpt	Converts input Alignment to a Taxa Optimized Alignment. May return same Alignment if already optimized for Taxa. This is mainly for testing purposes.
-taxaJoinStrict true | false	If true, taxa names are only considered the same if they match exactly. If false, taxa names match if all specified levels match. For example… B73 matches B73:XXX matches

	B73:XXX:YYY. But B73:XXX does not match B73:YYY.
-union	This joins (union) input datasets based taxa. This should follow a -combine specification.
-intersect	This joins (intersect) input datasets based taxa. This should follow a -combine specification.
-separate	This separates an input into its components if possible. For example, alignments separated by chromosome (locus).
-export <filename1,filename2,…>	Exports input dataset to specified filename(s). If no -exportType follows this parameter, the exported format will be determined by the type of input (i.e. Alignments will default to Hapmap format). Exportable datasets, other that Alignment, only have one format option. Therefore, there is no need to specify -exportType. Specify none, one, or multiple filenames matching the number of input data sets. If no filenames, the files will be named the same as the input data sets. If only one specified for multiple data sets, a count starting with 1 will be added to each resulting file. If multiple filenames (separated with commas but no spaces), there should be one for each input. When exporting Hapmap files, if the extension is .hmp.txt.gz, the file will be gzipped.
-exportType <type>	Defines format that previously specified -export should use. Type can be Hapmap, HapmapDiploid, Plink, zipBLOB (only Tassel 3), gzipBLOB (only Tassel 3), Flapjack, Phylip_Seq, Phylip_Inter, Text.
-impute	Imputes Genotypic Data.
-imputeMethod <method>	Specifies the impute method to use. Method can be Length (default), MajorAllele, SimilarWindow, or IBDProb. This should follow the -impute flag.
-imputeMinLength <num>	Specifies the minimum length for the impute Length method (default value: 30). This should follow the -impute flag.
-imputeMaxMismatch <num>	Specifies the maximum mismatch for the impute Length method (default value: 1). This should follow the -impute flag.
-imputeMinProb <num>	Specifies the minimum probability for the impute IBDProb method (default value: 0.001). This should follow the -impute flag.
-filterAlign	Filters an alignment by sites.
-filterAlignMinCount	Specifies the minimum count (default: 1) for the

<num>	previously specified -filterAlign.
-filterAlignMinFreq <num>	Specifies the minimum frequency (default: 0.01) for the previously specified -filterAlign.
-filterAlignMaxFreq	Specifies the maximum frequency (default 1.0) for the previously specified -filterAlign.
-filterAlignStart <num>	Specifies the starting site index (default value: 0) for the previously specified -filterAlign.
-filterAlignEnd <num>	Specifies the end site index (default value: last site in alignment) for the previously specified -filterAlign.
-filterAlignLocus <name>	Specifies the to be used with the starting and ending physical positions if defined. Defaults to first Locus in the Alignment.
-filterAlignStartPos <num>	Specifies the starting physical position (default is first site) for the previously specified -filterAlign.
-filterAlignEndPos <num>	Specifies the end physical position (default is last site) for the previously specified -filterAlign.
-filterAlignExtInd	Indicates that the last specified -filterAlign should extract indels. This is not done by default.
-filterAlignRemMinor	Indicates that the last specified -filterAlign should remove minor SNP states. This is not done by default.
-filterAlignSliding	Indicates that the last specified -filterAlign should use sliding windows. This in not done by default.
-filterAlignHapLen <num>	Specifies the haplotype length (default value: 3) if using sliding windows.
-filterAlignStepLen <num>	Specifies the step length (default value: 3) if using sliding windows.
-includeTaxa <taxon1,taxon2,…>	Filters input alignment to only include specified taxa. The taxa should be separated with commas and no spaces.
-includeTaxaInFile <filename>	Filters input alignment to only include taxa specified in file. The taxa cannot have spaces. Individual taxa should be separated by white space.
-excludeTaxa <taxon1,taxon2,…>	Filters input alignment to exclude specified taxa. The taxa should be separated with commas and no spaces.
-excludeTaxaInFile <filename>	Filters input alignment to exclude taxa specified in file. The taxa cannot have spaces. Individual taxa should be separated by white space.
-includeSiteNames	Filters input alignment to only include

<siteName1,siteName2,…>	specified site names. The site names should be separated with commas and no spaces.
-includeSiteNamesInFile <filename>	Filters input alignment to only include site names specified in file. The site names cannot have spaces. Individual site names should be separated by white space.
-excludeSiteNames <taxon1,taxon2,…>	Filters input alignment to exclude specified site names. The site names should be separated with commas and no spaces.
-excludeSiteNamesInFile <filename>	Filters input alignment to exclude site names specified in file. The site names cannot have spaces. Individual site names should be separated by white space.
-excludeLastTrait	This removes last column of Phenotype data. For example… Can be used to remove last column of population structure for use with MLM or GLM.
-numericalGenoTransform <type>	Performs genotype to numerical transform. <type> can be collapse or separated.
-newCoordinates <map filename>	This converts alignment to new coordinates specified in given map file.
-synonymizer	Runs the Synonymizer using the input dataset.
Analysis	
-glm	This takes a Phenotype dataset as input that is usually the intersection of sequence data, trait data, and population structure (optional).
-glmOutputFile <filename>	This sends GLM results to specified filename.
-glmMaxP <number>	This restricts the output file to entries with P values no larger than number specified.
-glmPermutations <number>	This sets the number of permutations. Default is to not do run permutations.
-mlm	This takes a Phenotype dataset as input (usually the intersection of sequence data, trait data, and population structure (optional)) and a Kinship matrix.
-mlmVarCompEst <method>	Defines the Variance Component Estimation for the previously specified -mlm. Method can be P3D (default) or EachMarker.
-mlmCompressionLevel <level>	Defines the Compression Level for the previously specified -mlm. Level can be Optimum (default), Custom, or None.
-mlmCustomCompression	This specifies the compression when compression

<number>	level is Custom. Default value is 1.0.
-mlmOutputFile <filename>	This sends MLM results to specified filename.
-mlmMaxP <number>	This restricts the output file to entries with P values no larger than number specified.
-diversity	Creates a Diversity Analysis step that uses an Alignment as input
-diversityStartBase <number>	This sets start base for the previously specified –diversity. Default is 0.
-diversityEndBase <number>	This sets end base for the previously specified –diversity. Default is last site.
-diversitySlidingWin	This uses sliding window analysis for the previously specified –diversity.
-diversitySlidingWinStep <number>	This sets the sliding window step size for the previously specified –diversity. Default is 100.
-diversitySlidingWinSize <number>	This sets the sliding window size for the previously specified –diversity. Default is 500.
-diversityTypeSites <type1,type2,…>	This sets type of sites the previous –diversity should analysis. Type options are ALL and INDEL. Types should be separated with commas and no spaces. Default is ALL.
-ld	Creates LinkageDisequilibriumPlugin. Uses Alignment from previous step to analysis linkage disequilibrium.
-ldPermNum <number>	This sets permutation number for the previously specified -ld. Default is 1000.
-ldRapidAnalysis true | false	Sets whether to use rapid analysis for the previously specified -ld. Default is true.
-ldWinSize <number>	Sets the window size for the previously specified -ld. Default is 50.
-ldType <type>	Sets the ld type for the previously specified -ld. Options are All, SlidingWindow (Default), and SiteByAll.
-ldTestSite <number>	Sets the test site for when LD type is set to SiteByAll.
-ck	Calculates Kinship from Marker Data.
-ckModelHets <type>	Sets how to model heterozygotes. Choose default type RelateHomo (Related to Homozygotes) or IndepState (Independent allele state).
-ckRescale true | false	Set whether to rescale results between 2 and 0. Default is true.
-tree <clustering method>	This creates a tree using given clustering method: Neighbor (default) or UPGMA. When exporting, use –exportType Text to get text version.

-treeSaveDistance true | false	This saves the distance matrix of a tree. Default is true.
-distanceMatrix	Calculate the distance matrix of given Alignment.
-distMatrixRanges	Calculates genetic distances for given taxon in specified physical position ranges.
-distMatrixRangesLocus <locus>	Locus that specified physical positions corresponds.
-distMatrixRangesTaxon <taxon>	Taxon of interest.
-distMatrixRangesPos <pos1,pos2,pos3,…>	Specified physical positions that define ranges. A comma should separate each one with no spaces.
-distMatrixRangesPosFile <filename>	File with list of physical positions that define ranges. Individual positions should be separated by white space.
-gs	Predicts phenotypes using ridge regression for genomic selection.
-genotypeSummary <types>	This generates summaries for alignment datasets. Types should be a comma-separated list (with no spaces) of the following (overall, site, taxa, all). Example -genotypeSummary overall,site
Results	
-td_csv <filename>	Writes (comma delimited) TableReport from previous plugin in current pipeline to specified filename.
-td_tab <filename>	Writes (tab delimited) TableReport from previous plugin in current pipeline to specified filename.
-td_gui	Displays TableReport from previous plugin in current pipeline in GUI.
-ldd <output type>	Creates LinkageDiseqDisplayPlugin. If output type is gui, this graphically displays results from a LinkageDisequilibriumPlugin. If output type is png, gif, bmp, jpg, or svg, then an image of that type is written to the output file specified with -o.
-ldplotsize <num>	Optionally specify LD plot size. Example: 1000 will produce a 1000 x 1000 plot. Default: 500. This should follow the -ldd flag within the current pipeline segment.
-ldplotlabels true | false	Optionally specify whether to show the LD Plot labels. DEFAULT: true. This should follow the -ldd flag within the current pipeline segment.

-o <output file>	This should follow the -ldd flag within the current pipeline segment.