TASSEL 5.0 Pipeline Command Line Interface:

Guide to using Tassel Pipeline

Terry Casstevens (*tmc46@cornell.edu*)
Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853-2703
March 9, 2017

Prerequisites

- Java JDK 8.0 or later (http://java.sun.com/javase/downloads/index.jsp).

Source Code

```
git clone https://bitbucket.org/tasseladmin/tassel-5-source.git
```

Install

```
git clone https://bitbucket.org/tasseladmin/tassel-5-standalone.git
```

OR

https://bitbucket.org/tasseladmin/tassel-5-standalone/downloads/?tab=tags

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 ...
```

Setting Logging to Debug or Standard *(With optional filename)*

```
./run_pipeline.pl -debug [<filename>] …
./run_pipeline.pl -log [<filename>] …
```

Examples

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

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

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

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>~~	**NOTE: This flag is no longer required. The pipeline will automatically run the necessary forks.** ~~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 to 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 ./run_pipeline.pl -fork1 -h genotype.hmp.txt -fork2 -includeTaxaInFile taxaList1.txt -input1 -export file1 -fork3 -includeTaxaInFile taxaList2.txt -input1 -export file2
-inputOnce<id>	This specifies a pipeline segment as a one-time input to a -combine. As such, this flag should follow -combine. After the -combine has received data from this input, it will use it for every iteration. Whereas -combine 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.
-printMemoryUsage	This prints memory used. Can be used in multiple places in the pipeline. ./run_pipeline.pl -fork1 -h mdp_genotype.hmp.txt -printMemoryUsage -KinshipPlugin -endPlugin -printMemoryUsage
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 <PHYLIP file>	Loads PHYLIP file.
-r <phenotype file>	Same at -t
-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)
-h5 <HDF5 file>	Loads HDF5 Alignment file (.hmp.h5).
-plink -ped <ped filename> -map <map filename>	Loads Plink format given ped and map files.
-fasta <filename>	Loads FASTA file.
-table	Loads a Table (i.e. exported from LD, MLM).
-vcf <filename>	Loads VCF file.
-importGuess <filename>	Uses Tassel Guess function to load file.
-hdf5Schema <hdf5 filename>	This inspects the HDF5 file for it's internal structure / schema. ./run_pipeline -hdf5Schema file.h5 -export schema.txt
-projection <filename>	./run_pipeline.pl -vcf file.vcf -projection file.pa -export output.hmp.txt
-sortPositions	Sorts genotype positions during import (Supports Hapmap, Plink, VCF)
-convertTOPMtoHDF5 <TOPM filename>	This converts TOPM file into a HDF5 formated TOPM file. New files extension will be .topm.h5.

	./run_pipeline.pl -convertTOPMtoHDF5 file.topm.bin
-retainRareAlleles <true | false>	Sets the preference whether to retain rare alleles. Notice this has no meaning for Nucleotide data. Only data that has more than 14 states at a given site (not including Unknown) are affected. If true, states more rare than the first 14 by frequency are changed to Rare (Z). If false, they are changed to Unknown (N).
-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 <chromosomes…>	This separates an input into its components if possible. For example, alignments separated by chromosome (locus). For alignments, optionally specify list of chromosomes (separated by commas and no spaces) to separate. Specifying nothing returns all chromosomes. Example: run_pipeline.pl -fork1 -h file.hmp.txt -separate 3,6 -export
-homozygous	This converts any heterozygous values to unknown. ./run_pipeline.pl -h file.hmp.txt -homozygous -export
-mergeGenotypeTables	Merges multiple Alignments regardless of taxa or site name overlap. Undefined taxa / sites are set to UNKNOWN. Duplicate taxon / site set to last Alignment processed. Example: run_pipeline.pl -fork1 -h file1.hmp.txt -fork2 -h file2.hmp.txt -combine3 -input1 -input2 -mergeGenotypeTables -export files merged.hmp.txt
-mergeAlignmentsSameSites -input <files> -output <filename>	Merges Alignments assuming all sites are the same in all Hapmap files. Input files separated by commas without spaces. The resulting file may have incorrect major/minor alleles, strand, center, etc. It uses values from first specified input file. Checks that Site Name, Chromosome, and Physical Position match for each site. Example: run_pipeline.pl -fork1 -mergeAlignmentsSameSites -input file1.hmp.txt,file2.hmp.txt -output temp

-export <file1,file2,…>	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. Genotype Tables will default to Hapmap format, Distance Matrix with default to SqrMatrix). Other exportable datasets 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, HDF5, VCF, Plink, Phylip_Seq, Phylip_Inter, Fasta, Text, ReferenceProbablity, Depth, SqrMatrix, SqrMatrixRaw (for MultiBLUP), SqrMatrixBin (for MultiBLUP), Phenotype, PlinkPhenotype, Table.
-exportIncludeAnno true | false	Indicates whether to include annotations in exported file if format allows.
-exportIncludeDepth true | false	Indicates whether to include depth in exported file if format allows.
-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 whitespace.
-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 whitespace.

-includeSiteNames <siteName1,siteName2,…>	Filters input alignment to only include 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 whitespace.
-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 whitespace.
-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.
-subsetSites <num>	This filters an alignment to include a random subset of sites. If <num> is >=1, it specifies the total number of sites to keep. If it is a decimal, it specifies the fraction of sites to keep. Adding the flag "-step" immediately after <num> tells the plugin to space the selected sites evenly instead of randomly.
-subsetTaxa <num>	This filters an alignment to include a random subset of taxa. If <num> is >=1, it specifies the total number of taxa to keep. If it is a decimal, it specifies the fraction of taxa to keep. Adding flag "-step" immediately after <num> tells the plugin to space the selected taxa evenly instead of randomly.
-step	This tells the previously specified -subsetTaxa or -subsetSites plugin to select sites/taxa evenly across the alignment instead of randomly.
~~-numericalGenoTransform <type>~~	https://bitbucket.org/tasseladmin/tassel-5-source/wiki/UserManual/NumericalGenotype/NumericalGenotype ./run_pipeline.pl -h mdp_genotype.hmp.txt -NumericalGenotypePlugin -endPlugin -export output -exportType ReferenceProbablity
-newCoordinates <map filename>	This converts alignment to new coordinates specified in given map file.

-synonymizer	Runs the Synonymizer using the input dataset.
Filter	
-filterAlign	Filters an alignment by sites.
-filterAlignMinCount <num>	Specifies the minimum count (default: 1) for the previously specified -filterAlign.
-filterAlignMinFreq <num>	Specifies the minimum frequency (default: 0.0) 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 Locus 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 is 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.
Analysis	
	GLM Flags are deprecated… Please use run pipeline.pl -FixedEffectLMPlugin
~~-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 <number>	This specifies the compression when compression 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.
-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.
-ldHetTreatment <type>	Sets the LD Heterzygous Treatment Method. Type can be Haplotype (For Inbred Lines), Homozygous (Default - Uses only homozygous site - heterozygotes set to missing), or Genotype (Not Implemented Yet).
~~-ck~~	~~Calculates Kinship from Marker Data.~~ **Deprecated: Please use ./run_pipeline.pl -KinshipPlugin**
-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 whitespace.
-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.