User Manual for

TASSEL

-Trait Analysis by aSSociation, Evolution and Linkage

Version 3

The Buckler Lab at Cornell University

(December 22, 2011)

www.maizegenetics.net/tassel

Disclaimer: While the Buckler Lab at Cornell University has performed extensive testing and results are, in general, reliable, correct or appropriate results are not guaranteed for any specific set of data. It is strongly recommended that users validate TASSEL results with other software.

Further help: Additional help is available beyond this document. Users are welcome to report bugs, request new features through the TASSEL website. Questions are also welcome to our current team members. For more quick and precise answers, please address your questions to the most pertinent person:

Tassel User Group (recommended)	**http://groups.google.com/group/tassel** tassel@googlegroups.com
General Information	**Ed Buckler (Project leader)** esb33@cornell.edu
Data import, GDPC, Pipeline	**Terry Casstevens** tmc46@cornell.edu
Statistical analysis	**Peter Bradbury** pjb39@cornell.edu **Zhiwu Zhang** zz19@cornell.edu

Contributors: Yogesh Ramdoss, Michael E. Oak, and Karin J. Holmberg, N. Stevens, and Yang Zhang.

The TASSEL project is supported by the National Science Foundation and the USDA-ARS.

Main Web Site: http://www.maizegenetics.net/tassel
Open source code: http://sourceforge.net/projects/tassel
Modified version of the PAL library is used: http://www.cebl.auckland.ac.nz/pal-project
Database access is achieved by GDPC middleware http://www.maizegenetics.net/gdpc

Table of Contents

INTRODUCTION

While TASSEL has changed considerably since its initial public release in 2001, its primary function continues to be providing tools to investigate the relationship between phenotypes and genotypes[1]. As indicated by its title – Trait Analysis by aSSociation, Evolution and Linkage – TASSEL has multiple functions, including association study, evaluating evolutionary relationships, analysis of linkage disequilibrium, principal component analysis, cluster analysis, missing data imputation and data visualization.

One of the design elements driving TASSEL development has been the need to analyze ever larger sets of data[2]. For example, the MLM (mixed linear model) function for association analysis originally used an EM (expectation-maximization) algorithm, which is a common method for solving mixed models but is relatively slow. Subsequently developers implemented the EMMA algorithm to increase computing speed[3]. Model compression was added to that to improve speed and statistical power for association study[4]. Another technique that optimizes variance components once and then uses the estimates to test markers now provides the ability to screen the large numbers of markers used in genome-wide association studies (GWAS). The method was independently described by Zhang et al. and Kang et al. in 2010. This method was named P3D by Zhang et al.[4] and EMMAX by Kang et al.[5]

TASSEL was designed for a wide range of users, including those not expert in statistics or computer science. A GWAS using the mixed linear model method to incorporate information about population structure[6-8] and cryptic relationships[9] can be performed by in a few steps by "clicking" on the proper choices using a graphic interface. All the processes necessary for the analysis are performed automatically, including importing phenotypic and genotype data, imputing missing data (phenotype or genotype), filtering markers on minor allele frequency, generating principal components and a kinship matrix to represent population structure and cryptic relationships, optimizing compression level and performing GWAS.

The command-line version of TASSEL, called the Pipeline, provides users the ability to program tasks using a script instead of the graphic user interface (GUI). This feature allows researchers to define tasks using a few lines of code and provides the ability to use TASSEL as part of an analysis pipeline or to perform simulation studies.

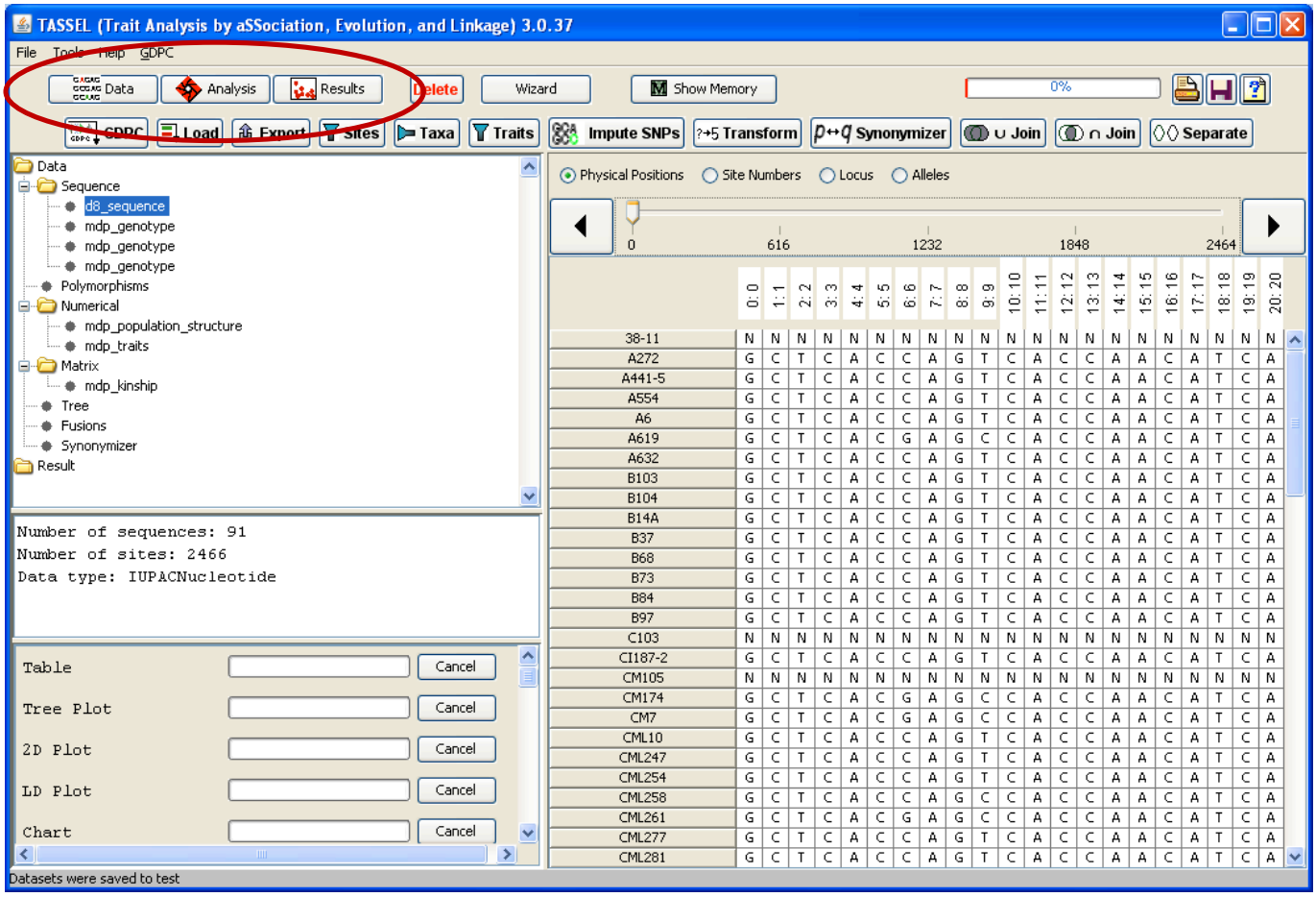
Due to the increasing availability of open data sources, TASSEL utilizes a data browser from the Genomic Diversity and Phenotype Connection (GDPC) project[10] to provide an interface to relational databases. As a result, TASSEL users can access any data source that provides a GDPC service. Using this middleware, which provides a common graphical interface, TASSEL users can avoid writing SQL queries to access data. Currently, GDPC provides connections to Panzea, Gramene, Germinate, and GRIN (USDA's Germplasm Resources Information Network).

TASSEL is written in Java, thereby enabling its use with virtually any operating system. It can be installed using Java Web Start technology by simply clicking on a link at www.maizegenetics.net/tassel. A stand-alone version of TASSEL can also be downloaded to use in pipeline mode or in any situation where the user wishes to start the software from a command line.

1 Getting Started

A quick way to get started using TASSEL is to load the tutorial data and try performing analyses. However, because some of the necessary steps may not be intuitive, we recommend that new users follow the tutorial at end of this manual. The objective of this section is to provide information necessary to install and start TASSEL software and to provide a brief overview of the interface.

Most functions are organized into three modes (Data, Analysis and Results) which correspond to the first three buttons on the TASSEL interface as shown below. Clicking one of these buttons changes the functions represented by the second row of buttons. Those three modes are described in detail in the subsequent sections of this manual. The screen shot shows TASSEL after the tutorial files have been loaded.



1.1 Installation

The graphic version of TASSEL can be installed in one of the three ways: using Java Web Start, as a stand-alone application, or using the source code

1.1.1 Web start

TASSEL can be installed using Java Web Start technology, which automatically checks for the most recent version of TASSEL each time the application is executed. In addition, Java Web Start will ensure that the correct version of the Java Runtime Environment is running, thus avoiding complicated

installation and upgrade procedures. Users should use Web Start unless they have a specific reason to use one of the other installation methods.

To begin, Java Web Start (JWS) must be installed (prior to the installation of TASSEL). JWS is included as part of Java Runtime Environment (JRE) 5.0 and above. PC's and Mac's will most likely have JWS already installed. If you need to install Java, the most recent version is available at http://www.java.com. The easiest way to tell if it is installed on your computer is to try running TASSEL from the following link:

http://www.maizegenetics.net/tassel

If you will be using TASSEL frequently and would prefer to launch the application from your desktop rather than by revisiting the website, Java Web Start can be used to manually launch TASSEL each time and/or to create a shortcut. Access the Java Application Cache Viewer by going to **Start** > **Settings** > **Control Panel** > **Java**. From the **General** tab, click on **Settings** in the **Temporary Internet Files** section and then click on **View Applications…** and the Java Application Cache Viewer will appear. (Another way of achieving this is by going to **Start** > **Run** and typing in **javaws**). The TASSEL icon should now be visible and can be used to launch the application. Shortcuts can be created from the menu of the Java Application Cache Viewer: **Application** > **Install Shortcuts.**

1.1.2 Stand-alone

Downloading a "stand-alone" version is recommended for anyone who has a slow Internet connection. While Java Web Start is a very good way of deploying software, it does not ask the user before attempting to download updates. Thus, a slow Internet connection may start a download process that requires an unreasonable amount of time to complete. If you are not interested in disabling your network connection each time before starting TASSEL, we recommend downloading the stand-alone version which does not attempt to update the program. However, given that TASSEL is a Java application, a Java Runtime Environment (version 1.6.0 or greater) is still required. To get the stand-alone version, download tassel3.0_standalone.zip from the TASSEL web site. To run the stand-alone version, double-click on the JAR file (sTASSEL.jar). Alternatively, from a command prompt (in Windows go to **Start** > **Run** and type in "cmd" or "command"), change into the tassel3.0_standalone directory and execute this command:

start_tassel.bat (For Windows)
start_tassel.pl (For UNIX)

1.1.3 Open source code

Open source code for the TASSEL software package is available at: http://sourceforge.net/projects/tassel. The package uses a number of other libraries that are included in the TASSEL distribution. These include a modified version of the PAL library (http://www.cebl.auckland.ac.nz/pal-project/), the COLT library (http://dsd.lbl.gov/~hoschek/colt/), and jFreeChart (http://www.jfree.org/jfreechart/). GDPC middleware (http://www.maizegenetics.net/gdpc) provides database access.

1.2 Panels

TASSEL is organized into five main panels. (1) The Control Panel at the top contains menus and buttons to control functions. (2) The Data Tree Panel is located beneath the Control Panel on the left side. This panel organizes data sets and results. Data set(s) displayed in the Data Tree Panel must first be selected before a desired function or analysis can be performed. To select multiple data sets, press the CTRL key while selecting the data sets. (3) The Report Panel is located below the Data Tree Panel. It displays

information about a selected data set from the Data Tree Panel, such as the type of data and how it was created. (4) The Progress Monitoring Panel below the Report Panel shows the progress of running tasks and has buttons that can be used to cancel tasks. (5) The Main Panel occupies the right side of the viewing area. It displays the content of a selected data set from the Data Tree Panel.

Functions in TASSEL are accessed by buttons and menus on the Control Panel.

The three buttons on the top left are the Mode Selectors (Data, Analysis and Results). The buttons below the Mode Selectors changed when a new Mode Selector is clicked. The modes are described in section 2-4. To the right of the Mode Selectors are the Progress Bar, and the Delete, Print, Save and Help buttons.

2 Data Mode 

Data mode serves the purpose of importing and managing data. Data mode is the default mode when TASSEL starts. Click on the Data button to switch to this mode.

Tassel has two ways of importing data. One way is via GDPC to import data from databases. The other way is via flat files formatted as genotypes (e.g. hapmap, flapjack, and plink), phenotypes (trait data), population structure and kinship matrices.

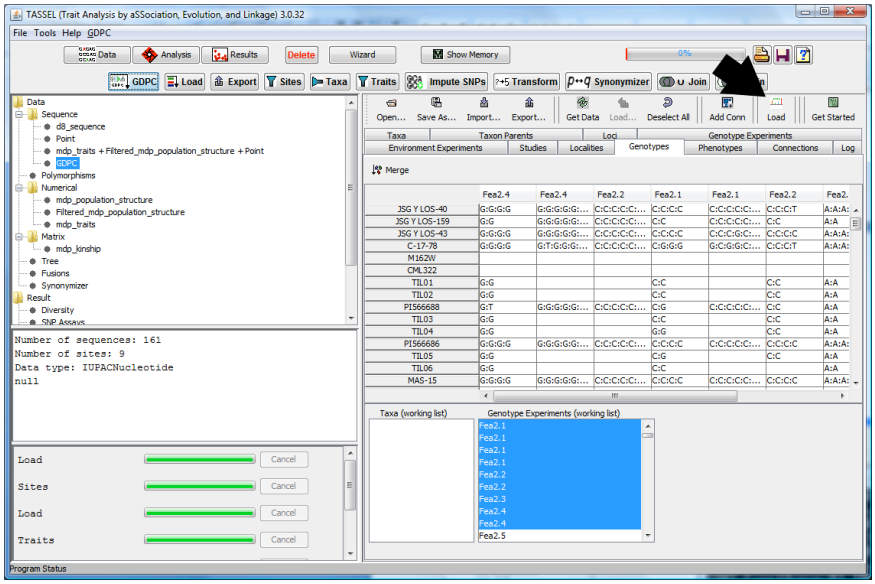
The preliminary data manipulations include filtering data by site or taxa, joining data and data transformation.

2.1 GDPC 

Genotype and phenotype data generated from numerous genomic research projects are still valuable resources for the public, even after results are published. Some of these data have been migrated to several databases and can be accessed using Genotype Data and Phenotype Connection (GDPC). GDPC is middleware that eliminates the need for end users of data to understand various database schemas and write SQL queries to extract data. Instead, the GDPC browser provides a single, easy-to-use interface which can extract genotype and phenotype data from a variety of sources[10]. Currently, GDPC has connections to the following databases:

- Gramene diversity for maize, wheat and rice[11-14]
  http://www.gramene.org/db/diversity/diversity_view
- Panzea[15-17]
  http://www.panzea.org
- GRIN
  http://www.ars-grin.gov

GDPC can be used within TASSEL or as a stand-alone application. To display GDPC in TASSEL, click on the GDPC button in Data mode.



Data is available for import once the user has defined the desired filters and data is visible in either the **Genotypes** or **Phenotypes** tab. To load data, activate either the **Genotypes** or **Phenotypes** tab (depending on the data you wish to import) and then click the **Load** button Load.

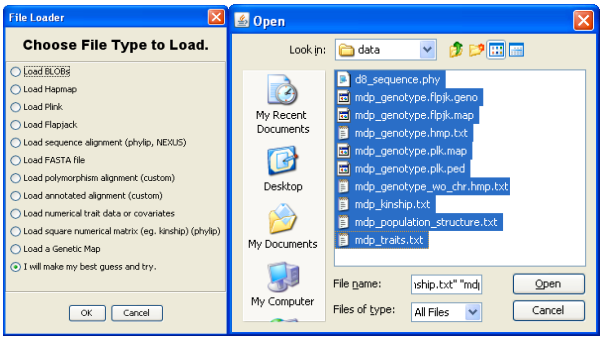
For additional information about GDPC, please see http://www.maizegenetics.net/gdpc

2.2 Load Load

This function provides options to import files for genotypes, phenotypes, populations structure, and kinship matrices. Several common sequence formats are accepted for genotype data, including BLOB, Hapmap, Plink, and Flapjack, and a general format for polymorphism data. Some file types used by TASSEL version 2 are also supported for backward compatibility. Phenotype and population structure can be imported as numerical trait data or covariates. Kinship must be loaded as a square numerical matrix.

Users can either specify the file type or use the "Guess" option to let the program determine the file type. As an example, we describe how the "Guess" function can be used to import all the files from the tutorial data set. The tutorial data can be downloaded from the TASSEL website or using this link: http://www.maizegenetics.net/tassel/docs/TASSELTutorialData3.zip.

To use the data, the zip file must be uncompressed and saved in a folder that the user specifies. To import data click the LOAD button. The File Loader dialog box will then pop up to let user choose the files and specify a format. For the files in the tutorial data set, the default (Guess) function will load all the files correctly. Multiple files can be imported simultaneously by highlighting them first (holding Shift or Control key while clicking) and then clicking the Open button.



2.2.1 BLOB

A Binary Large Object (BLOB) is a collection of binary data stored as a single entity. In TASSEL, BLOBs are used to compress large data sets into more manageable sizes. For sequence data, three types of BLOBs are used: SNP value BLOB, position BLOB and SNP ID BLOB. The three BLOBs are used to store individual SNP values, SNP position within the genome and the SNP identifiers respectively.

A BLOB is composed of two components, a header and a body. The header for each BLOB is 1024 bytes long, while the length of the body depends on the type of BLOB and on the amount of data being stored.

For a more detailed description on the structure and information contained within the header and body, refer to the GDPDM BLOB Specifications.
(http://www.maizegenetics.net/gdpdm/docs/20100526/GDPDMBlobSpecification_20100526.pdf)

2.2.2 Hapmap

Hapmap is a text based file format for storing sequence data. All the information for a series of SNPs as well as the germplasm lines is stored in one file. The first row contains the header labels, and each additional row contains all the information associated with a single SNP. The first 11 columns describe attributes of the SNP, while the following columns describe the SNP value for a single germplasm line. The first 12 columns of the first row should look like this, where “Line 1” is the beginning of germplasm line names.

rs#	alleles	chrom	pos	strand	assembly#	center	protLSID	assayLSID	panelLSID	QCcode	Line 1

While all 11 header columns are required, not all 11 of the columns need to be filled in for TASSEL to correctly interpret the data. The only required fields are “chrom”, Chromosome name, and “pos”, Position.

For TASSEL to correctly read Hapmap data, the data must be in order of chromosome and position within each chromosome, and the file should be TAB delimited. If some of the data is missing the correct number of TABs must still be present, so that TASSEL can properly assign data to columns.

2.2.3 Plink

Plink is a whole genome association analysis toolset, which comes with its own text based data format. The data is stored in a set of two files, a .map file and a .ped file.

The .ped file contains all the SNP values and has six mandatory header columns for Family ID, Individual ID, Paternal ID, Maternal ID, Sex and Phenotype. TASSEL only requires that the Individual ID field be filled in. Each row of the .ped file describes a single germplasm line. Notice in Plink, an unknown character is represented with a '0'. However in TASSEL an unknown character is represented with a 'N', and '0' is used to represent heterozygous indel. TASSEL will automatically convert between the '0' and the 'N'. Any exported Plink files will represent the heterozygous indel with a '+' (insertion) and a '-' (deletion).

The .map file describes all the SNPs in the associated .ped file, where each row provides information on one SNP. The .map file must contain exactly four columns: Chromosome, rs#, Genetic distance and Position. TASSEL does not require the Genetic distance field to be filled in.

Both files should be TAB delimited.

For a more detailed description on the data format, please visit the Plink basic usage and data formats webpage: (http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml).

2.2.4 Flapjack

Flapjack is a software tool for graphical genotyping and haplotype visualization. The program is capable of outputting data in its own text based data format. Like Plink, the data is stored in a set of two files, a .map file and a .geno genotype file.

The genotype file contains all the SNP values. Each column in the first row contains a SNP ID, except for the first column, which is blank. The first column of the following rows contains the germplasm line names. TASSEL requires that all fields be filled out in order for data to be read correctly.

The .map file describes all the SNPs associated with the genotype file. Each row describes a single SNP. There are three columns in the .map file for Flapjack, SNP ID, Chromosome and Position, all of which are required for TASSEL to run correctly.

Both files should be TAB delimited.

For a more detailed description on the Flapjack data file format, please visit the Flapjack data import website: (http://bioinf.scri.ac.uk/flapjack/help/gui.dialog.DataImportDialog.shtml).

2.2.5 Polymorphism

A general format that accepts almost any type of marker data can also be used. Any alphanumeric character is allowed. Diploid data can be represented by separating alleles with a colon (":"), for example A:A, A:B, or B:B. All loci in a file **must** have the same ploidy level. The first line starts with the symbol <Marker> followed by the marker names. Subsequent lines must start with the name of the individual or taxon genotyped followed by the marker scores in the same order as the header. Comments can be inserted at the beginning of the file as long as any comment lines begin with the symbol "#". Columns are TAB delimited. Numeric values are allowed but, by default, will be treated as classification variables not as covariates in analyses.

Example 1:

```
<Marker>    m1    m2    m3    m4    m5
33-16 A     B     B     A     A
38-11 A     A     B     A     A
4226  A     B     B     A     B
```

In some cases, a user may wish to have marker values treated as numerical covariates. If the first line of the file is "<Numeric>", then the data will be imported as numeric data but used as marker data in GLM and MLM.

Example 2:

```
<Numeric>
<Marker>    m1    m2    m3    m4    m5
33-16 0     1     1     0     0
38-11 0     0     1     0.3   0
4226  0     1     1     0.5   0
```

Note to TASSEL 2.1 users: The polymorphism format specified in TASSEL v2.1 is still supported to provide backward compatibility.

2.2.6 Phylip

The Phylip format used by TASSEL version 2.1 will continue to be supported. Details on Phylip format are described at the following website: http://evolution.genetics.washington.edu/phylip/doc/sequence.html

2.2.7 Numerical data

This type of format is used for trait and covariate data such as population structure. Similar to sequence alignment genotype data, numerical data also consists of two parts: a header that defines data structure and a body containing the main data. Tabs should be used as delimiters. However, any white space character such as blank will be treated as a delimiter as well. As a result, embedded blanks in names will cause data to be imported incorrectly. We suggest representing missing values using "NA", or "NaN". However, any text value (e.g. "?") will be interpreted as missing data. There are several formats for numerical data to fit the requirement for modeling. Trait data (dependent variables) can be imported by starting the first line with "<Trait>" and following that with the trait names. Additional classifiers may also be included in subsequent header rows by starting the row with "<Header name=xxx>" followed by a name for each column of data. For instance, to define environments, start the second header row with "<Header name=env>".

Comment lines may be inserted at the beginning of the file as long as each comment line begins with the character "#".

2.2.7.1 Trait format

This format does not require users to provide information on number of rows and columns. The file stats with key word <Trait> followed by names of columns. The column for line should not be labeled.

Example 1, simple list of trait values:

```
<Trait>     EarHT dpoll EarDia
811   59.5  NA    NA
33-16 64.75 64.5  NA
38-11 92.25 68.5  37.897
4226  65.5  59.5  32.21933
4722  81.13 71.5  32.421
A188  27.5  62    31.419
…
```

Example 2, traits data collected in multiple environments:

```
<Trait>     EarHT PlantHT     EarHT PlantHt
<Header name=env> Loc1  Loc1  Loc2  Loc2
811   59.5  NA    NA    NA
33-16 64.75 121.5 NA    NA
38-11 92.25 153.8 37.897      83.4
4226  65.5  130.1 32.21933    82.1
4722  81.13 165.7 32.421      90.1
A188  27.5  110.2 31.419      79.6
…
```

2.2.7.2 Covariate Format

Covariate data uses the same format as trait data except that the first line must be "<Covariate>". This line tells TASSEL that the variables in this file will be used as covariates not as dependent variables. This is the format to use for population structure covariates.

Example:

```
<Covariate>
<Trait>     Q1    Q2    Q3
33-16 0.014 0.972 0.014
38-11 0.003 0.993 0.004
4226  0.071 0.917 0.012
4722  0.035 0.854 0.111
A188  0.013 0.982 0.005
…
```

2.2.7.3 TASSEL version 2.1 formats

Version 2.1 formats for numeric data will continue to be supported to provide backward compatibility. However, that format does not identify covariates as such. As a result, any covariates imported using this format will need to be properly identified using the "Trait filter" function described later in the manual.

2.2.7.4 Repeated measurements

A format for repeated measurements may be implemented in the future.

2.2.8 Square Numerical Matrix

Kinship can be calculated externally from pedigrees by using SAS Proc Inbreeding[18] or from markers by using software packages such as SPAGedi[19]. The following format is provided to import the resulting kinship estimates:

If n represents the number of taxa, the format for kinship files is as follows:

n
Taxa1Name r11 r12 … r1n
Taxa2Name r21 r22 … r2n
…
TaxanName rn1 rn2 … rnn

Here rij (i, j=1,2, …, n) is the element in the kinship matrix located at row i and column j.

Missing values are not allowed for kinship matrix.

Important note: The current format is different from the format used in TASSEL version 2.0 or lower.

2.2.9 Genetic Map

A Genetic Map is a list of markers with chromosome and map position and, optionally, physical position. It can be used by GLM and MLM to provide genetic positions in the output files. It is not used as part of the analysis. The input format is:

First line: <Map> (as is, including the brackets)
Following lines: marker name, chromosome name, genetic position, physical position (actual data)

Example:
<Map>
marker1 c1 21.3 2456873
marker2 c1 52.1 52345691

There is no header line as such. Marker name, chromosome name, and genetic position are required. Physical position is optional and not used at this time. It is there because it is anticipated that information from this map may be used to convert between physical and genetic position at some time in the future.

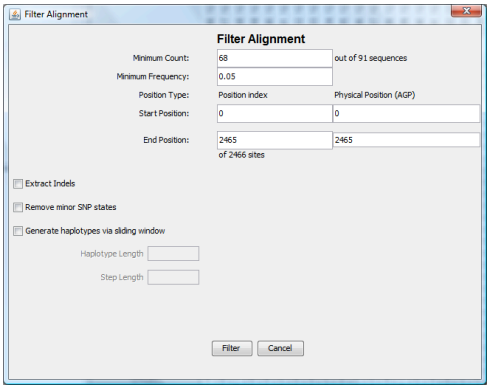
2.3 Export 

Options are provided to export sequence data: BLOB, Hapmap, Plink, Flapjack, Phylip (Sequential or Interleaved). Phenotypes and covariate data is exported as numerical trait data. Table Reports are exported as a tab delimited table.

This button has the same function as the "Save selected as" on the File menu. For numerical data, the function of Export is similar to the Table function in Results mode.

2.4 Sites 

The alignment can be filtered in several ways. Monomorphic sites can be eliminated, and regions of a sequence can be eliminated.



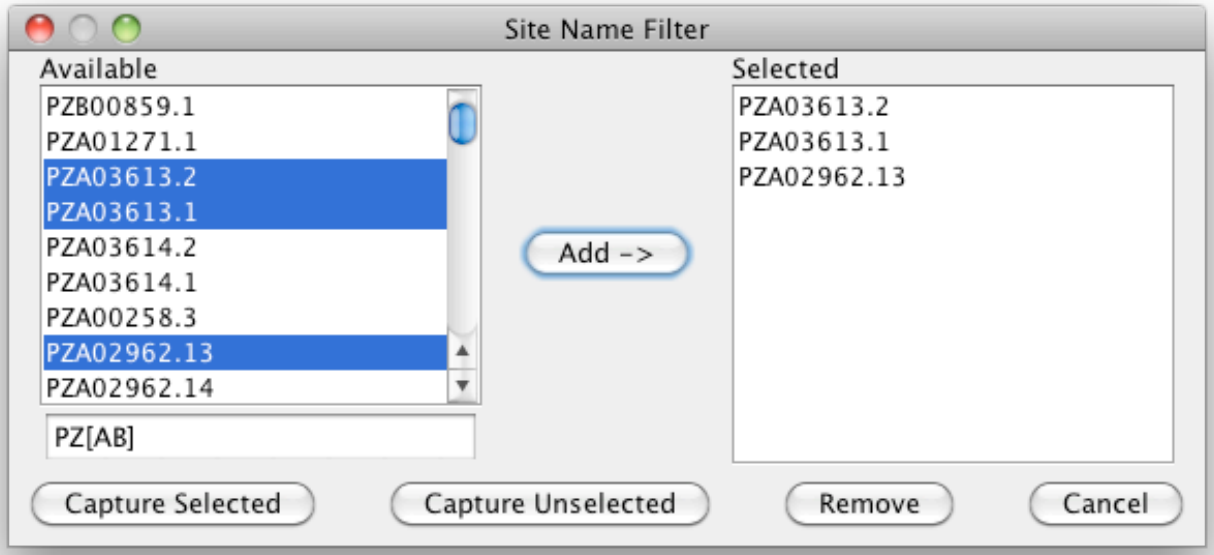
Minimum Count - the minimum number of taxa in which the site must have been scored to be included in the filtered data set (GAP or missing data do not count).
Minimum Frequency - the minimum frequency of the minority polymorphisms for the site to be included in the filtered data set.
Start Position, End Position – establishes the range of sites for filtering.
Extract Indels - if selected, indels are extracted from the alignment. If not selected, only point substitutions are extracted.
Remove minor SNP states – converts tertiary and rarer states to missing data ("?"), thereby forcing sites to have only two types of segregating sites at a locus. This may help remove sequencing errors.
Generate haplotypes via sliding window – creates haplotypes from an ordered set of SNPs.

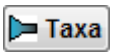
2.5 Site Names 

First select genotypic data from the data tree. The resulting dialog displays the site names associated with the selected data. By using either the CTRL or SHIFT key in conjunction with the mouse, the user can select or deselect site names. Once desired site names have been moved to the "Selected" window using the "Add ->" button, the "Capture Selected" or "Capture Unselected" buttons will create a new data set containing only the desired taxa.

Using the search box…

- * is the wildcard.
- * is always implied at end of search string.
- Search string is case sensitive. For example: use [Aa]bc to match taxa beginning with Abc or abc.
- PZ[AB] Will match anything starting with PZA or PZB.

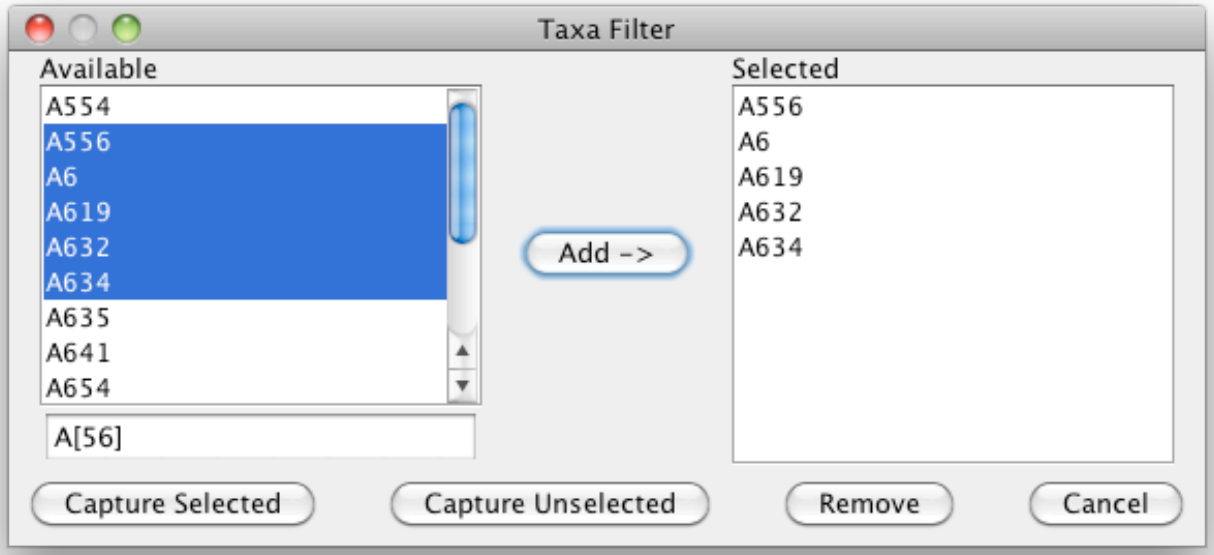


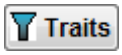
2.6 Taxa 

First select genotypic, phenotypic, or population structure data from the data tree. The resulting dialog displays the taxa associated with the selected data. By using either the CTRL or SHIFT key in conjunction with the mouse, the user can select or deselect taxa. Once desired taxa have been moved to the "Selected" window using the "Add ->" button, the "Capture Selected" or "Capture Unselected" buttons will create a new data set containing only the desired taxa.

Using the search box…

- * is the wildcard.
- * is always implied at end of search string.
- Search string is case sensitive. For example: use [Aa]bc to match taxa beginning with Abc or abc.
- A[56] Will match anything starting with A5 or A6



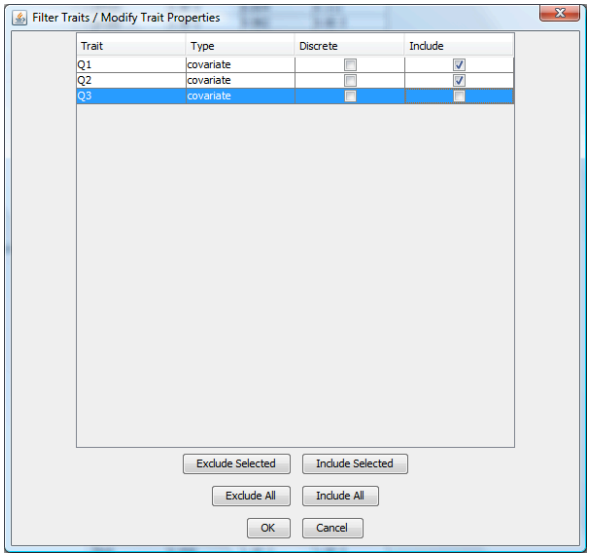
2.7 Traits 

Clicking the "Traits" button on the "Data" toolbar launches the Trait Filter dialog. This dialog is used with numerical data sets to (1) change the trait type, (2) view, but not change whether the trait is discrete or continuous and (3) drop one or more traits from the data set. In addition, the dialog can be used to view the trait properties without changing them. If the "OK" button is clicked, a new data set is created that incorporates the changes, the original data set remains unchanged, and the dialog closes. If the "Cancel" button is clicked no data set is created, the original data set remains unchanged, and the dialog closes.

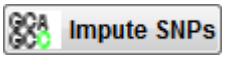
Allowable trait types are data, covariate, factor and marker. Generally, data and covariate traits will be continuous (not discrete) and factor will be discrete. Markers in a numerical data set will be continuous. Discrete valued markers are better imported as sequence or polymorphisms.

Clicking "Exclude All" unchecks the "Include" box for all traits. Clicking "Include All" checks the "Include" box for all traits. The "Exclude Selected" and "Include Selected" buttons do the same thing for traits that have been highlighted by selecting them with the mouse.

Important: Once a numerical data set has been joined with genotypes, it can no longer be modified using the trait filter function.

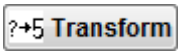


2.8 Impute SNPs



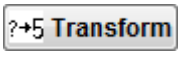
This function is used to impute missing genotypes. A sequence data type is required to use the function.

2.9 Transform

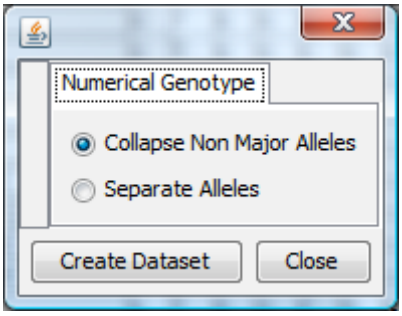


This suite of functions allows multiple data manipulation on genotype and phenotype (numerical) data. When a genotype data set is selected, the data are transformed to numbers. When a numerical data set is selected, mathematical transformation, data imputation and principal component analysis (PCA) can be performed. The Transform columns tags will be displayed in a Data dialog box with three tabs: Trans, Impute and PCA.

2.9.1 Genotype Numericalization



Two options are provided to transform genotype from character to numerical as shown in the following dialog box.



2.9.1.1 Collapse Non Major Alleles

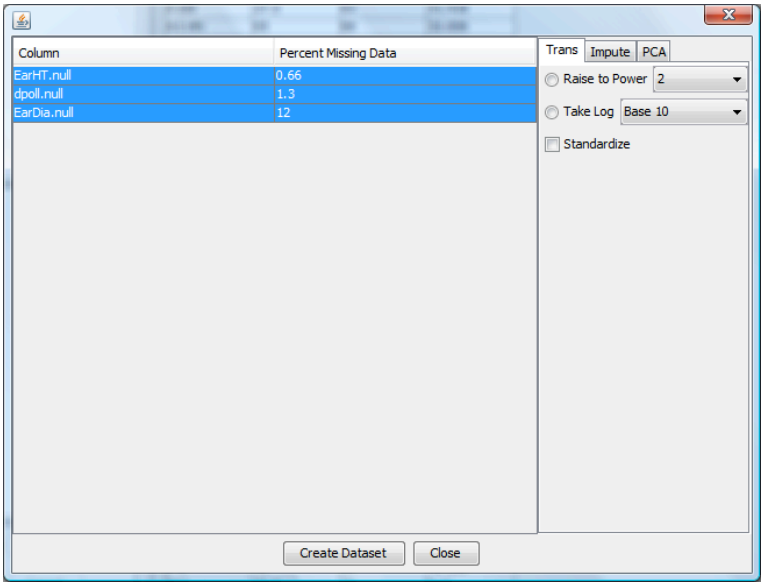
This function assigns 1 to the major allele and 0 to any other alleles. The converted genotypes are saved in a new numerical data set.

2.9.1.2 Separate Alleles

This function assigns an indicator (1 for present and 0 for absent) for each allele. The converted genotypes are saved in a new numerical data set.

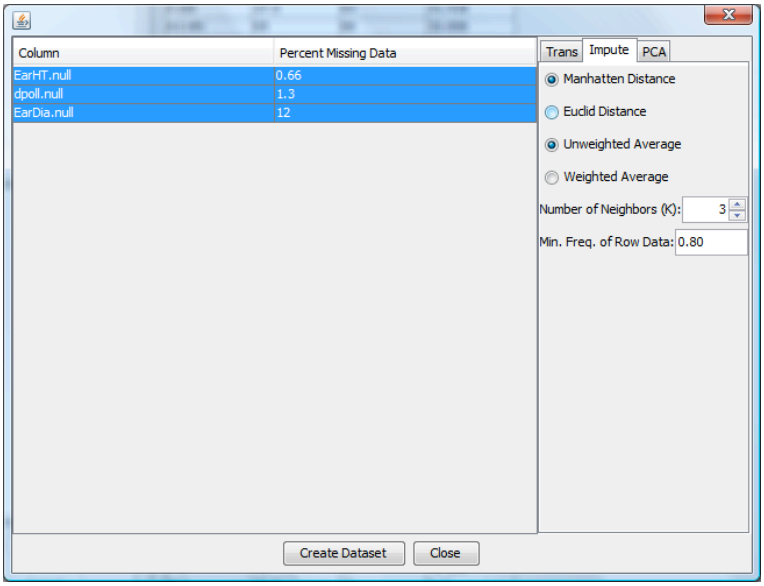
2.9.2 Transform and/or Standardize Data

The **Trans** dialog box is the default selection, as shown below. In the **Column** list, select the column(s) you wish to transform. Then select the type of transformation you wish to execute. Selecting the **Standardize** checkbox will transform data by subtracting the column mean from the value of the trait and then dividing by the column's standard deviation. Clicking on the **Create Data set** button will result in the placement of a dataset containing only the selected columns in the Data Tree.



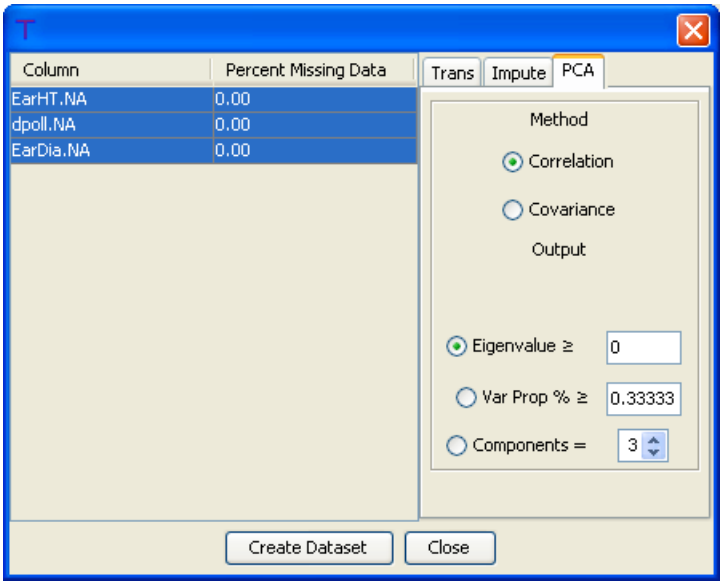
2.9.3 Impute Phenotype

The k-nearest-neighbor algorithm[20] is used to impute missing phenotype data. If data is missing for a taxon for one of the traits, the algorithm finds other taxa (neighbors) that are most like it for the non-missing traits. It uses the average of the neighbors to impute the missing data. Click on the **Impute** tab to display the following:



2.9.4 PCA

Principal component analysis (PCA) can only be performed on a numerical data set without missing values. Two methods are available: correlation or covariance. This determines whether a correlation or covariance matrix will be used as the basis for the analysis. The default, correlation, is a reasonable choice for genetic data. The number of PCA axes in the output data set can be controlled by selecting either of the minimum eigen value associated with each axis, the minimum percent of the variance captured by an axis or the number of axes. The resulting axes will be sorted by the amount of variance each captures.

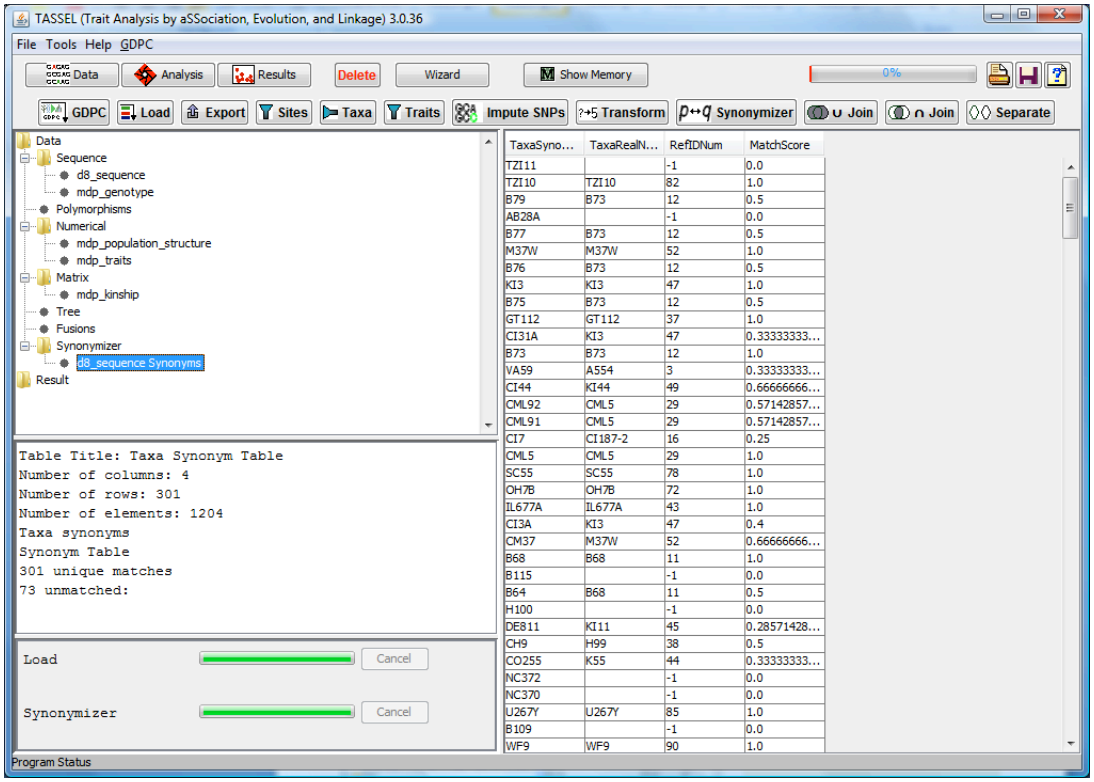


2.10 Synonymize Taxa Names p↔q Synonymizer

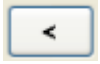
• This button makes taxa names uniform to permit the joining of data sets.

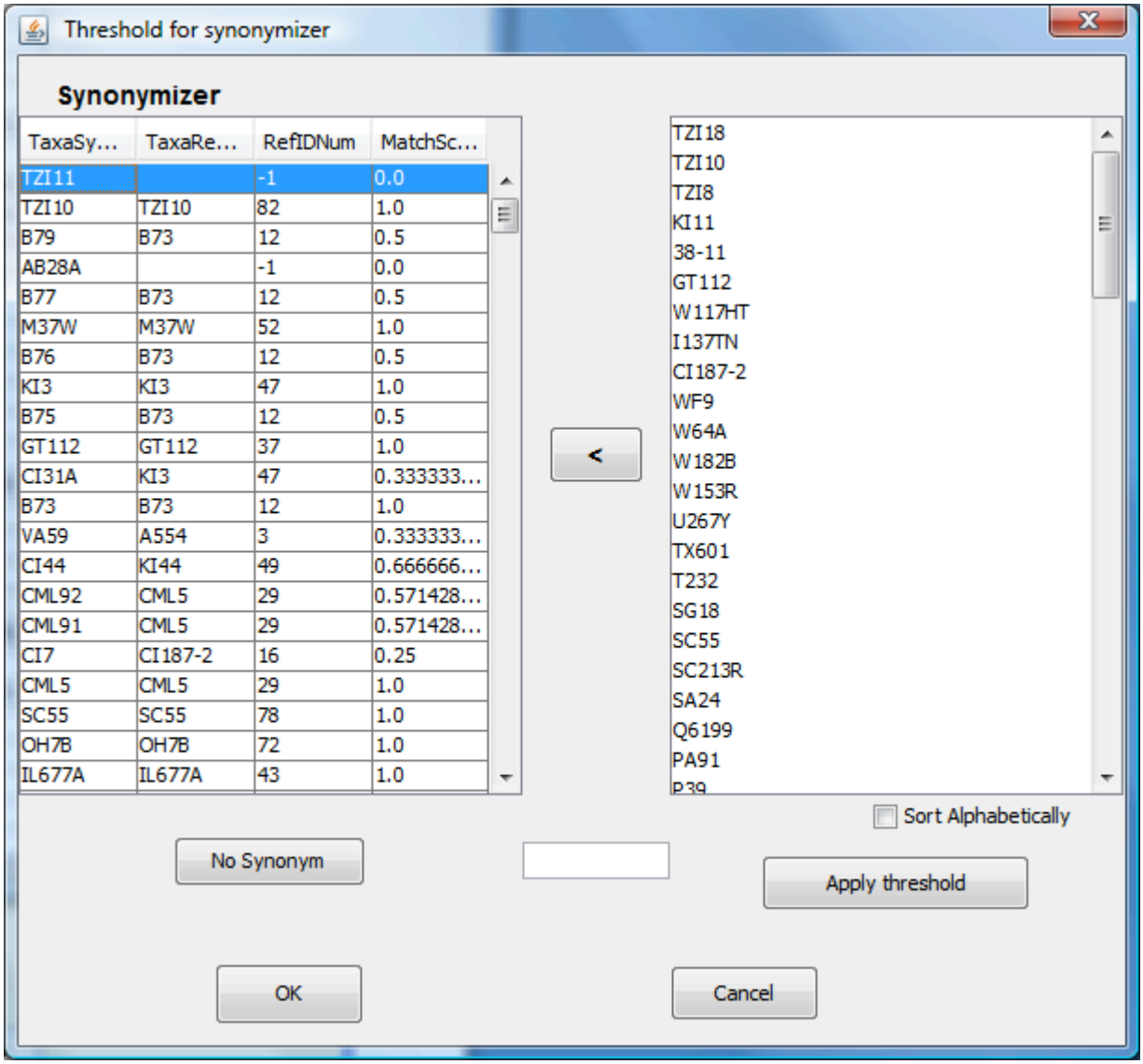
The join functions that generate fused data sets work by matching taxa names. Consequently, if multiple names exist for a given taxon (an added suffix, alternative spellings, different naming conventions, etc.) then the two data sets will not join correctly. To help remedy this, the Synonymizer function allows the taxa names of one data set to replace similar taxa names in the second data set. It relies on an algorithm that calculates the degree of similarity between names, using the name from the first set which is most similar to that in the second data set.

When using the Synonymizer, keep in mind that order of selection matters. Always select the data set with the names you wish to use (the "real" name) *first*, and then, while holding down the CTRL key, click on the second data set with the taxa names you wish to change (the "synonym"). Then click on the **Synonymizer** button. A synonym data set will be placed on the Data Tree panel under **Synonyms**. Each name in the data set selected second is now listed in the **TaxaSynonym** column. Next to this column is a **TaxaRealName** column listing the highest scoring match derived from the "real" name data set. The **MatchScore** column gives an indication of the amount of similarity between the two names (where 0 is no similarity and 1.0 is identity).



Caution! Before the synonyms are applied, we strongly encourage the user to check the match score, especially for those taxa with low match scores. To do that, the user selects the synonym file and clicks the "Synonymizer" button. The incorrect matches, usually the ones with the lowest match scores, can be rejected at this point. Sorting on the match score column first makes this a fairly easy process.

In the event that some of the taxa are not interpreted correctly, matches can be modified manually. Select the taxa you wish to modify on the left side, and then choose a replacement taxa from the right side. Click the arrow button  to substitute the taxa. Taxa with no synonym can be identified by selecting then clicking "No Synonym". Click **OK** to save the changes.



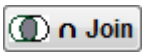
Once it has been determined that the taxa names were matched correctly, the synonyms can be applied. With the synonyms selected, hold down the CTRL key while clicking on the second/synonym data set (the data set whose names you would like to change). Then once again click on the **Synonymizer** button to apply the new names to the data set.

2.11 Union Join 

This button joins multiple data sets by a union of their taxa. Missing data will be inserted if taxa are missing from one data set.

Select multiple data sets using the CTRL key in conjunction with mouse clicks, and then click on the union button to join the data sets.

Because this function uses taxa names to join data sets, any variation in taxa names can prevent proper joining. Taxa names can be made uniform by using the "Synonymizer".

2.12 Intersection Join 

This button joins multiple data sets by the intersection of their taxa. Taxa must be present in both data sets to be included.

Select multiple data sets using the CTRL key in conjunction with mouse clicks, and then click on the intersection button to join the data sets.

Because this function uses taxa names to join data sets, any variation in taxa names can prevent proper joining. Taxa names can be made uniform by using the "Synonymizer".

3 Analysis Mode 

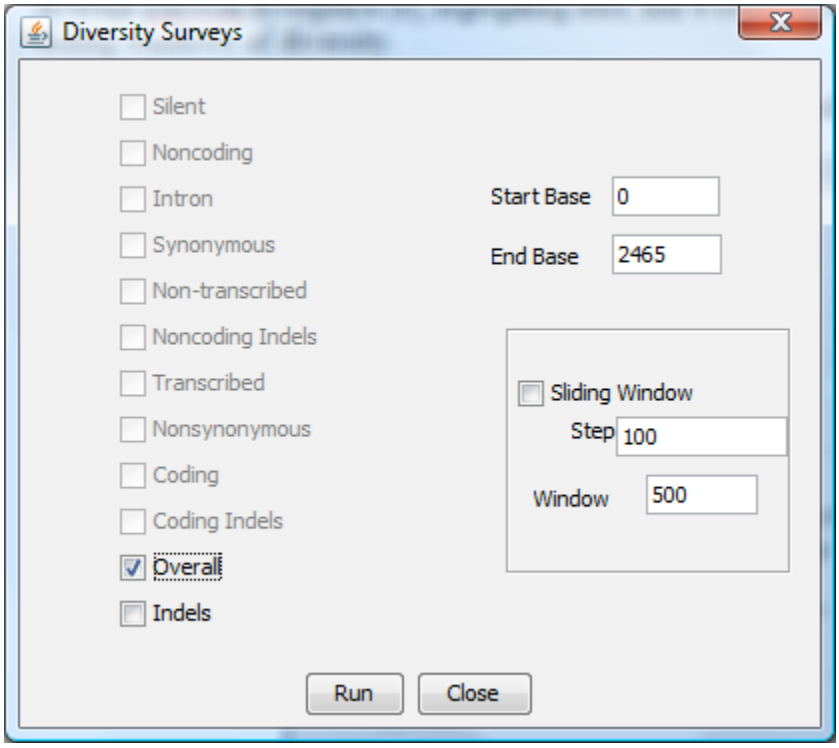
Analysis mode consists of the following options:

3.1 Diversity 

This button executes a basic diversity analysis.

Average pairwise divergence (π), segregating sites, and θ estimates ($4N\mu$) can be calculated, as well as sliding windows of diversity.

To run a diversity analysis, click on a raw sequence alignment, and then select **Analysis → Diversity**.



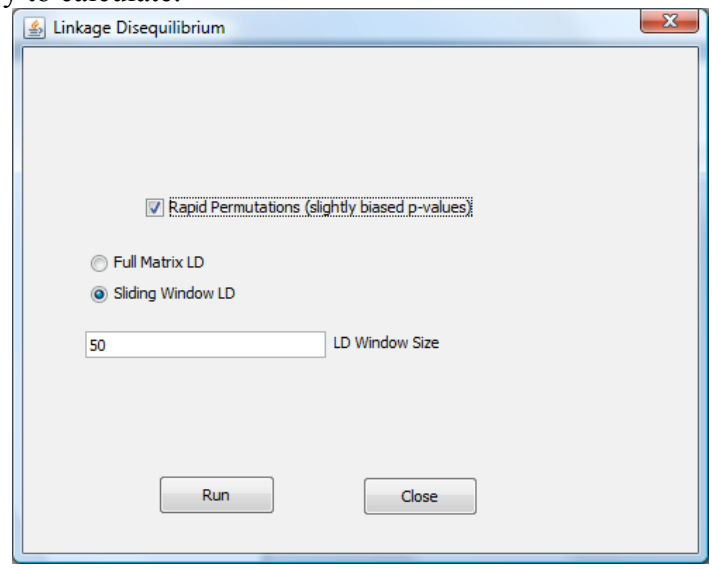
In the resulting Diversity Surveys dialog box, the various site classes available for analysis are listed on the left. If the sequence has no annotation, then only the "Overall" and "Indels" options will be active.
A sliding window of diversity can also be calculated across the region. To produce a sliding window, check the box next to "Sliding Window," and then enter the desired step size and size of the sliding window.

Results can be plotted using **Results → Chart** or viewed in a table via **Results → Table**.

3.2 Linkage Disequilibrium 

This button generates a linkage disequilibrium data set from SNP data.

NOTE: It is important to use only filtered data sets (apply **Data → Sites** first) when estimating linkage disequilibrium, as a raw alignment with numerous invariant bases will take a very long time and consume a large amount of memory to calculate.



Linkage disequilibrium between any set of polymorphisms can be estimated by clicking on a filtered set of polymorphisms and then using **Analysis → Link. Diseq.** At this time, *D'*, *r2* and *P*-values will be estimated. The current version calculates LD between haplotypes with known phase only (unphased diploid genotypes are not supported; see PowerMarker or Arlequin for genotype support).

> **D'** is the standardized disequilibrium coefficient, a useful statistic for determining whether recombination or homoplasy has occurred between a pair of alleles.
>
> r^2 represents the correlation between alleles at two loci, which is informative for evaluating the resolution of association approaches.

D' and *r2* can be calculated[21] when only two alleles are present. If multiple alleles are present, a weighted average of *D'* or *r2* is calculated between the two loci[22]. This weighted average is determined by calculating *D'* or *r2* for all possible combinations of alleles, and then weighting them according to the allele's frequency. *Note: It is not entirely certain that this procedure fully accounts for allele number effects.*

P-values are determined by two methods. If only two alleles are present at both loci, then a two-sided Fisher's Exact test is calculated. *Note: Previous editions of TASSEL used a one-sided test, but TASSEL version 1.0.8 and later use a two-sided test.*

If more than two alleles are present, permutations are used to calculate the proportion of permuted gamete distributions that are less probable then the observed gamete distribution under the null hypothesis of independence[21] .

When calculating linkage disequilibrium, users have the option of employing "**Rapid Permutations."** If this option is selected, the algorithm will compute either a fixed number of permutations or run until 10 permutations are found that are more significant than the observed P-value. While this slightly reduces P-values, it also saves a large amount of computational time. If an unbiased p-value is desired, then the user must unselect the **"Rapid Permutations"** check box.

"Full Matrix LD" calculates LD for every combination of sites in the alignment. "Sliding Window LD" calculates LD for sites within a window of sites surrounding the current site. The LD Window Size determines the width of the window on one side of the current site.

Linkage disequilibrium results can be plotted using **Results → LD Plot** or viewed in a table via (**Results → Table**).

3.3 Cladogram 

This function generates a tree or cladogram data set.

TASSEL produces neighbor-joining trees using only simple parsimony substitution models.

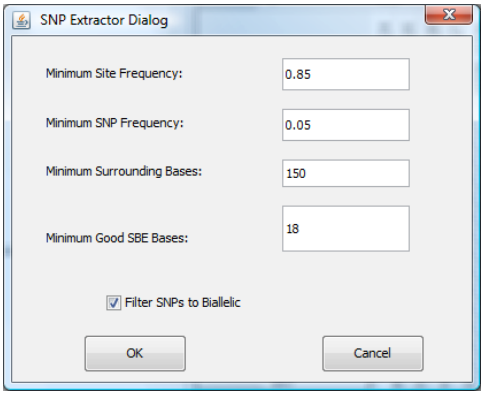
To retrieve cladogram data, first select genotypic data from the Data Tree panel and then click on the **"Analysis"** button, followed by the "**Cladogram**" button. The resulting tree data and the corresponding matrix will appear as separate data sets on the Data Tree panel.

Results can be plotted using **Results → Tree Plot**.

3.4 SNP Extract 

"SNP Extract" extracts SNPs from a raw sequence alignment into a useful format for export. Additionally, this function provides information for designing genotyping assays.

Below is a detailed explanation of the SNP Extractor Dialog:



Minimum Site Frequency: the minimum frequency for which the site must have a good base

Minimum SNP Frequency: the minimum frequency of the minority polymorphisms for the site to be included in the resulting data set

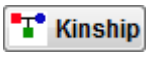
Minimum Surrounding Bases: the minimum number of good bases on at least one side of the SNP

Minimum Good SBE Bases: the minimum number of good bases on at least one side of SNP

Filter SNPs to Biallelic: converts tertiary and rarer states to missing data ("?"), thereby forcing sites to have only two types of segregating sites at any particular locus. This helps to remove bad sequence effects.

Results are displayed on the Data Tree panel and include SNPs along with their context. Additional information is also provided, including: the location of the nearest polymorphisms on either side, polymorphism information content ("**PIC**") and "**Haplotype PIC**". "**Overall score**" is essentially an estimate of the ability to design a single-base pair extension reaction in the region.

These results can be exported by using a table (**Results → Table**).

3.5 Kinship 

The function generates a kinship matrix from a set of random SNPs. To do so, first highlight SNP data then click on the "**Analysis**" button, followed by the "**Kinship**" button. The resulting kinship data will be added as a data set on the Data Tree panel.

When a genotype file is selected, the kinship matrix is generated by first using the TASSEL Cladogram function to calculate a distance matrix. Each element d_{ij} of the distance matrix is equal to the proportion of the SNPs which are different between taxon i and taxon j. The distance matrix is converted to a similarity matrix by subtracting all values from 2 then scaling so that the minimum value in the matrix is 0 and the maximum value is 2. Kinship can be derived from a set of random SNP data (a minimum of several hundred SNPs spread over the whole genome is recommended).

Warning: This method currently works correctly only for homozygous inbred lines. The method will be modified in the near future to work with heterozygous taxa. At that point, this warning will be removed.

Users may also load their own kinship data using **Data → Load**. Kinship matrices can be calculated using the SPAGeDi software package (http://www.ulb.ac.be/sciences/ecoevol/spagedi.html). Comparisons of methods for calculating kinship can be found in the literature (*e.g.* Stich et al. 2008).

3.6 General Linear Model 

This function performs association analysis using a least squares fixed effects linear model.

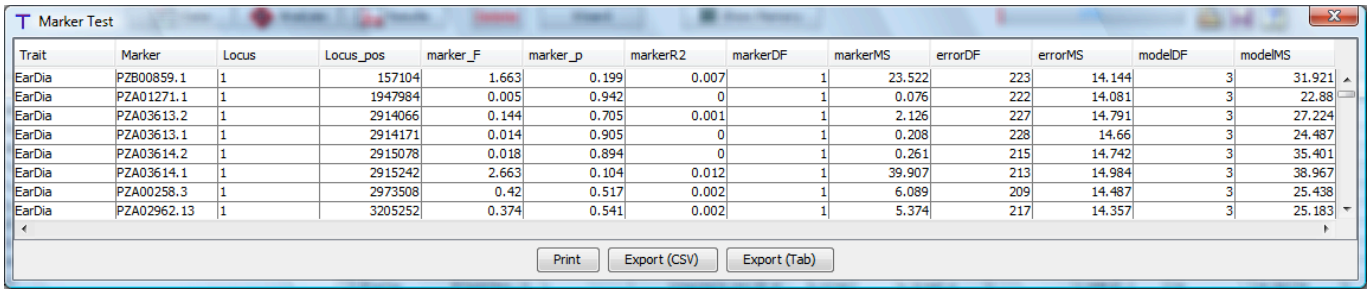
TASSEL utilizes a fixed effects linear model to test for association between segregating sites and phenotypes. The analysis optionally accounts for population structure using covariates that indicate degree of membership in underlying populations. A main effects only model is automatically built using all variable in the input data. A separate model is built and solved for each trait and marker combination. Any factors, covariates, reps or locations are included in every model as main effects. How the data is

used must be defined either in the input data files or using the **Trait Filter** after the data has been imported but before it has been joined with a genotype.

General Linear Model (GLM) can be run using a numeric data set only, numeric data joined to genotype data. If only numeric data is selected, best linear unbiased estimates (BLUEs or least square means) will be generated for the taxa for each trait. [Note: only factors and covariates intended to control field variation should be included at this stage. Population structure covariates which are intended to control for marker effects should only be included when markers are also in the analysis.] If numeric data with genotypes are analyzed, each trait by marker combination will be tested and two reports will be produced, one containing trait by marker F-tests and the other containing allele estimates.

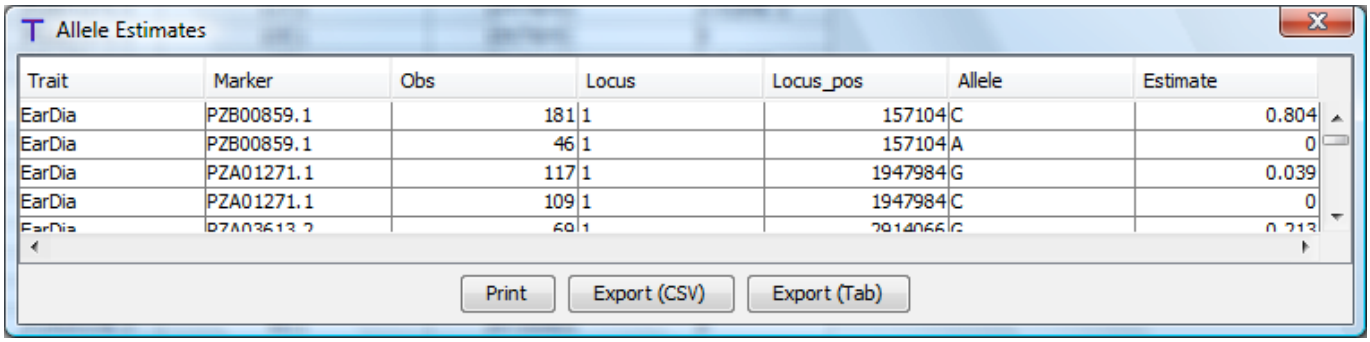
To run GLM, select a data set and then click the GLM button. A dialog box will pop-up to allow the user to indicate that a permutation test should be run and to allow the number of permutations to be changed. The permutation test will be run using the method suggested by Anderson and Ter Braak (2003), which calculates the predicted and residual values of the reduced model (contained all terms except markers) then permutes the residuals and adds them to the predicted values. When the GLM options dialog is closed, the user is presented with a dialog allowing the output to be saved to a file rather than stored in memory and displayed by TASSEL. This option is useful when the output is expected to be very large and risks exceeding available RAM.

The following table shows an example of the Marker Test output as viewed with Results/Table:



Trait	Marker	Locus	Locus_pos	marker_F	marker_p	markerR2	markerDF	markerMS	errorDF	errorMS	modelDF	modelMS
EarDia	PZB00859.1	1	157104	1.663	0.199	0.007	1	23.522	223	14.144	3	31.921
EarDia	PZA01271.1	1	1947984	0.005	0.942	0	1	0.076	222	14.081	3	22.88
EarDia	PZA03613.2	1	2914066	0.144	0.705	0.001	1	2.126	227	14.791	3	27.224
EarDia	PZA03613.1	1	2914171	0.014	0.905	0	1	0.208	228	14.66	3	24.487
EarDia	PZA03614.2	1	2915078	0.018	0.894	0	1	0.261	215	14.742	3	35.401
EarDia	PZA03614.1	1	2915242	2.663	0.104	0.012	1	39.907	213	14.984	3	38.967
EarDia	PZA00258.3	1	2973508	0.42	0.517	0.002	1	6.089	209	14.487	3	25.438
EarDia	PZA02962.13	1	3205252	0.374	0.541	0.002	1	5.374	217	14.357	3	25.183

In addition to displaying the F-statistics and p-values for the requested F-tests, the table also contains markerR2, mean squares (MS) and degrees of freedom (DF) for the marker effect, for the model (corrected for the mean), and for error. If taxa are replicated (across reps or environments), then the markers are tested using the taxa within marker mean square. If taxa are unreplicated, then the residual mean square is used. MarkerR2 is the marginal R2 for the marker calculated as SS Marker (after fitting all other model terms) / SS Total, where SS stands for sum of squares. The following table shows an example of the Allele Estimates output as viewed with Results/Table:



Trait	Marker	Obs	Locus	Locus_pos	Allele	Estimate
EarDia	PZB00859.1	181	1	157104	C	0.804
EarDia	PZB00859.1	46	1	157104	A	0
EarDia	PZA01271.1	117	1	1947984	G	0.039
EarDia	PZA01271.1	109	1	1947984	C	0
EarDia	PZA03613.2	69	1	2914066	G	0.213

For each marker and trait combination, each marker allele is listed along with the number of observations for taxa carrying that allele (Obs), the locus (usually chromosome) and locus position of that marker, the allele, and the estimate of the effect of that allele. Because of the way that GLM codes alleles, the last allele estimate for a marker is always zero and the other allele estimates are relative to that.

3.7 Mixed Linear Model 

This conducts association analysis via a mixed linear model (MLM).

A mixed model is one which includes both fixed and random effects. Including random effects gives MLM the ability to incorporate information about relationships among individuals. When a genetic marker based kinship matrix (K) is used jointly with population structure (Q), the "Q+K" approach improves statistical power compared to "Q" only[9]. MLM can be described in Henderson's matrix notation[23] as follows:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + e$$

where **y** is the vector of observations; **β** is an unknown vector containing fixed effects, including genetic marker and population structure (Q); **u** is an unknown vector of random additive genetic effects from multiple background QTL for individuals/lines; **X** and **Z** are the known design matrices; and **e** is the unobserved vector of random residual. The **u** and **e** vectors are assumed to be normally distributed with null mean and variance of

$$\mathrm{Var}\begin{pmatrix}\mathbf{u}\\ \mathbf{e}\end{pmatrix} = \begin{pmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{pmatrix}$$

where $\mathbf{G} = \square\mathbf{K}$ with $\square$ as the additive genetic variance and **K** as the kinship matrix. Homogeneous variance is assumed for the residual effect which means $\mathbf{R}=\mathbf{I}\square$, where $\square$ is the residual variance. The proportion of genetic variance over the total variance is defined as heritability (h^2).

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}$$

When K is derived from pedigrees, the elements of K equal 2*Probability(IBD), where IBD means that two alleles drawn at random are identical by descent. Generally, K calculated from markers is an IBS matrix. The resulting multiplier is then not σ_a^2 but some unknown constant times σ_a^2. Some methods for calculating K, such as those implemented in SPaGEDI, actually use markers to develop an estimate of the IBD relationship matrix. For those values of K, the resulting variance estimate can be considered an estimate of σ_a^2 as long as the assumptions of the method used to derive K are not violated for the population being analyzed. One implication is that two different K matrices may give very different estimates of σ_a and heritability yet produce the same model fit and test of marker association.

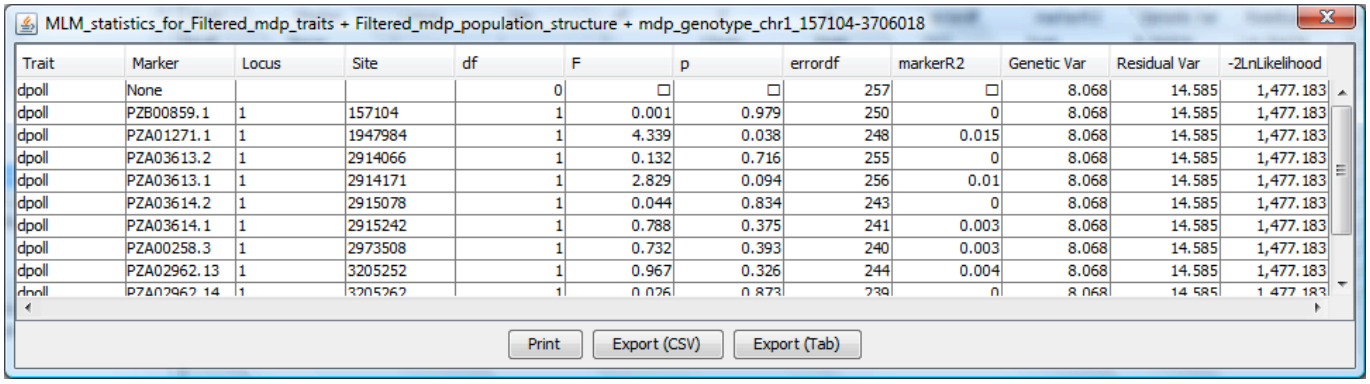
TASSEL implements several methods to improve statistical power and reduce computing time. The Restricted Maximum Likelihood (REML) estimates of $\acute{o}_a^2$ and $\acute{o}_e^2$ are obtained through the Efficient Mixed-Model Association (EMMA) algorithm[24] which is much faster than the expectation and maximization (EM) algorithm[25].

TASSEL also implements a method called compression which reduces the dimensionality of the kinship matrix to reduce computational time and improve model fitting. When MLM is used without compression (compression = 1), each taxon belongs to its own group. At the other extreme, GLM can be interpreted as maximum compression (compression = n) with all taxa in a single group. In that case, it is not possible to estimate the random effect independently of error and $ó^2_a$ is absorbed into $ó^2_e$. Between these two extremes, taxa can be grouped using cluster analysis based on kinship. When n individuals are compressed into s clusters (groups), the kinship among individuals is replaced with the kinship among groups. At some grouping levels, dependent on the trait and population being analyzed, this compressed MLM has improved statistical power compared to the regular MLM[4]. The optimum grouping with the best model fit for MLM without fitting genetic markers has the best statistical power for an association test of markers[4]. TASSEL allows users to specify the compression level (average number of individuals per group), or to have the program determine the optimum grouping.

Similar to GLM, MLM performs an association test for each combination of traits and markers. TASSEL provides users several options: 1) to estimate genetic and residual variance for each combination; 2) to get these estimates once for each trait without fitting genetic markers and then to use those estimates to test markers; 3) to use a prior heritability estimate provided by the user. The second option, named P3D (population parameters previously determined), has the same statistical power as the first option[4]. Using the P3D method or using a prior heritability can be much faster than calculating heritability for each marker.

Using MLM is very similar to using GLM. The difference is that in addition to choosing the joint data set (or numerical data set), kinship data must also be highlighted before clicking the MLM button to show the MLM option dialog. The option of "No Compression" is the regular MLM which is equivalent to "Custom level=1". For data sets with large numbers of taxa, the optimal compression option may be considerably slower than no compression or user supplied compression. This is because the algorithm solves the model once for each of a series of compression levels in order to determine the optimal one.

All MLM analyses create two output tables, model statistics and model effects. If compression is used, the analysis creates three tables.

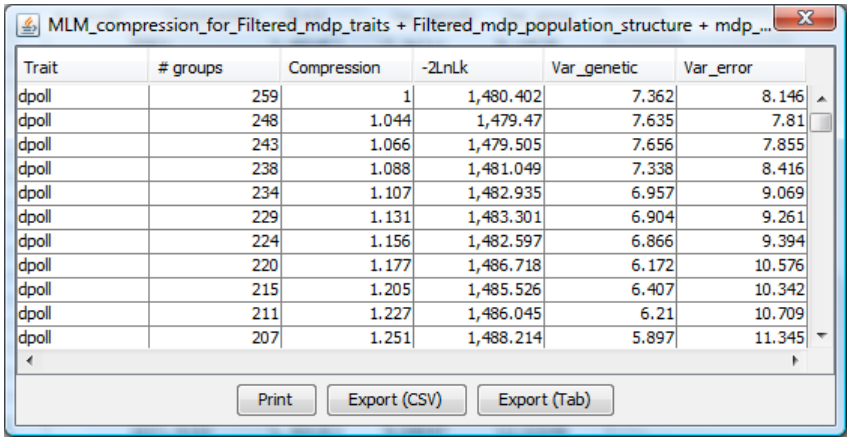


Trait	Marker	Locus	Site	df	F	p	errordf	markerR2	Genetic Var	Residual Var	-2LnLikelihood
dpoll	None			0			257		8.068	14.585	1,477.183
dpoll	PZB00859.1	1	157104	1	0.001	0.979	250	0	8.068	14.585	1,477.183
dpoll	PZA01271.1	1	1947984	1	4.339	0.038	248	0.015	8.068	14.585	1,477.183
dpoll	PZA03613.2	1	2914066	1	0.132	0.716	255	0	8.068	14.585	1,477.183
dpoll	PZA03613.1	1	2914171	1	2.829	0.094	256	0.01	8.068	14.585	1,477.183
dpoll	PZA03614.2	1	2915078	1	0.044	0.834	243	0	8.068	14.585	1,477.183
dpoll	PZA03614.1	1	2915242	1	0.788	0.375	241	0.003	8.068	14.585	1,477.183
dpoll	PZA00258.3	1	2973508	1	0.732	0.393	240	0.003	8.068	14.585	1,477.183
dpoll	PZA02962.13	1	3205252	1	0.967	0.326	244	0.004	8.068	14.585	1,477.183
dpoll	PZA02962.14	1	3205262	1	0.026	0.873	239	0	8.068	14.585	1,477.183

The statistics table shows the results of the tests for each trait. The first line is for the model with no markers. Following that is a single line for each marker tested. The columns labeled "df", "F", and "p" are the degrees of freedom, F, and p-value from the F distribution for the test of the marker. The column "errordf" is the degrees of freedom used for the denominator of the F-test. The column labeled "markerR2" is the R2 for the marker calculated based on a formula for R2 for a generalized least squares GLS) model as shown here.

$$R^2 = \frac{\left(\hat{Y}_{full} - \hat{Y}_{reduced}\right)^T V^{-1} \left(\hat{Y}_{full} - \hat{Y}_{reduced}\right)}{\left(Y - \bar{Y}\right)^T V^{-1} \left(Y - \bar{Y}\right)}$$

The columns "Genetic Var", Residual Var", and "-2LnLikelihood" list σ2a, σ2e, and minus two times the model likelihood, respectively. When the P3D option is used, all of the values are the same for a given trait because they are only calculated once. A second table lists the estimated effects of each allele for each marker similar to the output for GLM. The compression results table shown below shows the likelihood, genetic variance, and error variance for each compression level tested during the optimization process. The meaning of groups and compression is discussed above in the description of the compression method. The compression level with the lowest value of -2LnLk is used for testing markers.



Trait	# groups	Compression	-2LnLk	Var_genetic	Var_error
dpoll	259	1	1,480.402	7.362	8.146
dpoll	248	1.044	1,479.47	7.635	7.81
dpoll	243	1.066	1,479.505	7.656	7.855
dpoll	238	1.088	1,481.049	7.338	8.416
dpoll	234	1.107	1,482.935	6.957	9.069
dpoll	229	1.131	1,483.301	6.904	9.261
dpoll	224	1.156	1,482.597	6.866	9.394
dpoll	220	1.177	1,486.718	6.172	10.576
dpoll	215	1.205	1,485.526	6.407	10.342
dpoll	211	1.227	1,486.045	6.21	10.709
dpoll	207	1.251	1,488.214	5.897	11.345

3.8 Ridge Regression 

This function performs ridge regression to predict phenotypes from genotypes. It is one of the methods used for genomic selection (GS).

The input dataset must contain one or more phenotypes and numeric marker data. Optionally, it may also contain factors and covariates. The analysis is run by selecting the input dataset then clicking the "GS" button. Because no additional user input is needed, the analysis will run immediately after the button is clicked. All traits will be analyzed separately using all of the genotypes, factors, and covariates in the dataset. The output will consist of two new datasets for each trait. One of the datasets will contain genomic estimated breeding values (GEBVs) for each taxon and the other will contain BLUPs for each marker in the genotype file. The output datasets will appear in the "Numerical" folder, which holds the input data as well. The output datasets can in turn be used for subsequent analysis. For example, it could be joined with the input data so that the predicted values could be graphed against the original values.

Understanding the input data requirements is important to ensure that the results of the analysis will be correct and useful. Genotypes must be numeric with one column for each marker. It is expected that the markers are bi-allelic, with the homozygotes coded as 1 and -1 and the heterozygotes coded as 0. However, any reasonable coding scheme will work. For instance, missing data could be replaced by a probability resulting from imputation. If any genotype data is missing, it will be imputed as the average of

the marker scores across all taxa for that marker. If a user prefers to use a different method of imputation, then the missing genotypes must be imputed before importing the data into TASSEL.

GEBVs will be calculated for all taxa in the dataset, including any lines that have missing phenotype data. A typical use of genomic selection is to predict GEBVs for a set of unphenotyped lines based on the performance of a training set. To do that a dataset containing both the genotypes to be predicted and the genotypes of the training set can be joined with a dataset containing the phenotypes of the training set using a union join. All taxa in the phenotype set should have genotypes. If an individual without genotype data is included, all the marker data for that individual will be imputed, which is not a generally useful thing to do.

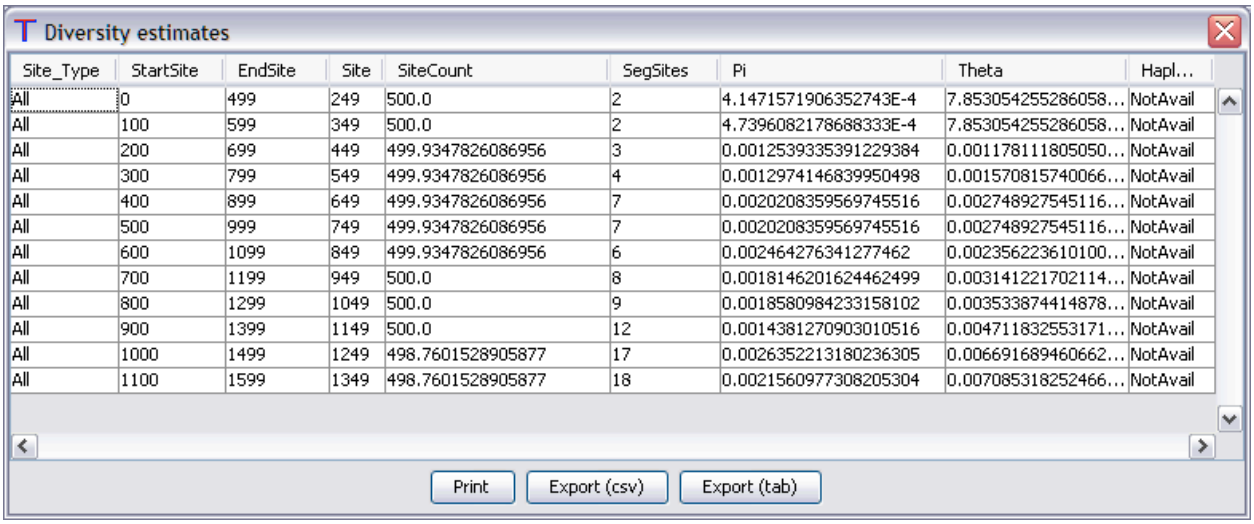
4 Result Mode 

Results mode consists of the functions to present data as table or graphics.

4.1 Table 

Allows data to be displayed in a spreadsheet view and exported into a flat file.

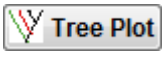
To create a table, select a data set from the Data Tree panel, then click on the "Results" button followed by the "Table" button (**Results → Table**). Shown below is an example in which diversity estimates are displayed.



Site_Type	StartSite	EndSite	Site	SiteCount	SegSites	Pi	Theta	Hapl...
All	0	499	249	500.0	2	4.1471571906352743E-4	7.853054255286058...	NotAvail
All	100	599	349	500.0	2	4.7396082178688333E-4	7.853054255286058...	NotAvail
All	200	699	449	499.9347826086956	3	0.0012539335391229384	0.001178111805050...	NotAvail
All	300	799	549	499.9347826086956	4	0.0012974146839950498	0.001570815740066...	NotAvail
All	400	899	649	499.9347826086956	7	0.0020208359569745516	0.002748927545116...	NotAvail
All	500	999	749	499.9347826086956	7	0.0020208359569745516	0.002748927545116...	NotAvail
All	600	1099	849	499.9347826086956	6	0.002464276341277462	0.002356223610100...	NotAvail
All	700	1199	949	500.0	8	0.0018146201624462499	0.003141221702114...	NotAvail
All	800	1299	1049	500.0	9	0.0018580984233158102	0.003533874414878...	NotAvail
All	900	1399	1149	500.0	12	0.0014381270903010516	0.004711832553171...	NotAvail
All	1000	1499	1249	498.7601528905877	17	0.0026352213180236305	0.006691689460662...	NotAvail
All	1100	1599	1349	498.7601528905877	18	0.0021560977308205304	0.007085318252466...	NotAvail

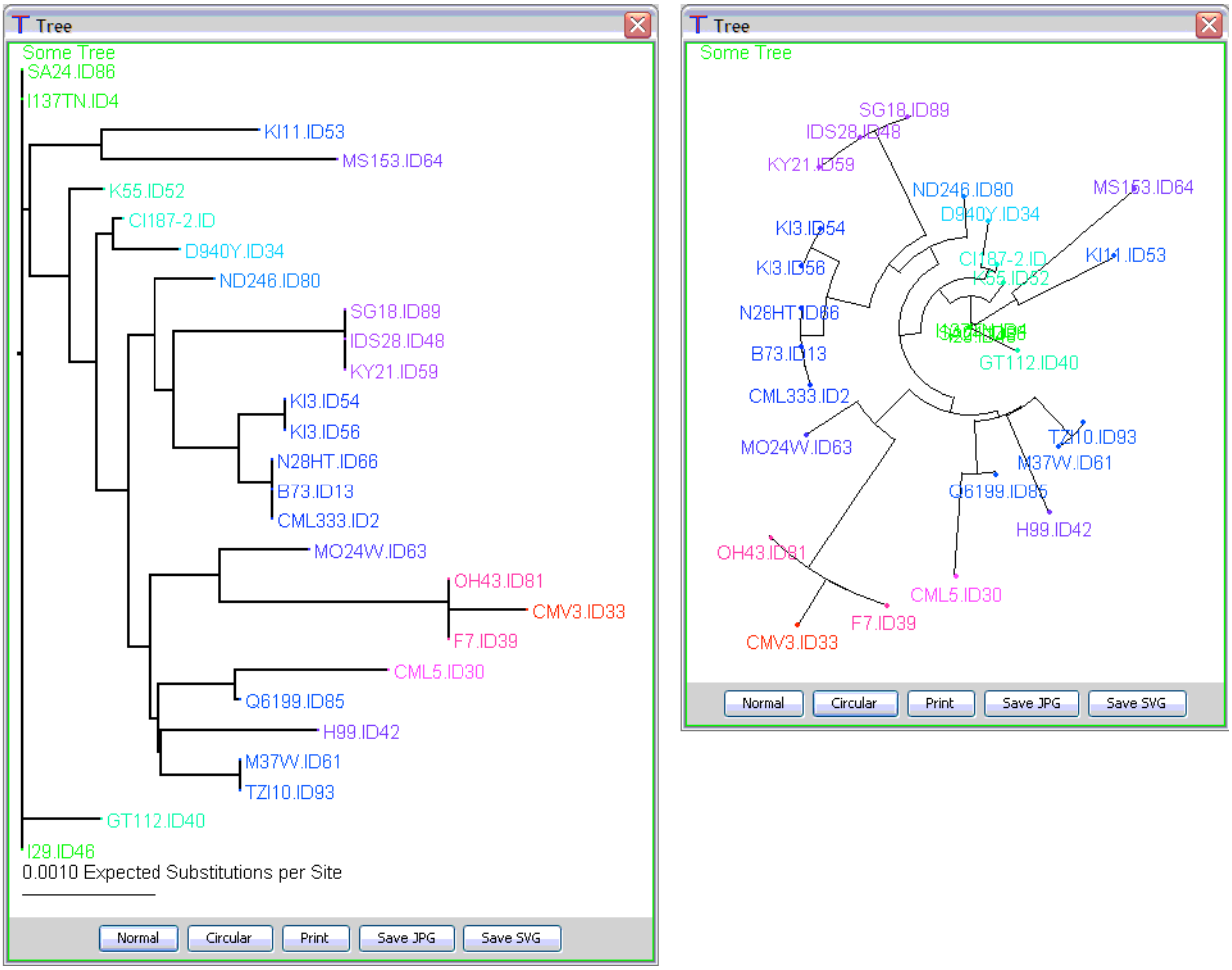
Data can be sorted by clicking on the column header of interest. A secondary sort can be done by holding down the CTRL key and clicking on a second column.

Data can be exported to flat files that are either comma-separated (Comma Separated Values = CSV) or tab-delimited. Both these formats can then be imported into a spreadsheet program such as Excel. Tables can also be printed.

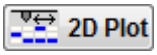
4.2 Tree Plot 

Displays the results of cladogram analysis.

After running **Analysis → Cladogram**, select the desired data set and then click **Tree Plot** in the **Results** mode (Results → Tree Plot). Trees can be visualized in either a **Normal** or **Circular** layout.



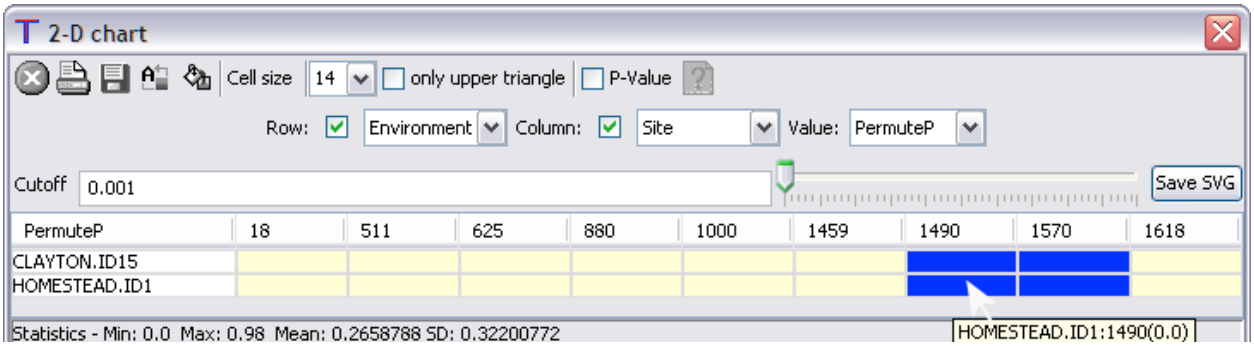
These images can be printed, saved in JPEG format, or saved as a Scalable Vector Graphics (SVG) file.

4.3 2D Plot 

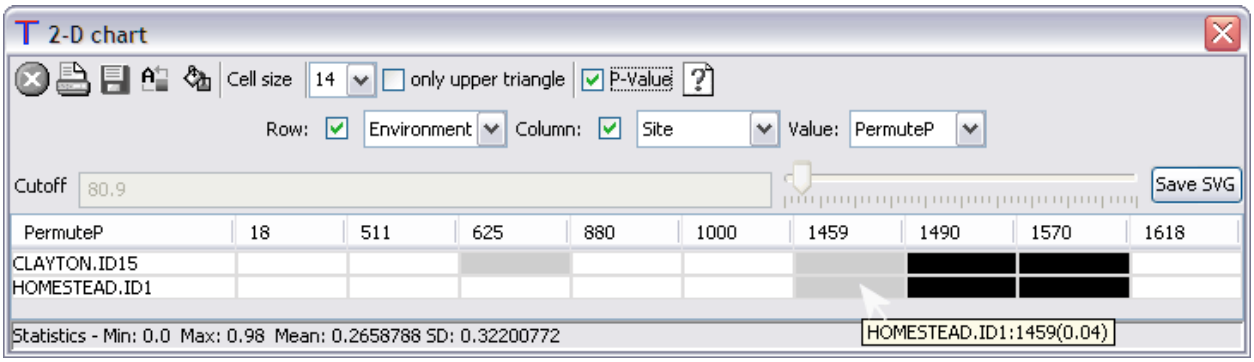
Displays 2D plots and determines color thresholds.

This function is useful for plotting associations in multiple environments.

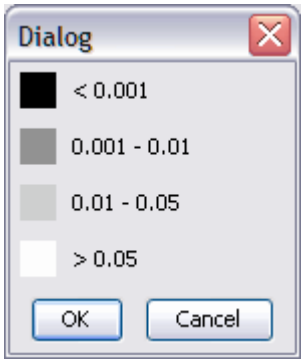
First, select the desired result set. Using the drop down boxes provided, populate rows with "Environment," columns with "Site," and value with "PermuteP." The cutoff value for coloring can be chosen either by inputting a value in the text box or by using the slider tool to the right of the text box. Users can "mouse over" any box to view the value associated with that box, as shown here:



If P-value coloring is desired, simply check the P-value box as shown below:



By checking the P-value box, Cutoff selection tools will be disabled and fields will instead be colored according to the following grayscale:



This key can be shown by clicking on the "?" icon next to the P-value check box.

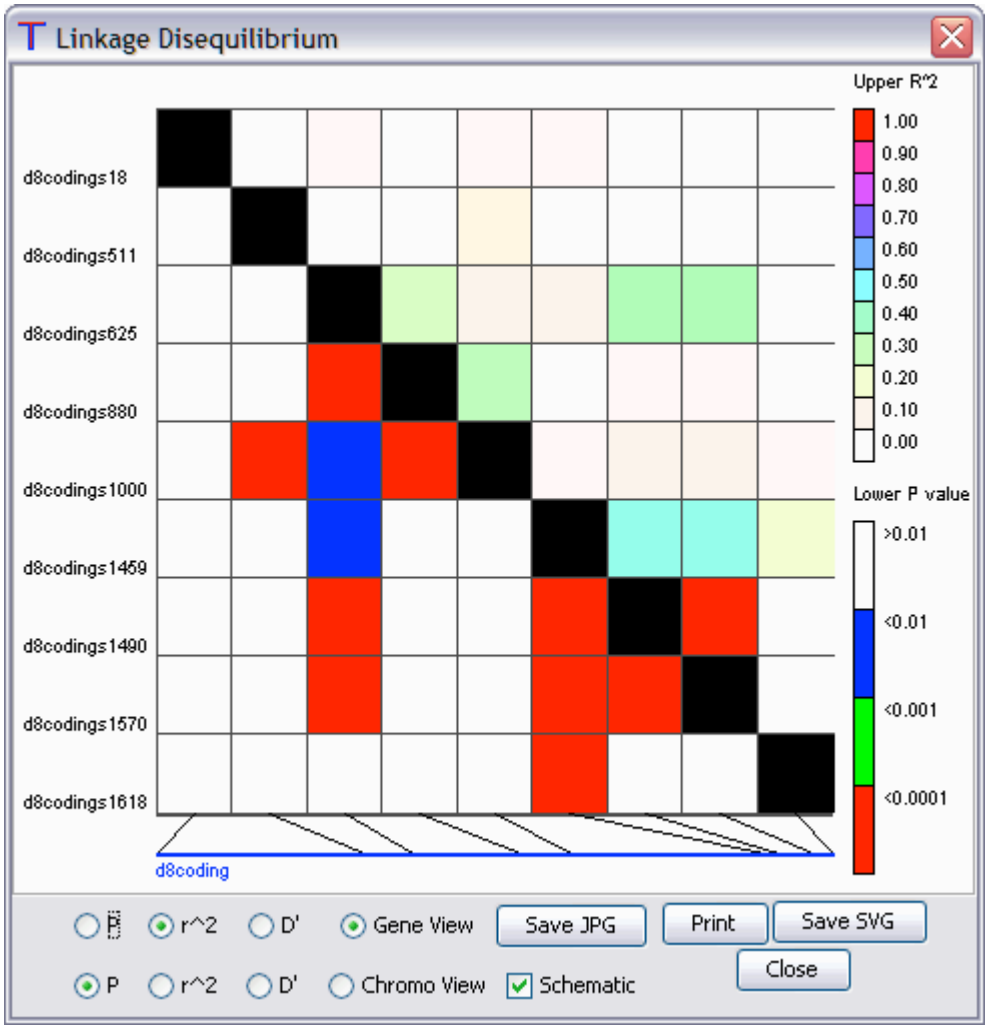
4.4 LD Plot 

Displays the results from the linkage disequilibrium analysis.

After selecting the desired result from the Data Tree panel, click on the "LD Plot" button while in "Results" mode (**Results → LD Plot**).

The graph that is generated displays LD between all possible pairs of sites. The black diagonal represents LD between each site and itself. The default setting graphs r^2 in the upper right and ***p*-values** in the lower left. This default can be modified by clicking on the radio buttons in the lower left. The left side of the

graph contains a text description of the gene (or chromosome) and the site within the gene (or genetic position within the chromosome). At the bottom of the graph is a display of the position of each site along the gene or chromosome. This display can be hidden by deselecting the "**Schematic**" checkbox. Legends describing the color scheme appear on the right hand side of the graph.



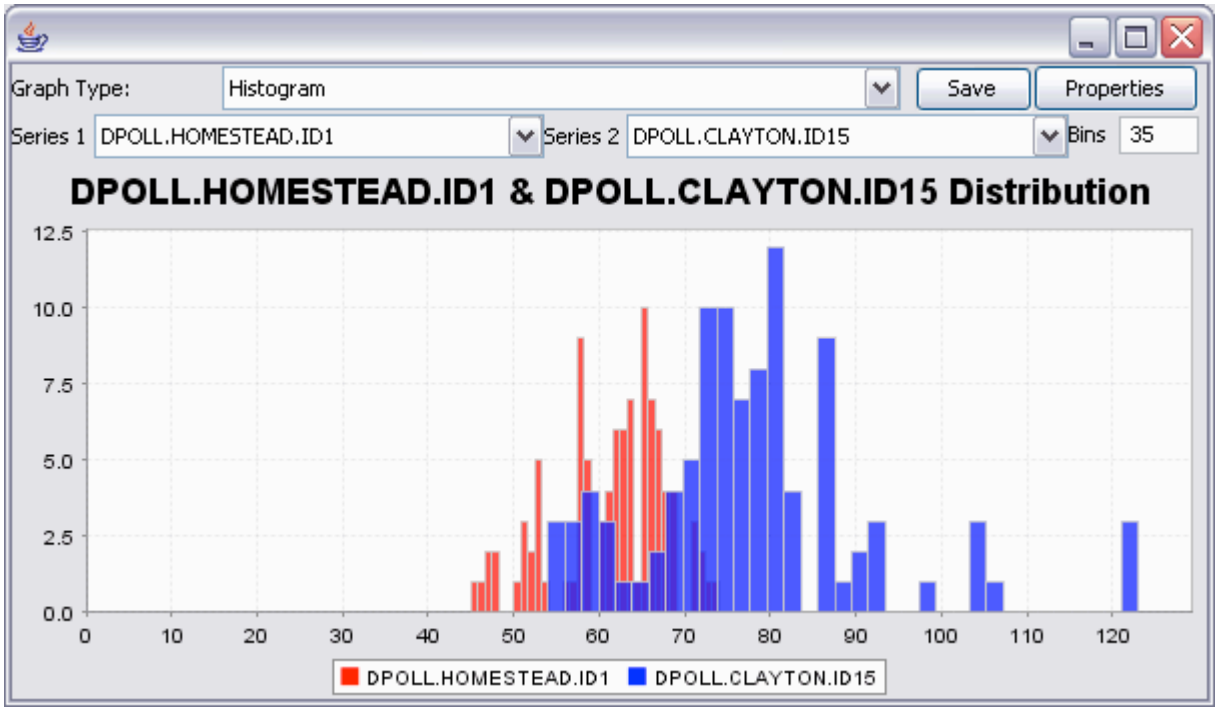
LD plots can be printed, saved in JPEG format, or saved as a Scalable Vector Graphics (SVG) file. An SVG file is useful for creating publication quality graphics which can be easily sized using an editor such as Adobe Illustrator, Corel Draw, or OpenOffice.org Draw 2.0+.

4.5 Chart Chart

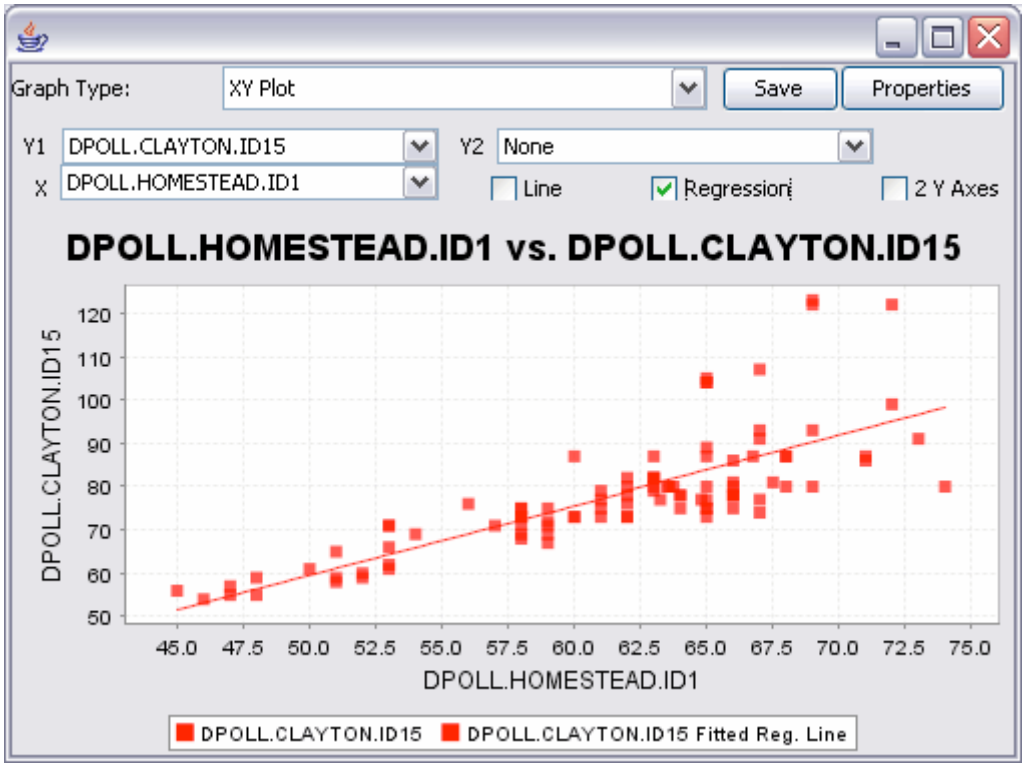
Chart provides a variety of graphs for visualizing numeric data.

This feature can be used to display histograms, XY plots, bar charts and/or pie charts. Any numeric table data can be charted, including LD results, phenotypic data, diversity results, and association results.

Histograms: Use the graph type combo box to select the desired graph type (Histogram) from the list of options. Up to two different series of data can be plotted together. Users may specify the number of bins to be used in the histogram.



Scatter plots: Use the graph type combo box to select the desired graph type (XY Plot) from the list of options. Select data to be plotted in X and Y axes using the appropriate drop down boxes. If two data series are plotted simultaneously on the Y axis, the "2 Y Axes" checkbox will provide an axis for each.



5 Menus

The menus in TASSEL include File, Tools, GDPC, and Help menus. The File menu is mainly used to save the entire data tree which includes the data loaded into TASSEL and the data created within TASSEL. A previously saved data tree can be loaded to TASSEL. This function provides the users the capability to save their intermediate results. The tools menu contains contingency test and option to set preference.

GDPC (Genomic Diversity and Phenotype Connection) is a software package to retrieve data from open database sources such as SNPs and phenotypic data. It can also be started using the "GDPC" button in data mode. Its use is described earlier in the manual.

5.1 File Menu

Individual data sets on the data tree and the entire data tree can be saved. An individual data set is saved in the genotype format for sequence data or numerical format for phenotype, covariate, and kinship. The data tree is saved in a binary format.

5.1.1 Save Data Tree

This feature allows you to save the entire contents of the Data Tree panel to a default location. This is helpful when the user does not wish to recreate a Data Tree panel that is already well populated with information the next time they initializes the program. To save a Data Tree, select **File > Save Data Tree**.

5.1.2 Open Data Tree

To restore a Data Tree that was saved previously saved, select **File > Open Data Tree**.

5.1.3 Save Data Tree As…

To save the contents of a Data Tree to a specific location or to give it a specific name, select **File > Save Data Tree As….**

5.1.4 Open Data Tree…

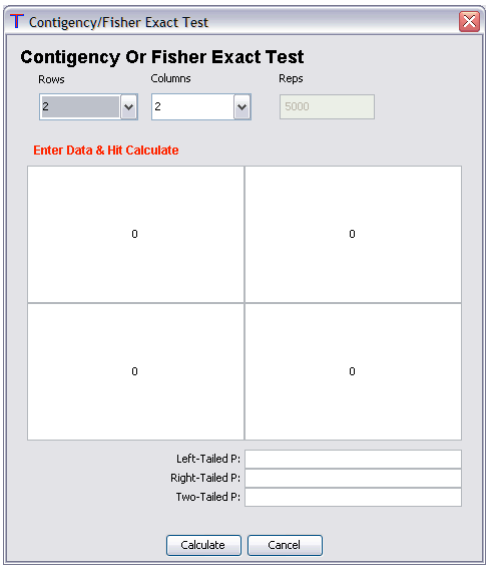
To restore a Data Tree from a specific location, select **File** > **Open Data Tree…**

NOTE: The information outlined above for saving a Data Tree is applicable to files that are, in general, version specific. When a new version of TASSEL is released, a data tree saved with a previous version might not load to the version. For longer term storage, the best practice is to save individual data sets rather than the entire data tree.

5.1.5 Save Selected As…

To export data to one of the supported file types, select **File** > **Save Selected As…**

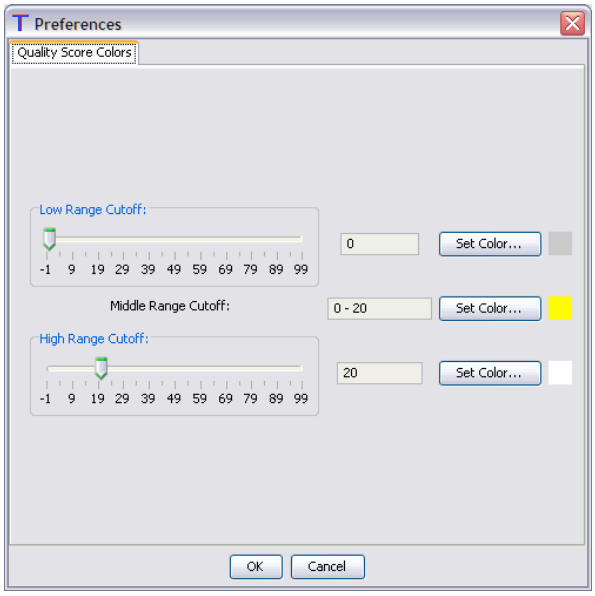
5.2 Contingency Test



This utility calculates a chi-square contingency test or Fisher exact test (when using only the 2 x 2 table of observations) using the same algorithm as is used in determining linkage disequilibrium.

5.3 Preferences

The **Quality Score Colors** tab, found in the **Preferences** dialog box, allows the user to set cutoff values for visualizing quality score values on a sequence alignment or a set of called SNPs.



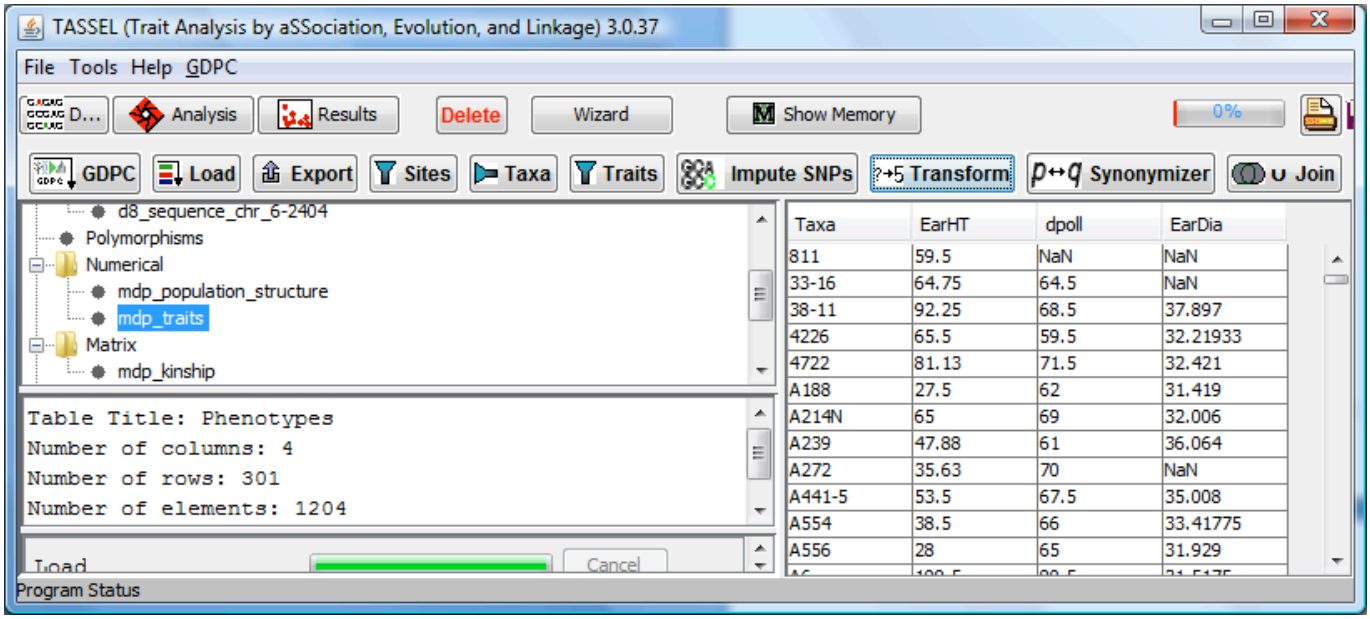
To set a desired threshold, simply adjust the slider on the left side of the dialog. Ns, "-" (dashes), and alignments without any quality score information have a default value of -1 (minus one).

6 Tutorial

This tutorial reviews several common scenarios for using TASSEL in order to help the user better understand its capabilities for data manipulation and association analyses. The TASSEL software package includes a tutorial data set that can be downloaded from the TASSEL website (please unzip all files to a directory of your choice). This tutorial data set contains data for phenotype, genotype, population structure, and kinship.

6.1 Missing Phenotype Imputation

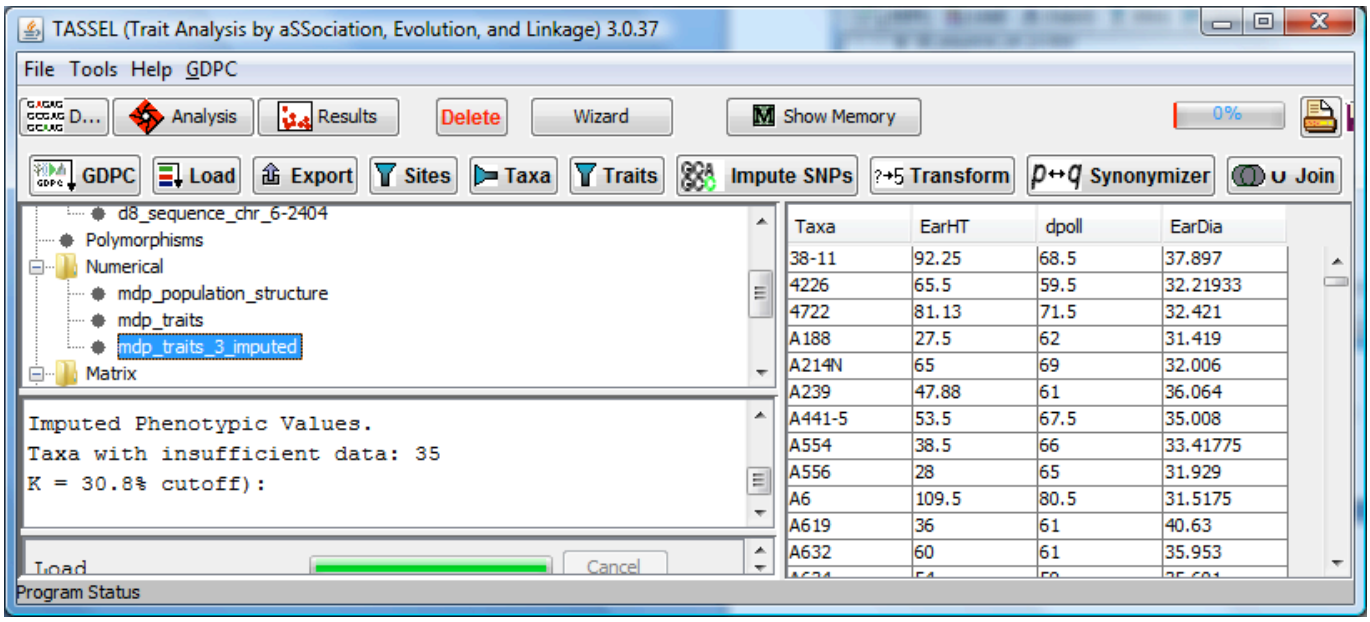
The phenotype file **mdp_traits** will be used to demonstrate the process of imputing missing data. Note that the data set below contains missing values (NaN).



Taxa	EarHT	dpoll	EarDia
811	59.5	NaN	NaN
33-16	64.75	64.5	NaN
38-11	92.25	68.5	37.897
4226	65.5	59.5	32.21933
4722	81.13	71.5	32.421
A188	27.5	62	31.419
A214N	65	69	32.006
A239	47.88	61	36.064
A272	35.63	70	NaN
A441-5	53.5	67.5	35.008
A554	38.5	66	33.41775
A556	28	65	31.929

To impute missing data, first select the **mdp_traits** data set in the Data Tree Panel and then click the **Transform** button (**Data → Transform**). The "Transform Column Data" window will open. Click on the **Impute** tab in this window. Finally, click on the **Create Data set** button to create the new data set with missing values imputed.

Note that missing values are now filled.

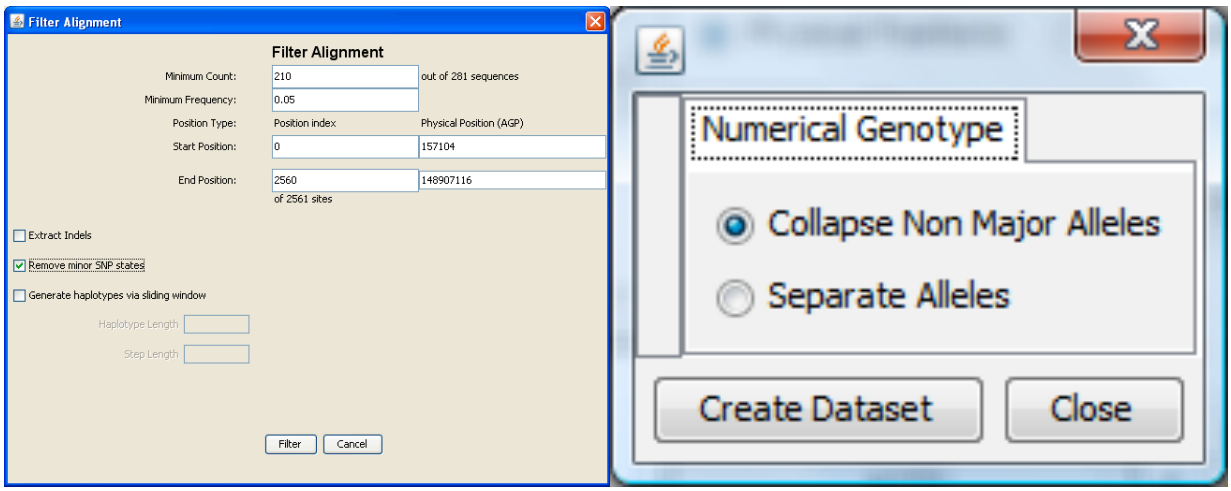


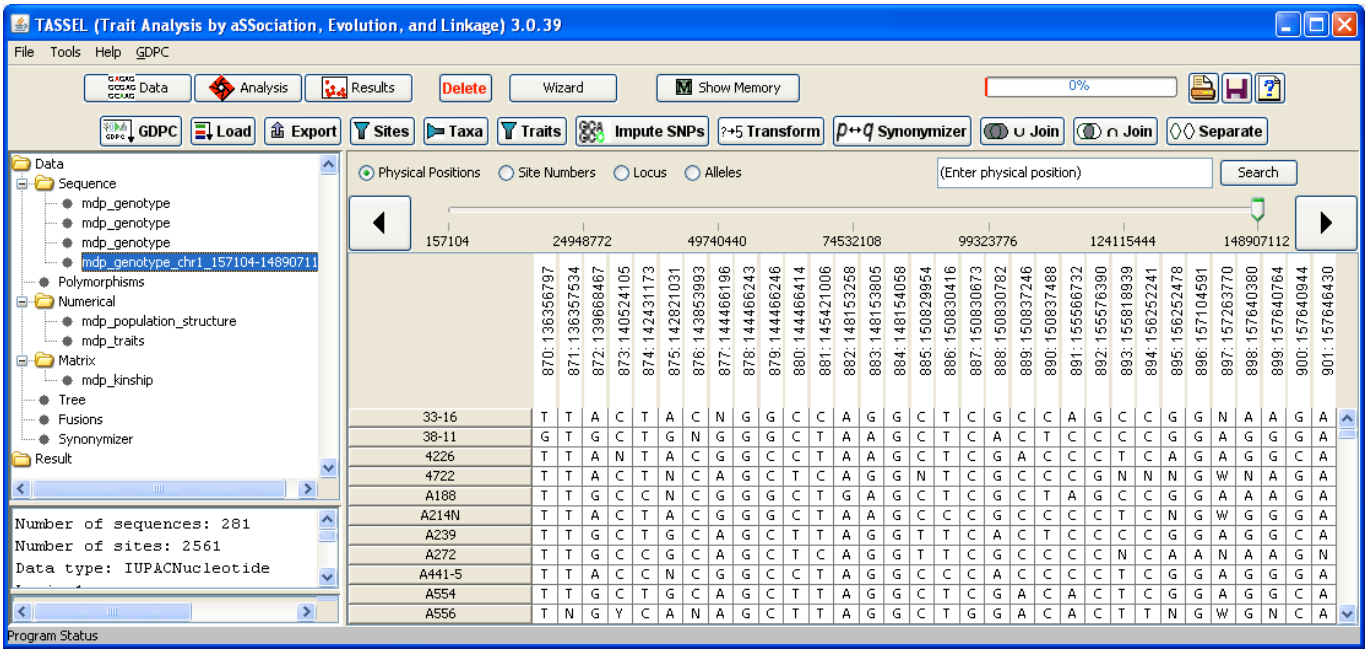
Taxa	EarHT	dpoll	EarDia
38-11	92.25	68.5	37.897
4226	65.5	59.5	32.21933
4722	81.13	71.5	32.421
A188	27.5	62	31.419
A214N	65	69	32.006
A239	47.88	61	36.064
A441-5	53.5	67.5	35.008
A554	38.5	66	33.41775
A556	28	65	31.929
A6	109.5	80.5	31.5175
A619	36	61	40.63
A632	60	61	35.953

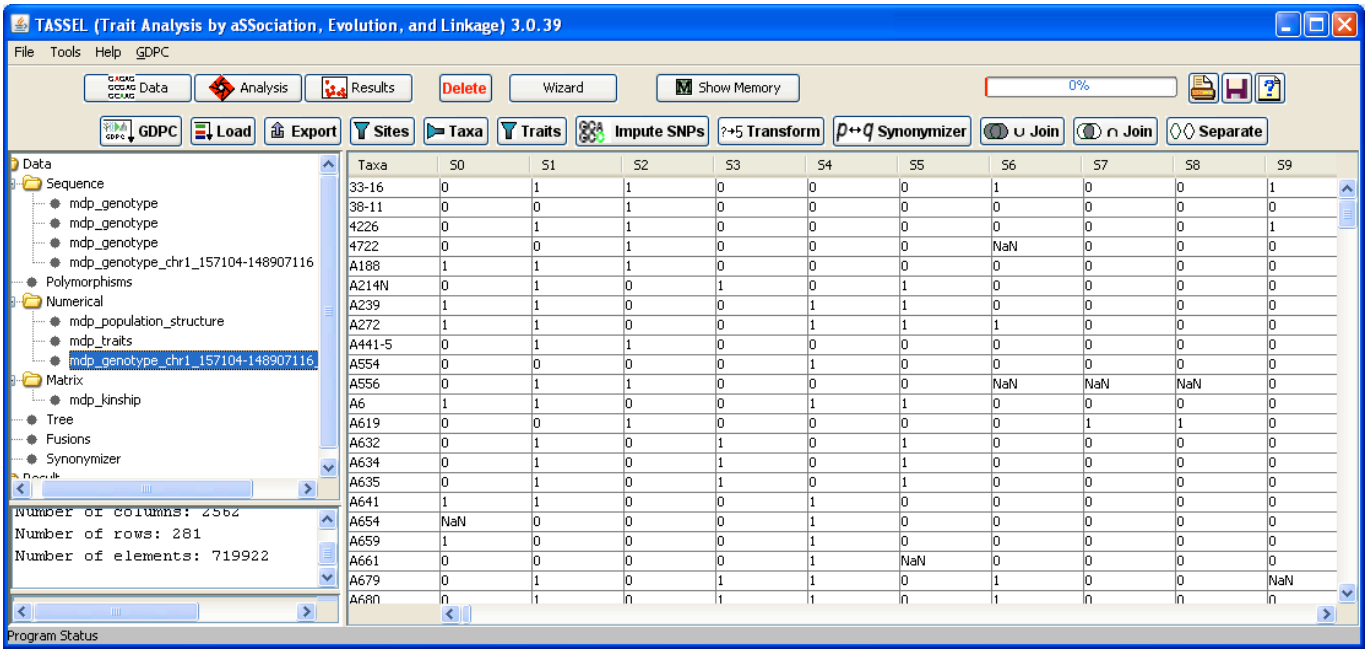
6.2 Principal Component Analysis

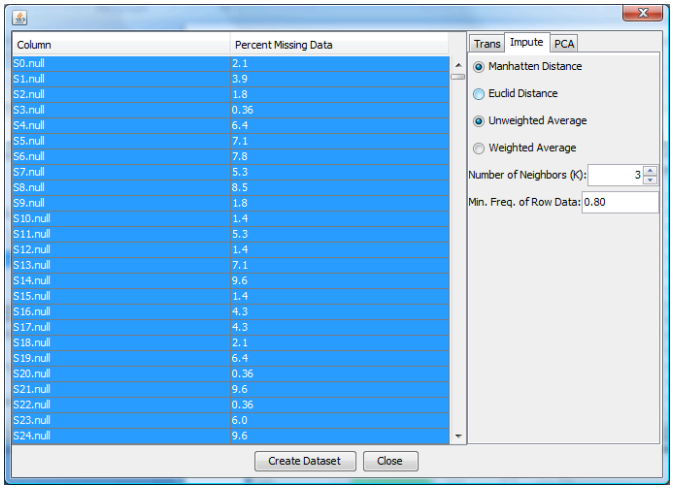
Principal component analysis (PCA) is a statistical tool that transforms a set of correlated variables into a smaller number of uncorrelated variables called principal components (PCs). The first PC captures as much of the variation as possible, and the succeeding PCs account for a decreasing fraction of the remaining variance. Another application of PCA is to use PCs derived from genetic markers to represent population structure[8]. This method requires much less computing time than maximum likelihood estimation. As most marker data are characters, numericalization must be performed first. A common approach for converting character marker scores is to set one of the homozygotes to 0, the other homozygote to 2, and the heterozygote to 1. For haploids, the conversion can be simply performed by coding one allele as 0 and the other as 1. The TRANSFORM function in TASSEL converts the major allele to 0. All the other alleles are collapsed to a single class and coded as 1. PCA requires that all variables should have variation and should not have missing values. As a result, filtering genotype to eliminate monomorphic markers and imputing missing values may be necessary. Imputing missing values can be done before or after numericalization. Here we demonstrate how to generate PCs from the genotype file in the tutorial data.

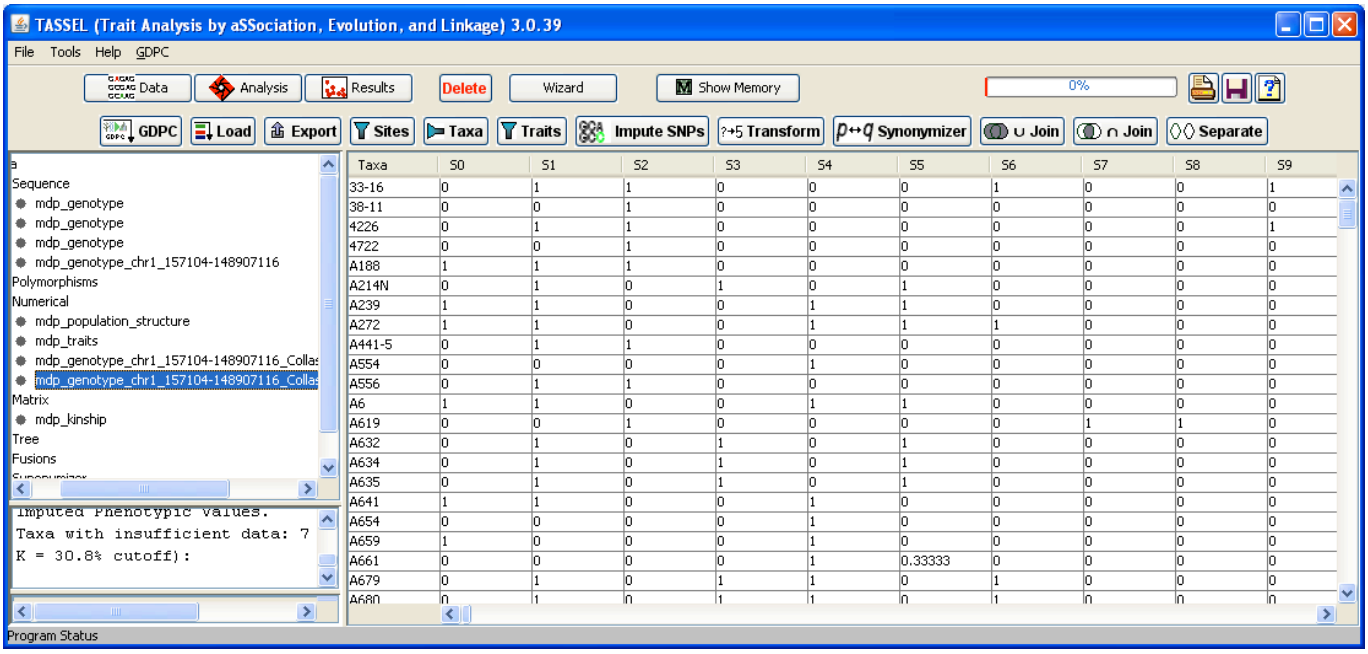
1. Remove monomorphic sites: Make sure TASSEL is in **Data** mode. Highlight the genotype and click **Site**. Set the minimum frequency to 0.05 and have "Remove minor SNP status" checked. Click **Filter**.
2. Numericalization: Highlight the filtered genotype and click **Transform**. Use the default option of "Collapse non major alleles." Click **Create data set**.
3. Imputation of missing values: Highlight the numerical genotype and click **Transform** and then click **Impute** Tab. Use the default options. Click **Create data set**.
4. PCA: Highlight the imputed numerical genotype, click **Transform,** and then click **PCA** Tab. Change the default option to "Components=3" by choosing **Components** and type 3 in the text box. Click **Create data set**.

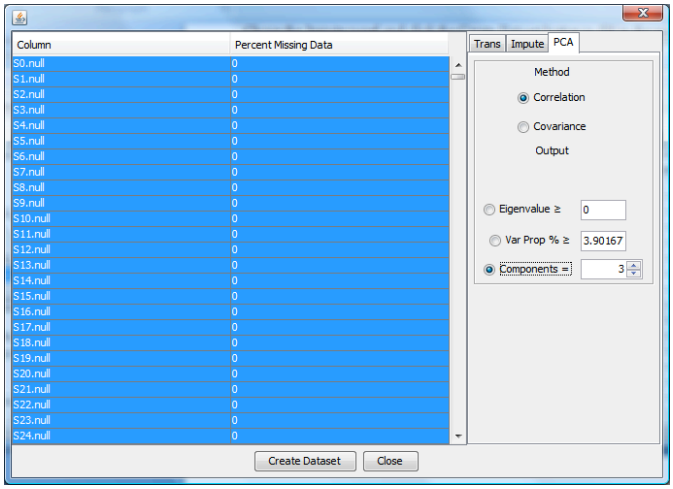



TASSEL (Trait Analysis by aSSociation, Evolution, and Linkage) 3.0.39
File Tools Help GDPC
Data Analysis Results Delete Wizard Show Memory 0%
GDPC Load Export Sites Taxa Traits Impute SNPs ?→5 Transform p↔q Synonymizer ∪ Join ∩ Join Separate
Data
Sequence
mdp_genotype
mdp_genotype
mdp_genotype
mdp_genotype_chr1_157104-14890711
Polymorphisms
Numerical
mdp_population_structure
mdp_traits
Matrix
mdp_kinship
Tree
Fusions
Synonymizer
Result
Number of sequences: 281
Number of sites: 2561
Data type: IUPACNucleotide
Physical Positions Site Numbers Locus Alleles
(Enter physical position) Search
157104 24948772 49740440 74532108 99323776 124115444 148907112
870:136356797 871:136357534 872:139668467 873:140524105 874:142431173 875:142821031 876:143853993 877:144466196 878:144466243 879:144466246 880:144466414 881:145421006 882:148153258 883:148153805 884:148154058 885:150829954 886:150830416 887:150830673 888:150830782 889:150837246 890:150837488 891:155566732 892:155576390 893:155818939 894:156252241 895:156252478 896:157104591 897:157263770 898:157640380 899:157640764 900:157640944 901:157646430
33-16 T T A C T A C N G G C C A G G C T C G C C A G C C G G N A A G A
38-11 G T G C T G N G G G C T A A G C T C A C T C C C C G G A G G G A
4226 T T A N T A C G G C C T A A G C T C G A C C C T C A G A G G C A
4722 T T A C T N C A G C T C A G G N T C G C C C G N N N G W N A G A
A188 T T G C C N C G G G C T G A G C T C G C T A G C C G G A A A G A
A214N T T A C T A C G G G C T A A G C C C G C C C C T C N G W G G G A
A239 T T G C T G C A G C T T A G G T T C A C T C C C C G G A G G C A
A272 T T G C C G C A G C T C A G G T T C G C C C C N C A A N A A G N
A441-5 T T A C C N C G G C C T A G G C C C A C C C C T C G G A G G G A
A554 T T G C T G C A G C T T A G G C T C G A C A C T C G G A G G C A
A556 T N G Y C A N A G C T T A G G C T G G A C A C T T N G W G N C A
Program Status

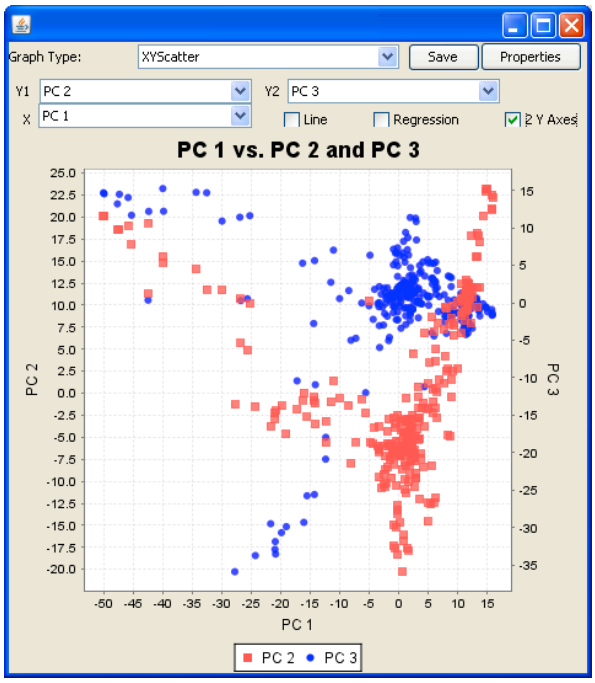

TASSEL (Trait Analysis by aSSociation, Evolution, and Linkage) 3.0.39
File Tools Help GDPC
Data Analysis Results Delete Wizard Show Memory 0%
GDPC Load Export Sites Taxa Traits Impute SNPs ?→5 Transform p↔q Synonymizer ∪ Join ∩ Join Separate
Data
Sequence
mdp_genotype
mdp_genotype
mdp_genotype
mdp_genotype_chr1_157104-148907116
Polymorphisms
Numerical
mdp_population_structure
mdp_traits
mdp_genotype_chr1_157104-148907116
Matrix
mdp_kinship
Tree
Fusions
Synonymizer
Number of columns: 2562
Number of rows: 281
Number of elements: 719922
Taxa S0 S1 S2 S3 S4 S5 S6 S7 S8 S9
33-16 0 1 1 0 0 0 1 0 0 1
38-11 0 0 1 0 0 0 0 0 0 0
4226 0 1 1 0 0 0 0 0 0 1
4722 0 0 1 0 0 0 NaN 0 0 0
A188 1 1 1 0 0 0 0 0 0 0
A214N 0 1 0 1 0 1 0 0 0 0
A239 1 1 0 0 1 1 0 0 0 0
A272 1 1 0 0 1 1 1 0 0 0
A441-5 0 1 1 0 0 0 0 0 0 0
A554 0 0 0 0 1 0 0 0 0 0
A556 0 1 1 0 0 0 NaN NaN NaN 0
A6 1 1 0 0 1 1 0 0 0 0
A619 0 0 1 0 0 0 0 1 1 0
A632 0 1 0 1 0 1 0 0 0 0
A634 0 1 0 1 0 1 0 0 0 0
A635 0 1 0 1 0 1 0 0 0 0
A641 1 1 0 0 1 0 0 0 0 0
A654 NaN 0 0 0 1 0 0 0 0 0
A659 1 0 0 0 1 0 0 0 0 0
A661 0 0 0 0 1 NaN 0 0 0 0
A679 0 1 0 1 1 0 1 0 0 NaN
Program Status

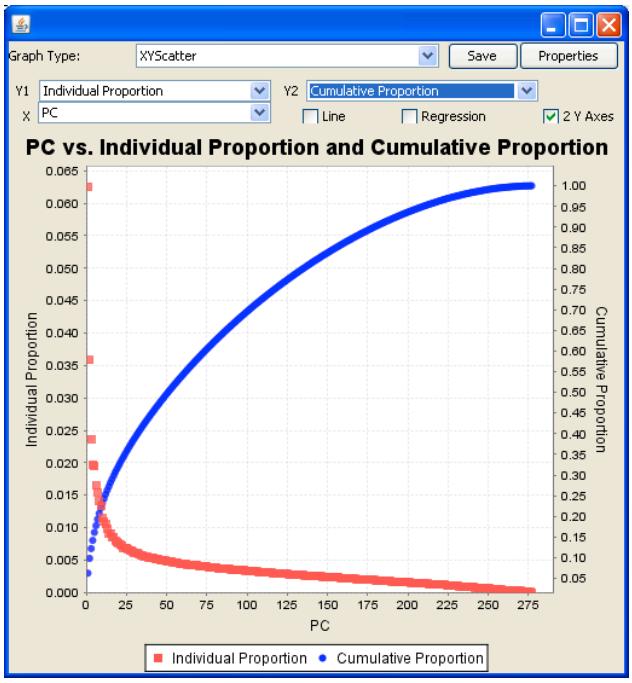

Column
Percent Missing Data
S0.null 2.1
S1.null 3.9
S2.null 1.8
S3.null 0.36
S4.null 6.4
S5.null 7.1
S6.null 7.8
S7.null 5.3
S8.null 8.5
S9.null 1.8
S10.null 1.4
S11.null 5.3
S12.null 1.4
S13.null 7.1
S14.null 9.6
S15.null 1.4
S16.null 4.3
S17.null 4.3
S18.null 2.1
S19.null 6.4
S20.null 0.36
S21.null 9.6
S22.null 0.36
S23.null 6.0
S24.null 9.6
Trans
Impute
PCA
Manhatten Distance
Euclid Distance
Unweighted Average
Weighted Average
Number of Neighbors (K): 3
Min. Freq. of Row Data: 0.80
Create Dataset
Close


TASSEL (Trait Analysis by aSSociation, Evolution, and Linkage) 3.0.39
File
Tools
Help
GDPC
Data
Analysis
Results
Delete
Wizard
Show Memory
0%
GDPC
Load
Export
Sites
Taxa
Traits
Impute SNPs
Transform
Synonymizer
∪ Join
∩ Join
Separate
Sequence
mdp_genotype
mdp_genotype
mdp_genotype
mdp_genotype_chr1_157104-148907116
Polymorphisms
Numerical
mdp_population_structure
mdp_traits
mdp_genotype_chr1_157104-148907116_Colla
mdp_genotype_chr1_157104-148907116_Colla
Matrix
mdp_kinship
Tree
Fusions
Imputed Phenotypic Values.
Taxa with insufficient data: 7
K = 30.8% cutoff):
Taxa S0 S1 S2 S3 S4 S5 S6 S7 S8 S9
33-16 0 1 1 0 0 0 1 0 0 1
38-11 0 0 1 0 0 0 0 0 0 0
4226 0 1 1 0 0 0 0 0 0 1
4722 0 0 1 0 0 0 0 0 0 0
A188 1 1 1 0 0 0 0 0 0 0
A214N 0 1 0 1 0 1 0 0 0 0
A239 1 1 0 0 1 1 0 0 0 0
A272 1 1 0 0 1 1 1 0 0 0
A441-5 0 1 1 0 0 0 0 0 0 0
A554 0 0 0 0 1 0 0 0 0 0
A556 0 1 1 0 0 0 0 0 0 0
A6 1 1 0 0 1 1 0 0 0 0
A619 0 0 1 0 0 0 0 1 1 0
A632 0 1 0 1 0 1 0 0 0 0
A634 0 1 0 1 0 1 0 0 0 0
A635 0 1 0 1 0 1 0 0 0 0
A641 1 1 0 0 1 0 0 0 0 0
A654 0 0 0 0 1 0 0 0 0 0
A659 1 0 0 0 1 0 0 0 0 0
A661 0 0 0 0 1 0.33333 0 0 0 0
A679 0 1 0 1 1 0 1 0 0 0
Program Status


Column
Percent Missing Data
S0.null 0
S1.null 0
S2.null 0
S3.null 0
S4.null 0
S5.null 0
S6.null 0
S7.null 0
S8.null 0
S9.null 0
S10.null 0
S11.null 0
S12.null 0
S13.null 0
S14.null 0
S15.null 0
S16.null 0
S17.null 0
S18.null 0
S19.null 0
S20.null 0
S21.null 0
S22.null 0
S23.null 0
S24.null 0
Trans
Impute
PCA
Method
Correlation
Covariance
Output
Eigenvalue ≥ 0
Var Prop % ≥ 3.90167
Components = 3
Create Dataset
Close

Three items will be added to the data tree after running PCA. The first are the PCs. The second are the eigenvalues. And, the last are the eigenvectors. Here we use the Chart Function in the Result mode to graph the first three PCs, the individual eigenvalue contributions (sometimes called a skree plot) and the cumulative eigenvalue contributions. The eigenvalues are of interest because they equal the variance explained by each of the PCs.

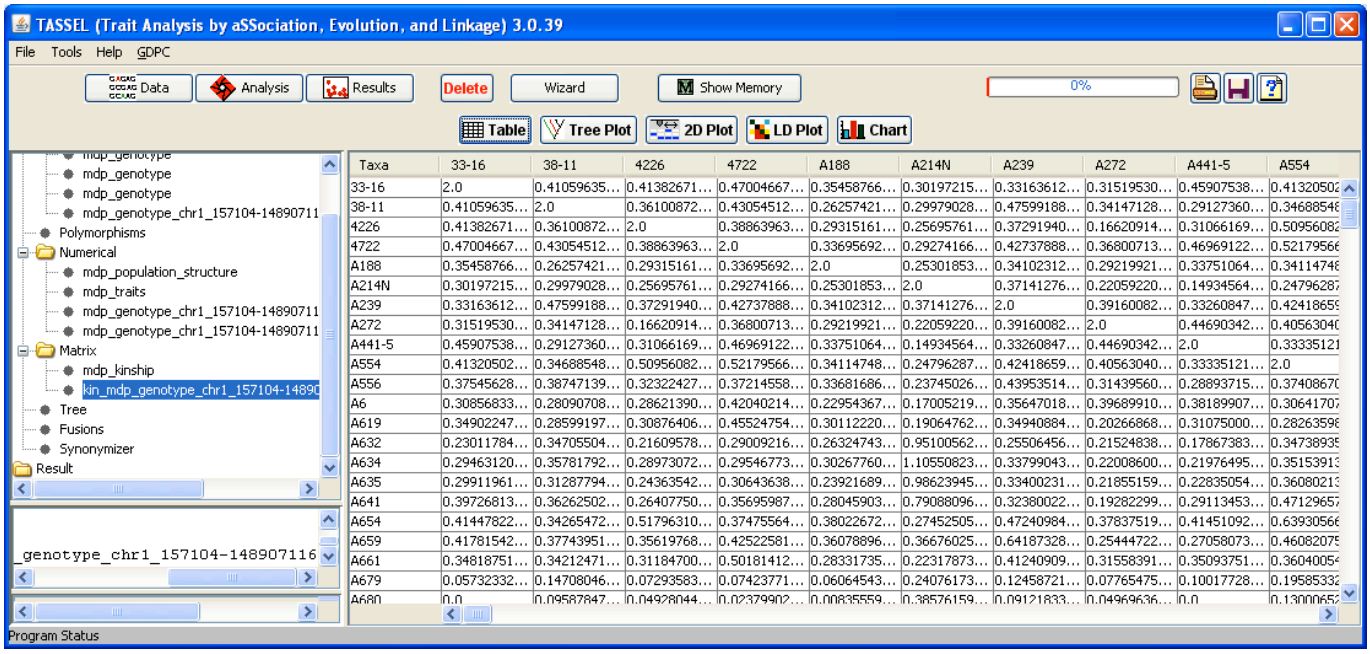




6.3 Estimation of Kinship using genetic markers

While PCs can be used to capture major population subdivisions, kinship can be used to capture more subtle relationships. This section shows how to create a kinship matrix based on the same SNP data used to calculate PC's.

1. Remove monomorphic sites: Highlight the genotype and click **Site** in **Data** mode. Set the threshold on MAF to 0.05, check "Remove minor SNP status," then click **Filter**.
2. Estimate kinship: Highlight the filtered genotype and click **Kinship** in **Data** mode. A kinship matrix will be added to the data tree under Matrix category.

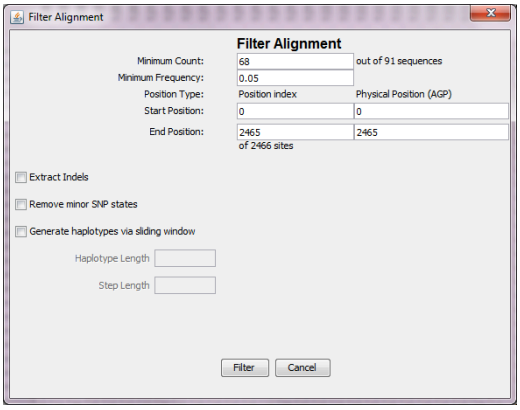


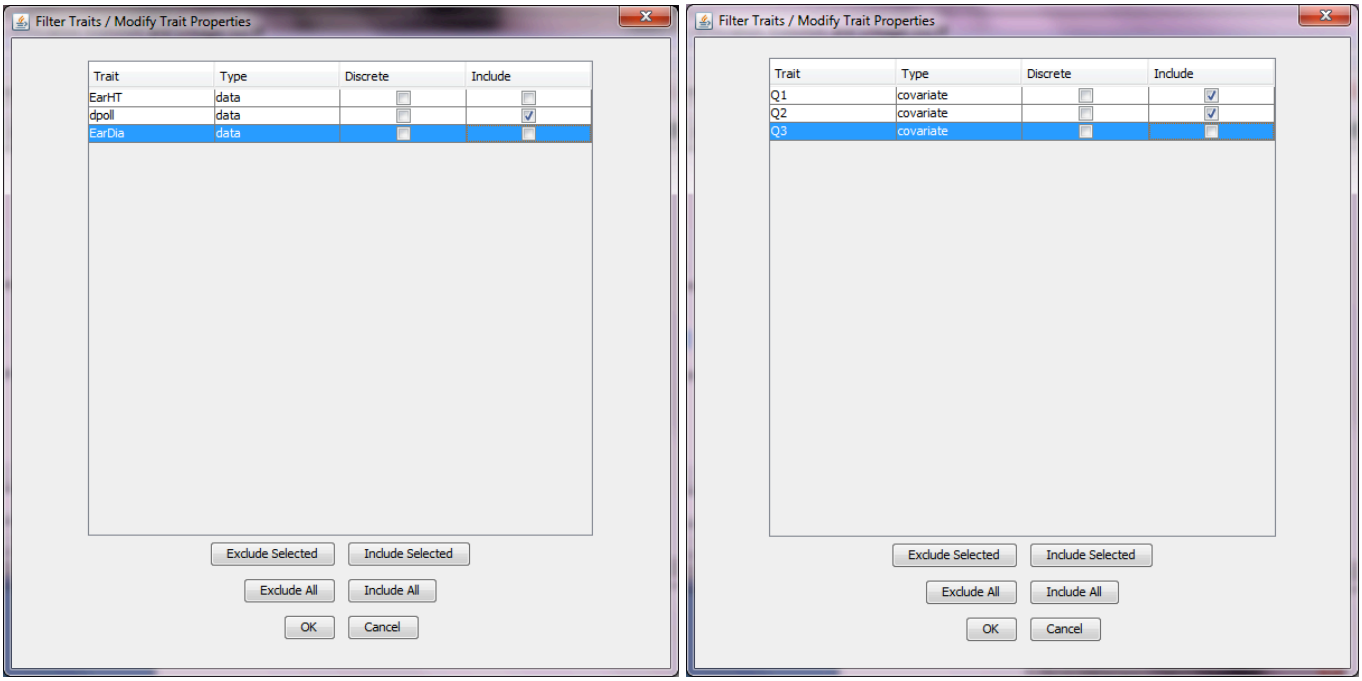
6.4 Association analysis using GLM

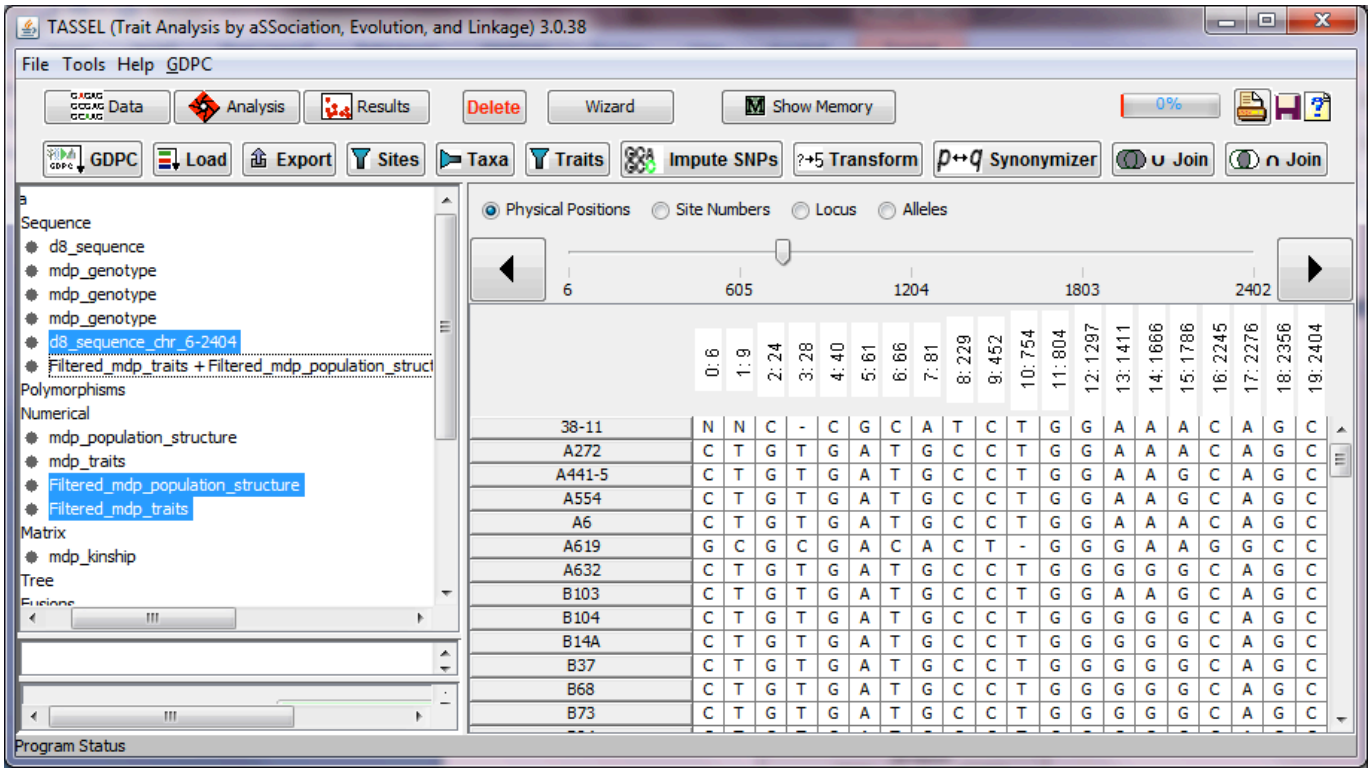
We use three files from the tutorial data set to perform association analysis using the **GLM**. The first file is the dwarf8 gene sequence with 2466 sites on 91 maize inbred lines. The second one is the population structure of 282 maize inbred lines. The last one is phenotypes for three traits, for 282 maize inbred lines. The statistical model is:

Flowering time = Population structure + Marker effect + residual

1. Remove monomorphic sites: Highlight the genotype and click **Site** in **Data** mode. Set the threshold on MAF as 0.05, then click **Filter**.
2. Trait selection: Highlight the phenotype and click **Trait** in **Data** mode. Uncheck all the traits except flowering time (DPOLL). Make sure that the Type is set to Data. Click **OK** to create a filtered phenotype.
3. Covariate selection: The population structure is presented as the proportion of each population. There are three populations represented as Q1, Q2, and Q3. They sum to 100%. This creates linear dependency if we use all of them as covariates. We can eliminate the dependency by removing one of them. In this demonstration, we exclude the last one. Highlight the filtered phenotype and click **Trait** in **Data** mode. Uncheck the last population (Q3). Make sure that the Type is set to Covariate. Then click **OK** to create a filtered population structure data.
4. Joining data: Highlight the three filtered data sets by holding the Control key while selecting the individual data. Then click Intersection (∩) **Join** on **Data** mode to create a combined data set.
5. Association analysis: Highlight the joint data set then click **GLM** in **Analysis** mode to perform association analysis. Two reports will be added to the data tree.



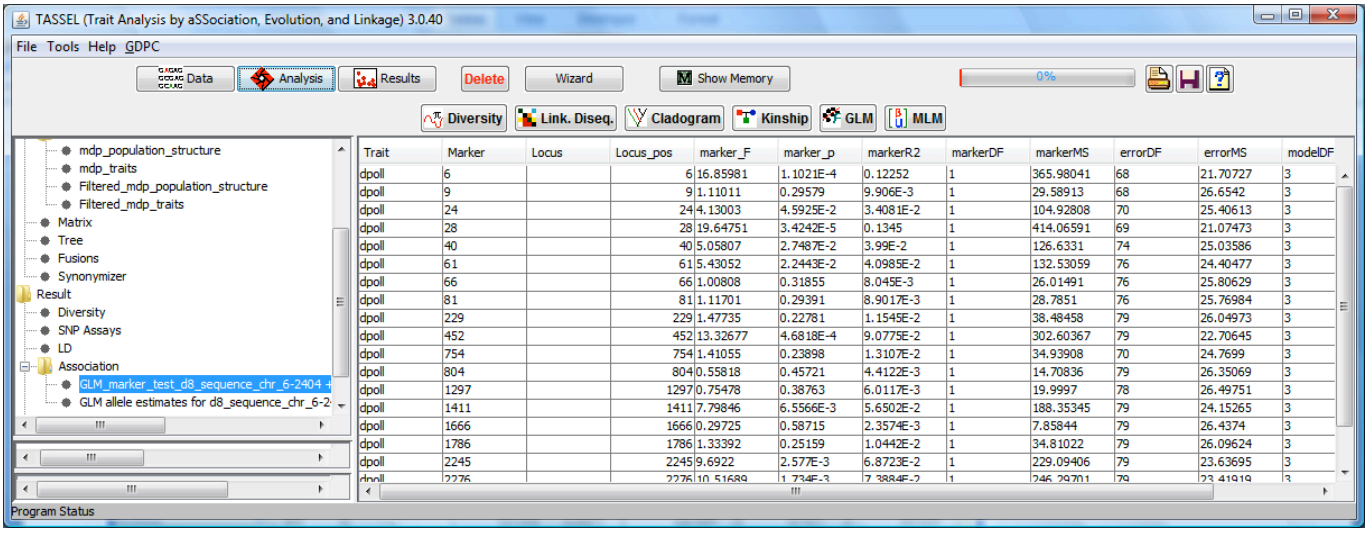




One of the reports added to data tree is labeled “GLM_Marker_Test_” followed by the name of the joint data. In addition to the information for traits and markers, the data set contains the following statistics:

marker_F: F value from the F test on marker;
marker_p: P value from the F test on marker;
markerR2: R^2 for the marker after fitting other model terms (population structure);

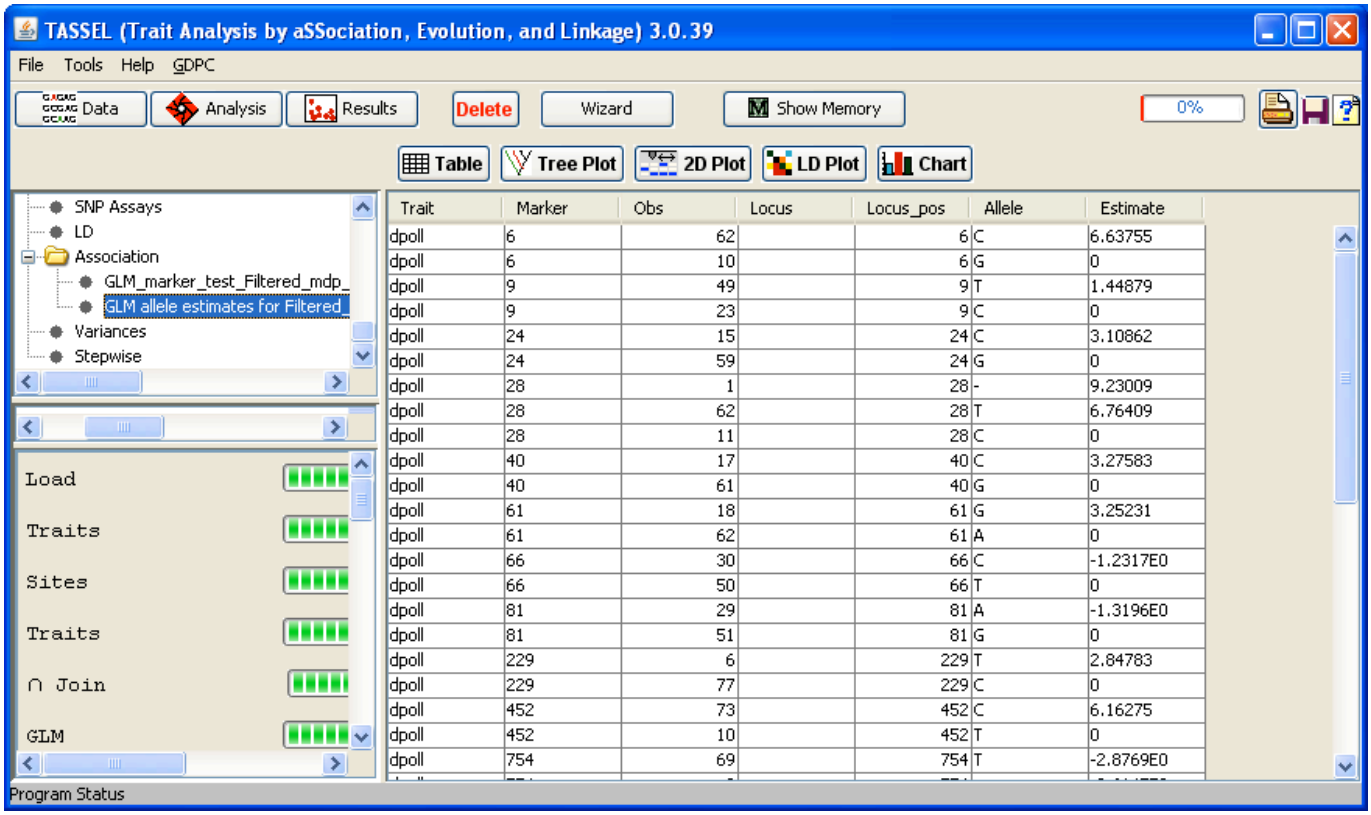
markerDF: Degree freedom of marker;
markerMS: Mean square of marker;
errorDF: Degree freedom of residual error;
errorMS: Mean square of residual error;
modelDF: Degree freedom of model;
modelMS: Mean square of model.



Trait	Marker	Locus	Locus_pos	marker_F	marker_p	markerR2	markerDF	markerMS	errorDF	errorMS	modelDF
dpoll	6		6	16.85981	1.1021E-4	0.12252	1	365.98041	68	21.70727	3
dpoll	9		9	1.11011	0.29579	9.906E-3	1	29.58913	68	26.6542	3
dpoll	24		24	4.13003	4.5925E-2	3.4081E-2	1	104.92808	70	25.40613	3
dpoll	28		28	19.64751	3.4242E-5	0.1345	1	414.06591	69	21.07473	3
dpoll	40		40	5.05807	2.7487E-2	3.99E-2	1	126.6331	74	25.03586	3
dpoll	61		61	5.43052	2.2443E-2	4.0985E-2	1	132.53059	76	24.40477	3
dpoll	66		66	1.00808	0.31855	8.045E-3	1	26.01491	76	25.80629	3
dpoll	81		81	1.11701	0.29391	8.9017E-3	1	28.7851	76	25.76984	3
dpoll	229		229	1.47735	0.22781	1.1545E-2	1	38.48458	79	26.04973	3
dpoll	452		452	13.32677	4.6818E-4	9.0775E-2	1	302.60367	79	22.70645	3
dpoll	754		754	1.41055	0.23898	1.3107E-2	1	34.93908	70	24.7699	3
dpoll	804		804	0.55818	0.45721	4.4122E-3	1	14.70836	79	26.35069	3
dpoll	1297		1297	0.75478	0.38763	6.0117E-3	1	19.9997	78	26.49751	3
dpoll	1411		1411	7.79846	6.5566E-3	5.6502E-2	1	188.35345	79	24.15265	3
dpoll	1666		1666	0.29725	0.58715	2.3574E-3	1	7.85844	79	26.4374	3
dpoll	1786		1786	1.33392	0.25159	1.0442E-2	1	34.81022	79	26.09624	3
dpoll	2245		2245	9.6922	2.577E-3	6.8723E-2	1	229.09406	79	23.63695	3
dpoll	2276		2276	10.51689	1.734E-3	7.3884E-2	1	246.29701	79	23.41919	3

Clicking "marker_p" will sort the table by P value. The smallest P value is 1.1021×10^{-4} for SNP at position 6. The threshold is 5×10^{-4} at a significance level of 1% after Bonferroni multiple test correction (0.01/20). The denominator in the Bonferroni correction is the total number of SNPs tested. The association was significant.

The other data added to the data tree is labeled "GLM_Allele_Estimates_" followed by the name of the joint data. For the most significant SNP at position 6, there were two genotypes (CC and GG). There are 62 lines with genotype CC and 10 lines with allele GG. For the trait dpoll (days to pollination), the difference between the two homozygotes was 6.63755 days.



Trait	Marker	Obs	Locus	Locus_pos	Allele	Estimate
dpoll	6	62		6	C	6.63755
dpoll	6	10		6	G	0
dpoll	9	49		9	T	1.44879
dpoll	9	23		9	C	0
dpoll	24	15		24	C	3.10862
dpoll	24	59		24	G	0
dpoll	28	1		28	-	9.23009
dpoll	28	62		28	T	6.76409
dpoll	28	11		28	C	0
dpoll	40	17		40	C	3.27583
dpoll	40	61		40	G	0
dpoll	61	18		61	G	3.25231
dpoll	61	62		61	A	0
dpoll	66	30		66	C	-1.2317E0
dpoll	66	50		66	T	0
dpoll	81	29		81	A	-1.3196E0
dpoll	81	51		81	G	0
dpoll	229	6		229	T	2.84783
dpoll	229	77		229	C	0
dpoll	452	73		452	C	6.16275
dpoll	452	10		452	T	0
dpoll	754	69		754	T	-2.8769E0

6.5 Association analysis using MLM

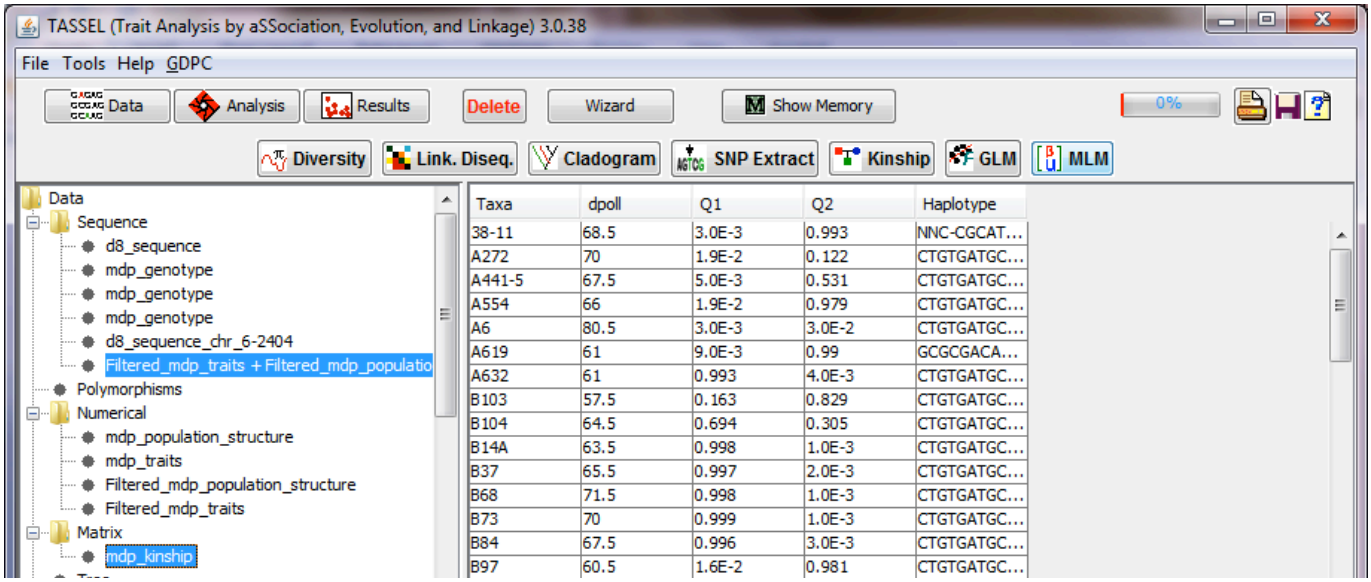
Running MLM in tassel is similar to running GLM. The difference is that in addition to the joint data (or numerical data), MLM requires kinship data to define the relationship between individuals. The kinship matrix times a parameter equals the covariance matrix between individuals. Here we use kinship file from the tutorial data set to fit the following statistical model.

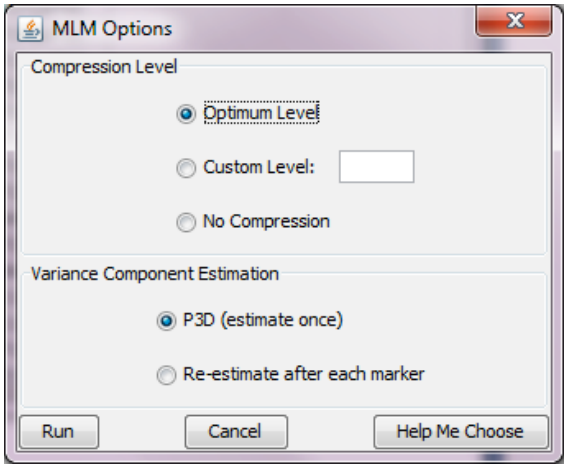
Flowering time = Population structure + Marker effect + Individuals + residual

Individuals and the residual are fit as random effects. The other terms are treated as fixed effects.

With respect to the marker effect, we will demonstrate the analysis using two sets of markers. One is the dwarf8 gene sequence used in the GLM tutorial. The other is a set of 3093 SNPs spread across the maize genome.

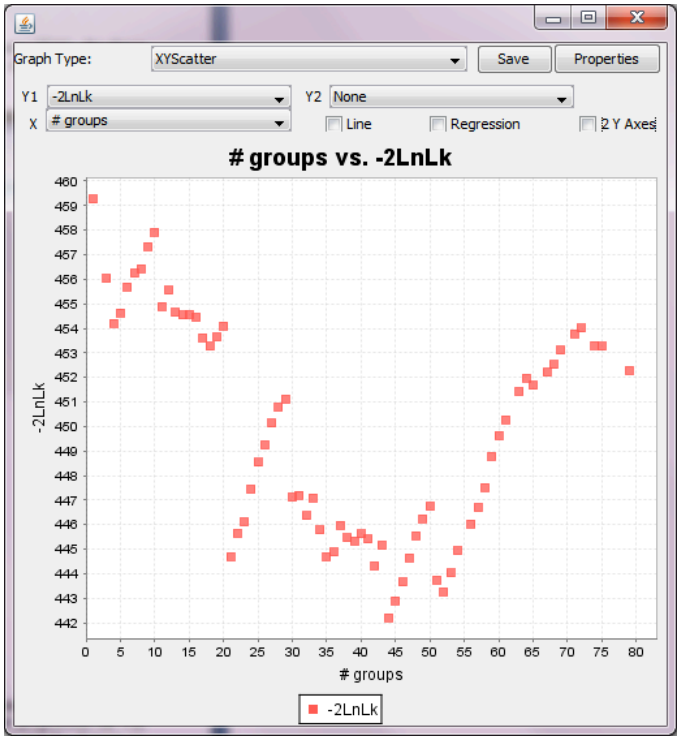
For the dwarf8 gene sequence, use the joint data set created by following the tutorial for GLM. Solve the mixed linear model by highlighting the joint data set and the kinship data then clicking the **MLM** button in **Analysis** mode.

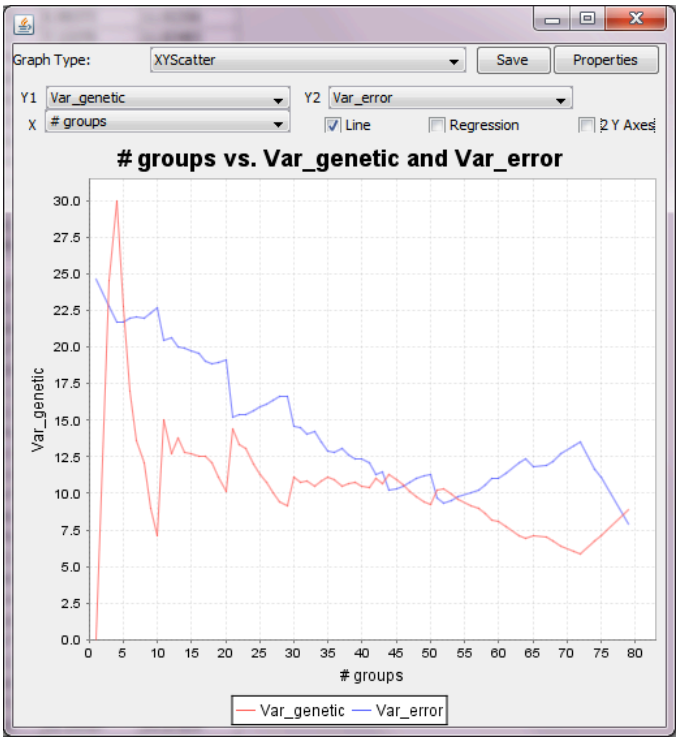




An MLM option dialog will pop up as shown above. Choose the default options, which use P3D and compression at the optimum compression level. After the Run button is clicked, the progress bar will start moving. The time required will depend on sample size, number of traits, number of markers, and the options chosen in the MLM option dialog. After the progress bar is reset to zero, indicating completion of MLM, three reports will be added to the data tree. The first two are similar to the reports created by GLM. The most significant SNP is still the same, however the strength of association is weaker, with a P value of 7.199×10^{-4} (vs. 1.1021×10^{-4} from GLM) which does not pass the Bonferroni multiple test threshold (5×10^{-4}).

The third report contains the MLM specific statistics, including -2 Log Likelihood, genetic variance and residual variance components under different level of compression. These statistics are illustrated by the Chart function on the Result mode as follows.

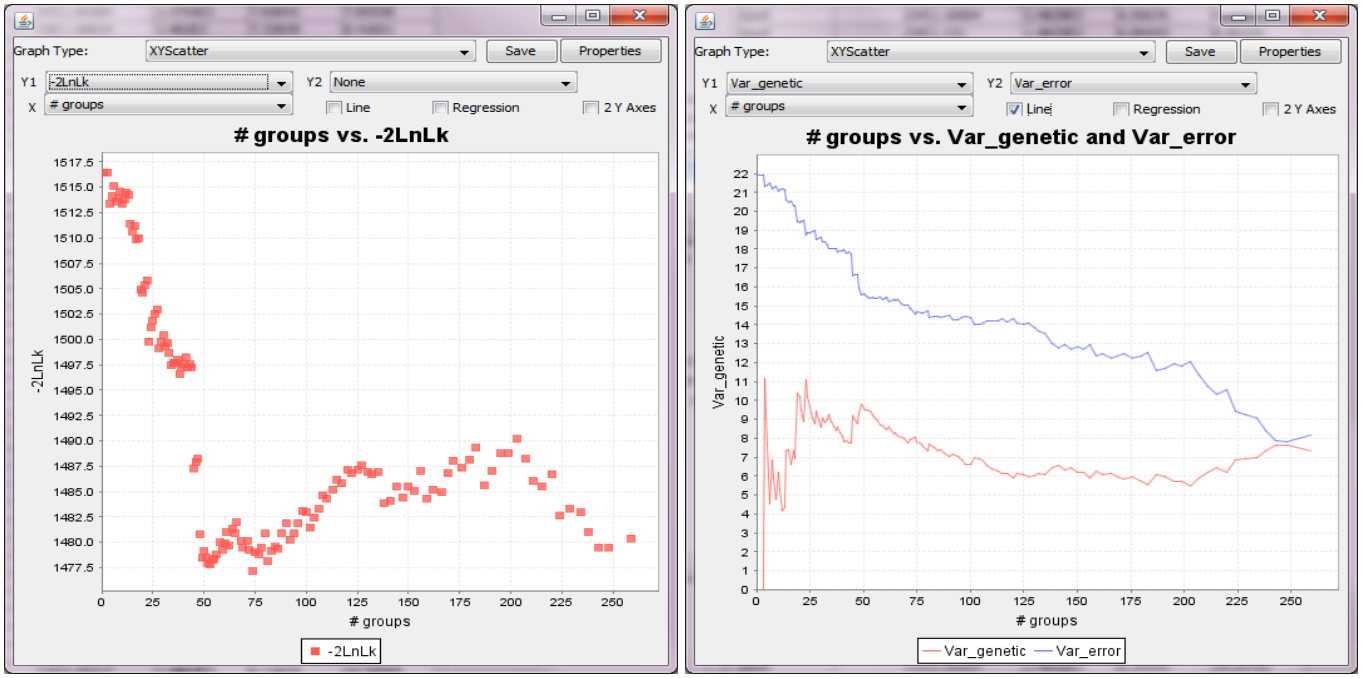




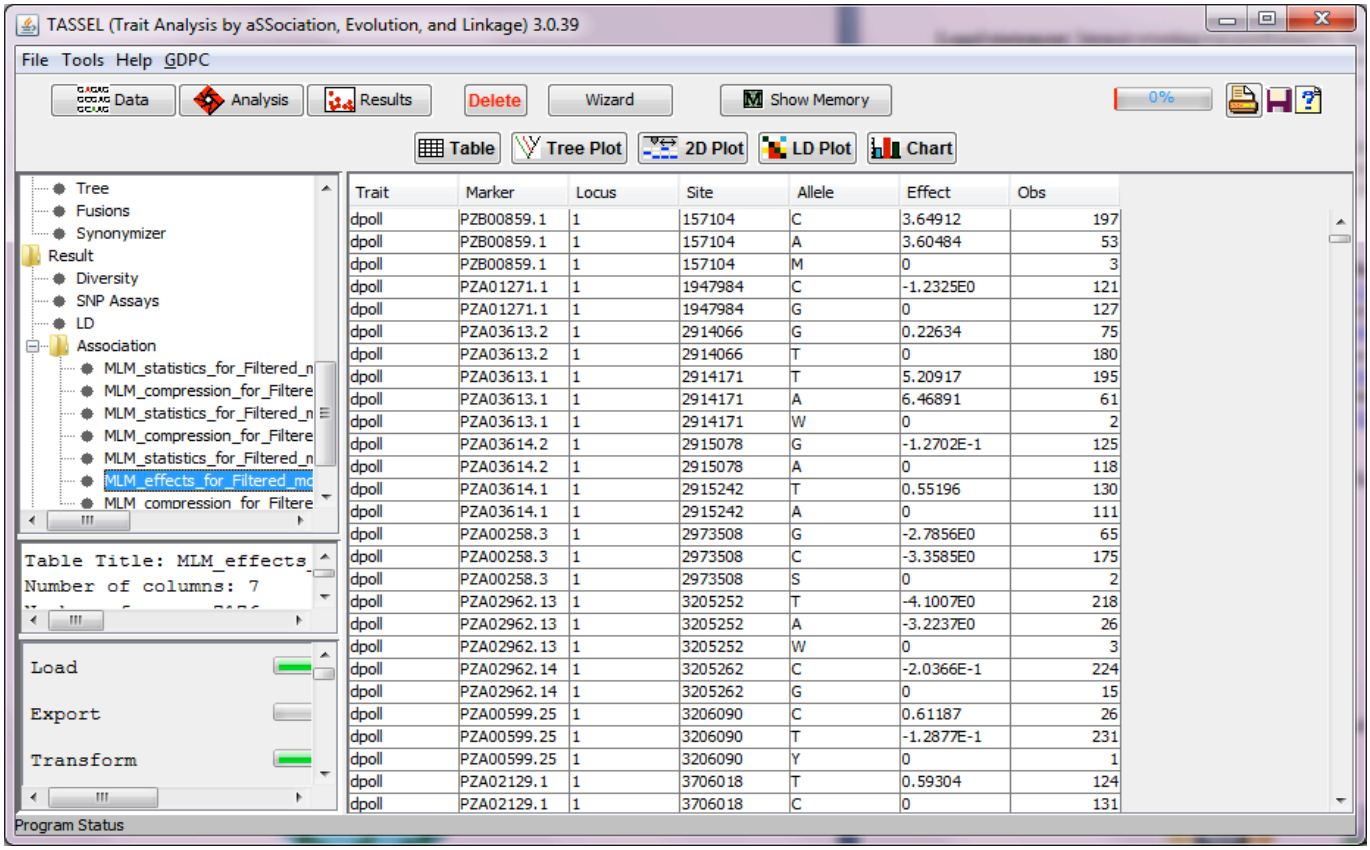
In the example, 79 are included in the final analysis. When they are clustered into 44 groups, the -2 Log Likelihood reaches a minimum, which indicates the best model fit. The screening of SNPs was performed at this optimum compression level.

Note: When two or more individuals are clustered into one group, the variance component for the random effect is not equivalent to the one without compression. Consequently, the heritability derived should not be interpreted as the individual based heritability.

To perform a Genome-Wide Association Study (GWAS) on the 3093 SNPs, we need to create a new joint data set containing the filtered phenotype, population structure, and the genome–wide genotype. Highlight the new joint file and the kinship data and click the **MLM** button. Choose the default options on the MLM option dialog. The analysis will take a minute or two. The output report labeled "MLM_compression" indicates that 259 lines were used in the analysis. With 74 groups, the statistics from the best are as graphed below.



The strongest associated SNP is at 193565357 bp on chromosome 3. The P value is 1.3027×10^{-4}. The threshold is 3.2331×10^{-5} at significant level of 1% after Bonferroni multiple test correction (0.01/3093). The association was not significant. As illustrated below, the output labeled "GLM_Allele_Estimates" shows the marker effects assigned to genotypes for each SNP (The GLM is also the same). For example, the first SNP at 157104 bp on chromosome 1 had three genotypes (AA, CC and AC) coded as A, C, and M based on the IUPAC code, see Appendix (Nucleotide Codes).



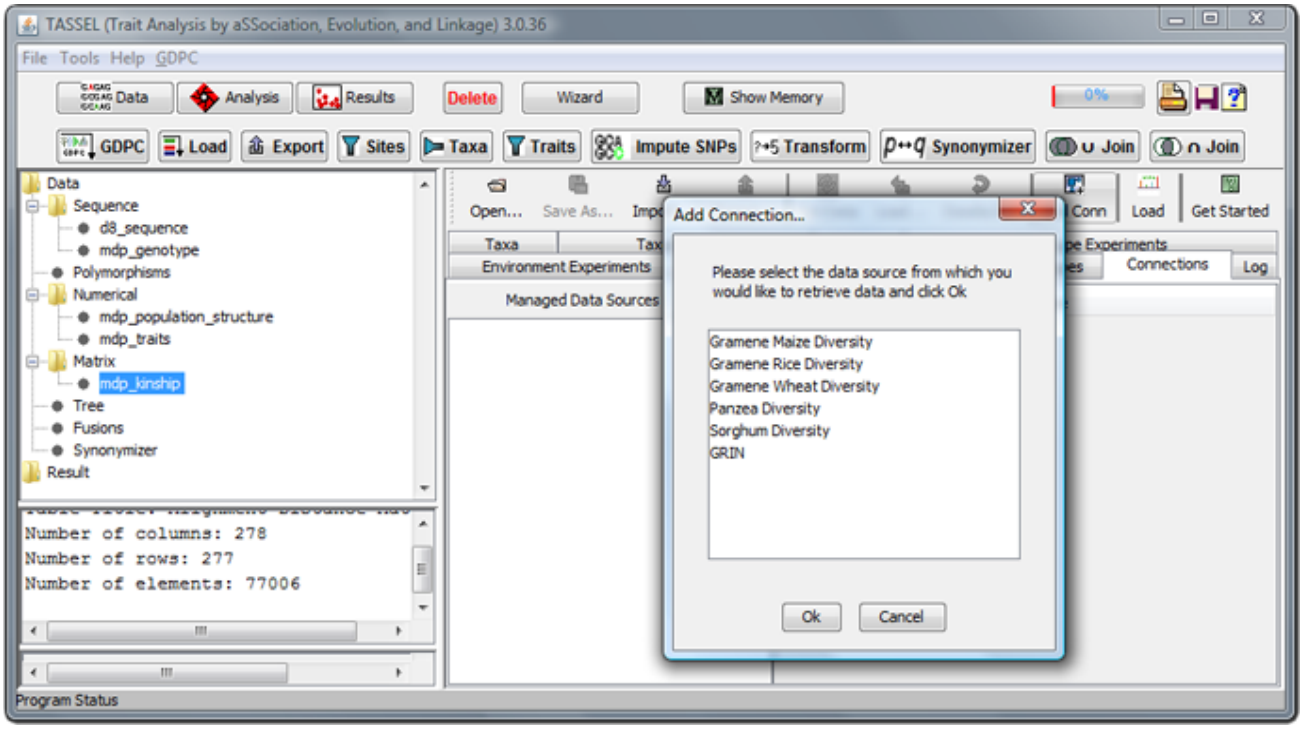
Trait	Marker	Locus	Site	Allele	Effect	Obs
dpoll	PZB00859.1	1	157104	C	3.64912	197
dpoll	PZB00859.1	1	157104	A	3.60484	53
dpoll	PZB00859.1	1	157104	M	0	3
dpoll	PZA01271.1	1	1947984	C	-1.2325E0	121
dpoll	PZA01271.1	1	1947984	G	0	127
dpoll	PZA03613.2	1	2914066	G	0.22634	75
dpoll	PZA03613.2	1	2914066	T	0	180
dpoll	PZA03613.1	1	2914171	T	5.20917	195
dpoll	PZA03613.1	1	2914171	A	6.46891	61
dpoll	PZA03613.1	1	2914171	W	0	2
dpoll	PZA03614.2	1	2915078	G	-1.2702E-1	125
dpoll	PZA03614.2	1	2915078	A	0	118
dpoll	PZA03614.1	1	2915242	T	0.55196	130
dpoll	PZA03614.1	1	2915242	A	0	111
dpoll	PZA00258.3	1	2973508	G	-2.7856E0	65
dpoll	PZA00258.3	1	2973508	C	-3.3585E0	175
dpoll	PZA00258.3	1	2973508	S	0	2
dpoll	PZA02962.13	1	3205252	T	-4.1007E0	218
dpoll	PZA02962.13	1	3205252	A	-3.2237E0	26
dpoll	PZA02962.13	1	3205252	W	0	3
dpoll	PZA02962.14	1	3205262	C	-2.0366E-1	224
dpoll	PZA02962.14	1	3205262	G	0	15
dpoll	PZA00599.25	1	3206090	C	0.61187	26
dpoll	PZA00599.25	1	3206090	T	-1.2877E-1	231
dpoll	PZA00599.25	1	3206090	Y	0	1
dpoll	PZA02129.1	1	3706018	T	0.59304	124
dpoll	PZA02129.1	1	3706018	C	0	131

6.6 Importing Data from a Database (via GDPC)

GDPC, middleware that is integrated into TASSEL, allows the user to import data from a database. To display GDPC in TASSEL, click on the **GDPC** button in **Data** mode. General rules for working with databases include: 1) Establish a connection with the database; 2) Define a query; 3) once the desired data is in GDPC, load the data from GDPC into TASSEL.

6.6.1 Connecting with a Database

To establish a connection with a database, click the **Add Conn** button followed by the button of the database you wish to add. Then click **Ok**. In the example below, we chose Panzea.



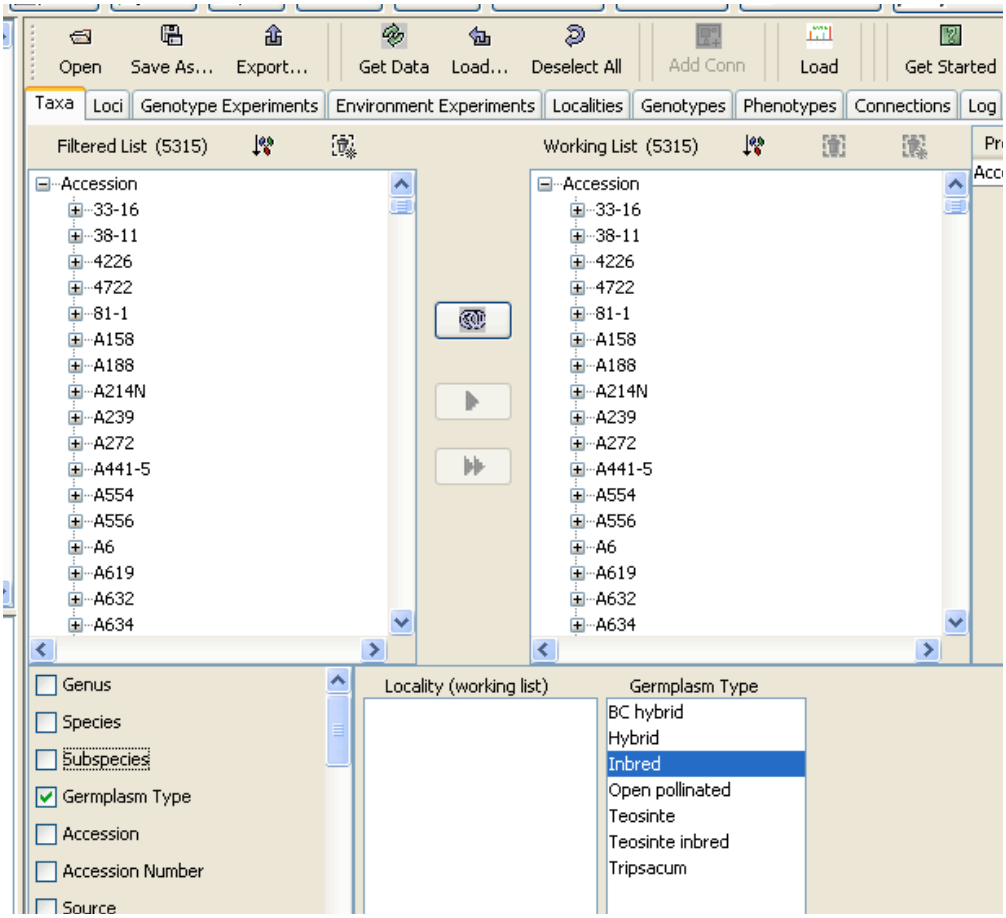
To connect to more than one database, simply repeat the process outlined above.

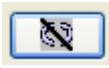
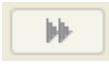
In the figures of following sections, only the GDPC area will be displayed if other areas are deemed irrelevant.

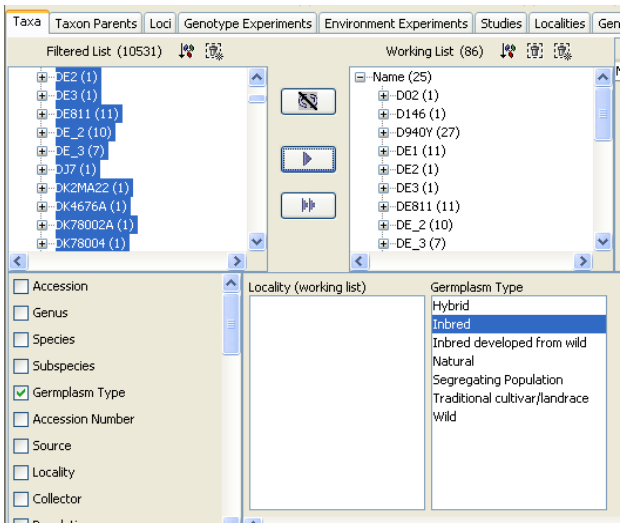
6.6.2 Data Query

GDPC is equipped with several tabs to query data, namely **Taxa**, **Taxon Parents**, **Loci**, **Genotype Experiments**, **Environment Experiment**, and **Localities**. Within each tab, any retrieved data will be displayed in the "Filtered List." Choose attributes by checking the desired boxes (located beneath the Filtered List). After an attribute is selected, values of that attribute from the database are displayed.

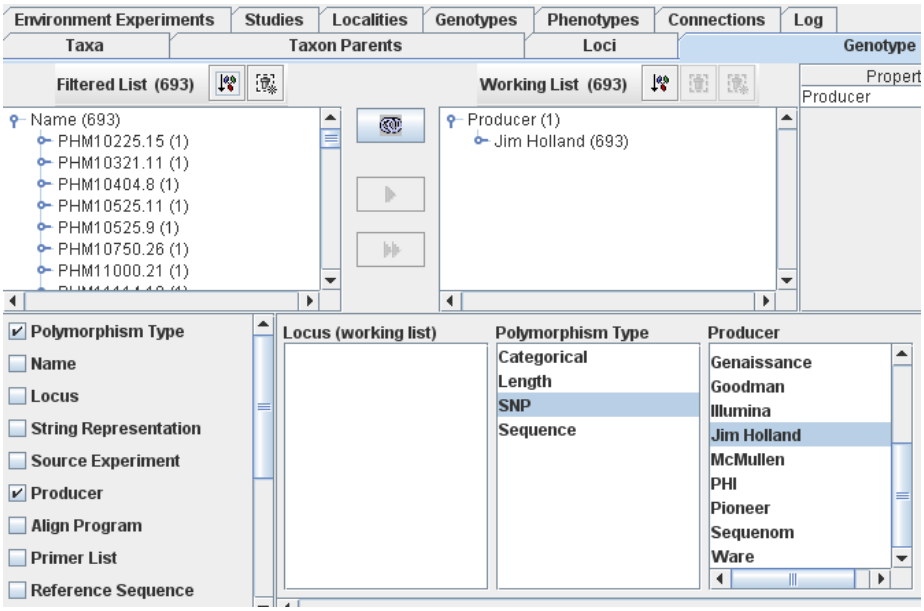
Here, using the Taxa tab, choose **Germplasm type** (field) and then select. After clicking the **Get Data** button, the subset of taxa from the database that meets these criteria will appear in both the **Filtered List** and the **Working List**.



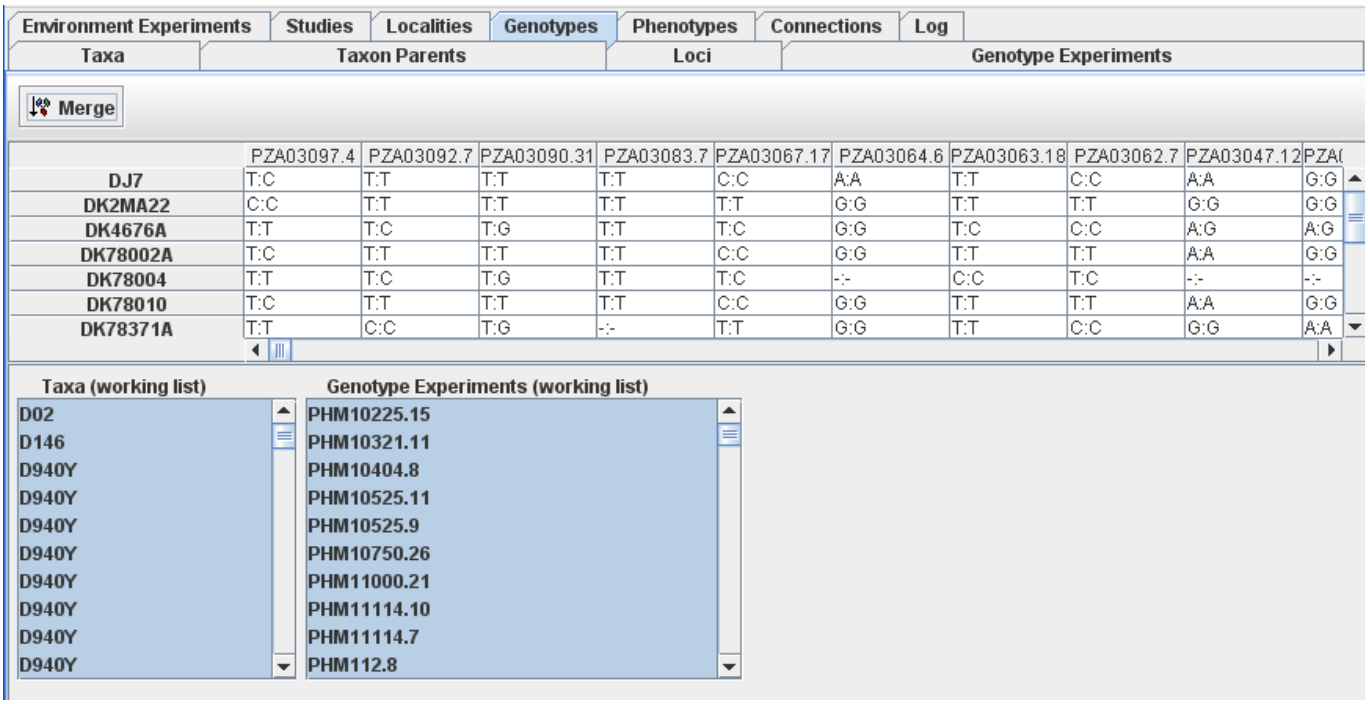
Items listed in the **Working List** can be modified by the user. To do so, first break the link between the Filtered List and the Working List by clicking on the Link/Unlink button  . The button will now appear as  . This activates the **Add selected items**  , **Add all items**  , **Remove**  and **Remove all**  buttons. Remove all items from the working list, then select items with a name starting with the letter D. Click on the Add selected items button to move them to the Working List. The resulting Working List is shown as follows:



To filter data by polymorphism type, first click on the **Genotype Experiments** tab, check the **Polymorphism Type** and **Producer** checkbox (field), and then select **SNP** and **Jim**. Finally, click the **Get Data** button to reveal the subset of data that meets these criteria. Results for this example are shown below:

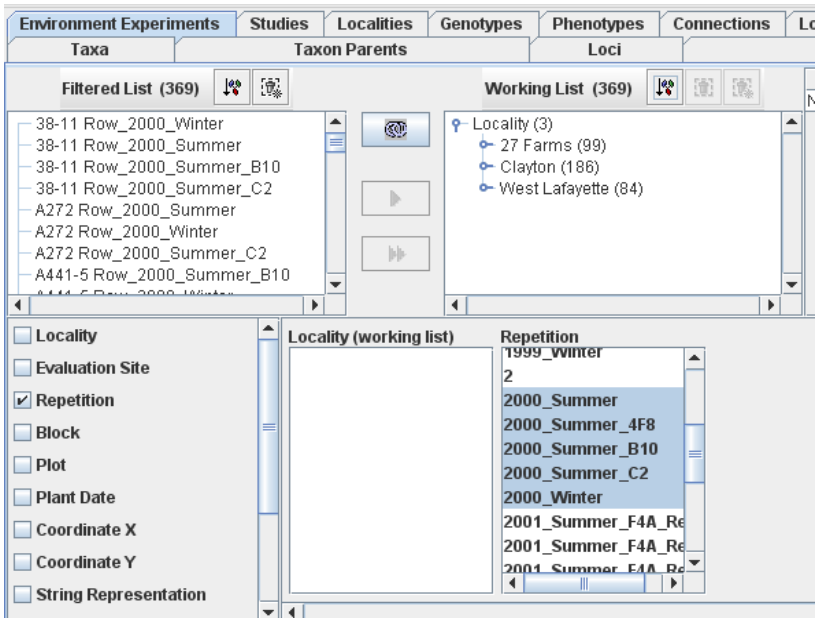


Genotype data can be extracted from the database by clicking on the **Genotypes** tab, followed by the **Get Data** button. After a moment, genotype data will be displayed as follows:

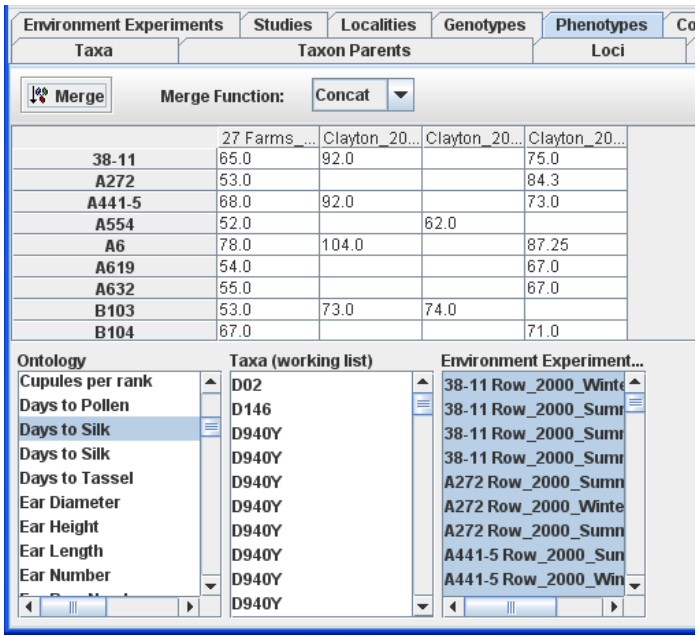


Users can either save this genotype data in several formats or upload it to TASSEL. However, before outlining these procedures, let us finish the query by exploring phenotypes. To get data from experiments conducted in 2000, first select the **Environment Experiments** tab, followed by the **Repetition** checkbox.

Select the desired repetitions in 2000 as the values to be used for filtering, then click the **Get Data** button. The subset of data that meets these criteria is returned as follows:



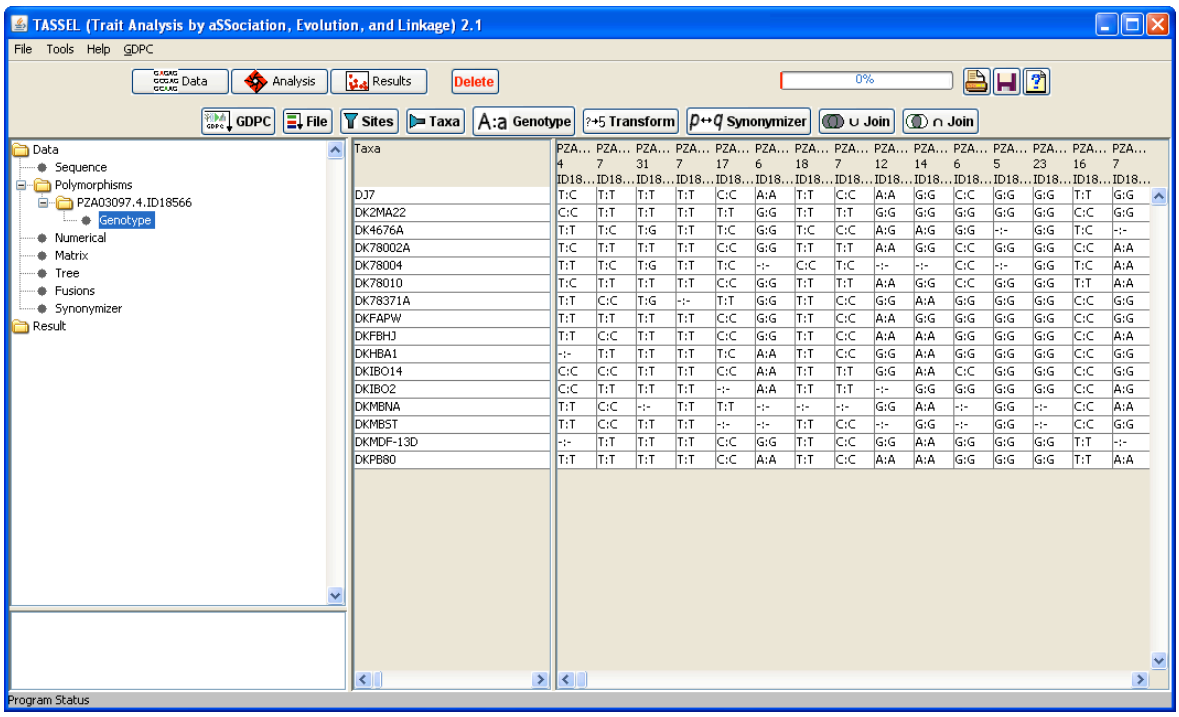
Now extract phenotype data by clicking on the **Phenotypes** tab. Traits can only be extracted one at a time. Choose **Days to Silk** from the **Ontology** field. Make sure no Taxa are selected and all Environment Experiments are selected that were retrieved in the previous step. Click the **Get Data** button, then the **Merge** button, leaving only **Accession** checked under the Taxa Properties section. Leave **Locality** and **Repetition** checked under the Environment Experiments Properties section. Data are merged as follows:



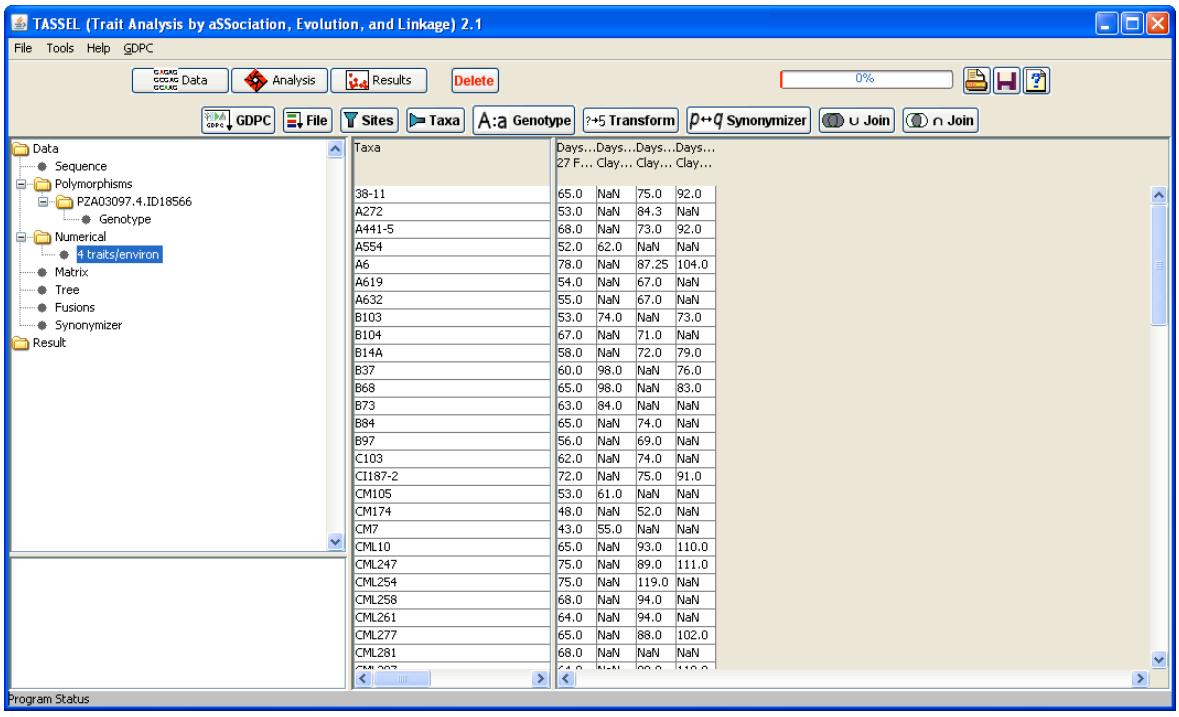
6.6.3 Importing GDPC data into TASSEL

Genotype and phenotype data must be loaded in separate steps. To load genotype data, first click on **GDPC** in **Data** mode. Then click on the **Genotypes** tab, followed by the **Load** button. The genotype data

is then loaded into TASSEL and labeled as "Genotype." To view the uploaded data, click on "Genotype" within the Data folder in TASSEL. Results will look as follows:

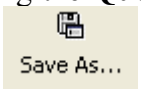


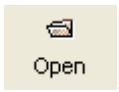
To load phenotype data from GDPC into TASSEL, first click on the **GDPC** button in **Data** mode. Then choose the **Phenotypes** tab, followed by the **Load** button. The phenotype data is then loaded into TASSEL and labeled as "**4 traits/environ.**" To view the uploaded data, select "**4 traits/environ**" from the Phenotypes folder in TASSEL. Results will appear as follows:



6.6.4 Saving GDPC Query Results

All query results, including both genotype and phenotype queries, can be saved as either Tab-delimited text files or XML files. Results are exported as tab-delimited text files by first choosing the **Query** Tab and then clicking on the **Export** button 

, or by clicking the **Save As** button 

to save results in XML format. Location and file name must be specified in both situations. Data in XML format can be imported back into GDPC by clicking on the **Open** button 

.

7 Appendix

7.1 Nucleotide Codes (Derived from IUPAC)

Code	Meaning
A	**A:A**
C	**C:C**
G	**G:G**
T	**T:T**
R	**A:G**
Y	**C:T**
S	**C:G**
W	**A:T**
K	**G:T**
M	**A:C**
+	**+:+ (insertion homozygous)**
0	**+:-**
-	**-:- (deletion homozygous)**
N	**Unknown**

7.2 TASSEL Tutorial Data sets

The data set contains 9 files and can be downloaded at:
http://www.maizegenetics.net/tassel/docs/TASSELTutorialData3.zip

File#	File name	Type	Format
1	d8_sequence.phy	Genotype	Phylip Alignment
2	mdp_genotype.hmp.txt	Genotype	Hapmap Alignment
3 4	mdp_genotype.flpjk.geno mdp_genotype.flpjk.map	Genotype	Flapjack Alignment
5 6	mdp_genotype.plk.ped mdp_genotype.plk.map	Genotype	Plink Alignment
7	mdp_kinship.txt	Kinship	Numerical square matrix
8	mdp_population_structure.txt	Population structure	Numerical trait data
9	mdp_traits.txt	Phenotype	Numerical trait data

File #1 is the sequence of dwarf8 gene with 2466 sites on 91 maize inbred lines. The data was described by the paper on the association between Dwarf8 and flowering time[26].

File #2-6 are 3093 SNPs on 281 maize association inbred lines. The data was presented in three formats (Hapmap, Plink and Flapjack). The data was created by the PANZEA project funded by NSF. Details of the data can be found at http://www.panzea.org.

File #3 and 4 are in pair for the format of Flapjack.

File #5 and 6 are in pair for the format of Plink.

File #7 is kinship created by Yu et al.[9].

File #8 is population structure of 282 maize inbred line[27].

File #9 is phenotype on three traits, including flowering time, on 282 maize inbred lines[9].

7.3 Biography of TASSEL

2001	First public release
December, 2004	Score-able SNP Extractor Updated Main Panel
February, 2005	StepClade update
March, 2005	Fixed handling of ?s and non-standard characters Added Sliding Haplotype functionality Changed LD Fisher's Exact p-value to use two-sided p-value
April, 2005	Added Ability to visualize sequence quality scores "Synonymize"/match taxa names between data sets GLM analysis improvements
June, 2005	Code change preventing large data sets from being shown in JTable Update of GDPC which allows automatic restoration of last data source connection
October, 2005	Data transformation utilities added K-Nearest Neighbor Data Imputation added
January, 2006	Association analysis with Mixed Linear Model Taxa name " synonymizer" added Basic heterozygosity handling added Many ease-of-use improvements
March, 2006	Fixed problem loading genotype data Mixed Linear Model changes: • Output NaN if non-converged • Fixed problem loading genotype data • Detection of duplicate ID in kinship • Correction on progressive bar with MLM • Starting values of NaN from previous marker are no longer used
September, 2006	MLM: Significant speed improvement (~10x faster) GLM: Added User-defined F-tests, Output taxa or marker means
October, 2006	Principle Components Analysis
September 2007	Architecture restructure and pipeline version for advanced users
April 2008	Genetic marker data numerical transformation
June 2008	MLM implemented P3D algorithm, increased speed in order of magnitude

	of at least ten times.
May 2009	EMMA implemented
November 2009	TASSEL Version 3 release (redesigned for large genomic data and large samples)
April 2010	Compression of MLM implemented

7.4 Frequently Asked Questions

1. What do I do if TASSEL misbehaves?

A: TASSEL is an open source software project hosted on SourceForge and has a bug tracking list at http://sf.net/projects/tassel where you can notify the developer community of problems. In order for a bug to be fixed, we must be able to replicate the problem. Thus, it is important to document the steps that were taken that produced the error. If the data you are working with is not too sensitive, please include the files which were used in the faulty procedure. If you would rather not post your data file on SourceForge, you may email it to one of the software developers.

2. Where do I turn for more information?

A: If you are having difficulty with a certain aspect of TASSEL, you can either email one of the software developers listed at www.maizegenetics.net or you may check the TASSEL forum on SourceForge http://sf.net/projects/tassel), as another user may have already addressed a similar question. There is also a TASSEL discussion group at http://groups.google.com/group/tassel.

3. How do I join the fun: TASSEL on SourceForge?

A: TASSEL is an open source project distributed under the GNU general public license. This means that the source code is available and the user is free to modify the code to suit their particular needs. We welcome input from developers and those who wish to become involved in the improvement of this software. The project is hosted on SourceForge (http://sf.net/projects/tassel), thereby allowing anyone to access the most recent changes to the code. This setup makes it convenient for anyone to add special functionality to TASSEL if they so desire. It also serves as a good platform for anyone who wishes to become involved in a bioinformatics software development project.

4. How do I change the amount of memory used? What do I do when the "Exception java.lang.OutOfMemoryError" appears?

A: If you are working with very large data sets or are running memory intensive procedures, there may be occasions when TASSEL runs out of memory. For most routine usage, however, TASSEL memory is sufficient. Memory issues usually result from attempting to execute a procedure like LD on a raw sequence alignment instead of selected SNPs. You may also experience a memory issue if you are not sufficiently specific when retrieving information through GDPC. By default, TASSEL is allocated up to 512 Mb of memory on your computer. If more is available on your computer, you can increase the amount allocated by downloading the "stand-alone" version of TASSEL and opening a command line window (in Windows use Start > Run and type in "cmd" or "command"). To run TASSEL from a command line, "cd" to go to the directory containing the stand-alone jar file then start TASSEL by typing the following:

java -Xms256M –Xmx768M –jar sTASSEL.jar

Where "-Xms###M" specifies the starting memory available and "–Xmx###M" specifies the maximum memory available to the Java Virtual Machine. You may set the values higher or lower as your hardware dictates. Alternatively, you can modify the start_tassel.bat or start_tassel.pl file that comes with the standalone distribution.

5. When I click on the most current version of TASSEL web start, a previous version appears. What should I do?

A: The previous version of TASSEL web start was cached in your machine. To replace it with the most current version, click the Start button in Windows, followed by Run. Type **javaws** and then click OK. In the window that opens, keep the most current version of TASSEL and delete the rest.

6. What should I substitute for missing values in TASSEL?

A: For numerical data in version 3 format, use NA or NaN. For numerical data in version 2 format, use "-999" for missing values. For SNP data, use "N". For SSR data, use "?". Kinship does not allow missing values.

7. Is it possible to change data names in the Data Tree?

A: Yes. Click on the desired data name in the Data Tree, wait for one second, and then click it again or immediately hit the F2 key. Rename the data set and then hit Enter to save the change.

8. How can I create a TASSEL icon on desktop?

A: Click "Start" on Microsoft Windows and select "Control Panel", then double click Java to show "java Control Panel". In "Temporary Internet Files" section, click "View" button show "Java Cache Viewer". Move mouse over TASSEL application and click right button and select "Install Shortcuts".

9. Why do I get empty squares in MLM association analysis?

A: The empty square means null information. The major reasons include non-convergence in the estimation of variance componentsor that the statistic in question was not calculated. For example, marker F, p, and R^2 are not calculated when no marker is included in the model.

10. Why should I exclude one column of the population structure?

A: For some methods of calculating population structure, such as the software STRUCTURE, the population proportions sum to one. This produces linear dependence between the population covariates. While the algorithm used by GLM tolerates that dependency, MLM will fail because the design matrix will not be invertibleExcluding one column eliminates linear dependence between columns. Using PC axes to represent population structure does not result in linear dependency because all PC columns are guaranteed to be independent.

11. Can kinship replace population structure?

A: Sometimes. For some traits and populations, the K-only model may be as good as or better than the Q+K model. For others, Q+K may be superior. The Q-only model is not as effective for controlling population structure as the alternatives. Unfortunately, no general guidelines exist for predicting which model will perform best. As a result, an investigator may wish to fit all three models and compare the results. If eliminating false positives is very important, then it may make sense to accept the most conservative model. However, if the objective is to identify candidates for further study and the cost of following up on a false lead is low, the most liberal model may be preferred.

12. Why do TASSEL and SPAGeDi give different kinship estimates?

A: First, many algorithms exist to calculate kinship and their estimates will differ from one another. Secondly, the algorithm in TASSEL treats each genotype as a haplotype. It is not recommended that TASSEL be used to generate a kinship matrix from heterozygous genotype. In the near future, the TASSEL kinship algorithm will be modified to handle heterozygous diploids.

13. Can I get Marker R square using SAS Proc Mixed or TASSEL MLM?

A: SAS Proc Mixed does not produce an R^2 statistic. MLM in TASSEL does. The user manual describes how it is calculated.

14. Does MLM find more associations than GLM?

A: Sometimes. MLM has higher statistical power than GLM and may detect more true associations.. When the tested genetic markers are confounded with kinship structure , GLM does not correct for that as effectively as MLM and may produce more false positives

15. Do I need multiple test correction for the p value from Tassel?

A: Yes.

16. Can TASSEL handle diploid genotype data?

A: While TASSEL accepts most common sequence alignment formats which handle polyploid genotype data including haploid and diploid, some analyses are not appropriate for heterozygous data. GLM or MLM fit SNPs one at a time, treating each distinct genotype as a separate class. This has the effect of fitting an additive plus dominance model. Separating the two effects is under consideration. Because handling heterozygotes as a third marker class is not appropriate for kinship or LD those analyses should not be used for that type of data at the present time. Work to improve handling heterzygotes is ongoing.

17. How to cite TASSEL?

A: The paper that describes TASSEL[1] as a software package and the papers that introduce specific methods implemented in TASSEL should be cited as appropriate, such as the unified ("Q+K") approach, EMMA, compression of mixed linear model and P3D. For example,:

A. Linkage disequilibrium (D', R^2 and P value) were calculated by TASSEL[1].
B. Association analyses were performed with the mixed linear model approach[9] implemented by TASSEL[1].
C. GWAS was performed with the compressed mixed linear model approach[4,9] carried by TASSEL[1] which also implemented the EMMA[3] and P3D[4] algorithms to reduce computing time.

REFERENCES

1. Bradbury, P.J. et al. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633-2635 (2007).
2. Zhang, Z., Buckler, E.S., Casstevens, T.M. & Bradbury, P.J. Software engineering the mixed model for genome-wide association studies on large samples. *Brief Bioinform* **10**, 664-75 (2009).
3. Kang, H.M. et al. Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics* **178**, 1709-1723 (2008).
4. Zhang, Z. et al. Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* **42**, 355-60 (2010).
5. Kang, H.M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**, 348-54 (2010).
6. Thornsberry, J.M. et al. Dwarf8 polymorphisms associate with variation in flowering time. *Nature Genetics* **28**, 286-289 (2001).
7. Pritchard, J.K., Stephens, M., Rosenberg, N.A. & Donnelly, P. Association mapping in structured populations. *American Journal of Human Genetics* **67**, 170-181 (2000).
8. Zhao, K. et al. An Arabidopsis example of association mapping in structured samples. *PLoS Genet* **3**, e4 (2007).
9. Yu, J.M. et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**, 203-208 (2006).
10. Casstevens, T.M. & Buckler, E.S. GDPC: connecting researchers with multiple integrated data sources. *Bioinformatics* **20**, 2839-2840 (2004).
11. Ware, D. et al. Gramene: a resource for comparative grass genomics. *Nucleic Acids Research* **30**, 103-105 (2002).
12. Ware, D.H. et al. Gramene, a tool for grass Genomics. *Plant Physiology* **130**, 1606-1613 (2002).
13. Jaiswal, P. et al. Gramene: development and integration of trait and gene ontologies for rice. *Comparative and Functional Genomics* **3**, 132-136 (2002).

14. Yamazaki, Y. & Jaiswal, P. Biological ontologies in rice databases. An introduction to the activities in gramene and oryzabase. *Plant and Cell Physiology* **46**, 63-68 (2005).
15. Zhao, W. et al. Panzea: a database and resource for molecular and functional diversity in the maize genome. *Nucleic Acids Research* **34**, D752-D757 (2006).
16. Canaran, P., Stein, L. & Ware, D. Look-Align: an interactive web-based multiple sequence alignment viewer with polymorphism analysis support. *Bioinformatics* **22**, 885-886 (2006).
17. Du, C.G., Buckler, E. & Muse, S. Development of a maize molecular evolutionary genomic database. *Comparative and Functional Genomics* **4**, 246-249 (2003).
18. SAS, I.I. SAS. Statistical Analysis Software for Windows, 9.0 ed. *Cary, NC. USA.* (2002.).
19. Hardy, O.J. & Vekemans, X. SPAGEDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes* **2**, 618-620 (2002).
20. Cover, T. & Hart, P. Nearest neighbor pattern classification. *Proc IEEE Trans Inform Theory* **13**(1967).
21. Weir. Genetic Data Analysis II. *Sunderland, MA.* (1996).
22. Farnir, F. et al. Extensive genome-wide linkage disequilibrium in cattle. *Genome Res* **10**, 220-7 (2000).
23. Henderson, C.R. Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics* **31**, 423-447 (1975).
24. Kang, H.M. et al. Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709-23 (2008).
25. Laird, N.M. & Ware, J.H. Random-Effects Models for Longitudinal Data. *Biometrics* **38**, 963-974 (1982).
26. Thornsberry, J.M. et al. Dwarf8 polymorphisms associate with variation in flowering time. *Nat Genet* **28**, 286-9 (2001).
27. Flint-Garcia, S.A. et al. Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J* **44**, 1054-64 (2005).
28. Anderson, M.J. & Ter Braak, C.J.F. Permutations tests for multi-factorial analysis of variance. Journal of Statistical Computation and Simulation 73, 85-113 (2003)

INDEX