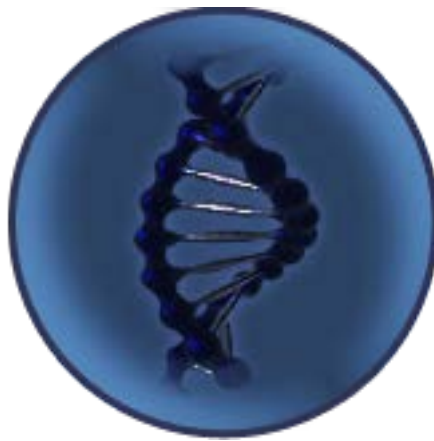


TRES: Toolbox for Ranking and Evaluation of SNPs



Software's website

<http://mlkd.csd.auth.gr/bio/tres/>

Version 1.0

Last update: 18/11/2015

0. Contents

0.	Contents.....	2
1.	Overview	3
2.	Requirements & Installation	4
2.1	How to Install and Run TRES?.....	4
2.2	How to RESERVE more MEMORY for the TRES?	4
3.	Data & Input File Format	6
4.	Application Description.....	7
4.1	SNP Selection Tab.....	7
4.2	Compare Tab	10
4.3	Converter – Splitter Tab	10
4.4	Reduced Dataset Generator Tabs	12
4.5	About Tab	13
5.	How to Use TRES – A Step by Step Analysis Process	13
5.1	Downloading, Installing and Running TRES.....	14
5.2	Splitting the Dataset into Train and Test	15
5.3	Converting the PED file Dataset into ARFF.....	16
5.4	SNP Selection.....	16
5.5	Reduced Dataset Generation	18
5.6	Hints on using GENECLASS2	20
6.	Contact Information.....	21

1. Overview

TRES is a collection of algorithms built in user friendly and computationally efficient software that can manipulate and analyze datasets even in the order of millions of genotypes in a matter of seconds.

Firstly and more importantly, It offers a variety of established methods for evaluating and ranking SNPs on user defined groups of populations and produces a set of pre-defined number of top ranked loci. Moreover, dataset manipulation algorithms enable users to convert datasets in different file formats, split the initial dataset into train and test and finally create datasets containing only selected SNPs occurring from the SNP selection analysis for later on evaluation in dedicated software such as GeneClass.

Although the application has been implemented in Windows Operating System it can be executed in **all operating systems** due to the fact that it is a JAVA application. We have tested the application also in Ubuntu and Mac.

In this user guide, we describe in detail every aspect of our application. Throughout this guide we provide important hints concerning not only the use of our application but also the analysis process. This guide is organized as follows: In the second chapter, we describe in detail the installation process, the requirements of the application and the way to allocate more memory for the application in order to analyze larger datasets. The third chapter describes the input data files and their exact structure. Later, the fourth chapter provides a detailed description of the application. The fifth chapter describes step by step a complete analysis scenario using TRES. Finally, in chapter six we provide contact information for any comment, suggestion or information about the application.

2. Requirements & Installation

TRES is an application developed exclusively in Java. So, in order to use it, you need to have **Java installed in your computer**. A simple way to find out if you have Java is to Google the following phrase “do I have java installed” and follow the instructions that appear in Java webpage. In case you don’t have Java, you can download it from <http://java.com/en/download/manual.jsp>

2.1 How to Install and Run TRES?

In order to **install and run TRES**, follow the steps below:

1. Download the TRES.zip from the website
(<http://mlkd.csd.auth.gr/bio/tres/downloads.html>)
2. Unzip the file to a folder of your choice. **IMPORTANT!!!! All files contained in TRES.zip should be extracted in the same folder.**
3. Double click on the TRES.jar icon.

2.2 How to RESERVE more MEMORY for the TRES?

If the dataset is too big, the user should reserve enough memory for the TRES to handle it. Even if the dataset is not too big, it is a good practice to reserve enough memory for the application. 1 Gigabyte of memory is a part of RAM that can easily be reserved in almost every average laptop with 4GB RAM, and it is adequate for TRES to handle most of datasets (e.g. A dataset with 500 individuals and 60000 thousand SNPs).

After the extraction of the zip file the user should follow the steps below:

1. Go to the command line (cmd) and get to the path where TRES.jar is.

TRES: Toolbox for Ranking and Evaluation of SNPs

2. Write the following command
 - a. For **windows**: `java -jar -Xmx1024m TRES.jar`
 - b. For **Ubuntu**: `java -jar -Xmx1024m ./TRES.jar`
 - c. For **Mac**: `java -Xmx1024m -jar TRES.jar`

TRES will now run. The only difference is that now TRES is using 1 Gigabyte Memory.

The user can change the parameter “1024” to reserve more or less memory.

USEFUL LINKS:

FAQ about JAVA: http://www.java.com/en/download/faq/whatis_java.xml

3. Data & Input File Format

TRES receives as input ARFF (Attribute – Relation File Format) files. An ARFF file is an ASCII text file that describes a list of instances (individuals) sharing a set of attributes (SNPs). It is probably the most popular file format in the field of data mining and machine learning, as it is used by the Weka machine learning library. A detailed description of ARFF files can be found in the following link. (<http://www.cs.waikato.ac.nz/ml/weka/arff.html>).

In the field of biology, there are many file formats for SNP datasets such as .PED files, HapMap files, VCF files and others. We used ARFF as a reference format and we provide a converter which converts PED files to ARFF since there are many reliable converters for example PGDSpider which convert any well known file format (HapMap, VCF, etc) to PED file format (e.g PGDSpider under the PED)A detailed description of PED files can be found in the following link (<http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml>). The application can optionally take as input a MAP file which contains information about SNPs. A detailed description of MAP files can be found in the same link with the PED files.

The following figure presents a valid ped file. More specifically, each ped file should have strictly the following format:

```
PI None 0 0 0 0 2 2 2 2 1 1 1
PI None 0 0 0 0 1 2 1 2 1 2 1
PI None 0 0 0 0 2 2 2 2 1 1 1
WB None 0 0 0 0 2 2 2 2 1 1 1
WB None 0 0 0 0 2 2 2 2 1 1 1
WB None 0 0 0 0 2 2 2 2 1 1 1
XC None 0 0 0 0 2 2 2 2 1 1 1
XC None 0 0 0 0 1 1 1 1 2 2 2
XC None 0 0 0 0 1 2 N N 1 2 1
```

Figure 1: A valid PED file

IMPORTANT NOTES!!!!

1. The first column is the **population**
2. The following **5 columns are ignored by TRES.**
3. Each allele should be represented by **only ONE number OR character**
4. Valid characters are **A,G,C,T,1,2,3,4,5,6,7,8,9**
5. MISSING VALUES should be: **0, n or N**

For a valid ped file please look at the sample dataset in TRES website. TRES currently filters data for non biallelic markers and excludes automatically from the analysis those loci with three or more alleles. Prior to using TRES, users are also advised to filter their data for linked markers.

4. Application Description

In this chapter we are going to present TRES functionalities in detail. The application consists of six tabs (figure 1):

SNP Selection: The tab where the SNP evaluation and ranking is performed.

Compare: Compare the results given from two or more evaluation methods

Converter - Splitter: User can convert a .PED file to .ARFF file and Split a PED file into two files (train and test)

Reduced Dataset Generator: The tab where user can create Genepop and Ped files based on selected SNPs

About: A tab with information about the application and the development team

4.1 SNP Selection Tab

This is the first tab of the program. In Figure 2 you can see a screenshot of the application.

Open *.arff

The "***Open *.arff***" button (1) opens a file chooser and the user can choose the SNP dataset from a specific location in the computer.

Open *.map

The dataset should be in ARFF format. The "***Open *.map***", button (2), also opens a file chooser and the user can select a map file that escorts the SNP Dataset.

The arff file and the map file should have the same name.

Groups (# individuals)

When the loading of the dataset is finished, the groups (subpopulations) of the data are presented in the **“Groups (# individuals)”** area (4). Next to the name of each subpopulation there is a number. This indicates the number of individuals that every subpopulation consists of. SNP selection will be executed on all subpopulations i.e. the process will return the SNPs which distinguish best all populations. **It is important** to mention that the user can also choose which subgroups (classes) will be used for the evaluation (figure 2 spot 4), by selecting only those. For instance, if the dataset comprises of five populations (ASW, CHB, CHD, GIH and JPT), the user can evaluate the SNPs that better distinguish specific pairs of samples or all groups together.

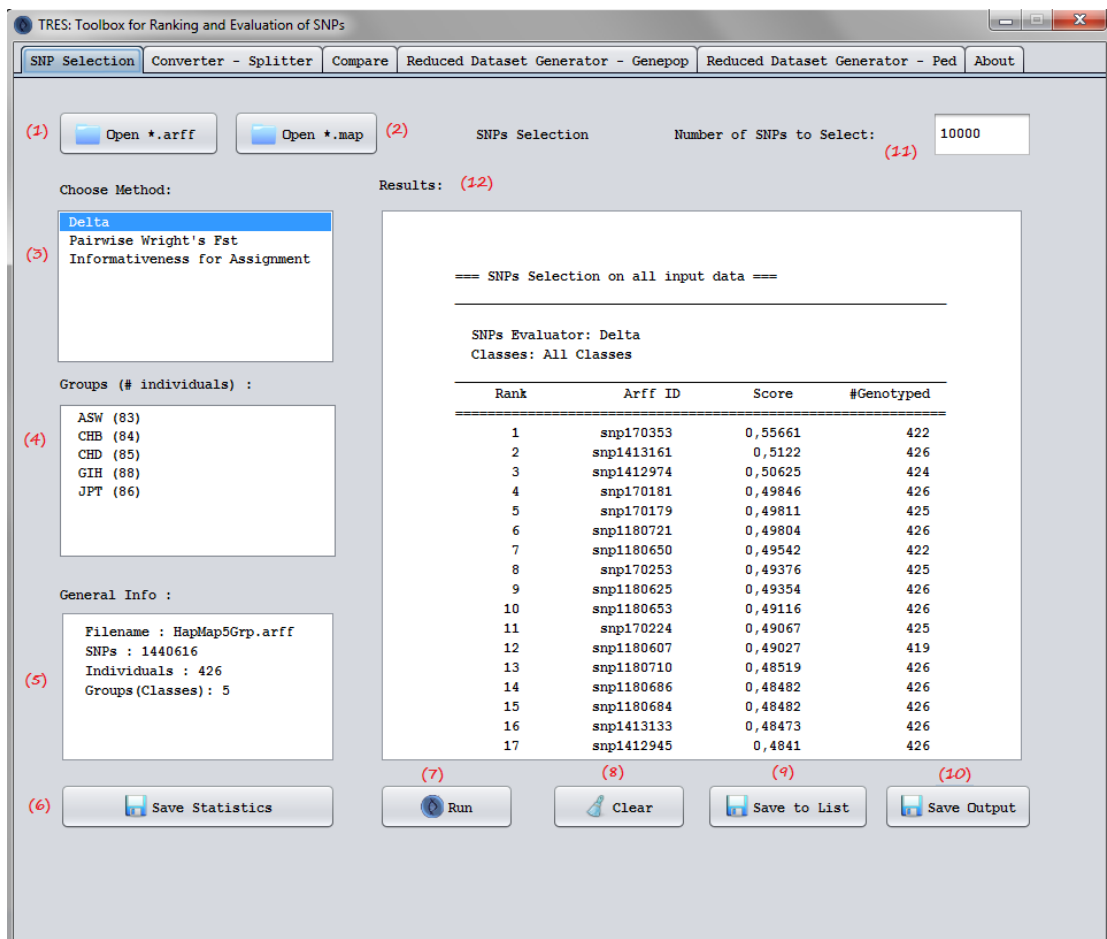


Figure 2: Screenshot of the SNP Selection Tab

General Info

Meanwhile information about the dataset is presented in the “**General Info**” area (5). Such information includes the filename, the number of SNPs, the number of individuals and finally the number of subpopulations.

Choose Method

After loading the dataset the user can choose one of the four SNP evaluation methods offered by TRES from the “**Choose Method**” area (3). The current available options are “*Delta*”, “*Pairwise WFst*”, “*Informativeness for Assignment (In)*”.

Number of SNPs to Select

In the “**Number of SNPs to Select**” area (11) the user can choose the top – k number of SNPs to be presented after the evaluation.

Results

Results are presented in the “**Results**” area (12). The results contain a list of the selected SNPs sorted by the evaluation score. On the top of the area, the application informs about the evaluation method and the subpopulations that have been considered in the evaluation. The rest information is presented in two ways depending on the existence or not of the MAP file.

Presence of Map file

In case a Map file is provided, the following information is available in the results: SNPid in the dataset which is also the position in the dataset, the score that is assigned to it by the evaluation method, the number of Individuals that have been genotyped, the SNP / Marker ID, the chromosome that it belongs, the Genetic Distance and lastly the Physical Location.

Absence of Map file

In case of absence of the map file information is limited to SNPid in the dataset, evaluation score and number of individuals genotyped.

Other Buttons

The buttons at the bottom of the application are:

Run (7): Begins the evaluation.

Clear (8): Clears the results area.

Save to List (9): Save the results in a text file. The file contains only the SNPs selected from the evaluation in a list. No other information is presented in this file. This file can be used as input in other applications such as GeneClass.

Save Output (10): Save the results in a text file. The file contains the same information which is presented in the results area.

4.2 Compare Tab

Another functionality of TRES is the comparison of methods. It is offered in the *compare* tab. The user can specify two or more evaluating methods and choose the number of top – k SNPs. The application returns the common SNPs between those methods and the corresponding percentage. For instance, the user can select to compare Delta and PairwiseWF_{ST}, for the top – 500 SNPs from a dataset of 10000 SNPs. The application will evaluate the SNPs with the two methods and then will find the common elements of those two sets of 500 SNPs. If hypothetically the application finds 200 common SNPs then the percentage is going to be $200/500 = 0.4$ or 40%.

4.3 Converter – Splitter Tab

The tab *Converter - Splitter* (Figure 3) contains two functions. The first “**Ped Converter**” converts a PED file into ARFF file. The “**Choose PED file**” button activates a file chooser in order to choose a PED file from the user’s computer. The “*Convert*” button activates a save window in order to specify the name of the arff file and the location to be saved.

TRES: Toolbox for Ranking and Evaluation of SNPs

The second function **“Ped Splitter”** splits a ped file into two new files. The **“Choose PED file”** button activates a file chooser in order to choose a PED file from the user’s computer. In the **“Choose percentage for Train Dataset”** area the user can specify the splitting percentage of the initial PED file. The **“Split”** Button begins the process of PED file splitting. The produced files are located in the same folder with the initial PED file. In scientific studies, when the purpose is to evaluate the model which has been constructed there is no correct percentage for training/test split. Common ratios are 80/20 and 70/30. When performing a split, we want to have a higher proportion in the training in order to correctly adjust the model, then a smaller percentage to test on. It depends on the researcher to decide.

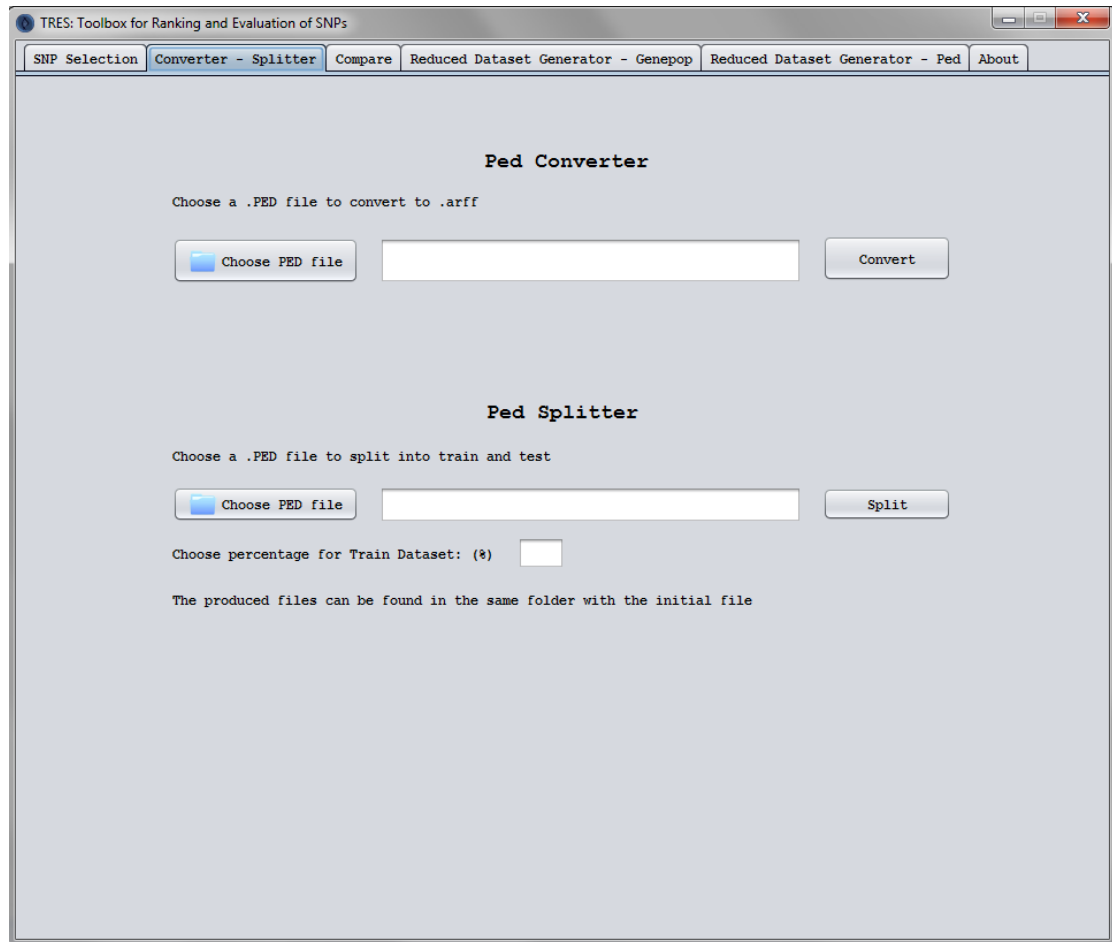


Figure 3: Converter - Splitter Tab

The names of the new files are the same with the initial file plus **“Train”** and **“Test”** ending respectively. For instance, if the full dataset is called `fugu.ped`, the new

TRES: Toolbox for Ranking and Evaluation of SNPs

files will be called `fuguTrain.ped` and `fuguTest.ped` for the train and test file respectively.

The application will inform you with an error message if:

- Your input is not a ped file
- Found string in the percentage box
- Found number X which is $X < 0$ or $X > 100$

4.4 Reduced Dataset Generator Tabs

The tab “Reduced Dataset Generator -Genepop” (figure 4) offers the ability to the user to create Genepop files using only a subset of SNPs. In this case the user can directly use the generated datasets in GeneClass software, in order to evaluate the assignment accuracy of the SNP subset.

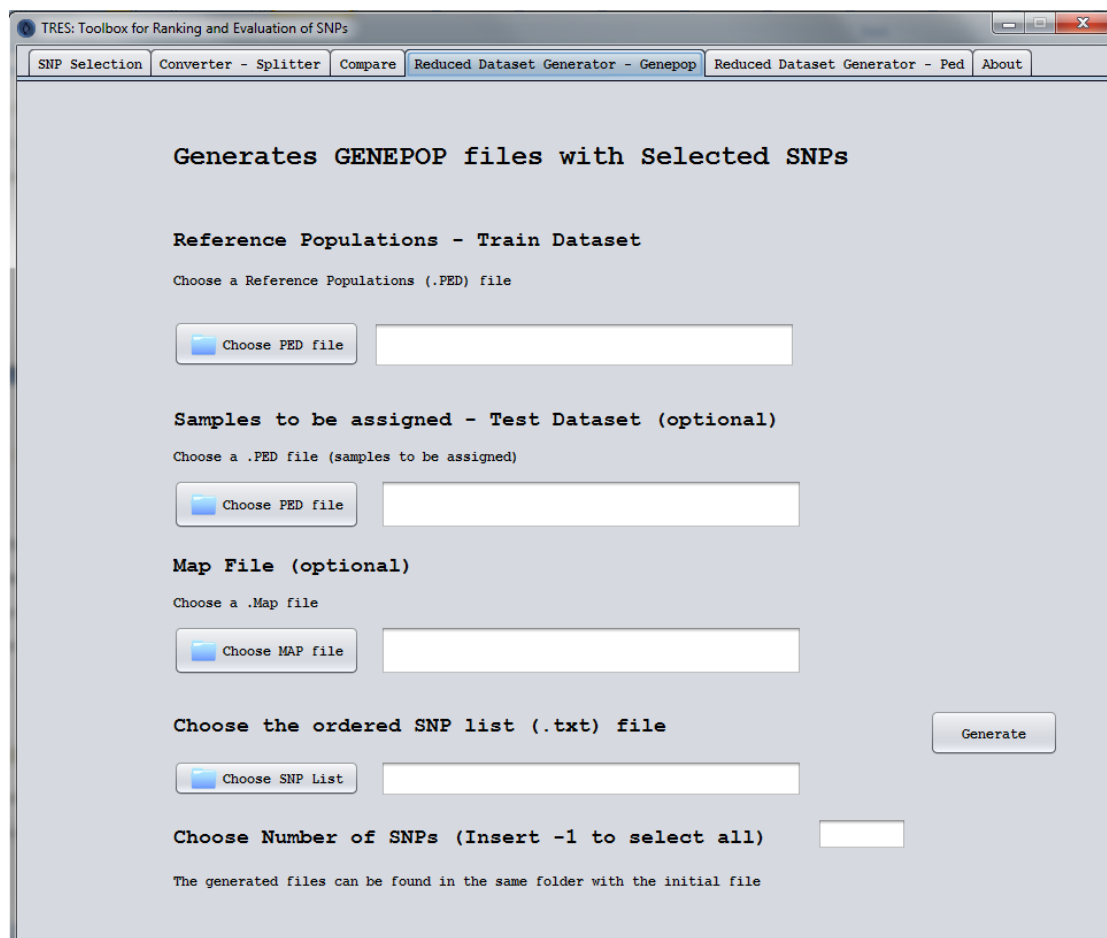


Figure 4: The Genepop File Generator Tab

TRES: Toolbox for Ranking and Evaluation of SNPs

The user can choose the reference population file (train file) and the “samples to be assigned” file (test file) from the two “Choose PED file” buttons. Those datasets have occurred by the ped splitter. The user should also give the application the list of the ordered SNPs which has been provided by the SNP Selection tab. Finally the user can provide the number of top-k informative SNPs that will be contained in the newly created files. Similarly, user can also create Ped reduced datasets in “Reduced Dataset Generator – Ped” tab.

4.5 About Tab

The “About Tab” has information about the Scientific Groups that have been involved in various ways in the application development.

5. How to Use TRES – A Step by Step Analysis Process

In this paragraph we are going to present an example of using TRES step by step. The scenario is the following:

The user has a SNP dataset in ped file format (TRESSampleData.ped available at the software website). The dataset contains 52 individuals genotyped at 873SNPs. The user wants to evaluate the SNPs with the Delta method. The assignment success of the SNP selection will be evaluated in GeneClass software, so the user wants to split the dataset into train and test with a 70 – 30 percentage split.

The steps are the following:

5.1 Downloading, Installing and Running TRES

- Download the TRES.zip file from the applications website.
<http://mlkd.csd.auth.gr/bio/tres/>
- Extract the zip file to a folder of your choice.
- Double click on the TRES icon in order to run TRES. **IMPORTANT!!!! All files contained in the TRES.zip should be extracted in the same folder.**

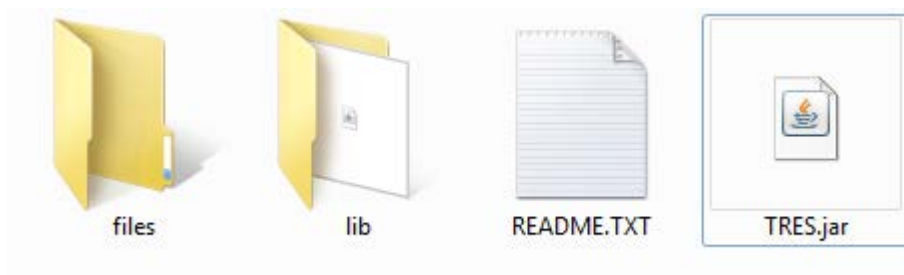
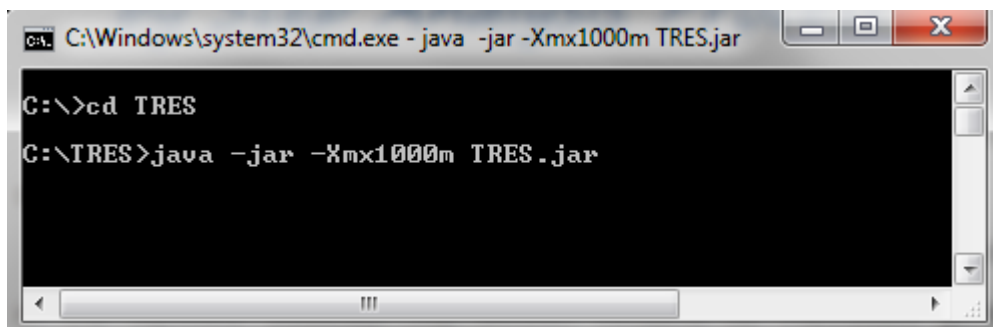


Figure 5: All files and folders from the TRES.zip file must be placed in the same folder.

- If you want to reserve more memory for the application, open the command line (cmd), go to the folder where the TRES.jar is and write the appropriate command (see section 2.2) depending on your operating system (Figure 6). For instance, if the application is extracted to a folder named “tres” which is located in “c:” then the command line you should type the following:

A screenshot of a Windows command prompt window. The title bar reads 'C:\Windows\system32\cmd.exe - java -jar -Xmx1000m TRES.jar'. The command prompt shows the following text:

```
C:\>cd TRES
C:\TRES>java -jar -Xmx1000m TRES.jar
```

Figure 6: Reserving 1GB RAM for TRES

5.2 Splitting the Dataset into Train and Test

- Once the TRES application starts go to the “**Converter - Splitter**” tab.
- Click on the “**Choose PED file**” button on the Ped Splitter section (Figure 7). Once you click the button, a file chooser opens in order to select the dataset (ped file). Find the dataset in your hard disk and choose open.
- In the appropriate box fill in the value “70”, which is the train file percentage.
- Then click on the “**Split**” button to start the file splitting. The newly created files will be placed in the same folder with the original dataset with the names TRESSampleDataTrain.ped, containing approximately 70% of the initial individuals and the TRESSampleDataTest.ped which contains ~30%. We note the percentages approximately because it is not always possible to split the dataset into the given percentage. For instance, 70% of 45 individuals are 31,4. So the train file contains 31 individuals and the test file contains 14 individuals which correspond to 68,9% and 31,1% respectively. TRES will inform you when the conversion has successfully finished.

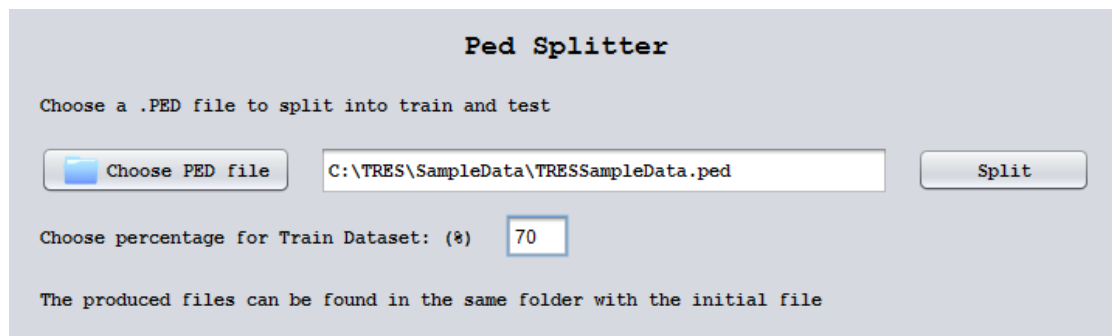


Figure 7: Ped Splitter Section

5.3 Converting the PED file Dataset into ARFF

- Click on the “**Choose PED file**” button (Figure 8). Once you click the button, a file chooser opens in order to select the dataset (ped file). Find the dataset in your hard disk and choose open.

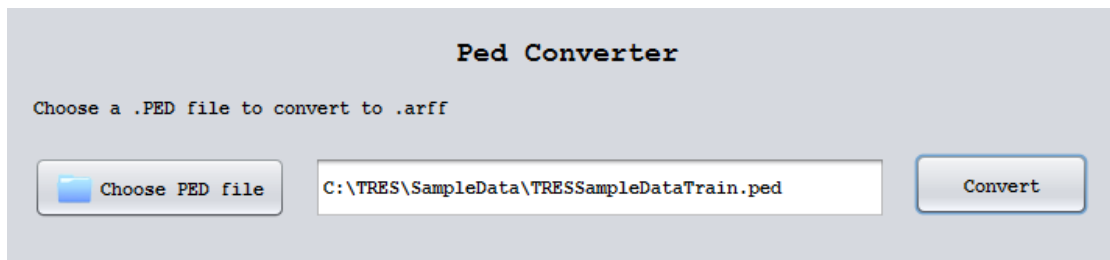


Figure 8: Converter section

- Then click on the “**Convert**” button. A new window opens in order to choose the name and the save location of the converted file. Once you choose destination and name for the converted file, the conversion will start. The TRES will inform you when the conversion has successfully finished.

5.4 SNP Selection

- In order to start the snp selection process with method delta, go to the “**SNP Selection**” Tab.
- Click on the “**Open arff**” button and find the converted file on your disk. You can also choose a map file optionally. **IMPORTANT!** The arff file and the map file should have the same name.
- All the information about the dataset is shown in the “**General Info**” area. The dataset named TRESSampleDataTrain.ped contains 35 individuals (~70% of the initial dataset) divided in 3 classes (populations - groups). The dataset contains 873 SNPs.

TRES: Toolbox for Ranking and Evaluation of SNPs

- The names of the Groups are shown in the “**Groups**” area (Figure 9). Each group is mentioned with the number of individuals belonging to this group. In this dataset there are 3 groups; PopA with 11 individuals, PopB with 15 individuals and PopC with 9 individuals.
- In the “**Number of SNPs to Select**” area write the number of the top – k SNPs that the evaluation will return (e.g. 100).
- Choose the evaluation method from the “**Choose Method**” area (e.g. Delta).
- Choose the populations that are going to be considered in the evaluation from the “**Groups**” area (e.g. PopA and PopB).
- Next click the “**Run**” button in order to start the evaluation.
- The results are presented in the “**Results**” area.

The screenshot shows the TRES software interface with the following components:

- SNPs Selection:** A text input field containing the value "100".
- Choose Method:** A dropdown menu with "Delta" selected. Other options include "Pairwise Wright's Fst" and "Informativeness for Assignment".
- Groups (# individuals):** A list box containing "grp_A (11)", "grp_B (15)", and "grp_C (9)".
- General Info:** A text area showing "Filename : TRESSampleDataTrain.", "SNPs : 873", "Individuals : 35", and "Groups (Classes) : 3".
- Results:** A text area displaying the following table:

Rank	Arff ID	Score	#Genotyped
1	snp339	0,74242	26
2	snp553	0,71678	24
3	snp489	0,62727	26
4	snp698	0,61818	26
5	snp654	0,60909	26
6	snp684	0,60606	26
7	snp332	0,60303	26
8	snp379	0,6	26
9	snp136	0,58788	26
10	snp442	0,57576	26
11	snp347	0,5697	26
12	snp704	0,56364	26
13	snp385	0,56061	26
14	snp870	0,55455	26
15	snp865	0,55152	26
16	snp864	0,55152	26
17	snp501	0,54242	26
18	snp291	0,53333	26
- Buttons:** "Open *.arff", "Open *.map", "Save Statistics", "Run", "Clear", "Save to List", "Save Output".

Figure 9: SNP Selection Tab

- You can save the results by clicking the “**Save Output**” Button.

- For the next step you should save the list of the SNPs using the **“Save to List”** button.

5.5 Reduced Dataset Generation

- In order to generate files for the GeneClass, go to the Reduced Dataset Generation tab (figure 10)
- Click on the **“Choose PED files”** button to choose the train and test dataset that have been produced from the Splitting process.
- Click on the **“Choose SNP List”** button in order to select the ordered list of SNPs that have been produced by the SNP selection process.
- Lastly choose the number of the top-k SNPs that the Genepop files will contain.
- The generated files can be found in the same folder with the Initial file with almost the same name with the initial datasets. The only difference is that the newly created datasets contain the number of containing SNPs

TRES: Toolbox for Ranking and Evaluation of SNPs

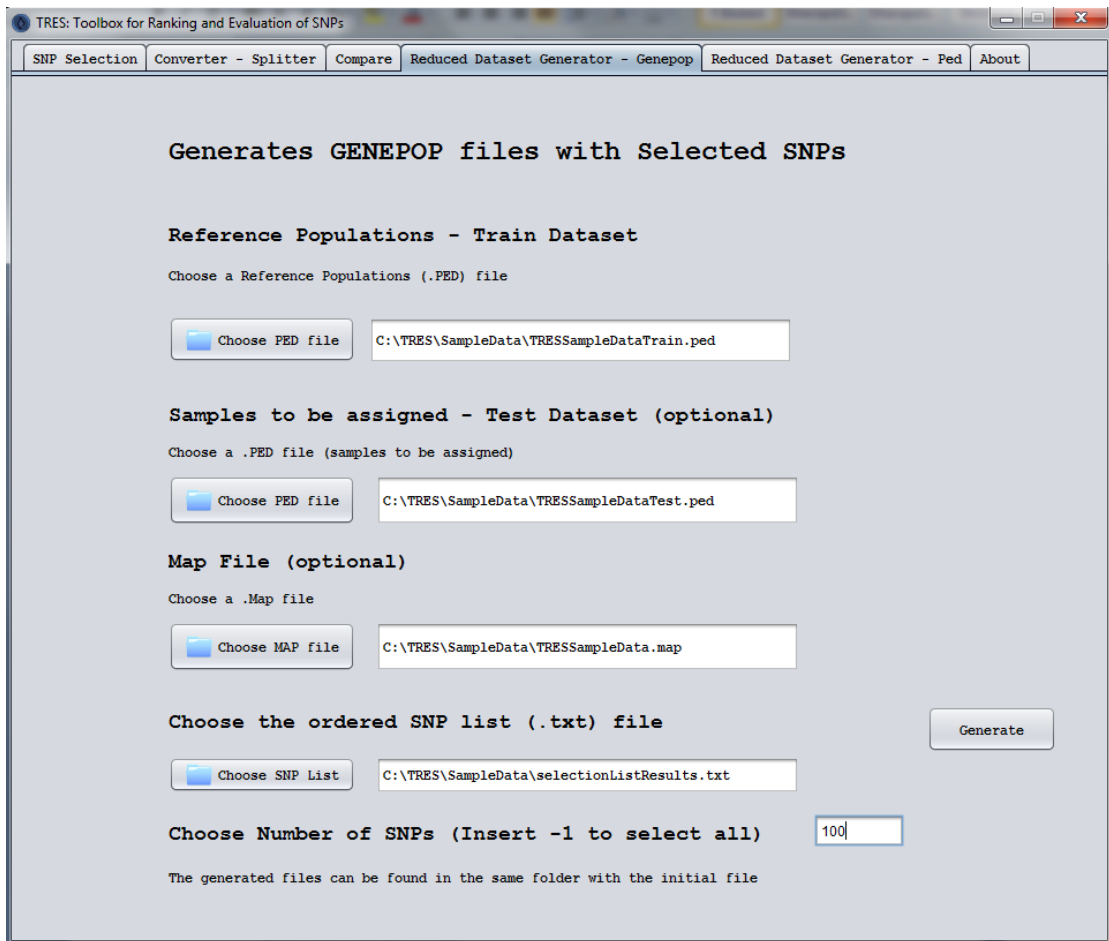


Figure 10: Generating Genepop Files

- You can also save information about allele and genotype frequencies related to each population by clicking in the “**save statistics**” button (figure 11).

	#Genotyped	1st Allele	2nd Allele	Genotypes		
SNP339		freq: T	freq: C	CC	CT	TT
grp_A	11	0,091	0,909	0,818	0,182	0
grp_B	15	0,833	0,167	0	0,333	0,667
Total	26	0,519	0,481	0,346	0,269	0,385

	#Genotyped	1st Allele	2nd Allele	Genotypes		
SNP553		freq: A	freq: C	AC	AA	CC
grp_A	11	0,091	0,909	0,182	0	0,818
grp_B	13	0,808	0,192	0,231	0,692	0,077
Total	24	0,479	0,521	0,208	0,375	0,417

Figure 11: Statistics for the first two SNPs of the selection

5.6 Hints on using GENECLASS2

A complete guide to GENECLASS is provided in software's website.

(<http://www1.montpellier.inra.fr/URLB/GeneClass2/Help.pdf>)

We would also like to mention that GenoDive is a similar program to GENECLASS2 that is available for Macintosh Computers

(<http://www.bentleydrummer.nl/software/software/GenoDive.html>).

6. Contact Information

For any comments, suggestions or any other information, don't hesitate to contact Ioannis Kavakiotis (ikavak@csd.auth.gr).

The following groups and institutions were involved in various ways in the development process.

Machine Learning and Knowledge Discovery Group



(<http://mlkd.csd.auth.gr>)

Department of Informatics

Aristotle University of Thessaloniki

Population Genetics of Animal Organisms

School of Biology

Department of Genetics, Development & Molecular Biology

Aristotle University of Thessaloniki

Animal Breeding and Genomics Centre

Wageningen University