



Università degli Studi di Sassari

**SCUOLA DI DOTTORATO DI RICERCA**  
**Scienze dei Sistemi Agrari e Forestali**  
**e delle Produzioni Alimentari**



Indirizzo Produttività delle Piante Coltivate

Ciclo XXIV

Development and use of statistical and bioinformatic tools for the  
analysis of linkage disequilibrium in plant populations

dr. Marco Vargiu

<i>Direttore della Scuola:</i>	prof. Giuseppe Pulina
<i>Referente di Indirizzo</i>	prof. Antonino Spanu
<i>Docente Guida</i>	dr. Domenico Rau

Triennio accademico 2009- 2011

# **Development and application of statistical methods and bioinformatic tools for the analysis of plant genomes and populations**

## General index

- *User's manual of Nuragene software.* **page 3**
- *An example of the application of Nuragene software.* **page 47**
- *Bibliography* **page 81**

User's manual

## NURAGEN ver 1.0

### **authors**

Domenico Rau, Marco Vargiu

Scuola di dottorato in

Scienze e biotecnologie dei Sistemi Agrari e Forestali e delle Produzioni alimentari

Università di Sassari

Dipartimento di Scienze Zootecniche - Via Enrico de Nicola

07100 Sassari

Italia

E-mail: [dmrau@uniss.it](mailto:dmrau@uniss.it); [mavargi@tin.it](mailto:mavargi@tin.it);

AD 2011

## Summary

---

➤	1 Overview .....	6
➤	2 Installation .....	8
➤	3 Getting started .....	8
	3.2 Loading an input file .....	8
	3.3 How to perform calculations.....	10
	3.3.1 Calculate $r_d$ index.....	10
	3.3.2 Test of significance of the $r_d$ index.....	10
	3.3.3 Calculate $r_d$ index with the option “Standardize for population size”.....	12
	3.3.4 Calculate Brown and Feldman’s variance components for the entire dataset (the VarCom – All) .....	13
	3.3.5 Test of significance of the Brown and Feldman’s variance components for the entire dataset (VarCom – All).....	14
	3.3.6 Genomic sliding window analysis of the variance components – VarCom - All16	
	3.3.7 Calculate the variance components overall loci and for pair of populations (VarCom-PWPops).....	18
	3.3.8 Significance of the variance components for pair of populations (VarCom – PWPops).....	19
	3.3.9 Correlation between variance components and geographical and ecological variables (VarCom – PWPops).....	20
	3.3.10 Calculate the variance components for pair of loci (VarCom – PWLoc1).....	21
	3.3.11 Significance of the variance components for pair of loci (VarCom-PWLoc1).....	23
	3.3.12 Correlation between variance components and genetic distance (VarCom-PWLoc1).....	23
	3.3.13 Significance of the correlations between variance components and geographical and ecological variables (VarCom-PWLoc2).....	24
	3.4 Batch .....	26
	3.4.1 Massive input file .....	26
	3.4.2 Post processing search.....	27
	3.5 Options .....	28
➤	4 Input/output file.....	28
	4.1 Input file.....	28
	4.1.1 Flat Style.....	28
	4.1.2 Flat Style with loci position.....	29
	4.1.3 EasyPop output files.....	30
	4.1.4 Input file for geographical coordinates .....	30
	4.1.5 Input file for ecological variables.....	31
	4.1.6 Input file “Easypop + grp defs”.....	31
	4.2 Output file .....	32
	4.2.1 $r_d$ – Test of significance.....	33
	4.2.2 $r_d$ – Standardize for sample size .....	33
	4.2.3 VarCom – All. Tests of significance .....	34
	4.2.4 VarCom – All. Sliding window analysis.....	34
	4.2.5 VarCom - PWLoc1. Tests of significance.....	35
	4.2.6 VarCom – PWloc1. Correlation with genetic distances.....	35

4.2.7 VarCom – PW Pops. Test of significance.....	36
4.2.8 VarCom – PW Pops. Correlation with geographical and ecological variables. .	36
4.2.9 VarCom – PW Loci2. Correlation between variance components and geographical and ecological variable .....	37
➤ 5 Methods .....	38
5.1 Randomizations.....	38
5.1.1 Mode 1- shuffling of alleles (among individuals, within populations). .....	38
5.1.2 Mode 2- permutation of individuals among populations. ....	39
5.1.3 Mode 3 - shuffling alleles and permuting individual simultaneously .....	41
➤ 6 Statistics.....	42
6.1 $r_d$ .....	42
6.2 Components of the variance for the number of heterozygous loci in several populations. ....	43
6.3 Correlation and partial correlation analyses.....	45
➤ Handling missing data .....	46

## ➤ 1 Overview

### *Purpose.*

This program has been written to facilitate the analysis of multi-locus population genetic data. In particular, it allows calculation and testing linkage disequilibrium indices, useful for analysis in single and multiple populations.

In particular, Nuragen implement the method of Brown and Feldman (1981) whereby the structure of multilocus associations among and within several populations can be partitioned into its components. The components are measured by their contributions to the variance in the number of heterozygous loci in two randomly chosen gametes. The single-locus components are the average and the variation among populations in gene diversity (MH and VH) and the variance among populations in allele frequency (WH). The two-locus components include the mean (MD) and variance (VD) of disequilibria, the covariance of allele frequencies over populations (WC), and various interactions (AI and CI).

Nuragen allows calculation of such components at several levels (see Table 1 here below for more details). The ultimate objective is to give a contribution in the understanding of the patterns and cause of LD in multiple populations. In addition, there are randomization routines that allow one to test various null-hypotheses. Moreover, additional functions allow the filtering of the data based on different criteria and on the components values and their significance. Finally, a batch routine allows the analyses of dataset obtained by EasyPop (Balloux 2006), a forward-in-time population genetic simulator.

**Table** – Calculations implemented in Nuragen software.

<i>Statics computed</i>	<i>Description</i>	<i>Utility</i>
$r_d$ (Burt et al. 1996)	<i>A standardized measure of multilocus linkage disequilibrium</i>	<i>Useful to quantify the degree of LD within each single population.</i>
<i>Standardization of <math>r_d</math> for sample size</i>	<i>Given a set of populations, the software allows the resampling</i>	<i>Useful to compare and rank a set of populations with</i>

	<i>without re-immission of the same number of individuals from each population. It is possible to set the number of replicates.</i>	<i>different sample size based on LD.</i>
<p><i>Brown and Feldman (1981) variance components.</i></p> <p><b>Single-locus components</b>  Mean gene diversity (MH)  Variance of diversity (VH)  Wahlund's effect (WH)</p> <p><b>Two-locus components</b>  Mean disequilibrium (MD)  Covariance of allele frequencies over populations (WC)  Interaction between MD and WC (AI)  Variance of disequilibrium (VD)  Covariance of interaction (CI)</p> <p>Total variance  (MH+VH+WH+MD+WC+AI)</p> <p>Average Variance  (MH+MD+AI+VD+CI)</p>	<i>Let K the number of heterozygous loci between two randomly chosen gametes in a population, this method partitions the total and average variances of K in a mixed pool of several populations into single-locus and two-locus components (Brown and Feldman 1981).</i>	<i>Useful to describe the population structure of genetic diversity and of multilocus associations in a set of populations and overall loci</i>
<i>Brown and Feldman's (1981) method - genomic sliding window</i>	<i>When markers positions are specified, the Brown and Feldman's (1981) method can be applied performing a sliding window analysis of the genome.</i>	<i>Useful to surf the genome searching for genomic regions with a peculiar structure of LD.</i>
<i>Brown and Feldman's (1981) method - pair of populations</i>	<i>Brown and Feldman's (1981) method is applied for all possible <math>np = p(p-1)/2</math> pairs of populations separately, given p the number of populations in the dataset.</i>	<i>Useful to understand the relationships between populations based on the population structure of the multilocus associations.</i>
<i>Brown and Feldman's (1981) method - pair of loci</i>	<i>Brown and Feldman's (1981) method is applied for all possible <math>nm = m(m-1)/2</math> pairs of markers separately, given</i>	<i>Useful to search and to identify loci pair with outlying behaviour.</i>



	<i>m</i> the number of markers in the dataset.	
<i>Correlation and partial correlation between the Brown and Feldman's variance components and the geographical and ecological variables</i>	<i>The significance of the associations is tested by Mantel test. Available both for pair of populations and for pair of loci options.</i>	<i>It allows inferring the relative impact of geographical distances and ecological differences among population's sampling sites in shaping the population structure of multilocus associations.</i>
<i>Correlations between variance components and genomic distance between pair of loci</i>	<i>When markers positions is specified, significance of the associations is tested by Mantel test.</i>	<i>Useful to investigate the genomic structure of the multilocus associations.</i>

## ➤ 2 Installation

Nuragen is written in VB.Net. Versions are available to run on Macintosh (68K and PowerPC) and PC computers. To install the program simply unzip it and double click on "Setup\_Nuragene.msi" and follow the instructions.

## ➤ 3 Getting started

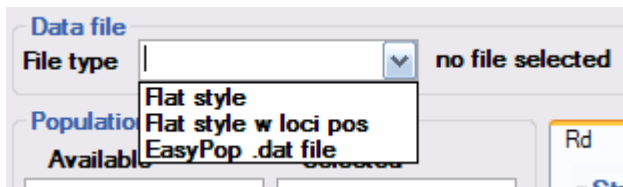
### 3.2 Loading an input file

*Note.* To perform the analyses implemented in Nuragene software, the gametic phase must be known or estimated or the investigated organism be haploid. The software can accept input files with dominant markers such as AFLPs or similar (SSAP, SAMPLE, etc...). Loci must have two possible alleles, 0 (band absent) or 1 (band present). However, Nuragene can manage biallelic SNPs provided that substitutions at variable positions are recoded (by the user) as 0 and 1.

First, you must select and must choose the file type. Nuragen accepts three different file

formats named:

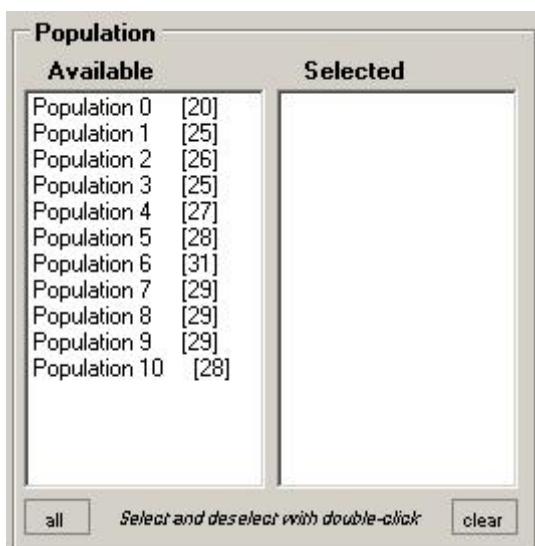
- flat style (see 4.1.1);
- flat style with pos loci (see 4.1.2);
- EasyPop . dat file (see 4.1.3);



If file loading is successful, in the log area you should read the file path, the name of the file and the number of loci and of populations in your dataset.



In the area “Population”, all available populations are listed together with the [sample size] of each population.



It is possible select (or deselect) the populations to analyze by double clicking.

### 3.3 How to perform calculations

Different analyses can be invoked

#### 3.3.1 Calculate $r_d$ index

To perform this calculation you must:

1. select ONE population;
2. select the “ $R_d$ ” panel .
3. click “Start computation” button.

The screenshot shows a software interface for calculating the  $r_d$  index. The 'Rd' panel is active, showing options for standardization, significance testing, and randomization. The 'Standardize for sample size' section has a checkbox, 'Items n°' set to 1, and 'Random cycles' set to 10. The 'Test of significance' section has a checkbox, 'Random cycles' set to 1, and 'N° of CPU' set to 1. The 'Randomization' section has radio buttons for 'mode 1', 'mode 2', and 'mode 3'. A 'Start computation' button is located at the bottom of the panel.

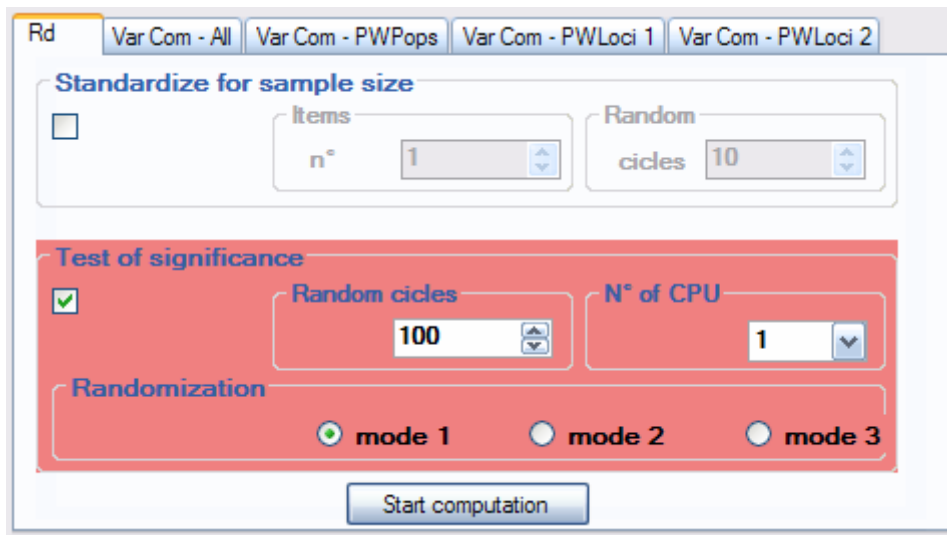
The result will be displayed in the log area.

#### 3.3.2 Test of significance of the $r_d$ index.

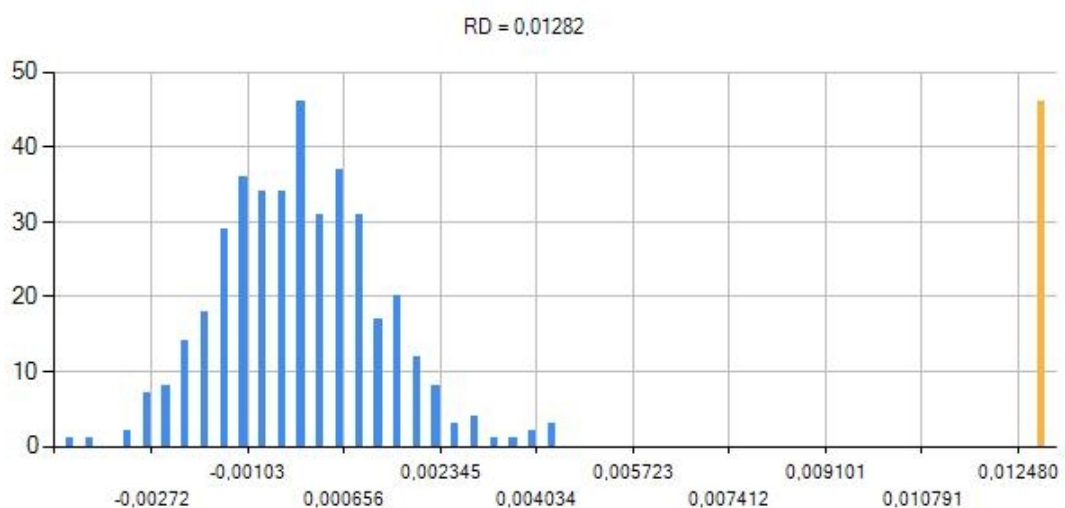
In this case, the user must:

1. Select ONE population;
2. Select the “ $R_d$ ” panel;
3. Activate the option “test of significance” (the area will become red coloured);
4. Specify the number of randomizations;

5. Specify the number of CPU involved in the calculations (the software automatically detects the number of available CPU in you computer);
6. choose and select the type of randomization (mode1, mode 2 or mode 3) desired to build the null distribution rd values;
7. Click “Start computation” button.



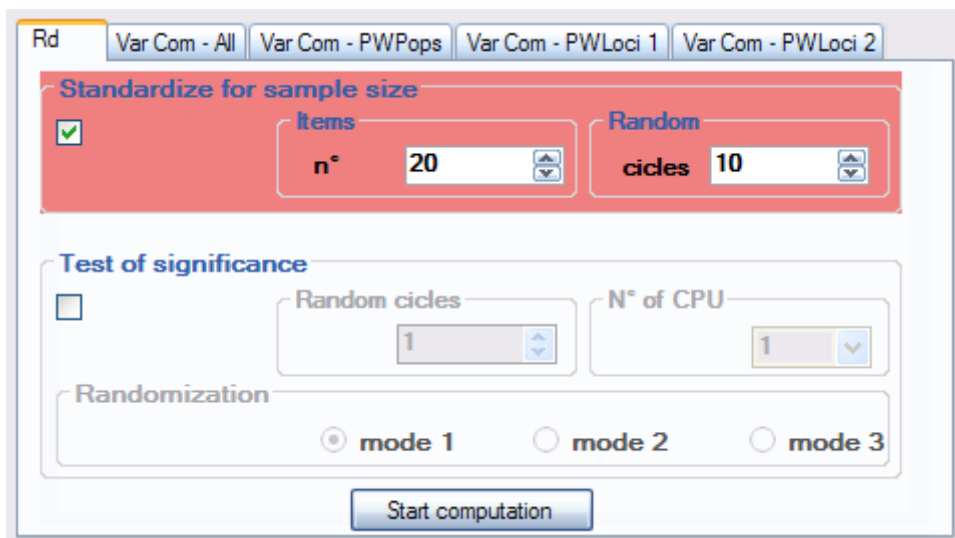
At the end of the calculations the software automatically displays a graph that shows the distribution frequency of the  $r_d$  values obtained by randomization (in blue). In the same graph, the observed  $r_d$  values is reported (in orange) to allow direct comparison with the null-distribution. At the top of the graph, the observed  $r_d$  value is reported.



### 3.3.3 Calculate $r_d$ index with the option “Standardize for population size”

In this case, the user must:

1. Select the “Rd” panel;
2. activate the option “Standardize for sample size” (the area will become red coloured). Activating this option the software automatically selects ALL populations in your datafile for subsequent calculations.
3. Specify the number of items (individuals) to sample from each population (the maximum number of individuals allowed is the sample size of the smallest population in your dataset);
4. Specify the number of sampling events;
5. Click the “Start computation” button.



At the end of the calculation the log area will show  $r_d$  average values and variances for each population. Moreover, the software will automatically display a graph where all populations are compared simultaneously.

	Mean	Variance
Population 0	0,0128178686050451	0
Population 1	0,029673870858585	3,9955695727234E-05
Population 2	0,0319601743057089	1,79049724868116E-05
Population 3	0,0126016810512499	8,90954346632397E-06
Population 4	0,0272251394133835	8,82740646054798E-06
Population 5	0,00637205756475374	2,1280508289676E-06
Population 6	0,0159023343249073	1,33024069586219E-05
Population 7	0,010670773749429	5,38955088417501E-06
Population 8	0,0101025732335647	1,94969482010478E-06

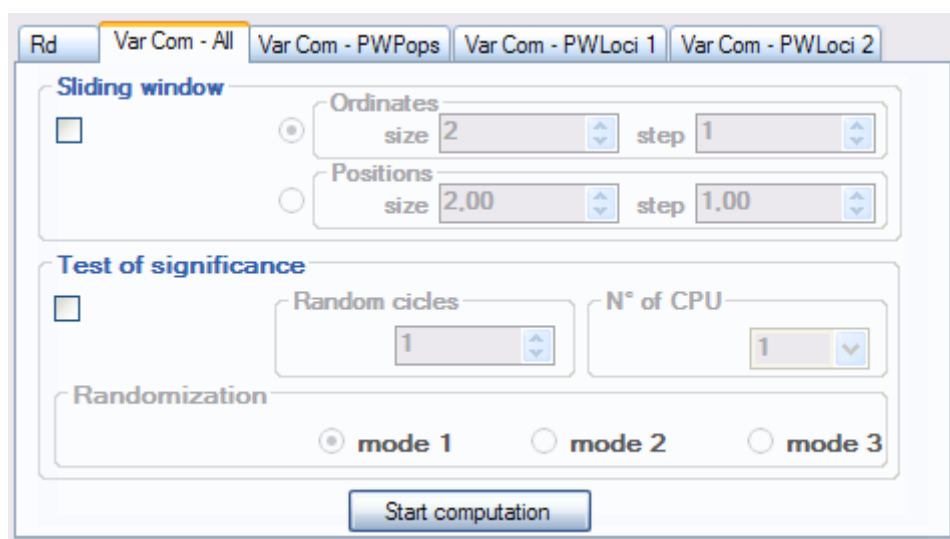
A graph showing simultaneously the distribution or  $r_d$  for each population is produced.

### 3.3.4 Calculate Brown and Feldman's variance components for the entire dataset (the VarCom – All)

This calculation is performed considering all individuals, all populations, and all loci in your dataset.

In this case, the user must:

1. Selects AT LEAST two populations;
2. select the “VarCom-All” panel;
3. Click “Start computation” button.



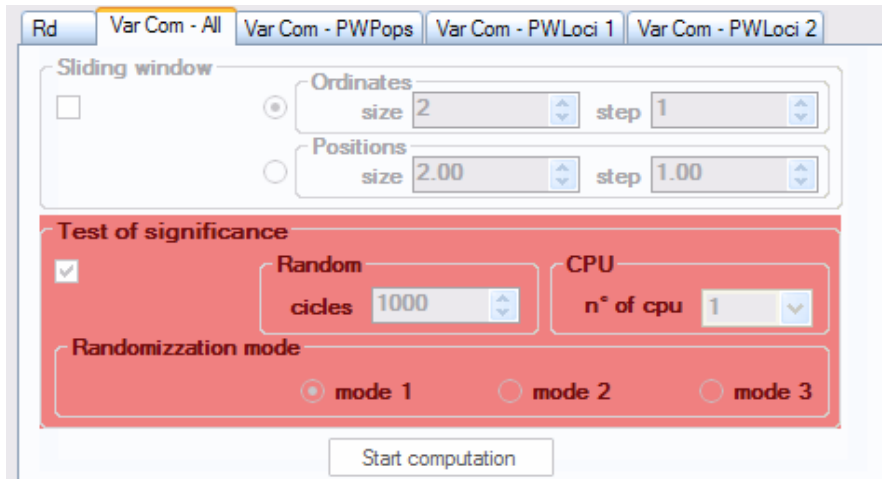
The log area will display the results for all the components and for the variance of the average and of the total population.

```
Logs
MH: 18,4824621552395
VH: 2,16974196780586
WH: 2,41735482507549
MD: 10,1287247921575
AI: 3,22247548768719
WC: 13,5330261156
VD: 39,799095549076
CI: 6,5673389189624
mh + md + ai + vd + ci: 78,2000969031226
mh + vh + wh + md + wc + ai: 49,954
(md + ai + vd + ci) / mh: 3,23104325853869
md / mh: 0,548018154025336
ai / mh: 0,174353149522
vd / mh: 2,15334381397846
ci / mh: 0,355328141012893
```

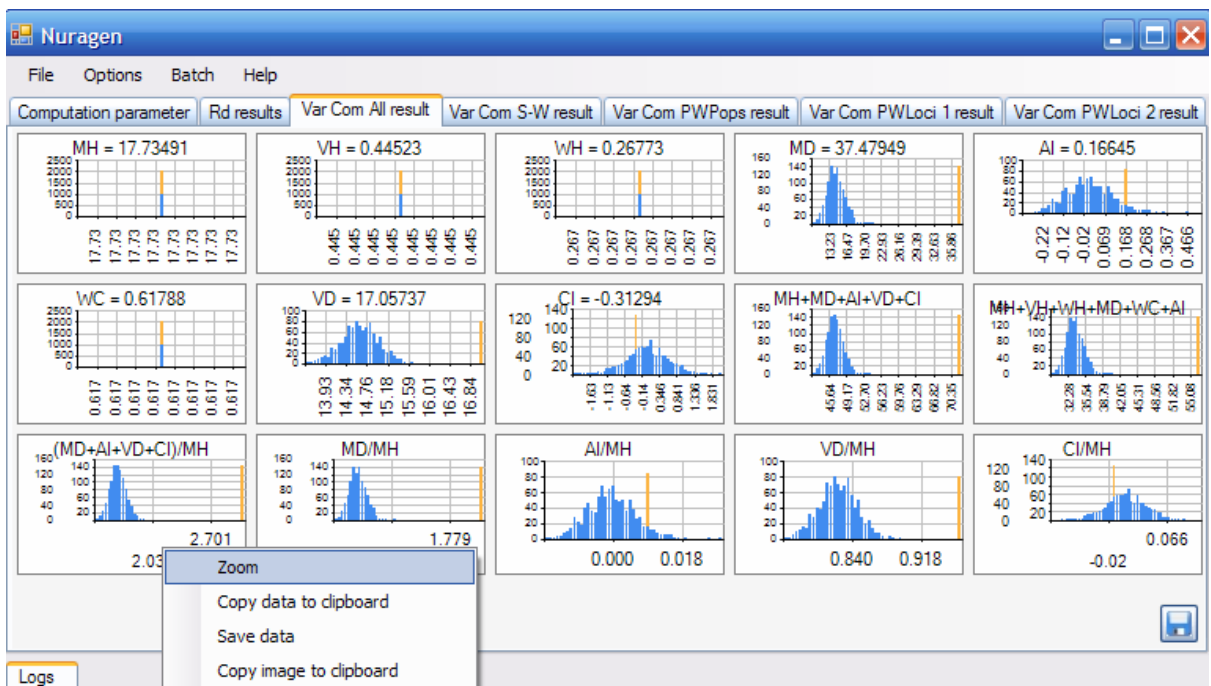
### **3.3.5 Test of significance of the Brown and Feldman's variance components for the entire dataset (VarCom – All)**

In this case the user must:

1. Select at least two populations;
2. Select the “Var Com - All” panel;
3. Activate the “Test of of significance” option;
4. Specify the number of randomizations;
5. Specify the number of CPUs involved in the calculations (the software automatically detects the number of available CPU in your computer);
6. choose and select the type of randomization (mode1, mode 2 or mode 3) desired to build the null-distribution of the Var-Com - All values;
7. Click the “Start computation” button.



At the end of the calculation, the software automatically displays 15 graphs (figure here below). They represent the null distributions of the eight variance components plus the null distribution for the average (MH+MD+AI+VD+CI) and of the total population (MH+VH+WH+MD+WC+AI). LD of the average population and the components MD, AI, VD and CI are also present after standardization by dividing for MH. The distributions of the randomized values are in blue. A vertical orange bar indicates the position of the observed value.

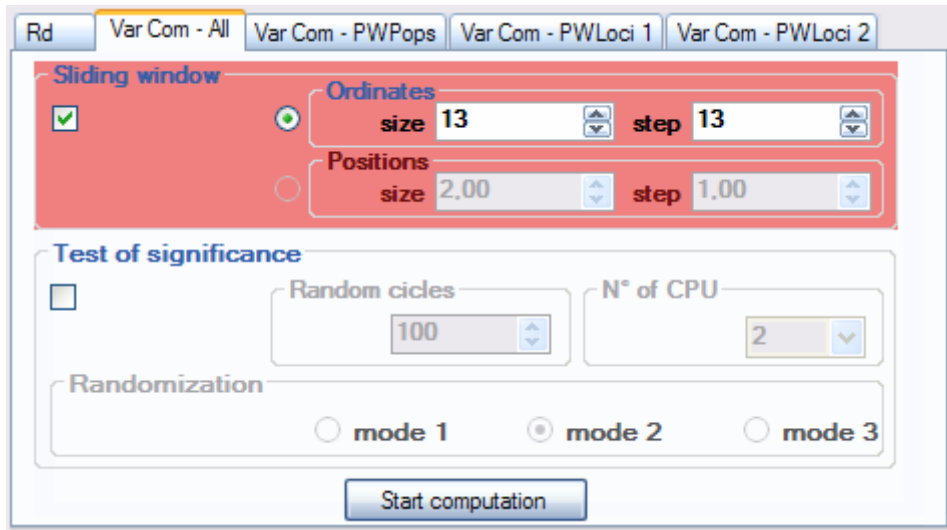




Right-clicking on each graph it is possible to choose among four options: zoom the graph, copy the data that have generated the graph on the clipboard, save the data as a .csv file and copy the image on the clipboard.

### **3.3.6 Genomic sliding window analysis of the variance components – VarCom - All**

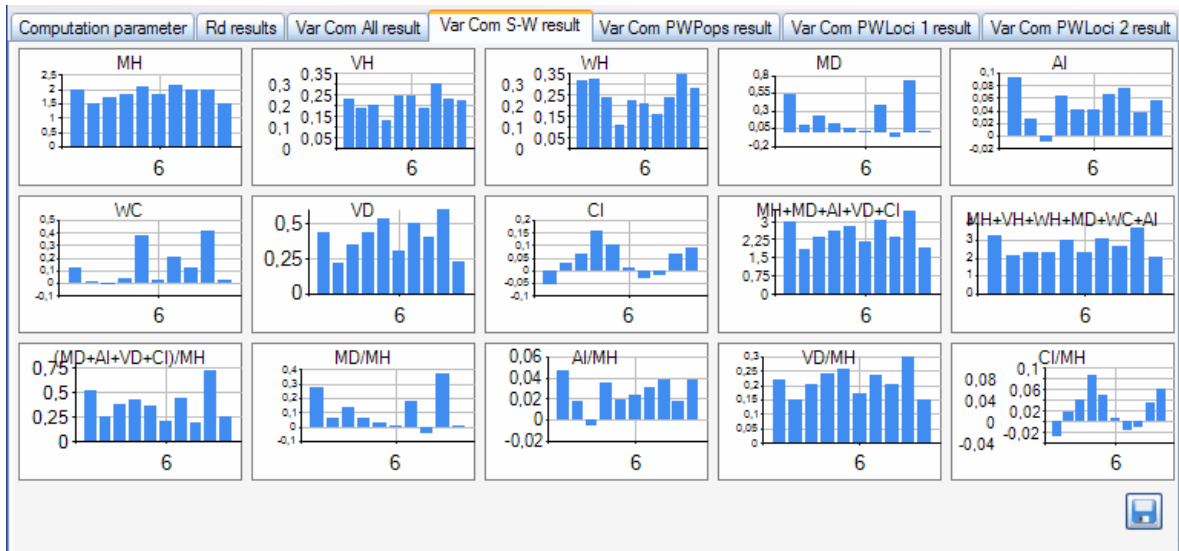
1. Select AT LEAST two populations;
2. Select the “Var Com - All” panel;
3. activate the “Sliding window” option (the area will become red coloured).
4. Choose between two different modes of performing the sliding window analysis.
  - ordinates: windows are based just upon the order of the loci in your input data file. As an example, if you choose a window of size 20 and with a step of 5, the calculation of the components will be performed for the first window considering the first 20 loci (columns) in your datasets, for the second windows considering the loci from the sixth to the 26th, etc.. This option allows performing analyses with windows that have a number of markers, but that not necessarily have same genomic size.
  - positions: windows are based upon the genomic position in bp. As an example, if you specify a window of 20 this means that computations will be performed considering intervals of 20 bp (not of 20 markers!). This analysis allows performing analyses with windows of constant genomic size but (in general) with variable number of markers.
5. Specify the window size (this cannot exceed 1/10 of the number of markers or of the total length of the genomic region investigated);
6. Specify the step size (this cannot exceed the window size);
7. Click the "Start computation" button.



At the end of the computations, the software automatically displays 15 graphs were for each component the results for all the windows are shown.

By clicking the button on the right-low corner, it is possible to save raw data as .csv file.

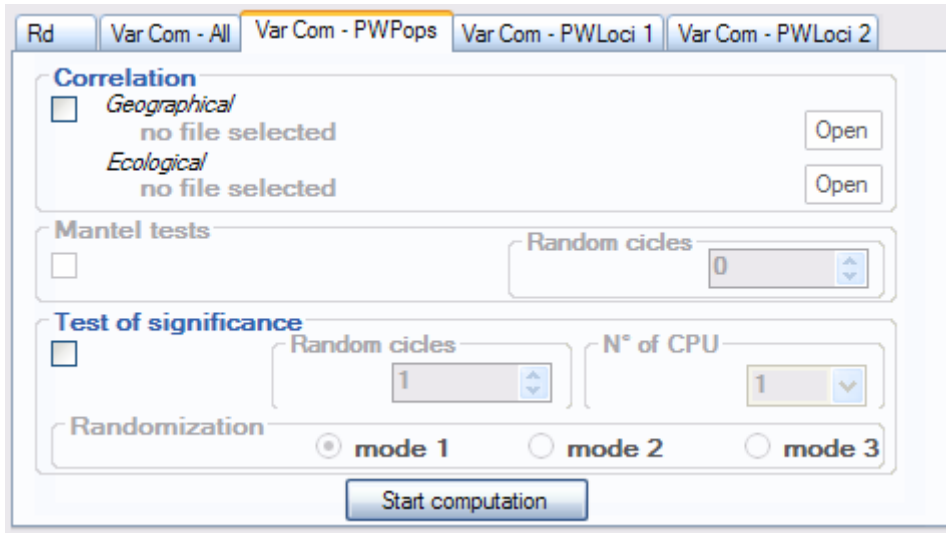
**Figures** - An example of dataset of 169 markers analyzed with the sliding window option (window size=13 markers, step=13, not overlapping windows).



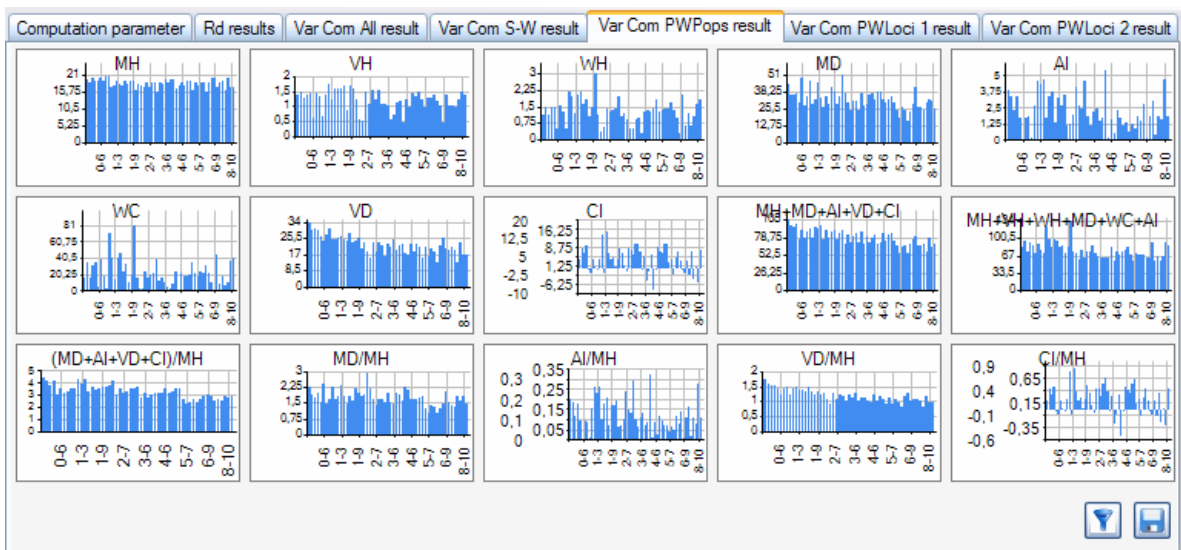
**Note.** The number of blue vertical bars is equal to the number of sliding windows.

### 3.3.7 Calculate the variance components overall loci and for pair of populations (VarCom-PWPops).

1. Select AT LEAST two populations;
2. Selects the “Var Com - PWPops” panel;
3. click the “Start computation” button.



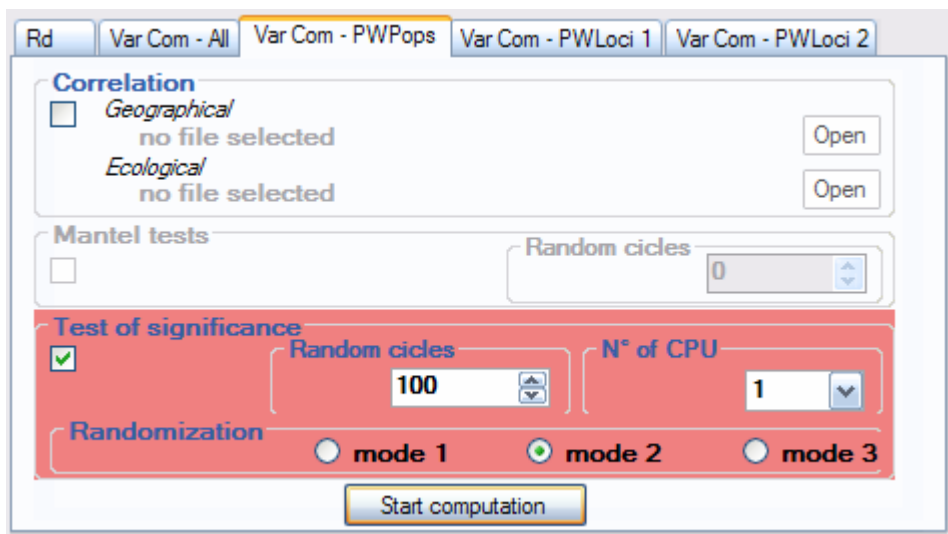
At the end of the calculations, the software will automatically shows the graphs of the variance components for all possible  $np = n(n-1)/2$  pair of population.



**Note.** Each vertical blue bar represents the value of the component for a pair of population.

### 3.3.8 Significance of the variance components for pair of populations (VarCom – PWPops).

1. Selects AT LEAST two populations;
2. Selects the “Var Com - PWPops” panel;
3. Activate the “Test of significance” option (the area will become red coloured);
4. Specify the numbers of randomizations;
5. Specify the number of CPU involved in calculation (the software automatically detects the number of available CPU in your computer);
6. choose and select the type of randomization (mode1, mode 2 or mode 3) desired to build the null-distribution of the variance components;
7. Click the “Start computation” button.

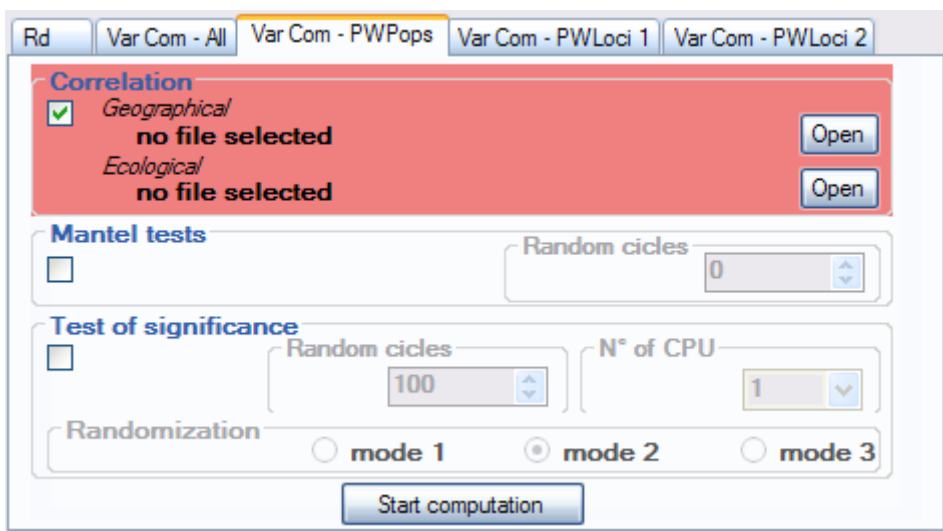


At the end of the calculations, the software automatically displays the graphs of the variance components for all possible  $np = p(p-1)/2$  population pairs (where  $p$  is the number of populations in your dataset) (as in 3.3.10). P values are reported in the .csv. output file.

### 3.3.9 Correlation between variance components and geographical and ecological variables (VarCom – PWPops)

For this analysis, the user must prepare three input files (a file with the genetic information, a file with the geographical coordinates for each sampling sites, a file with the ecological data for each sampling sites, but see (the section for preparation of the input files). The software will compute three kinds of matrices (genetic, geographic, and ecologic) between all possible  $np = p(p-1)/2$  pair of populations, where  $p$  is the number of populations in your dataset. The correlations among matrices can be then computed and tested for significance.

1. selects AT LEAST two populations;
2. Selects the “Var Com - PWPops” panel;
3. Activate the “Correlation” option (the area will became red coloured);
4. load the input file containing the geographical coordinates;
5. load the input file containing the ecological data;
6. If you want test the correlations for their significance, activate the “test of significance” option and specify the number randomizations.
7. Click “Start computation” button.

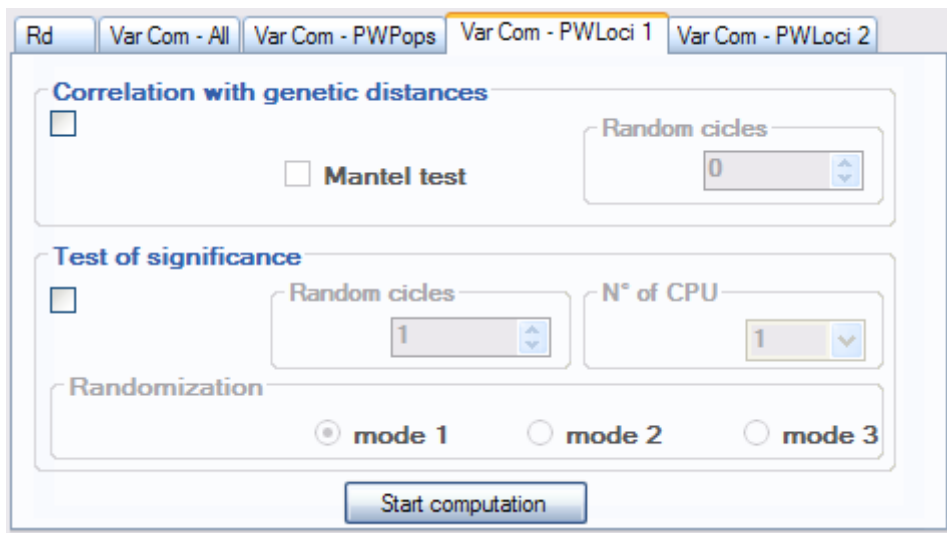


At the end of the calculation the software automatically displays the graphs of the variance components for all possible  $np = p(p-1)/2$  population pairs (where  $p$  is the number of

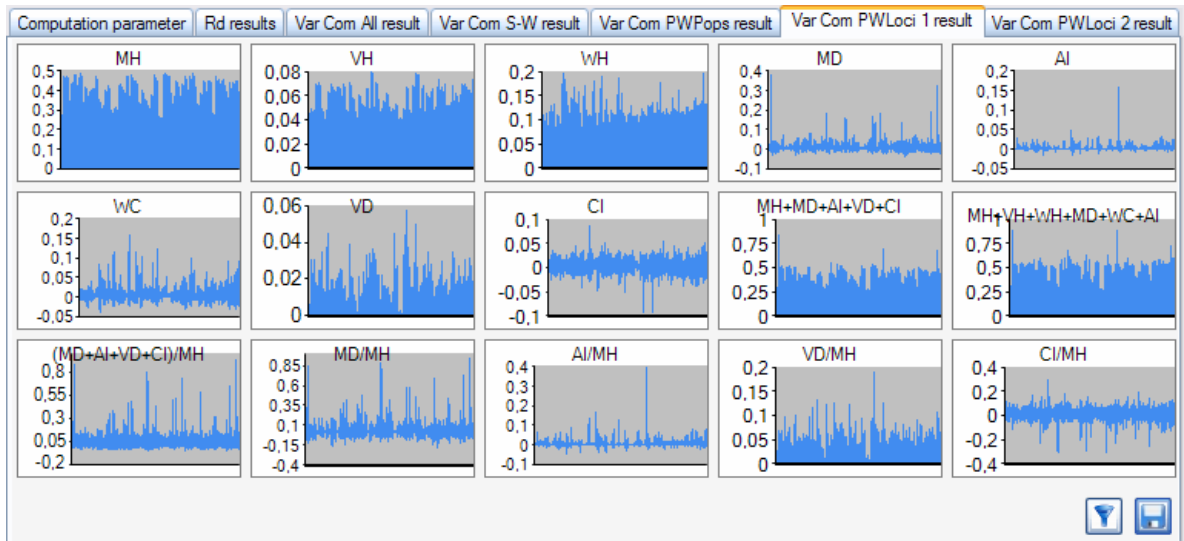
populations in your dataset) (as in 3.3.10). The data with correlation coefficients and their significance can be saved as a .csv file.

### 3.3.10 Calculate the variance components for pair of loci (VarCom –PWLoc1)

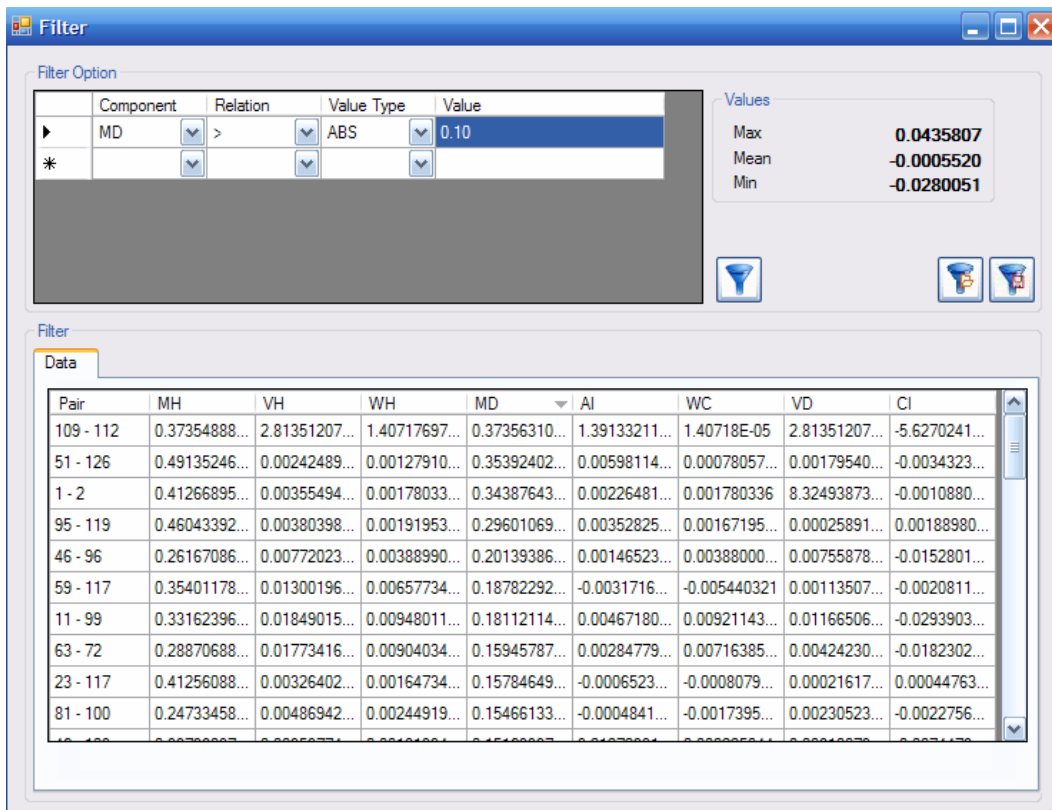
1. Select AT LEAST two populations;
2. Selects the “Var Com – PWLoc1 1” panel;
3. click the "Start computation" button.



At the end of the calculation, the software automatically displays the graphs of the variance components for all of the possible  $nm = n(n-1)/2$  pair of loci (where n is the number of loci in your dataset).

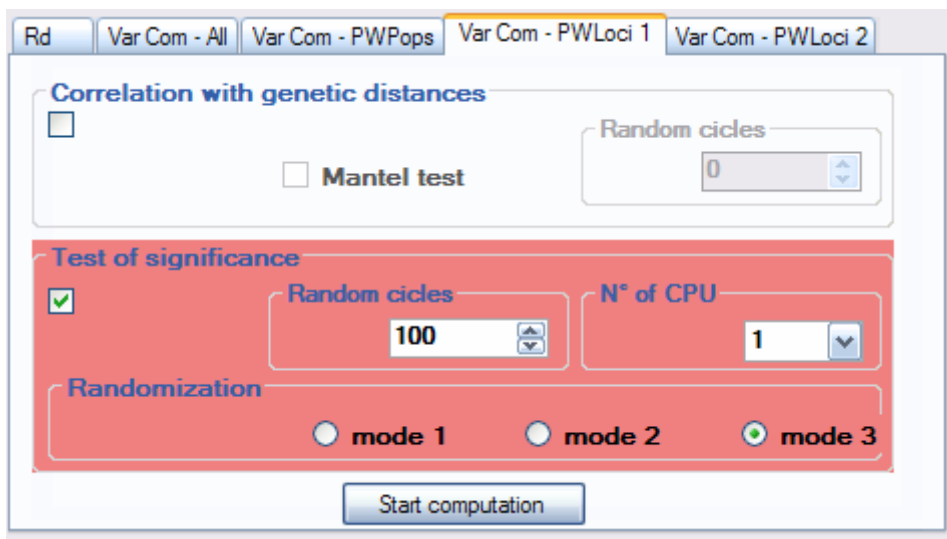


Note. Each vertical blue bar represents the value of the component for a pair of loci. On the right-lower corner, there are two buttons. The first allows filtering data based on the values of the components (see screenshot here below); the second allows saving raw data as a .csv file. This filter is useful to look for pairs that have some specified variance component values. Filter options can be saved.



### 3.3.11 Significance of the variance components for pair of loci (VarCom-PWLoc1).

1. Select AT LEAST two populations;
2. Selects the “Var Com – PWLoc1 1” panel;
3. activate the “Test of significance” option (the area will become red coloured);
4. Specify the number of randomizations;
5. Specify the number of CPU involved in calculation (the software automatically detects the number of available CPU in your computer);
6. Choose and select the type of randomization (mode1, mode 2 or mode 3) desired to build the null distribution of the variance components;
7. Click “Start computation” button.



At the end of the calculation, the software automatically displays the graphs of the variance components for all possible  $n(n-1)/2$  pair of loci (where  $n$  is the number of loci in your dataset) (as for 3.3.7). The results of the randomization tests can be saved as a .csv file.

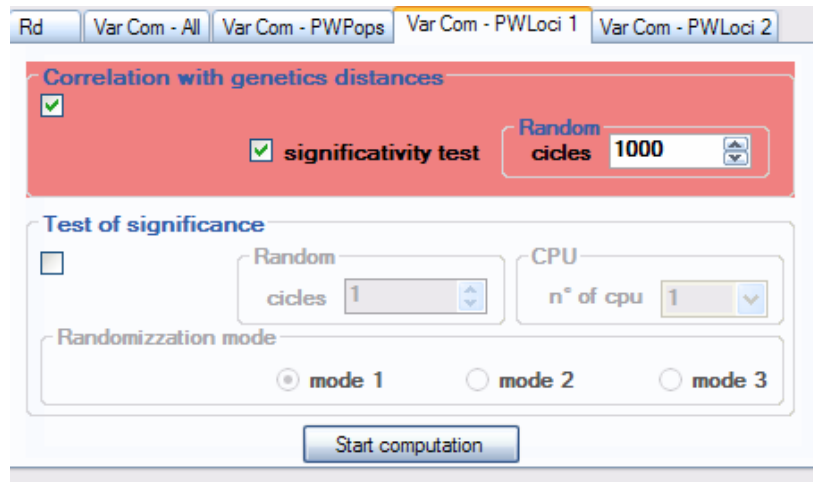
### 3.3.12 Correlation between variance components and genetic distance (VarCom-PWLoc1).

This calculation is possible only with “flat style + position loci” input file. In this case you



must:

1. Select at least two populations;
2. Select the “Var Com – PWLoc1 1” panel;
3. Activate the “Correlation with genetic distance” option;
4. Click “Start computation”.



The results can be saved as a.csv file.

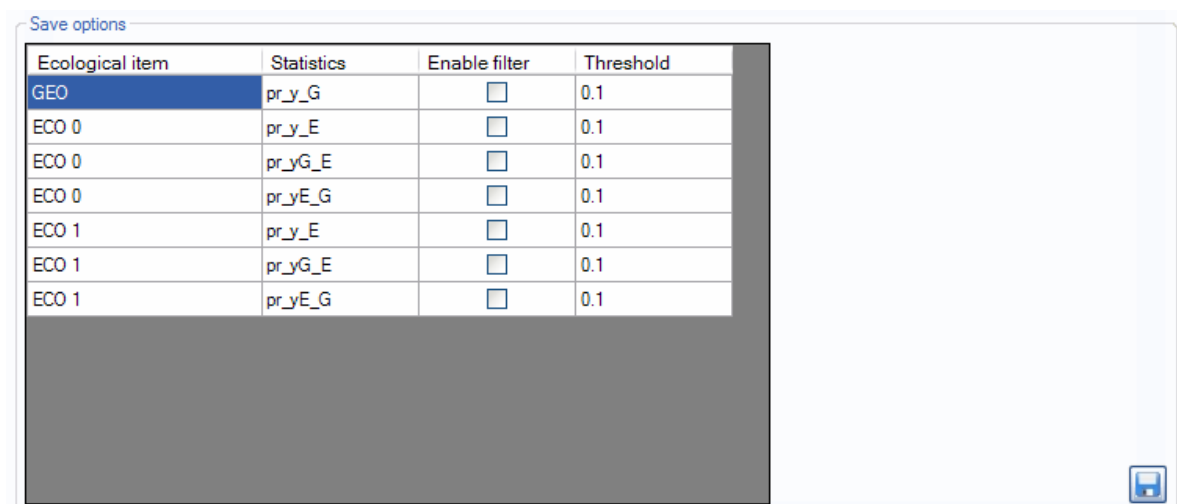
### 3.3.13 Significance of the correlations between variance components and geographical and ecological variables (VarCom-PWLoc12)

For this analysis user must prepare three input files (a file with the genetic information, a file with the geographical coordinates for each sampling sites, a file with the ecological data for each sampling sites (but see the input/out file section). The software will compute three kinds of triangular matrices (genetic, geographic, and ecologic). Each triangular matrix has  $n(n-1)/2 \times p(p-1)/2$  elements, where  $n$  is the number of loci and  $p$  is the number of populations. Specifically, for a given pair, the variance components are calculated for all possible pairs of populations. Thus, the same pair could be associated with different geographical and ecological distance. The correlations among matrices can be then computed and tested for their significance.

1. Select at least two populations;
2. Selects the “Var Com – PWLoc1 2” panel;
3. Load the input file containing the geographical coordinates
4. Load the input file containing the ecological variables.
5. If you want to test the correlations for their significance, activate the “Mantel tests” option and specify the number of randomizations.
6. Click the "Start computation" button.



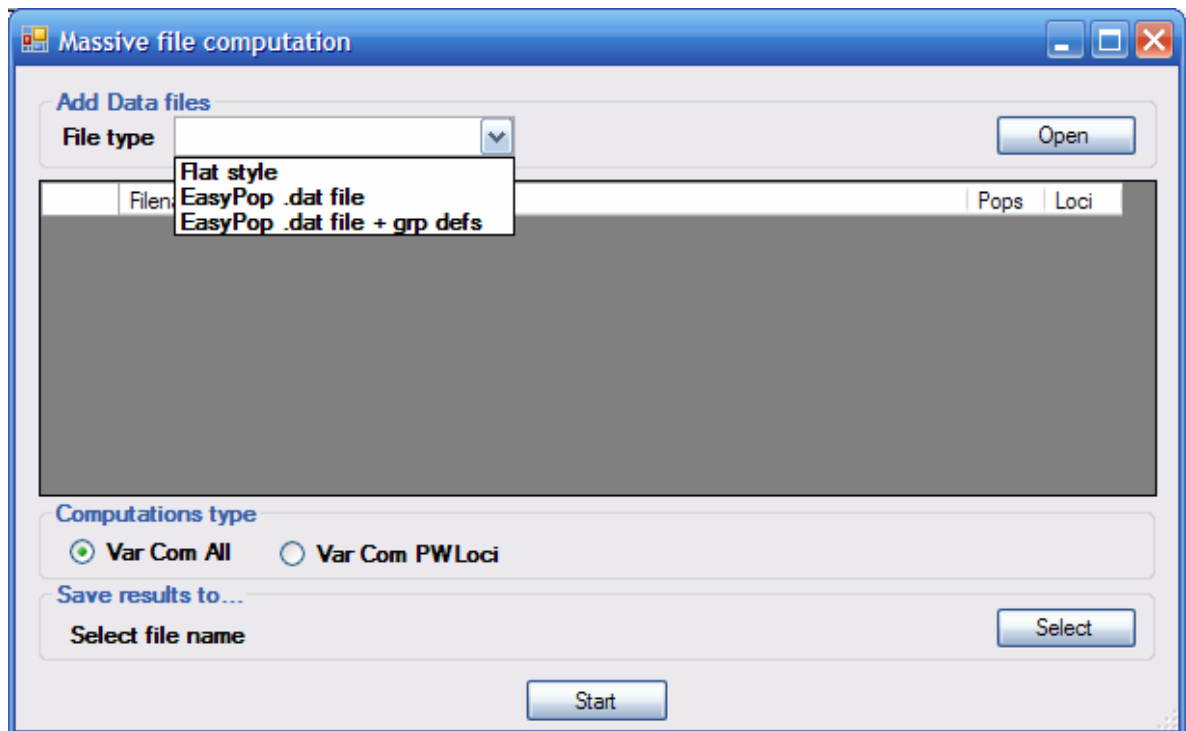
At the end of the calculations, a window will appear that allow filtering your correlation data specifying a desired P value.



## 3.4 Batch

### 3.4.1 Massive input file

Selecting Batch → Massive input file it is possible to automatically process multiple “flat style” or “EasyPop .dat files and calculate the variance components (overall or for loci pairs). Therefore, this function is particularly useful to process datafiles obtained by simulation with EasyPop software. Indeed, the user can simulate several replicates of a given neutral scenarios with EasyPop and process outputs with Nuragene: this will allow building null-hypotheses (neutral expectations) for the variance components under a variety of population structure models, mating systems, mutational model and migration regimes.



If one select the “EasyPop + grp defs” input file type, the user can sample groups of populations, individuals within populations and loci from a (large) dataset obtained with EasyPop. This function is useful to study sampling effects on the variance components. The instruction for sampling must be specified in a separated file named “grp defs” (see 4.1.6).

### 3.4.2 Post processing search

This option is useful to look for pairs of loci that satisfy certain conditions in term of significance of the variance components.

First, the user must perform tests of significance for the variance components under the modality VarCom-PWloci1. Thus, it will be possible:

1) chose the file(s) to process;

2) specify the conditions for filtering. These must include:

2.1) the sign of the component that can be a) positive, b) negative, c) either positive or negative.

2.2) the significance of the component that can be a) not significant, b) not significant >, c) not significant <, d) significant, e) significant >, f) significant <

2.3) Define a significance threshold (P values).

File	Value	Test results	p value
Shuffling_MH+VH+WH+MD+WC+AI.csv	Positive	Significant >	0.001
Shuffling_MH+MD+AI+VD+CI.csv	Positive	Significant >	0.001
Shuffling_MD.csv	Positive	Significant >	0.001

Suspected loci: single loci candidates balancing selection.txt

Check's results

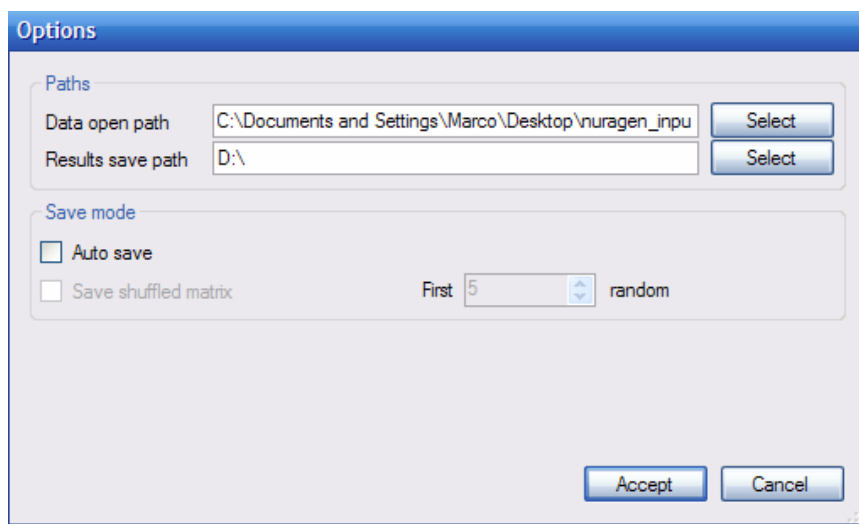
Locus	Locus
1	2
1	81
2	81
5	16
5	54
5	63

At the end of the search, the software displays a list of pair of loci. If the user has specified

a list of “suspected” loci (based on other evidence, for example based on the results of neutrality tests), the suspected loci will appear in red. To specify the “suspected” loci you must create and load a .txt file with one row that contains locus names separated by comma.

### 3.5 Options

The user can select the default data open path and the default Results save path together with the preferred saving mode.



## ➤ 4 Input/output file

### 4.1 Input file

#### 4.1.1 Flat Style

This input file must be a .txt file with the following characteristics:

- There are n rows and m columns;
- each of the n rows specify a haplotype;
- each of the m columns represents a marker.

- each row must have the same numbers of elements, 0 or 1, <TAB> delimited;
- a blank line separates different populations.

Note: there is no limit to the number of row and columns. The only restriction is given by the available memory in your computer.

*Example: input file “flat style” Two populations of three individuals each and characterized by three markers.*

1	0	1	} Population 1
1	1	0	
0	1	1	
1	0	1	} Population 2
0	0	1	
0	0	0	
1	1	0	

#### 4.1.2 Flat Style with loci position

Essentially this input file has the same format than the "Flat style" but the first row specify the position <TAB> delimited (in bp or in cM) of the markers.

*Es. input file “flat style with loci position”*

125	589	795	} Loci position
1	0	1	
1	1	0	} Population 1
0	1	1	
1	0	1	
0	0	1	} Population 2
0	0	0	
1	1	0	
1	1	0	

### 4.1.3 EasyPop output files

At the moment Nuragene can input the output .dat file generated by EasyPop when this is used to simulate populations of diploid individuals with loci with a maximum of two alleles. Specifically, when used with diploid individuals and with loci with two alleles EasyPop generates a data file with individuals in row and loci in columns. At each locus, the diploid genotype is specified by a four digit code. The first two digits specify the first allele and the third and the fourth digit the second allele. Thus, in general, for a locus with two alleles four genotypes are possible: 0101, 0102, 0201 and 0202

Nuragen do not work with heterozygotes and automatically recodes genotypes at a locus into 0 and 1: 0101 → 1, 0102 → 1, 0201 → 1 0202 → 0

Please note that Nuragen has been designed to work with haploid individuals, with diploid organism characterized by very high degree of selfing (such as *Hordeum*, *Phaseolus*, etc..) or in the case where haplotype phase is known or estimated.

### 4.1.4 Input file for geographical coordinates

This input file must be a .txt file with the following characteristics.

- $n$  rows and two columns
- the  $i^{\text{th}}$  row specify the x and y coordinates, respectively of the  $i^{\text{th}}$  population;
- x and y coordinates must <TAB> delimited;
- decimals must be comma separated “,”.
- the number of rows must be equal to the number populations in your genetic input data file;

*Example - input file “geographical coordinates”*

10,25	20,32
84,38	12,25

Diagram illustrating the input file structure for geographical coordinates. The table shows two rows and two columns of data. Arrows point to the columns:  $x_1$  and  $x_2$  for the first column, and  $y_1$  and  $y_2$  for the second column.

### 4.1.5 Input file for ecological variables

This input file must be a .txt file with the following characteristics.

- the file contains n rows and m columns
- the number of columns (m) is equals to the number of ecological variables considered in the analysis.
- Columns must be <tab> delimited.
- the first row contains variables names.
- the remaining n-1 rows must be equal to the number populations in your genetic input data file;
- decimals are comma “,” separated.
- 

**Example** - Input file for “ecological variables.” Two sites, characterized by three ecological variables.

---

Temperature	Humidity	Altitude
10.3	40	1200
20.8	70	200

### 4.1.6 Input file “Easypop + grp defs”

Choosing this option you must prepare two files. The first file is an EasyPop .dat file (see 4.1.3).

The second file is a .txt file with the same name of the EasyPop.dat file but with the \_ suffix (EasyPop.dat\_).

Example

NAME.dat	(file Easypop format)
NAME.dat_	(file grp defs)

This below is an example of “name.dat\_” file. This file contains all the information for sampling arranged in a single column.



5	← number of (simulated) groups with EasyPop
100	← number of simulated populations with EasyPop
4	← number of groups to extract
11	← number of populations to extract
3	← number of populations first group.
3	← number of populations second groups
3	← number of populations third groups
2	← number of populations fourth groups
20	← number of individuals population n.1
21	← number of individuals population n.2
25	← number of individuals first population n.3
24	← number of individuals first population n.4
26	← number of individuals first population n.5
25	← number of individuals first population n.6
21	← number of individuals first population n.7
20	← number of individuals first population n.8
25	← number of individuals first population n.9
26	← number of individuals first population n.10
25	← number of individuals first population n.11
127	← number of loci

In the example above reported, it is specified that the simulation with EasyPop has been conducted with 5 groups of populations of 100 populations each. However,, from this hierarchical island structure one wants (randomly) extract 4 groups and 11 populations with 3, 3, 3, and 2 populations per each of the four groups, respectively. User must specify the sample size of each population. The last line indicates the number of loci to extract.

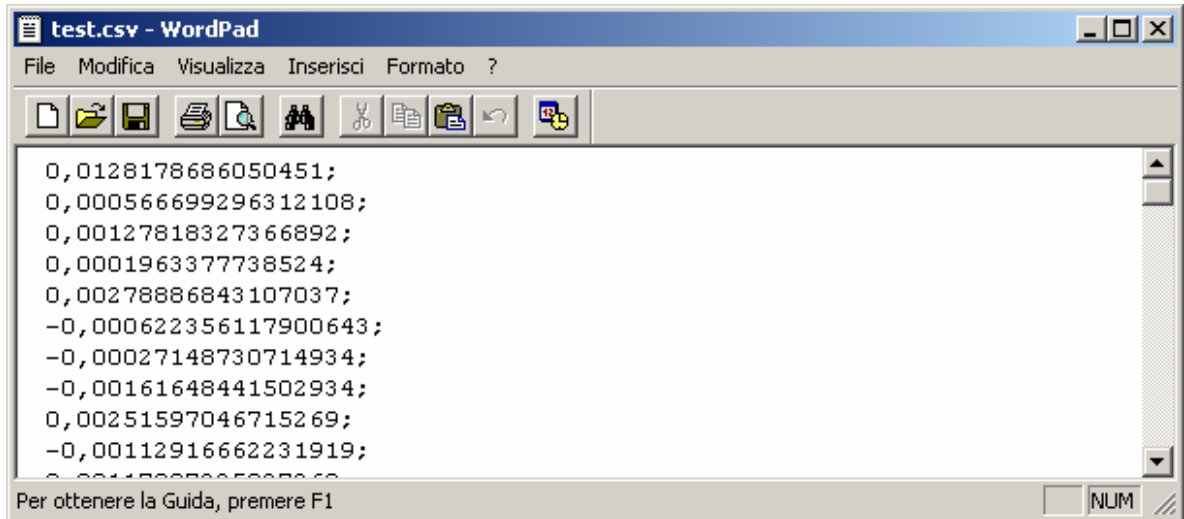
## 4.2 Output file

Nuragen allows saving the results (with or without filtering) of calculations as .CSV file. Here below the different output types are illustrated.

### 4.2.1 $r_d$ – Test of significance

The files has only one column of  $1 + n$  values, where  $n$  is the number of randomizations.

The first value is the observed  $r_d$ , the following  $n$  values are the simulated ones.

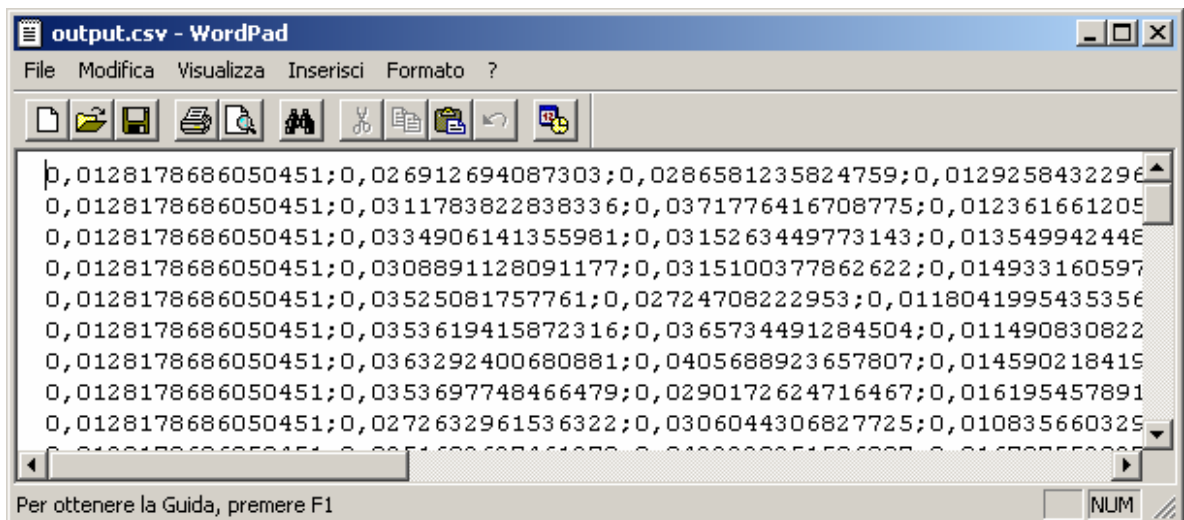


### 4.2.2 $r_d$ – Standardize for sample size

This file is of  $m$  columns and  $1 + n$  rows, where  $m$  is the number of populations and  $n$  is the number of randomizations.

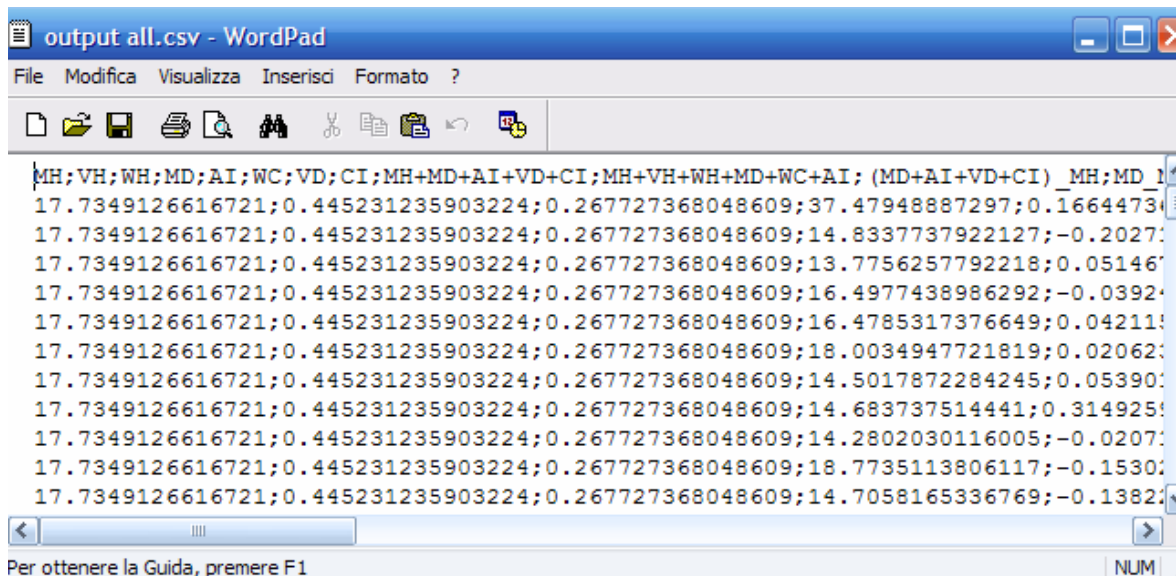
The first number of each column is the observed  $r_d$  values of a population.

For each column separately and below each of the observed  $r_d$  values, there are the  $n$  randomized  $r_d$  values.



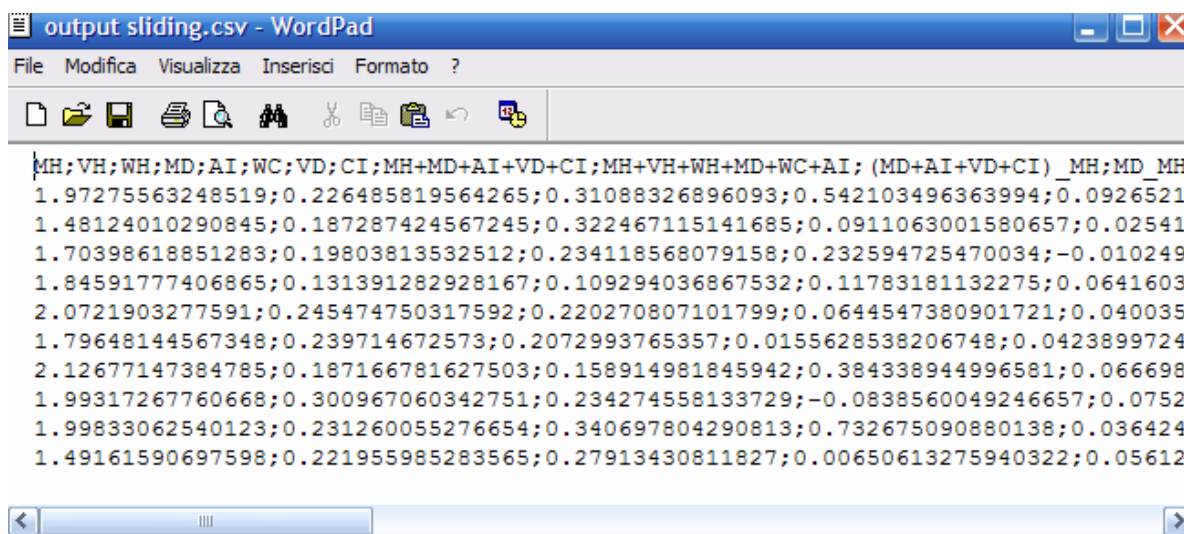
### 4.2.3 VarCom – All. Tests of significance

The file has 15 columns (as many as the number of the variance components and for their derivatives). Each column contains the name of the component, the observed value of the component, and the n randomized values of the component.



### 4.2.4 VarCom – All. Sliding window analysis

The file consists of 15 columns (as many as the number of variance components and for their derivatives) and of 1+ n rows where n is the number of the analysed windows. The first row contains the component's name. Each row contains the 15 values calculable for a window.

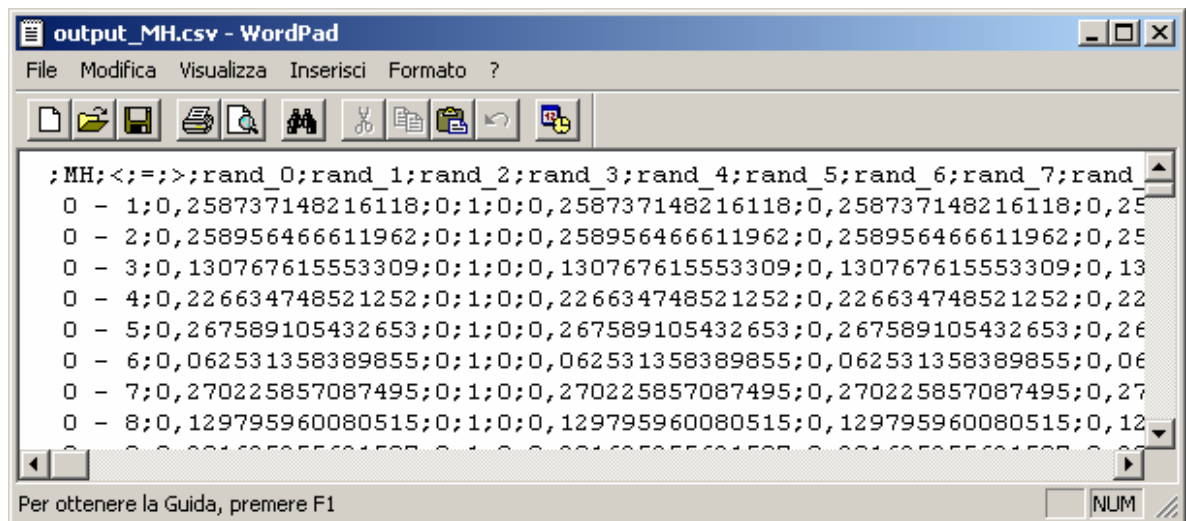


#### 4.2.5 VarCom - PWLoc1. Tests of significance.

The output consists of 15 files, one for each of the variance components and for their derivatives. Each file contains 5 + m columns, where m is the number of randomizations. The first five columns are:

- 1) The list of the  $nL = L(L-1)/2$  loci pairs.
- 2) The observed value of the component for all loci pairs.
- 3) The proportion of replicates in which the RANDOMIZED values are HIGHER THAN the OBSERVED values;
- 4) The proportion of replicates in which the RANDOMIZED values are EQUAL TO the OBSERVED values.
- 5) The proportion of replicates in which the RANDOMIZED values are LOWER THAN the OBSERVED values.

Here below, as an example, it is reported the output for the MD component.



```
;MH;<;=;>;rand_0;rand_1;rand_2;rand_3;rand_4;rand_5;rand_6;rand_7;rand_8
0 - 1;0,258737148216118;0;1;0;0,258737148216118;0,258737148216118;0,25
0 - 2;0,258956466611962;0;1;0;0,258956466611962;0,258956466611962;0,25
0 - 3;0,130767615553309;0;1;0;0,130767615553309;0,130767615553309;0,13
0 - 4;0,226634748521252;0;1;0;0,226634748521252;0,226634748521252;0,22
0 - 5;0,267589105432653;0;1;0;0,267589105432653;0,267589105432653;0,26
0 - 6;0,062531358389855;0;1;0;0,062531358389855;0,062531358389855;0,06
0 - 7;0,270225857087495;0;1;0;0,270225857087495;0,270225857087495;0,27
0 - 8;0,129795960080515;0;1;0;0,129795960080515;0,129795960080515;0,12
0 - 9;0,001605055601507;0;1;0;0,001605055601507;0,001605055601507;0,00
```

#### 4.2.6 VarCom – PWloc1. Correlation with genetic distances.

This file contains 16 columns: the first is the list of all of the possible  $nL = L(L-1)/2$  pair of loci. The remaining 15 are for components of variance and for their derivatives.

Overall there are 1+ n+2 rows:

- column labels;
- n rows that contain all randomized values for all of the variance components;
- correlation values (Pearson's r);
- Significance values (P) of the correlations (tested by non-parametric Mantel test).

#### **4.2.7 VarCom – PWPops. Test of significance**

The output consists of 15 files, one for each of the variance component and for their derivatives.

Each file consists of c columns, where  $c = p(p-1)/2$  is the number of possible pairs of populations and of  $1 + r$  rows where r is the number of randomizations. The first row specifies labels and the remaining rows are the randomized Pearson's r values.

#### **4.2.8 VarCom – PWPops. Correlation with geographical and ecological variables.**

The output consists of 16 files. The first 15 are the results for each of the variance component and for their derivatives. The last file, denoted with the suffix “\_collect”, is a short summary of the results of the correlation analysis.

Each of the 15 files has  $3 + c$  columns where c is the number of ecological variables investigated.

- The first column is the list of all possible  $nP = (P-1)P/2$  populations;
- The second columns is the observed value of the component;
- The third column is the geographical distance between pairs of populations;
- The remaining c columns are the distances between pair of populations for the c ecological variables considered

In this file the following statistics are reported:

- The Person's r simple correlation coefficient between the variance component and

the geographical distance and its significance level

- The Person's r simple correlation coefficient between the variance component and the ecological distance and its significance level.

- The correlation between the variance component and the geographical distance partialled respect to the ecological distance and its significance level.

- The correlation between the variance component and the ecological distance partialled respect to the geographical distance and its significance level.

#### **4.2.9 VarCom – PWLoc2. Correlation between variance components and geographical and ecological variable**

The output consists of 16 files. The first 15 are the results for each of the variance component and their derivatives. The last file, denoted with the suffix “\_collect”, is a short summary of the results of the correlation analysis.

The first column list all possible  $nL = L(L-1)/2$  pair of loci. Then, for each ecological variable, eight values are calculated. These are:

- the Person's r simple correlation coefficient between the variance component and the geographical distance and its significance level.

- the Person's r simple correlation coefficient between the variance component and the ecological distance and its significance level.

- The correlation between the variance component and the geographical distance partialled respect to the ecological distance and its significance level.

- The correlation between the variance component and the ecological distance partialled respect to the geographical distance and its significance level.

## ➤ 5 Methods

### 5.1 Randomizations

#### 5.1.1 Mode 1- shuffling of alleles (among individuals, within populations).

A null hypothesis, which is interesting and relatively easy to test, is that of complete panmixia. To do this, one compares the observed dataset to datasets in which an infinite amount of sex and recombination has been imposed on the data by randomly shuffling the alleles amongst individuals, independently for each locus (*e.g.*, Burt et al. 1996). Second, if populations have been defined, then alleles are only shuffled amongst individuals of the same population. This option can be useful because linkage disequilibrium can arise from combining samples from genetically distinct populations. With this option the linkage disequilibrium due to population differentiation is maintained in all the randomized datasets, so if the observed dataset still shows significant associations, they cannot be due solely to population differentiation (at least at the scale defined).

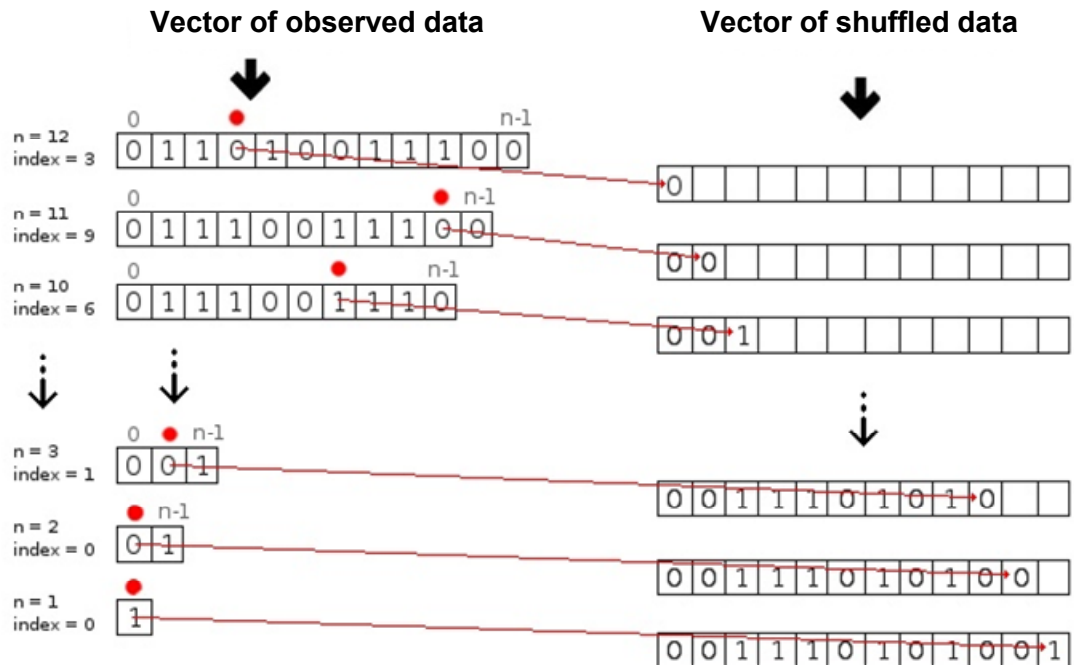
**Figure** – Conceptual representation of shuffling of alleles.



For each locus separately, Nuragen uses the procedure as follow:

- 1) The scoring at a locus is represented by a vector of dimension  $n$ .
- 2) A random integer number (index) is generated with a value between 0 and  $n-1$ ;
- 3) The allele that occupies the position “index” in the observed vector is moved in a new vector (the vector of shuffled data) in position 0 (zero).
- 4) The allele at position “index” in the vector of observed data is then eliminated.

- 5) Steps 2-4 are repeated until the size of the vector of the observed data has zero (0) elements.

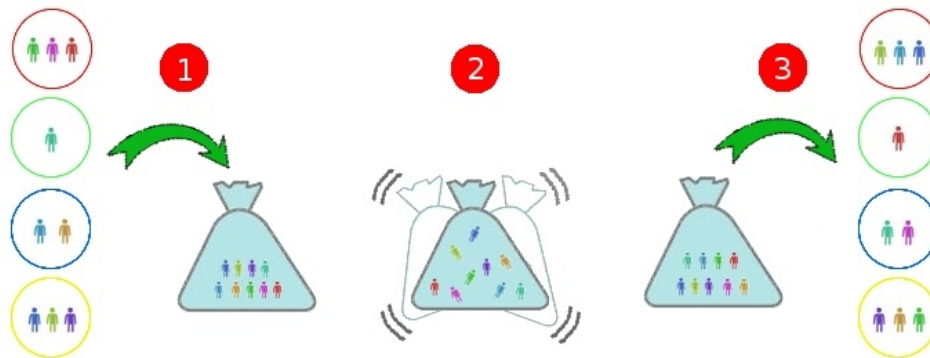


### 5.1.2 Mode 2- permutation of individuals among populations.

The null hypothesis of interest here is no population differentiation, and this is tested by comparing statistics calculated for the observed populations to that for datasets in which individuals have been randomly permuted across populations. This randomization procedure maintains any linkage disequilibrium and (for diploids) deviations from Hardy-Weinberg proportions that may exist in the observed dataset. Independent randomization of individual loci or alleles across populations would not do this. If such disequilibria do exist, tests which did not take them into account could give spuriously low p-values.

**Figure** – Conceptual representation of shuffling of alleles.

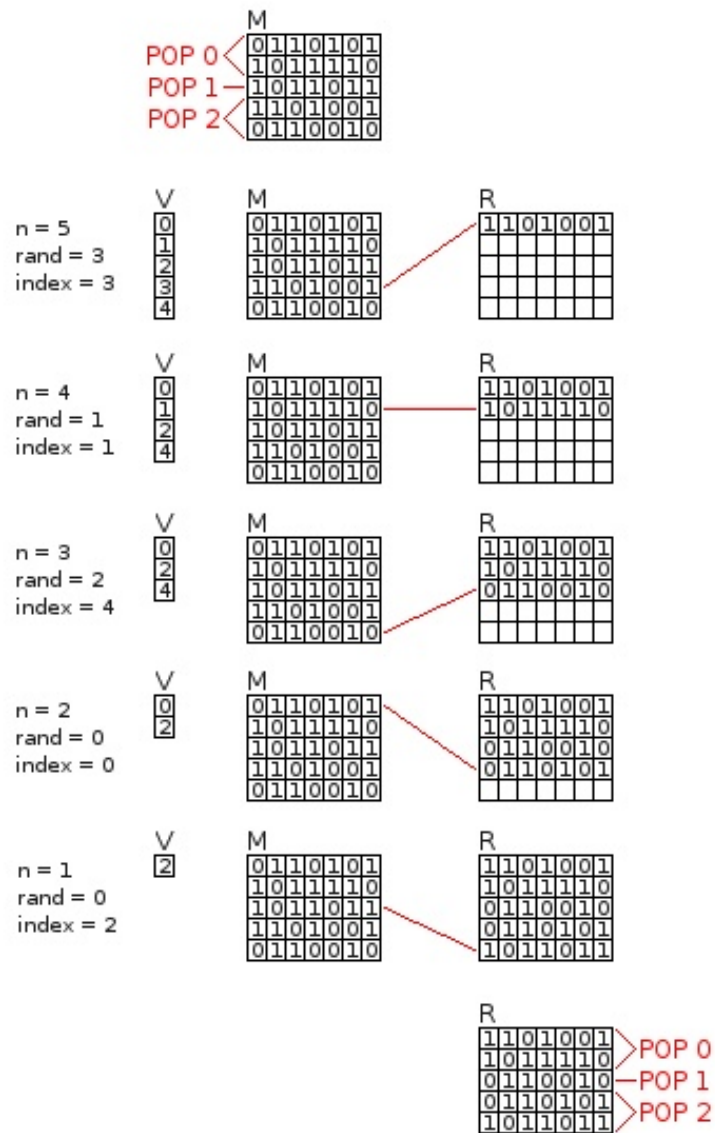




The procedure applied by Nuragen is articulated in the following steps:

- 1) All data were used to build a matrix of  $m$  rows (where  $m$  is the number of individuals) and  $n$  columns (where  $n$  is the number of loci);
- 2) a new vector ( $V$ ) of  $m$  elements (numbered from 0 to  $m-1$ ) is created;
- 3) A random integer number is generated with value between 0 and  $n-1$ ;
- 4) In the vector  $V$ , the value (index) that occupies the position specified by the random number ( $rand$ ) is taken;
- 5) The corresponding row of the matrix is copied in the first row of a new matrix (the matrix of permuted individuals among populations);
- 6) The element of  $V$  that occupies the position “ $rand$ ” is eliminated;
- 7) Steps 3-6 are repeated until the dimension of vector  $V$  reach zero elements;
- 8) The new matrix is partitioned in accord with sample sizes of the original populatio

**Figure –** Permutating individuals across populations



### 5.1.3 Mode 3 - shuffling alleles and permutating individual simultaneously

In this case, each randomization consists of two events: first, full migration among population is simulated as in Mode 2; second, panmixia is simulated as in Mode 1.

## ➤ 6 Statistics

### 6.1 $r_d$

The index of association (IA) is the traditional measure of multilocus linkage disequilibrium (Brown et al. 1980). In brief, the "distance" (number of loci at which they differ) between all pairs of individuals is calculated, and the variance of these distances compared to that expected if there is no linkage disequilibrium. In essence, one is asking whether two individuals being the same at one locus makes them more likely to be the same at another. However, IA increases with the number of loci. That is, if there is linkage disequilibrium, then the value will usually depend upon the number of loci included in the analysis, which makes comparisons among studies difficult (Brown et al. 1980; Maynard Smith et al. 1993).

To avoid this problem, a slightly modified statistic that should largely remove this dependency on number of loci is (Burt et al., 1996):

$$\bar{r}_d = \frac{\sum \sum cov_{j,k}}{\sum \sum \sqrt{var_j \cdot var_k}}$$
$$= \frac{V_D - \sum var_j}{2 \sum \sum \sqrt{var_j \cdot var_k}}$$

The first equation has a form similar to a correlation coefficient (hence the symbol  $r$ , with the subscript  $d$  referring to distances, and will have a maximum value of 1.

## 6.2 Components of the variance for the number of heterozygous loci in several populations.

Consider the hypothetical population consisting of an equiproportional mixture (**bulk**) of the  $J$  populations. The variance in number of heterozygous loci is (Brown and Feldman 1991):

$$\begin{aligned} {}^T\sigma_K^2 &= {}^T h_1 + {}^T h_2 - {}^T h_1^2 - {}^T h_2^2 \\ &+ \sum_i \sum_k 2^T D_{ik} [2^T p_i^T q_k + {}^T D_{ik}]. \end{aligned}$$

Where  $h_1$  and  $h_2$  indicate the genetic diversity (Nei, 1978) at locus 1 and 2, respectively;  $p_i$  and  $q_k$  are the frequencies of allele  $i$  at locus 1 and of alleles  $k$  at locus 2;  $D_{ik}$  is a measure of linkage disequilibrium calculated as  $g_{ik} - p_i q_k$  (i.e. difference between observed and expected gametic frequencies).

This expression can be split into six components, of which three are single-locus contributions and three arise from the pairwise summation term.

The *single-locus* components are:

$${}^T h_1 + {}^T h_2 - {}^T h_1^2 - {}^T h_2^2 = MH + VH + WH.$$

$MH$  arises from the averages of the population gene diversity,

$$MH = \bar{h}_1 + \bar{h}_2 - \sum_i [{}^i h_1^2 + {}^i h_2^2]/J.$$

$VH$  arises from the variation among populations in gene diversity,

$$VH = \text{var}(h_1) + \text{var}(h_2).$$

$WH$  arises from the Wahlund effect or variance of allele frequency among populations,

$$WH = d_1(1 - 2\bar{h}_1 - d_1) + d_2(1 - 2\bar{h}_2 - d_2).$$

The *two-locus* components are:

$$\sum_i \sum_k 2^T D_{ik} [2^T p_i^T q_k + {}^T D_{ik}] = MD + WC + AI.$$

$MD$  arises from the average of the population disequilibria

$$MD = \sum_i \sum_k 2D_{ik} [2^T p_i^T q_k + D_{ik}].$$

*WC* arises from the Wahlund effect in two loci or covariance of allele frequencies;

$$WC = \sum_i \sum_k 2 \text{cov}(p_i, q_k) [2^T p_i^T q_k + \text{cov}(p_i, q_k)].$$

*AI* arises from the average interaction of these effects,

$$AI = \sum_i \sum_k 4D_{ik} \text{cov}(p_i, q_k).$$

Now consider the **average of the variances belonging to each of the J populations**. This can also be split into five components, one being a single locus effect and the remaining four being two locus effects.

$$\sigma_K^2 = \sum_i {}^t \sigma_K^2 / J = MH + MD + AI + VD + CI.$$

The single-locus component *MH* and the two-locus components *MD* and *AI* are as defined above. The remaining two-locus components are *VD*, which arises from variation among populations

in disequilibria:

$$VD = \sum_i \sum_k 2 \text{var}(D_{ik}) = \sum_i \sum_k 2 \left[ \sum_i {}^t D_{ik}^2 / J - D_{ik}^2 \right],$$

and *CI*, which arises from covariation in the interaction of disequilibria and Wahlund's covariance among populations:

$$CI = \sum_i \sum_k 4 \text{cov}[(p_i q_k), D_{ik}].$$

When *m* loci are scored, the single-locus components-namely, *MH*, *VH*, and *WH*-are evaluated for each locus and summed, whereas the two-locus components are evaluated over the  $m(m-1)/2$  possible pairs of loci and summed. The gametic phase must be known or all  $D_{ik}$  must be estimated.

### 6.3 Correlation and partial correlation analyses

The goal is to compute correlation between the *Ymatrix* and *X1* (or *X2*) or a partial correlations between the *Ymatrix*, *X1* and *X2* or *Ymatrix*, *X2* and *X1*. The *Ymatrix* can be either a pairwise population or a pairwise loci *matrix* for a variance component. *X1* and *X2* are the geographical and ecological distances matrices between pairs of populations.

At the moment, the matrix of the geographical distances is obtained simply by calculating, the Euclidean (linear) distance between pairs of populations directly using the UMT coordinates provided in the input file.

Finally, in the current version, the matrix of ecological distance is obtained by calculating, for each pair of populations, the absolute difference among the level of the ecological variable.

We tested correlations and partial correlations by the Mantel non-parametric test. This consists in testing the significance of the correlation between two or more matrices by a permutation procedure allowing getting the empirical null distribution of the correlation coefficient taking into account the auto-correlations of the elements of the matrix.

The correlation of the two matrices is classically defined as

$$r_{XY} = \frac{SP(\mathbf{X}, \mathbf{Y})}{\sqrt{SS(\mathbf{X}) \cdot SS(\mathbf{Y})}}$$

the ratio of the cross product of  $\mathbf{X}$  and  $\mathbf{Y}$  over the square root of the product of sums of squares. The denominator of the above equation is insensitive to permutation, such that only the numerator will change upon permutation of rows and columns. The only quantity that will actually change between permutations is the Hadamard product of the two matrices noted as

$$Z_{XY} = \mathbf{X} * \mathbf{Y} = \sum_{i=1}^N \sum_{j=1}^i x_{ij} y_{ij}$$

which is the only variable term involved in the computation of the cross-product.

The Mantel testing procedure applied to two matrices will then consist in computing the quantity  $Z_{XY}$  from the original matrices, permute the rows and column of one matrix while keeping the other constant, and each time recompute the quantity , and compare it to the original  $Z_{XY}$  value (Smouse et al. 1986).

In the case of three matrices, say **Y**, **X1** and **X2**, the procedure is very similar. The partial correlation coefficients are obtained from the pairwise correlations as,

$$r_{Y.X_1.X_2} = \frac{r_{YX_1} - r_{X_1X_2}r_{YX_2}}{\sqrt{(1 - r_{X_1X_2}^2)(1 - r_{YX_2}^2)}}$$

The other relevant partial correlations can be obtained similarly (see e.g. Sokal and Rohlf 1981). The significance of the partial correlations are tested by keeping one matrix constant and permuting the rows and columns of the other two matrices, recomputing each time the new partial correlations and comparing it to the observation (Smouse et al. 1986). Applications of the Mantel test in anthropology and genetics can be found in Smouse and Long (1992).

### ➤ **Handling missing data**

At the moment, Nuragen do not work with missing data.

## **A case of study: the Sardinian Barley landraces**

In this section, we show how the use of Nuragen could help the investigation of linkage disequilibrium (LD) in plant populations.

In particular, we were interested in populations of a Sardinian barley landraces to answer to two main questions:

- Do ecological factors have a role in determining the population structure of LD?.
- Is there evidence for pair of loci under epistatic divergent or balancing selection?

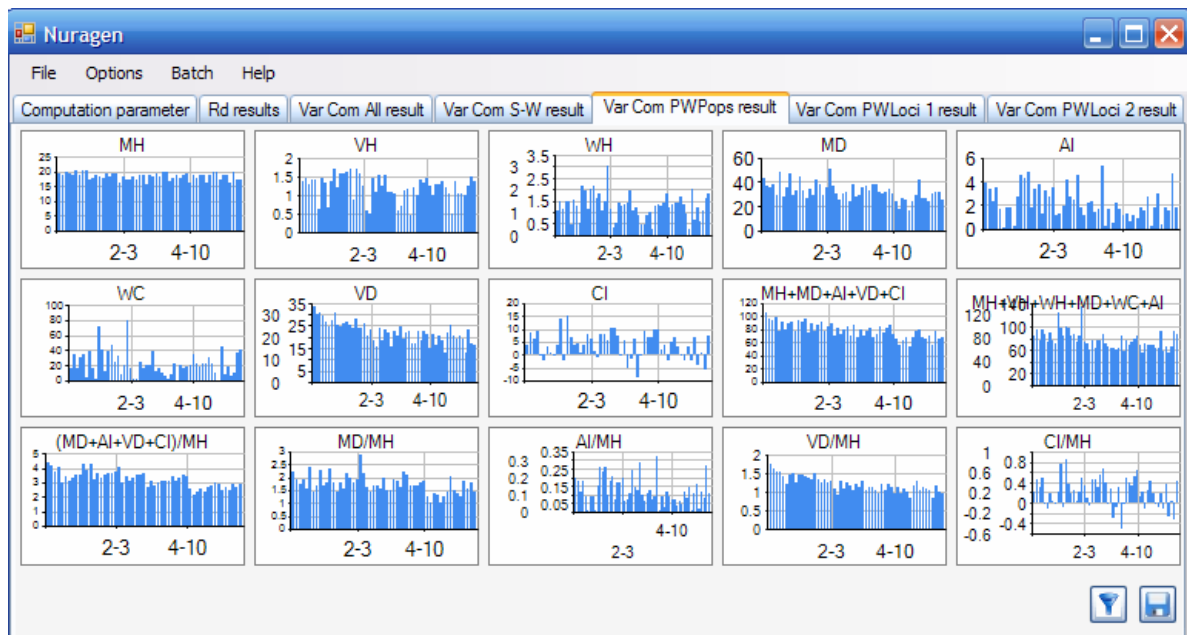


## Do ecological factors have a role in determining the population structure of Linkage disequilibrium?.

First, using the option VarCom–PWPOps, we calculated the eight variance components (and their standardizations) were calculated for all of the  $n_p = p(p-1)/2$  possible pair of populations, where  $p$  is the number of populations ( $p= 11$  in our case,  $n_p = 55$ ).

Results showed that each component takes very different values depending on the pair of populations compared (Figure 1).

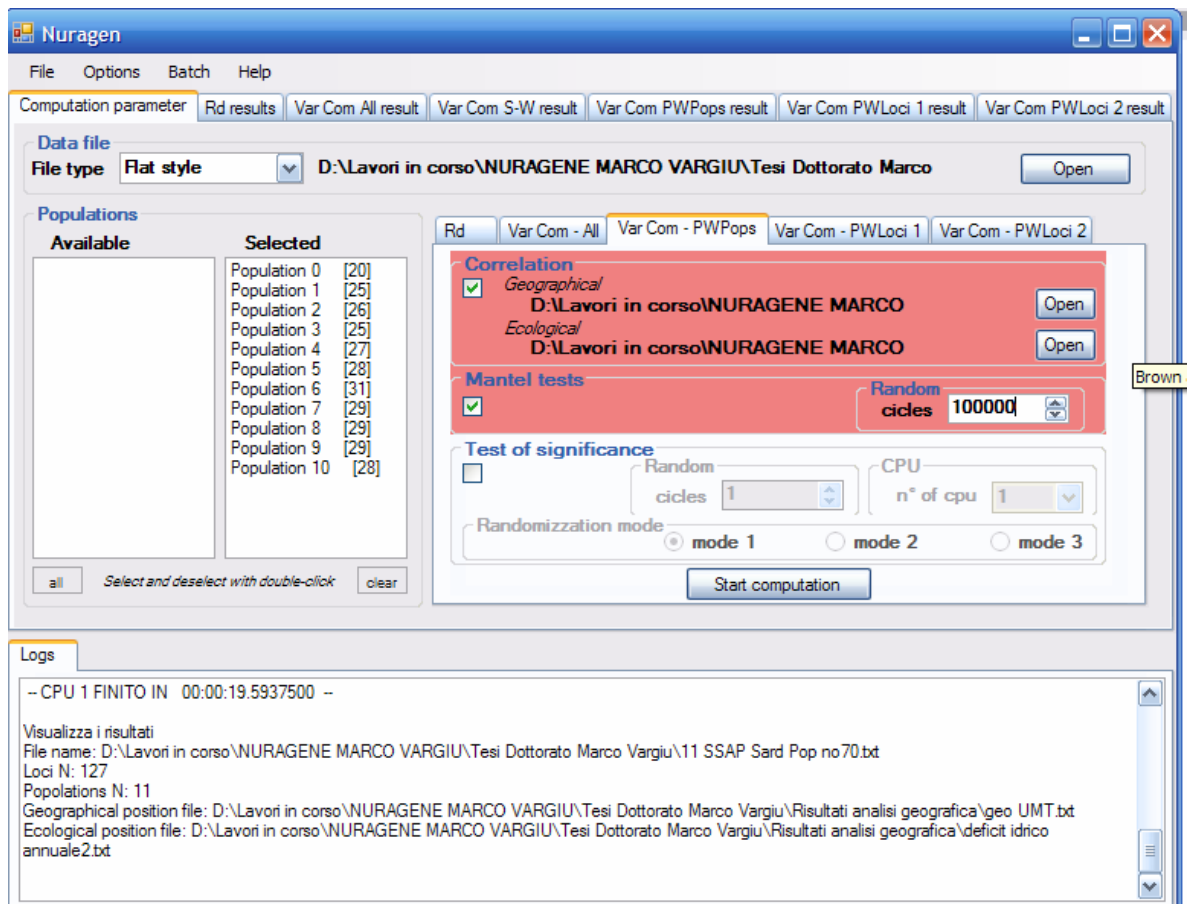
**Figure 1** – Values of each of the eight components (and their standardized values) for all of the 55 pairs of barley populations.



Now, the question is: do such differences between populations depend from their geographical distances ? Alternatively, do ecological differences among sampling sites are more important than geographical distances?. Are both variables relevant?.

With this aim, we then conduct a correlation analysis applying partial Mantel tests considering each component as dependent variable and geographical and ecological variable information as independent variables (Figure 2).

**Figure 2** – Performing a correlation analysis considering geographical and ecological variables.



Geographical coordinates were in UMT system, while as ecological variable we used the water deficit (in mm of water) calculated for each sites. Annual water deficit was calculated as difference between annual rainfall (mm) and the potential Evapo-transpiration (ET<sub>p</sub>, mm). We determine ET<sub>p</sub> using the formula of Thornthwaite-Mather (1955). Climatic data (rainfall, temperature and solar radiation were provided by ARPAS, Regione Autonoma della Sardegna. The period considered was 1960-1990.

We chose water deficit as a “synthetic” variable of the level and interplay between rainfall levels, temperature and solar radiation. The results of the analysis are reported in Table 1.

	<b>Annual water deficit (mm)</b>
<i>COR</i>	-784.5
<i>Vi</i>	-784.5
<i>ORO</i>	-678
<i>NXM</i>	-591.4
<i>SEN</i>	-591.4
<i>STU</i>	-591.4
<i>CUM</i>	-493.4
<i>N2</i>	-493.4
<i>PIR</i>	-307.5
<i>SIS3</i>	-307.5
<i>SOR</i>	-263.9

**Table 1** – Results of partial correlations analysis: *geo\_wd*, correlations between variance components and geographical distance partialled with respect to water deficit (*geo\_wd*); *wd\_geo*, correlations between variance components and annual water deficit partialled with respect to geographical distance. Significances of the partial correlations were obtained by a three way-mantel test with 100000 replicate. The analysis has been performed with Nuragen both with all loci (A) and only with shared polymorphisms among populations (B). In **red**: significant values,  $P < 0.05$ .

	A) All 127 polymorphic loci				B) 47 shared polymorphisms			
	<i>Geo</i>	<i>wd</i>	<i>wd</i>	<i>geo</i>	<i>wd</i>	<i>wd</i>	<i>geo</i>	<i>wd</i>
	<i>P<sub>geo wd</sub></i>	<i>P<sub>wd geo</sub></i>	<i>P<sub>geo wd</sub></i>	<i>P<sub>wd geo</sub></i>	<i>P<sub>geo wd</sub></i>	<i>P<sub>wd geo</sub></i>	<i>P<sub>geo wd</sub></i>	<i>P<sub>wd geo</sub></i>
<b>Single locus</b>								
<i>MH</i>	-0.135	0.417	0.138	<b>0.001</b>	-0.080	0.223	0.795	<b>0.028</b>
<i>VH</i>	0.243	0.449	0.038	<b>0.001</b>	0.126	0.365	0.165	<b>0.004</b>
<i>WH</i>	0.207	0.247	0.053	<b>0.049</b>	0.206	0.346	0.039	<b>0.008</b>
<b>Two – locus</b>								
<i>MD/MH</i>	-0.311	-0.204	<b>0.992</b>	0.924	-0.231	-0.263	<b>0.966</b>	<b>0.965</b>
<i>WC</i>	0.101	0.168	0.201	0.141	0.129	0.256	0.124	<b>0.046</b>
<i>AI/MH</i>	-0.055	0.066	0.674	0.312	0.185	0.166	0.073	0.132
<i>VD/MH</i>	0.001	0.481	0.550	$<10^{-5}$	0.094	0.389	0.295	<b>0.002</b>
<i>CI/MH</i>	0.476	0.215	<b>0.000</b>	0.073				

Overall, ecological differences among population's sampling sites have a stronger impact than geographical distances on the structure of genetic diversity of the landraces populations.

This is true for all of the three single locus components, *MH*, *VH* and *WH*.

However, the highest and strongest correlation is between the two-locus components *VD*, the variance of disequilibrium, and the difference in water deficit ( $r = 0.550; P < 10^{-5}$ ). This means that when difference among sites in water deficit increases, the variance of LD (i.e. the variation in frequency of the most frequent gametic type) also increases. Moreover, no correlation between *VD* and geographical distance has been detected. *VD* is expected to be high when drift and founder effect or epistatic divergent selection are operating. Thus, our results suggest that epistatic divergent selection for water deficit could have had a significant role in this system. Alternatively, it is conceivable that geographical distances are completely unrepresentative of the degree of the genetic isolation of the populations and the actual pattern of migration and drift exactly parallel that of ecological distance. However, this seems unlikely: indeed, it should be noted that *WH* and *CI* are correlated

with geographical distance indicating that, at least partially, migration rate is (inversely) related to geographic distance among sampling sites.

Under isolation by distance, the component MD is expected to be inversely correlated with geographic distance: the repetitive disequilibrium should be lower when drift or founder effect increase. Effectively, we observed a significant negative correlation between MD and geographical distance suggesting a more prominent role for drift and limited migration than for selection.

To estimate the probability that the observed pattern of correlations is due to neutral evolution:

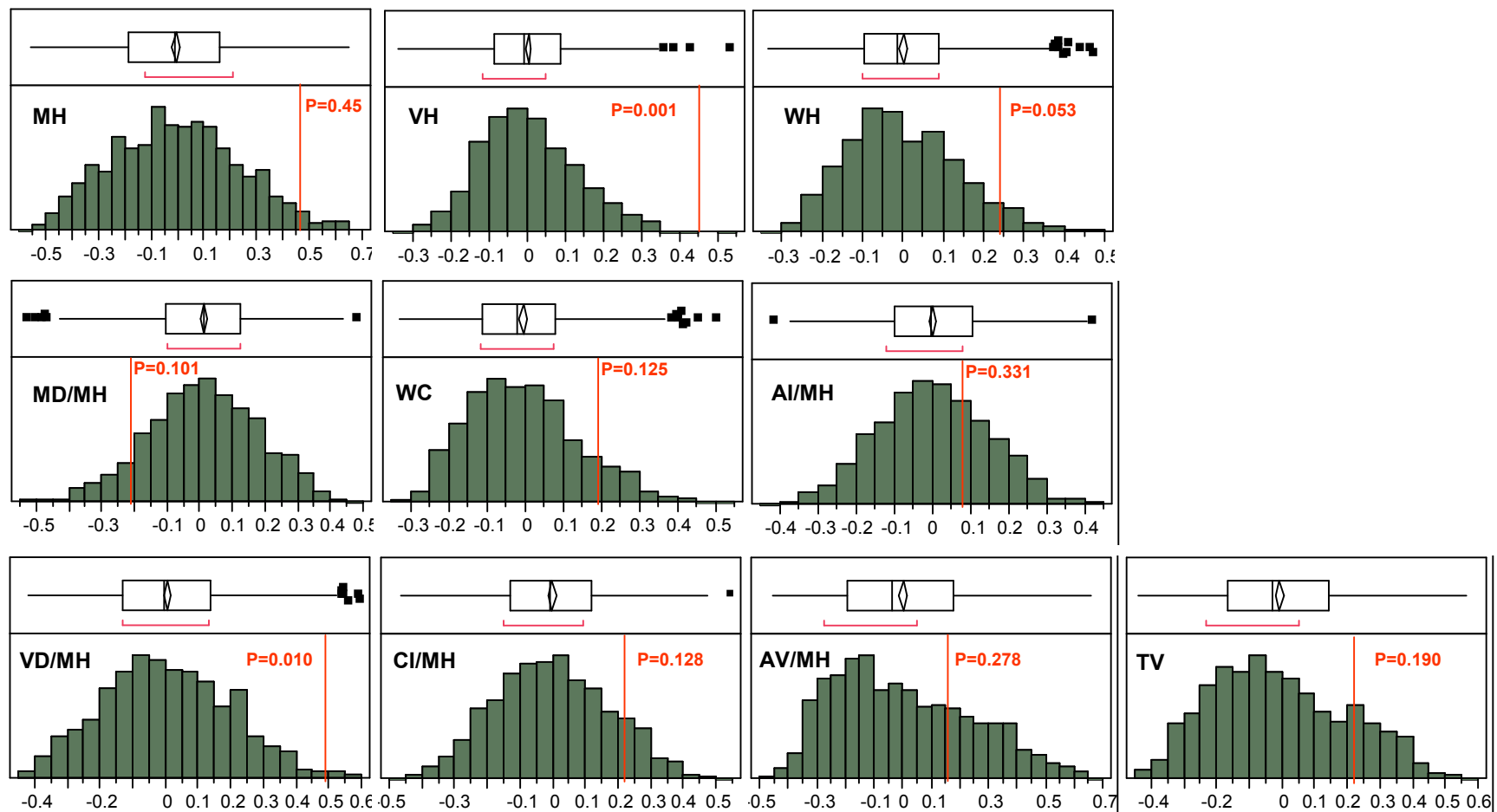
- 1) We performed 1000 simulations (using a hierarchical island model with 5 demes, 50 pops per demes, 100 individuals per population) using EasyPop (Balloux et al. 2006). For these simulations we adjusted migration rate in order to obtain ( $F_{ST}=0.21$ ,  $F_{SC}=0.071$  and  $F_{CT} = 0.149$ ; Rodriguez et al. submitted). We set a proportion of autogamy of 0.99, a mutation rate of  $10^{-5}$ , a distance between adjacent markers of 1 cM. We simulated 200 loci per each simulation. From each simulation we extract 11 populations;
- 2) For each simulation, the observed ecological data were then randomly assigned to the 11 populations;
- 3) For each of the 1000 simulations, correlations between variance components and ecological data were calculated using Nuragen.

Remarkably, this analysis confirmed the impact of water deficit on the single locus components and on the variance of disequilibrium (i.e., when a neutral scenario is considered, correlations are lower than when the observed SSAP data are considered,  $0.05 < P < 0.001$ ).

This analysis also confirms this tendency also for the WC and CI components, despite significance is not reached ( $0.10 < P < 0.12$ ) (Figure 3).

We also repeated the analysis considering shared polymorphism among populations. In this case, MD is stronger (and negatively) correlated with differences in water deficit than with geographical distance; WC is positively correlated with water deficit distance but not with geographical distance while AI does not show any significant correlation but has a positive sign. All this again suggests epistatic diversifying selection (Table 1,B).

**Figure 3** – Results of the simulations to test the significance of the correlations between variance components and geographical and ecological variables. Green histograms: distribution of the simulated correlation values. In red: observed correlation values.



## **Is there evidence for pair of loci under epistatic divergent or balancing selection?**

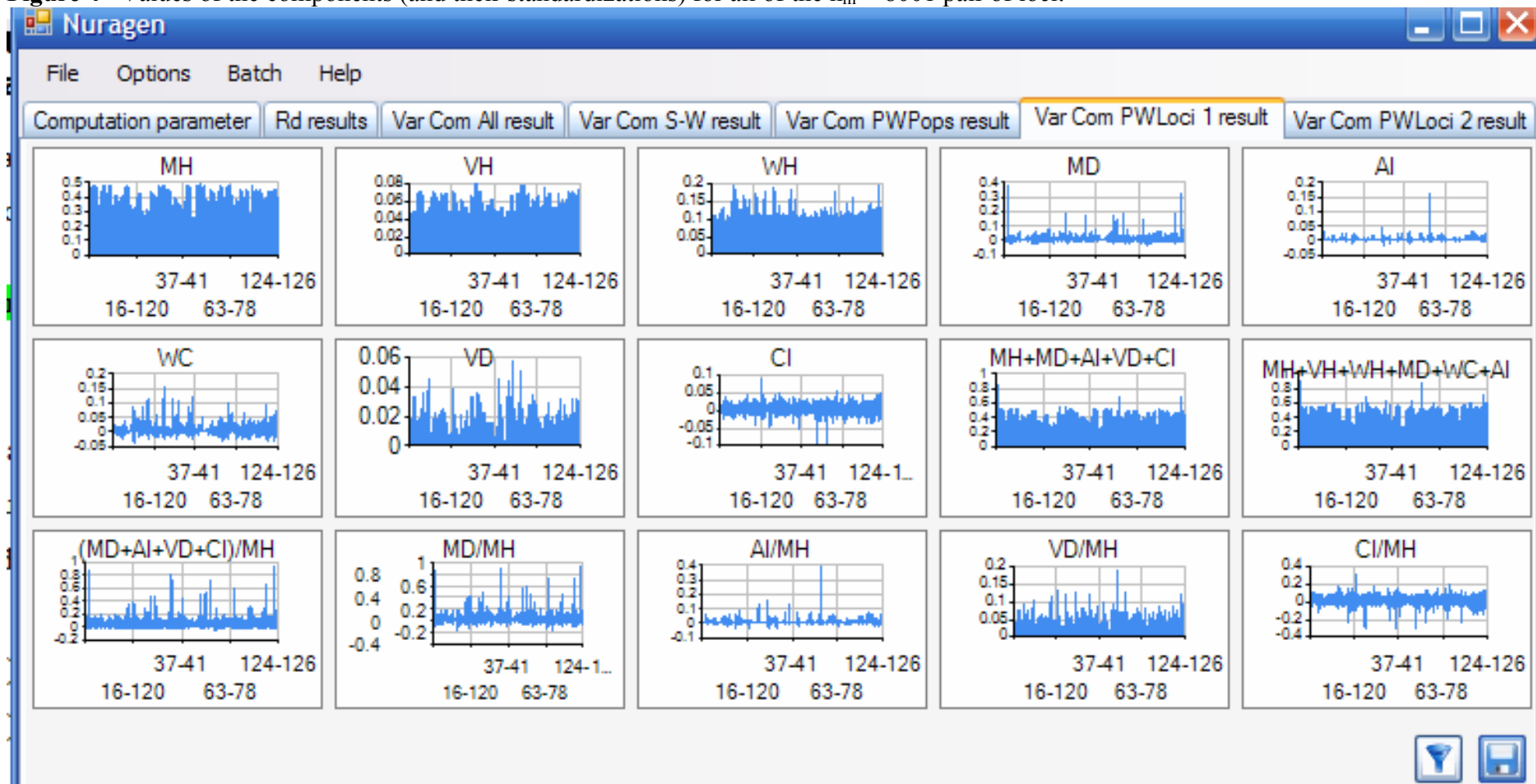
The analysis of the variance components was carried out for all of the  $n_m = m(m-1)/2$  possible pair of loci, with  $m$  the number of loci ( $m=127$  and  $n_m = 8001$  in our case; Figure 4). This analysis clearly shows that, for a given component, pair of loci can have very different values

In particular, it is evident that some pairs could have outlying behaviour.

Based on this observation, to identify pairs that could be under the effect of selection, we follow an approach articulated in three steps

- 1) We conduct a multivariate outlier analyses in the space of the eight variance components.
- 2) We tested the identified outliers for the patterns of the components expected under epistatic (balancing or divergent) selection.
- 3) We then defined indices of epistatic balancing and divergent selection and performed a model-based simulation to estimate the probability of our data.
- 4) Our results we then compared against well-established methods to detect.
  - i. single locus  $F_{ST}$  outliers for selection
  - ii. pair of loci under epistatic selection.
- 4) Finally, we also validate the candidates identified by studying the association between markers and phenotypic traits and between markers and known QTLs.

Figure 4– Values of the components (and their standardizations) for all of the  $n_m = 8001$  pair of loci.



### ***1) outlier analysis***

First, we conducted a multivariate outlier analysis in the space of the eight variance components. The analysis was conducted both for the SSAP dataset than for the simulated datasets. Both for the observed and simulated dataset we calculated Mahalanobis distance of each point from the multivariate mean (centroid) and obtained the distributions of the distances. Distances were computed with a jack-knife method, i.e., for each observation, estimates of the mean, standard deviation, and correlation matrix do not include the observation itself.

The Mahalanobis distance takes into account the correlation structure of the data as well as the individual scales. Multivariate distances are useful for spotting outliers in many dimensions. However, if the variables are highly correlated in a multivariate sense, then a point can be seen as an outlier in multivariate space without looking unusual along any subset of dimensions. Said another way, when data are correlated, it is possible for a point to be unremarkable when seen along one or two axes but still be an outlier by violating the correlation.

The idea here is that demographic processes are the main cause of the type and strength of the correlation among the different variance components. Thus, a locus pair that violates such correlation structure among variance components, can be seen as a pair upon which, superimposed to the demographic effects, selection might also be acting. Moreover, simulating a neutral scenario it is possible to calculate the probability that in the observed dataset a given distance from the centroid is due to demographic process only.

We simulated a neutral scenario using EasyPop software (Balloux et al. 2006). To allow for heterogeneous affinities between sampled populations, we simulated the genetic diversity by using an explicit hierarchical island model of population structure (Slatkin and Voelm, 1991), in which populations samples are assigned to different groups (defined *a priori*).

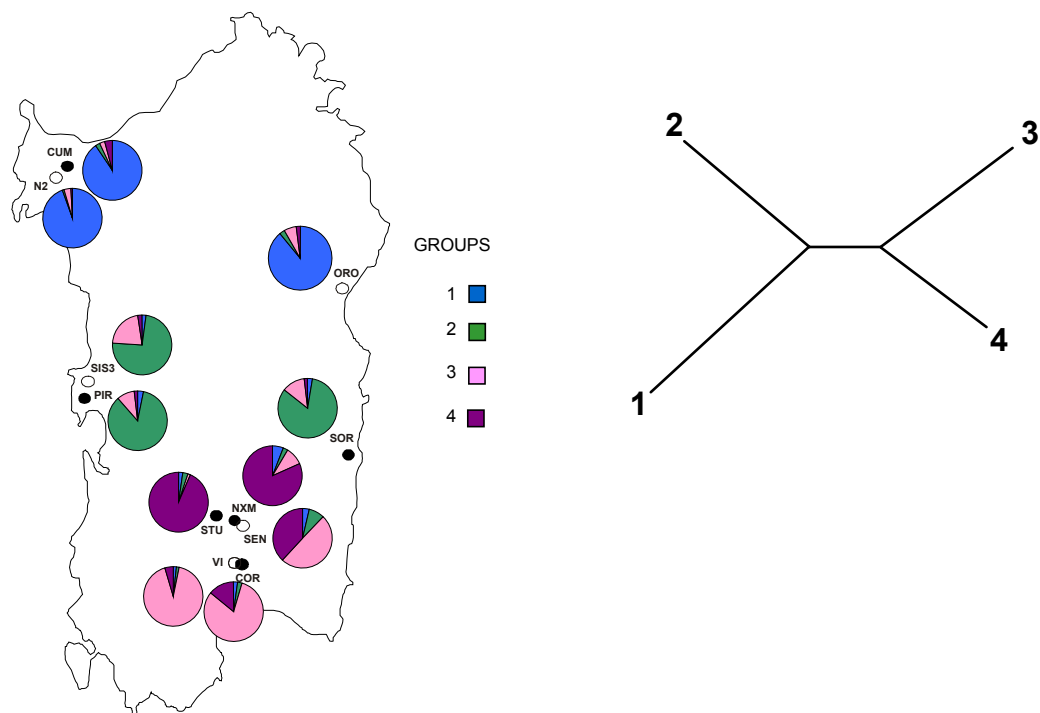
Based on the results of Rodriguez et al. (2011, submitted; figure here below), we consider this model for our system.

Simulations were conducted with 5 groups of 50 demes, with 100 individual per deme and 1000 loci using EasyPop software. The migration rates within and between groups were adjusted such as to have  $F_{SC} = 0.0714$  and  $F_{CT} = 0.149$ , implying an  $F_{ST}$  of 0.210



(Rodriguez et al. 20011, in press). We set 0.99 of autogamy, with recombination rate between adjacent markers of 1 cM, with mutation rate of  $10^{-5}$ .

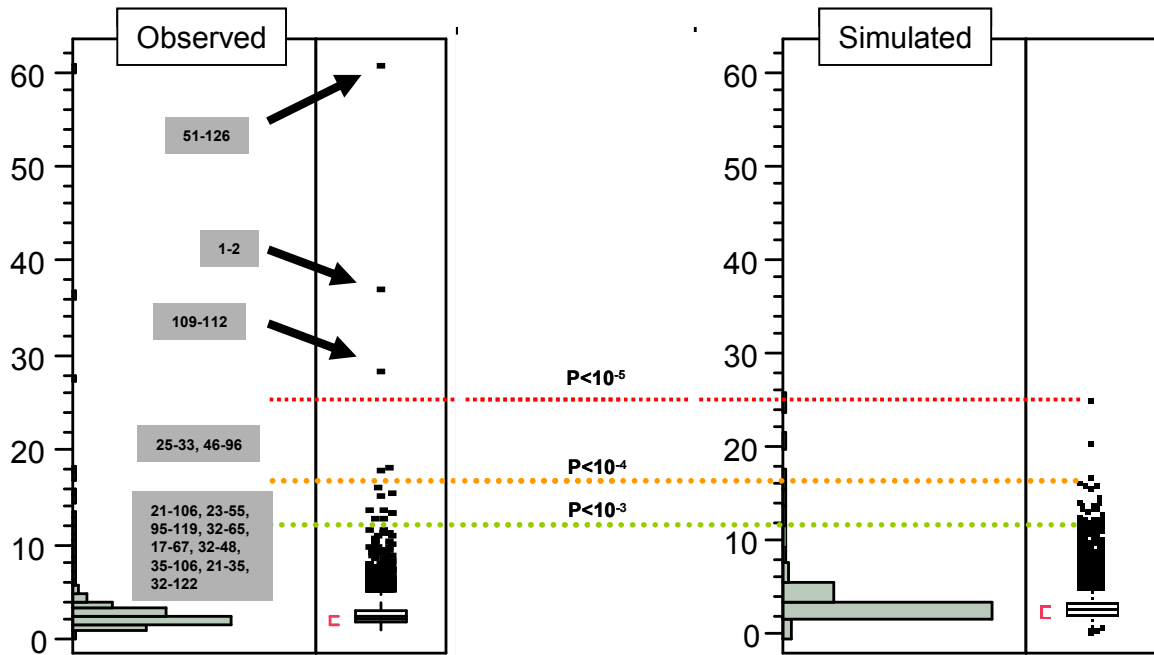
**Figure** - Map of Sardinia illustrating the distribution of the four genetic groups. Pie charts were obtained by using the mean value of q (membership coefficient) per genetic group within populations. The neighbor-joining tree on the side was obtained by the net nucleotide distances (allele frequency divergence) among Structure groups (from Rodriguez et al., 2011, submitted).



The simulated datasets produced by EasyPop were then processed with Nuragen to extract 11 population samples allocated as follows: 4 groups are chosen at random (among the 5 available group), and for each of these 5 selected groups, population samples are allocated to 3 randomly selected demes (among the 50 available demes) for three groups and to 2 randomly selected demes for the fourth randomly selected group. The number of individuals per population in each group was [26, 28, 27] , [20, 31, 29] , [25, 25, 29] , [29, 28]. We conduct 10 simulations. For each simulated dataset, after eliminating monomorphic loci, we consider 100 loci. We then have 1000 loci and 499500 pairs.

The distribution of the multivariate distances was then obtained both for observed than for simulated data. (Figure 5)

**Figure 5** – distribution of the jack-knife Mahalanobis distances both for the observed and simulated dataset. P value is defined as the proportion of simulated distances lower than the observed one.



We found 14 pairs for which the distance from the centroid is higher than expect under a neutral scenario, with  $P < 0.001$ . This number drop to 5 for  $P < 0.0001$  and to 3 for  $P < 10^{-5}$ . Correcting significance P for the number of locus pair (Adjusted  $P = 0.05/8001 = 6.25 \cdot 10^{-6}$ ) only three pairs (51-126, 1-2 and 109-112) were still significant.

However, being outlier does not necessarily mean satisfy certain conditions (as for example the conditions needed for epistatic balancing or divergent selection). Moreover, the same distance from the centroid could indicate very different patterns for variance components.

To shed some light on the behaviours our fourteen outliers, we defined a method specifically devoted to disentangle LD due to epistatic balancing selection from epistatic divergent selection.

Specifically, we contrast the ratio  $(WC+AI)/MH$  (a measure of between population LD standardized for within population diversity) against the ratio  $MD/(VH+WH)$  (a measure

of the repetition of LD across populations standardized for the between population variation in gene diversity and alleles frequencies). This has been done both for the SSAP dataset (observed dataset) and for 1000 loci simulated under the hierarchical model of population structure that gave 499500 pairs of loci. Results are presented in Figure (6).

**Figure 6**– Comparison between observed (red) and simulated (black) data. Observed values are 8001 SSAP markers pairs. Simulated data are 499500 loci pairs resulting from the simulation of 1000 loci evolved under a “neutral” hierarchical island model of population structure with Easypop software (Balloux et al. 2006). Outlier pairs are indicated by a two-numbers code.

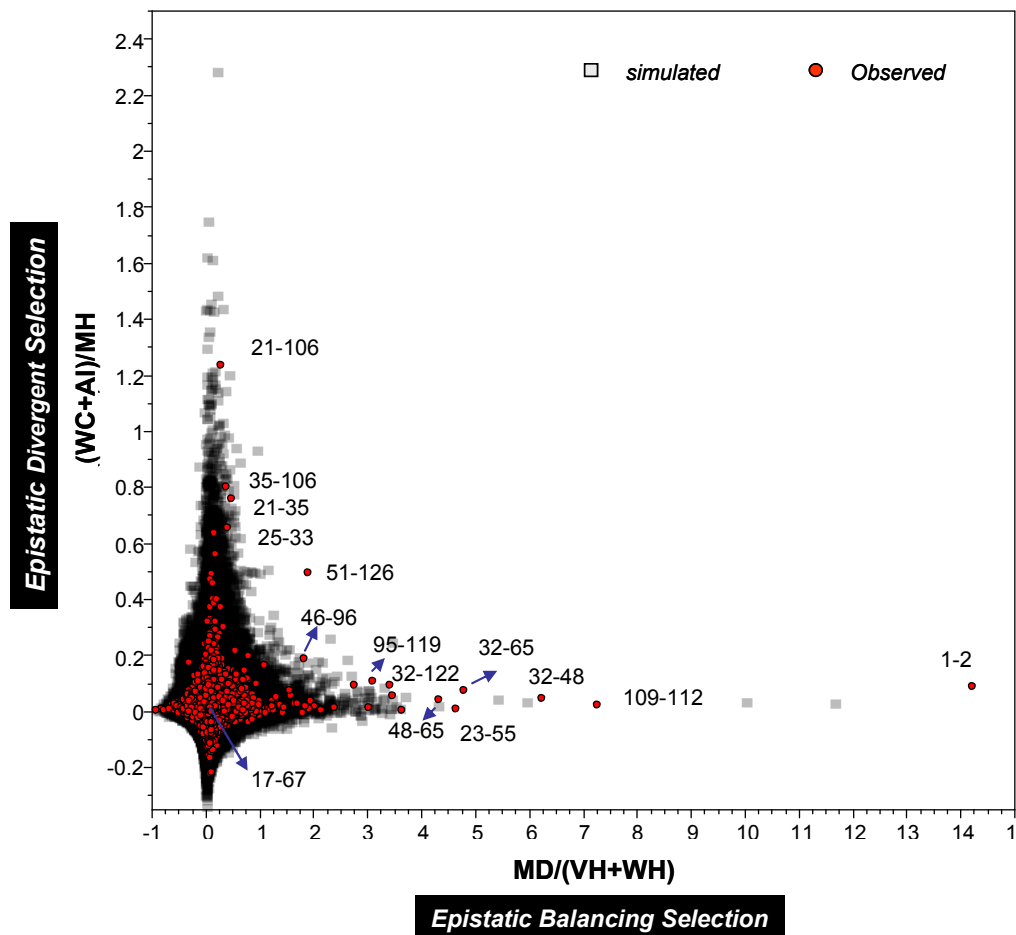
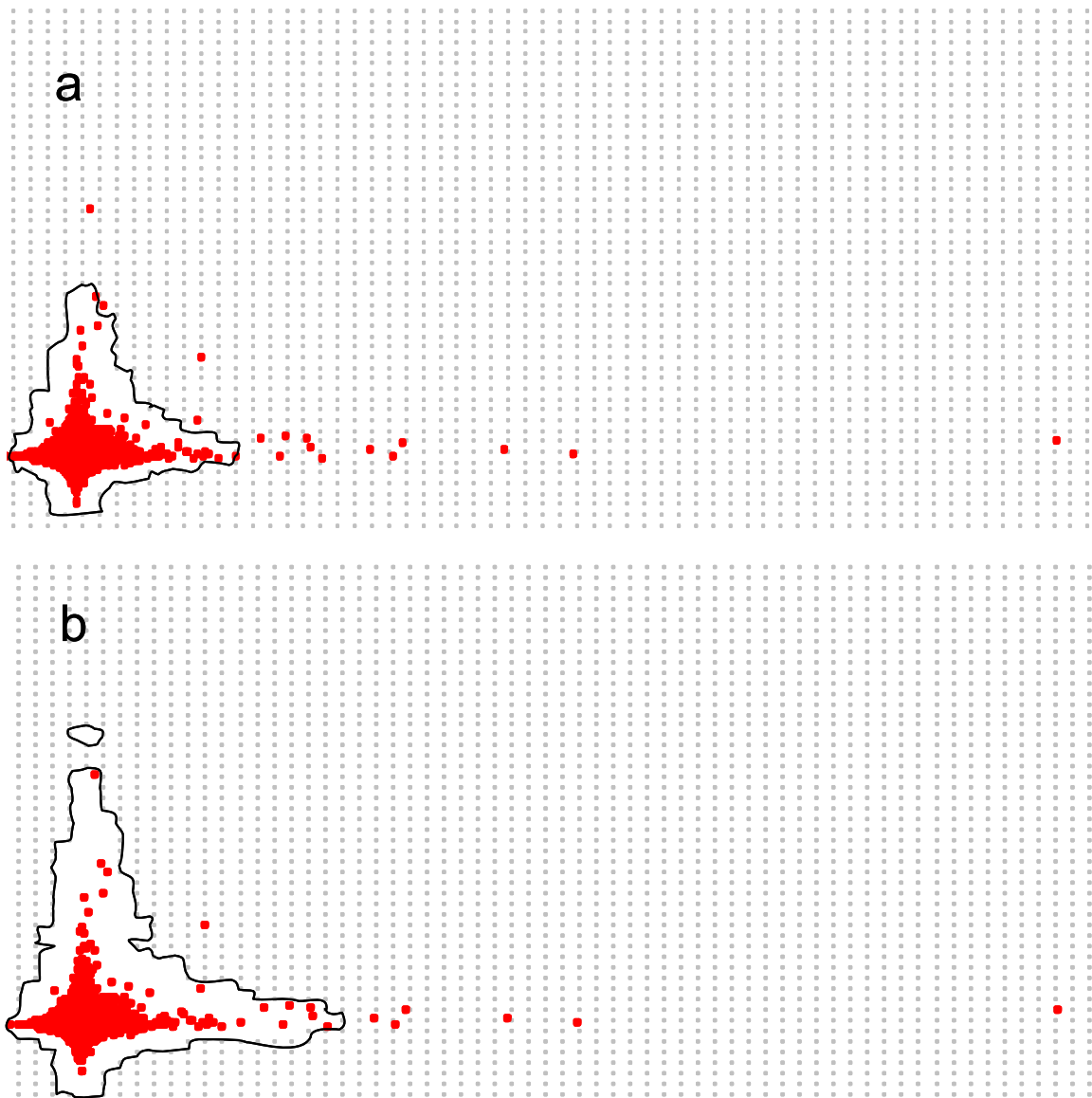


Figure 6 – density grids. Gray dots indicates combinations of MD/(VH+WH) and (WC+AI)/MH values that by simulation of a neutral hierarchical island scenario were obtained with frequency at least lower than 0.001 (a) or 0.0001 (b). Thus, observed values (red points) outside the continuous black line have values greater than the simulated ones with a probability (at least) lower than 0.001 (a) or 0.0001 (b).



Overall, among the 14 outliers, 6 have an MD/(VH+WH) ratio very high and outside the expectation obtained by simulation (1-2, 109-112, 32-65, 32-48, 23-55, 48-65; Figure 5). This suggests epistatic balancing selection for this loci pairs. Among this six pairs, 1-2 seems the best candidates. Three pairs, namely 32-122, 95-119, 46-96 have a pattern suggestive of balancing selection but they have an increasingly between-population component of LD. The pair 51-126 is peculiar in that it is intermediate between balancing selection and divergent selection.

The pairs that has the highest probability of being under epistatic divergent selection is 21-106 followed by, 35-106, 21-35 and 25-33. However, this three pairs are intermingled with simulated values more frequently than outlier for balancing selection. Finally, the pair 17-67 does not show any repetitive pattern of LD.

Thus, data suggest epistatic selection for 6 pairs (and particularly for one of them, 1-2) and , with less support, divergent selection for one pair (21-106).

## ***2) testing the significance of the variance components***

We further investigated the behaviour of our candidate pairs by testing the significance of the variance components. We conduct the analysis distinguishing between pairs that putatively under balancing or divergent selection.

### *Epistatic balancing selection*

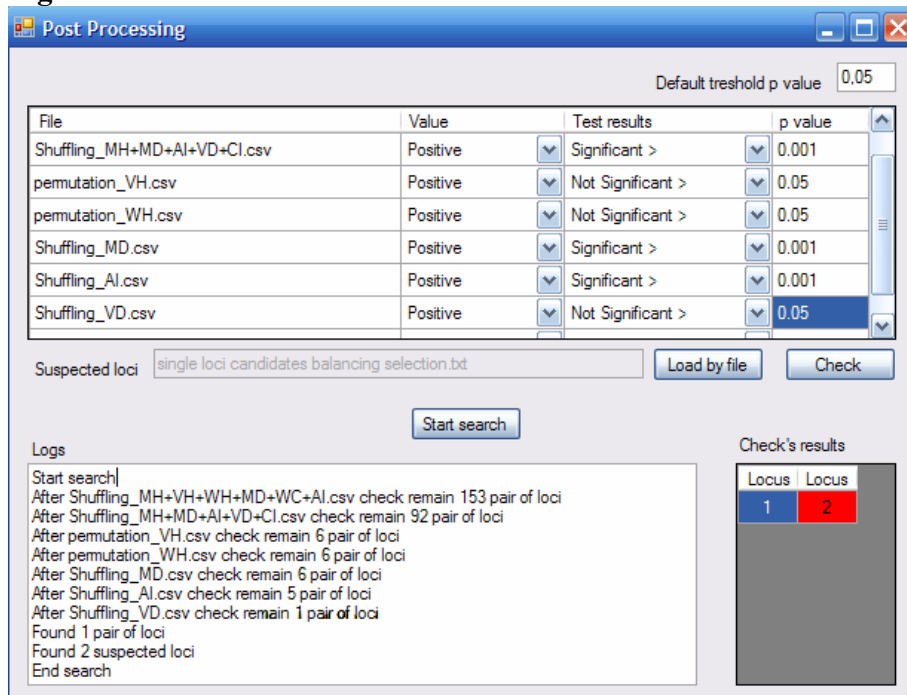
The analysis was conducted in three rounds. For each round we specified a series of conditions. The first four conditions are the same for all of the three rounds.

The first two conditions are that a pair should be in significant LD both in the total (TV) and in the average population (AV), i.e. TV and AV must be significant at shuffling of alleles across individuals within populations. The third and fourth conditions are that populations are not significantly differentiated for gene diversity and alleles frequencies (i.e. VH and WH not significant at permutation of individuals across populations).

It should be noted that testing the two-locus components by shuffling of alleles among individuals within populations, means that LD due to population differentiation is maintained in all the randomized datasets. Thus, if the observed dataset still shows significant associations, they cannot be due solely to population differentiation that, on the other hand, we asked that must be not significant (by permutation of individuals across populations).

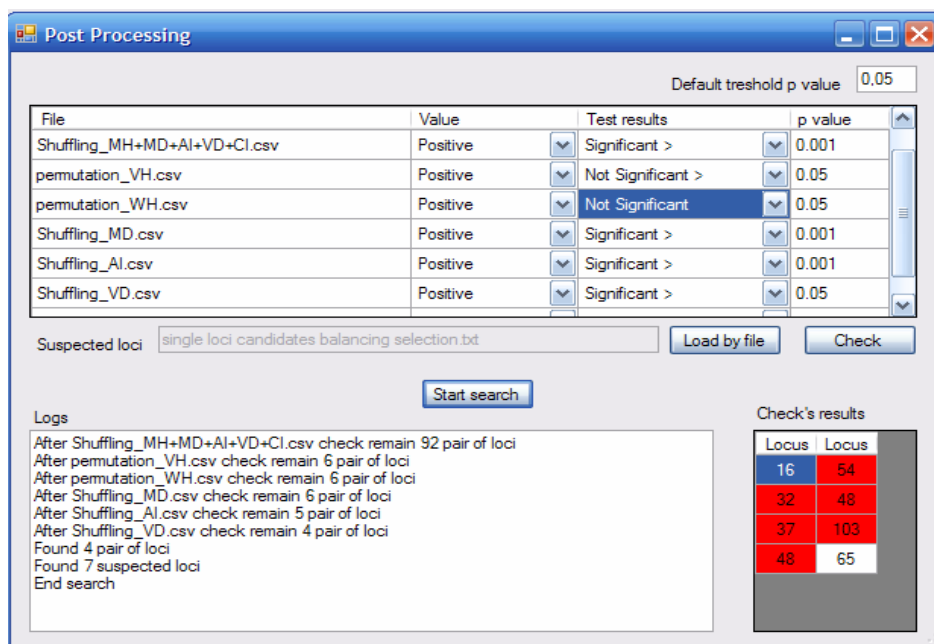
- in round 1, we asked that MD is the only significant source of LD. However, no locus pair that satisfies this condition has been detected in our dataset ;
- in round 2, we asked that AI is significant and positive; this means that if a between population component of LD is present (WC is different from zero) this must be repetitive of the LD within populations (MD). Additionally, we also asked that the variance of LD (VD) is not significant, i.e. that the strength of LD is not significantly different across populations. We found that only one locus pair satisfies these conditions: the locus 1-2 (Figure 7);

**Figure 7**– Results based on the conditions of round 2



In round 3 we investigate the remaining four candidates pairs (109-112, 32.48, 32-65 and 23-55) (Figure 8). All show significant VD indicating lower stable associations across populations than 1-2. However, the pair 32-48 has not significant population differentiation (figure) the pair 23-55, do not pass the filter because of a significant VH, while 109-112 and 32-65 show significant VH and WH ( $P < 0.05$ ) (not shown).

**Figure 8**– Filtering results based on the conditions of round 3

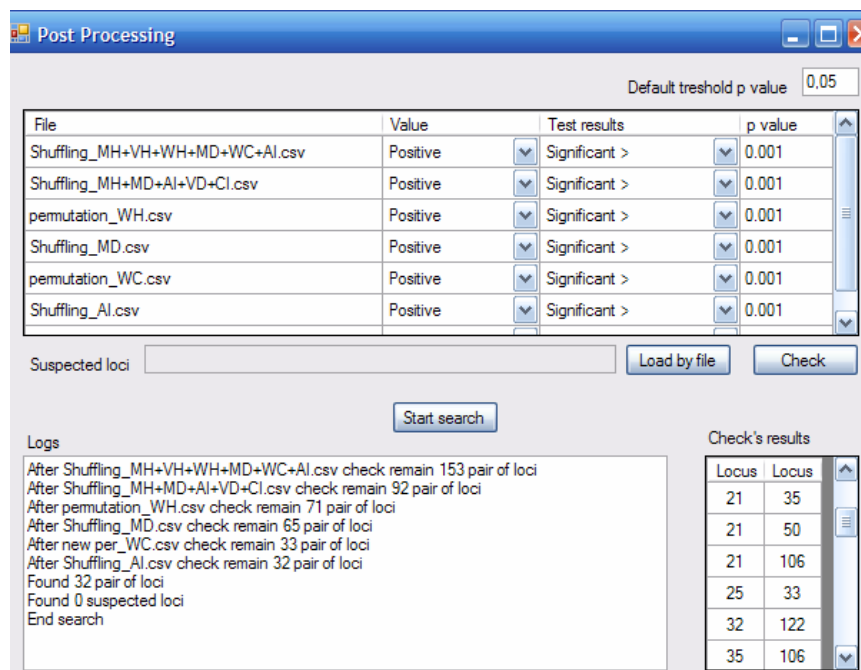


Concluding, overall our analysis suggests that the best candidate for balancing selection is the pair 1-2.

### *Epistatic divergent selection*

We look for pair of loci for which there is divergence in allele's frequency between populations (WH), repetitive LD (significant MD), significant covariance of allele's frequencies among populations (WC) and positive and significant interaction between MD and WC (AI). Moreover, when epistatic divergent selection is operating it might be expected the VD is also significant, i.e. that the frequency of the most frequent gametic pairs varies among populations. It should be noted, that this is not the only pattern that epistatic divergent selection can produce. Indeed, epistatic divergent selection could lead to patterns that are without any repetitive component of LD among populations.

**Figure 9** – Looking for pair with significant MH, MD, WC and AI.



Imposing the rules explicated in figure, we found 32 loci pairs (Figure 9). Among these we found the pairs:

1-81, 2-81, 7-104, 14-21, 15-21, **21-35**, 21-50, **21-106**, **25-33**, 32-122, **35-106**, 46-67, 49-69, 46-126, 51-96, 51-118, 51-126, 65-97, 76-88, 78-85, 78-92, 78-98, 81-101, 81-105, 84-122, 95-119, 96-126, 98-117, 108-112, 115-116, 118-124 and 118-126.



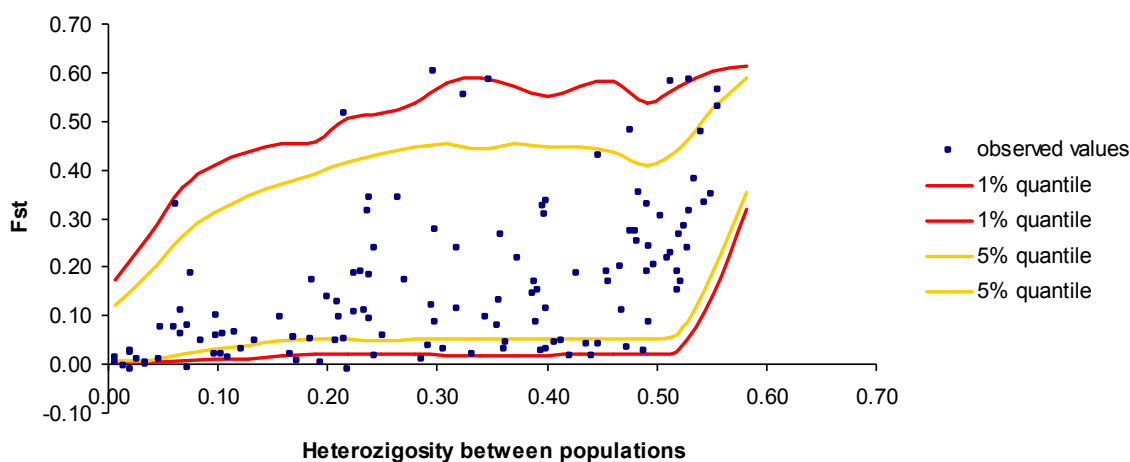
If we filter for significant VD ( $P < 0.05$ ), we obtained the following list of 24 pairs:

14-21, 15-21, **21-35, 21-106, 25-33**, 32-122, **35-106**, 46-67, 49-69, 46-126, 51-96, 51-118, 51-126, 76-88, 78-85, 78-92, 78-98, 81-105, 95-119, 96-126, 108-112, 115-116, 118-124 and 118-126.

Thus, all the four candidates for epistatic divergent selection (in the order 21-106, 35-106 - 21-35, 25-33) have “significant” patterns that are compatible with epistatic divergent selection.



**Figure 11** – Loci putatively under selection by assuming a hierarchical model of population structure using Arlequin software.



**Table 2** – List of loci putatively under balancing (A) or diversifying (B) selection detected by Mcheza – workbench for dominant selection (assuming an island model of population structure, I) or Arlequin vers. 3.5.1.2 software (assuming a hierarchical island model of population structure, H).

<i>Balancing selection</i>				<i>Divergent selection</i>			
<i>Mcheza</i>	$P^I$	<i>Arlequin</i>	$P^I$	<i>Mcheza</i>	$P^I$	<i>Arlequin</i>	$P^I$
(I)		(H)		(I)		(H)	
1	0.005(O) <sup>2</sup>	1	0.019	8	0.005(O) <sup>2</sup>	8	0.022
2	0.008(O)	2	0.016	9	0.012(F)		
5	0.015(F)	5	0.34	12	0.002(O)		0.051
6	0.022(F)	6	0.015	17	0.025(F)	17	0.031
	0.045	13	0.045			21	0.010
16	0.010(F)	16	0.020	25	<0.001(O)	25	0.007
	0.074	20	0.033	33	<0.001(O)	33	0.008
23	<0.001(O)	23	0.005	59	0.005(O)	59	0.022
24	0.008(O)	24	0.001	106	0.007(O)	106	0.001
28	0.013(F)	28	0.034		0.030	108	0.012
	0.153	30	0.027	116	0.015(F)		
	0.112	32	0.032	118	<0.001(O)	118	0.015
	0.195	37	0.038				
	0.087	38	0.004				
44	0.006(O)	44	0.008				
45	0.002(O)	45	0.017				
48	<0.001(O)	48	0.002				
	0.035	54	0.013				
	0.027	55	0.017				
62	0.003(O)	62	0.027				
	0.082	66	0.011				

	0.264	72	0.047
	0.035	75	0.016
80	0.008(O)	80	0.010
83	0.005(O)	83	0.035
86	0.006(O)		0.069
89	<0.001(O)	89	0.004
103	<0.001(O)	103	0.009
	0.052	109	0.008
112	0.015(F)	112	0.032

<sup>1</sup>  $Pr(\text{simulated } F_{ST}) > \text{observed } F_{ST}$ ; simulated values are obtained assuming an island (McHeza software) or a hierarchical island (Arlequin) model of population structure.

<sup>2</sup> F = not significant after correction for a false discovering rate of 0.10. T = significant after correction for a false discovering rate of 0.10.

For  $P < 0.05$ , McHeza found a higher number of loci under balancing selection loci (18) than Arlequin (29), however this difference lowered and reversed when  $P < 0.01$  (13 *versus* 9, respectively). Seven SSAP markers are indicated under balancing selection with high probability both by McHeza and by Arlequin software.

The number of loci putatively under divergent selection is similar number at  $P < 0.05$  (10 *versus* 9 for McHeza and Arlequin, respectively). However, when  $P < 0.01$  the number of loci is 7 and 3, respectively. Three SSAP markers are inferred under divergent selection with high probability both by McHeza and by Arlequin software (Table 2).

Thus, when a more stringent P value is used, the adoption of a hierarchical island model leads to a reduced number of loci putatively under selection (either putatively under balancing or divergent selection).

However, statistically convincingly signatures of both balancing and divergent selection were found in the set of population analysed.

***b) Ohta's analysis for epistatic selection***

Total linkage disequilibrium was subdivided into within- and between-populations components according to the method given by Ohta (1982a,b), similar to Nei's subdivision of the total diversity into the same components. The method was developed to distinguish between the different evolutionary forces that may generate disequilibria. A series of five different components was calculated and compared (Ohta, 1982a):  $D_{IS}^2$  and  $D'_{IS}$  are within-subpopulation components,  $D_{ST}^2$  and  $D'_{ST}$  are among-populations components and  $D_{IT}^2$  is the total population component of disequilibria. Analytical calculations have shown that when migration is limited among subpopulations, e.g. when disequilibria are created mainly by genetic drift, then  $D_{IS}^2$  is less than  $D_{ST}^2$  and  $D'_{IS}$  is greater than  $D'_{ST}$  (Ohta, 1982b). On the other hand, if systematic effects contribute to maintain allelic associations in the different populations then  $D_{IS}^2$  is greater than  $D_{ST}^2$  and  $D'_{IS}$  is less than  $D'_{ST}$ . The comparison of the values of the different components allows therefore a distinction between stochastic and systematic causes of gametic disequilibria. It is worth noting that  $D'_{IS}$  is a between population component, although it has been given by Ohta (1982a) the subscript  $IS$  (Kremer and Zanetto, 1996).

Thus, In order to understand if selection is acting to shape the structure of linkage disequilibrium, we applied the Ohta test to all of 8001 possible pair of loci to search for LD pattern consistent with epistatic selection. Results are summarized in Table 3.

Table 3 – Dual relationships and average values of Ohta's disequilibrium coefficients.

Dual relationship	N.of locus pairs	Loci pairs	$D_{IS}^2$	$D_{ST}^2$	$D'_{IS}$	$D'_{ST}$	$D_{IT}^2$	$D'_{IS}/D'_{ST}$	$D_{ST}^2/D_{IS}^2$
I. $D_{IS}^2 < D_{ST}^2$	7995	Average	0.003	0.088	0.090	0.001	0.091	69.154	30.310
II. $D_{IS}^2 > D_{ST}^2$	1	48 <sup>HH</sup> - 65	0.016	0.016	0.010	0.009	0.018	1.153	0.988
III. $D_{IS}^2 > D_{ST}^2$	5	1 <sup>HH</sup> - 2 <sup>HH</sup> 23 <sup>HH</sup> - 55 <sup>(0)HH</sup> 32 <sup>H</sup> - 48 <sup>HH</sup> 32 <sup>H</sup> - 65 109 <sup>H</sup> 112 <sup>HH</sup>	0.126 0.033 0.018 0.021 0.065	0.029 0.022 0.013 0.019 0.032	0.022 0.020 0.008 0.009 0.035	0.131 0.025 0.012 0.011 0.050	0.153 0.045 0.020 0.020 0.084	0.169 0.810 0.681 0.852 0.705	0.233 0.671 0.761 0.918 0.492
IV. $D_{IS}^2 < D_{ST}^2$	1	52 - 126	0.111	0.183	0.140	0.183	0.323	0.762	1.647

Note- I, H: significant adopting an island model or a hierarchical island of populations structure respectively. When I is between parentheses means that  $Pr(\text{simulated } F_{ST}) > \text{observed } F_{ST}$  is lower than 0.05 (whether or not this is considered an outlier for McHeza).

Among the 8001 pair of loci only 5 pairs of loci satisfy the III Ohta's dual relationships. i.e. the conditions expected if epistatic balancing selection is operating.

The  $D'_{IS^2} / D'_{ST^2}$  and  $D_{ST^2} / D_{IS^2}$  ratios are lower (i.e. indices of epistatic selection are stronger) for the pair 1-2 than for the other four pairs (Table 3). Interestingly, both loci composing this pair were indicated as putatively under balancing selection (PBS) whatever is the model of population structure used to simulate the neutral expectations (island or hierarchical island).

A similar situation is observed for the pair 23-55 is where, however, the locus 55 despite has  $P < 0.05$  is not considered a putatively under balancing selection by McHeza.

The pairs 32-48 and 109-112 are composed by loci that are always indicated as putatively under balancing selection if a hierarchical island model is used, while results are not concordant if an island model of population structure is used.

The fifth pair (32-65) contains one locus (32) that is indicated as putatively under balancing selection if a hierarchical island model is assumed by Arlequin while the second locus (65) has  $P = 0.03$  and  $P = 0.061$  if an island or a hierarchical island model of population structure is used, respectively.

One pair is of loci (48-65) satisfies the II dual relationship (that expected if epistatic section is operation not uniformly among populations). Locus 48 has always a strong support for balancing selection (Table 2) while locus 65 has  $P = 0.03$  and  $P = 0.061$  if an island or a hierarchical island model of population structure is used, respectively.

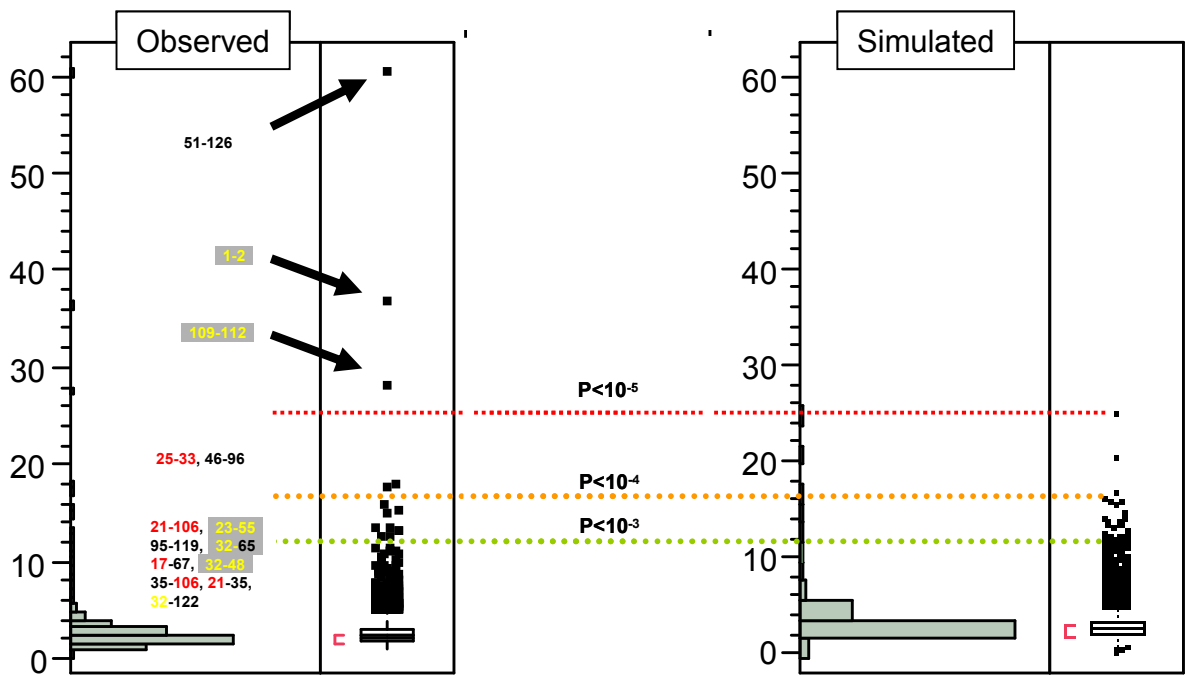
One pair of (51-126) has also an intermediate pattern but no one of these loci is identified as putatively under selection.

Thus, overall, it seems that the inference based on Ohta's LD analysis is more concordant with single locus  $F_{ST}$  outlier analysis based on a hierarchical structure of genetic variation than based on an island model.

c) comparing our list of candidate pairs with the results of the  $F_{ST}$ -based method and Ohta analysis.

Based on Ohta's analysis of LD in subdivided population there are five pairs putatively under epistatic balancing selection (Table 3). These five pairs were all identified by our outlier analysis with  $P < 0.001$  (Figure 12). Among these five pairs, the pair 1-2, that has the strongest signal of epistatic balancing selection (i.e. has the lowest  $D'_{IS^2} / D'_{ST^2}$  and  $D_{ST^2} / D_{IS^2}$  values) has the highest P value. As already above mentioned, in four of these pairs both loci were identified as under balancing selection by  $F_{ST}$  outlier analysis.

**Figure 12** – Comparison between the outlier analysis in the space of the eight variance components and others population genetic tests for selection. Gray background: pair under epistatic balancing selection based on Ohta analysis of LD in subdivided populations. Yellow (red): loci under balancing (divergent) selection based on  $F_{ST}$  outlier analysis.



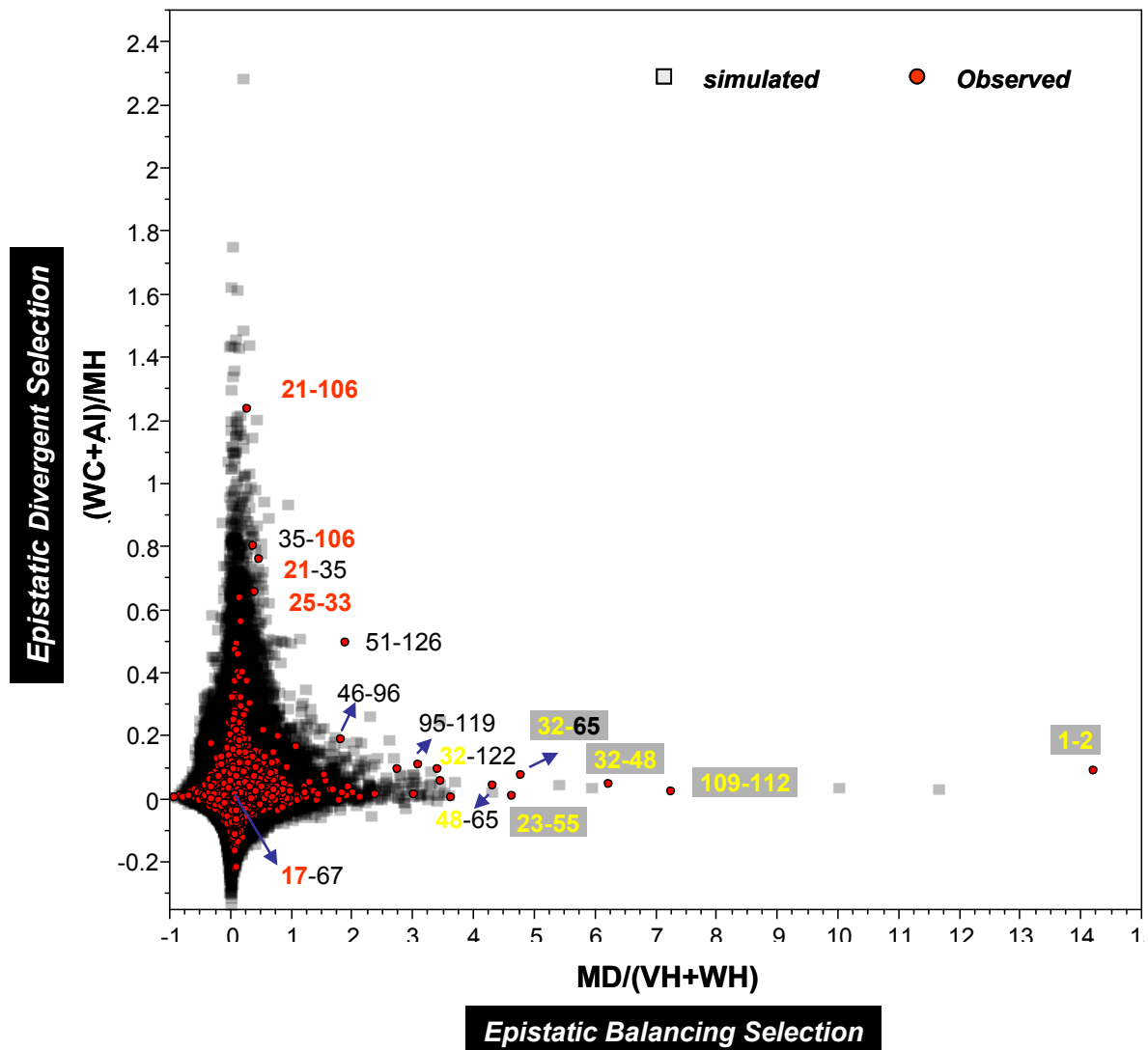
The pairs 46-96, 95-119 and 32-122 do not satisfy Ohta conditions for epistatic selection; however, they rank 9, 10 and 11 for  $D'_{IS^2} / D'_{ST^2}$  and 8, 11 and 13 for  $D_{ST^2} / D_{IS^2}$ . The pair 51-126 is the only characterized by the IV Ohta's dual relationship and our analysis, the most

distant from the centroid. The pair 48-65 (the only that satisfies the II Ohta's dual relationship) has P value in our analysis is 0.00174 (not shown in figure).

Thus, it is clear that the pool of 14 outliers identified in the space of the variance components is enriched of loci putatively under selection (balancing or divergent). On the other hand, such analysis is not able to distinguish between different outlier types. However, we found that this is possible by contrasting the ratios  $RB = MD/(WH+VH)$  and  $RD = (WC+WH)/MH$  (figure) and comparing observed and simulated data (Figure 13). Indeed, the first ratio is higher when low difference in genetic diversity (VH) and in allele's frequencies (WH) is associated with repetitive disequilibrium among populations. The second is higher when the between population components of linkage disequilibrium (WC and AI) is higher compared to the within population genetic diversity (MH). The difference among the two indices can be also useful in ranking the pairs for epistatic balancing (RB-RD) or epistatic divergent (RD-RB) selection.

Figure 13—Comparison between observed (red) and simulated (black) data. Observed values are 8001 SSAP markers pairs. Simulated data are 499500 loci pairs resulting from the simulation of 1000 loci evolved under a “neutral” hierarchical island model of population structure with Easypop software (Balloux et al. 2006). Outlier pairs are indicated by a two-numbers code. Loci putatively under divergent selection based on  $F_{ST}$  outlier analysis are indicated in red, those putatively under balancing selection are in yellow.





It is evident that pairs of loci putatively under divergent or balancing selection are well resolved, with pairs characterized by intermediate patterns occupying an intermediate position in the scatter plot. The position of the pairs 17-67 in this analysis (the only for which MD and AI components are not significant) suggests that this approach could be able to solve more than two types of patterns due to selection.

The fact that our method can distinguish between epistatic balancing and epistatic divergent selection seems a clear advantage compared to the method of Ohta. Moreover, our method allows an easier identification of pairs with intermediate pattern. Most important, by simulation we were able to evaluate the probability that a pair is under epistatic selection.

Multilocus tests of selection can be divided into two classes: model-based approaches that rely on assumptions about population structure and model-free approaches that are based on empirical distributions of summary statistics (Storz 2005). The advantage of empirical approaches is that outlier detection is not biased by model-based assumptions about population structure or history. However, choosing an empirical cut-off for identifying outliers in distributions of summary statistics may involve its own set of assumptions and biases (Storz 2005). For example, if the normal probability density function is used to identify outlier loci in an empirical distribution of ratios (e.g. Harr *et al.* 2002; Kauer *et al.* 2003a, Kauer *et al.* 2003b; Schöfl & Schlötterer 2004), loci with values that fall more than  $|1.96|$  standard deviations (SDs) from the mean of the distribution are typically interpreted as significant outliers (i.e., candidates for selection) at a two-tailed  $\alpha$ - level of 0.05 (Storz 2005). The problem is that if a given data set contains a large number of divergently selected loci, the empirical distribution will be characterized by an especially high SD. This will result in an overly conservative neutrality test because only the most extreme values would fall outside the 95% confidence interval. Conversely, a data set that contains only neutral loci will be characterized by a comparatively low SD, so outlier-detection tests would be more likely to identify false positives. Thus, type I and type II error rates are strongly dependent on the true number of selected loci that are included in the analyzed data set (Storz 2005).

Considering the outlier analysis on SSAP data, and the distribution of jackknife Mahalanobis distances (table here below), if we apply the criterion that pair of loci with values that fall more than  $|1.96|$  standard deviations (SDs) from the mean of the distribution can be interpreted as significant outliers (i.e., candidates for selection) at a two-tailed  $\alpha$  level of 0.05, we found 211 pairs, i.e. a quite huge number of candidates. Based on our simulation, we found 18 outliers with  $P < 0.001$  and 7 with  $P < 0.0001$ . To obtain the same number of candidates treating the observed distribution as a normal distribution, one needs to consider pairs that fall more than  $|5.75|$  and  $|8.68|$  standard deviations from the mean, a very high threshold. Thus, it seems that the use of model and simulations could be effective in reducing the number of false positives.

**Table** – distribution of jackknife Mahalanobis distances for 8001 SSAP pair of loci.

<b>Distribution of Jackknife Mahalanobis distance for the 8001 SSAP pair of loci</b>	<b>Quantiles</b>		
Mean	2.479	100.00%	maximum 60.11
Std Dev	1.431	99.50%	8.726
Std Err Mean	0.016	97.50%	5.395
upper 95% Mean	2.511	90.00%	3.733
lower 95% Mean	2.448	75.00%	quartile 2.898
N	8001	50.00%	median 2.191
Mean + 1.96 Std Dev	5.284	25.00%	quartile 1.72
Mean – \1.96 Std Dev	-0.325	10.00%	1.411
		2.50%	1.16
		0.50%	0.961
		0.00%	minimum 0.707

Finally, it should be remembered that, when diversifying selection is strong and gene flow is high, adaptive divergence in phenotypic trait values may occur primarily as a result of covariance in allele frequencies among QTL even in the absence of appreciable shifts in allele frequencies at individual loci (Le Corre and Kremer, 2003). Parallel differentiation of QTL allele frequencies will increase the level of trait divergence beyond that predicted by the additive effects of each locus considered separately, and the contributions of covariances to trait divergence will increase as a positive function of QTL number (Latta 1998, 2003; McKay & Latta 2002).

Thus, pair of loci that are outside the expectation but that do not include “outlier  $F_{ST}$ ” also deserve attention.

## VALIDATION

In this section, we address two main questions:

- 1) Are “candidates” associated with morphological traits?
- 2) Are they close to known QTLs?

To answer to the first question we studied the association between the 127 SSAP markers with 24 morphological traits (see appendix). Phenotypic data were available for 280 of the 297 barley lines used in this study. Data were from a one-year trial conducted in 1999 year at the experimental farm of Dipartimento di Scienze Agronomiche e Genetica Vegetale Agraria of Sassari University. For each trait, data subjected to the analysis are from four replicates. Association study then was performed by using the logistic regression method implemented in Tassel software using the population structure information obtained in Rodriguez et al. 2011 (Table 4).

To answer to the second question, we consider the map position available for 52 SSAP markers (Rodriguez et al. 2004). For each of these mapped markers we look for the closest known QTLs using Graingene web Database. Results of these analyses are reported in Table 5.

**Table 4**—Result of the markers-traits association study and associated QTLs.

Pair	Traits	P	R <sup>2</sup>	QTLs	Pair	Traits	P	R <sup>2</sup>	QTLs	Pair	Traits	P	R <sup>2</sup>	QTLs
1	NCA	*	0.040	Unmapped	21	IF NNC	* *	0.065 0.097	Unmapped	51	LUN HI	0.029 0.040	0.136 0.036	Grain protein Diastatic power
2	NCA	*	0.038	Unmapped	106	NNC	* *	0.036 0.063	Lodging grain protein	126	LUN NCF	0.039 0.043	0.137 0.120	???
109	NCF	**	0.130	Unmapped	35	PMS IF	* *	0.036 0.063	Unmapped					
112	PMS GP PP NCF	* * * 0.06	0.043 0.033 0.022 0.119	Unmapped	106	NNC	*	0.081	Lodging grain protein					
32	PMS NCF	* *	0.036 0.127	Unmapped	21	IF NNC	* *	0.065 0.097	Unmapped					
48	LAR	*	0.154	Unmapped	35	PMS IF	* *	0.036 0.063	Unmapped					
23	NCF	*	0.127	Unmapped	25	-	-	-	-					
55	PP GP	* *	0.023 0.027	Unmapped	33	-	-	-	Yield, lodging, height, grain protein					
32	PMS	*	0.036	Unmapped										

	NCF	*	0.127	
65	GP	**	0.045	Unmapped
	NCF	**	0.130	
	PP	*	0.031	
	NCP	*	0.033	
	PMS	*	0.040	
	LUN	*	0.140	
	PCSP	*	0.071	
48	LAR	*	0.154	
65	GP	**	0.047	
	NCF	**	0.130	
	PP	*	0.031	
	NCP	*	0.033	
	PMS	*	0.040	
	LUN	*	0.140	
	PCSP	*	0.071	

**Table 5**– Association between SSAP markers and QTLs in the Steptoe x Morex map. The association between SSAP markers and phenotypic traits of DH lines from the SxM cross is also reported (two last columns).

Chr.	Marker code	Name on the SxM map	Trait	QTL peak	SSAP-QTL distance (cM)	ANOVA (P,R <sup>2</sup> )	
2(2H)	106	BARE1-5M59c	Lodging	MWG557	2.7	0.001	0.08
			Grain protein			<0.001	0.14
3(3H)	33	BAGY2CM61k	Yield	ABG399	1.2	<0.001	0.26
			Lodging		1.2	<0.001	0.23
			Grain protein		1.2	n.s.	
			Height	BCD828	0.6	<0.001	0.16
	51		Grain protein				
7(5H)		BAGY2CM61v	Diastatic power	Ubi2	2.7	n.s.	
7(5H)	126	BARE1-5M59q	-	-	-	-	-

Loci 1 and 2 are associated with the same traits (number of culms per plant, NCA). Interestingly, 1-2 is the pair for which indices for balancing selection were the highest. Loci 109 and 112 are also associated to a similar traits, the number of fertile culms per plant, NCF), despite with lower significance level. Interestingly, association with NCF was also observed for the single locus 23, 32 and 65 where the R<sup>2</sup> is higher than for other associated traits.

Unfortunately, no one of these loci was mapped, thus we cannot know their relative distance and their position in relation to known QTLs. However, incidentally, it should be noted that maintenance of high levels of disequilibria across populations for these, can only be attributed to systematic forces. Indeed, even in the case of tight linkage, random causes

such as genetic drift will result in negligible values of  $D_{ST}^2$  (Ohta, 1982b). Thus, close linkage cannot be an explanation of the inequality  $D_{ST}^2 > D_{IS}^2$  that must be satisfied under epistatic selection. Moreover, it should be noted that repetitive disequilibrium often appears where complex loci whose peptide products are nearly identical and act in the same biochemical system are tightly linked and show allelic associations. This is the case for three loci of the Rh system and four loci of the HLA system in humans (Hedrick et al., 1978), or for three loci of the esterases in the plant species *Hordeum vulgare* (Clegg et al., 1972; Brown & Feldman, 1981) and in *Quercus petrae* (Kremer and Zanetto, 1996). If epistasis is synergistic, clustering of genes in the same chromosomal region as a coadapted complex or “supergene” would provide a large selective advantage (Charlesworth and Charlesworth 1975; Joron 2006).

However, is it conceivable that two loci (like those of the pair 1-2 and with lesser extent 109-112) that might subtend the traits “number of culms per plants”, are under balancing selection in our system? Interestingly, it is of note that Sardinian barley landraces are maintained by the farmers both for seed production and for the production of green biomass for sheep grazing (Papa et al. 1998). There are several examples in literature that suggest as in highly variable environment such as the Sardinian ones, the number of culms is a key factor for the adaptation of the barley plant (i.e., for its fitness). Indeed, modulating the number of culms, the plant is capable to increase (in favorable conditions) or reduce (in adverse conditions) its “investment”. Moreover, the number of culms is (positively) correlated to the number of leaves and, indeed, with the biomass production. This suggests that sheep grazing could have exerted a selective pressure on this trait.

Regarding the pairs putatively under epistatic divergent selection, both loci of the pair 21-106 were associated with the number of nodes per culm, a trait that is known to be relevant for determining lodging resistance. Noteworthy, locus 106, that is mapped on the Steptoe x Morex map, is at 2.7 cM from QTLs for lodging and grain protein and, noteworthy, it is also correlated with the lodging score of the SxM DH lines ( $P < 0.001$ ,  $R^2 = 0.08$ , Table 5). Again, it must be noted that another single locus (33) is close to a QTL for plant height, lodging and grain yield with correlation with the phenotypic traits of the DH lines ranking from  $R^2 = 0.16$  to  $R^2 = 0.26$ . Interestingly, also the pair 51-126 is associated with a trait that is related to the resistance to lodging: the length of the last node of the culm. In accord,

deserve to be noticed, that Rodriguez et al. (2008) in a multi environmental trials in Sardinia have clearly shown that lodging and plant height are the most important traits in explaining the strong interaction genotype x environment for yield.

## **Conclusion**

In this Thesis we have presented a new software for the analysis of population structure of LD. We also show that it can be useful to solve several questions related to the biological significance of LD in plant populations. In particular, we have shown that this software, in junction with other available programs, could be effective for the identification of pair of loci under epistatic balancing and divergent selection.



## References

- Balloux 2006. EasyPop. A program for forward in time simulation.
- Beaumont MA, Balding DJ (2004). Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol* 13: 969–980.
- Beaumont MA, Nichols RA (1996). Evaluating loci for use in the genetic analysis of population structure. *Proc R Soc Lond B* 263: 1619–1626.
- Brown et al. 1980; Maynard Smith et al. 1993; Haubold et al. 1998
- Brown, A. H. D., Feldman, M. W. & Nevo, E. (1980) *Genetics*
- Brown, A.H.D. & Feldman, M.W (1981) *Proc. Natl Acad. Sci. USA* 78, pp. 5913-5916.
- Burt A, Carter DA, Koenig GL, White TJ, Taylor JW (1996) Molecular markers reveal cryptic sex in the human pathogen *Coccidioides immitis*. *Proc Natl Acad Sci USA* 93:770–773
- Cavalli-Sforza LL (1966). Population structure and human evolution. *Proc R Soc Lond B Biol Sci* 164: 362–379.
- Charlesworth D and Charlesworth B, 1975. Theoretical genetics of Batesian mimicry II.
- Clegg, M. T., Allard, R. W. & Kahler, A. L. (1972) *Proc. Natl. Acad. Sci. USA* 69, 2474-2478
- Egan SP, Nosil P, Funk DJ (2008). Selection and genomic differentiation during ecological speciation: Isolating the contributions of host association via a comparative genome scan of *Neochlamisus bebbianae* leaf beetles. *Evolution* 62: 1162–1181.
- Evolution of supergenes. *J Theor Biol.* 55: 305-324
- Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity*.
- Foll M, Gaggiotti OE (2008). A genome scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics* 180: 977–993
- Harr B, Kauer M, Schlötterer C (2002) Hitchhiking mapping — a population-based fine mapping strategy for adaptive mutations in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 12949–12954.
- Hedrick, P., Jain, S. & Holden, L. (1978) *EvoL Biol.* 11, 101-184.
- Joron M, Papa R, Beltran M, Chamberlain N, Mavarez J, *et al.*, 2006. A conserved supergene locus controls colour pattern diversity in *Heliconius* butterflies. *PLoS Biol.* 4: e303
- Kauer M, Dieringer D, Schlötterer C (2003a) A microsatellite variability screen for positive selection associated with the ‘out of Africa’ habitat expansion of *Drosophila melanogaster*. *Genetics*, **165**, 1137–1148.
- Kauer M, Zangerl B, Dieringer D, Schlötterer C (2003b) Chromosomal patterns of microsatellite variability contrast sharply in African and non-African populations of *Drosophila melanogaster*. *Genetics*, **160**, 247–256.

- Kremer A and A. Zanetto (1997) Geographical structure of gene diversity in *Quercus petraea* (Matt.) Liebi. II: Multilocus patterns of variation *Heredity* 78 476—489
- Latta RG (1998) Differentiation of allelic frequencies at quantitative trait loci affecting locally adaptive traits. *American Naturalist*, **151**, 283–292.
- Latta RG (2003) Gene flow, adaptive population divergence and comparative population structure across loci. *New Phytologist*, **161**, 51–58
- Le Corre V, Kremer A (2003) Genetic variability at neutral markers, quantitative trait loci and trait in a subdivided population under selection. *Genetics*, **164**, 1205–1219
- Lewontin RC, Krakauer J (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* 74: 175–195.
- Mañkinen HS, Cano JM, Merila J (2008). Identifying footprints of directional and balancing selection in marine and freshwater three-spined stickleback (*Gasterosteus aculeatus*) populations. *Mol Ecol* 17: 3565–3582
- McKay JK, Latta RG (2002) Adaptive population divergence: markers, QTL and traits. *Trends in Ecology and Evolution*, **17**, 285–291.
- Ohta M. 1982a. Linkage disequilibrium due to random genetic drift in finite subdivided populations. *Proc. Natl. Acad. Sci.*, 79, 1940—1944,
- Ohta, M. 1982b. Linkage disequilibrium with the islandmodel. *Genetics*, 101, 139—155.
- Riebler A, Held L, Stephan W (2008). Bayesian variable selection for detecting adaptive genomic differences among populations. *Genetics* 178: 1817–1829.
- Rodriguez M, Rau D, Papa R, Attene G (2008) Genotype by environment interactions in barley (*Hordeum vulgare* L.): different responses of landraces, recombinant inbred lines and varieties to Mediterranean environment *Euphytica* 163:231–247
- Rodriguez, O’Sullivan D., Leigh F., Papa R. E Attene G, et al. 2004. Integrating retrotransposon markers in barley map. *Molecular breeding*.
- Schlotterer C (2002). A microsatellite-based multilocus screen for the identification of local selective sweeps. *Genetics* 160: 753–763.
- Schöfl G, Schlotterer C (2004) Patterns of microsatellite variability among X chromosomes and autosomes indicate a high frequency of beneficial mutations in non-African *D. simulans*. *Molecular Biology and Evolution*, **21**, 1384–1390.
- Slatkin M, Voelm L (1991) FST in a hierarchical island model. *Genetics* 127, 627.-629
- Smouse, P. E., and J. C. Long. 1992. Matrix correlation analysis in Anthropology and Genetics. *Y. Phys. Anthop.* 35:187-213.
- Smouse, P. E., J. C. Long and R. R. Sokal. 1986. Multiple regression and correlation extensions of the Mantel Test of matrix correspondence. *Systematic Zoology* 35:627-632.

Sokal, R. R., and F. J. Rohlf. 1981. *Biometry*. 2nd edition. W.H. Freeman and Co., San Francisco, CA.

Storz JF (2005). Using genome scans of DNA polymorphism to infer adaptive population divergence. *Mol Ecol* 14: 671–688.

Tang K, Thornton KR, Stoneking M (2007). A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol* 5: e171.

Thornton KR, Jensen JD (2007). Controlling the false-positive rate in multilocus genome scans for selection. *Genetics* 175: 737–750.