USER'S MANUAL
BIO-R (Biodiversity Analysis with R)

Angela Pacheco[1], Gregorio Alvarado[1], Francisco Rodríguez[1], José Crossa[1], and Juan Burgueño[1]

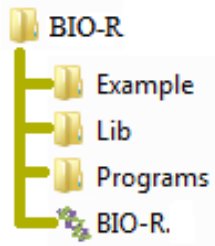
[1]Biometrics and Statistics Unit,Genetic Resource Program,CIMMYT,Batán México.

2016

1. BIO-R description

BIO-R is a set of R programs that do the biodiversity analysis, in order to calculate heterozygosity, diversity among and within groups, shannon index, number of effective allele, percent of polymorphic loci, Rogers distance, Nei distance, cluster analysis and multidimensional scaling 2D plot and 3D plot; you can included external groups for colored the dendogram or MDS plots, and additionally you can obtain a Core Subset. BIO-R was designed because is necessary to do biodiversity analysis easily. BIO-R contains a graphical JAVA interface that helps the user to easily.

When you install BIO-R, you will can start the program directly of the programs window or you will search the folder in C:\, the folder contains the following files:



- **Examples.** Folder where you can find examples files in .csv format.

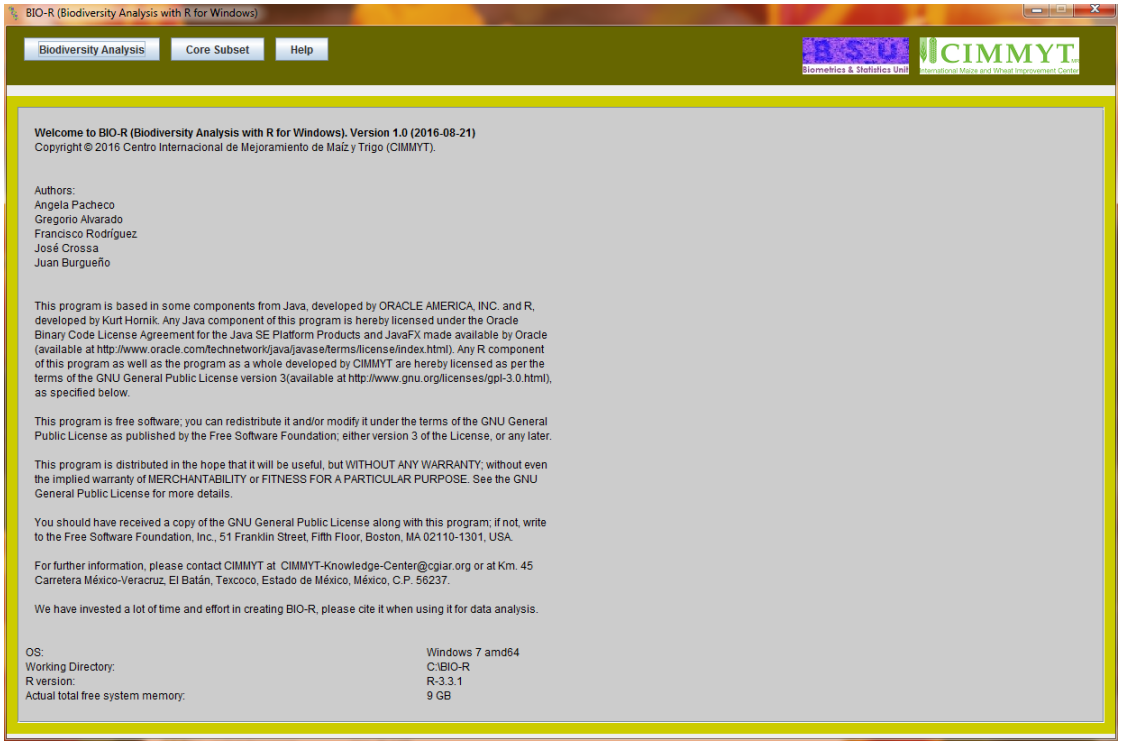
- **lib.** Folder necessary for run the application.
- **Programs.** Contains all the necessary R packages that will run automatically.
- **BIO-R application.** JAVA interface used to choose analysis options.

2. BIO-R requirements

- WINDOWS operating system.
- JAVA updated.

3. Installing the necessary packages and loading the R-version

When JAVA interface BIO-R is opened, a screen like the one below appears. This screen is signal that everything works good, no need to install anything else, all is included in the software.The first step is to open the input file to be analyzed. This step is as follows:

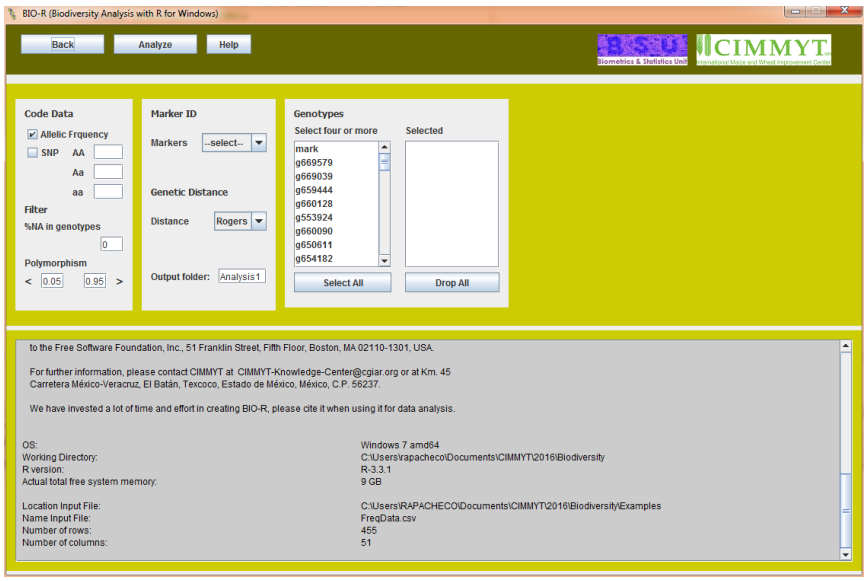


4. Biodiversity Analysis parameters

Input files for diversity analysis can be opened by clicking on *"Biodiversity Analysis"* and choosing a comma separated file (*.csv) with the data saved wherever you want. The input data is a matrix with allelic frequency or SNP's of one of the alleles for each markers, where the columns are the genotypes and the rows are the markers, the first column must be the names of markers, it is recommended that both, the names of the markers as the names of the genotypes, are short and without strange characters (e.g. \,_,*,-,etc). Specifically, the names of the genotypes must start with a letter. If there are missing values in the input file, these have to be indicated by "." or "NA". Blank spaces are not allowed. Data should look like this:

	A	B	C	D	E	F	G	H	I	J	K	L
1	mark	g669579	g669039	g659444	g660128	g553924	g660090	g650611	g654182	g660981	g553426	g650375
2	1	0.666667	NA	0.571429	NA	1	0	0.777778	1	NA	0	NA
3	2	NA	1	1	1	NA	1	1	1	1	NA	1
4	3	1	1	0.52	0.5	0.8	NA	0.7	0	NA	1	0.263158
5	4	1	NA	0.5	1	NA	NA	0.939394	0.176471	NA	0.666667	0.428571
6	5	1	NA	1	NA	1	NA	1	1	1	1	1
7	6	0.666667	0	0.714286	1	0.333333	NA	0.363636	1	NA	1	0.583333
8	7	1	NA	0	1	1	NA	0	NA	NA	1	0
9	8	1	NA	1	1	1	1	1	NA	1	1	1
10	9	NA	1	0.6	0.217391	0.708333	NA	0.833333	1	NA	0.846154	1
11	10	1	1	1	1	1	NA	1	1	1	1	1
12	11	1	NA	0.25	0.375	0.333333	0.896552	0.444444	NA	1	NA	0
13	12	0.076923	1	0.333333	0.8	0.5	0	0.9	1	0	0.875	0
14	13	0.571429	NA	0.666667	0.5	1	0	0.333333	1	0.333333	0.75	0.857143
15	14	1	NA	1	1	0.923077	1	1	1	1	1	1

To analyze the data, first you must choose your parameters.



4. BIODIVERSITY ANALYSIS PARAMETERS

Firstly, you need indicated if your data set have a allele frequency (select **Allelic Frequency**) or SNP information (select **SNP**). If you select **SNP** option, you have to write the numbers that identify dominant homozygote (**AA**), heterozygote (**Aa**) and recessive homozygote (**aa**).

Then you can do a **filter**, by percent of missing values in each genotype and/or percent of polymorphism, the values must be between [0,1].

In **Markers**, you must select the column that identify the markers.

In **Distance**, you must select the method to will used for calculated the distances: Rogers distance or Nei distance.

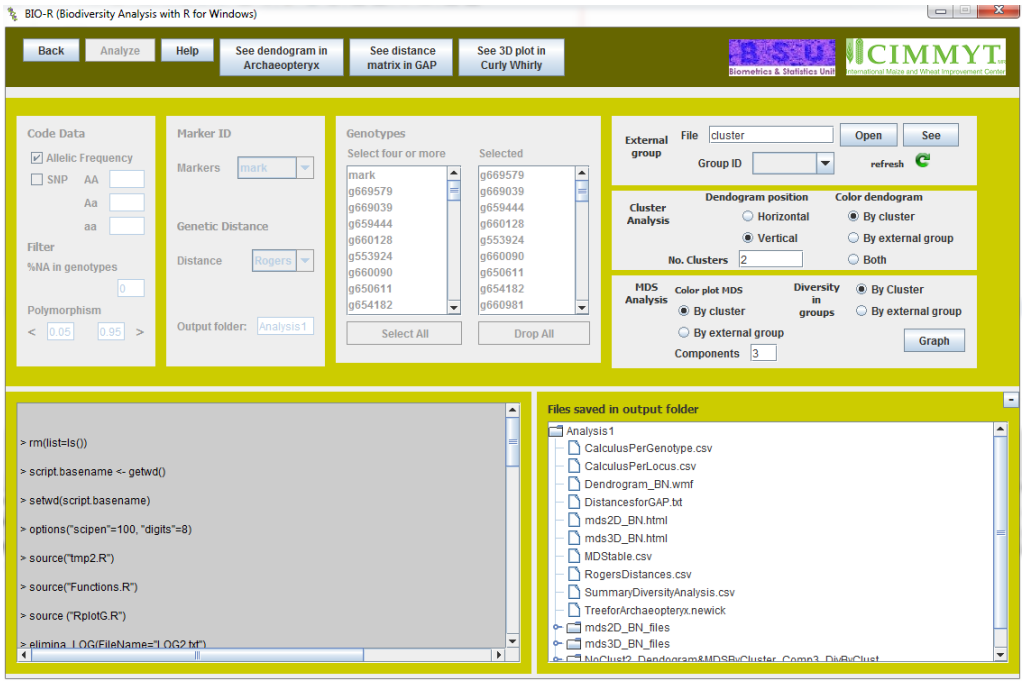
In the field **Output folder**, you must type the name of the output folder where results will be saved; it will be created inside the Output_BIO-R folder. You can change the name to separate outputs of different data sets. Is necessary to change the name of the output folder for each analysis.

In the frame **Genotypes**, you must select the columns that represent the genotypes for analyze.

When you have selected all the necessary parameters just press **"Analyze"** to start with the analysis.

After having conducted an analysis you have the option to plot the dendrogram and MDS graphic, and calculated the diversity in groups, just pressing the **"Graph"** button; the software will automatically generate a new folder to store the new graphics. For do this you must to select the next parameters:

4. BIODIVERSITY ANALYSIS PARAMETERS



If you want consider an **External group**, you must **"Open"** a *.csv file, this should contain in the first column the names of genotypes, in **Group ID**, you need to select the variable that will used like a groups.

	A	B	C
1	Gen	numcolor	faccolor
2	g669579	11	I
3	g669039	3	A
4	g659444	8	F
5	g660128	11	I
6	g553924	4	B
7	g660090	8	F
8	g650611	6	D
9	g654182	7	E
10	g660981	14	L
11	g553426	15	M
12	g650375	15	M
13	g623127	5	C
14	g650594	15	M
15	g607289	6	D
16	g668799	9	G
17	g660829	11	I
18	g660255	14	I

If you have chosen to open a file, you can see your file by clicking in **"See"** button.

The refresh button is useful for reset the parameters and use the default.

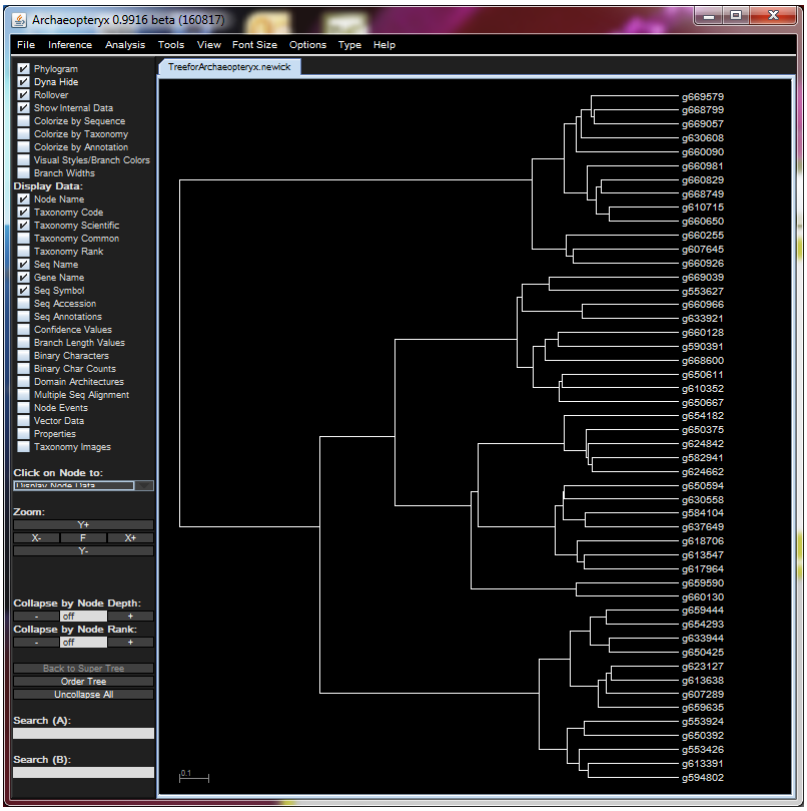
4. BIODIVERSITY ANALYSIS PARAMETERS

In **Cluster Analysis** you must write the number of groups that you need to divide the population in **No. Clusters**, choose the position of the dendogram (**Horizontal** or **Vertical**) and will choose as colored, either according to the groups that formed the cluster analysis (**By cluster**), by groups that can be added (**By external group**) or both (**Both**).

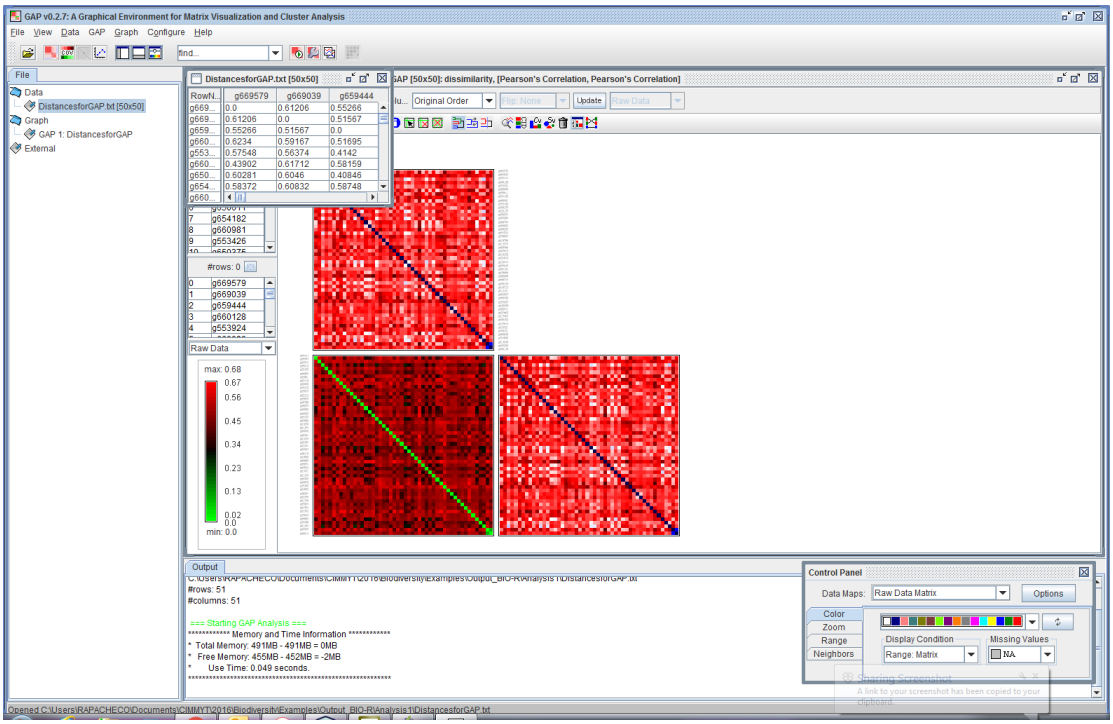
In **MDS Analysis** choose **By cluster** if you want colored MDS graph according to the groups that formed the cluster analysis or choose **By external group** if you want colored MDS graph by the groups added, by default the software obtain 3 components for the MDS analysis, but if you want more, you must change the number in parameter **"Components"**.

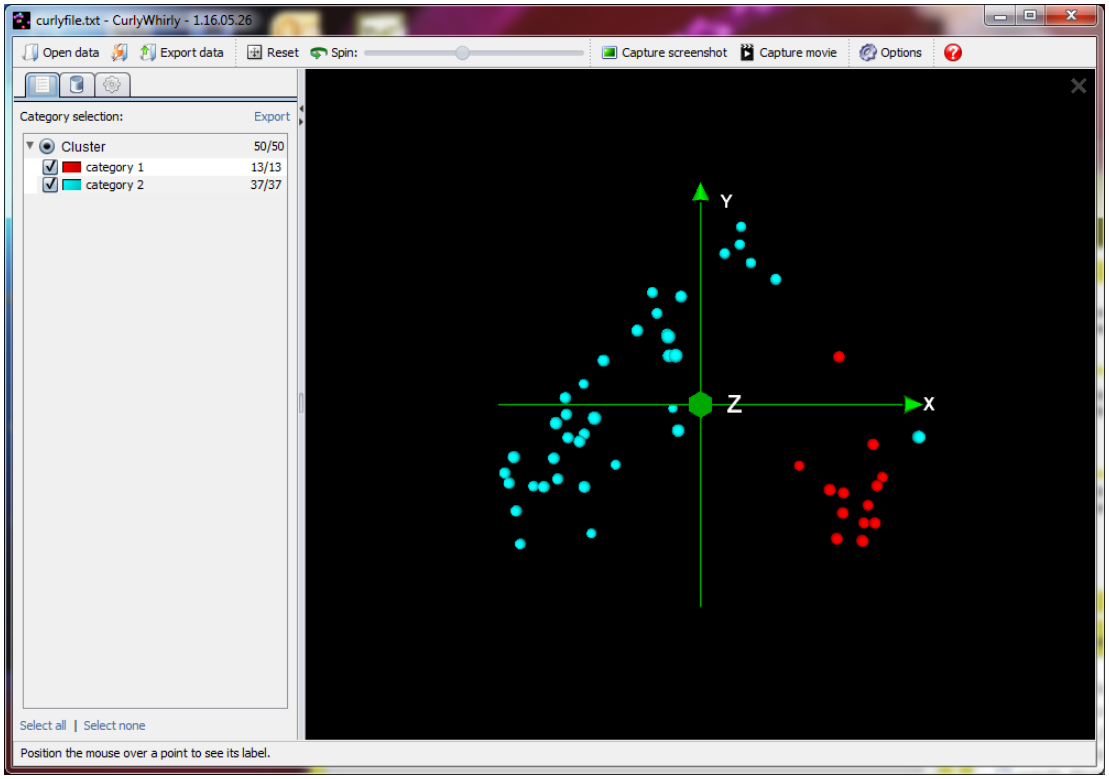
In **Diversity in groups**, the results that you can find are the diversity for each group and within groups, if select **By cluster** reports diversity for these groups, if select **By external group** reports diversity for specific groups you assigned in the file.

In the bottom appears 3 new buttons, **See dendogram in Archaeopteryx**, **See distance matrix in GAP** and **See 3D plot in Curly Whirly**, with each one you can see different graphs in different programs, in which you can edit the graphs.



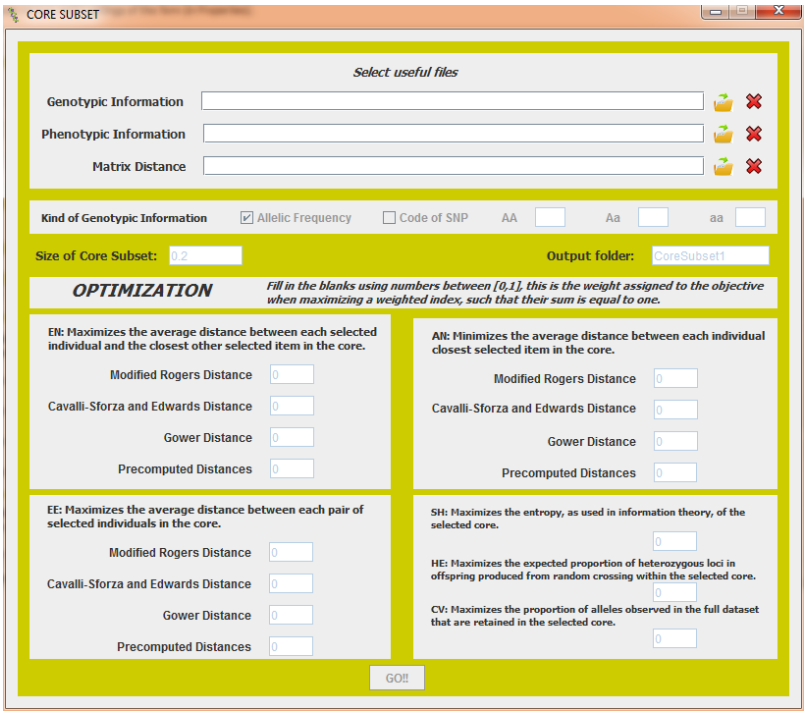
4. BIODIVERSITY ANALYSIS PARAMETERS





5. Core Subset parameters

The **"Core Subset"** obtain sampling core subsets from genetic resources while maintaining as much as possible the genetic diversity of the original collection. Parameters needed to obtain are as follows:



First you need to select the *.csv files containing:

- The **genotypic information**, which should have the same format as if we were to do an analysis of diversity. You need indicated if your data set have a allele frequency (select **Allelic Frequency**) or SNP information (select **Code of SNP**). If you select **Code of SNP** option, you have to write the numbers that identify dominant homozygote (**AA**), heterozygote (**Aa**) and recessive homozygote (**aa**).

5. CORE SUBSET PARAMETERS

	A	B	C	D	E	F	G	H	I	J	K	L
1	mark	g669579	g669039	g659444	g660128	g553924	g660090	g650611	g654182	g660981	g553426	g650375
2	1	0.666667	NA	0.571429	NA	1	0	0.777778	1	NA	0	NA
3	2	NA	1	1	1	NA	1	1	1	1	NA	1
4	3	1	1	0.52	0.5	0.8	NA	0.7	0	NA	1	0.263158
5	4	1	NA	0.5	1	NA	NA	0.939394	0.176471	NA	0.666667	0.428571
6	5	1	NA	1	NA	1	NA	1	1	1	1	1
7	6	0.666667	0	0.714286	1	0.333333	NA	0.363636	1	NA	1	0.583333
8	7	1	NA	0	1	1	NA	0	NA	NA	1	0
9	8	1	NA	1	1	1	1	1	NA	1	1	1
10	9	NA	1	0.6	0.217391	0.708333	NA	0.833333	1	NA	0.846154	1
11	10	1	1	1	1	1	NA	1	1	1	1	1
12	11	1	NA	0.25	0.375	0.333333	0.896552	0.444444	NA	1	NA	0
13	12	0.076923	1	0.333333	0.8	0.5	0	0.9	1	0	0.875	0
14	13	0.571429	NA	0.666667	0.5	1	0	0.333333	1	0.333333	0.75	0.857143
15	14	1	NA	1	1	0.923077	1	1	1	1	1	1

- The **phenotypic information**, the format is: in the first column ”ID” is listed 1 to the number of genotypes in the second ”NAME” column is the identifier name genotype and the following columns such phenotypic information as we have.

	A	B	C	D	E	F
1	ID	NAME	GY	PHT	EHT	AD
2	1	g669579	9.0185	252.9729	147.4754	58.6003
3	2	g669039	7.8369	253.7993	140.2147	58.4904
4	3	g659444	9.1269	246.6514	141.4851	59.4014
5	4	g660128	10.1272	257.9942	152.9806	61.597
6	5	g553924	7.7505	230.2895	134.2098	61.1845
7	6	g660090	9.2025	249.6351	138.0064	60.6466
8	7	g650611	8.5484	247.3041	151.4287	60.7849
9	8	g654182	7.7515	237.277	134.5529	57.6743
10	9	g660981	9.6193	252.0673	146.331	59.1013
11	10	g553426	10.8189	248.6024	147.7987	59.5835
12	11	g650375	6.9872	241.7385	128.9937	61.1103
13	12	g623127	7.9416	247.1738	137.7584	59.4763
14	13	g650594	9.1361	244.5358	139.7396	59.3819
15	14	g607289	10.8519	238.3231	134.7169	60.9595

- A precalculated **matrix distance**, which in the first column ”ID” is listed 1 to the number of existing genotypes, in the second column ”NAME” is the name of ID genotype and the following columns the distance matrix was formed which it is square and symmetrical and those columns whose titles match the names of the ”ID” column.

  The format of the distance matrix that makes this software, is ready for use in the Core Subset.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ID	NAME	1	2	3	4	5	6	7	8	9	10	11
2	1	g669579	0	0.49923	0.45553	0.49411	0.47082	0.36025	0.49175	0.47322	0.33464	0.47634	0.44676
3	2	g669039	0.49923	0	0.42714	0.47224	0.46401	0.48837	0.49146	0.49059	0.44182	0.46646	0.46337
4	3	g659444	0.45553	0.42714	0	0.41451	0.33729	0.46087	0.3386	0.47481	0.4515	0.3467	0.4352
5	4	g660128	0.49411	0.47224	0.41451	0	0.39667	0.4673	0.42047	0.47984	0.46228	0.41574	0.44434
6	5	g553924	0.47082	0.46401	0.33729	0.39667	0	0.46397	0.38216	0.45048	0.45831	0.39268	0.43321
7	6	g660090	0.36025	0.48837	0.46087	0.4673	0.46397	0	0.49417	0.46183	0.3348	0.46071	0.4486
8	7	g650611	0.49175	0.49146	0.3386	0.42047	0.38216	0.49417	0	0.50179	0.50552	0.39775	0.46167
9	8	g654182	0.47322	0.49059	0.47481	0.47984	0.45048	0.46183	0.50179	0	0.44825	0.499	0.32917
10	9	g660981	0.33464	0.44182	0.4515	0.46228	0.45831	0.3348	0.50552	0.44825	0	0.47631	0.42466
11	10	g553426	0.47634	0.46646	0.3467	0.41574	0.39268	0.46071	0.39775	0.499	0.47631	0	0.44109
12	11	g650375	0.44676	0.46337	0.4352	0.44434	0.43321	0.4486	0.46167	0.32917	0.42466	0.44109	0
13	12	g623127	0.48397	0.46887	0.28238	0.40165	0.3472	0.48168	0.33048	0.47224	0.48183	0.36489	0.44415
14	13	g650594	0.43172	0.48595	0.39758	0.44952	0.44231	0.44081	0.41548	0.48904	0.45126	0.3976	0.43414
15	14	g607289	0.49468	0.44739	0.27659	0.4178	0.3613	0.48481	0.34845	0.49683	0.47558	0.37783	0.47703
16	15	g668799	0.31149	0.50029	0.43788	0.49134	0.45601	0.32944	0.4902	0.48743	0.33731	0.46113	0.44176

In the field **Size of Core Subset**, you must type numbers between [0,1] for indicate the percent of size for the new subset. If larger than one the value is used as the absolute core size after rounding.

In the field **Output folder**, you must type the name of the output folder where results will be saved; it will be created inside the Output_BIO-R folder. You can change the name to separate outputs of different data sets. Is necessary to change the name of the output folder for each analysis.

In the **OPTIMIZATION** section you need to fill in the blanks using numbers between [0,1], this is the weight assigned to the objective when maximizing a weighted index, such that their sum is equal to one.

A full explanation can be found by clicking HERE

6. Procedures

6.1. Percent of polymorphic loci

A gene is defined as polymorphic if the frequency one of its alleles is less than or equal to 0.95 or 0.99

$$P_j = q \leq 0.95 \quad \text{o} \quad P_j = q \leq 0.99$$

where P_j is the polymorphic rate and q is the frequency allele. This measure provides the criteria to determine whether a gene has variation.

From here we can calculate the proportion of polymorphic loci:

$$P = \frac{n_{P_j}}{n_{total}}$$

where n_{P_j} is the number of polymorphic loci and n_{total} is the total number of loci. Expresses the percentage of loci in a population variables.

6.2. Number of effective allele

The number of alleles that may be present in a population

$$A_e = \sum_{l=1}^{L} \frac{1}{1 - h_l} = \sum_{l=1}^{L} \frac{1}{\sum p_i^2}$$

Where p_i is the frequency in the ith allele in one locus and $h_l = 1 - \sum p_i^2$ is the heterozygosity in a locus. This measure of diversity can provide useful information for establishing collection strategies. For example, we estimate the number effective alleles in a sample. Then the check in a or different shows throughout the collection. If the figure obtained the second time is less than the first, this could mean that our strategy collection needs revision.

6.3. Expected Heterozygosity

Genetic diversity of Nei is the probability that, in a single locus, any pair of alleles, chosen at random population, different from each other. It can be calculated expected heterozygosity, which is the average of all loci is a estimate the degree of genetic variability in the population

$$H_e = \frac{1}{L} \sum_{j=1}^{L} h_j$$

where h_j is the heterozygosity per locus and L is the total number of loci. The values for H_e are between 0 and 1 and a minimum of 30 loci to be analyzed in 20 individuals per population, to reduce the risk of statistical bias.

6.4. Shannon's Index

Shannon's index accounts for both abundance and evenness of the species present.

$$SH = -\sum_{a=1}^{A} \hat{p}_a log(\hat{p}_a)$$

where $\hat{p}_a$ is the estimated frequency of the allele a on the whole sample and A is the total number of alleles in the sample.

6.5. Diversity within groups

$$H_{sl} = 1 - \sum p_{si}^2$$

where p_{si} is the frequency in the ith allele in one locus in the sth subpopulation and if you calculate the mean of this you obtain:

$$H_s = \frac{1}{L} \sum_{j=1}^{L} H_{slj}$$

6.6. Diversity among groups (Wright's statistics)

$$F_{ST} = \frac{D_{ST}}{H_e}$$

where $D_{ST} = H_e - H_s$ is the diversity among individuals within subpopulation and H_e is the expected heterozygosity, so, F_{ST} measure the degree of genetic differentiation among populations, depending of allele frequencies.

Then we say that if F_{ST} is 0-0.05 the genetic differentiation is small.

Then we say that if F_{ST} is 0.05-0.15 the genetic differentiation is middle.

Then we say that if F_{ST} is 0.15-0.25 the genetic differentiation is big.

Then we say that if F_{ST} is >0.25 the genetic differentiation is very big.

6.7. Genetic distance

6.7.1. Modified Rogers distance

$$MR_{xy} = \sqrt{\frac{\sum_{l=1}^{L}\sum_{a=1}^{n_l}(p_{lax} - p_{lay})^2}{2L}}$$

where p_{lax} is the estimated frequency of the allele a, within the locus l, at the genotype x; L the number of loci, and n_l the number of alleles within the lth locus.

6.7.2. Nei distance

$$NeiD_{xy} = -ln\left(\frac{\sum_{l=1}^{L}\sum_{a=1}^{n_l} p_{lax}p_{lay}}{\sum_{l=1}^{L}\sum_{a=1}^{n_l} p_{lax}^2 \sum_{l=1}^{L}\sum_{a=1}^{n_l} p_{lay}^2}\right)$$

where p_{lax} is the estimated frequency of the allele a, within the locus l, at the genotype x; L the number of loci, and n_l the number of alleles within the lth locus.

6.8. Specificity of a marker in each allele

The so-defined mutual information equals the average allele specificity, defined in this context as the information gained about an accession's identity, by the random extraction and identification of a the allele.

$$S_i = \sum_{j=1}^{N} \frac{p_{ij}}{Np_i} log_2\left(\frac{p_{ij}}{p_i}\right)$$

where:

N: is the total number of accessions.

p_{ij}: the allele frequency of $i-th$ allele within the accession j.

$p_i = \frac{1}{N}\sum_{j=1}^{N} p_{ij}$: the average frequency of the $i-th$ allele across accessions.

6.9. Rarity of an accession

The rarity of an accession is defined as the average specificity of the alleles it contains.

$$R_j = \sum_{i=1}^{k} p_{ij} S_i$$

where:

k: is the total number of alleles, in our case always is 2.

6.10. Multidimensional scaling analysis (MDS)

After calculating the distance matrix proceeds to represent the distances among the objects in a parsimonious (and visual) way (i.e., a lower k-dimensional space). The goal of an MDS analysis is to find a spatial configuration of objects when all that is known is some measure of their general (dis)similarity. The spatial configuration should provide some insight into how the subject(s) evaluate the stimuli in terms of a (small) number of potentially unknown dimensions.

Classical MDS algorithms typically involve some linear algebra. The classical MDS algorithm rests on the fact that the coordinate matrix X can be derived by eigenvalue decomposition from the scalar product matrix $B = XX'$.

The problem of constructing B from the proximity matrix P is solved by multiplying the squared proximities with the matrix $J = I - n^{-1}11'$. This procedure is called double centering. The following steps summarize the algorithm of classical MDS:

- Set up the matrix of squared proximities $P^{(2)}$.

- Apply the double centering: $B = \frac{-1}{2} J P^{(2)} J$ using the matrix $J = I - n^{-1}11'$, where n is the number of objects.
- Extract the m largest positive eigenvalues $\lambda_1 \cdots \lambda_m$ of B and the corresponding m eigenvectors $e_1 \cdots e_m$.
- A $m-$dimensional spatial configuration of the n objects is derived from the coordinate matrix $X = E_m \Lambda_m^{1/2}$, where E_m is the matrix of m eigenvectors and Λ_m is the diagonal matrix of m eigenvalues of B, respectively.

6.11. Cluster Analysis

The best numerical classification strategy is the one that produces the most compact and well separated groups, i.e.,minimum variability within each group and maximum variability among groups.

6.11.1. Ward's Method

In Ward's minimum variance method, the distance between two clusters is the ANOVA sum of squares between the two clusters added up over all the variables. At each generation, the within-cluster sum of squares is minimized over all partitions obtainable by merging two clusters from the previous generation. The sums of squares are easier to interpret when they are divided by the total sum of squares to give the proportions of variance (squared semipartial correlations).

Ward's method joins clusters to maximize the likelihood at each level of the hierarchy under the assumptions of multivariate normal mixtures, spherical covariance matrices, and equal sampling probabilities.

Ward's method tends to join clusters with a small number of observations and is strongly biased toward producing clusters with approximately the same number of observations. It is also very sensitive to outliers.

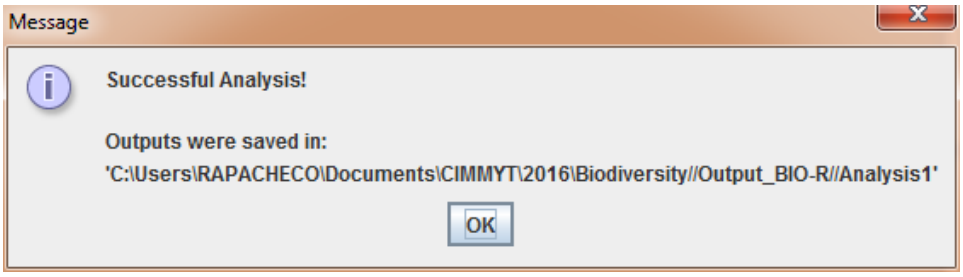
Distance for Ward's method is:

$$D_{KL} = \frac{\| \bar{x}_K - \bar{x}_L \|^2}{\frac{1}{N_K} + \frac{1}{N_L}}$$

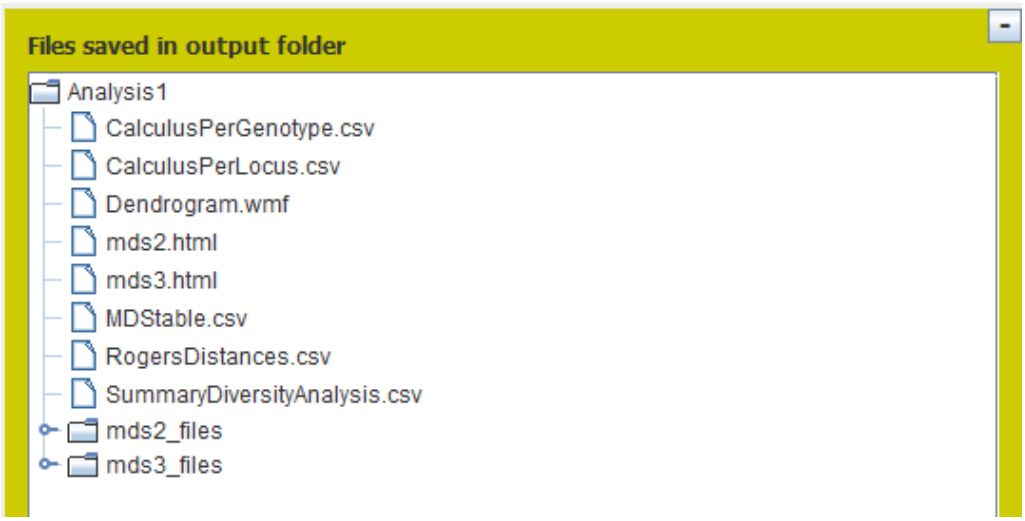
where $\bar{x}_K$ and $\bar{x}_L$ is the mean vector for the Kth and Lth cluster respectively; N_K and N_L is the number of observation in the Kth and Lth cluster respectively and $\| x \|$ is the square root of the sum of the squares of the elements of x.

7. Results

If everything it's good, a window like below appears:

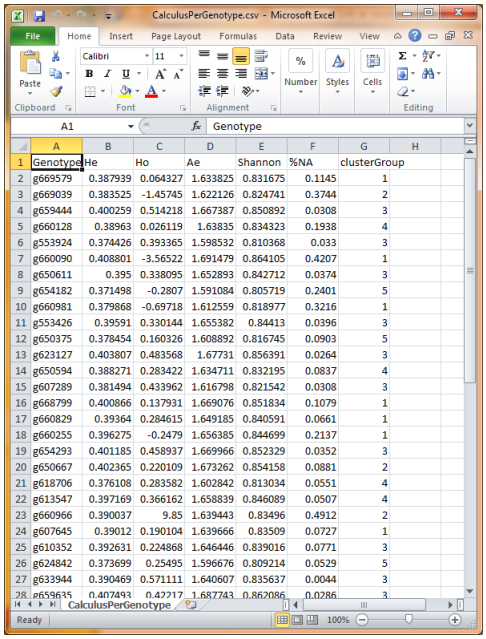


And you can see and open the results files from this window:



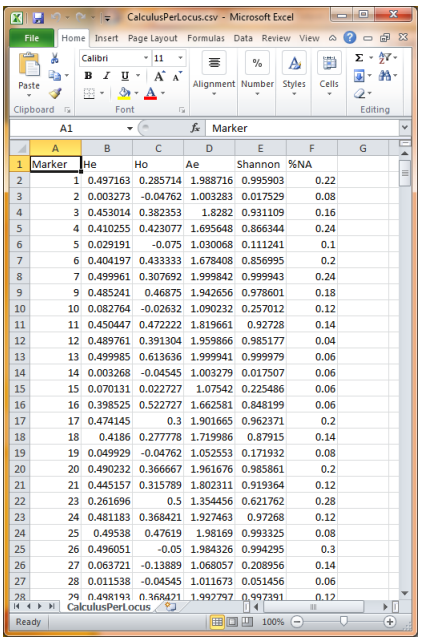
In file "CalculusPerGenotype.csv" you can find expected heterozygosity(He), observed heterozygosity (Ho), number of effective allele (Ae), Shannon Index (Shannon), Proportion of

missing values (%NA) and number of the cluster group (clusterGroup) for each genotype.



Genotype	He	Ho	Ae	Shannon	%NA	clusterGroup
g669579	0.387939	0.064327	1.633825	0.831675	0.1145	1
g669039	0.383525	-1.45745	1.622126	0.824741	0.3744	2
g659444	0.400259	0.514218	1.667387	0.850892	0.0308	3
g660128	0.38963	0.026119	1.63835	0.834323	0.1938	4
g553924	0.374426	0.393365	1.598532	0.810368	0.033	3
g660090	0.408801	-3.56522	1.691479	0.864105	0.4207	1
g650611	0.395	0.338095	1.652893	0.842712	0.0374	3
g654182	0.371498	-0.2807	1.591084	0.805719	0.2401	5
g660981	0.379868	-0.69718	1.612559	0.818977	0.3216	1
g553426	0.39591	0.330144	1.655382	0.84413	0.0396	3
g650375	0.378454	0.160326	1.608892	0.816745	0.0903	5
g623127	0.403807	0.483568	1.67731	0.856391	0.0264	3
g650594	0.388271	0.283422	1.634711	0.832195	0.0837	4
g607289	0.381494	0.433962	1.616798	0.821542	0.0308	3
g668799	0.400866	0.137931	1.669076	0.851834	0.1079	1
g660829	0.39364	0.284615	1.649185	0.840591	0.0661	1
g660255	0.396275	-0.2479	1.656385	0.844699	0.2137	1
g654293	0.401185	0.458937	1.669966	0.852329	0.0352	3
g650667	0.402365	0.220109	1.673262	0.854158	0.0881	2
g618706	0.376108	0.283582	1.602842	0.813034	0.0551	4
g613547	0.397169	0.366162	1.658839	0.846089	0.0507	4
g660966	0.390037	9.85	1.639443	0.83496	0.4912	2
g607645	0.39012	0.190104	1.639666	0.83509	0.0727	1
g610352	0.392631	0.224868	1.646446	0.839016	0.0771	3
g624842	0.373699	0.25495	1.596676	0.809214	0.0529	5
g633944	0.390469	0.571111	1.640607	0.835637	0.0044	3
g659635	0.407493	0.42217	1.687743	0.862086	0.0286	3

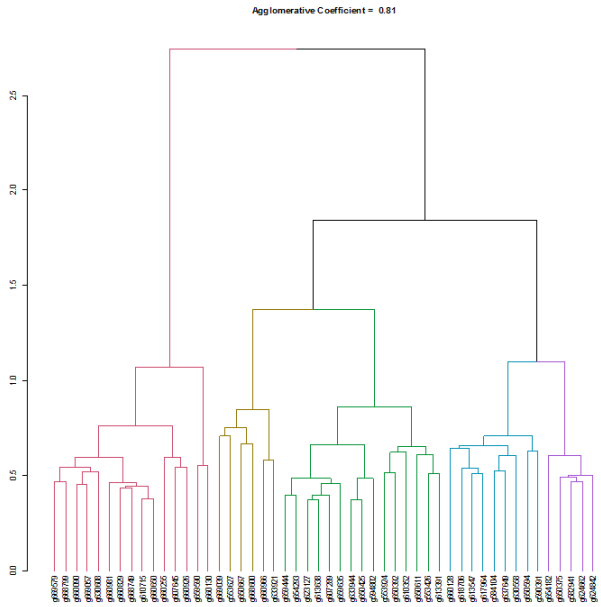
In file ”CalculusPerLocus.csv” you can find expected heterozygosity(He), observed heterozygosity (Ho), number of effective allele (Ae), Shannon Index (Shannon) and proportion of missing values (%NA) for each Marker.



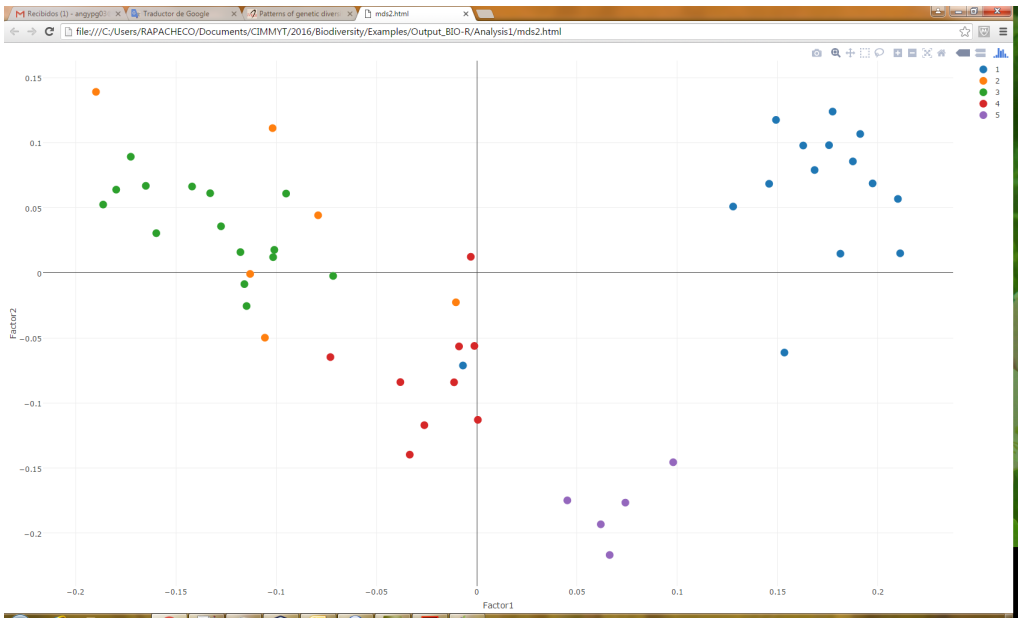
Marker	He	Ho	Ae	Shannon	%NA
1	0.497163	0.285714	1.988716	0.995903	0.22
2	0.003273	-0.04762	1.003283	0.017529	0.08
3	0.453014	0.382353	1.8282	0.931109	0.16
4	0.410255	0.423077	1.695648	0.866344	0.24
5	0.029191	-0.075	1.030068	0.111241	0.1
6	0.404197	0.433333	1.678408	0.856995	0.2
7	0.499961	0.307692	1.999842	0.999943	0.24
9	0.485241	0.46875	1.942656	0.978601	0.18
10	0.082764	-0.02632	1.090232	0.257012	0.12
11	0.450447	0.472222	1.819661	0.92728	0.14
12	0.489761	0.391304	1.959866	0.985177	0.04
13	0.499985	0.613636	1.999941	0.999979	0.06
14	0.003268	-0.04545	1.003279	0.017507	0.06
15	0.070131	0.022727	1.07542	0.225486	0.06
16	0.398525	0.522727	1.662581	0.848199	0.06
17	0.474145	0.3	1.901665	0.962371	0.2
18	0.4186	0.277778	1.719986	0.87915	0.14
19	0.049929	-0.04762	1.052553	0.171932	0.08
20	0.490232	0.366667	1.961676	0.985861	0.2
21	0.445157	0.315789	1.802311	0.919364	0.12
22	0.261696	0.5	1.354456	0.621762	0.28
23	0.481183	0.368421	1.927463	0.97268	0.12
24	0.49538	0.47619	1.98169	0.993325	0.08
25	0.496051	-0.05	1.984326	0.994295	0.3
26	0.063721	-0.13889	1.068057	0.208956	0.14
27	0.011538	-0.04545	1.011673	0.051456	0.06
28	0.498193	0.368421	1.992797	0.997391	0.12

7. RESULTS

The ”Dendogram.wmf” file is a dendogram result of the cluster analysis.

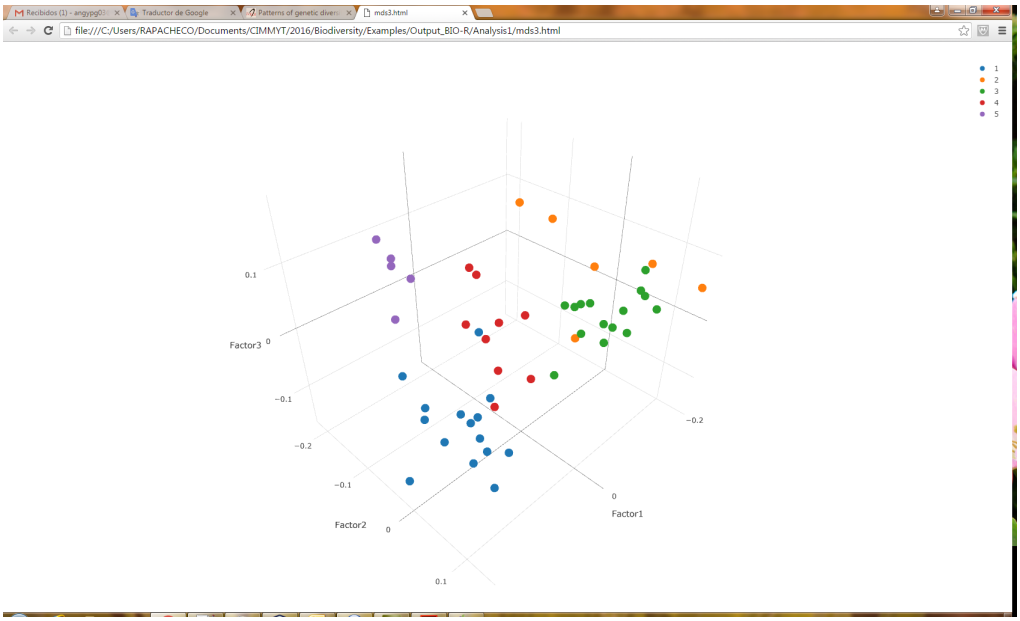


The ”mds2.html” and ”mds3.html” are multidimensional scaling graph in 2D and 3D respectively, these graphics are interactively, for that reason the format is *.html and need the folders ”mds2_files” and ”mds3_files” like a complement.

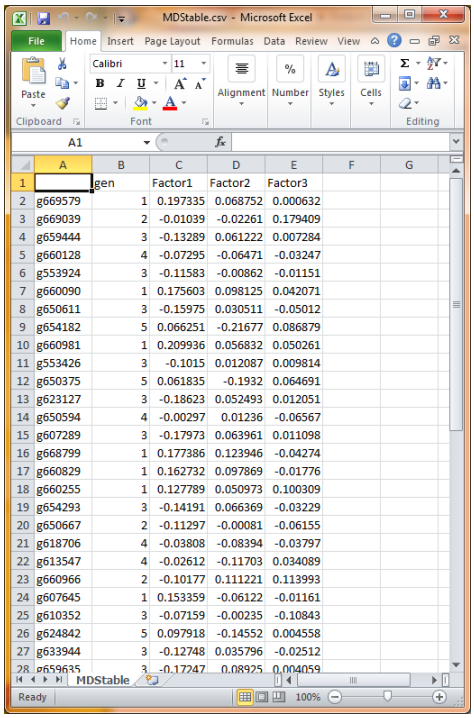


In these graphs you can make them disappear or appear groups were formed just by clicking

on the legend on the right side; also when approaching your mouse at some point, immediately the name of that point and coordinate axis values appears.



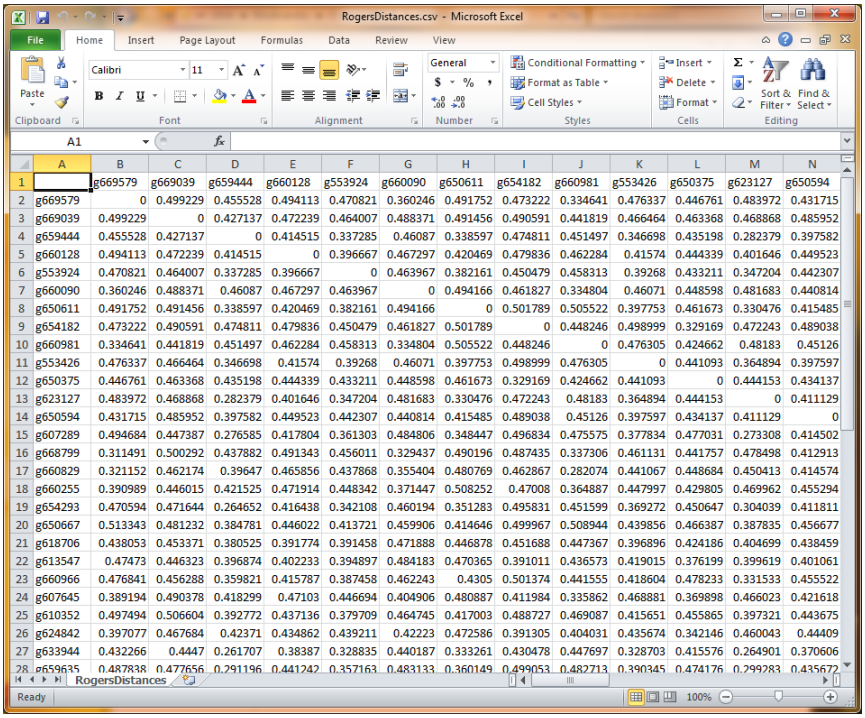
The data in ”MDStable.csv” are the values that were used to make the graphics.



	A	B	C	D	E
1		gen	Factor1	Factor2	Factor3
2	g669579	1	0.197335	0.068752	0.000632
3	g669039	2	-0.01039	-0.02261	0.179409
4	g659444	3	-0.13289	0.061222	0.007284
5	g660128	4	-0.07295	-0.06471	-0.03247
6	g553924	3	-0.11583	-0.00862	-0.01151
7	g660090	1	0.175603	0.098125	0.042071
8	g650611	3	-0.15975	0.030511	-0.05012
9	g654182	5	0.066251	-0.21677	0.086879
10	g660981	1	0.209936	0.056832	0.050261
11	g553426	3	-0.1015	0.012087	0.009814
12	g650375	5	0.061835	-0.1932	0.064691
13	g623127	3	-0.18623	0.052493	0.012051
14	g650594	4	-0.00297	0.01236	-0.06567
15	g607289	3	-0.17973	0.063961	0.011098
16	g668799	1	0.177386	0.123946	-0.04274
17	g660829	1	0.162732	0.097869	-0.01776
18	g660255	1	0.127789	0.050973	0.100309
19	g654293	3	-0.14191	0.066369	-0.03229
20	g650667	2	-0.11297	-0.00081	-0.06155
21	g618706	4	-0.03808	-0.08394	-0.03797
22	g613547	4	-0.02612	-0.11703	0.034089
23	g660966	2	-0.10177	0.111221	0.113993
24	g607645	1	0.153359	-0.06122	-0.01161
25	g610352	3	-0.07159	-0.00235	-0.10843
26	g624842	5	0.097918	-0.14552	0.004558
27	g633944	3	-0.12748	0.035796	-0.02512
28	g659635	3	-0.17247	0.08925	0.004059

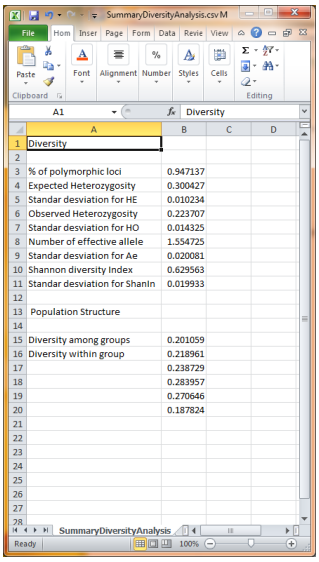
7. RESULTS

In file ”RogersDistances.csv” is saved the square matrix of Rogers’s distance.



	g669579	g669039	g659444	g660128	g553924	g660090	g650611	g654182	g660981	g553426	g650375	g623127	g650594
g669579	0	0.499229	0.455528	0.494113	0.470821	0.360246	0.491752	0.473222	0.334641	0.476337	0.446761	0.483972	0.431715
g669039	0.499229	0	0.427137	0.472239	0.464007	0.488371	0.491456	0.490591	0.441819	0.466464	0.463368	0.468868	0.485952
g659444	0.455528	0.427137	0	0.414515	0.337285	0.46087	0.338597	0.474811	0.451497	0.346698	0.435198	0.282379	0.397582
g660128	0.494113	0.472239	0.414515	0	0.396667	0.467297	0.420469	0.479836	0.462284	0.41574	0.444339	0.401646	0.449523
g553924	0.470821	0.464007	0.337285	0.396667	0	0.463967	0.382161	0.450479	0.458313	0.39268	0.433211	0.347204	0.442307
g660090	0.360246	0.488371	0.46087	0.467297	0.463967	0	0.494166	0.461827	0.334804	0.46071	0.448598	0.481683	0.440814
g650611	0.491752	0.491456	0.338597	0.420469	0.382161	0.494166	0	0.501789	0.505522	0.397753	0.461673	0.330476	0.415485
g654182	0.473222	0.490591	0.474811	0.479836	0.450479	0.461827	0.501789	0	0.448246	0.498999	0.329169	0.472243	0.489038
g660981	0.334641	0.441819	0.451497	0.462284	0.458313	0.334804	0.505522	0.448246	0	0.476305	0.424662	0.48183	0.45126
g553426	0.476337	0.466464	0.346698	0.41574	0.39268	0.46071	0.397753	0.498999	0.476305	0	0.441093	0.364894	0.397597
g650375	0.446761	0.463368	0.435198	0.444339	0.433211	0.448598	0.461673	0.329169	0.424662	0.441093	0	0.444153	0.434137
g623127	0.483972	0.468868	0.282379	0.401646	0.347204	0.481683	0.330476	0.472243	0.48183	0.364894	0.444153	0	0.411129
g650594	0.431715	0.485952	0.397582	0.449523	0.442307	0.440814	0.415485	0.489038	0.45126	0.397597	0.434137	0.411129	0
g607289	0.494684	0.447387	0.276585	0.417804	0.361303	0.484806	0.348447	0.496834	0.475575	0.377834	0.477031	0.273308	0.414502
g668799	0.311491	0.500292	0.437882	0.491343	0.456011	0.329437	0.490196	0.487435	0.337306	0.461131	0.441757	0.478498	0.412913
g660829	0.321152	0.462174	0.39647	0.465856	0.437868	0.355404	0.480769	0.462867	0.282074	0.441067	0.448684	0.450413	0.414574
g660255	0.390989	0.446015	0.421525	0.471914	0.448342	0.371447	0.508252	0.47008	0.364887	0.447997	0.429805	0.469962	0.455294
g654293	0.470594	0.471644	0.264652	0.416438	0.342108	0.460194	0.351283	0.495831	0.451599	0.369272	0.450647	0.304039	0.411811
g650667	0.513343	0.481232	0.384781	0.446022	0.413721	0.459906	0.414646	0.499967	0.508944	0.439856	0.466387	0.387835	0.456677
g618706	0.438053	0.453371	0.380525	0.391774	0.391458	0.471888	0.446878	0.451688	0.447367	0.396896	0.424186	0.404699	0.438459
g613547	0.47473	0.446323	0.396874	0.402233	0.394897	0.484183	0.470365	0.391011	0.436573	0.419015	0.376199	0.399619	0.401061
g660966	0.476841	0.456288	0.359821	0.415787	0.387458	0.462243	0.4305	0.501374	0.441555	0.418604	0.478233	0.331533	0.455522
g607645	0.389194	0.490378	0.418299	0.47103	0.446694	0.404906	0.480887	0.411984	0.335862	0.468881	0.369898	0.466023	0.421618
g610352	0.497494	0.506604	0.392772	0.437136	0.379709	0.464745	0.417003	0.488727	0.469087	0.415651	0.455865	0.397321	0.443675
g624842	0.397077	0.467684	0.42371	0.434862	0.439211	0.42223	0.472586	0.391305	0.404031	0.435674	0.342146	0.460043	0.44409
g633944	0.432266	0.4447	0.261707	0.38387	0.328835	0.440187	0.333261	0.430478	0.447697	0.328703	0.415576	0.264901	0.370606
g659635	0.487838	0.477656	0.291196	0.441242	0.357163	0.483133	0.360149	0.499053	0.482713	0.390345	0.474176	0.299283	0.435672

Finally in file ”SummaryDiversityAnalysis.csv” you can find the calculus for the average of percent of polymorfic loci, expected heterozygosity(He), observed heterozygosity (Ho), number of effective allele (Ae), Shannon Index (Shannon) and their standar desviation for each one, diversity among groups and diversity within group are parameters of population structure.

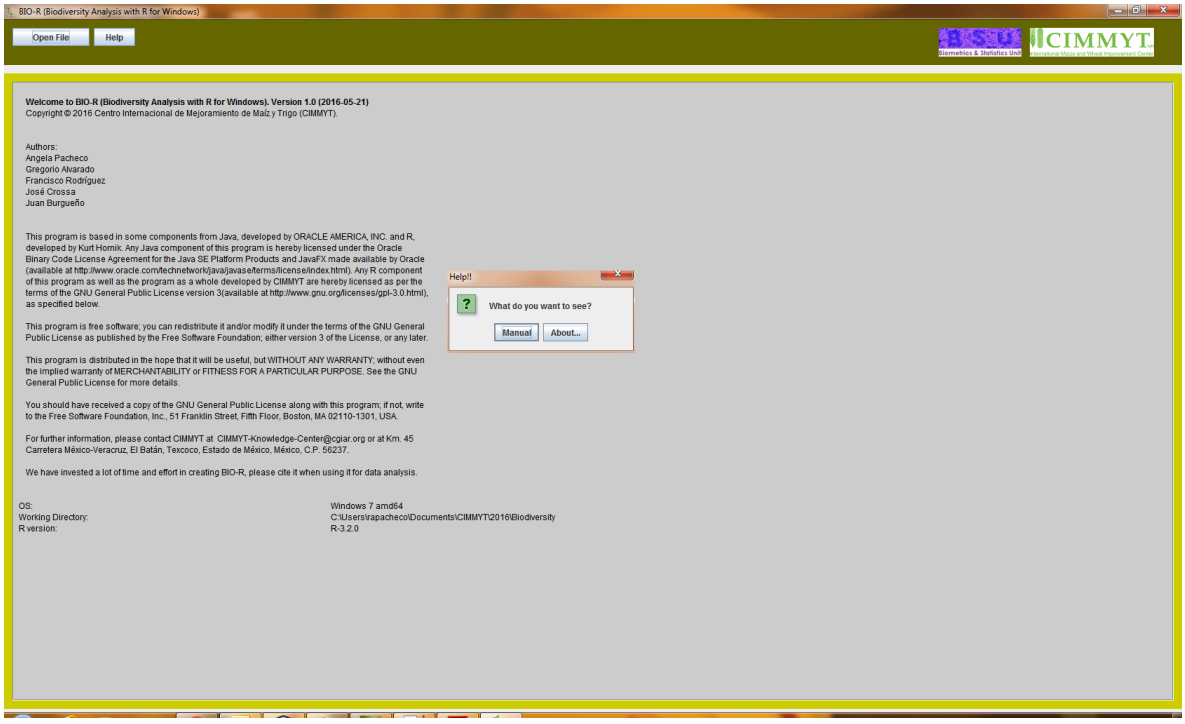


Diversity	
% of polymorphic loci	0.947137
Expected Heterozygosity	0.300427
Standar desviation for HE	0.010234
Observed Heterozygosity	0.223707
Standar desviation for HO	0.014325
Number of effective allele	1.554725
Standar desviation for Ae	0.020081
Shannon diversity Index	0.629563
Standar desviation for ShanIn	0.019933
Population Structure	
Diversity among groups	0.201059
Diversity within group	0.218961
	0.238729
	0.283957
	0.270646
	0.187824

8. Help button

In the help button you can find two option:

- Manual : Is a help for you whenever you want.
- About : If you require cite BIO-R in your work you can find how cite here,if you need more information about license you can find the GNU and Oracle licenses here.



9. REFERENCES

de Vicente, M.C., Lopez, C. y Fulton, T. (eds.). 2004. Analisis de la Diversidad Genetica Utilizando Datos de Marcadores Moleculares: Modulo de Aprendizaje. Instituto Internacional de Recursos Fitogeneticos (IPGRI), Roma, Italia.