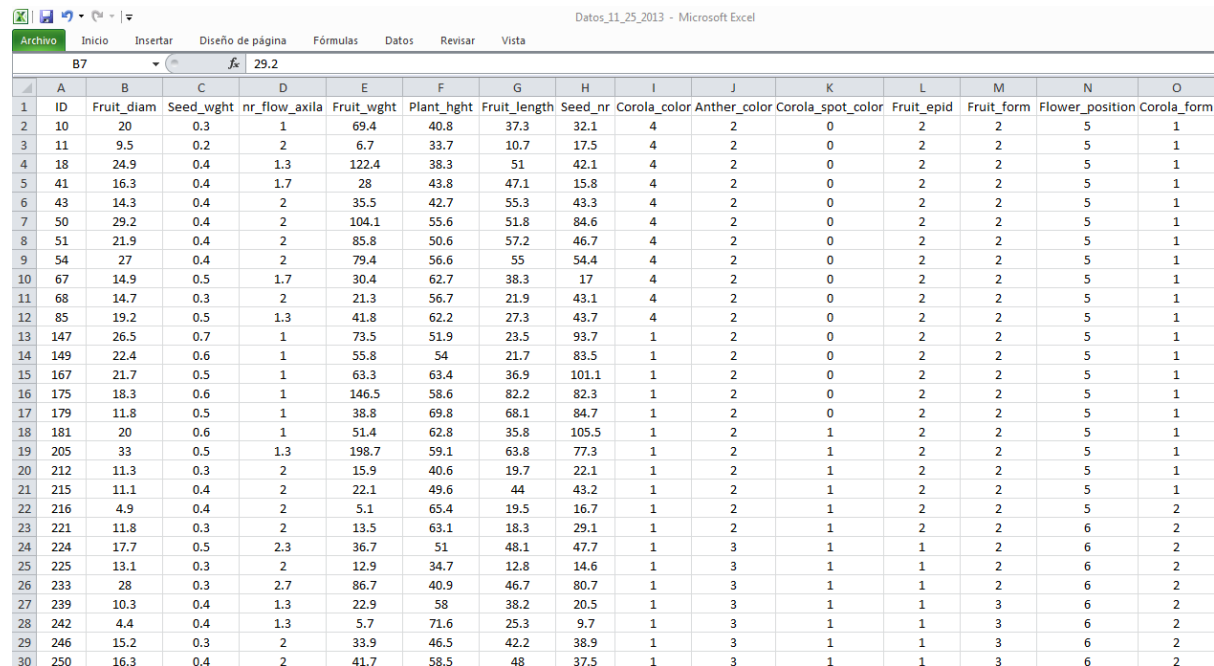


Guide to use Explora

1) Data preparation and running the Script

The main purpose of Explora is to select a set of promising accessions from large genebank collections that meet the interests of the user. Explora allows 1) selection of germplasm from big characterization and/or evaluation datasets; 2) to consider more than one trait and trade-offs between different traits of interests; and 3) maximizing diversity for specific traits of interests.

Data file preparation



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	ID	Fruit_diam	Seed_wght	nr_flow_axila	Fruit_wght	Plant_hght	Fruit_length	Seed_nr	Corola_color	Anther_color	Corola_spot_color	Fruit_epid	Fruit_form	Flower_position	Corola_form
2	10	20	0.3	1	69.4	40.8	37.3	32.1	4	2	0	2	2	5	1
3	11	9.5	0.2	2	6.7	33.7	10.7	17.5	4	2	0	2	2	5	1
4	18	24.9	0.4	1.3	122.4	38.3	51	42.1	4	2	0	2	2	5	1
5	41	16.3	0.4	1.7	28	43.8	47.1	15.8	4	2	0	2	2	5	1
6	43	14.3	0.4	2	35.5	42.7	55.3	43.3	4	2	0	2	2	5	1
7	50	29.2	0.4	2	104.1	55.6	51.8	84.6	4	2	0	2	2	5	1
8	51	21.9	0.4	2	85.8	50.6	57.2	46.7	4	2	0	2	2	5	1
9	54	27	0.4	2	79.4	56.6	55	54.4	4	2	0	2	2	5	1
10	67	14.9	0.5	1.7	30.4	62.7	38.3	17	4	2	0	2	2	5	1
11	68	14.7	0.3	2	21.3	56.7	21.9	43.1	4	2	0	2	2	5	1
12	85	19.2	0.5	1.3	41.8	62.2	27.3	43.7	4	2	0	2	2	5	1
13	147	26.5	0.7	1	73.5	51.9	23.5	93.7	1	2	0	2	2	5	1
14	149	22.4	0.6	1	55.8	54	21.7	83.5	1	2	0	2	2	5	1
15	167	21.7	0.5	1	63.3	63.4	36.9	101.1	1	2	0	2	2	5	1
16	175	18.3	0.6	1	146.5	58.6	82.2	82.3	1	2	0	2	2	5	1
17	179	11.8	0.5	1	38.8	69.8	68.1	84.7	1	2	0	2	2	5	1
18	181	20	0.6	1	51.4	62.8	35.8	105.5	1	2	1	2	2	5	1
19	205	33	0.5	1.3	198.7	59.1	63.8	77.3	1	2	1	2	2	5	1
20	212	11.3	0.3	2	15.9	40.6	19.7	22.1	1	2	1	2	2	5	1
21	215	11.1	0.4	2	22.1	49.6	44	43.2	1	2	1	2	2	5	1
22	216	4.9	0.4	2	5.1	65.4	19.5	16.7	1	2	1	2	2	5	2
23	221	11.8	0.3	2	13.5	63.1	18.3	29.1	1	2	1	2	2	6	2
24	224	17.7	0.5	2.3	36.7	51	48.1	47.7	1	3	1	1	2	6	2
25	225	13.1	0.3	2	12.9	34.7	12.8	14.6	1	3	1	1	2	6	2
26	233	28	0.3	2.7	86.7	40.9	46.7	80.7	1	3	1	1	2	6	2
27	239	10.3	0.4	1.3	22.9	58	38.2	20.5	1	3	1	1	3	6	2
28	242	4.4	0.4	1.3	5.7	71.6	25.3	9.7	1	3	1	1	3	6	2
29	246	15.2	0.3	2	33.9	46.5	42.2	38.9	1	3	1	1	3	6	2
30	250	16.3	0.4	2	41.7	58.5	48	37.5	1	3	1	1	3	6	2

Entering characterization and evaluation data: Explora reads csv text file. A file with characterization and evaluation data can be saved as csv file from Excel. List the accession codes in the first column of the spreadsheet as demonstrated in the image above. All continuous variables should be listed in the consecutive columns. Categorical variables should be added in the following columns. Headers go the first line. Please verify that the decimal separator in Excel is a “point” and the thousand separator a “comma”. If this is not the case, this can be changed in Excel options (advanced) that are found in Excel’s file menu.

Run the Explora script in R

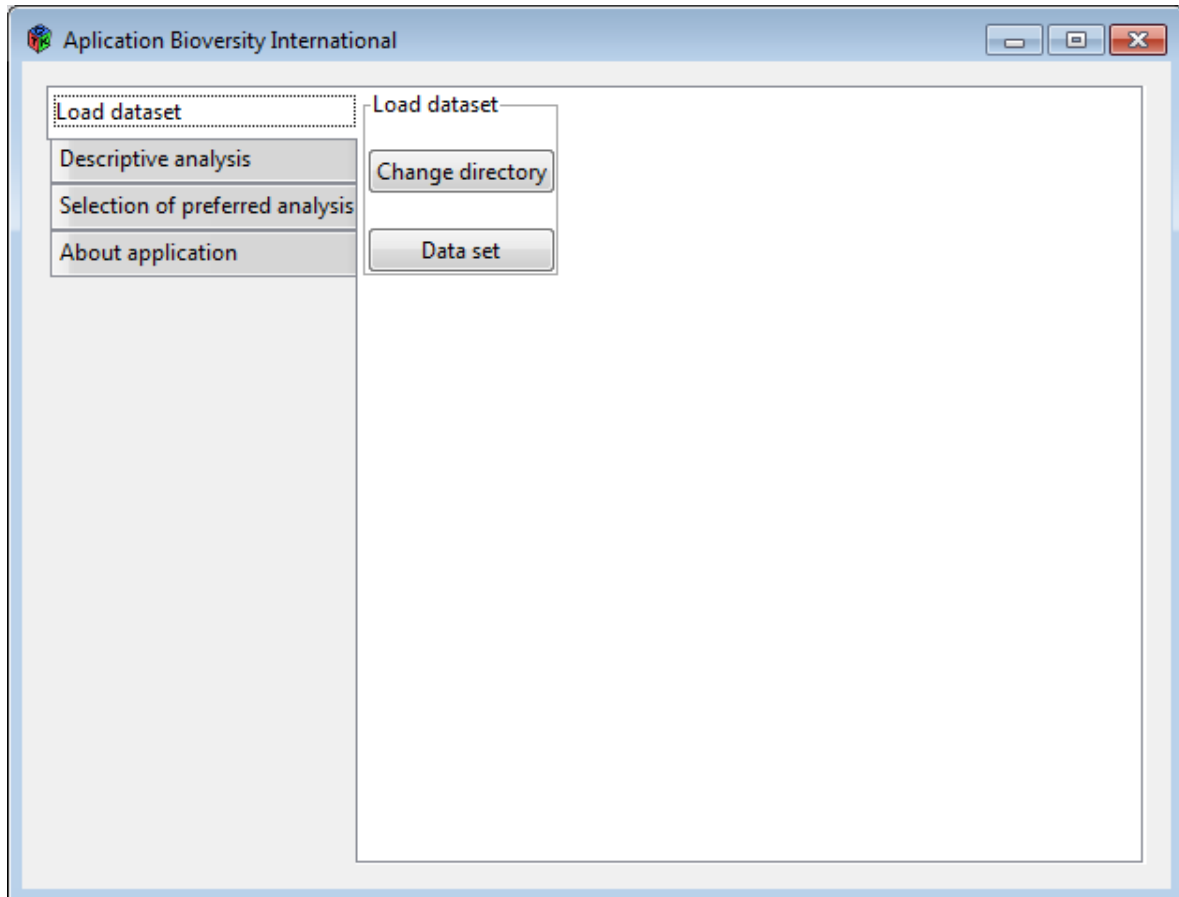
After running the R script, the Explora window can be opened from the taskbar



2) Exercise to run Explora

Use the dataset "Datos_11_25_2013.csv" to run Explora

Window 1: Load dataset



- Select in *Change directory* the folder where your file with characterization and evaluation data is stored
- Select in *Data set* the data file.

Window 2: Descriptive analysis

Aplication Bioversity International

Load dataset
Descriptive analysis
Selection of preferred analysis
About application

Descriptive analysis

Dataset for analyzed

Number of continuos variables:

Number of categorical variables:

Descriptive analysis for continuous variables

Descriptive analysis for nominal variables

Level of correlation: 0.0

Correlation analysis

Number accessions final dataset: 10

Select number accessions

Thresholds analysis:

Select variables

- Confirm the dataset to be analyzed in **Dataset for analyzed**
- Indicate the number of continuous and categorical variables; in this case **7** continuous and **7** categorical variables
- **Optional descriptive analyses:** select descriptive analyses and correlation matrix. These analyses are saved in a csv file in an automatically generated folder named "Results" in the selected directory. The number of correlations can be restricted by indicating the minimum Pearson coefficient in the *Level of correlation*.
- **Select number of promising accessions in final subset**
- **Optional threshold analysis:** Accessions that have undesirable high or low values for specific traits can be excluded of further selection by applying thresholds.

Window 3: Selection of preferred analysis

Aplication Bioversity International

Load dataset
Descriptive analysis
Selection of preferred analysis
About application

Type selection of preferred
Enter the number of solutions: 10000
Select the number of solutions

Optimization analysis:
Select variables

Enter the percentage of solutions (%): 1
Select the percentage of solutions

Enter the number of final solutions for (Maximum variation or Principal components): 10
Select the number of final solutions

Select the type selection of preferred:
Run

- **Select the number of solutions:** Explora generates a number of subsets with random selections of accessions. Out of these subsets the user can select in the following steps (see below) the solution with most promising accessions according to his/her interest.
- **Select variables:** a menu of different objective functions for continuous and categorical variables is presented. The user can determine for each variable of his/ her interest the desired objective function. The user can indicate optionally for each variable a weight of importance for the final selection of the subset that fits best to his/her interests. If there is no preference, please indicate 1 for each variable.

Objective functions for continuous variables

- maximize coefficient of variation for a specific subset;
- maximize average for a specific subset;
- maximize minimum for a specific subset;
- maximize maximum for a specific subset;
- minimize average for a specific subset;
- minimize minimum for a specific subset; or
- minimize maximum for a specific subset.

Objective functions for categorical variables

- Maximize proportion of a determined category of preference in final set of accessions; or
 - Maximize Shannon index.
- **Select the percentage of solutions:** This is the percentage of subsets that correspond best to the objective functions. In the following steps, the user can select his/her optimal solution from these subsets through stepwise clustering (see below). To calculate the best solutions, we standardized the values per objective functions between 0 and 1, and then we calculated for each subset the sum, and selected those with the highest sum value for further analysis to identify the most optimal solution.
- **Select the number of final solutions:** from the percentage of optimal subsets, a number of solutions is selected for a principal component analysis. This analysis is one of the methods that helps the user to find an optimal solution. The optimal subsets with the highest coefficient of variance across the objective functions are selected for this exercise.
- Choose from four methods to find the subset of accessions that most suit your interest.
 - **Maximum variation:** choose the optimal subset of accessions with the highest coefficient of variance (CV) across all variables. The subset that has the highest CV score will be saved in the "Results" folder in a csv file with the details of the corresponding accessions.
 - **Principal component analysis:** graphs of the first three component plots are saved in the "Results" folder. These graphs show the ordination of the subsets in function of the selected objective functions. The user can select manually the subset that he/she finds most interesting. This subset will be saved in the "Results" folder in a csv file with the details of the corresponding accessions.
 - **Weighted Sum Model (WSM):** this value is calculated with the weights indicated for each object function. The subset that has the highest (WSM) score will be saved in the "Results" folder in a csv file with the details of the corresponding accessions
 - **Stepwise clustering** to come to the best subset of accessions. K-means clustering is used to present the user two clusters in each step. For each cluster we present the median values of each objective function and the median weighted sum model value. On the basis of this information the user can decide which cluster to choose. This goes on until the user decides to stop or when a cluster has less than 10 solutions. From this final number of subsets, the one with the highest weighted sum model score is saved in the "Results" folder.