Basic Explora User Manual

Once unzipped locally, Explora may be built and reloaded from within the RStudio tool or, alternatively, from the command line:

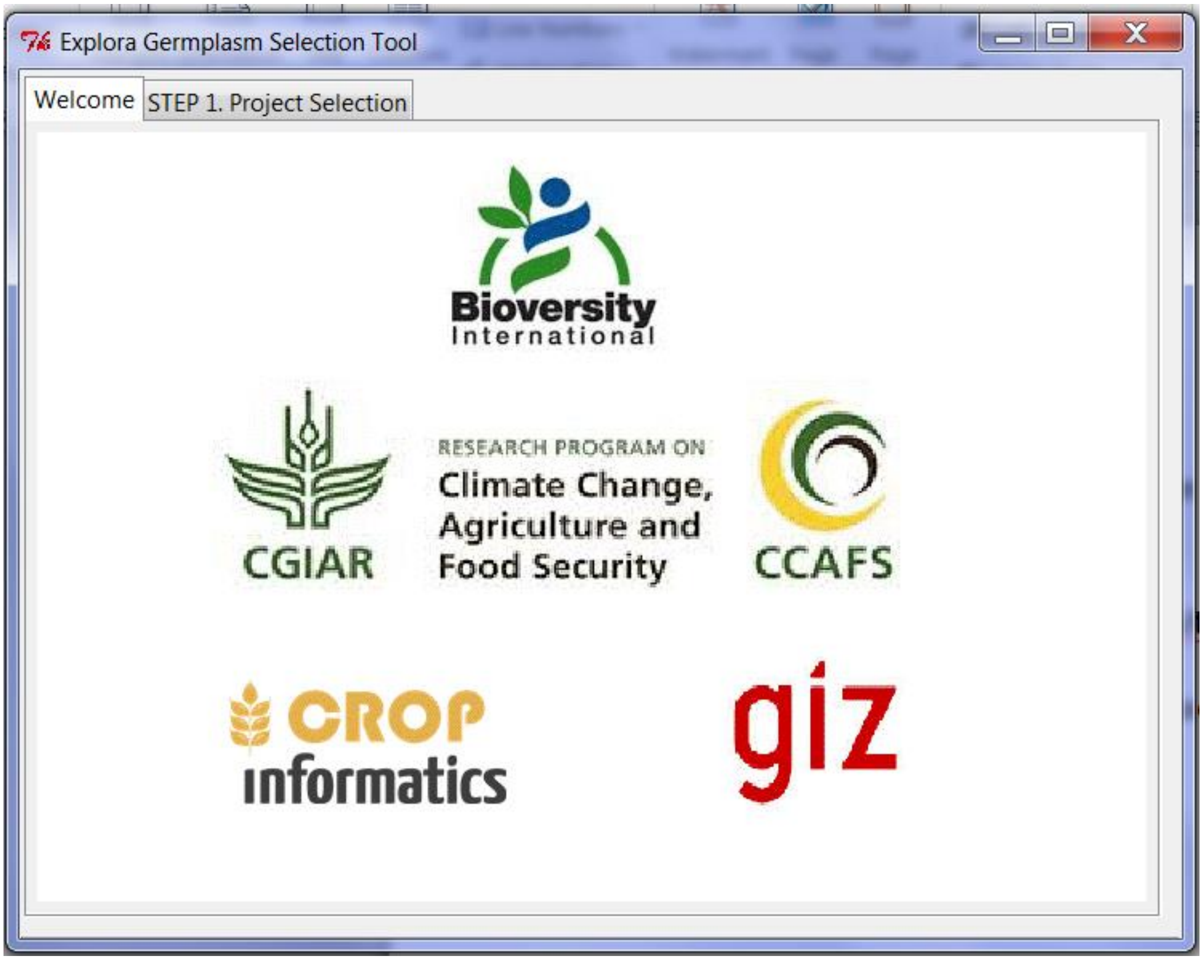
```
Rcmd.exe INSTALL --no-multiarch --with-keep.source explora
```

Once rebuilt, it may be loaded by “Load All” (Ctrl+Shift+L) from within RStudio or run from the R cmd shell:

```
> library(explora)
> devtools::load_all(".")
```

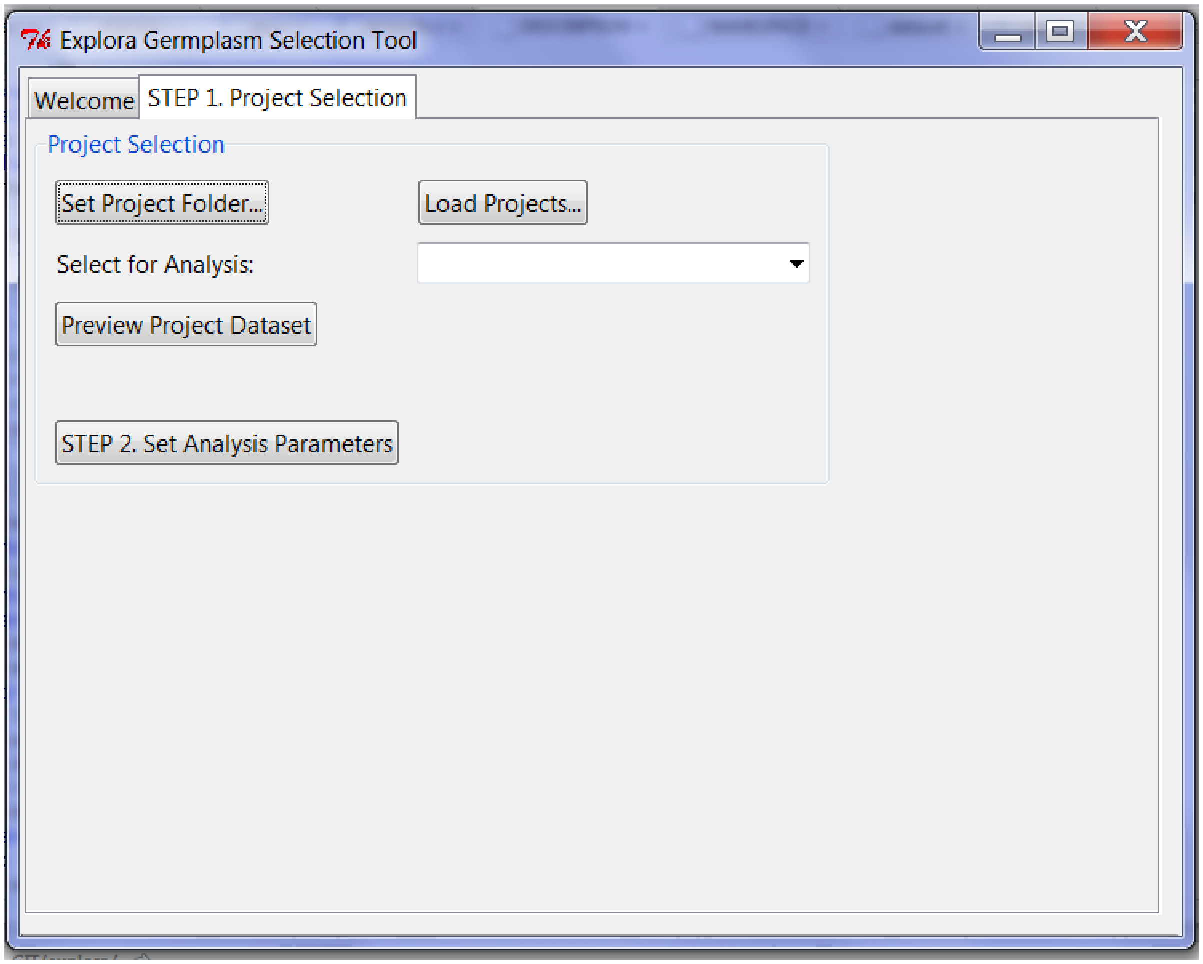
Upon execution, the logo banner tab is seen (Figure 1).

Figure 1) Explora Banner page


Explora Germplasm Selection Tool
Welcome
STEP 1. Project Selection
Bioversity International
CGIAR
RESEARCH PROGRAM ON Climate Change, Agriculture and Food Security
CCAFS
CROP informatics
giz

Clicking on the "*Step 1. Project Selection*" tab exposes the project management view (Figure 2).

Figure 2) Project Selection page



Step 1. Project Selection Tab

Set Project Folder...

Clicking on this button gives a local directory selection dialog box. Pick (or create) the local file directory within which you will place all your analysis project datasets and results.

Load Projects

After selecting a project folder, you can use this button to get another file system dialog box, which you can use to navigate within your computer to find and select CSV data files that are suitable for Explora analysis. In a nutshell, such files should have the following:

1) A header line, labelling each column of data
2) Data consists of accession records, one per line
3) Column 1 should be the one containing a unique accession identifier, one per row

4) The next 2..n columns should be the C = n-1 columns of continuous variable data
5) The next n+1..m columns should contain all the N = m-n columns of nominal variable data

Unfortunately, Explora doesn't yet have a protocol for automatically discriminating between Continuous and Nominal data columns – the user has to ensure that Continuous columns come after the accession identification (column 1) and preceed Nominal columns, then tell Explora (see Step 2. Below) how many of each kind of variable that there are.

Once selected, a given input dataset file is copied into a subdirectory of project, with the given root <filename> of the data file used as the project name ("<filename>.explora"). This subdirectory will contain all the analysis results for that data file, as they are generated. In the process of copying the file over, the first column ("identification") header is relabelled to read 'accession'.

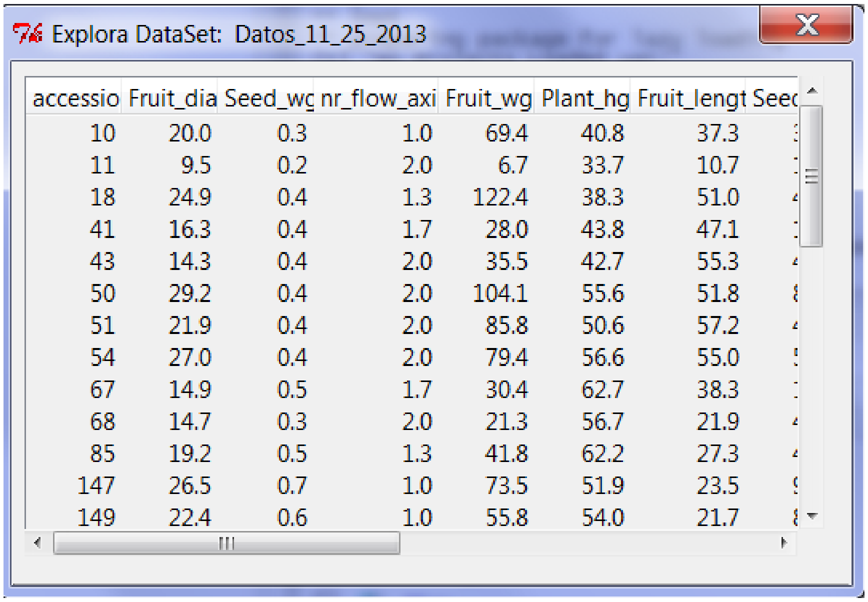
Select for Analysis

After any or all of your target datasets have been loaded into the project space, you select one of them at a time for analysis, using the available drop down menu.

Preview Project Dataset

Clicking on this button gives a modeless window displaying the data file which was loaded, as a table (Figure 3). Once you have finished inspecting the input data, the window may be closed by clicking on the OS-specific "Close Window" widget on the window frame (usually, something like a red and white "X" glyph on the top right hand corner of the window).

Figure 3) Preview Dataset Window



Explora DataSet: Datos_11_25_2013

accessio	Fruit_dia	Seed_wg	nr_flow_axi	Fruit_wg	Plant_hg	Fruit_lengt	Seec
10	20.0	0.3	1.0	69.4	40.8	37.3	3
11	9.5	0.2	2.0	6.7	33.7	10.7	1
18	24.9	0.4	1.3	122.4	38.3	51.0	4
41	16.3	0.4	1.7	28.0	43.8	47.1	1
43	14.3	0.4	2.0	35.5	42.7	55.3	4
50	29.2	0.4	2.0	104.1	55.6	51.8	8
51	21.9	0.4	2.0	85.8	50.6	57.2	4
54	27.0	0.4	2.0	79.4	56.6	55.0	5
67	14.9	0.5	1.7	30.4	62.7	38.3	1
68	14.7	0.3	2.0	21.3	56.7	21.9	4
85	19.2	0.5	1.3	41.8	62.2	27.3	4
147	26.5	0.7	1.0	73.5	51.9	23.5	9
149	22.4	0.6	1.0	55.8	54.0	21.7	8

Step 2. Set Analysis Parameters Button

By clicking this button, the dataset is loaded into memory and a new "Set Analysis Parameters" tab view is displayed (Figure 4a).

Step 2. Set Analysis Parameters Tab

Data Analysis Tag

The results of every analysis run is stored in its own subdirectory of the project subdirectory. The name of this subdirectory is taken from the specified "Data Analysis Tag" string, which defaults to a kind of system time stamp (of the time when the project was first loaded for analysis) but may be reset to any string of characters meaningful to the end user and compatible with the operating system's directory naming coventions.

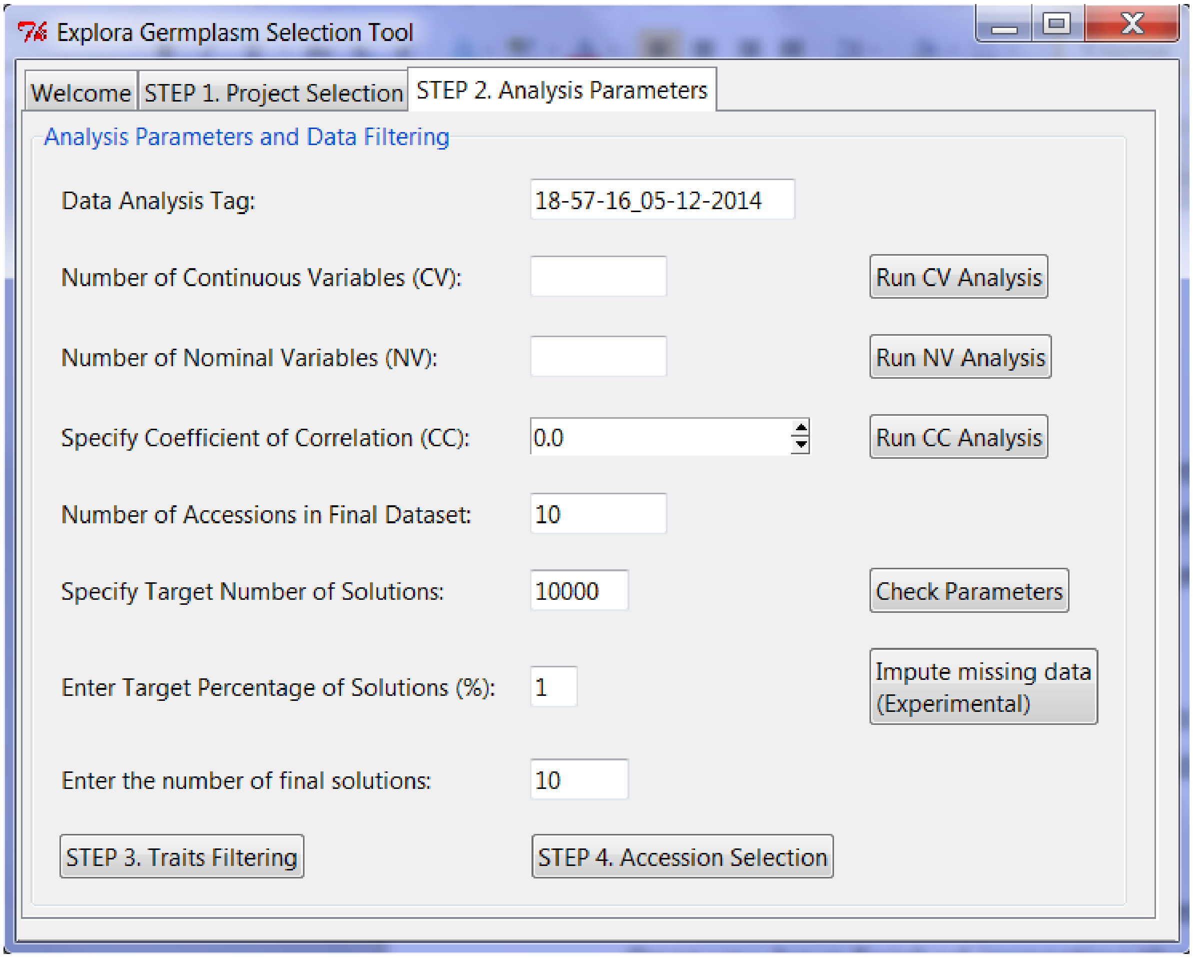
Number of Continuous Variables & Run CV Analysis

This parameter slot should be set to the 'C' number of continuous values in the input data set, as noted previously (see above). Once specified, users may click the "Run CV Analysis" button to trigger computation of some basic statistical inferences on the continuous data, the results of which are displayed in a modeless window (Figure 4b).

Number of Nominal Variables & Run NV Analysis

Comparable operations and results as per Continuous variables, except for Nominal variables.

Figure 4a) Analysis Parameters page



Explora Germplasm Selection Tool
Welcome
STEP 1. Project Selection
STEP 2. Analysis Parameters
Analysis Parameters and Data Filtering
Run CV Analysis
Run NV Analysis
Run CC Analysis
Check Parameters
Impute missing data (Experimental)
STEP 3. Traits Filtering
STEP 4. Accession Selection

Data Analysis Tag: 18-57-16_05-12-2014

Number of Continuous Variables (CV):

Number of Nominal Variables (NV):

Specify Coefficient of Correlation (CC): 0.0

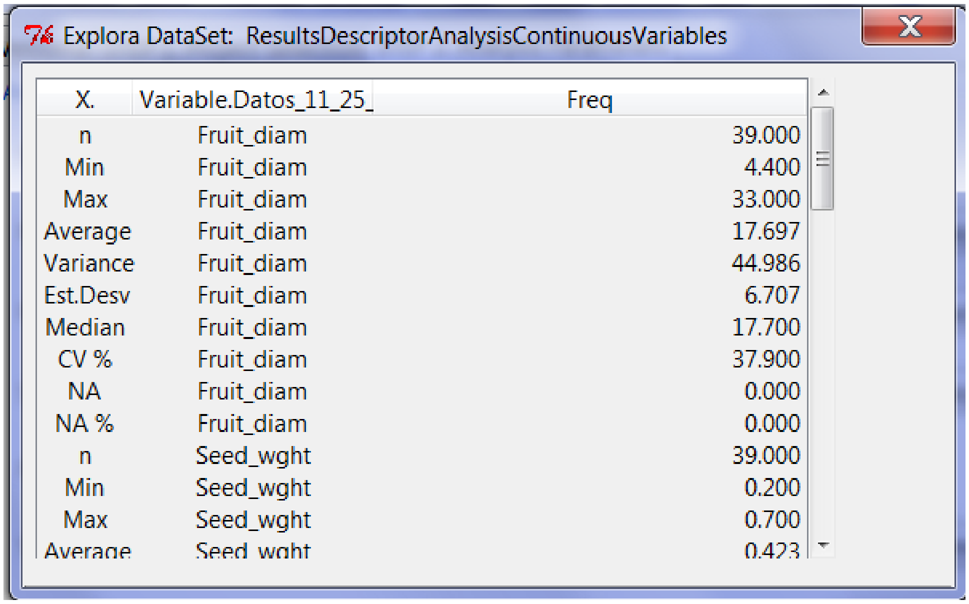
Number of Accessions in Final Dataset: 10

Specify Target Number of Solutions: 10000

Enter Target Percentage of Solutions (%): 1

Enter the number of final solutions: 10

Figure 4b) Continuous Variable Statistics Window



Explora DataSet: ResultsDescriptorAnalysisContinuousVariables

X.	Variable.Datos_11_25_	Freq
n	Fruit_diam	39.000
Min	Fruit_diam	4.400
Max	Fruit_diam	33.000
Average	Fruit_diam	17.697
Variance	Fruit_diam	44.986
Est.Desv	Fruit_diam	6.707
Median	Fruit_diam	17.700
CV %	Fruit_diam	37.900
NA	Fruit_diam	0.000
NA %	Fruit_diam	0.000
n	Seed_wght	39.000
Min	Seed_wght	0.200
Max	Seed_wght	0.700
Average	Seed_wght	0.423

Coefficient of Correlation Analysis

Users may select a given level of correlation then click the corresponding button, to generate the results of correlation analysis on the data. The results are also posted to a modeless window.

Check Parameters

The following set of parameters need to fall within valid ranges of values. After specifying the values of such parameters, clicking this button triggers a validation of the values, with errors reported back to the user.

Number of Accessions in Final Dataset

Here you set the target number of accessions you wish to have at the end of the optimization selection of members from the input dataset of accessions.

Target Number of Solutions

Several of the optimization algorithms perform a sampling of the space of possible subsets of accessions that could be presented as optimized solutions. Here you specify how many such samples should be generated for analysis.

Target Percentage of Solutions

Here you set the minimum target percentage of solutions you wish to choose at the end of specific optimization analysis procedures.

Number of Final Solutions

Here you set the target number of final solutions you wish to have at the end of the optimization selection from the input dataset of accessions.

Impute Missing Data

The current release of Explora has a limited capacity to deal with missing trait data values (which are generally noted as “NA” cell entries in the CSV data file). In this respect, the software has three measures strategies:

1) Columns and rows detected to be completely filled with missing data (‘NA’) values are deleted during data project creation.

2) Missing data is detected during "Check Parameters" feature of the parameter setting step and reported to the user.

3) Upon receiving the “Check Parameters” notification about missing data, users may (optionally) run a data imputation process on the data set to infer in missing data by clicking the “Impute Missing Data” button. The initial implementation is based on the Amelia R statistical package (see http://cran.r-project.org/web/packages/Amelia/vignettes/amelia.pdf) which can impute values across “cross-sectional” data, among other things. One can perhaps the accession measurements as being taken as a normal multi-variate trait data measured at a particular point in time (or rather, time invariant). However, Amelia’s fitness-to-purpose for trait data is not yet established. Moreover, the Explora implementation using Amelia doesn't (yet) leverage other parameters to guide the processing (e.g. Amelia ought to be told what variables are continuous versus nominal versus ordinal, but not yet done... Other tweaking of the procedure may also help in the future...).

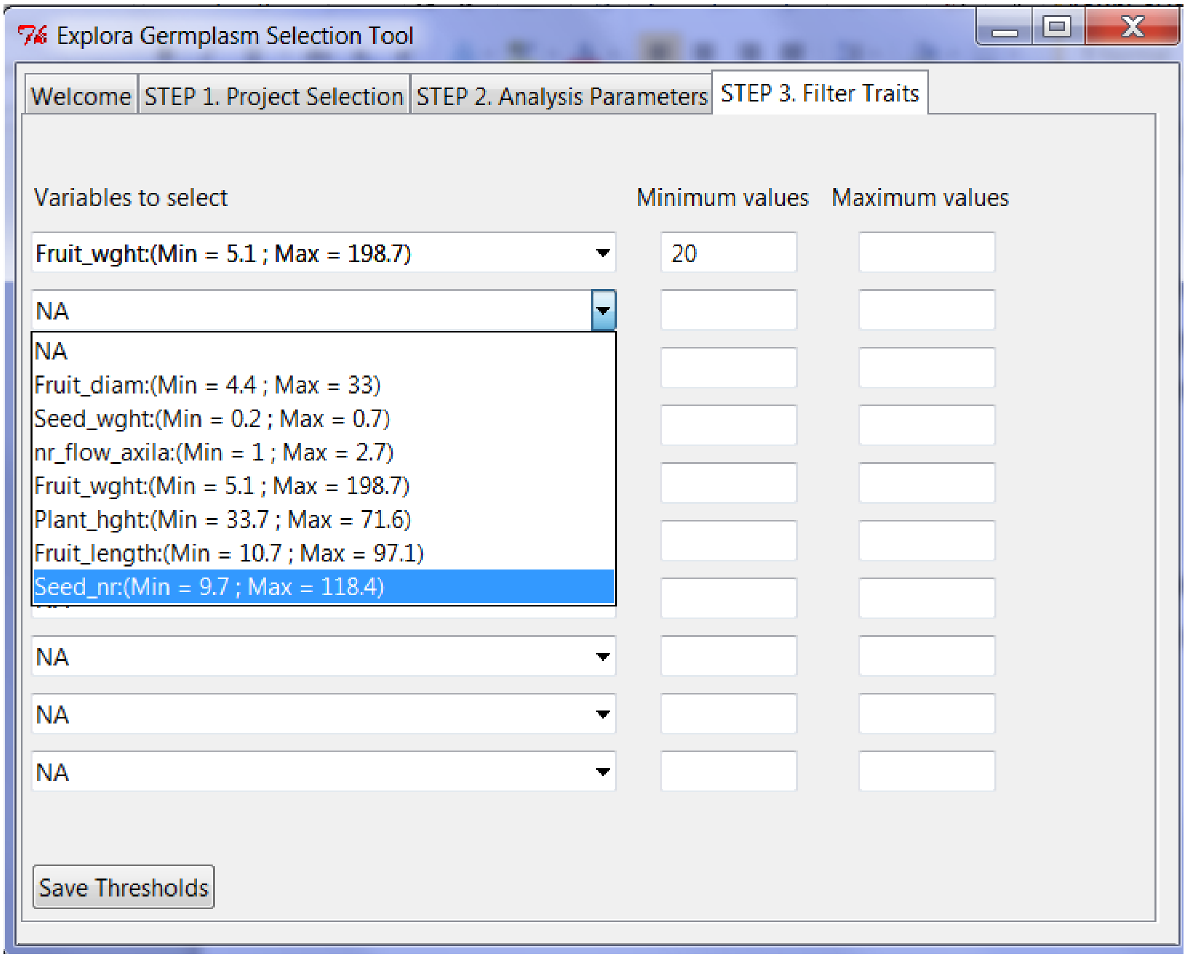
Step 3. Traits Filtering Button

By clicking this button, a new “*Step 3. Traits Filtering*” tab view is displayed (Figure 5). This view provides a means to pre-filter the acceptable range of values of the continuous variables in the input dataset. Only accessions satisfying these constraints are propagated to subsequent (Step 4) optimization selection.

Step 4. Accession Selection Button

By clicking this button, a new “*Step 4. Accession Selection*” tab view is displayed (Figure 6a). This view is the main one guiding the optimization analysis to select suitably optimized subsets of accessions from the input dataset.

Figure 5) Traits Filtering Tab View



Explora Germplasm Selection Tool
Welcome
STEP 1. Project Selection
STEP 2. Analysis Parameters
STEP 3. Filter Traits
Variables to select
Minimum values
Maximum values
NA
Fruit_diam:(Min = 4.4 ; Max = 33)
Seed_wght:(Min = 0.2 ; Max = 0.7)
nr_flow_axila:(Min = 1 ; Max = 2.7)
Fruit_wght:(Min = 5.1 ; Max = 198.7)
Plant_hght:(Min = 33.7 ; Max = 71.6)
Fruit_length:(Min = 10.7 ; Max = 97.1)
Seed_nr:(Min = 9.7 ; Max = 118.4)
NA
NA
NA
Save Thresholds

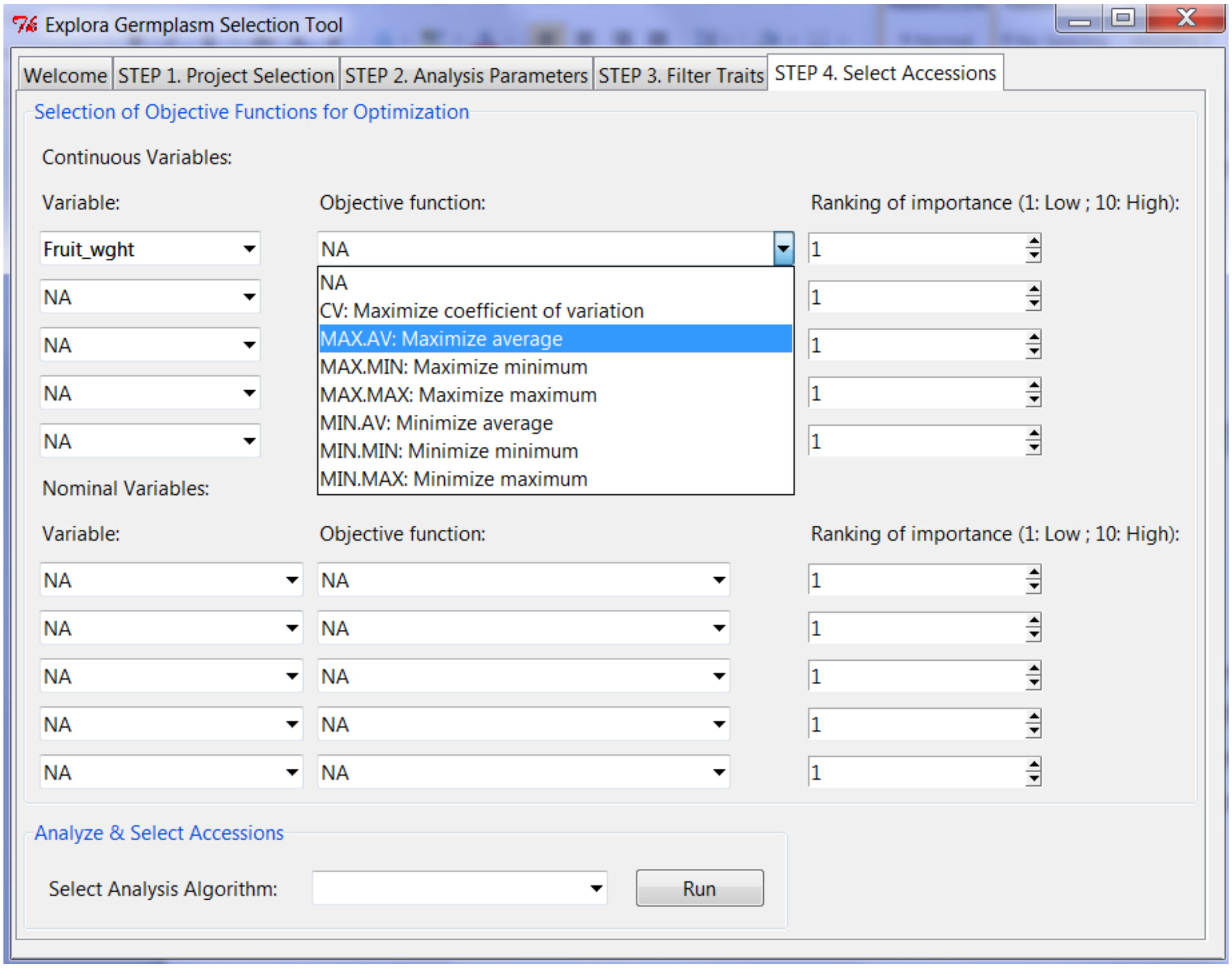
Fruit_wght:(Min = 5.1 ; Max = 198.7) | 20

NA

Step 3. Traits Filtering Tab

This tab view (Figure 5) allows users to pre-filter the list of input accessions to be analyzed in a subsequent optimization selection based on constraining the ranges of continuous trait values allowed. First, one selects a target continuous variable from a dropdown menu. Each entry shows the input data minimum and maximum values. The proposed minimum and maximum values may be specified in the boxes alongside a selected trait. If either box is left blank, then the range defaults to the dataset maximum and/or minimum values. After all such ranges of trait values have been entered, the resulting thresholds must be saved, by clicking the button provided at the bottom of the view for this purpose.

Figure 6a) Accession Selection Tab View: Selection of Objective Functions


Explora Germplasm Selection Tool
Welcome
STEP 1. Project Selection
STEP 2. Analysis Parameters
STEP 3. Filter Traits
STEP 4. Select Accessions

Selection of Objective Functions for Optimization

Continuous Variables:

Variable:	Objective function:	Ranking of importance (1: Low ; 10: High):
Fruit_wght	NA	1
NA		1
NA		1
NA		1
NA		1

NA
CV: Maximize coefficient of variation
MAX.AV: Maximize average
MAX.MIN: Maximize minimum
MAX.MAX: Maximize maximum
MIN.AV: Minimize average
MIN.MIN: Minimize minimum
MIN.MAX: Minimize maximum

Nominal Variables:

Variable:	Objective function:	Ranking of importance (1: Low ; 10: High):
NA	NA	1
NA	NA	1
NA	NA	1
NA	NA	1
NA	NA	1

Analyze & Select Accessions

Select Analysis Algorithm:

Step 4. Accession Selection Tab

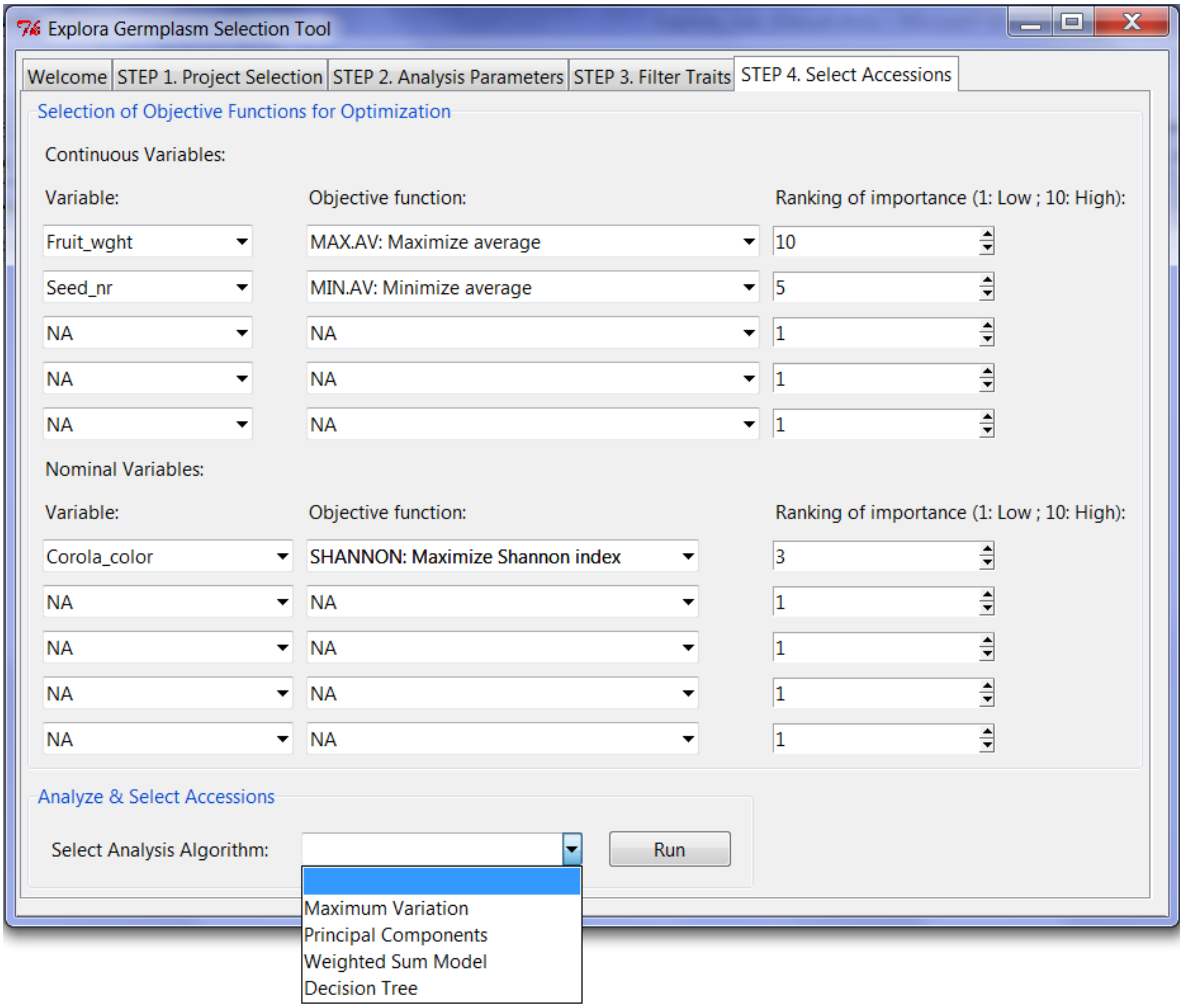
This tab view (Figure 6a) has two parts to it:

1. Selection of Objective Functions
2. Analysis and Selection of Accessions

For the selection of objective functions, one selects a trait from a dropdown menu, then selects the objective function to apply to this trait. For some optimization computations, a ranking of importance is also used and may be specified here as indicated. Continuous variables and nominal variables each have their own types of applicable objective functions (not elaborated further here). For some algorithms (like PCA), you should have specified at least 2 distinct objective functions to apply the optimization.

After selecting objective functions, a selection algorithm may be selected from the given dropdown menu (Figure 6b) and executed by clicking the Run button. The results for each algorithm are saved under the corresponding tagged Project data directory. The selected subset of accessions is displayed in modeless windows as they are generated.

Figure 6b) Accession Selection Tab View: Running of Algorithms to an Optimized Set of Accessions



Explora Germplasm Selection Tool
Welcome
STEP 1. Project Selection
STEP 2. Analysis Parameters
STEP 3. Filter Traits
STEP 4. Select Accessions

Selection of Objective Functions for Optimization

Continuous Variables:

Variable:	Objective function:	Ranking of importance (1: Low ; 10: High):
Fruit_wght	MAX.AV: Maximize average	10
Seed_nr	MIN.AV: Minimize average	5
NA	NA	1
NA	NA	1
NA	NA	1

Nominal Variables:

Variable:	Objective function:	Ranking of importance (1: Low ; 10: High):
Corola_color	SHANNON: Maximize Shannon index	3
NA	NA	1
NA	NA	1
NA	NA	1
NA	NA	1

Analyze & Select Accessions

Select Analysis Algorithm: