UPOV

International Union for the Protection of New Varieties of Plants

E

Working Group on Biochemical and Molecular Techniques and DNA-Profiling in Particular

BMT/18/11

Eighteenth Session
Hangzhou, China, October 16 to 18, 2019

Original: English
Date: September 10, 2019

REPORT ON DEVELOPMENTS OF A SOFTWARE TOOL FOR MARKER SELECTION USING THE TRAVELING SALESMAN ALGORITHM

Document prepared by an expert from the Seed Association of the Americas (SAA)

BACKGROUND

Previous work has demonstrated that small sets of SNPs can be very effective to distinguish plant varieties when SNPs are selected largely based on informativeness i.e. abilities to individually discriminate among varieties while collectively allowing most, if not all chromosomes to be sampled (Liu et al, Yoon et al, Jones). Here a different approach to SNP selection for variety ID is investigated using an algorithm to select SNPs based solely on their discrimination abilities regardless of known map positions. The Uniqueness program is based on a computer algorithm called the Traveling Salesman Problem. (Lawler et al). The Traveling Salesman Problem is described as the most efficient city-to-city route a salesperson should travel while visiting each city exactly one time. This concept was applied to determine the minimum number of SNPs needed to differentiate all plant varieties in any genotypic database. The algorithm randomly places a marker into a set and then determines whether discrimination power has been improved compared to previous best performing sets. Then the process is iterated thousands of times. Then the user systematically increases the number of SNPs until all varieties are differentiated.

USE OF THIS APPLICATION

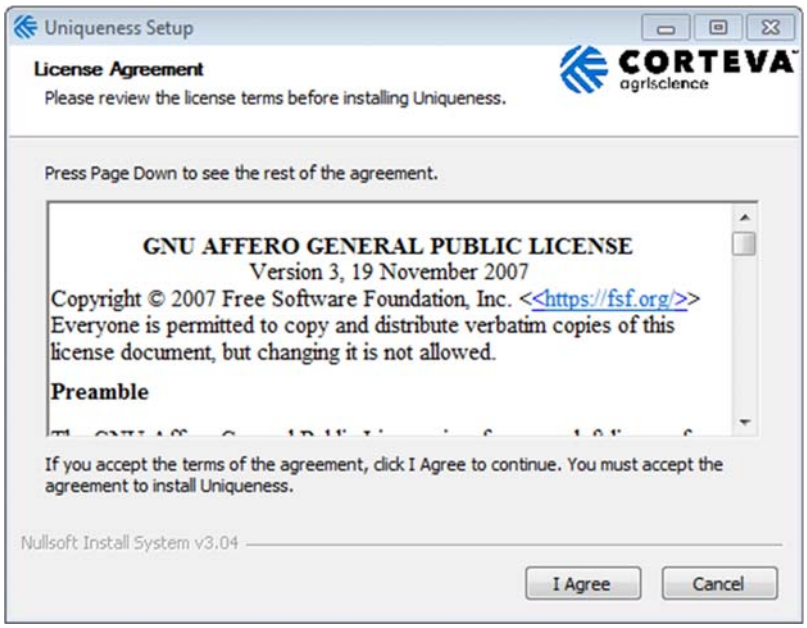
"Uniqueness" was designed to be a molecular marker selection tool that allows users to select a minimum number of markers (from a database of any size) to uniquely identify genotypes for use in Intellectual Property / Seed Testing related activities. However, uses are not limited to molecular marker selection only, rather any application where selection of small yet robust data variables are needed.

- Select small powerful marker sets for cost effective high throughput assays
  Alternatively, Uniqueness can perform a similarity analysis to select the most similar rows among columns.
- Can be used to select similar markers or other data variables that are common among columns.

OPEN SOURCE CODE

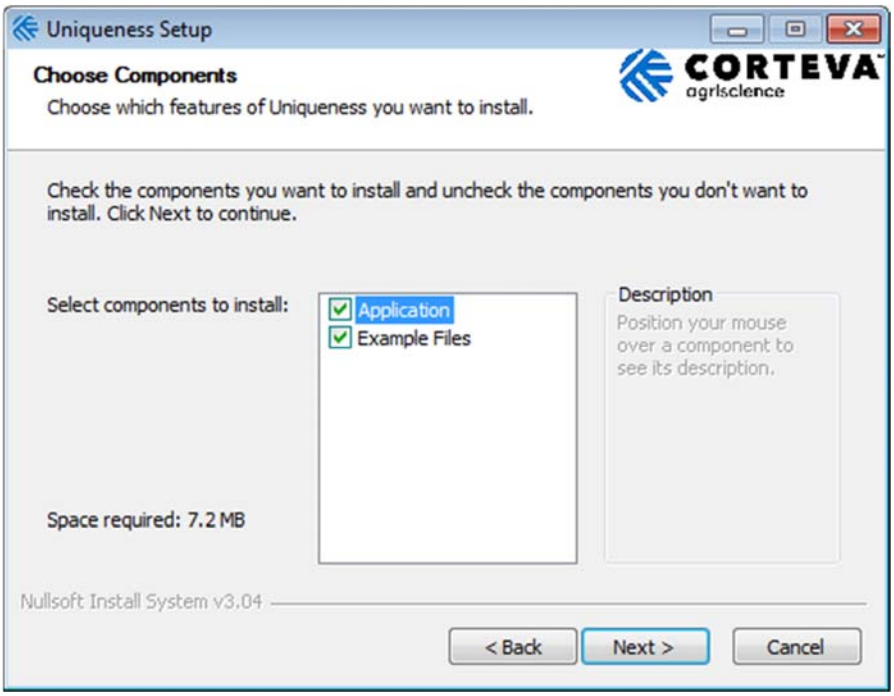
"Uniqueness" is written and positioned to be open source. The program can be downloaded at: https://github.com/corteva

Select ‘I Agree’ only if you agree to the terms of use for this application.

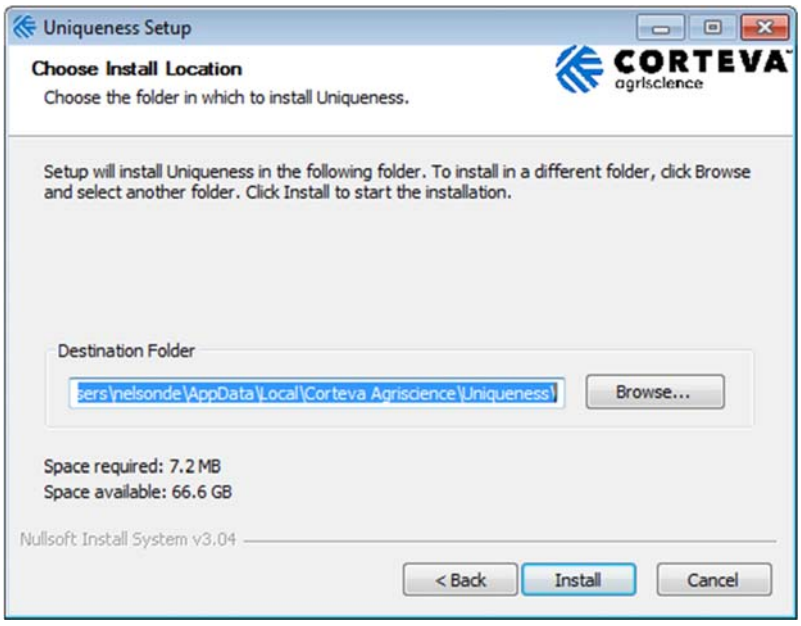


INSTALLATION

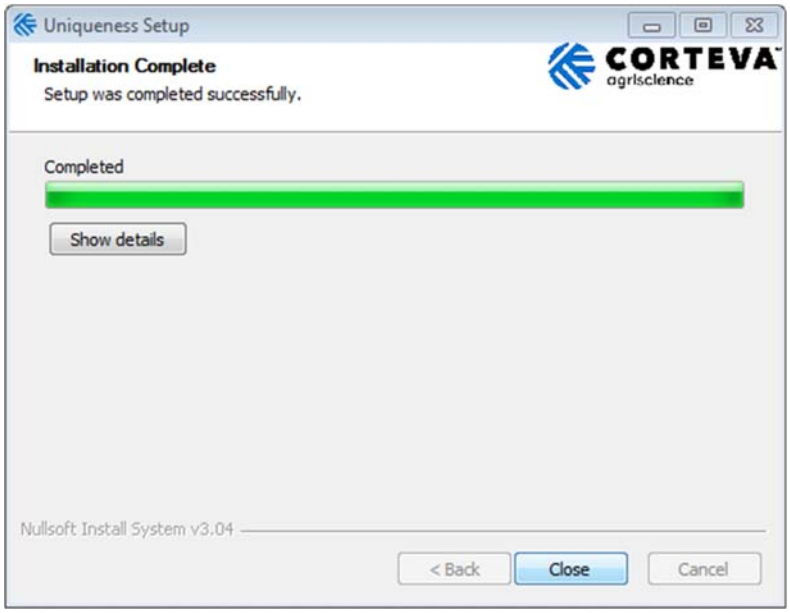
Keep ‘Application’ checked to install Uniqueness. The example file ‘example input.csv’ is loaded when the program is installed. Uncheck the example file if you do not wish to have this file.



User has the option to use the default application folder or select a different folder using the browse option.



Close the installation window. Open the application in your start up option or select from the folder destination selected.



FILE LAYOUT

File types

The file types best suited for this application is tab or comma delimited (.txt or .csv).

Design

The first row of the file, regardless of the format must contain headers. Note: When the output file is created the column headers will be returned the same as the input file.

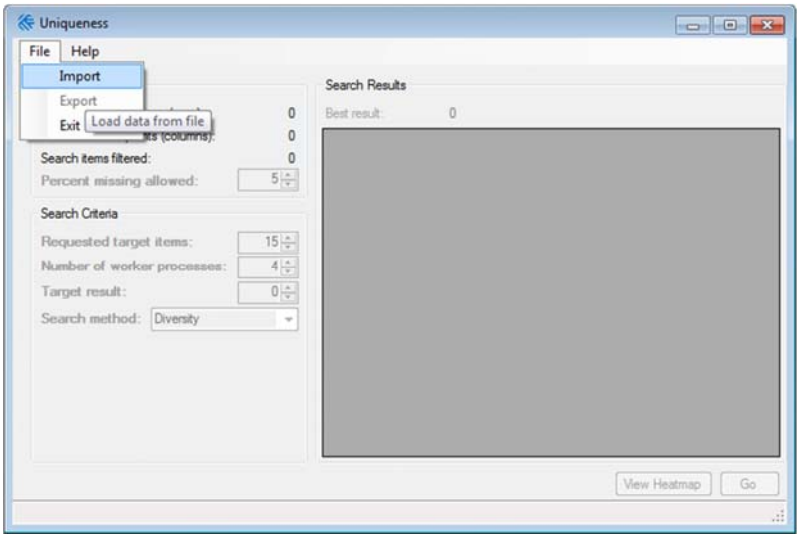
Unique Id	Name	Required	Sprite	Probst	KS4694	MN1302	Iroquois	Simpson
1	BARC_1.01_Gm_01_1013695_A_G	0	1	1	1	1	1	1
2	BARC_1.01_Gm_01_1059407_T_C	0	4	4	4	4	2	4
3	BARC_1.01_Gm_01_10768239_C_T	0	2	2	2	2		2
4	BARC_1.01_Gm_01_11835461_T_C	0	4	2	2	2		4
5	BARC_1.01_Gm_01_11904297_T_C	0	4	4	4	4	4	4

- First Column must contain a '**Unique Id'** for each row
- Second column should be the '**Name'**, i.e. marker name
- Third Column signifies if the row is '**Required'** in the final results
  - 0 = Not Required
  - 1 = Required
- Remaining columns contains the Data Set.

There is no limit in place with regards to the number of rows or columns, however please keep in mind the speed of the returned results is dependent on your computer memory.

IMPORT FILE

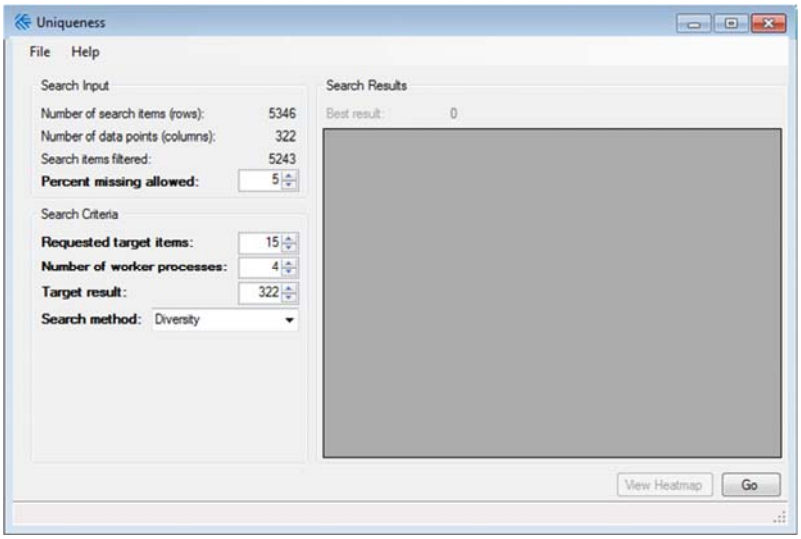
User must import file from the desired file path. If it is not in the correct format, as indicated above, the file will not import.



The User Interface will remain disabled until the file is imported. On import, the program automatically computes percent missing data for each row.

SEARCH INPUT

Results from the data input file.

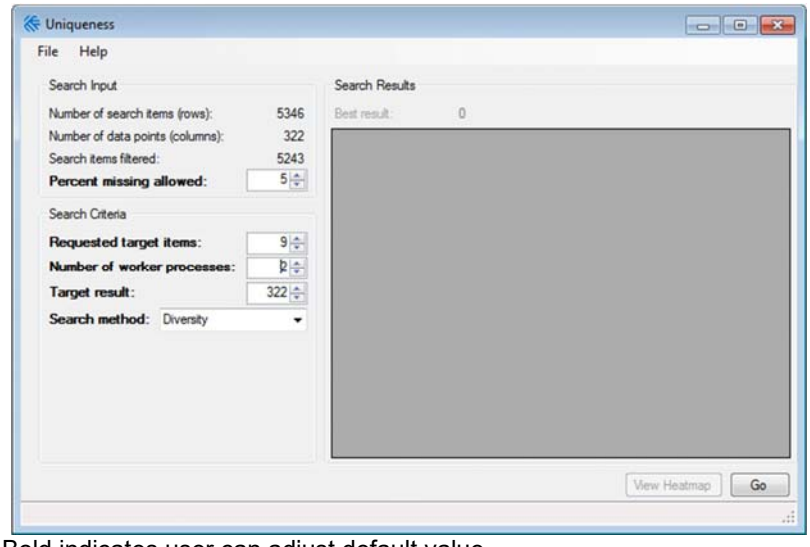


* Bold variables indicate where the user can adjust default values.

- 'Number of search items' (rows): This field indicates the original number of rows included in the data set.
- 'Search items filtered' (rows): When the user adjusts the percent missing allowed, the data set is auto-filtered. This field tells the user the number of rows available for evaluation.
- 'Number of data points' (columns): This field informs the number of columns being evaluated.
- 'Percent missing allowed': This number can manually be adjusted to review the percent missing in the data set.
    - The minimum number is 0.
    - The maximum number is 100.

SEARCH CRITERIA

Allows user to adjust parameters used for searching.



* Bold indicates user can adjust default value.

- ‘Requested target items’: This value corresponds to the number of rows desired by the user.
    - Default set to 15
    - For marker selection, this would be the fewest number of markers to achieve differentiation of fields (columns).

- ‘Number of worker process’: Automatically read from your pc. This can be cut to allow your computer to do additional tasks or increased to review data set faster.
    - Default: varies by computer

- ‘Target Result’: Allow the user to manually adjust the number of results to find
    - Defaults to the number of columns in data set.

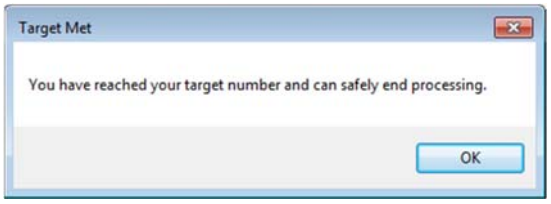
- ‘Search Method’ Diversity or Similarity: Allows the user to adjust the application to search the data set for similarity or differences of target items.
    - Defaults to ‘Diversity’ to select ‘Requested target items’ (rows) to uniquely differentiate ‘Target result’ (columns).
    - ‘Similarity’ performs the opposite analysis selecting the most similar rows among columns.

RUNNING THE APPLICATION

Select ‘Go’ to begin evaluation.

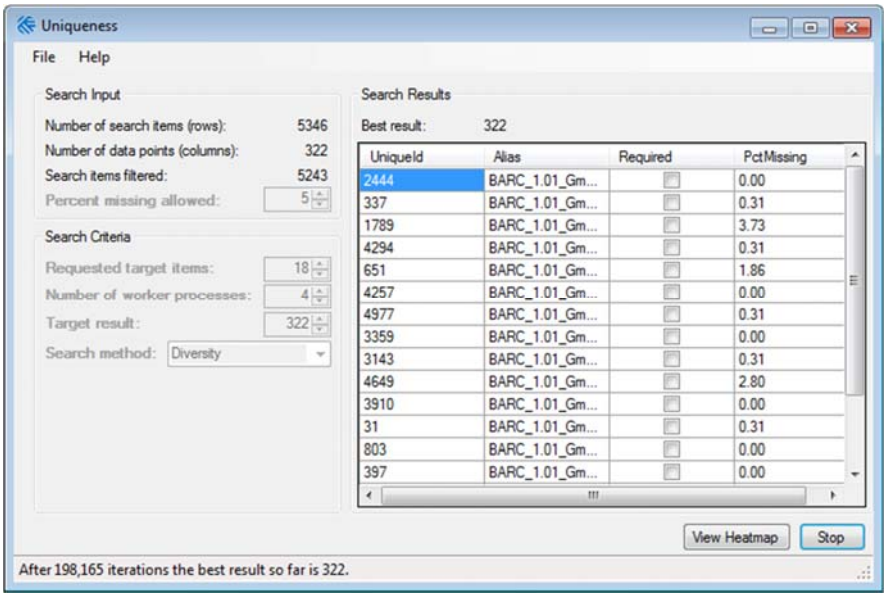
To stop the application from processing, click the ‘Stop’ button (formerly the ‘Go’ button).

When the target results have met the number of data points, a notification will display indicating you can get your results. Even after this notification displays, the program continues processing until you click the ‘Stop’ button.



RESULTS GRID

The results grid will display the records found in the data set.



Best Result: Total number of differentiated columns based on search criteria

- Uniqueid: First column in the data set used to identify the record
- Alias: Name supplied in the input data set, i.e. marker name.
- Required: Checked if the row was required in the input data set.
- PctMissing data: Computed percent missing data for each row (record) in the data set.

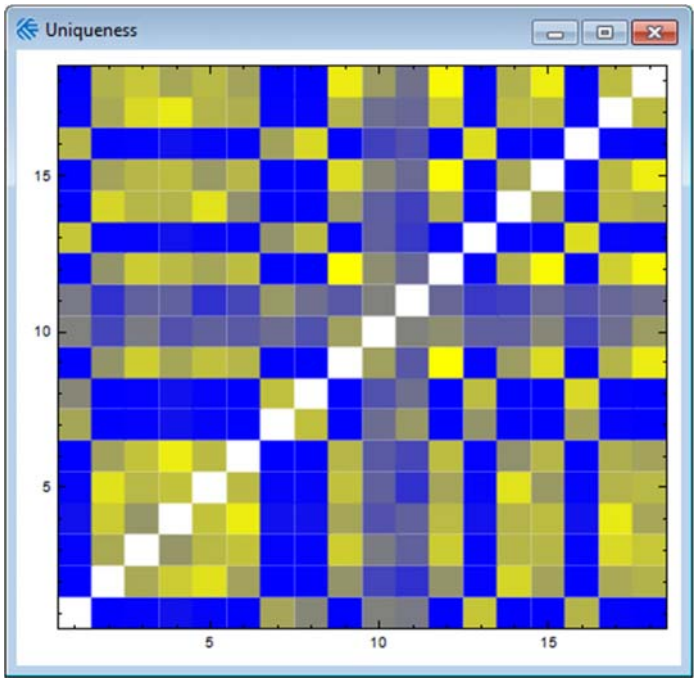
HEAT MAP

The heat map will indicate the differences in the results grid.

After 'Go' is selected View Heatmap will be enabled. The heatmap will open in a new window.

When the user hovers over the cell in the heatmap a difference score will display between two row's aliases.

The color key displays from yellow to blue with yellow displaying less diversity and blue showing more.



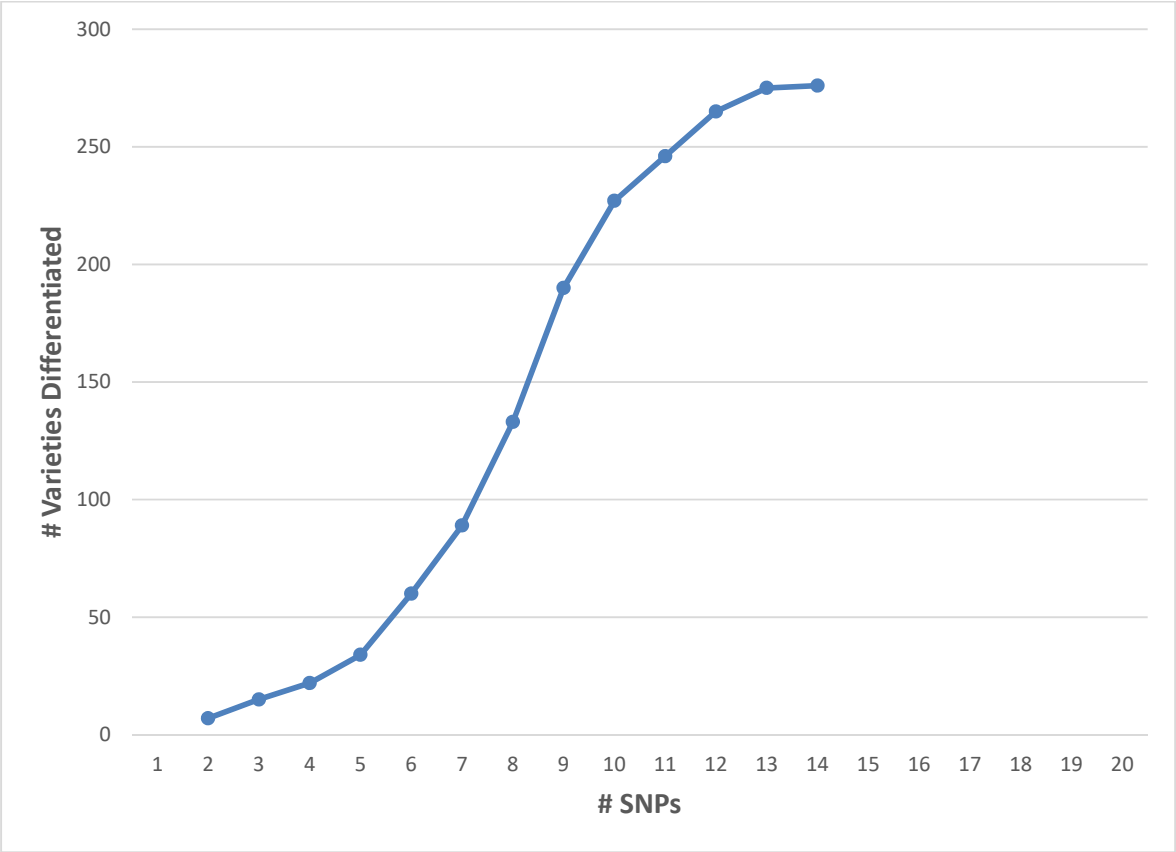
EXPORT RESULTS FILE

Original file headers will be included in the output file. The application will default back to original file path. The file name will default to the original file name with the date, time (24 hr with seconds) underscore "result" as the file name.

If the user has not stopped the application the export functionality will not work.

PERFORMANCE RESULTS

Example. Differentiating power of successive SNPs selected using Uniqueness in soybeans using 276 varieties across 5346 SNP loci.



Uniqueness performed very well in selecting 14 SNPs to differentiate all 276 soybean varieties. A key point in the analysis is that it's assumed that all the varieties deemed necessary to differentiate are contained in the database used. Therefore, if varieties or genetics change among the varieties one desires to identify, the analysis and SNP selection process may need to be repeated to reflect these germplasm diversity changes.

REFERENCES

- Gale, Z. Jiang, H., Westcott M. 2005. An optimization method for the identification of minimal sets of discriminating gene markers: Application to cultivar indentification in wheat. Jour. Bioinformatics and Computational Biology. 3:269-279.

- Jones, L. 2013. Varietal Identification in Maize: Are Sixteen SNP Markers Sufficient? UPOV BMT/12/15. Ottawa, Canada.
  Nucl. Acids Res. (2010) 38 (suppl 1): D843-D846.

- Lawler, E. L., Lenstra, J. K., Rinnooy Kan, A. H. G., and Shmoys, D. B. 1985. *The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization.* John Wiley and Sons, New York.

- Liu, Z., Li, J., Fan, X., Htwe, N. M. P. S., Wang, S., Huang, W., ... Qiu, L. (2017). Assessing the numbers of SNPs needed to establish molecular IDs and characterize the genetic diversity of soybean cultivars derived from Tokachi nagaha. *Crop Journal*, *5*(4), 326-336.

- Yoon, M.S., Q.J. Song, I.Y. Choi, J.E. Specht, D.L. Hyten, P.B. Cregan. 2007. BARCSoySNP23: a panel of 23 selected SNPs for soybean cultivar identification. Theor. Appl. Genet. 114, 885-899.

[End of document]