

# LociScan V2.0 User Manual

## Contents

1.	Preface .....	2
1.1	Purpose of the Manual .....	2
1.2	Project Background .....	2
2.	Software Overview .....	2
2.1	Target .....	2
2.2	Operating Environment .....	3
2.3	Function Structure .....	3
2.4	Main Interface .....	3
3.	Instructions .....	6
3.1	Installation and Initialization of GUI Version .....	6
3.2	Installation and Initialization of CLI Version .....	7
3.3	Input of Training Data .....	8
3.4	Input of Loci Combinations to be Evaluated .....	10
3.5	Output of Loci Combination Screening Results .....	11
3.6	Output of Loci Combination Evaluation Results .....	12
3.7	Output of Data Cleaning Results .....	13
3.8	Main Function Modules .....	13
3.8.1	Loci Combination Screening .....	13
3.8.2	Loci Combination Evaluation .....	13
3.8.2	Data Cleaning .....	13
3.9	Operating Procedures .....	13
3.9.1	Loci Combination Screening - GA .....	13
3.9.1.1	GA Operation in GUI Version .....	14
3.9.1.2	GA Operation in CLI Version .....	17
3.9.2	Loci Combination Screening - EA .....	19
3.9.2.1	EA Operation in GUI Version .....	20
3.9.2.2	EA Operation in CLI Version .....	21
3.9.3	Loci Combination Evaluation .....	23
3.9.3.1	Evaluation Operation in GUI Version .....	23
3.9.3.2	Evaluation Operation in CLI Version .....	24
3.9.4	Data Cleaning .....	25
3.9.4.1	Cleaning Operation in GUI Version .....	26
3.9.4.2	Cleaning Operation in CLI Version .....	26

### Open access

LociScan can be downloaded at:

<https://gitee.com/caurwx/lociscan> or <https://github.com/caurwx1/LociScan>

### Developer contact information

For permission or any technical questions, please email to:

[caurwx@163.com](mailto:caurwx@163.com) or [caurwx@gmail.com](mailto:caurwx@gmail.com)

# **1. Preface**

## **1.1 Purpose of the Manual**

This manual includes instructions and operation procedures for LociScan V2.0, to help the users quickly understand and manage the software. The software provides methods to screen out the best VDP loci combinations out of a massive pool in a quick, simple and effective way, mainly for plant variety discrimination.

## **1.2 Project Background**

Along with the development of molecular marker technology, the third generation molecular markers, single nucleotide polymorphism (SNP) and Insertion and Deletion of nucleotides (InDel), have been gradually explored in the application of plant variety discrimination. Optimization of tens of thousands of loci in terms of both quality and quantity as per the demands of practical operation can not only reduce the application cost, but also improve the data analysis efficiency of the technology.

The regular molecular marker loci screening methods usually carry out analysis on the genetic background of the samples and select the best molecular marker loci combinations by taking genetic diversity as the appraisal index. Taking into consideration the features of molecular marker technology used for plant variety discrimination, this software raised a loci screening model for plant variety discrimination based on genetic algorithm (GA), introduced it into the solutions of loci combination screening problem by making full use of the advantages of GA in optimizing solutions, defined the statistical methods for plant variety discrimination power(VDP) according to the characteristics of plant variety discrimination, designed the fitness functions and restriction conditions of GA, and developed a GA model fit for loci combination screening of plant variety discrimination, which improves the plant variety discrimination capability of the screened loci combinations.

# **2. Software Overview**

## **2.1 Target**

LociScan V2.0 is developed under the framework of .Net Framework 4.0, and the software can quickly and effectively screen out the best loci combinations by molecular markers such as SSR/STR, SNP and INDEL. The graphical user interface (GUI) version adopts traditional user interface design to facilitate quick master of the software by the users. The command line interface (CLI) version provides command line operation to facilitate handling of large scale training data. The software only supports reading and output of data in .CSV format. The core function is to screen out loci combinations with the highest VDP value out of a massive pool in a

quick and simple way, so as to reduce the number of molecular marker loci, cut the cost of molecular tests and improve the efficiency of plant variety discrimination.

## 2.2 Operating Environment

### (1) Software Environment

GUI version: Windows 7 or higher version, with .Net Framework 4.0 or higher version already installed.

CLI version: CentOS 7, Ubuntu 16.04, Debian 9, Raspbian 9, Fedora 28 or higher version Linux systems recommended for the server, with MonoDevelop source code IDE already installed; macOS 10.9 or higher version macOS systems with MonoDevelop source code IDE already installed; Windows 7 or higher version with .Net Framework 4.0 or higher version already installed.

### (2) Hardware Environment

CPU: 1GHz or higher.

RAM: 1GB or larger.

Display: 1024×768 or higher, 32-bit true color

## 2.3 Function Structure

The software is composed of three modules, loci combination screening, loci combination evaluation and data cleaning. The function structure is shown as follows:

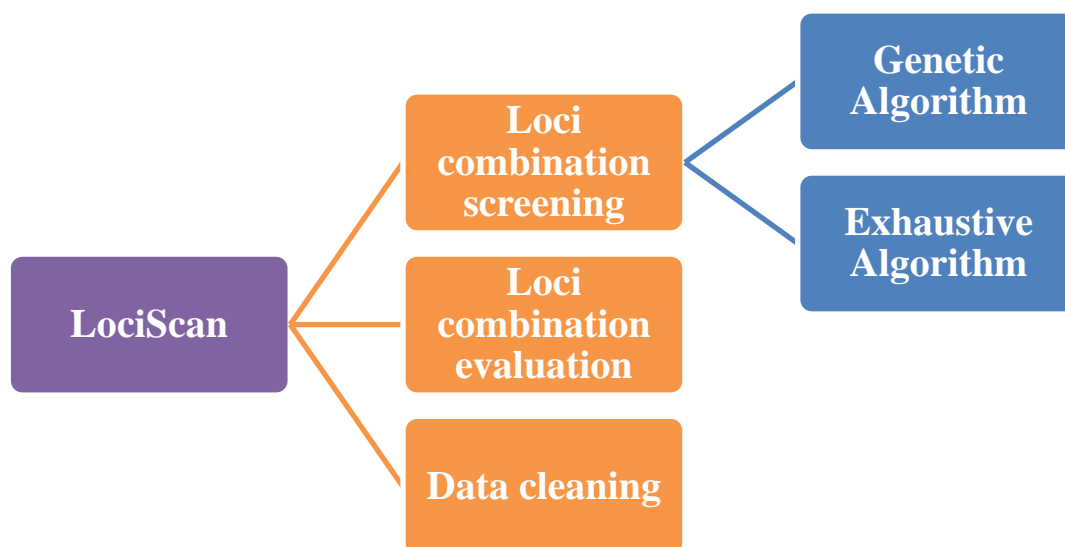


Figure 1 Software Function Structure

## 2.4 Main Interface

Click **Loci combination screening** to enter the loci combination screening page (Figure 2). Click **Loci combination evaluation** to enter the loci combination evaluation page (Figure 3). Click **Data**

cleaning to enter the data cleaning page (Figure 4).

The screenshot shows the LociScan V2.0 software window. At the top, there are three radio buttons for 'Operation': 'Loci combination screening' (selected), 'Loci combination evaluation', and 'Data cleaning'. Below this, there are fields for 'Marker Type' (SSR/STR, SNP/InDel, and Missing Allele Format) and 'Data Structure' (Row: Loci, Col: Samples or Row: Samples, Col: Loci). A 'Data File Path(.csv)' field with an 'Open' button is present. A red warning message states: 'Warning: Please do not enter a data set that contains non-polymorphic loci (that is, loci with only one genotype); otherwise the program will keep running with no end.' The main settings area is divided into two tabs: 'Genetic Algorithm' and 'Exhaustive Algorithm'. Under 'Genetic Algorithm', there are input fields for 'Crossover Rate [0, 1] = 0.90', 'Mutation Rate [0, 1] = 0.01', 'Population Size [3, ∞) = 100', and 'Generation Size [1, ∞) = 2000'. There are checkboxes for 'Keep Best in Generation' (checked) and 'Stop when FF=1'. Below these are dropdown menus for 'Mutate Method: WholeRandom' and 'GA Method: normal-GA'. At the bottom of this section are 'Result Format' (All Generation, Last Generation) and 'File Format' (Unit Results, Merged Results) options. The 'Exhaustive Algorithm' section has a 'Fitness Function (FF): R-VDP' dropdown, 'Variety Threshold Setting' (Ratio of Different Loci (0,1] < 0.05 or Number of Different Loci [1,∞) < 1), 'Number of Target Loci' (Fixed Genome Size [2, ∞) = 3 or Batch Genome Size [2, ∞) from 2 to 10), and 'Times of Calculation Cycles: 1'. At the bottom right are 'OK', 'Open Result Folder', and 'Exit' buttons. The footer contains copyright information: 'Copyright©2023 BAAFS.MRC, Designer: YANGYANG, E-mail: caurwx@163.com'.

Figure 2 Loci Combination Screening Page

**LociScan V2.0**

Operation: ☐ Loci combination screening ☒ **Loci combination evaluation** ☐ Data cleaning

Marker Type: ☐ SSR/STR (e.g.188/266) ☒ SNP/InDel (e.g.AT,CG or AB) Missing Allele Format: \*, ??, -, ---

Data Structure: ☒ Row: Loci; Col: Samples ☐ Row: Samples; Col: Loci

Data File Path(.csv)

Loci File Path(.csv)

Fitness Function (FF):

Variety Threshold Setting:

☐ Ratio of Different Loci (0,1] <

☒ Number of Different Loci  $[1, \infty)$  <

Copyright©2023 BAAFS.MRC, Designer: YANGYANG, E-mail: caurwx@163.com

Figure 3 Loci Combination Evaluation Page

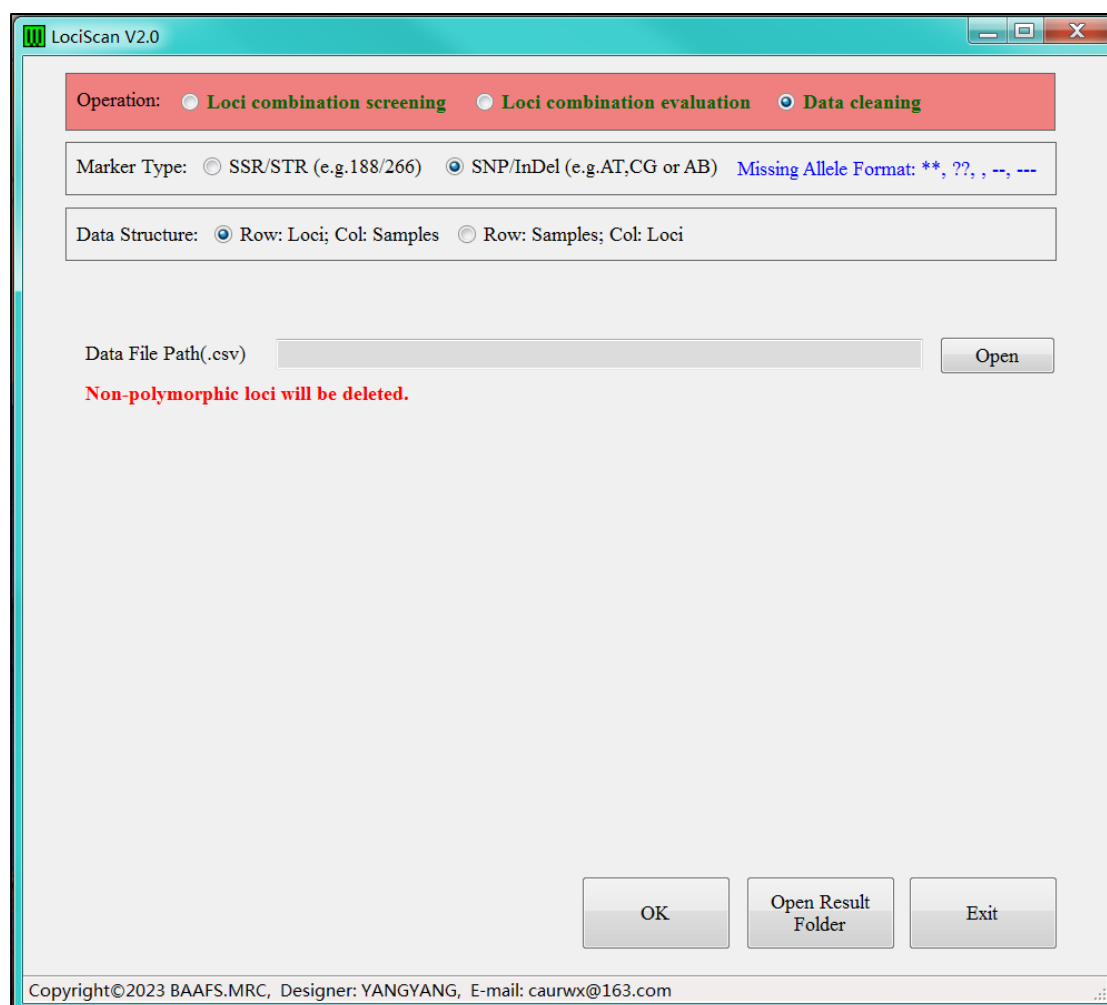


Figure 4 Data Cleaning Page

### 3. Instructions

#### 3.1 Installation and Initialization of GUI Version

- (1) As shown in Figure 5, double click the LociScan\_V2.0.exe application program, and enter into the main interface if permission is granted after automatic detection, or enter into the registration page if permission is not granted.

名称	修改日期	类型	大小
GeneticAlgorithm.dll	2023/3/28 15:26	应用程序扩展	32 KB
GeneticAlgorithm.pdb	2023/3/28 15:26	Program Debug Database	70 KB
LociScan_V2.0.exe	2023/4/10 9:45	应用程序	84 KB
LociScan_V2.0.exe.config	2021/11/5 10:22	XML Configuration File	1 KB
LociScan_V2.0.pdb	2023/4/10 9:45	Program Debug Database	116 KB
LociScan_V2.0.vshost.exe	2023/4/10 9:45	应用程序	23 KB
LociScan_V2.0.vshost.exe.config	2021/11/5 10:22	XML Configuration File	1 KB
LociScan_V2.0.vshost.exe.manifest	2010/3/17 22:39	MANIFEST 文件	1 KB

Figure 5 Running Program

- (2) The registration page is shown in Figure 6. E-mail the registration code to [caurwx@163.com](mailto:caurwx@163.com) to get the verification code and finish the registration.



Figure 6 Registration Page of GUI Version

### 3.2 Installation and Initialization of CLI Version

- (1) Taking CentOS 7 system as example, write the following command to install MonoDevelop and check if it is successful with the last line:

```
rpmkeys --import "http://keyserver.ubuntu.com/pks/lookup?op=get&search=0x3FA7E0328081BFF6A14DA29AA6A19B38D3D831EF"
su -c 'curl https://download.mono-project.com/repo/centos7-stable.repo | tee /etc/yum.repos.d/mono-centos7-stable.repo'
yum install mono-devel
mono -v
```

- (2) Please refer to <https://www.mono-project.com/download/stable/> for instructions on installing MonoDevelop in other Linux and macOS systems.
- (3) Switch the command terminal to the document file of the program and run the following command. The user will be prompted to receive a registration code for the first time as shown in Figure 7. E-mail the registration code to [caurwx@163.com](mailto:caurwx@163.com) to get the verification code and finish the registration.

```
mono LociScan-C.exe GASetting.txt
```

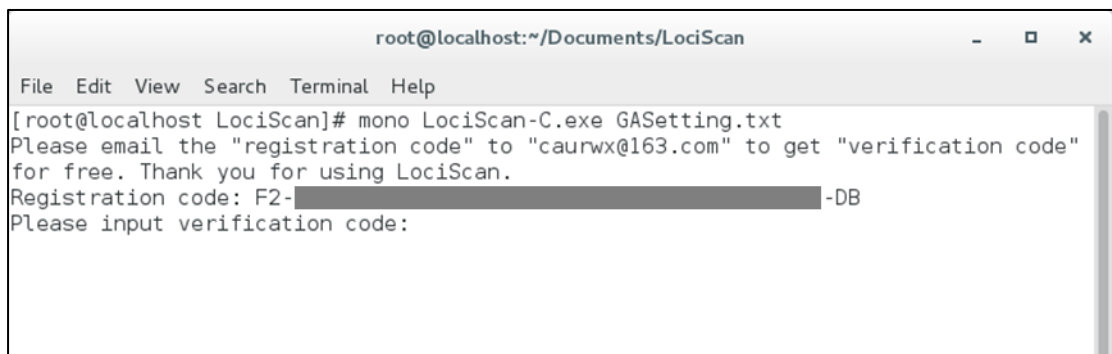


Figure 7 Registration Page of CLI Version

### 3.3 Input of Training Data

#### (1) Preparation of SNP or InDel Data

- The software supports SNP marker allelic data in the general format of A\T\C\G or simplified format of A\B and InDel marker allelic data in the general format of A\B. It only supports species with diploid chromosome and the actual genotype data are shown in Figure 8. Taking the general format as example, the homozygous types are AA, TT, CC and GG, and the heterozygous types are AT, AC, AG, TC, TG and CG. As for the simplified format, the homozygous types are AA and BB, and the heterozygous type is AB.
- In the data structure of Figure 8, rows represent loci and columns represent samples. The first row shows the serial number of samples and the other rows the corresponding genotype data of loci. The first column shows the serial number of loci and the other columns the corresponding genotype data of samples.
- The file format of the input data is .csv and the missing data are indicated with ??, --, --- or \*\*. The example of input data is shown in Figure 8.
- Before opening the data file, please close all programs which open the same data file.

```

probeset_id,S01,S02,S03,S04,S05,S06,S07,S08,S09,S10,S11,S12,S13,S14,S15,S16,S17,S18,S19,S20,
L001,GG,GG,AG,GG,AG,GG,GG,AG,GG,GG,??,AG,GG,GG,AG,AG,GG,GG,AG,AG,--,--,GG,--,AG,GG,--,GG
L002,GG,AG,GG,GG,AG,AA,AG,AG,AG,GG,AG,AA,AG,GG,AG,AG,AG,AA,--,AG,AG,AA,GG,--,GG,GG,GG,--,GG
L003,GG,GG,GG,GG,GG,GG,TT,GG,GG,GG,TT,GG,TT,GG,GG,GG,GG,GG,GG,GG,GG,TG,GG,GG,GG,GG,GG,GG
L004,AA,GG,AA,GG,AA,AA,AA,AA,GG,AA,AA,GG,AA,AA,AA,AA,AA,AA,GG,GG,AA,AA,AA,AA,GG,AA
L005,CC,CC,TT,CC,CC,TT,--,CC,CC,TC,CC,CC,CC,TC,*,TT,CC,CC,TT,TC,TT,TT,CC,CC,CC,CC,TC,CC
L006,GG,TT,TG,TT,--,TT,TT,TG,GG,TG,TT,TG,TT,--,TT,TG,TT,TT,TG,TG,TG,TT,TG,TT,TG,TT,TG
L007,--,AA,GG,AA,AA,GG,GG,GG,GG,GG,GG,GG,--,GG,GG,GG,GG,AA,AA,GG,AA,AA,GG,GG,GG,GG,GG
L008,AG,GG,AG,GG,--,GG,GG,--,AA,GG,GG,AG,GG,AG,GG,GG,GG,GG,AG,--,AA,AG,GG,GG,AA,GG,GG
L009,CC,CC,CC,TC,CC,TC,CC,CC,TC,CC,TC,CC,CC,CC,CC,CC,TC,CC,TC,CC,CC,TC,CC,CC,TC,CC
L010,CC,--,TC,--,CC,CC,TT,TC,TC,TC,TC,TT,--,TC,CC,--,TT,CC,--,TC,CC,TT,--,CC,TC
L011,AA,AC,AA,AA,AA,AA,--,AA,AA,AA,AA,CC,AA,AC,AA,AA,AA,AA,AA,AA,AA,AA,CC,AA,AA,AA,CC
L012,GG,--,GG,AA,AA,AA,AG,--,AA,--,--,AA,GG,--,GG,GG,AA,AA,GG,--,AG,--,AA,AA,--,GG,--,AG
L013,AA,GG,GG,GG,GG,GG,GG,AA,GG,AA,GG,GG,AA,GG,GG,GG,GG,GG,AA,GG,GG,AA,GG,GG,AA,AA
L014,CC,CC,TT,CC,CC,TT,CC,CC,TT,CC,CC,CC,TT,CC,CC,TT,CC,CC,TT,CC,TT,CC,CC,TT,CC,CC,TT
L015,TT,GG,TT,GG,GG,TG,GG,GG,TT,GG,GG,--,TT,TG,TT,TT,GG,--,--,GG,TT,GG,TG,TT,GG,GG,TG
L016,AG,AG,AG,AA,AG,AG,AA,--,AA,AG,AG,AG,AA,AG,AG,AG,AA,AG,AG,AG,AG,AG,AA,AA,AG,AG
L017,CC,AA,AA,AA,AA,AA,AA,CC,AA,AA,AA,AA,AA,AA,AA,AA,AA,AA,AA,AA,AA,AA,AA,AA,AA,CC

```

Figure 8 Formats of Input SNP Data

#### (2) Preparation of SSR or STR Data

- The software supports SSR allelic data in the general format of positive integers, which represent the quantity of bases in the section. It only supports species with diploid chromosome. The actual genotype data are shown in Figure 9. Determine if it is a hybrid locus according to the value of base difference set by the user. If the value is set as larger than 0bp, then 322/323 shall be heterozygous type and 322/322 homozygous type. If the value is set as larger than 1bp, then 322/325 shall be heterozygous type and 322/323 or 322/322 homozygous type.
- In the data structure of Figure 9, rows represent samples and columns represent loci. The first row shows the serial number of loci and the other columns the corresponding genotype data of samples. The first column shows the serial number of samples and the other rows the corresponding genotype data of loci.
- The file format of the input data is .csv and the missing data are indicated with /, ?/?, \*/\*, -/- or ./.. The example of input data is shown in Figure 9.
- Before opening the data file, please close all Excel programs which open the same data file.



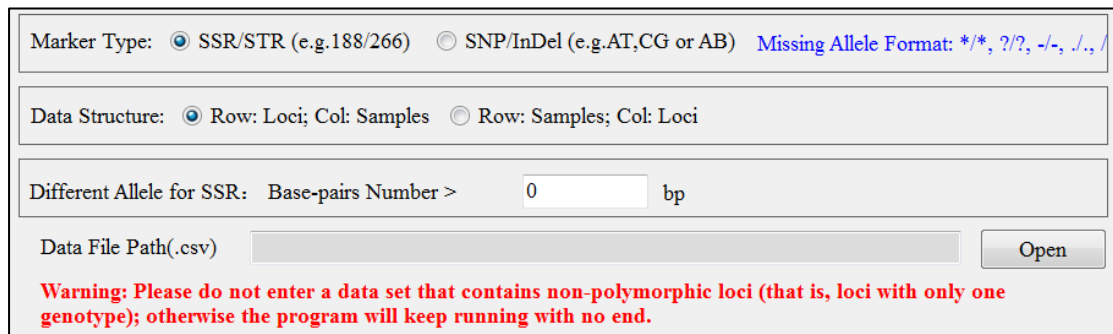
Sample_id	L01	L02	L03	L04	L05	L06	L07	L08	L09	L10	L11	L12	L13	L14	L15	L16	L17	L18	L19	L20	L21	L22	L23	L24	L25
S001	323/325	282/282	353/353	288/288	341/341	410/410	382/382	319/319	290/290	183/183	269/269	206/206	169/169	2											
S002	322/322	255/255	348/348	292/292	336/336	410/410	364/364	273/273	252/252	173/173	299/299	212/212	173/173												
S003	354/354	240/240	270/270	290/290	362/362	410/430	380/380	275/275	248/248	183/183	276/276	206/206	154/154												
S004	335/335	273/273	360/360	317/317	341/341	301/301	260/260	173/173	305/305	206/206	169/169	233/233	413/413												
S005	350/350	255/255	360/360	335/335	336/336	410/410	364/364	290/290	201/201	280/280	206/206	150/150	229/229												
S006	350/350	255/255	360/360	336/336	410/410	279/279	290/290	201/201	206/206	150/150	229/229	393/393	284/284												
S007	250/250	360/360	290/290	362/362	410/410	382/382	319/319	290/290	183/183	265/265	206/206	173/173	237/237												
S008	352/352	248/248	360/360	292/292	362/362	410/410	382/382	319/319	290/290	183/183	265/265	206/206	173/173												
S009	352/352	248/248	360/360	317/317	341/341	410/410	382/382	301/301	290/290	183/183	265/265	206/206	173/173												
S010	322/322	252/252	255/255	348/348	292/292	336/336	410/410	364/364	273/273	252/252	173/173	299/299	212/212												
S011	350/350	248/248	255/255	386/386	317/317	341/341	410/410	404/404	301/301	262/262	185/185	265/265	206/206												
S012	350/350	250/250	360/360	292/292	362/362	410/410	382/382	319/319	290/290	183/183	265/265	212/212	173/173												
S013	352/352	240/240	246/246	360/360	290/290	362/362	410/410	364/364	319/319	252/252	183/183	265/265	206/206												

Figure 9 Formats of Input SSR Data

### (3) Parameter Selection

- As shown in Figure 10, select in Marker Type section the corresponding molecular marker type of the training data (for CLI version, modify the corresponding line of **#2.Marker Type** in the parameter setting file). For example, select **SNP/InDel** (e.g.AT, CG or AB) when the molecular marker type is SNP as shown in Figure 8, and **SSR/STR** (e.g.188/266) when the type is SSR as shown in Figure 9.
- As shown in Figure 10, select in Data Structure section the corresponding data structure of the training data (for CLI version, modify the corresponding line of **#3. Data Structure** in the parameter setting file). For example, select **Row:Loci Col:Samples** when the data structure is shown as in Figure 8, where rows represent loci and columns represent samples, and select **Row: Samples Col:Loci** when the data structure is shown as in Figure 9, where rows represent samples and columns represent loci.
- When selecting **SSR/STR** (e.g.188/266) for Marker Type section, we need to set the threshold for the difference in the number of base pairs in Different Allele for SSR section as shown in Figure 10 (for CLI version, modify the corresponding line of **#4. Different Allele for SSR: Base-pairs Number > n bp** in the parameter setting file). For example, when the default value is set as  $>0$  bp, it means there is difference between the two loci with genotype 200/200 and 201/201 respectively, but there is no difference between the two loci with genotype 200/200 and 200/200. When the value is set as  $>1$ bp, it means there is difference between the two loci with genotype 200/200 and 202/202 respectively, but there is no difference between the two loci with genotype 200/201 and 201/202, or 200/200 and 201/201. The rest can be deduced in the same manner.
- After setting the above parameter, click Open in Data File Path section and select the corresponding training data file (for CLI version, modify the corresponding line of **#5.Data File Path** in the parameter setting file). Only the .csv format is supported for now. It becomes impossible to change the parameters mentioned in step a) and b) when such training data file has been selected. If you would like to reset the two parameters, you have to relaunch LociScan, or click Cancel in the file selection dialogue box triggered by the Open button.
- Please note that it is recommended to use Data Cleaning function of LociScan to carry out the following content verification before introducing the training data file into LociScan for loci combination screening: loci without polymorphism must be ruled out of the dataset, that is to say, loci of different samples with exactly the same genotype must be deleted from the dataset; otherwise the program will get stuck in an infinite

dead loop when the loci combination screening function is operated.



Marker Type: ☒ SSR/STR (e.g.188/266) ☐ SNP/InDel (e.g.AT,CG or AB) Missing Allele Format: \*/\*, ?/? , -/-, ./., /

Data Structure: ☒ Row: Loci; Col: Samples ☐ Row: Samples; Col: Loci

Different Allele for SSR: Base-pairs Number >  bp

Data File Path(.csv)

**Warning: Please do not enter a data set that contains non-polymorphic loci (that is, loci with only one genotype); otherwise the program will keep running with no end.**

Figure 10 Parameters to be Set When Inputting Training Data

### 3.4 Input of Loci Combinations to be Evaluated

#### (1) Data Preparation for Loci Combinations to be Evaluated

- a) The software supports the following loci combination format: each row represents one loci combination and different combinations are separated by line break. The first column shows the serial number of the loci combination and the other columns the names of loci in the combination, with different loci separated by comma. As shown in Figure 11, SNP-LC001 is the serial number of the loci combination, and L056, L081, etc. are the serial number of loci.

```
SNP-LC001,L056,L081,L105,L116,L126,L134,L150,L162,L164,L300
SNP-LC002,L002,L098,L158,L210,L230,L228,L243
SNP-LC003,L034,L068,L123,L159,L179
SSR-LC001,L04,L12,L13,L22,L27,L34,L37
SSR-LC002,L03,L06,L18,L20,L23,L40
SSR-LC003,L01,L04,L08,L15,
```

Figure 11 Data Format of Loci Combinations to be Evaluated

- b) The format of the file with loci combinations to be evaluated shall be .csv, editable with text editors or Excel. Unnecessary separation signs in the file will be automatically neglected by the software as shown in Figure 12.

```
SNP-LC001,L056,L081,L105,L116,L126,L134,L150,L162,L164,L300
SNP-LC002,L002,L098,L158,L210,L230,L228,L243,,
SNP-LC003,L034,L068,L123,L159,L179,,,,
SSR-LC001,L04,L12,L13,L22,L27,L34,L37,,,
SSR-LC002,L03,L06,L18,L20,L23,L40,,,,
SSR-LC003,L01,L04,L08,L15,,,,,
SSR-LC001,L04,L12,L13,,,,,,
```

Figure 12 Data Format of Loci Combinations to be Evaluated

- c) Please close all the programs which open the same file before opening the data file.
- d) Please note that the names of loci must be identical with those in the training data of section 3.3.

#### (2) Parameter Setting

When entering the loci combination evaluation page by clicking **Loci combination evaluation** in Operation section, a parameter setting window shall appear as shown in Figure 13. Please refer to Section 3.3 for the input and parameter setting of the training data and prepare loci

combination file as mentioned above. Then click Open in Loci File Path section to select the corresponding file (for CLI version please modify the corresponding line of #6.Loci File Path in the parameter setting file).

Marker Type: <input checked="" type="radio"/> SSR/STR (e.g.188/266) <input type="radio"/> SNP/InDel (e.g.AT,CG or AB) Missing Allele Format: */*, ?/? , -/-, ./., /	
Data Structure: <input checked="" type="radio"/> Row: Loci; Col: Samples <input type="radio"/> Row: Samples; Col: Loci	
Different Allele for SSR: Base-pairs Number > <input type="text" value="0"/> bp	
Data File Path(.csv)	<input type="text"/> <input type="button" value="Open"/>
Loci File Path(.csv)	<input type="text"/> <input type="button" value="Open"/>

Figure 13 Parameter Setting Area for Loci Combination Evaluation

### 3.5 Output of Loci Combination Screening Results

The loci combination screening results will be output in .csv file, in the same file directory of the main program for the GUI version LociScan and in the file directory set by the user for the CLI version LociScan (please modify the corresponding line of #7.Output Path in the parameter setting file). The name of the file starts with 'LociScan' and is composed of the original name of the training data and the ending time of data analysis. The results file can be opened by text editors or Excel.

Figure 14 shows the file of loci combination screening results when using GA and selecting All Generation in Result Format section. The rows represent the results of populations with different number of evolved generations. Each line is composed of the following content: the serial number of evolved generation of the loci combination in the first column, the smallest fitness function value in the population of that generation in the second column, the largest fitness function value in the population of that generation in the third column and the names of the loci in the combination with the largest fitness function from the fourth column to the end of the row.

```
1,0.137931034482759,0.689655172413793,L044,L065,L131
2,0.103448275862069,0.689655172413793,L044,L065,L131
3,0.172413793103448,0.689655172413793,L044,L065,L131
4,0.206896551724138,0.689655172413793,L044,L065,L131
5,0.206896551724138,0.689655172413793,L044,L065,L131
6,0.206896551724138,0.689655172413793,L044,L065,L131
7,0.206896551724138,0.689655172413793,L044,L065,L131
8,0.172413793103448,0.689655172413793,L044,L065,L131
9,0.275862068965517,0.689655172413793,L044,L065,L131
10,0.241379310344828,0.689655172413793,L044,L065,L131
```

Figure 14 Loci Combination Screening Results in All Generation format Output by GA

Figure 15 shows the file of loci combination screening results when using GA and selecting Last Generation in Result Format section. The rows represent the results of different times of calculation. Each line is composed of the following content: the automatic serial number of calculation times of the loci combination in the first column, the smallest fitness function value in the first generation population of that time of calculation in the second column, the largest fitness function value in the last generation population in the third column and the names of the loci in the combination with the largest fitness function in the last generation population from

the fourth column to the end of the row.

```
R000G002,0.206896551724138,0.310344827586207,L065,L178
R000G003,0.379310344827586,0.689655172413793,L174,L178,L263
R000G004,0.586206896551724,1,L019,L059,L158,L185
R000G005,0.655172413793103,1,L062,L130,L151,L158,L262
R000G006,0.689655172413793,1,L119,L122,L169,L170,L239,L281
R000G007,0.793103448275862,1,L098,L102,L109,L151,L167,L178,L286
R000G008,0.793103448275862,1,L019,L110,L117,L163,L169,L186,L219,L298
R000G009,0.862068965517241,1,L022,L050,L092,L130,L161,L169,L245,L250,L297
R000G010,0.862068965517241,1,L016,L032,L075,L110,L127,L153,L198,L214,L234,L287
```

Figure 15 Loci Combination Screening Results in Last Generation format Output by GA

Figure 16 shows the file of loci combination screening results when using EA. It is composed of the following content: the serial number of the loci combination in the first column, the fitness function value in the second column, and the names of the loci in the combination from the third column to the end of the row. The first row represents the loci combination with the smallest fitness function and the second the loci combination with the largest fitness function.

```
R000G003-worst,0.0344827586206897,L056,L047,L040
R000G003-best,0.758620689655172,L158,L130,L019
```

Figure 16 Loci Combination Screening Results Output by Exhaustive Algorithm

The log of loci combination screening is output in .log file in the same directory with the main program for all version of LociScan. The name of the file is 'Stopwatch.log' and it can be opened with text editors. When carrying out loci combination screening with LociScan, the software will add one record into that log file for each time of calculation, as shown in Figure 17. The first column is the file name of the loci combination screening results acquired by that time of calculation, the second column is the time of calculation (unit: second), the third column is the fitness function adopted for the calculation and the fourth column is the model parameters set for the calculation. Taking GA as example, the model parameters shall be Crossover Rate, Mutation Rate, Population Size, Generation Size, Keep Best in Generation, Mutate Method and GA Method.

```
\LociScan_R000G002_GA_SNP data(row-loci col-sample) 20230403_163231.csv 0.617s R-VDP 0.90_0.01_100_2000_KB_wholeRandom_normal-GA
\LociScan_R000G003_GA_SNP data(row-loci col-sample) 20230403_163232.csv 0.992s R-VDP 0.90_0.01_100_2000_KB_wholeRandom_normal-GA
\LociScan_R000G004_GA_SNP data(row-loci col-sample) 20230403_163233.csv 1.243s R-VDP 0.90_0.01_100_2000_KB_wholeRandom_normal-GA
\LociScan_R000G005_GA_SNP data(row-loci col-sample) 20230403_163234.csv 1.318s R-VDP 0.90_0.01_100_2000_KB_wholeRandom_normal-GA
\LociScan_R000G006_GA_SNP data(row-loci col-sample) 20230403_163235.csv 1.564s R-VDP 0.90_0.01_100_2000_KB_wholeRandom_normal-GA
\LociScan_R000G007_GA_SNP data(row-loci col-sample) 20230403_163237.csv 1.598s R-VDP 0.90_0.01_100_2000_KB_wholeRandom_normal-GA
\LociScan_R000G008_GA_SNP data(row-loci col-sample) 20230403_163238.csv 1.662s R-VDP 0.90_0.01_100_2000_KB_wholeRandom_normal-GA
\LociScan_R000G009_GA_SNP data(row-loci col-sample) 20230403_163240.csv 1.714s R-VDP 0.90_0.01_100_2000_KB_wholeRandom_normal-GA
\LociScan_R000G010_GA_SNP data(row-loci col-sample) 20230403_163242.csv 1.784s R-VDP 0.90_0.01_100_2000_KB_wholeRandom_normal-GA
\LociScan_R000G003_GA_SNP data(row-loci col-sample) 20230404_085840.csv 0.9692s R-VDP 0.90_0.01_100_2000_KB_wholeRandom_normal-GA
```

Figure 17 Log of Loci Combination Screening

### 3.6 Output of Loci Combination Evaluation Results

The loci combination evaluation results will be output in .csv file, in the same file directory of the main program for the GUI version LociScan and in the file directory set by the user for the CLI version LociScan (please modify the corresponding line of #7.Output Path in the parameter setting file). The name of the file starts with 'LociCheck' and is composed of the original name of the training data and the ending time of data analysis. The results file can be opened by text editors or Excel.

As shown in Figure 18, the loci combination evaluation file is composed of the following content: the serial number of the loci combination in the first column, the fitness function value in the

second column, and the names of the loci in the samples which cannot be identified by the combination from the third column to the end of the row. Different rows represent the evaluation results of different loci combinations. When the second column is 1, it means the loci combination can discriminate all the samples. When the second column is -2, it means the names of the loci in the combination are not identical with those of the loci in the training data and the loci combination evaluation has failed.

SNP-LC001,0.0344827586206897,S01,S02,S03,S04,S05,S06,S07,S08,S09,S10,S11,S12,S13,S14,S15,
SNP-LC002,1
SNP-LC003,0.482758620689655,S01,S03,S04,S05,S06,S07,S10,S11,S12,S14,S17,S18,S20,S23,S27
SSR-LC001,-2
SSR-LC002,-2
SSR-LC003,-2

Figure 18 Loci Combination Evaluation File

### 3.7 Output of Data Cleaning Results

The data cleaning results will be output in .csv file, in the same file directory of the main program for the GUI version LociScan and in the file directory set by the user for the CLI version LociScan (please modify the corresponding line of #7.Output Path in the parameter setting file). The name of the file starts with 'DataClean' and is composed of the original name of the training data and the ending time of data analysis. The results file can be opened by text editors or Excel. The data structure of the file is kept the same as that of the original file and the loci of different samples with no genotype difference will be deleted, with only polymorphism loci and their corresponding genotype data left.

### 3.8 Main Function Modules

#### 3.8.1 Loci Combination Screening

Loci combination screening can screen out loci combinations which satisfy the plant variety discrimination demands out of a massive pool in a quick and simple way, thus providing scientific and accurate basis for the optimization of both quality and quantity of molecular marker loci combinations and further technical support for applying molecular markers in plant variety discrimination.

The screening model adopts GA and EA. GA screens out effective combinations through parameters including hybridization rate, mutation rate, population size, number of evolved generations, etc. and shows in the result the VDP and loci serial numbers of each combination. EA searches for the real best solution through traversing all the loci combinations and shows in the result the VDP and loci serial numbers of both the best and worst combinations.

#### 3.8.2 Loci Combination Evaluation

Loci combination evaluation refers to the process of evaluating the VDP of the loci combinations input by the user with the genotype data of the training samples and listing out the samples which cannot be discriminated.

#### 3.8.2Data Cleaning

Data cleaning refers to the process of checking the polymorphism of all loci with the genotype data of the training samples and then deleting the loci with no genotype difference among all the samples and their corresponding genotype data.

### 3.9 Operating Procedures

#### 3.9.1 Loci Combination Screening - GA

Loci combination screening based on GA refers to the process of using GA to screen out the most

effective loci combinations with specified loci number. The 8 kinds of loci combination appraisal indices targeting at plant variety discrimination are used as the fitness functions and the output results shall include the corresponding fitness function values of the loci combinations. Only .csv file is supported for the training data (input data) at present.

### 3.9.1.1 GA Operation in GUI Version

Loci combination screening based on GA is operated in GUI version LociScan as follows:

- (1) Select **Loci combination screening** in the Operation section.
- (2) Select the corresponding molecular marker type for the training data in Marker Type section as shown in Figure 10, i.e. **SNP/InDel** (e.g.AT, CG or AB) , the default value, or **SSR/STR** (e.g.188/266). See section 3.3 for the details about the format of the training data.
- (3) Select the data structure of the training data in Data Structure section as shown in Figure 10, where the default value **Row:Loci; Col:Samples** means that the rows represent loci and column samples, and **Row: Samples; Col:Loci** means the opposite. See section 3.3 for the details about the data structure of the training data.
- (4) Click Open button in Data File Path section as shown in Figure 10, select the data file to be analyzed in the popup dialog box 'Open File' and then click Open button to read the data file.
- (5) Select the module algorithm in TabPages Section as shown in Figure 19, i.e. **Genetic Algorithm** or **Exhaustive Algorithm**. Please select **Genetic Algorithm**.
- (6) In TabPages with Genetic Algorithm (Figure 19), set the value of Crossover Rate between 0 and 1, which represents the rate of the individual carrying out crossover operator during the evolution of the next generation population. The default value is 0.90.
- (7) In TabPages with Genetic Algorithm (Figure 19), set the value of Mutation Rate between 0 and 1, which represents the rate of the individual carrying out mutation operator during the evolution of the next generation population. The default value is 0.01.
- (8) In TabPages with Genetic Algorithm (Figure 19), set the value of Population Size as a positive integer equaling or larger than 3, which represents the size of the population divided. The default value is 100 and the number of divided populations shall be determined according to the data size. The more populations there are, the better the result of searching for the best combinations is, but the longer the calculation time is.
- (9) In TabPages with Genetic Algorithm (Figure 19), set the value of Generation Size as a positive integer equaling or larger than 1, which represents the number of generations evolved. The default value is 2000 and the more generations there are, the better the result of searching for the best combinations is, but the longer the calculation time is.
- (10) In TabPages with Genetic Algorithm (Figure 19), tick or untick the box **Keep Best in Generation** to confirm if the best individual of that generation is kept for the next generation or not. The box is ticked by default, meaning that the best individual of that generation will be kept for the next generation.
- (11) In TabPages with Genetic Algorithm (Figure 19), tick or untick the box **Stop When FF = 1** to confirm if the program shall automatically stop searching when loci combination with fitness function value (FF) of 1 has been reached. The box is not ticked by default, meaning that the program shall not stop searching in that case.
- (12) In TabPages with Genetic Algorithm (Figure 19), set the value of Mutate Method to select the mutation method of the mutation operator of GA, i.e. **Whole Random**, the default value,



or Half Random.

- (13) In TabPages with Genetic Algorithm (Figure 19), set the value of GA Method, i.e. normal-GA or self-adaption-GA. Normal-GA, the default value, is described in the References. Self-Adaption-GA is an optimized algorithm in which the parameters of genetic operators and the calculation methods are dynamically adjusted during the solving process.
- (14) In Light Green color section as shown in Figure 20, set the value of Fitness Function, which represents the VDP method, including R-VDP, the default value, C-VDP, P-VDP, TDP, R-VDP-reciprocal, C-VDP-reciprocal, TDP-reciprocal and O-VDP. Please refer to the references about VDP for the detailed statistical methods of R-VDP [1], C-VDP, P-VDP and TDP, and those about LociScan for the detailed methods of O-VDP. For these five functions, we search for the largest value of their respective statistical method, and for the other three functions (i.e. R-VDP-reciprocal), the smallest value.
- (15) In Light Green color section as shown in Figure 20, set the method and threshold value of Variety Threshold Setting, which allows the user to define the loci difference threshold of the same variety, i.e. Ratio of Different Loci or Number of Different Loci, the default value. The former defines samples as the same variety when the ratio of different loci is smaller than the threshold value set by the user. The value is set as a decimal between 0 and 1 (including 1) and the default value is 0.05, which means that the samples will be identified as the same variety when the proportion of different loci out of all loci is less than 5%. The later defines samples as the same variety when the number of different loci is less than the threshold value set by the user. The value is set as a positive integer larger than 1 and the default value is 1, which means that the samples will be identified as the same variety when the number of different loci is less than 1.
- (16) In Aqua color section as shown in Figure 20, set the method and value of Number of Target Loci, which defines the number of loci included in the target loci combinations, i.e. Fixed Genome Size, the default value, or Batch Genome Size. The two values represent the number of loci in the target loci combinations and the number interval of loci in the target loci combinations respectively. The former means that a single round of calculation will be carried out as per one fixed number of loci. The value shall be a positive integer equaling or larger than 2 and the default value is 3, which means that target loci combinations with 3 loci will be screened out. The later means that a single round of calculation will be carried out as per multiple loci numbers gradually increased one by one within a certain number interval. The value shall be a positive integer equaling or larger than 2 and the default value is 2-10, which means the target loci combinations with 2-10 loci will be screened out.
- (17) In Result Format section as shown in Figure 21, set the format of the output content of the loci combination screening results based on GA, i.e. All Generation or Last Generation, the default value. See section 3.5 for the details about the output format of the loci combination screening results. It is recommended to set the value as Last Generation when the value of Generation Size set by the user the relatively large or multi rounds of calculation are set.
- (18) In Blue color section as shown in Figure 20, set the value of Times of Calculation Cycles, which allows repeated calculation under the same parameter values set in step (1)-(17). The default value is 1, which means that only one round of calculation shall be carried out under the above-mentioned values of parameters for loci combination screening.
- (19) In File Format section as shown in Figure 21, set the format of the output file of the loci

combination screening results, i.e. **Unit Results** or **Merged Results**, the default value. The former outputs the screening results of multi rounds of calculation as several individual files and the later as one merged file. The multi rounds of calculation will be triggered if **Batch Genome Size** is selected as the method of Number of Target Loci, or if the value of Cycle calculation times is set as larger than 1. In such case, **Merged Results** is recommended.

- (20) Click OK button and wait until LociScan pops up success or failure messages on the main interface after backstage calculation (Figure 22), with the calculation time used for loci combination screening in case of success and none in the other case.
- (21) Click **Open Result Folder** to check the analysis results in the corresponding "LociScan\_GA\_CSVFileName\_CreateDate\_CreateTime.csv" file. See section 3.5 for the details about the output format of the loci combination screening results.

Genetic Algorithm Exhaustive Algorithm

Crossover Rate  $[0, 1] =$  0.90

Mutation Rate  $[0, 1] =$  0.01

Population Size  $[3, \infty) =$  100

Generation Size  $[1, \infty) =$  2000

☒ Keep Best in Generation ☐ Stop when FF=1

Mutate Method: WholeRandom

GA Method: normal-GA

Figure 19 Parameter Setting of GA Model

Fitness Function (FF): R-VDP

Variety Threshold Setting:

☐ Ratio of Different Loci  $(0,1] <$  0.05

☒ Number of Different Loci  $[1,\infty) <$  1

Number of Target Loci:

☒ Fixed Genome Size  $[2, \infty) =$  3

☐ Batch Genome Size  $[2, \infty)$  from 2 to 10

Times of Calculation Cycles: 1

Figure 20 Parameter Setting of Fitness Function



Result Format:	<input type="radio"/> All Generation	<input checked="" type="radio"/> Last Generation
File Format:	<input type="radio"/> Unit Results	<input checked="" type="radio"/> Merged Results

Figure 21 Output Format Setting of Loci Combination Screening Results

Copyright©2023 BAAFS.MRC, Designer: YANGYANG, E-mail: caurwx@163.com Run done! Time: 0h0m0s969ms

Figure 22 Success or Failure Message of GUI Version LociScan

### 3.9.1.2 GA Operation in CLI Version

Loci combination screening based on GA is operated in CLI version LociScan as follows:

- (1) Open the parameter setting file GASetting.txt. Note that the annotation lines and the parameter lines in the file must not be deleted and shall only be modified according to the following steps.
- (2) Modify the corresponding line of the parameter **#1.Operation** as 1. 1 represents Loci combination screening, 2 Loci combination evaluation, and 3 Data Cleaning.
- (3) Modify the corresponding line of the parameter **#2.Marker Type** as 1 or 2. 1 represents that the molecular marker type of the training data is SSR/STR, and 2 SNP/InDel. See section 3.3 for the details about the format of the training data.
- (4) Modify the corresponding line of the parameter **#3.Data Structure** as 1 or 2. 1 means that the rows represent loci and the columns samples and 2 means the opposite. See section 3.3 for the details about the structure of the training data.
- (5) Modify the corresponding line of the parameter **#4.Different Allele for SSR: Base-pairs Number > n bp** as 0 or a positive integer. The parameter needs to be modified only when the value of the corresponding line of the parameter **#2.Marker Type** is 1. See section 3.3 for the details.
- (6) Input into the corresponding line of the parameter **#5.Data File Path (.csv)** the path of the genotype data file to be analyzed, including both the file name and extension.
- (7) Input NULL into the corresponding line of the parameter **#6.Loci File Path (.csv)**.
- (8) Input into the corresponding line of the parameter **#7.Output Path** the storage path of the result file, with no need to mention the file name and extension.
- (9) Modify the corresponding line of the parameter **#8.Algorithm** as 1. 1 represents Genetic Algorithm and 2 Exhaustive Algorithm.
- (10) Modify the corresponding line of the parameter **#9.Crossover Rate = n** as a decimal between 0 and 1 (including 0 and 1), which represent the rate of the individual carrying out crossover operator during the evolution of the next generation population. The default value is 0.90.
- (11) Modify the corresponding line of the parameter **#10.Mutation Rate = n** as a decimal between 0 and 1 (including 0 and 1), which represents the rate of the individual carrying out mutation operator during the evolution of the next generation population. The default value is 0.01.
- (12) Modify the corresponding line of the parameter **#11.Population Size = n** as a positive integer equaling or larger than 3, which represents the number of individuals included in the population of each generation. The default value is 100 and the number of divided populations shall be determined according to the data size. The more populations there are,

the better the result of searching for the best combinations is, but the longer the calculation time is.

- (13) Modify the corresponding line of the parameter **#12.Generation Size = n** as a positive integer equaling or larger than 1, which represents the number of generations evolved. The default value is 2000 and the more generations there are, the better the result of searching for the best combinations is, but the longer the calculation time is.
- (14) Modify the corresponding line of the parameter **#13.Keep Best in Generation** as true or false to confirm if the best individual of that generation is kept for the next generation or not. The default value is true, meaning that the best individual of that generation will be kept for the next generation.
- (15) Modify the corresponding line of the parameter **#14.Stop running when FF = 1** as true or false to confirm if the program shall automatically stop searching when loci combination with fitness function value (FF) of 1 has been reached. The default value is false, meaning that the program shall not stop searching in that case.
- (16) Modify the corresponding line of the parameter **#15.Mutate Method** as 0 or 1 to select the mutation method of the mutation operator of GA. 0 represents **Half Random** and 1 **Whole Random**. The default value is 1.
- (17) Modify the corresponding line of the parameter **#16.GAMethod** as 0 or 1. 0 represents **normal-GA** and 1 **self-adaption-GA**. Normal-GA is described in the References. Self-Adaption-GA is an optimized algorithm in which the parameters of genetic operators and the calculation methods are dynamically adjusted during the solving process. The default value is 0.
- (18) Modify the corresponding line of the parameter **#17.Fitness Function** as a number between 1 and 8 to set the fitness function used for screening loci combinations. 1 represents **R-VDP**, 2 **C-VDP**, 3 **P-VDP**, 4 **TDP**, 5 **R-VDP-reciprocal**, 6 **C-VDP-reciprocal**, 7 **TDP-reciprocal** and 8 **O-VDP**. The default value is 1. Please refer to the references about VDP for the detailed statistical methods of **R-VDP** [1], **C-VDP**, **P-VDP** and **TDP**, and those about LociScan for the detailed methods of **O-VDP**. For these five functions, we search for the largest value of their respective statistical method, and for the other three functions, the smallest value.
- (19) Modify the corresponding line of the parameter **#18.Variety Threshold Setting Type** as 1 or 2, which allows the user to define the extent of tolerance of different loci in one variety. 1 represents **Ratio of Different Loci** and 2 **Number of Different Loci**. The default value is 2. 1 defines samples as the same variety when the ratio of different loci is smaller than the threshold value set by the user and the corresponding line of the parameter **#19.Ratio of Different Loci < n** shall be modified accordingly. The value is set as a decimal between 0 and 1 (including 1) and the default value is 0.05, which means that the samples will be identified as the same variety when the proportion of different loci out of all loci is less than 5%. 2 defines samples as the same variety when the number of different loci is less than the threshold value set by the user and the corresponding line of the parameter **#20.Number of Different Loci < n** shall be modified accordingly. The value is set as a positive integer larger than 1 and the default value is 1, which means that the samples will be identified as the same variety when the number of different loci is less than 1.
- (20) Modify the corresponding line of the parameter **#21.Analysis method for Number of Target Loci** as 1 or 2, which defines the number of loci included in the target loci combinations. 1

represents Fixed Genome Size and 2 Batch Genome Size. The default value is 1. 1 means that a single round of calculation will be carried out as per one fixed number of loci and the corresponding line of the parameter #22.Fixed Genome Size = n shall be modified accordingly. The value shall be a positive integer equaling or larger than 2 and the default value is 3, which means that target loci combinations with 3 loci will be screened out. 2 means that a single round of calculation will be carried out as per multiple loci numbers gradually increased one by one within a certain number interval and the corresponding lines of the parameters #23.Batch Genome Size from n and #24.Batch Genome Size to m shall be modified accordingly. The value shall be a positive integer equaling or larger than 2 and the default value is 2-10, which means the target loci combinations with 2-10 loci will be screened out.

- (21) Modify the corresponding line of the parameter #25. Times of Calculation Cycles = n as a positive integer, which allows repeated calculation under the same parameter values set in step (2)-(20). The default value is 1, which means that only one round of calculation shall be carried out under the above-mentioned values of parameters for loci combination screening. When the value is 2, it means that two rounds of screening shall be carried out and so on.
- (22) Modify the corresponding line of the parameter #26.Result Format as a number between 1 and 4 to set the format of the output file of the loci combination screening results. 1 represents Unit Results with All Generation, 2 Unit Results with Last Generation, 3 Merged Results with All Generation and 4 Merged Results with Last Generation. 1 and 2 output the screening results of multi rounds of calculation as several individual files, and 3 and 4 as one merged file. 1 and 3 output the best loci combination of the populations of each generation, and 2 and 4 the best loci combination of only the populations of the last generation. The default value is 4. The multi rounds of calculation will be triggered if Batch Genome Size is selected for #21.Analysis method for Number of Target Loci, or if the value of #25. Times of Calculation Cycles = n is set as larger than 1. In such case, 2 and 4 are recommended.
- (23) Save the parameter setting file GASetting.txt.
- (24) Run the command mono LociScan-C.exe GASetting.txt and wait until LociScan pops up success or failure messages on the command line interface after backstage calculation (Figure 23), with the calculation time used for loci combination screening in case of success and none in the other case.
- (25) Switch to the results file directory set in #7.Output Path to check the analysis results in the corresponding "LociScan\_GA\_CSVFileName\_CreateDate \_CreateTime.csv" file. See section 3.5 for the details about the output format of the loci combination screening results.

```

root@localhost:~/Documents/LociScan
File Edit View Search Terminal Help
[root@localhost LociScan]# pwd
/root/Documents/LociScan
[root@localhost LociScan]# mono LociScan-C.exe GASetting.txt
Run done! Time: 0h0m11s27ms
[root@localhost LociScan]#

```

Figure 23 Success or Failure Message of CLI Version LociScan

### 3.9.2 Loci Combination Screening - EA

Loci combination screening - EA refers to the process of using EA to screen out the most effective

loci combinations with specified loci number. The 8 kinds of loci combination appraisal indices targeting at plant variety discrimination are used as the fitness functions and the output results shall include the corresponding fitness function value of the loci combination. Only .csv file is supported for the training data (input data) at present.

### 3.9.2.1 EA Operation in GUI Version

Loci combination screening based on EA is operated in GUI version LociScan as follows:

- (1) Select Loci combination screening in the Operation section.
- (2) Select the corresponding molecular marker type for the training data in Marker Type section as shown in Figure 10, i.e. SNP/InDel (e.g.AT, CG or AB) , the default value, or SSR/STR (e.g.188/266). See section 3.3 for the details about the format of the training data.
- (3) Select the data structure of the training data in Data Structure section as shown in Figure 10, where the default value Row:Loci; Col:Samples means that the rows represent loci and column samples, and Row: Samples; Col:Loci means the opposite. See section 3.3 for the details about the data structure of the training data.
- (4) Click Open button in Data File Path section as shown in Figure 10, select the data file to be analyzed in the popup dialog box Open File and then click Open button to read the data file.
- (5) Select the module algorithm in TabPages Section as shown in Figure 24, i.e. Genetic Algorithm or Exhaustive Algorithm. Please select Exhaustive Algorithm. Please note: the algorithm requires no parameter setting. It may occupy a large amount of computing resource, so it is necessary to keep a close eye on the computing resource and see if there is overflow after exhaustion of resource by the program. When program exception has occurred, please try to lower the value of the parameter Number of Target Loci.
- (6) In Light Green color section as shown in Figure 20, set the value of Fitness Function, which represents the VDP method, including R-VDP, the default value, C-VDP, P-VDP, TDP, R-VDP-reciprocal, C-VDP-reciprocal, TDP-reciprocal and O-VDP. Please refer to the references about VDP for the detailed statistical methods of R-VDP [1], C-VDP, P-VDP and TDP, and those about LociScan for the detailed methods of O-VDP. For these five functions, we search for the largest value of their respective statistical method, and for the other three functions, the smallest value.
- (7) In Light Green color section as shown in Figure 20, set the method and threshold value of Variety Threshold Setting, which allows the user to define the loci difference threshold of the same variety, i.e. Ratio of Different Loci or Number of Different Loci, the default value. The former defines samples as the same variety when the ratio of different loci is smaller than the threshold value set by the user. The value is set as a decimal between 0 and 1 (including 1) and the default value is 0.05, which means that the samples will be identified as the same variety when the proportion of different loci out of all loci is less than 5%. The later defines samples as the same variety when the number of different loci is less than the threshold value set by the user. The value is set as a positive integer larger than 1 and the default value is 1, which means that the samples will be identified as the same variety when the number of different loci is less than 1.
- (8) In Aqua color section as shown in Figure 20, set the method and value of Number of Target Loci, which defines the number of loci included in the target loci combinations, i.e. Fixed Genome Size, the default value, or Batch Genome Size. The two values represent the number of loci in the target loci combinations and the number interval of loci in the target

loci combinations respectively. The former means that a single round of calculation will be carried out as per one fixed number of loci. The value shall be a positive integer equaling or larger than 2 and the default value is 3, which means that target loci combinations with 3 loci will be screened out. The later means that a single round of calculation will be carried out as per multiple loci numbers gradually increased one by one within a certain number interval. The value shall be a positive integer equaling or larger than 2 and the default value is 2-10, which means the target loci combinations with 2-10 loci will be screened out.

- (9) In Blue color section as shown in Figure 20, set the value of Cycle calculation times, which allows repeated calculation under the same parameter values set in step (1)-(8). The default value is 1, which means that only one round of calculation shall be carried out under the above-mentioned values of parameters for loci combination screening.
- (10) In File Format section as shown in Figure 21, set the format of the output file of the loci combination screening results, i.e. **Unit Results** or **Merged Results**, the default value. The former outputs the screening results of multi rounds of calculation as several individual files and the later as one merged file. The multi rounds of calculation will be triggered if **Batch Genome Size** is selected as the method of Number of Target Loci, or if the value of Cycle calculation times is set as larger than 1. In such case, **Merged Results** is recommended.
- (11) Click OK button and wait until LociScan pops up success or failure messages on the main interface after backstage calculation (Figure 22), with the calculation time used for loci combination screening in case of success and none in the other case.
- (12) Click Open Result Folder to check the analysis results in the corresponding "LociScan\_EA\_CSVFileName\_CreateDate\_CreateTime.csv" file. See section 3.5 for the details about the output format of the loci combination screening results.

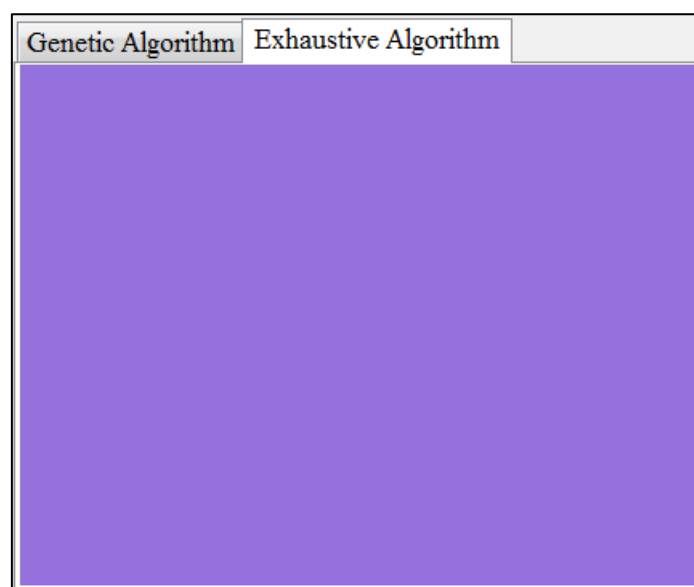


Figure 24 Parameter Setting of EA Model

### 3.9.2.2 EA Operation in CLI Version

Loci combination screening based on EA is operated in CLI version LociScan as follows:

- (1) Open the parameter setting file GASetting.txt. Note that the annotation lines and the parameter lines in the file must not be deleted and shall only be modified according to the following steps.

- (2) Modify the corresponding line of the parameter **#1.Operation** as 1. 1 represents Loci combination screening, 2 Loci combination evaluation, and 3 Data Cleaning.
- (3) Modify the corresponding line of the parameter **#2.Marker Type** as 1 or 2. 1 represents that the molecular marker type of the training data is SSR/STR, and 2 SNP/InDel. See section 3.3 for the details about the format of the training data.
- (4) Modify the corresponding line of the parameter **#3.Data Structure** as 1 or 2. 1 means that the rows represent loci and the columns samples and 2 means the opposite. See section 3.3 for the details about the structure of the training data.
- (5) Modify the corresponding line of the parameter **#4.Different Allele for SSR: Base-pairs Number > n bp** as 0 or a positive integer. The parameter needs to be modified only when the value of the corresponding line of the parameter **#2.Marker Type** is 1. See section 3.3 for the details.
- (6) Input into the corresponding line of the parameter **#5.Data File Path (.csv)** the path of the genotype data file to be analyzed, including both the file name and extension.
- (7) Input NULL into the corresponding line of the parameter **#6.Loci File Path (.csv)**.
- (8) Input into the corresponding line of the parameter **#7.Output Path** the storage path of the result file, with no need to mention the file name and extension.
- (9) Modify the corresponding line of the parameter **#8.Algorithm** as 2. 1 represents Genetic Algorithm and 2 Exhaustive Algorithm.
- (10) Modify the corresponding line of the parameter **#17.Fitness Function** as a number between 1 and 8 to set the fitness function used for screening loci combinations. 1 represents R-VDP, 2 C-VDP, 3 P-VDP, 4 TDP, 5 R-VDP-reciprocal, 6 C-VDP-reciprocal, 7 TDP-reciprocal and 8 O-VDP. The default value is 1. Please refer to the references about VDP for the detailed statistical methods of R-VDP [1], C-VDP, P-VDP and TDP, and those about LociScan for the detailed methods of O-VDP. For these five functions, we search for the largest value of their respective statistical method, and for the other three functions, the smallest value.
- (11) Modify the corresponding line of the parameter **#18.Variety Threshold Setting Type** as 1 or 2, which allows the user to define the extent of tolerance of different loci in one variety. 1 represents Ratio of Different Loci and 2 Number of Different Loci. The default value is 2. 1 defines samples as the same variety when the ratio of different loci is smaller than the threshold value set by the user and the corresponding line of the parameter **#19.Ratio of Different Loci < n** shall be modified accordingly. The value is set as a decimal between 0 and 1 (including 1) and the default value is 0.05, which means that the samples will be identified as the same variety when the proportion of different loci out of all loci is less than 5%. 2 defines samples as the same variety when the number of different loci is less than the threshold value set by the user and the corresponding line of the parameter **#20.Number of Different Loci < n** shall be modified accordingly. The value is set as a positive integer larger than 1 and the default value is 1, which means that the samples will be identified as the same variety when the number of different loci is less than 1.
- (12) Modify the corresponding line of the parameter **#21.Analysis method for Number of Target Loci** as 1 or 2, which defines the number of loci included in the target loci combinations. 1 represents Fixed Genome Size and 2 Batch Genome Size. The default value is 1. 1 means that a single round of calculation will be carried out as per one fixed number of loci and the corresponding line of the parameter **#22.Fixed Genome Size = n** shall be modified

accordingly. The value shall be a positive integer equaling or larger than 2 and the default value is 3, which means that target loci combinations with 3 loci will be screened out. 2 means that a single round of calculation will be carried out as per multiple loci numbers gradually increased one by one within a certain number interval and the corresponding lines of the parameters **#23.Batch Genome Size from n** and **#24.Batch Genome Size to m** shall be modified accordingly. The value shall be a positive integer equaling or larger than 2 and the default value is 2-10, which means the target loci combinations with 2-10 loci will be screened out.

- (13) Modify the corresponding line of the parameter **#25. Times of Calculation Cycles = n** as a positive integer, which allows repeated calculation under the same parameter values set in step (2)-(12). The default value is 1, which means that only one round of calculation shall be carried out under the above-mentioned values of parameters for loci combination screening. When the value is 2, it means that two rounds of screening shall be carried out and so on.
- (14) Modify the corresponding line of the parameter **#26.Result Format** as 2 or 4 to set the format of the output file of the loci combination screening results. 2 represents **Unit Results** and 4 **Merged Results**. The former outputs the screening results of multi rounds of calculation as several individual files, and the later as one merged file. The default value is 4. The multi rounds of calculation will be triggered if **Batch Genome Size** is selected for **#21.Analysis method for Number of Target Loci**, or if the value of **#25. Times of Calculation Cycles = n** is set as larger than 1. In such case, 4 is recommended.
- (15) Save the parameter setting file GASetting.txt, with no need to modify any other parameters.
- (16) Run the command **mono LociScan-C.exe GASetting.txt** and wait until LociScan pops up success or failure messages on the command line interface after backstage calculation (Figure 23), with the calculation time used for loci combination screening in case of success and none in the other case.
- (17) Switch to the results file directory set in **#7.Output Path** to check the analysis results in the corresponding "LociScan\_EA\_CSVFileName\_CreateDate\_CreateTime.csv" file. See section 3.5 for the details about the output format of the loci combination screening results.

### 3.9.3 Loci Combination Evaluation

Loci combination evaluation refers to the process of evaluating the fitness function values of the many loci combinations input by the user with training data. These fitness functions are 8 kinds of loci combination appraisal indices targeting at plant variety discrimination and the result will list out the names of the samples which cannot be identified. Only .csv format is supported at present.

#### 3.9.3.1 Evaluation Operation in GUI Version

Loci combination evaluation is operated in GUI version LociScan as follows:

- (1) Select **Loci combination evaluation** in the Operation section.
- (2) Select the corresponding molecular marker type for the training data in Marker Type section as shown in Figure 13, i.e. **SNP/InDel** (e.g.AT, CG or AB) , the default value, or **SSR/STR** (e.g.188/266). See section 3.3 for the details about the format of the training data.
- (3) Select the data structure of the training data in Data Structure section as shown in Figure 13, where the default value **Row:Loci; Col:Samples** means that the rows represent loci and column samples, and **Row: Samples; Col:Loci** means the opposite. See section 3.3 for the details about the data structure of the training data.



- (4) Click Open button in Data File Path section as shown in Figure 13, select the data file to be analyzed in the popup dialog box Open File and then click Open button to read the data file.
- (5) Click Open button in Loci File Path section as shown in Figure 13, select the loci combination data file to be evaluated in the popup dialog box Open File and then click Open button to read the data file. See section 3.4 for the details about the format of the loci combination data to be evaluated.
- (6) In Light Green color section as shown in Figure 20, set the value of Fitness Function, which represents the VDP method, including R-VDP, the default value, C-VDP, P-VDP, TDP and O-VDP. Please refer to the references about VDP for the detailed statistical methods of R-VDP [1], C-VDP, P-VDP and TDP, and those about LociScan for the detailed methods of O-VDP. For these five functions, we search for the largest value of their respective statistical method.
- (7) In Light Green color section as shown in Figure 20, set the method and threshold value of Variety Threshold Setting, which allows the user to define the loci difference threshold of the same variety, i.e. Ratio of Different Loci or Number of Different Loci, the default value. The former defines samples as the same variety when the ratio of different loci is smaller than the threshold value set by the user. The value is set as a decimal between 0 and 1 (including 1) and the default value is 0.05, which means that the samples will be identified as the same variety when the proportion of different loci out of all loci is less than 5%. The later defines samples as the same variety when the number of different loci is less than the threshold value set by the user. The value is set as a positive integer larger than 1 and the default value is 1, which means that the samples will be identified as the same variety when the number of different loci is less than 1.
- (8) Click OK button and wait until LociScan pops up success or failure messages on the main interface after backstage calculation (Figure 22), with the calculation time used for loci combination screening in case of success and none in the other case.
- (9) Click Open Result Folder to check the analysis results in the corresponding "LociCheck\_FitnessFuctionName\_CSVFileName\_CreateDate\_CreateTime.csv" file. See section 3.6 for the details about the output format of the loci combination evaluation results.

### 3.9.3.2 Evaluation Operation in CLI Version

Loci combination evaluation is operated in CLI version LociScan as follows:

- (1) Open the parameter setting file GASetting.txt. Note that the annotation lines and the parameter lines in the file must not be deleted and shall only be modified according to the following steps.
- (2) Modify the corresponding line of the parameter #1.Operation as 2. 1 represents Loci combination screening, 2 Loci combination evaluation, and 3 Data Cleaning.
- (3) Modify the corresponding line of the parameter #2.Marker Type as 1 or 2. 1 represents that the molecular marker type of the training data is SSR/STR, and 2 SNP/InDel. See section 3.3 for the details about the format of the training data.
- (4) Modify the corresponding line of the parameter #3.Data Structure as 1 or 2. 1 means that the rows represent loci and the columns samples and 2 means the opposite. See section 3.3 for the details about the structure of the training data.
- (5) Modify the corresponding line of the parameter #4.Different Allele for SSR: Base-pairs



**Number > n bp** as 0 or a positive integer. The parameter needs to be modified only when the value of the corresponding line of the parameter **#2.Marker Type** is 1. See section 3.3 for the details.

- (6) Input into the corresponding line of the parameter **#5.Data File Path (.csv)** the path of the genotype data file to be analyzed, including both the file name and extension.
- (7) Input into the corresponding line of the parameter **#6.Loci File Path (.csv)** the path of the loci combination data file to be analyzed, including both the file name and extension.
- (8) Input into the corresponding line of the parameter **#7.Output Path** the storage path of the result file, with no need to mention the file name and extension.
- (9) Modify the corresponding line of the parameter **#17.Fitness Function** as a number between 1 and 8 to set the fitness function used for the loci combinations input by the user. 1 represents **R-VDP**, 2 **C-VDP**, 3 **P-VDP**, 4 **TDP**, 5 **R-VDP-reciprocal**, 6 **C-VDP-reciprocal**, 7 **TDP-reciprocal** and 8 **O-VDP**. The default value is 1. Please refer to the references about VDP for the detailed statistical methods of **R-VDP** [1], **C-VDP**, **P-VDP** and **TDP**, and those about LociScan for the detailed methods of **O-VDP**. For these five functions, we search for the largest value of their respective statistical method, and for the other three functions, the smallest value.
- (10) Modify the corresponding line of the parameter **#18.Variety Threshold Setting Type** as 1 or 2, which allows the user to define the extent of tolerance of different loci in one variety. 1 represents **Ratio of Different Loci** and 2 **Number of Different Loci**. The default value is 2. 1 defines samples as the same variety when the ratio of different loci is smaller than the threshold value set by the user and the corresponding line of the parameter **#19.Ratio of Different Loci < n** shall be modified accordingly. The value is set as a decimal between 0 and 1 (including 1) and the default value is 0.05, which means that the samples will be identified as the same variety when the proportion of different loci out of all loci is less than 5%. 2 defines samples as the same variety when the number of different loci is less than the threshold value set by the user and the corresponding line of the parameter **#20.Number of Different Loci < n** shall be modified accordingly. The value is set as a positive integer larger than 1 and the default value is 1, which means that the samples will be identified as the same variety when the number of different loci is less than 1.
- (11) Save the parameter setting file **GASetting.txt**, with no need to modify any other parameters.
- (12) Run the command **mono LociScan-C.exe GASetting.txt** and wait until LociScan pops up success or failure messages on the command line interface after backstage calculation (Figure 23), with the calculation time used for loci combination screening in case of success and none in the other case.
- (13) Switch to the results file directory set in **#7.Output Path** to check the analysis results in the corresponding "LociCheck\_FitnessFunctionName\_CSVFileName\_CreateDate\_CreateTime.csv" file. See section 3.6 for the details about the output format of the loci combination evaluation results.

### 3.9.4 Data Cleaning

The purpose of data cleaning is to carry out verification and treatment on the content of the training data in order to guarantee the normal operation of loci combination screening. The method of verification is to evaluate by the training data if there exists genotypic variation in all the loci, that is to say, if there exists different genotype between different samples. If there is no

difference between the genotype data in the loci of different samples, then delete the loci and its corresponding data. Only loci with genotype difference and their data can be reserved in the output data. The supported format for the training data (input data) is .csv at present.

#### 3.9.4.1 Cleaning Operation in GUI Version

Data cleaning is operated in GUI version LociScan as follows:

- (1) Select **Data cleaning** in the Operation section.
- (2) Select the corresponding molecular marker type for the training data in Marker Type section as shown in Figure 10, i.e. **SNP/InDel** (e.g.AT, CG or AB) , the default value, or **SSR/STR** (e.g.188/266). See section 3.3 for the details about the format of the training data.
- (3) Select the data structure of the training data in Data Structure section as shown in Figure 10, where the default value **Row:Loci; Col:Samples** means that the rows represent loci and column samples, and **Row: Samples; Col:Loci** means the opposite. See section 3.3 for the details about the data structure of the training data.
- (4) Click Open button in Data File Path section as shown in Figure 10, select the data file to be analyzed in the popup dialog box Open File and then click Open button to read the data file.
- (5) Click OK button and wait until LociScan pops up success or failure messages on the main interface after backstage calculation (Figure 22), with the calculation time used for loci combination screening in case of success and none in the other case.
- (6) Click **Open Result Folder** to check the analysis results in the corresponding "DataClean\_CSVFileName\_CreateDate\_CreateTime.csv" file. See section 3.7 for the details about the output format of the data cleaning results.

#### 3.9.4.2 Cleaning Operation in CLI Version

Data cleaning is operated in CLI version LociScan as follows:

- (1) Open the parameter setting file GASetting.txt. Note that the annotation lines and the parameter lines in the file must not be deleted and shall only be modified according to the following steps.
- (2) Modify the corresponding line of the parameter **#1.Operation** as 3. 1 represents **Loci combination screening**, 2 **Loci combination evaluation**, and 3 **Data Cleaning**.
- (3) Modify the corresponding line of the parameter **#2.Marker Type** as 1 or 2. 1 represents that the molecular marker type of the training data is **SSR/STR**, and 2 **SNP/InDel**. See section 3.3 for the details about the format of the training data.
- (4) Modify the corresponding line of the parameter **#3.Data Structure** as 1 or 2. 1 means that the rows represent loci and the columns samples, and 2 means the opposite. See section 3.3 for the details about the structure of the training data.
- (5) Modify the corresponding line of the parameter **#4.Different Allele for SSR: Base-pairs Number > n bp** as 0 or a positive integer. The parameter needs to be modified only when the value of the corresponding line of the parameter **#2.Marker Type** is 1. See section 3.3 for the details.
- (6) Input into the corresponding line of the parameter **#5.Data File Path (.csv)** the path of the genotype data file to be analyzed, including both the file name and extension.
- (7) Input NULL into the corresponding line of the parameter **#6.Loci File Path (.csv)**.
- (8) Input into the corresponding line of the parameter **#7.Output Path** the storage path of the result file, with no need to mention the file name and extension.
- (9) Save the parameter setting file GASetting.txt, with no need to modify any other parameters.

- (10) Run the command `mono LociScan-C.exe GASetting.txt` and wait until LociScan pops up success or failure messages on the command line interface after backstage calculation (Figure 23), with the calculation time used for loci combination screening in case of success and none in the other case.
- (11) Switch to the results file directory set in `#7.Output Path` to check the analysis results in the corresponding “DataClean\_CSVFileName\_CreateDate\_CreateTime.csv” file. See section 3.7 for the details about the output format of the data cleaning results.

## References:

1. Yang Y, Tian H, Wang R et al. Variety Discrimination Power: An Appraisal Index for Loci Combination Screening Applied to Plant Variety Discrimination, *Frontiers in Plant Science* 2021;12:331.