

“LociScan V2.0” 用户手册

目录

1. 引言	2
1.1 编写目的	2
1.2 项目背景	2
2. 软件概述	2
2.1 目标	2
2.2 运行环境	2
2.3 功能结构	3
2.4 主界面展示	3
3. 使用说明	6
3.1 GUI 版本的安装和初始化	6
3.2 CLI 版本的安装和初始化	7
3.3 训练数据的输入	7
3.4 验证位点组合的输入	10
3.5 位点组合筛选结果的输出	11
3.6 位点组合验证结果的输出	12
3.7 数据清洗结果的输出	12
3.8 主要功能模块简介	13
3.8.1 位点组合筛选	13
3.8.2 位点组合验证	13
3.8.2 数据清洗	13
3.9 操作步骤	13
3.9.1 位点组合筛选——遗传算法	13
3.9.1.1 GUI 版本的遗传算法操作	13
3.9.1.2 CLI 版本的遗传算法操作	16
3.9.2 位点组合筛选——穷举算法	18
3.9.2.1 GUI 版本的穷举算法操作	18
3.9.2.2 CLI 版本的穷举算法操作	19
3.9.3 位点组合验证	21
3.9.3.1 GUI 版本的位点组合验证	21
3.9.3.2 CLI 版本的位点组合验证	22
3.9.4 数据清洗	22
3.9.4.1 GUI 版本的数据清洗	23
3.9.4.2 CLI 版本的数据清洗	23

公开获取方法

LociScan 可以通过以下链接下载:

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

开发者联系方式

授权使用和技术问题可以通过以下地址进行邮件联系

caurwx@163.com or caurwx@gmail.com

1. 引言

1.1 编写目的

本手册是“LociScan V2.0”配套使用说明及操作手册,方便用户更好地熟悉和使用该款软件。该软件提供在海量的位点组合中快速、简单、有效的筛选出品种识别率最优位点组合的方法,主要用于植物品种鉴定位点组合筛选。

1.2 项目背景

随着分子标记技术的发展,单核苷酸的多态性(以下简称为 SNP)和核苷酸的插入和缺失(以下简称为 InDel)作为第三代分子标记已经在逐步探索应用于植物品种鉴定工作。根据实际业务的需求,对成千上万的位点进行质量和数量的优化,不仅可以降低该技术的应用成本,而且可以提高其数据分析效率。

常规的分子标记位点筛选方法是分析样品的遗传背景信息,以遗传多样性为评价指标挑选最优的分子标记位点集合。本软件结合用于植物品种鉴定分子标记技术的特点,提出一种基于遗传算法的植物品种鉴定位点筛选方法,充分利用遗传算法在优化问题求解方面的优势,将其引入位点组合筛选问题的求解,根据植物品种鉴定的特点,定义植物品种识别率的统计方法,设计了遗传算法适应度函数和约束条件,构建了适用于植物品种鉴定位点组合筛选的遗传算法模型,提高了筛选出来的位点组合的植物品种鉴定能力。

2. 软件概述

2.1 目标

本软件“LociScan V2.0”是基于.Net Framework 4.0 框架研发的一款有效、快速筛选 SSR/STR、SNP 和 INDEL 等分子标记的最优位点组合的软件。本软件的图形化用户界面版本(GUI)采用传统的用户界面设计,便于用户快速掌握该软件的使用;本软件的命令行界面版本(CLI)采用命令行启动的方式设计,便于用户处理大规模的训练数据。本软件仅支持.csv 格式的读取,以.csv 格式输出数据;本软件的核心功能主要是在海量的位点组合中快速、简单的筛选出品种识别率最高的位点组合,以达到减少分子标记的位点数量、降低分子检测试验成本、提高植物品种鉴定工作效率的目的。

2.2 运行环境

(1) 软件环境

GUI 版本软件运行环境:Windows 7 或者更高版本的系统,需要保证电脑已安装.Net Framework 4.0 及其以上版本。

CLI 版本软件运行环境:服务器推荐 CentOS 7、Ubuntu 16.04、Debian 9、Raspbian 9 、Fedora 28 或更高版本的 Linux 系统,需要确保服务器已安装 MonoDevelop 源代码集成开发环境;macOS 10.9 或更高版本的 macOS 系统,需要确定电脑已安装 MonoDevelop 源代码集

成开发环境；Window 7 或更高版本的操作系统，需要确保电脑已安装 .Net Framework 4.0 及其以上版本。

(2) 硬件环境

CPU：建议使用频率 1GHz 以上的 CPU。

内存：建议使用 1GB 或更大的内存。

显示器：用 1024×768 像素及以上，真彩色 32 位颜色。

2.3 功能结构

该软件分为两大模块：位点组合筛选、位点组合评估和数据清洗，下面是其功能结构图：

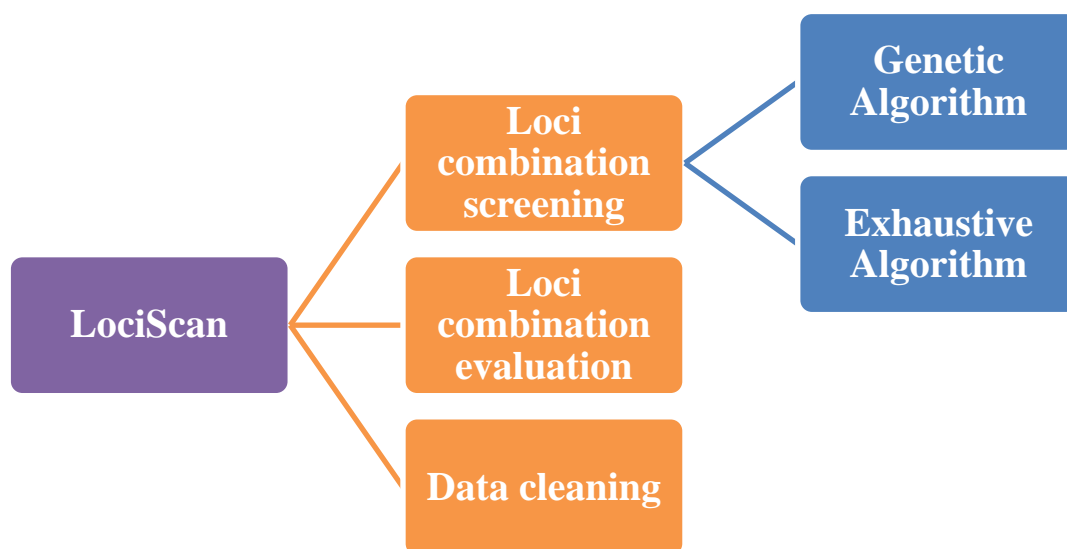


图 1 软件功能结构图

2.4 主界面展示

打开 LociScan 工具，在 Operation 区域默认选择 Loci combination screening 进入位点组合筛选界面（图 2），点击 Loci combination evaluation 进入位点组合评估界面（图 3），点击 Data cleaning 进入数据清洗界面（图 4）。

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)
Open

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.

Genetic Algorithm
Exhaustive Algorithm

Crossover Rate [0, 1] = 0.90

Mutation Rate [0, 1] = 0.01

Population Size [3, ∞) = 100

Generation Size [1, ∞) = 2000

☒ Keep Best in Generation
☐ Stop when FF=1

Mutate Method: WholeRandom

GA Method: normal-GA

Fitness Function (FF): R-VDP

Variety Threshold Setting:
☐ Ratio of Different Loci (0,1] < 0.05
☒ Number of Different Loci [1,∞) < 1

Number of Target Loci:
☒ Fixed Genome Size [2, ∞) = 3
☐ Batch Genome Size [2, ∞) from 2 to 10

Times of Calculation Cycles: 1

Result Format:
☐ All Generation
☒ Last Generation

File Format:
☐ Unit Results
☒ Merged Results

OK
Open Result Folder
Exit

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

图 2 位点组合筛选界面

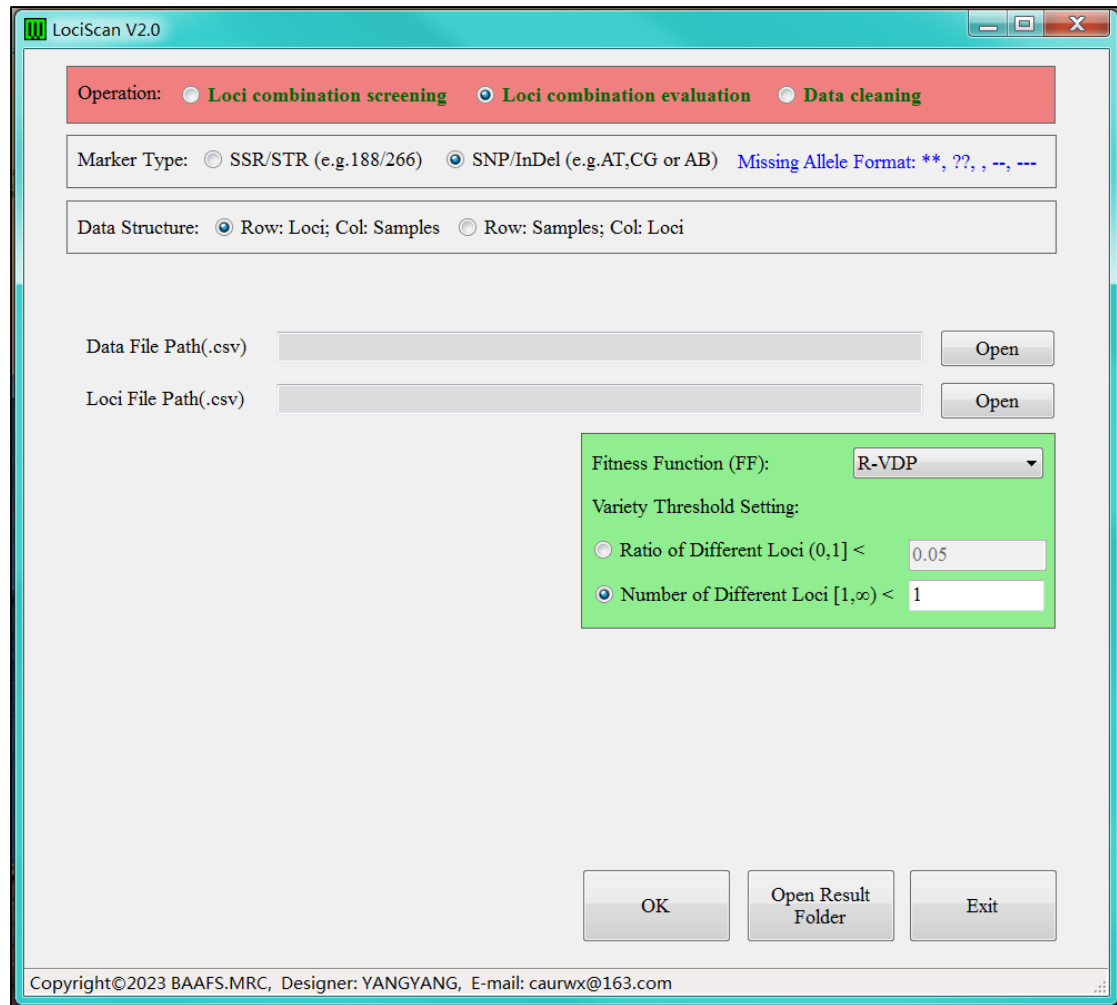


图 3 位点组合评估界面

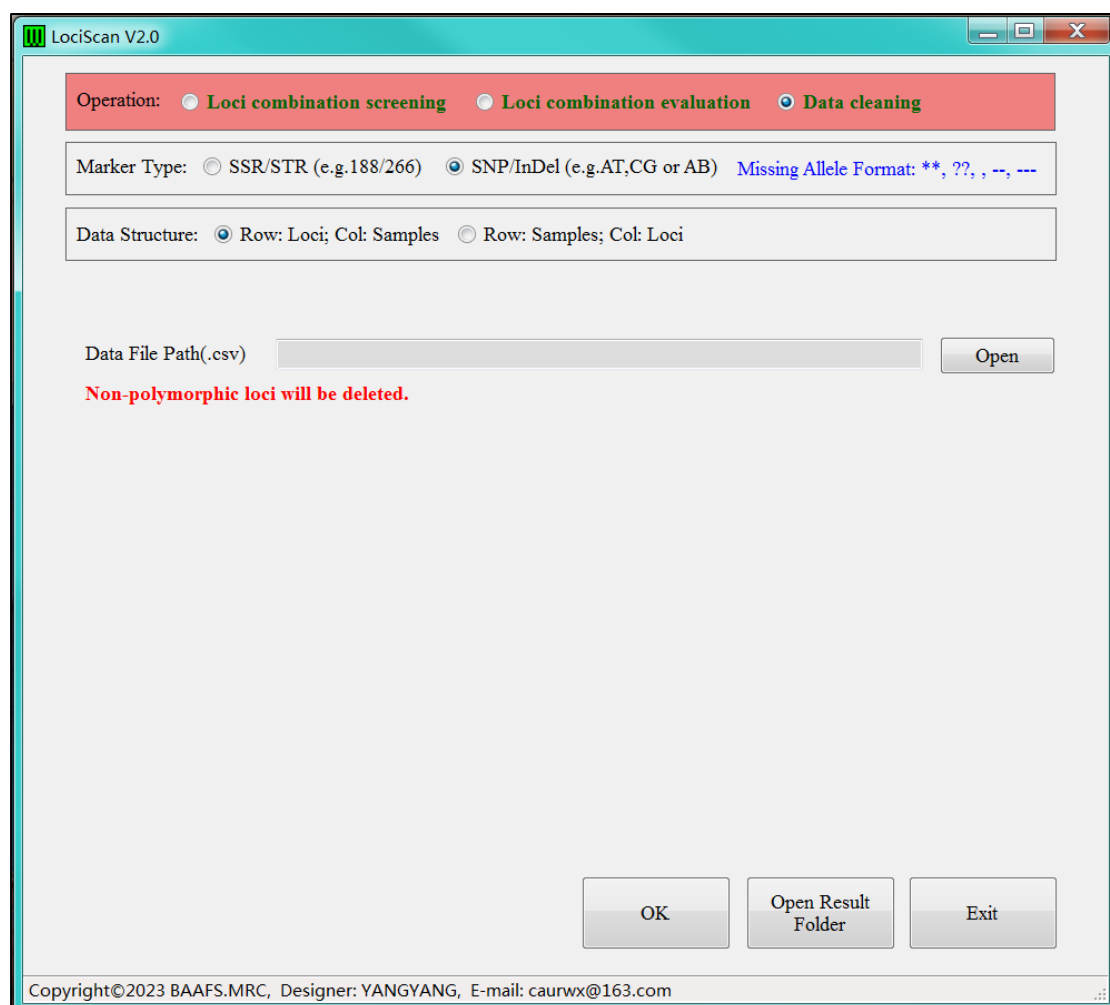


图 4 数据清洗界面

3. 使用说明

3.1 GUI 版本的安装和初始化

- (1) 如图 5 所示，首次运行需双击“LociScan_V2.0.exe”应用程序，自动检测是否获得已经获得许可，如果获得许可将直接进入程序主界面，如果未获得许可将进入下一步注册界面。

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

图 5 运行程序

- (2) 注册界面如图 6 所示，请将注册码发邮件到 caurwx@163.com 获得验证码，完成注册方可运行。

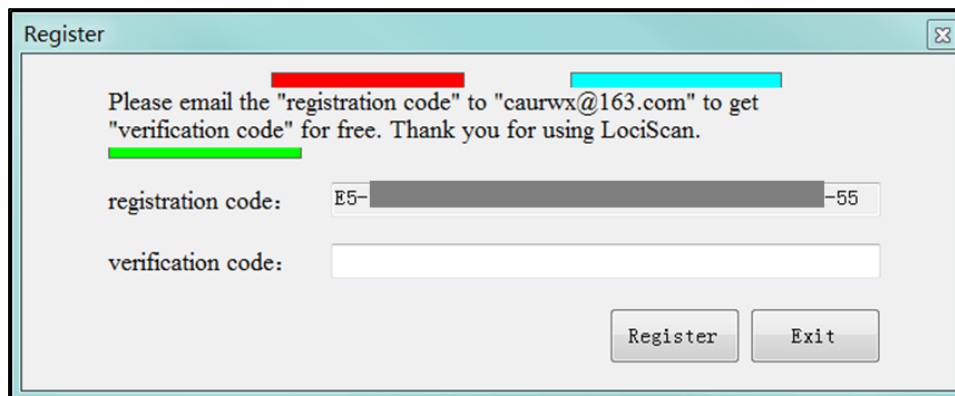


图 6 GUI 版本的注册界面

3.2 CLI 版本的安装和初始化

- (1) 以 CentOS 7 系统为例，采用以下命令安装 MonoDevelop，通过最后一行命令确认 MonoDevelop 是否安装成功：

```
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) MonoDevelop 在其他版本的 Linux 系统和 macOS 系统的安装步骤，具体可参考网站 <https://www.mono-project.com/download/stable/>。
- (3) 将命令终端跳转到本程序所在的文件目录，运行以下启动命令，首次运行程序将提示获得注册码，如图 7 所示，请将注册码发邮件到 caurwx@163.com 获得验证码，完成注册方可运行。

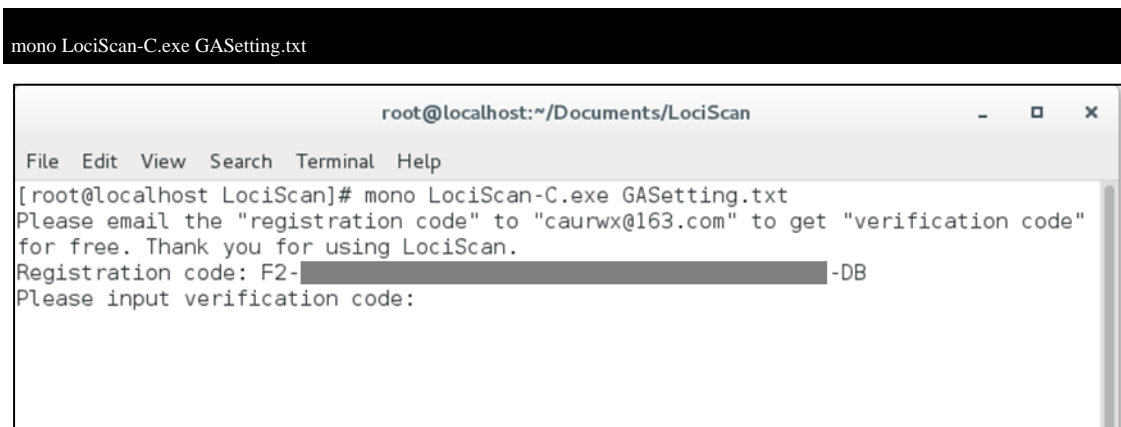


图 7 CLI 版本的注册界面

3.3 训练数据的输入

- (1) SNP 或 InDel 数据准备

- 本软件支持的 SNP 标记等位基因数据格式，一般格式为 A\T\C\G，或者简化格式的 A\B； 本软件支持的 InDel 标记等位基因数据格式，一般格式为 A\B。本软件仅支持物种的染色体倍性为二倍体，实际的基因型数据如图 8 所示。以普通格式为例，纯合体类型具体为 AA、TT、CC、GG，杂合体类型具体为 AT、AC、AG、TC、TG、CG；以简化格式为例，具体为 AA、AB、BB，其中 AA、BB 代表纯合类型，AB 代表杂合类型。
- 图 8 展示的数据结构是不同行代表不同的位点，不同列代表不同的样品。第一行是不同样品对应的样品编号，从第二行开始是不同位点对应的基因型数据。第一列是不同位点对应的位点编号，从第二列开始是不同样品对应的基因型数据。
- 输入数据的文件格式为 .csv，其中缺失数据用 ??、--、---、**、表示。输入的示例数据如图 8 所示。
- 请在打开数据文件前，关闭所有打开相同的数据文件的程序。

```

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,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,GG,TT,GG,GG,GG,TT,GG,TT,GG,GG,GG,GG,GG,GG,GG,GG,TG,GG,GG,GG,GG,GG
L004,AA,GG,AA,GG,AA,AA,AA,GG,AA,AA,GG,AA,AA,AA,AA,AA,AA,AA,GG,GG,AA,AA,AA,GG,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,CC,TC,CC
L006,GG,TT,TG,TT,--,TT,TT,TG,GG,TG,TT,TG,TT,--,TT,TG,TT,TT,TG,TG,TT,TG,TT,TG,TT,TG,TT,TG
L007,--,AA,GG,AA,AA,GG,GG,GG,GG,GG,GG,GG,GG,--,GG,GG,GG,GG,GG,AA,AA,GG,AA,AA,GG,GG,GG,GG
L008,AG,GG,AG,GG,--,GG,GG,--,AA,GG,GG,AG,GG,AG,AG,GG,AG,GG,GG,GG,AG,--,AA,AG,GG,GG,AA,GG,GG
L009,CC,CC,CC,TC,CC,TC,CC,CC,TC,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,--,TT,TC,CC,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,AA,CC,AA,AA,AA,CC
L012,GG,--,GG,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,GG,AA,GG,GG,GG,GG,GG,AA,GG,GG,AA,GG,GG,AA,GG,AA
L014,CC,CC,TT,CC,CC,TT,CC,CC,TT,CC,CC,CC,TT,CC,CC,CC,TT,TC,CC,TT,CC,TT,CC,CC,TT,CC,CC,TT,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,GG,TT
L016,AG,AG,AG,AA,AG,AG,AA,--,AA,AG,AG,AA,AG,AG,AG,AA,AG,AG,AA,GG,AG,AG,AG,AG,AA,AG,AA,AG,AG
L017,CC,AA,AA,AA,AA,AA,AA,CC,AA,AA,AA,AA,AA,AA,CC,AA,AA,AA,AA,CC,AC,AA,AA,AA,AA,AA,AA,CC

```

图 8 输入 SNP 数据格式

(2) SSR 或 STR 数据准备

- 本软件支持的 SSR 等位基因数据格式，一般格式为正整数，正整数代表该片段包含的碱基数量，且仅支持物种的染色体倍性为二倍体。实际的基因型数据格式如图 9 所示，根据用户设定的差异碱基数确定其是否为杂合位点，若碱基差异数设定为大于 0bp，322/323 为杂合体类型，322/322 为纯合体类型；若碱基差异数设定为大于 1bp，322/325 为杂合体类型，322/323 或 322/322 为纯合体类型。
- 图 9 展示的数据结构是不同行代表不同的样品，不同列代表不同的位点。第一行是不同位点对应的位点编号，从第二行开始是不同样品对应的基因型数据。第一列是不同样品对应的样品编号，从第二列开始是不同位点对应的基因型数据。
- 输入数据的文件格式为 .csv。其中缺失数据用 /、?/?、*/、-/-、./表示，输入的示例数据如图 9。
- 请在打开数据文件前，关闭所有打开相同的数据文件的 Excel 程序。

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	154/1												
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	413/4												
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	229/2												
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	284/2												
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	248/255	386/386	317/317	341/341	410/410	404/404	301/301	262/262	185/185	265/265	206/206	169/169												
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												

图 9 输入 SSR 数据格式

(3) 参数选择

- 在图 10 的 Marker Type 区域选择训练数据对应的分子标记类型（CLI 版本请调整参数配置文件的 #2.Marker Type 对应的参数行），例如图 8 展示分子标记类型是 SNP，则选择 SNP/InDel (e.g.AT, CG or AB)。图 9 展示分子标记类型是 SSR，则选择 SSR/STR (e.g.188/266)。
- 在图 10 的 Data Structure 区域选择训练数据对应的数据结构（CLI 版本请调整参数配置文件的 #3. Data Structure 对应的参数行），例如图 8 展示的数据结构是不同行代表不同位点，不同列代表不同样本，则选择 Row:Loci; Col:Samples。图 9 展示的数据结构是不同行代表不同样本，不同列代表不同位点，则选择 Row: Samples; Col:Loci。
- 当 Marker Type 区域选择为 SSR/STR (e.g.188/266) 时，需要在图 10 的 Different Allele for SSR 区域设置其碱基对数量差异阈值（CLI 版本请调整参数配置文件的 #4. Different Allele for SSR: Base-pairs Number > n bp 对应的参数行）。例如，默认值为 >0 bp，代表基因型分别为 200/200 与 201/201 的两个位点之间是有差异的，200/200 与 200/200 的两个位点是无差异的；当该阈值设置为 >1bp，代表基因型分别为 200/200 与 202/202 的两个位点之间是有差异的，200/201 与 201/202 的两个位点或 200/200 与 201/201 的两个位点之间均是无差异的，以此类推。
- 上述参数设置好后，再到 Date File Path 区域点击 Open 按钮，选择对应的训练数据文件，数据格式目前只支持.csv 后缀文件（CLI 版本请调整参数配置文件的 #5.Data File Path 对应的参数行）。训练数据对应的文件选中后，将不能修改步骤 a) 和步骤 b) 中提到的参数，若要重新设置这两个参数，需要启动 LociScan 程序，或点击 Open 按钮出现文件选择对话框后再点击 Cancel 按钮即可。
- 重点提示，训练数据文件在导入 LociScan 进行位点组合筛选前，建议采用 LociScan 的 Data Cleaning 功能做以下内容校验：数据集需要排除没有多态性的位点，即在该数据集的不同样品之间基因型全部一致的位点都必须全部删除，否则本程序在运行位点组合筛选功能时将会陷入无限死循环。

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.

图 10 输入训练数据时需要设置的参数区域

3.4 验证位点组合的输入

(1) 待验证的位点组合数据准备

- a) 本软件支持的位点组合数据格式，每一行代表一个位点组合，不同的位点组合用换行符进行分隔，具体内容为：第一列是位点组合的编号，从第二列开始是具体组成该位点组合的位点名称，不同的位点用逗号进行分隔。以图 11 为例，SNP-LC001 代表位点组合的编号，L056、L081……代表位点名称。

```
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,
```

图 11 待验证位点组合的数据格式

- b) 输入待验证位点组合的文件格式为.csv，可用文本编辑器进行编辑，也可用 Excel 进行文件编辑，本软件会自动忽略文件中出现的多余分隔符号，如图 12 所示。

```
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,,,,,,
```

图 12 待验证位点组合的数据格式

- c) 请在打开数据文件前，关闭所有打开相同的数据文件的程序。
d) 注意：位点名称需要和 3.3 训练数据中的位点名称保持一致。

(2) 参数设置

当在 Operation 区域点击 **Loci combination evaluation** 进入位点组合评估界面时，将出现图 13 的参数设置，训练数据文件的输入和参数配置参考 3.3 小节，位点组合文件的请按照上述格式进行编辑准备，然后点击 Loci File Path 区域的 Open 按钮，选择对应的文件（CLI 版本请调整参数配置文件的 #6.Loci File Path 对应的参数行）。

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"/>

图 13 输入验证位点组合时需要设置的参数区域

3.5 位点组合筛选结果的输出

位点组合筛选的结果以.csv 的文件格式输出，GUI 版本的 LociScan 的结果文件输出路径和主程序在同一文件目录下，CLI 版本的 LociScan 的结果文件路径由用户自行设置（CLI 版本请调整参数配置文件的#7.Output Path 对应的参数行）。结果文件名以 LociScan 开头、训练数据的原文件名和数据分析的结束时间三者组合而成，结果文件可用文本编辑器打开，也可用 EXCEL 打开。

当采用遗传算法且 Result Format 区域选择 All Generation 时，位点组合筛选结果文件如图 14 所示。不同行代表不同演化代数群体对应的结果。每一行包含以下主要内容：第一列是位点组合的演化代数序号，第二列是本代（generation）群体中最小的适应度函数值，第三列是本代群体中最大的适应度函数值，从第四列开始到本行结束是本代群体中适应度函数最大的位点组合对应的位点名称。

```
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
```

图 14 遗传算法输出为 All Generation 格式的位点组合筛选结果

当采用遗传算法且 Result Format 区域选择 Last Generation 时，位点组合筛选结果文件如图 15 所示。不同行代表不同次计算对应的结果。每一行包含以下主要内容：第一列是位点组合的第 N 次计算的自动编号，第二列是本次计算中第一代群体最小的适应度函数值，第三列是本次计算最后一代群体中最大的适应度函数值，从第四列开始到本行结束是本次计算最后一代群体中适应度函数最大的位点组合对应的位点名称。

```
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
```

图 15 遗传算法输出为 Last Generation 格式的位点组合筛选结果

当采用穷尽搜索算法时，位点组合筛选结果文件如图 16 所示，包含以下主要内容：第一列是位点组合的序号，第二列是位点组合的适应度函数值，从第三列开始到本行结束是本位点组合对应的位点名称。第一行代表适应度函数最小对应的位点组合结果，第二行适应度函数最大对应的位点组合结果。

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

图 16 穷举算法的位点组合筛选结果

位点组合筛选的日志文件以 .log 的文件格式输出，所有版本的 LociScan 输出的日志文件均存放在和主程序同一文件目录下。日志文件名为 Stopwatch.log，可用文本编辑器打开。当 LociScan 进行位点组合筛选分析时，每一次计算都会形成一条记录添加到该日志文件，每一条记录内容如图 17 所示：第一列为本次计算输出的位点组合筛选结果文件名，第二列为本次计算消耗的时间（单位：秒），第三列为本次计算采用的适应度函数，第四列为本次计算的采用的模型参数，以 GA 为例模型参数依次为 Crossover Rate、Mutation Rate、Population Size、Generation Size、Keep Best in Generation、Mutate Method、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
```

图 17 位点组合筛选的日志文件内容

3.6 位点组合验证结果的输出

位点组合验证的结果以 .csv 的文件格式输出，GUI 版本的 LociScan 的结果文件输出路径和主程序在同一文件目录下，CLI 版本的 LociScan 的结果文件路径由用户自行设置（CLI 版本请调整参数配置文件的 #7.Output Path 对应的参数行）。结果文件名以 LociCheck 开头、训练数据的原文件名和数据分析的结束时间三者组合而成，结果文件可用文本编辑器打开，也可用 EXCEL 打开。

如图 18 所示，位点组合验证结果文件包含以下主要内容：第一列是位点组合的编号，第二列是位点组合的适应度函数值，从第三列开始到本行结束是位点组合不能识别的样品位点名称。不同行代表不同位点组合对应的验证结果。当第二列为 1 时，代表该位点组合可以识别全部样品，当第二列为 -2 时，代表位点组合包含的位点名称与训练数据的位点名称不符，位点组合评价失败。

```
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
```

图 18 位点组合验证的文件内容

3.7 数据清洗结果的输出

数据清洗的结果均以 .csv 的形式输出，GUI 版本的 LociScan 的结果文件输出路径和主程序在同一文件目录下，CLI 版本的 LociScan 的结果文件路径由用户自行设置（CLI 版本请调整参数配置文件的 #7.Output Path 对应的参数行）。结果文件名以 DataClean 开头、训练数据的原文件名和数据分析的结束时间三者组合而成，结果文件可用文本编辑器打开，也可用 EXCEL 打开。结果文件的数据结构将与源文件保持一致，对不同样品之间没有基因型差异的

位点将会被直接删除，只保留有多态性的位点及其对应的基因型数据。

3.8 主要功能模块简介

3.8.1 位点组合筛选

位点组合筛选可以在海量的位点组合中快速、简单的筛选出满足植物品种鉴定需求的位点组合，为分子标记位点组合的质量和数量的优化提供科学、准确的基础依据，进而为促进分子标记应用于植物品种鉴定工作提供技术支持。

位点筛选算法分为遗传算法（Genetic Algorithm）和穷举算法（Exhaustive Algorithm）。遗传算法筛选通过输入杂交率、变异率、种群大小、演化代数等参数来筛选有效组合，结果显示每个组合的品种识别率和具体位点编号。穷举算法通过遍历所有的位点组合来搜索真正的最优解，结果显示最优组合和最劣组合的品种识别率和具体位点序号。

3.8.2 位点组合验证

位点组合验证指利用训练样品的基因型数据验证用户输入的位点组合的品种识别率，并列出不能识别的样品。

3.8.2 数据清洗

数据清洗指利用训练样品基因型数据校验所有位点的多态性，并删除在所有训练样品之间没有基因型差异的位点及其基因型数据。

3.9 操作步骤

3.9.1 位点组合筛选——遗传算法

位点组合筛选——遗传算法，指通过遗传算法筛选出指定位点数目且最有效的位点组合，适应度函数是以鉴定植物品种为目标的 8 种位点组合评价指标，输出结果将同时包含对应位点组合的适应度函数值。目前支持的训练数据（输入数据）格式为.csv。

3.9.1.1 GUI 版本的遗传算法操作

在 GUI 版本的 LociScan 软件中执行基于遗传算法的位点组合筛选，具体步骤如下：

- (1) 在 Operation 区域中选择 **Loci combination screening**。
- (2) 在图 10 的 Marker Type 区域选择训练数据对应的分子标记类型，包括 **SNP/InDel (e.g.AT, CG or AB)**和 **SSR/STR (e.g.188/266)**两种类型，默认值是 **SNP/InDel (e.g.AT, CG or AB)**。训练数据的数据格式具体内容参见 3.3 小节。
- (3) 在图 10 的 Data Structure 区域中选择训练数据的数据结构，**Row:Loci; Col:Samples** 表示行为位点列为样本，**Row: Samples; Col:Loci** 表示行为样本列为位点，默认值是 **Row:Loci; Col:Samples**。训练数据的数据结构具体内容参见 3.3 小节。
- (4) 在图 10 的 Data File Path 区域中，单击 Open 按钮，弹出 Open File 对话框，选择需要分析的数据文件，然后单击“打开”按钮等待数据文件的读取。
- (5) 在图 19 的 TabPages 区域选择模型算法，包括 **Genetic Algorithm** 和 **Exhaustive Algorithm** 两种方法，请选择 **Genetic Algorithm**。
- (6) 在 Genetic Algorithm 对应的 TabPages（图 19）中，设置 Crossover Rate 的值，其表示个体在演化下一代群体过程中执行交叉算子的概率，取值范围为 0-1，默认值为 0.90。
- (7) 在 Genetic Algorithm 对应的 TabPages（图 19）中，设置 Mutation Rate 的值，其表示个体在演化下一代群体过程执行变异算子的概率，取值范围为 0-1，默认值为 0.01。
- (8) 在 Genetic Algorithm 对应的 TabPages（图 19）中，设置 Population Size 的值，其表示划分的种群大小，取值范围为大于等于 3 的正整数，默认值为 100，视数据量划分类群的数量，种群数量越大寻找最优组合效果越好，但计算时间会延长。
- (9) 在 Genetic Algorithm 对应的 TabPages（图 19）中，设置 Generation Size 的值，其表示演化代数，取值范围为大于等于 1 的正整数，默认值为 2000，演化代数越多寻找最优组合效果越好，但是计算时间会延长。

- (10) 在 Genetic Algorithm 对应的 TabPages(图 19)中,勾选或不勾选 **Keep Best in Generation**, 其表示传递到下一代是否保留当代最优个体,默认值为勾选,即保留当代最优个体到下一代。
- (11) 在 Genetic Algorithm 对应的 TabPages (图 19) 中, 勾选或不勾选 **Stop When FF = 1**, 其表示当搜索到适应度函数值 (FF) 为 1 的位点组合时, 程序是否自动停止搜索, 默认值为不勾选, 即程序不自动停止搜索。
- (12) 在 Genetic Algorithm 对应的 TabPages (图 19) 中, 设置 **Mutate Method** 的值, 其表示遗传算法中变异算子的变异方法, 包括 **Whole Random** 和 **Half Random** 两种方法, 默认值为 **Whole Random**。
- (13) 在 Genetic Algorithm 对应的 TabPages 区域 (图 19) 中, 设置 **GA Method** 的值, 包括 **normal-GA** 和 **self-adaption-GA** 两种方法, 普通遗传算法如参考文献所述, 自适应遗传算法是在求解过程中动态调整遗传算子的参数和运算方式的优化算法, 默认值为 **normal-GA**。
- (14) 在图 20 的颜色为 LightGreen 区域中, 设置 **Fitness Function** 的值, 其表示品种识别率方法, 包括 **R-VDP**、**C-VDP**、**P-VDP**、**TDP**、**R-VDP-reciprocal**、**C-VDP-reciprocal**、**TDP-reciprocal** 和 **O-VDP**, 默认值为 **R-VDP**。**R-VDP**、**C-VDP**、**P-VDP** 和 **TDP** 的具体统计方法请参考 VDP 相关文献[1], **O-VDP** 请参考 LociScan 相关文献, 这五个函数均为搜索其对应统计方法的最大值, **R-VDP-reciprocal** 等三个函数则反之。
- (15) 在图 20 的颜色为 LightGreen 区域中, 设置 **Variety Threshold Setting** 的方法和阈值, 本参数是将同一品种的位点差异阈值交给用户来进行定义, 包括 **Ratio of Different Loci** 和 **Number of Different Loci** 两种方法, 默认值是 **Number of Different Loci**。前者表示采用“差异位点百分比”小于用户设定阈值方式进行定义同一品种, 取值范围为大于 0 且小于等于 1 的小数, 默认值是 0.05, 即差异位点占全部位点的比例小于 5%为同一品种; 后者表示采用“差异位点数”小于用户设定阈值方式进行定义同一品种, 取值范围为大于 1 的正整数, 默认值是 1, 即差异位点数小于 1 为同一品种。
- (16) 在图 20 的颜色为 Aqua 区域中, 设置 **Number of Target Loci** 的方法和参数值, 本参数是对搜索的目标位点组合包含的位点数目进行设定, 包括 **Fixed Genome Size** 和 **Batch Genome Size** 两种方法, 分别代表“目标位点组合的位点数”和“目标位点组合的位点数量区间”输入相应的值, 方法默认选择 **Fixed Genome Size**。前者是按照指定一个的位点数进行单轮计算, 取值范围为大于等于 2 的正整数, 默认值是 3, 即筛选位点数量是 3 的目标位点组合; 后者是按照位点数量区间从小到大逐个位点递增指定多个的位点数据进行单轮计算, 取值范围为大于等于 2 的正整数, 默认值是 2-10, 即筛选位点数量分别是 2、3、4……10 的目标位点组合。
- (17) 在图 21 的 **Result Format** 区域中设置基于 GA 的位点组合筛选结果的输出内容格式, 包括 **All Generation** 和 **Last Generation**, 默认值为 **Last Generation**。位点组合筛选结果的输出格式具体内容参见 3.5 小节。当用户设置 **Generation Size** 的值较大, 或者设置为多轮计算时, 建议采用 **Last Generation**。
- (18) 在图 20 的颜色为 Blue 区域中, 设置 **Times of Calculation Cycles** 的参数值, 本参数是对 (1)-(17)步骤设置的参数保持一致并进行重复计算, 默认值是 1, 即只计算一轮上述参数的位点组合筛选。
- (19) 在图 21 的 **File Format** 区域中设置位点组合筛选结果的输出文件格式, 包括 **Unit Results** 和 **Merged Results**, 前者对多轮计算的筛选结果输出为多个独立文件, 后者对多轮计算的筛选结果输出为一个合并文件, 默认值是 **Merged Results**。所谓的多轮计算是由 **Number of Target Loci** 的方法和 **Cycle calculation times** 的参数值决定, 当前者选择 **Batch**

Genome Size 方法时，或者后者设置的参数值大于 1 时，将会执行多轮计算，此时建议采用 Merged Results。

- (20) 点击“OK”按钮，等待 LociScan 的后台计算，运行成功或运行失败均将会在主界面右下角给出提示（图 22），运行成功时同时会显示位点组合筛选消耗的计算时间，运行失败将不显示计算时间。
- (21) 点击“Open Result Folder”，可以在对应的“LociScan_GA_CSVFileName_CreateDate_CreateTime.csv”文件中查看分析结果。位点组合筛选结果的输出格式具体内容参见 3.5 小节。

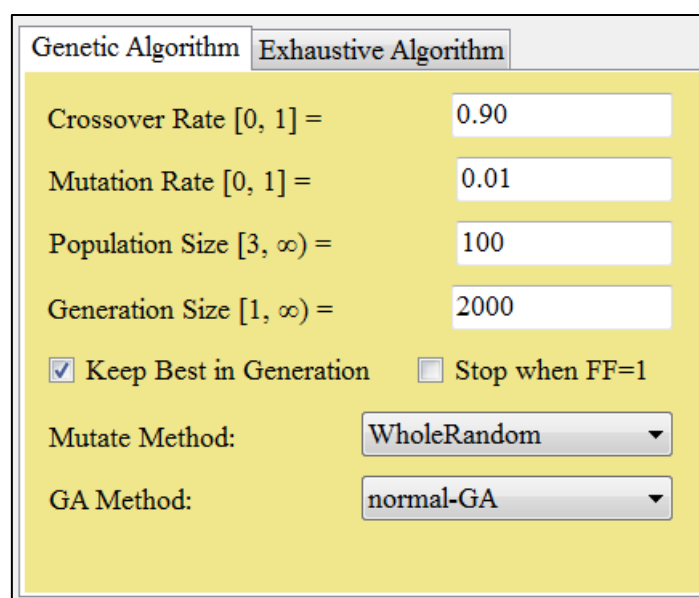


图 19 遗传算法模型参数设置区域

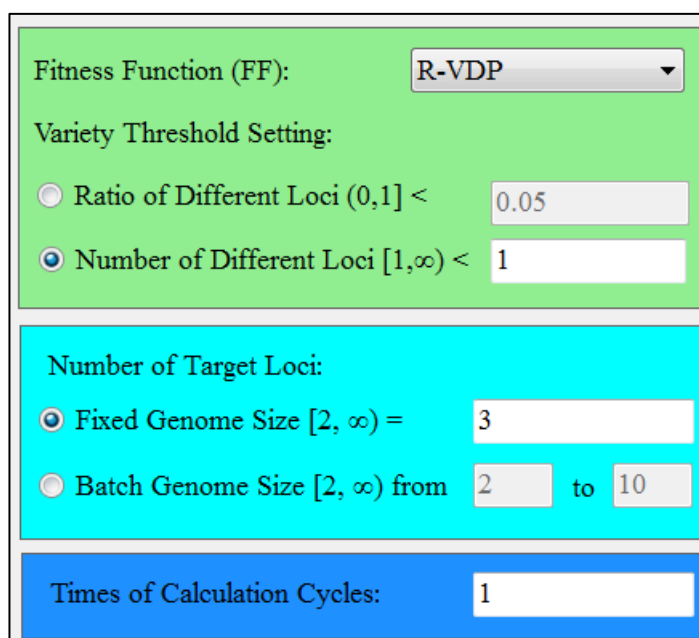


图 20 适应度函数参数设置区域

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

图 21 位点组合筛选结果的输出格式设置区域

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

图 22 GUI 版本的 LociScan 运行成功或运行失败的提示区域

3.9.1.2 CLI 版本的遗传算法操作

在 CLI 版本的 LociScan 软件中执行基于遗传算法的位点组合筛选，具体步骤如下：

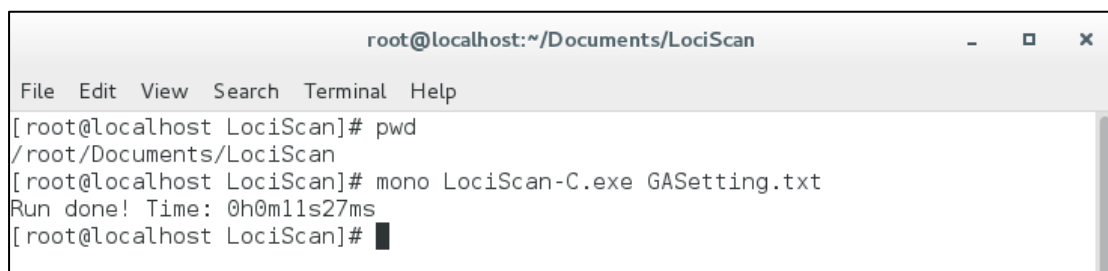
- (1) 打开参数配置文件 `GASetting.txt`，注意不要删除参数配置文件中的注释行和参数行，只根据以下操作步骤修改对应参数行内容。
- (2) 将#1.Operation 对应的参数行修改为 1。1 代表 Loci combination screening，2 代表 Loci combination evaluation，3 代表 Data Cleaning。
- (3) 将#2.Marker Type 对应的参数行修改为 1 或 2。1 代表训练数据对应的分子标记类型为 SSR/STR，2 代表 SNP/InDel。训练数据的数据格式具体内容参见 3.3 小节。
- (4) 将#3.Data Structure 对应的参数行修改为 1 或 2。1 表示行为位点、列为样本，2 表示行为样本、列为位点。训练数据的数据结构具体内容参见 3.3 小节。
- (5) 将#4.Different Allele for SSR: Base-pairs Number > n bp 对应的参数行修改为 0 或正整数，当#2.Marker Type 对应的参数行为 1 时才需要调整该参数，具体内容参见 3.3 小节。
- (6) 在#5.Data File Path (.csv) 对应的参数行输入需要分析的基因型数据文件路径，文件路径需要包含训练数据文件名和扩展名。
- (7) 在#6.Loci File Path (.csv) 对应的参数行输入 NULL。
- (8) 在#7.Output Path 对应的参数行输入结果文件的存放路径，文件路径不需要包含结果文件的文件名和扩展名。
- (9) 将#8.Algorithm 对应的参数行修改为 1。1 代表 Genetic Algorithm，2 代表 Exhaustive Algorithm。
- (10) 将#9.Crossover Rate = n 对应的参数行修改为大于等于 0 且小于等于 1 之间的小数，该参数表示个体在演化下一代群体过程中执行交叉算子的概率，取值范围为 0-1，默认值为 0.90。
- (11) 将#10.Mutation Rate = n 对应的参数行修改为大于等于 0 且小于等于 1 之间的小数，该参数表示个体在演化下一代群体过程执行变异算子的概率，取值范围为 0-1，默认值为 0.01。
- (12) 将#11.Population Size = n 对应的参数行修改为大于等于 3 的正整数，该参数表示每一代群体包含的个体数目，默认值为 100，视数据量划分类群的数量，种群数量越大寻找最优组合效果越好，但计算时间会延长。
- (13) 将#12.Generation Size = n 对应的参数行修改为大于等于 1 的正整数，该参数其表示演化新一代群体的次数，默认值为 2000，演化代数越多寻找最优组合效果越好，但是计算时间会延长。
- (14) 将#13.Keep Best in Generation 对应的参数行修改为 true 或 false，该参数表示传递到下一代是否保留当代最优个体，默认值为 true，即保留当代最优个体到下一代。
- (15) 将#14.Stop running when FF = 1 对应的参数行修改为 true 或 false，该参数表示当搜索到适应度函数值（FF）为 1 的位点组合时，程序是否自动停止搜索，默认值为 false，即

程序不自动停止搜索。

- (16) 将#15.Mutate Method对应的参数行修改为 0 或 1, 本参数表示遗传算法中变异算子的变异方法, 0 代表 Half Random, 1 代表 Whole Random, 默认值为 1。
- (17) 将#16.GAMethod对应的参数行修改为 0 或 1, 0 代表 normal-GA, 1 代表 self-adaption-GA 两种方法, 普通遗传算法如参考文献所述, 自适应遗传算法是在求解过程中动态调整遗传算子的参数和运算方式的优化算法, 默认值为 0。
- (18) 将#17.Fitness Function对应的参数行修改为 1-8 中某个值, 设置筛选位点组合采用的适应度函数方法, 1 代表 R-VDP、2 代表 C-VDP、3 代表 P-VDP、4 代表 TDP、5 代表 R-VDP-reciprocal、6 代表 C-VDP-reciprocal、7 代表 TDP-reciprocal 和 8 代表 O-VDP, 默认值为 1。R-VDP、C-VDP、P-VDP 和 TDP 的具体统计方法请参考 VDP 相关文献[1], O-VDP 请参考 LociScan 相关文献, 这五个函数均为搜索其对应统计方法的最大值, R-VDP-reciprocal 等三个函数则反之。
- (19) 将#18.Variety Threshold Setting Type对应的参数行修改为 1 或 2, 本参数是将同一品种存在差异位点的容忍度交给用户来进行定义, 1 代表 Ratio of Different Loci, 2 代表 Number of Different Loci 两种方法, 默认值是 2。该参数修改为 1 时, 表示采用“差异位点比率”小于用户设定阈值的方式进行定义同一品种, 需要同时修改#19.Ratio of Different Loci < n 对应的参数行, 取值范围为大于 0 且小于等于 1 的小数, 默认值是 0.05, 即差异位点占全部位点的比例小于 5%为同一品种; 该参数修改为 2 时, 表示采用“差异位点数”小于用户设定阈值的方式进行定义同一品种, 需要同时修改#20.Number of Different Loci < n 对应的参数行, 取值范围为大于 1 的正整数, 默认值是 1, 即差异位点数小于 1 为同一品种。
- (20) 将#21.Analysis method for Number of Target Loci对应的参数行修改为 1 或 2, 本参数是对搜索的目标位点组合包含的位点数目进行设定, 1 代表 Fixed Genome Size, 2 代表 Batch Genome Size 两种方法, 默认值是 1。该参数修改为 1 时, 表示按照指定一个的位点数据进行单轮计算, 需要同时修改#22.Fixed Genome Size = n 对应的参数行, 取值范围为大于等于 2 的正整数, 默认值是 3, 即筛选位点数量是 3 的目标位点组合; 该参数修改为 2 时, 表示按照位点数量区间从小到大逐个位点递增指定多个的位点数据进行单轮计算, 需要同时修改#23.Batch Genome Size from n 和#24.Batch Genome Size to m 对应的参数行, 取值范围为大于等于 2 的正整数, 默认值是 2-10, 即筛选位点数量分别是 2、3、4……10 的目标位点组合。
- (21) 将#25. Times of Calculation Cycles = n 对应的参数行修改为正整数, 本参数是对(2)-(20)步骤设置的参数保持一致并进行重复计算, 默认值是 1, 即只计算一轮上述参数的位点组合筛选。当参数行修改为 2 时, 即执行两轮上述参数的位点组合筛选, 以此类推。
- (22) 将#26.Result Format对应的参数行修改为 1-4 中某个值, 本参数设置位点组合筛选结果的输出文件格式, 1 代表 Unit Results with All Generation, 2 代表 Unit Results with Last Generation, 3 代表 Merged Results with All Generation, 4 代表 Merged Results with Last Generation, 1 和 2 对多轮计算的筛选结果输出为多个独立文件, 3 和 4 对多轮计算的筛选结果输出为一个合并文件, 1 和 3 将输出每一代群体的最优位点组合, 2 和 4 只输出最后一代群体的最优位点组合, 默认值是 4。所谓的多轮计算是由#21.Analysis method for Number of Target Loci 的方法和#25. Times of Calculation Cycles = n 的参数值决定, 当前者选择 Batch Genome Size 方法时, 或者后者设置的参数值大于 1 时, 将会执行多轮计算, 此时建议采用 2 和 4。
- (23) 保存参数配置文件 GASetting.txt。
- (24) 运行命令 mono LociScan-C.exe GASetting.txt, 等待 LociScan 的后台计算, 运行成功或运

行失败均将会在命令行界面给出提示（图 23），运行成功时同时会显示位点组合筛选消耗的计算时间，运行失败将不显示计算时间。

- (25) 切换到#7.Output Path 设定的结果文件存放路径对应的文件目录，可以在对应的“LociScan_GA_CSVFileName_CreateDate_CreateTime.csv”文件中查看分析结果。位点组合筛选结果的输出格式具体内容参见 3.5 小节。



```
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]#
```

图 23 CLI 版本的 LociScan 运行成功或运行失败的提示区域

3.9.2 位点组合筛选——穷举算法

位点组合筛选——穷举算法，指通过穷举算法筛选出指定位点数目且最有效的位点组合，适应度函数是以鉴定植物品种为目标的 8 种位点组合评价指标，输出结果将同时包含对应位点组合的适应度函数值。目前支持的训练数据（输入数据）格式为.csv。

3.9.2.1 GUI 版本的穷举算法操作

在 GUI 版本的 LociScan 软件中执行基于穷举算法的位点组合筛选，具体步骤如下：

- (1) 在 Operation 区域中选择 Loci combination screening。
- (2) 在图 10 的 Marker Type 区域选择训练数据对应的分子标记类型，包括 SNP/InDel (e.g.AT, CG or AB)和 SSR/STR (e.g.188/266)两种类型，默认值是 SNP/InDel (e.g.AT, CG or AB)。训练数据的数据格式具体内容参见 3.3 小节。
- (3) 在图 10 的 Data Structure 区域中选择训练数据的数据结构，Row:Loci; Col:Samples 表示行为位点列为样本，Row: Samples; Col:Loci 表示行为样本列为位点，默认值是 Row:Loci; Col:Samples。训练数据的数据结构具体内容参见 3.3 小节。
- (4) 在图 10 的 Data File Path 区域中，单击 Open 按钮，弹出 Open File 对话框，选择需要分析的数据文件，然后单击“打开”按钮等待数据文件的读取。
- (5) 在图 24 的 TabPages 区域选择模型算法，包括 Genetic Algorithm 和 Exhaustive Algorithm 两种方法，请选择 Exhaustive Algorithm。注意：本算法没有任何参数需要设置。由于本算法需要占用大量的计算资源，请时刻关注计算资源是否被程序耗尽而溢出，出现程序异常时，请尝试降低 Number of Target Loci 的参数值。
- (6) 在图 20 的颜色为 LightGreen 区域中，设置 Fitness Function 的值，其表示品种识别率方法，包括 R-VDP、C-VDP、P-VDP、TDP、R-VDP-reciprocal、C-VDP-reciprocal、TDP-reciprocal 和 O-VDP，默认值为 R-VDP。R-VDP、C-VDP、P-VDP 和 TDP 的具体统计方法请参考 VDP 相关文献[1]，O-VDP 请参考 LociScan 相关文献，这五个函数均为搜索其对应统计方法的最大值，R-VDP-reciprocal 等三个函数则反之。
- (7) 在图 20 的颜色为 LightGreen 区域中，设置 Variety Threshold Setting 的方法和阈值，本参数是将同一品种的位点差异阈值交给用户来进行定义，包括 Ratio of Different Loci 和 Number of Different Loci 两种方法，默认值是 Number of Different Loci。前者表示采用“差异位点百分比”小于用户设定阈值方式进行定义同一品种，取值范围为大于 0 且小于等于 1 的小数，默认值是 0.05，即差异位点占全部位点的比例小于 5%为同一品种；后者表示采用“差异位点数”小于用户设定阈值方式进行定义同一品种，取值范围为大于 1 的正整数，默认值是 1，即差异位点数小于 1 为同一品种。

- (8) 在图 20 的颜色为 Aqua 区域中，设置 Number of Target Loci 的方法和参数值，本参数是对搜索的目标位点组合包含的位点数目进行设定，包括 **Fixed Genome Size** 和 **Batch Genome Size** 两种方法，分别代表“目标位点组合的位点数”和“目标位点组合的位点数量区间”输入相应的值，方法默认选择 **Fixed Genome Size**。前者是按照指定一个的位点数进行单轮计算，取值范围为大于等于 2 的正整数，默认值是 3，即筛选位点数量是 3 的目标位点组合；后者是按照位点数量区间从小到大逐个位点递增指定多个的位点数据数据进行单轮计算，取值范围为大于等于 2 的正整数，默认值是 2-10，即筛选位点数量分别是 2、3、4……10 的目标位点组合。
- (9) 在图 20 的颜色为 Blue 区域中，设置 Cycle calculation times 的参数值，本参数是对(1)-(8)步骤设置的参数保持一致并进行重复计算，默认值是 1，即只计算一轮上述参数的位点组合筛选。
- (10) 在图 21 的 File Format 区域中设置位点组合筛选结果的输出文件格式，包括 **Unit Results** 和 **Merged Results**，前者对多轮计算的筛选结果输出为多个独立文件，后者对多轮计算的筛选结果输出为一个合并文件，默认值是 **Merged Results**。所谓的多轮计算是由 Number of Target Loci 的方法和 Cycle calculation times 的参数值决定，当前者选择 **Batch Genome Size** 方法时，或者后者设置的参数值大于 1 时，将会执行多轮计算，此时建议采用 **Merged Results**。
- (11) 点击“OK”按钮，等待 LociScan 的后台计算，运行成功或运行失败均将会在主界面右下角给出提示（图 22），运行成功时同时会显示位点组合筛选消耗的计算时间，运行失败将不显示计算时间。
- (12) 点击“Open Result Folder”，可以在对应的“LociScan_EA_CSVFileName_CreateDate_CreateTime.csv”文件中查看分析结果。位点组合筛选结果的输出格式具体内容参见 3.5 小节。

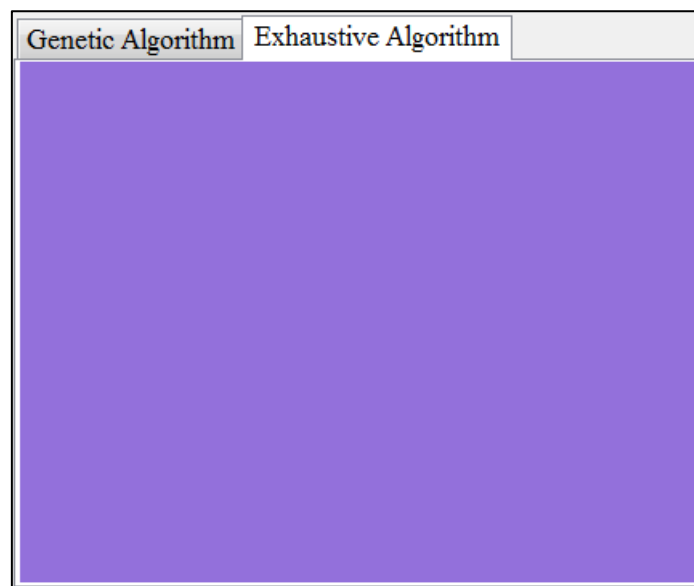


图 24 穷举算法模型参数设置区域

3.9.2.2 CLI 版本的穷举算法操作

在 CLI 版本的 LociScan 软件中执行基于穷举算法的位点组合筛选，具体步骤如下：

- (1) 打开参数配置文件 **GASetting.txt**，注意不要删除参数配置文件中的注释行和参数行，只根据以下操作步骤修改对应参数行内容。
- (2) 将 **#1.Operation** 对应的参数行修改为 1。1 代表 **Loci combination screening**，2 代表 **Loci**

combination evaluation, 3 代表 Data Cleaning。

- (3) 将#2.Marker Type 对应的参数行修改为 1 或 2。1 代表训练数据对应的分子标记类型为 SSR/STR, 2 代表 SNP/InDel。训练数据的数据格式具体内容参见 3.3 小节。
- (4) 将#3.Data Structure 对应的参数行修改为 1 或 2。1 表示行为位点、列为样本, 2 表示行为样本、列为位点。训练数据的数据结构具体内容参见 3.3 小节。
- (5) 将#4.Different Allele for SSR: Base-pairs Number > n bp 对应的参数行修改为 0 或正整数, 当#2.Marker Type 对应的参数行为 1 时才需要调整该参数, 具体内容参加 3.3 小节。
- (6) 在#5.Data File Path (.csv) 对应的参数行输入需要分析的基因型数据文件路径, 文件路径需要包含训练数据文件名和扩展名。
- (7) 在#6.Loci File Path (.csv) 对应的参数行输入 NULL。
- (8) 在#7.Output Path 对应的参数行输入结果文件的存放路径, 文件路径不需要包含结果文件的文件名和扩展名。
- (9) 将#8.Algorithm 对应的参数行修改为 2。1 代表 Genetic Algorithm, 2 代表 Exhaustive Algorithm。
- (10) 将#17.Fitness Function 对应的参数行修改为 1-8 中某个值, 设置筛选位点组合采用的适应度函数方法, 1 代表 R-VDP、2 代表 C-VDP、3 代表 P-VDP、4 代表 TDP、5 代表 R-VDP-reciprocal、6 代表 C-VDP-reciprocal、7 代表 TDP-reciprocal 和 8 代表 O-VDP, 默认值为 1。R-VDP、C-VDP、P-VDP 和 TDP 的具体统计方法请参考 VDP 相关文献[1], O-VDP 请参考 LociScan 相关文献, 这五个函数均为搜索其对应统计方法的最大值, R-VDP-reciprocal 等三个函数则反之。
- (11) 将#18.Variety Threshold Setting Type 对应的参数行修改为 1 或 2, 本参数是将同一品种存在差异位点的容忍度交给用户来进行定义, 1 代表 Ratio of Different Loci, 2 代表 Number of Different Loci 两种方法, 默认值是 2。该参数修改为 1 时, 表示采用“差异位点比率”小于用户设定阈值的方式进行定义同一品种, 需要同时修改#19.Ratio of Different Loci < n 对应的参数行, 取值范围为大于 0 且小于等于 1 的小数, 默认值是 0.05, 即差异位点占全部位点的比例小于 5%为同一品种; 该参数修改为 2 时, 表示采用“差异位点数”小于用户设定阈值的方式进行定义同一品种, 需要同时修改#20.Number of Different Loci < n 对应的参数行, 取值范围为大于 1 的正整数, 默认值是 1, 即差异位点数小于 1 为同一品种。
- (12) 将#21.Analysis method for Number of Target Loci 对应的参数行修改为 1 或 2, 本参数是对搜索的目标位点组合包含的位点数目进行设定, 1 代表 Fixed Genome Size, 2 代表 Batch Genome Size 两种方法, 默认值是 1。该参数修改为 1 时, 表示按照指定一个的位点数目进行单轮计算, 需要同时修改#22.Fixed Genome Size = n 对应的参数行, 取值范围为大于等于 2 的正整数, 默认值是 3, 即筛选位点数量是 3 的目标位点组合; 该参数修改为 2 时, 表示按照位点数量区间从小到大逐个位点递增指定多个的位点数据进行单轮计算, 需要同时修改#23.Batch Genome Size from n 和#24.Batch Genome Size to m 对应的参数行, 取值范围为大于等于 2 的正整数, 默认值是 2-10, 即筛选位点数量分别是 2、3、4……10 的目标位点组合。
- (13) 将#25. Times of Calculation Cycles = n 对应的参数行修改为正整数, 本参数是对(2)-(12)步骤设置的参数保持一致并进行重复计算, 默认值是 1, 即只计算一轮上述参数的位点组合筛选。当参数行修改为 2 时, 即执行两轮上述参数的位点组合筛选, 以此类推。
- (14) 将#26.Result Format 对应的参数行修改为 2 或 4, 本参数设置位点组合筛选结果的输出文件格式, 2 代表 Unit Results, 4 代表 Merged Results, 前者对多轮计算的筛选结果输出为多个独立文件, 后者对多轮计算的筛选结果输出为一个合并文件, 默认值是 4。所

谓的多轮计算是由#21.Analysis method for Number of Target Loci的方法和#25. Times of Calculation Cycles = n的参数值决定，当前者选择Batch Genome Size方法时，或者后者设置的参数值大于1时，将会执行多轮计算，此时建议采用4。

- (15) 其他参数不需要调整，保存参数配置文件 GASetting.txt。
- (16) 运行命令 mono LociScan-C.exe GASetting.txt，等待 LociScan 的后台计算，运行成功或运行失败均将会在命令行界面给出提示（图 23），运行成功时同时会显示位点组合筛选消耗的计算时间，运行失败将不显示计算时间。
- (17) 切换到#7.Output Path 设定的结果文件存放路径对应的文件目录，可以在对应的“LociScan_EA_CSVFileName_CreateDate_CreateTime.csv”文件中查看分析结果。位点组合筛选结果的输出格式具体内容参见 3.5 小节。

3.9.3 位点组合验证

位点组合验证指利用训练数据评价用户输入的多个组合位点的适应度函数值，适应度函数是以鉴定植物品种为目标的8种位点组合评价指标，位点组合验证结果将同时列出不能识别的样品名称，目前支持的数据格式为.csv。

3.9.3.1 GUI 版本的位点组合验证

在 GUI 版本的 LociScan 软件中执行位点组合验证，具体步骤如下：

- (1) 在 Operation 区域中选择 Loci combination evaluation。
- (2) 在图 13 的 Marker Type 区域选择训练数据对应的分子标记类型，包括 SNP/InDel (e.g.AT, CG or AB)和 SSR/STR (e.g.188/266)两种类型，默认值是 SNP/InDel (e.g.AT, CG or AB)。训练数据的数据格式具体内容参见 3.3 小节。
- (3) 在图 13 的 Data Structure 区域中选择训练数据的数据结构，Row:Loci; Col:Samples 表示行为位点列为样本，Row: Samples; Col:Loci 表示行为样本列为位点，默认值是 Row:Loci; Col:Samples。训练数据的数据结构具体内容参见 3.3 小节。
- (4) 在图 13 的 Data File Path 区域中，单击 Open 按钮，弹出 Open File 对话框，选择需要分析的数据文件，然后单击 Open 按钮等待数据文件的读取。
- (5) 在图 13 的 Loci File Path 区域中，单击 Open 按钮，弹出 Open File 对话框，选择需要验证的位点组合的数据文件，然后单击 Open 按钮等待数据文件的读取。待验证位点组合数据的具体格式见 3.4 小节。
- (6) 在图 20 的颜色为 LightGreen 区域中，设置 Fitness Function 的值，其表示品种识别率方法，包括 R-VDP、C-VDP、P-VDP、TDP 和 O-VDP，默认值为 R-VDP。R-VDP、C-VDP、P-VDP 和 TDP 的具体统计方法请参考 VDP 相关文献[1]，O-VDP 请参考 LociScan 相关文献，这五个函数均为搜索其对应统计方法的最大值。
- (7) 在图 20 的颜色为 LightGreen 区域中，设置 Variety Threshold Setting 的方法和阈值，本参数是将同一品种的位点差异阈值交给用户来进行定义，包括 Ratio of Different Loci 和 Number of Different Loci 两种方法，默认值是 Number of Different Loci。前者表示采用“差异位点百分比”小于用户设定阈值方式进行定义同一品种，取值范围为大于 0 且小于等于 1 的小数，默认值是 0.05，即差异位点占全部位点的比例小于 5%为同一品种；后者表示采用“差异位点数”小于用户设定阈值方式进行定义同一品种，取值范围为大于 1 的正整数，默认值是 1，即差异位点数小于 1 为同一品种。
- (8) 点击“OK”按钮，等待 LociScan 的后台计算，运行成功或运行失败均将会在主界面右下角给出提示（图 22），运行成功时同时会显示位点组合筛选消耗的计算时间，运行失败将不显示计算时间。
- (9) 点击“Open Result Folder”，可以在对应的“LociCheck_FitnessFuctionName_CSVFileName_CreateDate_CreateTime.csv”文件中查看分析结果。位点组合验证结果的输出格式具体

内容参见 3.6 小节。

3.9.3.2 CLI 版本的位点组合验证

在 CLI 版本的 LociScan 软件中执行位点组合验证，具体步骤如下：

- (1) 打开参数配置文件 `GASetting.txt`，注意不要删除参数配置文件中的注释行和参数行，只根据以下操作步骤修改对应参数行内容。
- (2) 将 `#1.Operation` 对应的参数行修改为 2。1 代表 Loci combination screening, 2 代表 Loci combination evaluation, 3 代表 Data Cleaning。
- (3) 将 `#2.Marker Type` 对应的参数行修改为 1 或 2。1 代表训练数据对应的分子标记类型为 SSR/STR, 2 代表 SNP/InDel。训练数据的数据格式具体内容参见 3.3 小节。
- (4) 将 `#3.Data Structure` 对应的参数行修改为 1 或 2。1 表示行为位点、列为样本, 2 表示行为样本、列为位点。训练数据的数据结构具体内容参见 3.3 小节。
- (5) 将 `#4.Different Allele for SSR: Base-pairs Number > n bp` 对应的参数行修改为 0 或正整数, 当 `#2.Marker Type` 对应的参数行为 1 时才需要调整该参数, 具体内容参见 3.3 小节。
- (6) 在 `#5.Data File Path (.csv)` 对应的参数行输入需要分析的基因型数据文件路径, 文件路径需要包含训练数据文件名和扩展名。
- (7) 在 `#6.Loci File Path (.csv)` 对应的参数行输入需要分析的位点组合数据文件路径, 文件路径需要包含位点组合数据文件名和扩展名。
- (8) 在 `#7.Output Path` 对应的参数行输入结果文件的存放路径, 文件路径不需要包含结果文件的文件名和扩展名。
- (9) 将 `#17.Fitness Function` 对应的参数行修改为 1-8 中某个值, 设置评估用户输入的位点组合采用的适应度函数方法, 1 代表 R-VDP、2 代表 C-VDP、3 代表 P-VDP、4 代表 TDP、5 代表 R-VDP-reciprocal、6 代表 C-VDP-reciprocal、7 代表 TDP-reciprocal 和 8 代表 O-VDP, 默认值为 1。R-VDP、C-VDP、P-VDP 和 TDP 的具体统计方法请参考 VDP 相关文献[1], O-VDP 请参考 LociScan 相关文献, 这五个函数均为搜索其对应统计方法的最大值, R-VDP-reciprocal 等三个函数则反之。
- (10) 将 `#18.Variety Threshold Setting Type` 对应的参数行修改为 1 或 2, 本参数是将同一品种存在差异位点的容忍度交给用户来进行定义, 1 代表 Ratio of Different Loci, 2 代表 Number of Different Loci 两种方法, 默认值是 2。该参数修改为 1 时, 表示采用“差异位点比率”小于用户设定阈值的方式进行定义同一品种, 需要同时修改 `#19.Ratio of Different Loci < n` 对应的参数行, 取值范围为大于 0 且小于等于 1 的小数, 默认值是 0.05, 即差异位点占全部位点的比例小于 5% 为同一品种; 该参数修改为 2 时, 表示采用“差异位点数”小于用户设定阈值的方式进行定义同一品种, 需要同时修改 `#20.Number of Different Loci < n` 对应的参数行, 取值范围为大于 1 的正整数, 默认值是 1, 即差异位点数小于 1 为同一品种。
- (11) 其他参数不需要调整, 保存参数配置文件 `GASetting.txt`。
- (12) 运行命令 `mono LociScan-C.exe GASetting.txt`, 等待 LociScan 的后台计算, 运行成功或运行失败均将会在命令行界面给出提示 (图 23), 运行成功时同时会显示位点组合筛选消耗的计算时间, 运行失败将不显示计算时间。
- (13) 切换到 `#7.Output Path` 设定的结果文件存放路径对应的文件目录, 可以在对应的“`LociCheck_FitnessFunctionName_CSVFileName_CreateDate_CreateTime.csv`”文件中查看分析结果。位点组合验证结果的输出格式具体内容参见 3.6 小节。

3.9.4 数据清洗

数据清洗是为了保障位点组合筛选的正常运行而对训练数据的数据内容进行校验和处理。校验方法是根据训练数据评估所有位点是否存在基因型变异, 即判断在不同样品之间存

在不同基因型。如果发现位点在不同样品之间的基因型数据没有任何差异，将直接删除该位点及其对应数据，输出数据只保留存在基因型变异的位点及其对应数据。目前支持的训练数据（输入数据）格式为.csv。

3.9.4.1 GUI 版本的数据清洗

在 GUI 版本的 LociScan 软件中执行数据清洗，具体步骤如下：

- (1) 在 Operation 区域中选择 Data cleaning。
- (2) 在图 10 的 Marker Type 区域选择训练数据对应的分子标记类型，包括 SNP/InDel (e.g.AT, CG or AB)和 SSR/STR (e.g.188/266)两种类型，默认值是 SNP/InDel (e.g.AT, CG or AB)。训练数据的数据格式具体内容参见 3.3 小节。
- (3) 在图 10 的 Data Structure 区域中选择训练数据的数据结构，Row:Loci; Col:Samples 表示行为位点列为样本，Row: Samples; Col:Loci 表示行为样本列为位点，默认值是 Row:Loci; Col:Samples。训练数据的数据结构具体内容参见 3.3 小节。
- (4) 在图 10 的 Data File Path 区域中，单击 Open 按钮，弹出 Open File 对话框，选择需要分析的数据文件，然后单击“打开”按钮等待数据文件的读取。
- (5) 点击“OK”按钮，等待 LociScan 的后台计算，运行成功或运行失败均将会在主界面右下角给出提示（图 22），运行成功时同时会显示位点组合筛选消耗的计算时间，运行失败将不显示计算时间。
- (6) 点击“Open Result Folder”，可以在对应的“DataClean_CSVFileName_CreateDate_CreateTime.csv”文件中查看分析结果。数据清洗的输出格式具体内容参见 3.7 小节。

3.9.4.2 CLI 版本的数据清洗

在 CLI 版本的 LociScan 软件中执行数据清洗，具体步骤如下：

- (1) 打开参数配置文件 GASetting.txt，注意不要删除参数配置文件中的注释行和参数行，只根据以下操作步骤修改对应参数行内容。
- (2) 将#1.Operation 对应的参数行修改为 3。1 代表 Loci combination screening，2 代表 Loci combination evaluation，3 代表 Data Cleaning。
- (3) 将#2.Marker Type 对应的参数行修改为 1 或 2。1 代表训练数据对应的分子标记类型为 SSR/STR，2 代表 SNP/InDel。训练数据的数据格式具体内容参见 3.3 小节。
- (4) 将#3.Data Structure 对应的参数行修改为 1 或 2。1 表示行为位点、列为样本，2 表示行为样本、列为位点。训练数据的数据结构具体内容参见 3.3 小节。
- (5) 将#4.Different Allele for SSR: Base-pairs Number > n bp 对应的参数行修改为 0 或正整数，当#2.Marker Type 对应的参数行为 1 时才需要调整该参数，具体内容参见 3.3 小节。
- (6) 在#5.Data File Path (.csv)对应的参数行输入需要分析的基因型数据文件路径，文件路径需要包含训练数据文件名和扩展名。
- (7) 在#6.Loci File Path (.csv)对应的参数行输入 NULL。
- (8) 在#7.Output Path 对应的参数行输入结果文件的存放路径，文件路径不需要包含结果文件的文件名和扩展名。
- (9) 其他参数不需要调整，保存参数配置文件 GASetting.txt。
- (10) 运行命令 mono LociScan-C.exe GASetting.txt，等待 LociScan 的后台计算，运行成功或运行失败均将会在命令行界面给出提示（图 23），运行成功时同时会显示位点组合筛选消耗的计算时间，运行失败将不显示计算时间。
- (11) 切换到#7.Output Path 设定的结果文件存放路径对应的文件目录，可以在对应的“DataClean_CSVFileName_CreateDate_CreateTime.csv”文件中查看分析结果。数据清洗结果的输出格式具体内容参见 3.7 小节。

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.