
DNA 序列特征性片段鉴定软件(IdenDSS)使用说明

Version 0.2

October 8, 2022

1 软件简介

IdenDSS 软件用于 DNA 序列特征性片段(DNA Signature Sequence, DSS)分子标记开发。DSS 是指与来源于其他分类单元相比(下称, 背景分类单元(background taxon)), 只出现在某个特定分类单元(下称, 目标分类单元(target taxon))中的 DNA 序列, 该分子标记可以用于物种鉴定, 在中药鉴定、海关检查等领域具有潜在价值。

目前 IdenDSS 包括四个模块: 建立索引(index), 鉴定(identify), 插件(plugin), 验证(validate)。为方便不同用户使用, 目前 IdenDSS 软件提供命令行版(兼容所有系统)和图形化界面版(Windows 系统), 用户可以根据需要自行选择。**建议有条件的用户在 Linux 系统下使用命令行版本以使用完整功能。**

2 软件安装

为方便用户下载，已将命令行版本、图形化界面版本和示例文件打包到百度网盘，下载地址如下：

链接：https://pan.baidu.com/s/1_4lXXQk21xulZSMGsThrw?pwd=fx8j

提取码：fx8j

2.1 硬件与软件要求

CPU：推荐使用 4 线程及以上 CPU

内存：推荐使用 16G 及以上内存

必要依赖软件：

BLAST 2.9.0+及以上版本：

下载地址：<https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST>

非必要依赖软件：

KMC：用于使用高通量数据进行 DSS 验证

下载地址：<https://github.com/refresh-bio/KMC/releases>

Primer3：用于进行引物设计

下载地址：<https://github.com/primer3-org/primer3/releases>

2.2 命令行版本安装

打开命令行，使用如下命令安装：`pip install IdenDSS`

```

hzy@hmp1 ~$ pip install IdenDSS
Looking in indexes: https://pypi.tuna.tsinghua.edu.cn/simple
Collecting IdenDSS
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/02/a2/1cd9a073d69248abb0a6573f84737c0ec00209b27995d490b85ebc4ba3dc/IdenDSS-0.2.1-py2.py3-none-any.whl (29 kB)
Requirement already satisfied: biopython>=1.78 in /home/anaconda3/lib/python3.9/site-packages (from IdenDSS) (1.79)
Requirement already satisfied: numpy>=1.20.0 in /home/anaconda3/lib/python3.9/site-packages (from IdenDSS) (1.20.3)
Requirement already satisfied: pandas>=1.0.0 in /home/anaconda3/lib/python3.9/site-packages (from IdenDSS) (1.3.4)
Requirement already satisfied: python-dateutil>=2.7.3 in /home/anaconda3/lib/python3.9/site-packages (from pandas==1.0.0->IdenDSS) (2.8.2)
Requirement already satisfied: pytz>=2017.3 in /home/anaconda3/lib/python3.9/site-packages (from pandas==1.0.0->IdenDSS) (2021.3)
Requirement already satisfied: six>=1.5 in /home/anaconda3/lib/python3.9/site-packages (from python-dateutil>=2.7.3->pandas==1.0.0->IdenDSS) (1.16.0)
Installing collected packages: IdenDSS
Successfully installed IdenDSS-0.2.1

[notice] A new release of pip available: 22.2.1 -> 22.2.2
[notice] To update, run: pip install --upgrade pip

```

图 2.1 IdenDSS 的安装

在命令行输入 `IdenDSS -h` 出现如下信息代表安装成功

```

hzy@hmp1 ~/Demo$ IdenDSS -h
usage: IdenDSS [-h] {index,identify,plugin,validate} ...

This script was for identifying DNA signature sequences(DSS)

optional arguments:
  -h, --help            show this help message and exit

Available:
  {index,identify,plugin,validate}
    index               Generate database for DSS identification
    identify            Identification DSS based on database
    plugin              Some plugin for DSS results
    validate            Validate DSS using HTS file. MUST BE DSS, NOT COMBINED DSS
hzy@hmp1 ~/Demo$

```

图 2.2 IdenDSS 安装成功

注:

1. 本使用手册默认使用命令行版本的用户具备基础的命令行使用知识。
2. 命令行版本需要用户先安装 python 3.8 及以上版本。

2.3 图形化界面版本安装

从如下地址下载 IdenDSS.exe，直接双击打开使用，界面如下：

https://github.com/Hua-CM/IdenDSS/releases/download/v0.2.3/IdenDSS_windows_x64.exe

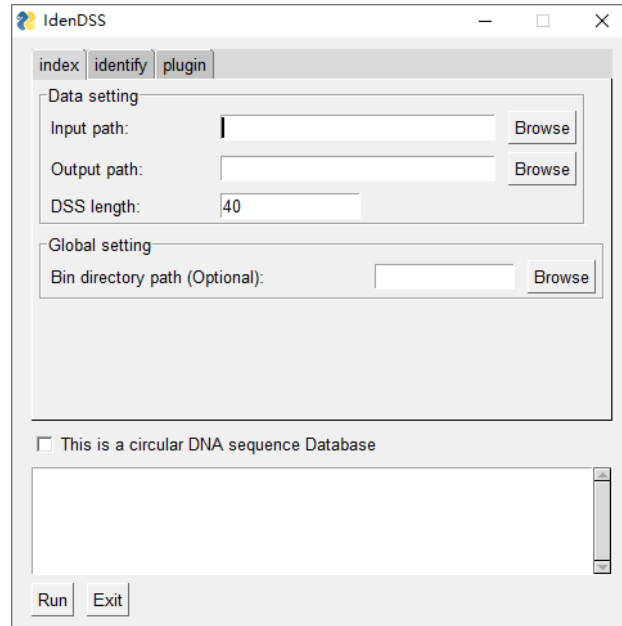


图 2.3 IdenDSS 图形化界面版本

2.4 示例文件

IdenDSS 使用简单，用户只需要提供两个输入文件：(1).用于开发 DSS 标记的 DNA 序列（fasta 格式）(下称**输入序列文件**)。该文件包含，且仅包含，目标分类单元的序列和背景分类单元的序列；(2). 目标分类单元和对应 DNA 序列的关系，**制表符分隔(TAB 键)**文本文件(下称**元信息文件**)，详见对示例文件的说明。

为便于用户使用，本软件以鉴定党参属(*Codonopsis*)三个物种的 DSS 为例，提供了对应的示例文件：**example.fasta** 和 **example_input.tsv**，下载地址如下：

<https://github.com/Hua-CM/IdenDSS/tree/v0.2/example>

1. example.fasta

```
hzy@hmp1 ~/Demo$ head example.fasta
>G140224LY0334
ACTTGGGCGAACGACGGGAATTGAACCCGCGAATGGTGGATTCACAATCCACTGCCTTGT
CCACTTGGCTACATCCGCCCTCTACTATGTTTTCCATAGAAAGAAAAAATGATATTCCA
ACTGTTATTGCATATTGAATATAATAACTGCACAAAAGATTGAGAAGAGGGTATAAATAT
AGAAATATAGAAAAAATAGGAAAGGAAAGTACGATACTCAAACAAAGCCCCCAAAAAAG
AAAAAGGGGGGACCAATAACGCCCTCTTGATAGAACAAGAAAGGGGTTATTGCTCCTTTT
CCTTCAAAAACCTCATATACACTCAGACCAAGATCTTAGCCATTTTTAGATGGGGCTTCGA
TAGCAGCTAGGTCTAGAGGGAAGTTATGAGCATTACGTTTCATGCATAACTTCCATACCAA
GGTTAGCCCGGTTAATAATATCAGCCCAAGTATTAATTACACGGCCTTGACTATCAACTA
CAGATTGGTTGAAATTGAAACCATTTAGGTTGAAAGCCATAGTACTGATACCTAAAGCAG
hzy@hmp1 ~/Demo$
```

图 2.4 IdenDSS 输入序列文件

example.fasta 中共记录了 44 条 DNA 序列，其中 17 条来自于分类单元党参 (*Codonopsis pilosula*) 的 2 个样本；26 条属于分类单元川党参 (*C. pilosula* subsp. *tangshen*) 的 2 个样本；1 条来自素花党参 (*C. pilosula* var. *modesta*) 的 1 个样本。

注: fasta 文件格式为生物医药领域通用的 DNA 序列存储格式，本软件默认使用者了解这种格式。

2. example_input.tsv

元信息文件共分为三列：

1. 第一列：目标分类单元名。名称由用户自定义，不可包含空格。
2. 第二列：样品名。每个分类单元往往会存在生物学重复，故本软件允许一个分单元有多个样品来源的数据。每个分类单元内部样品名要求无重复，使用字符或数字标注均可；
3. 第三列：对应样品的序列名，一个样品允许有多条序列，这些序列须存在于前述的 fasta 文件中。

```

hzy@hmpl ~/Demo$ cat example_input.tsv
Codonopsis_pilosula 1 G140224LY0334
Codonopsis_pilosula 2 dangshenscaffold_1
Codonopsis_pilosula 2 dangshenscaffold_1
Codonopsis_pilosula 2 dangshenscaffold_2
Codonopsis_pilosula 2 dangshenscaffold_3
Codonopsis_pilosula 2 dangshenscaffold_4
Codonopsis_pilosula 2 dangshenscaffold_5
Codonopsis_pilosula 2 dangshenscaffold_6
Codonopsis_pilosula 2 dangshenscaffold_7
Codonopsis_pilosula 2 dangshenscaffold_8
Codonopsis_pilosula 2 dangshenscaffold_9
Codonopsis_pilosula 2 dangshenscaffold_10
Codonopsis_pilosula 2 dangshenscaffold_11
Codonopsis_pilosula 2 dangshenscaffold_12
Codonopsis_pilosula 2 dangshenscaffold_13
Codonopsis_pilosula 2 dangshenscaffold_14
Codonopsis_pilosula 2 dangshenscaffold_15
Codonopsis_pilosula 2 dangshenscaffold_16
Codonopsis_angshen 1 chuandangshen2scaffold_1
Codonopsis_angshen 1 chuandangshen2scaffold_2
Codonopsis_angshen 1 chuandangshen2scaffold_3
Codonopsis_angshen 1 chuandangshen2scaffold_4
Codonopsis_angshen 1 chuandangshen2scaffold_5
Codonopsis_angshen 1 chuandangshen2scaffold_6
Codonopsis_angshen 1 chuandangshen2scaffold_7
Codonopsis_angshen 1 chuandangshen2scaffold_8
Codonopsis_angshen 1 chuandangshen2scaffold_9
Codonopsis_angshen 1 chuandangshen2scaffold_10
Codonopsis_angshen 1 chuandangshen2scaffold_11
Codonopsis_angshen 1 chuandangshen2scaffold_12
Codonopsis_angshen 1 chuandangshen2scaffold_13
Codonopsis_angshen 2 chuandangshenscaffold_1
Codonopsis_angshen 2 chuandangshenscaffold_2
Codonopsis_angshen 2 chuandangshenscaffold_3
Codonopsis_angshen 2 chuandangshenscaffold_4
Codonopsis_angshen 2 chuandangshenscaffold_5
Codonopsis_angshen 2 chuandangshenscaffold_6
Codonopsis_angshen 2 chuandangshenscaffold_7
Codonopsis_angshen 2 chuandangshenscaffold_8
Codonopsis_angshen 2 chuandangshenscaffold_9
Codonopsis_angshen 2 chuandangshenscaffold_10
Codonopsis_angshen 2 chuandangshenscaffold_11
Codonopsis_angshen 2 chuandangshenscaffold_12
Codonopsis_angshen 2 chuandangshenscaffold_13
Codonopsis_sumodesta 1 sudangshenscaffold_1

```

图 2.5 IdenDSS 元信息文件

如上图，example_input.tsv 记录的信息表明：G140224LY0334 属于 Codonopsis_pilosula 分类单元的样本 1；dangshenscaffold_1~dangshenscaffold_16 属于 Codonopsis_pilosula 分类单元样本 2；chuandangshen2scaffold_1~chuandangshen2scaffold_13 属于 Codonopsis_angshen 分类单元的样本 1；chuandangshenscaffold_1 ~ chuandangshenscaffold_13 属于 Codonopsis_angshen 分类单元样本 2；sudangshenscaffold_1 属于 Codonopsis_sumodesta 分类单元样本 1（当分类单元仅有一个样本时也需要标注）

以该 example_input.tsv 及其对应的 example_input.fasta 作为输入文件进行输

入，可以鉴定 *Codonopsis_pilosula*、*Codonopsis_angshe* 与 *Codonopsis_sumodesta* 这三个分类单元的 DSS 分子标记。更具体的而言，当目标分类单元为 *Codonopsis_pilosula* 时，背景分类单元为 *Codonopsis_angshe* 与 *Codonopsis_sumodesta*；当目标分类单元为 *Codonopsis_angshe* 时，背景分类单元为 *Codonopsis_pilosula* 与 *Codonopsis_sumodesta*；当目标分类单元为 *Codonopsis_sumodesta* 时，背景分类单元为 *Codonopsis_pilosula* 与 *Codonopsis_angshe*。

3 建立数据库索引

在使用 IdenDSS 鉴定 DSS 前需要先构建数据库索引

3.1 命令行界面操作

```
hzy@hmp1 ~/Demo$ IdenDSS index -h
usage: IdenDSS index [-h] -i INPUT [-l LENGTH] [-c] -o OUTPUT [--blast BIN_DIR]

optional arguments:
  -h, --help            show this help message and exit
  -i INPUT, --input INPUT
                        <File path> The genome fasta file
  -l LENGTH, --length LENGTH
                        <Int> DSS length <default=40>
  -c, --circular        the sequences are circular or not
  -o OUTPUT, --output OUTPUT
                        <File path> The genome fasta output file (a BLAST database would be constructed)
  --blast BIN_DIR       <Directory path> BLAST exec directory <If your BLAST software not in PATH>
```

图 3.1 建立索引模块(index)提示信息

必填选项说明：

- **-i/--input** 输入序列文件路径。
- **-l/--length** 用户需要鉴定的 DSS 长度。默认为 40 bp。
- **-o/--output** 建立的数据库的存储路径。

其他选项说明：

- **-c/--circular** 若使用的序列是环状 DNA（如植物叶绿体或动物线粒体 DNA 序列），用户需要启用该选项。默认不启动。
- **--blast** 若 BLAST 软件不在环境变量中，或用户希望使用指定版本的 BLAST 软件，可以使用该选项指定 BLAST 软件的目录。

以示例文件做为示范进行操作：

```
hzy@hmp1 ~/Demo$ mkdir database (1)
hzy@hmp1 ~/Demo$ IdenDSS index -i example.fasta -o database/example_db.fasta -l 40 (2)
IdenDSS database created success!
hzy@hmp1 ~/Demo$
```

图 3.2 建立索引

- (1) 创建一个用于存放数据库的文件夹

- (2) 建立数据库索引：输入文件为 `example.fasta`; 输出文件为 `example_db.fasta`, 索引长度为 40. 由于示例文件中的叶绿体不成环, 因此没有开启 `-c` 选项。

3.2 图形化界面操作

参数含义同命令行界面, 以示例文件作为示范进行操作:

- (1) 首先双击打开软件, 按照图 3.3 所示选取输入序列文件路径. 选取完毕, 如图 3.4. 需要注意: 本文档中所有的路径名称不能出现空格。
- (2) 使用同样的方法在 `output` 处选择存储数据库的文件夹路径(图 3.5, 图 3.6), 然后需要手动补全输出数据库名(图 3.7).
- (3) 根据实际情况指定 DSS 长度、选择是否指定 BLAST 软件路径和序列类型(图 3.8).
- (4) 指定 BLAST 软件路径示例(图 3.9).
- (5) 全部确认后点击 `RUN` 按钮(图 3.10). 出现如下信息表示构建索引成功(图 3.11).

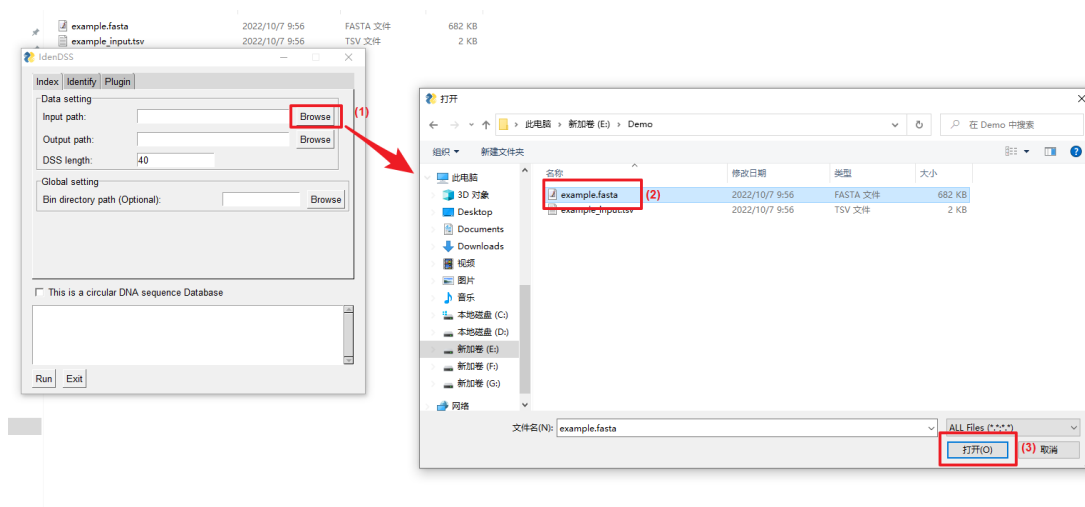


图 3.3 选择文件 1

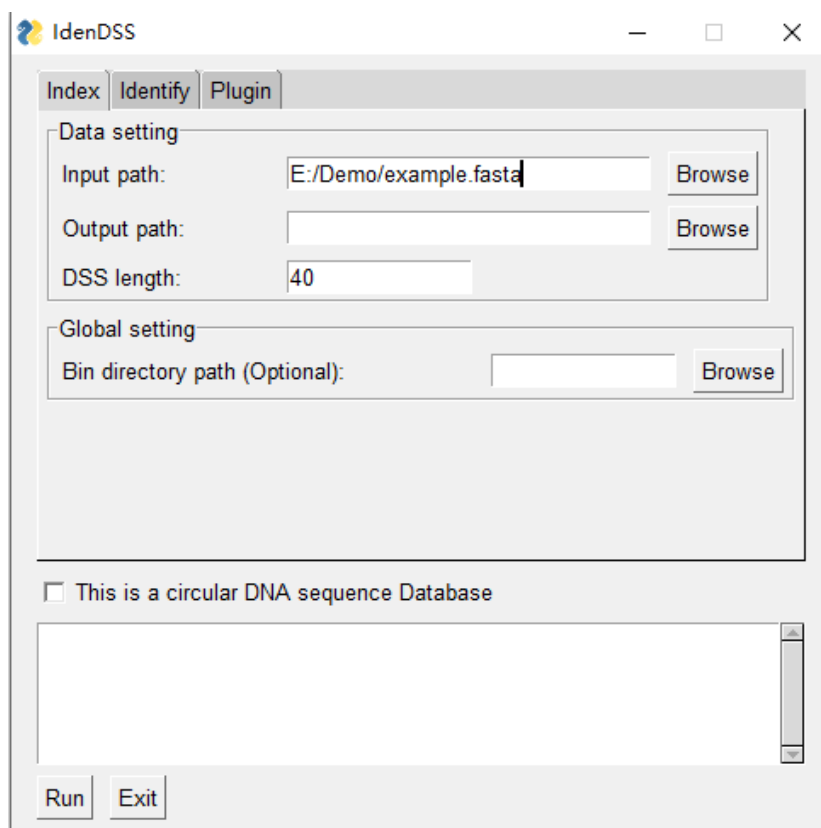


图 3.4 选择文件 2

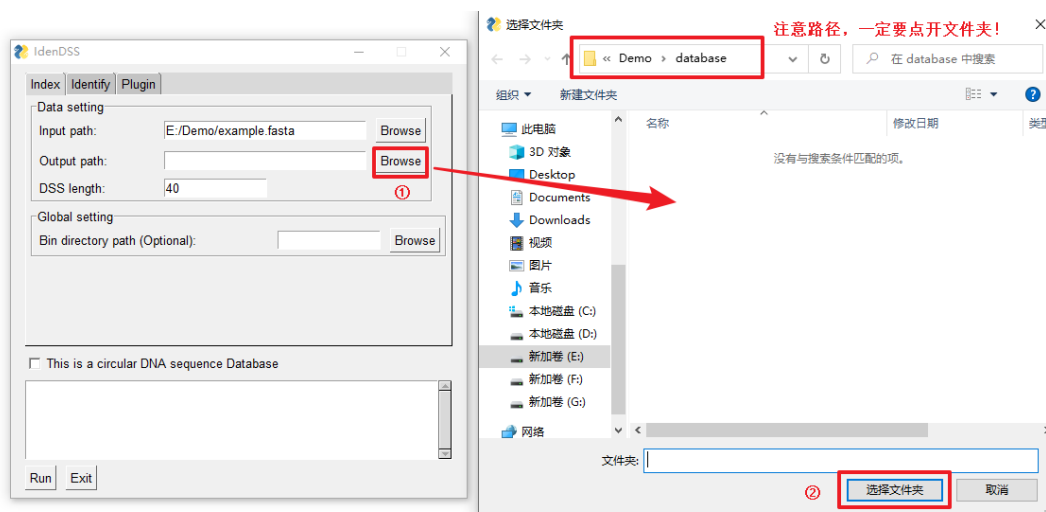


图 3.5 选择文件夹 1

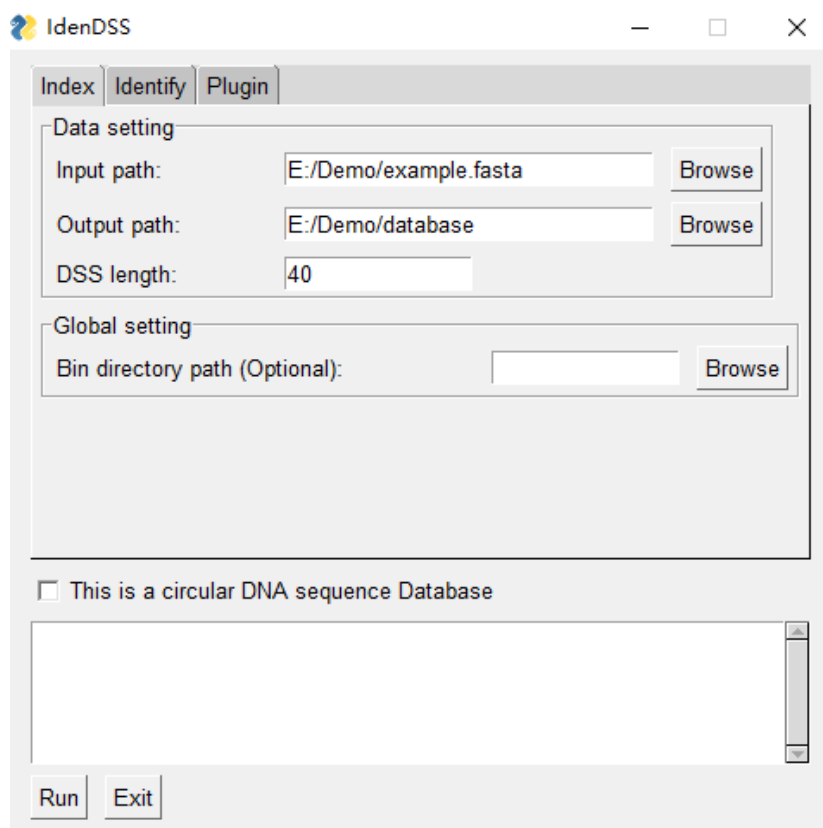


图 3.6 选择文件夹 2

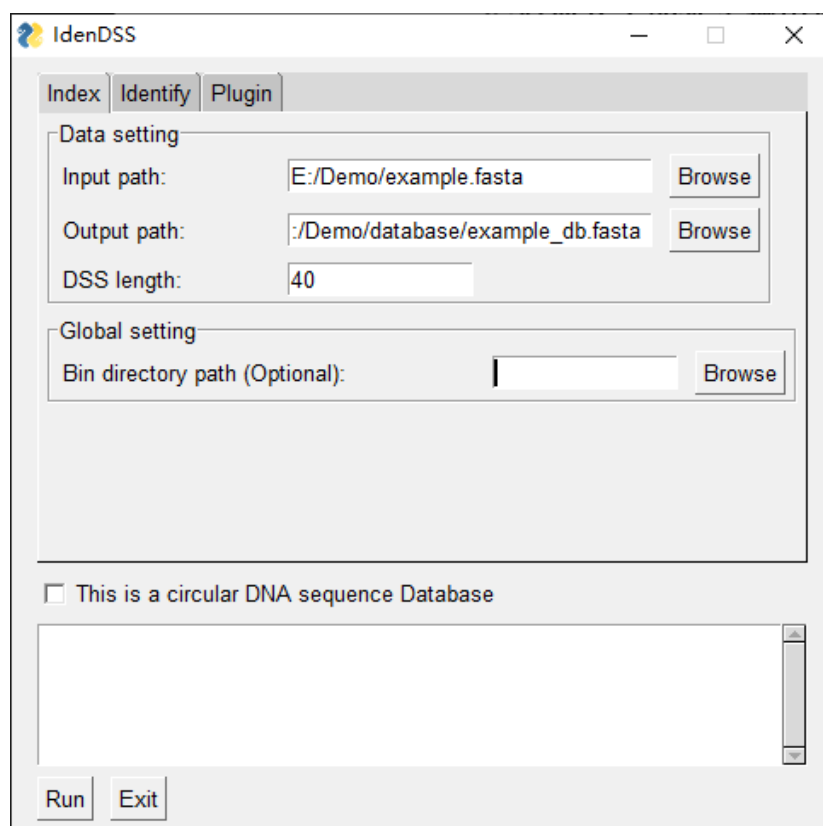


图 3.7 手动补全输出数据库名

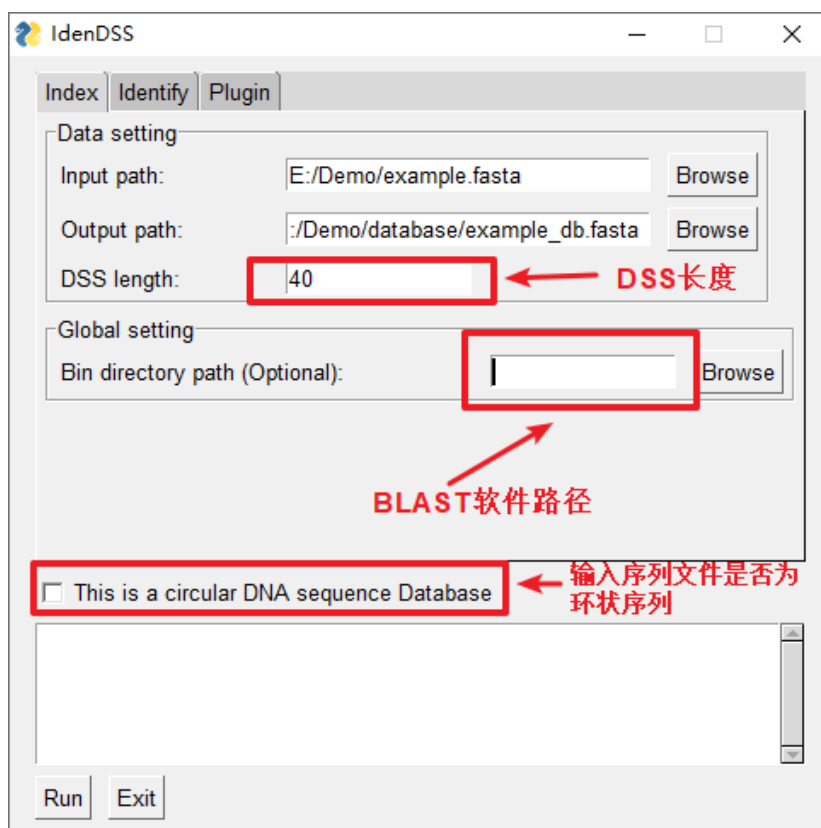


图 3.8 指定 DSS 长度

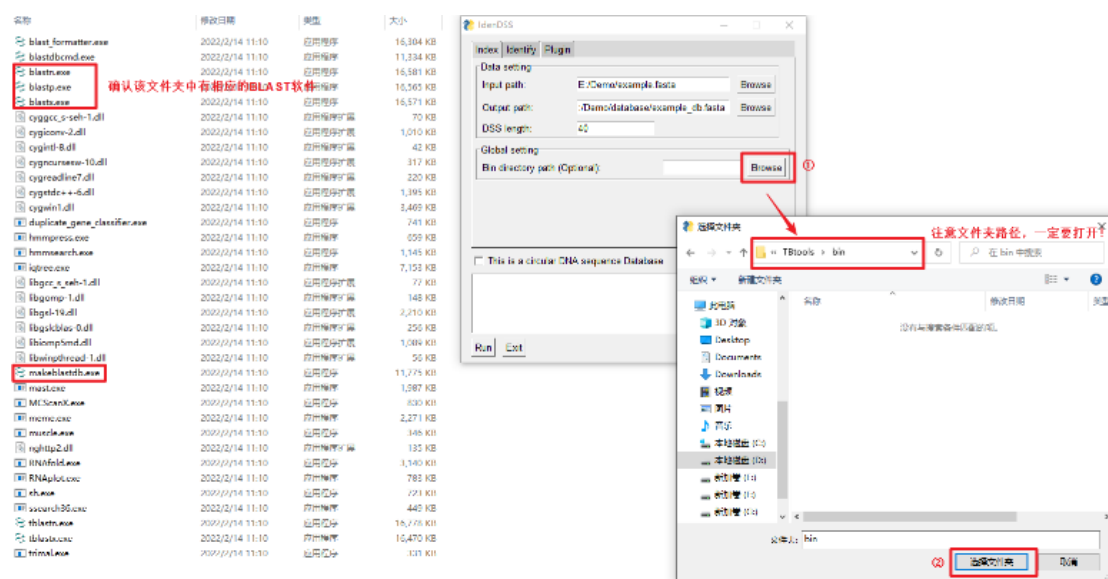


图 3.9 指定 BLAST 软件路径

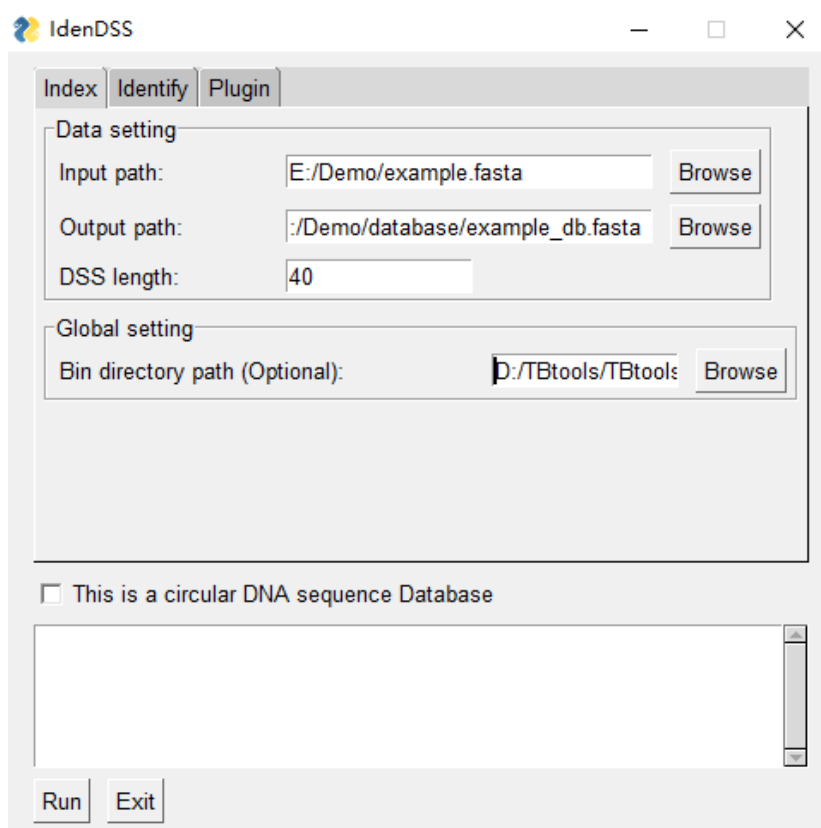


图 3.10 点击 Run 运行

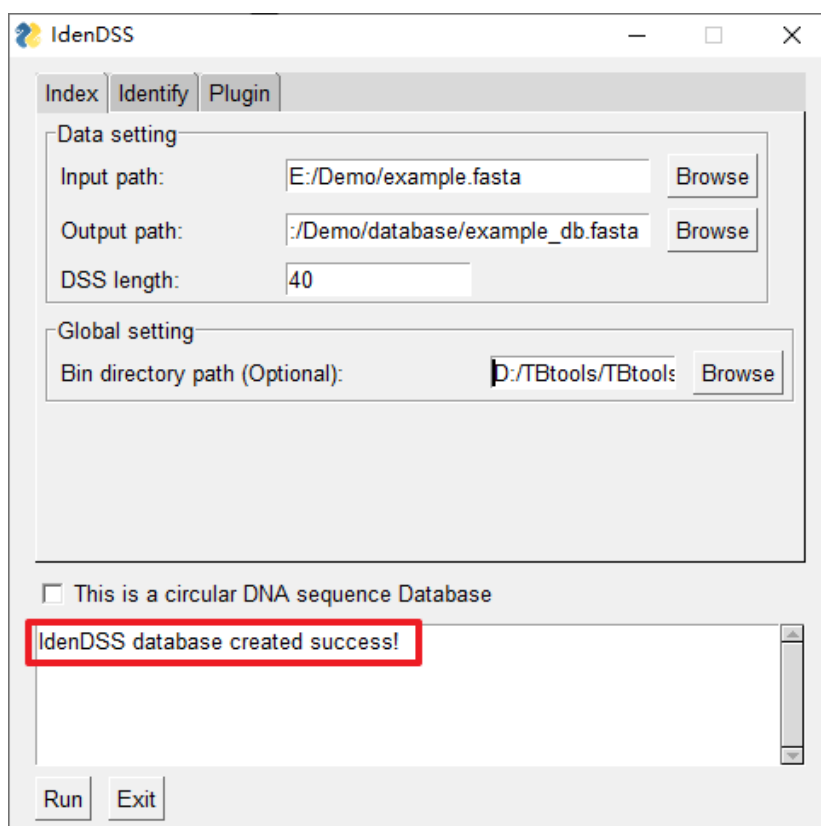


图 3.11 运行成功提示信息

3.3 其他注意事项

(1). 输入文件里可以同时包含成环序列和非成环序列（例如同时有成环叶绿体和不成环的叶绿体片段），若为此情况，则启动 `-c/--circular` 选项。

(2). 若需要同时鉴定多个长度的 DSS 标记，则直接建立最长长度的数据库索引即可。例如用户如需要分别鉴定一个目标分类单元 20bp、30bp、40bp、50bp 和 60bp 的 DSS 标记，则在建立数据库索引时设定 DSS 长度为 60 建立一次数据库索引即可。

4 鉴定 DSS

4.1 命令行界面操作

```
hzy@hmp1 ~/Demo$ IdenDSS identify -h
usage: IdenDSS identify [-h] -m META -d DATABASE [-l LENGTH] [-c] -o OUTPUT [-@ THREADS] [-t TMP] [--blast BIN_DIR]

optional arguments:
  -h, --help            show this help message and exit
  -m META, --meta META  <File path> The meta file
  -d DATABASE, --database DATABASE
                        <File path> Database fasta <Corresponding BLAST database file must in the same directory>
  -l LENGTH, --length LENGTH
                        <Int> DSS length <default=40>
  -c, --circular         the sequences are circular or not
  -o OUTPUT, --output OUTPUT
                        <File path> result directory
  -@ THREADS, --threads THREADS
                        <Int> Threads used in BLAST <Default 4>
  -t TMP, --tmp TMP      <Directory path> Temporary directory path (Default system temporary directory)
  --blast BIN_DIR        <Directory path> BLAST exec directory <If your BLAST software not in PATH>
```

图 4.1 鉴定(identity)模块提示信息

必填选项说明:

- **-m/--meta** 元信息文件路径.
- **-d/--database** 3.1 中建立的数据库路径.
- **-l/--length** 用户需要鉴定的 DSS 长度.默认为 40 bp.
- **-o/--output** 输出结果存储文件夹.

其他选项说明:

- **-c/--circular** 若使用的序列是环状 DNA (如植物叶绿体或动物线粒体 DNA 序列), 用户需要启用该选项.默认不启动.
- **-@/--threads** BLAST 软件并行的线程数。默认为 4.
- **-t/--tmp** 用户可以通过该选项指定 IdenDSS 运行过程中的临时文件存放路径.默认为系统临时文件夹.
- **--blast** 若 BLAST 软件不在环境变量中, 或用户希望使用指定版本的 BLAST 软件, 可以使用该选项指定 BLAST 软件的目录.

以示例文件进行操作：

```
hzy@hmpl ~/Demo$ mkdir results (1)
hzy@hmpl ~/Demo$ IdenDSS identify \
> -m example_input.tsv \
> -d database/example_db.fasta \
> -l 40 \
> -o results (2)
Begin to identify Codonopsis_angshen putative DSS, please wait for a moment!
Intraspeceis conserved k-mers, done.
Putative DSS identification, done
Parse results, done
Begin to identify Codonopsis_pilosula putative DSS, please wait for a moment!
Intraspeceis conserved k-mers, done.
Putative DSS identification, done (3)
Parse results, done
Begin to identify Codonopsis_sumodesta putative DSS, please wait for a moment!
Intraspeceis conserved k-mers, done.
Putative DSS identification, done
Parse results, done
All groups done!
hzy@hmpl ~/Demo$ ls results/ (4)
Codonopsis_angshen.txt Codonopsis_pilosula.txt Codonopsis_sumodesta.txt
```

图 4.2 使用鉴定模块鉴定 DSS

- (1) 创建一个用于存放结果的文件夹.
- (2) 鉴定 DSS.输入元信息文件为 `example_input.tsv`；输入数据库为 `example_db.fasta`；输出文件夹为 `results`，DSS 长度为 40. 由于示例文件中的叶绿体不成环，因此没有开启 `-c` 选项.
- (3) 软件运行过程中的提示信息.
- (4) 结果文件（具体信息含义见 4.3 节）.

4.2 图形化界面操作

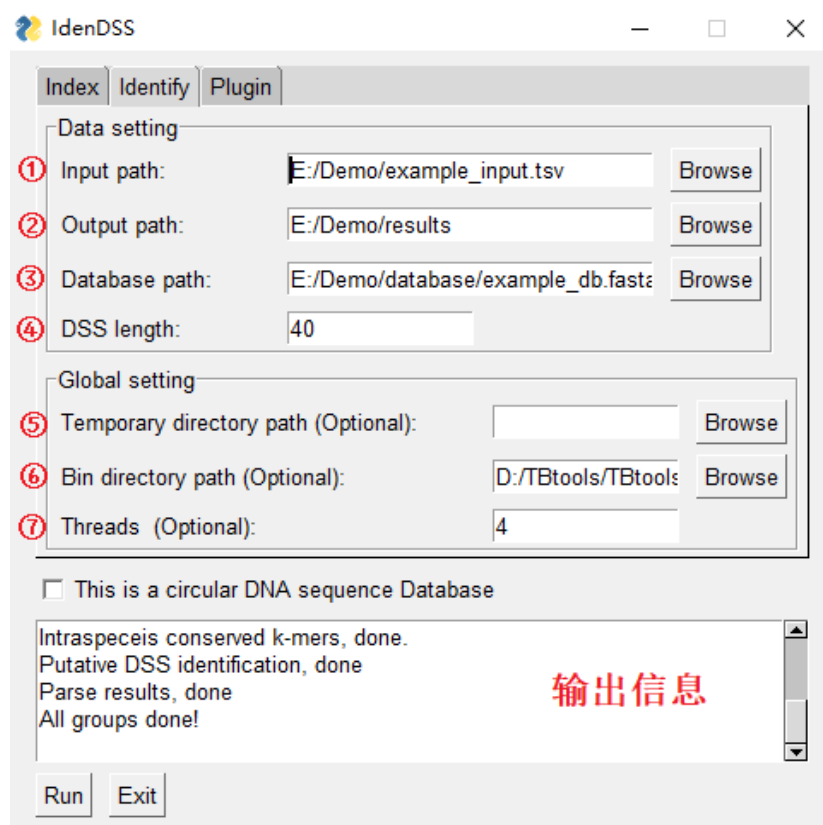


图 4.3 使用图形化界面鉴定模块鉴定 DSS

(1) 选取 Identify 标签

(2) 填选相关信息，操作同 3.2，各选项说明如下：

①对应-m/--meta，元信息文件路径.

②对应-o/--output，输出结果存储文件夹.

③对应-d/--database，3.2 中建立的数据库路径.

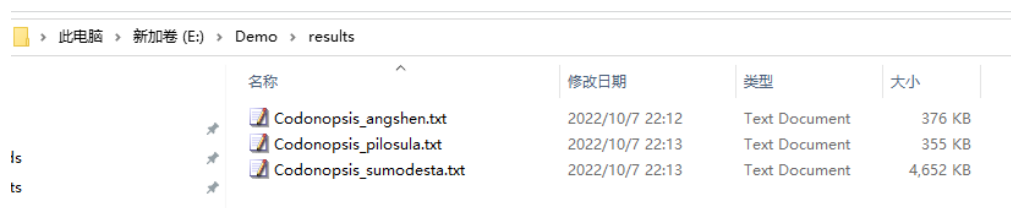
④对应-l/--length，用户需要鉴定的 DSS 长度.默认为 40 bp.

⑤对应-t/--tmp，用户可以通过该选项指定 IdenDSS 运行过程中的临时文件存放路径.默认为系统临时文件夹.

⑥对应--blast，若 BLAST 软件不在环境变量中，或用户希望使用指定版本的 BLAST 软件，可以使用该选项指定 BLAST 软件的目录.

⑦对应-@/--threads BLAST 软件并行的线程数。

(3) 点击 Run 按钮运行(图 4.3)。



名称	修改日期	类型	大小
Codonopsis_angshen.txt	2022/10/7 22:12	Text Document	376 KB
Codonopsis_pilosula.txt	2022/10/7 22:13	Text Document	355 KB
Codonopsis_sumodesta.txt	2022/10/7 22:13	Text Document	4,652 KB

图 4.4 图形化界面 DSS 鉴定结果

(4) 结果文件见图 4.4（具体信息含义见 4.3 节）。

4.3 结果文件解读

示例 1：目标分类单元鉴定到 DSS

以 Codonopsis_angshen.txt 文件为例，如图 4.5 所示：

```
hzy@hmp1 ~/Demo$ csvtk pretty -t -T -H results/Codonopsis_angshen.txt | head
```

group	assembly	seq	position	GC
Codonopsis_angshen	chuandangshen2scaffold_13	ATTCGAGGAAAAATATAAAATTCGCGGATTTCAATTCAG	3818-3857	30.0
Codonopsis_angshen	chuandangshen2scaffold_7	CGAACAAATGTAATGGAAACCGCTTTTAAAAACCTAAGGA	10160-10199	35.0
Codonopsis_angshen	chuandangshen2scaffold_11	TAAGTCCAATATATATATATATAATATATATATATATA	15709-15748	7.5
Codonopsis_angshen	chuandangshen2scaffold_5	GAATAATTTCCGAGGTTGGTCTGAATTTGTGAAAGATGGA	7089-7128	37.5
Codonopsis_angshen	chuandangshen2scaffold_1	GGAATCTCCGTGACTCTGCCCTGATAAGGGTCGCTACTTT	4826-4865	52.5
Codonopsis_angshen	chuandangshen2scaffold_11	ATTATCCAAGATACTTAAACGAAATACTGGACTATAGTTT	22846-22885	27.5
Codonopsis_angshen	chuandangshen2scaffold_11	GATACTATAAAATTATTCTCTTTCCTTGGTTCTAGCTCAT	25187-25226	30.0
Codonopsis_angshen	chuandangshen2scaffold_13	AAAATATAAATTCGCGGATTTCAATTCAGGGCCACTTA	3827-3866	32.5
Codonopsis_angshen	chuandangshen2scaffold_13	CTTTCACCTTCATCGGTGAATTCATCCTCCTGCAAATCCAA	8965-9004	42.5

图 4.5 目标分类单元鉴定到 DSS

第一列：group：用户指定的目标分类单元名称。

第二列：assembly：DSS 分子标记所在的 DNA 序列名。该列和第四列一起用于指示 DSS 分子标记在 DNA 序列上的位置。

第三列：seq：DSS 分子标记序列。

第四列：position：DSS 分子标记在第二列(assembly)所记录的 DNA 序列上的位置。

第五列：GC：DSS 分子标记中 GC 两种碱基所占的百分比。

示例文件中三个参属物种的均鉴定到 DSS。

示例 2：目标分类单元未鉴定到 DSS 时

若目标分类单元没有鉴定到 DSS，则结果文件如下图所示：

```
hzy@hmp1 ~/software$ head ~/software/IdenDSS/example/Monarda_didyma.txt
group    assembly      seq    position      GC
Monarda_didyma MN642631.1
hzy@hmp1 ~/software$
```

图 4.6 目标分类单元未鉴定到 DSS

该文件第三、四、五列均为空，说明没有在该分类单元中鉴定到可以用于区别目标分类单元和背景分类单元的 DSS 分子标记。

5 插件

现有插件模块包括了 4 个进一步对 DSS 鉴定结果进行处理的功能插件，它们均可以独立运行，具体如下：

- 设计引物

根据 DSS 鉴定结果，在其两端设计引物，便于后续开展实验。

- 识别限制性核酸内切酶酶切位点

用于识别 DSS 序列上是否存在限制性核酸内切酶酶切位点，便于筛选可以转化为 RFLP 标记的 DSS 标记。

- 合并 DSS

以长度为 40 bp 的 DSS 为例，假设一条序列上 1~40 位碱基、2~41 位碱基、3~42 位碱基均为 DSS，则我们可以将这三个 DSS 合并，称 1~42 位碱基为一个连续 DSS(combined DSS)。该功能对 DSS 结果中的连续 DSS 进行识别，并整理输出。

- 统计多个 DSS 鉴定结果文件

如果同时对多个目标分类单元开展了鉴定，用户可以使用该功能将各个目标分类单元的结果统计到一个文件中。

5.1 输入文件准备

插件模块的输入文件为一个包含 DSS 鉴定结果路径的文本文件，每行一个文件。示例如下（命名为 plugin_input.txt 用于后续示例）：

```
/home/hzy/Demo/results/Codonopsis_angshen.txt
/home/hzy/Demo/results/Codonopsis_pilosula.txt
/home/hzy/Demo/results/Codonopsis_sumodesta.txt
~
~
~
~
~
```

图 5.1 插件模块(plugin)输入文件

5.2 命令行界面操作

```
hzy@hmp1 ~/Demo$ IdenDSS plugin -h
usage: IdenDSS plugin [-h] -i INPUT -d DATABASE [-c] -o OUTPUT [-t TMP] [--primer] [--rflp] [--combine] [--statistic] [--bin BIN_DIR]

optional arguments:
  -h, --help            show this help message and exit
  -i INPUT, --input INPUT <File path> A meta file. One DSS result file path per line (combined result is OK)
  -d DATABASE, --database DATABASE <File path> Database fasta (The database used to identify DSS)
  -c, --circular         the sequences are circular or not
  -o OUTPUT, --output OUTPUT <Directory path> result directory
  -t TMP, --tmp TMP      <Dir path> Temporary directory path <Default system temporary directory>
  --primer              Design Primer (Need primer3_core in your PATH)
  --rflp                Identify restriction enzyme sits on DSS for putative RFLP method
  --combine             Generate the corresponding combined DSS file
  --statistic           Count the DSS number
  --bin BIN_DIR         <Directory path> Primer3 exec directory <If your Primer3 software not in PATH>
```

图 5.2 插件模块(plugin)提示信息

必填选项说明:

- `-i/--input` 输入文件路径(格式见 5.1 节)
- `-d/--database` 3.1 中建立的数据库路径.
- `-o/--output` 输出结果存储文件夹(注意输出路径为文件夹而不是文件).

插件选项说明:

- `--primer` 启动引物设计. 由于 **Primer3** 软件限制, 目前该功能仅限 **Linux** 系统下使用.
- `--rflp` 启动限制性核酸内切酶酶切位点识别.
- `--combine` 启动连续 DSS 识别.
- `--statistic` 启动 DSS 鉴定结果统计.

其他选项说明：

- `-c/--circular` 若使用的序列是环状 DNA（如植物叶绿体或动物线粒体 DNA 序列），用户需要启用该选项.默认不启动.
- `-t/--tmp` 用户可以通过该选项指定 IdenDSS 运行过程中的临时文件存放路径.默认为系统临时文件夹.
- `--bin` 若 Primer3 软件不在环境变量中，使用该选项指定 Primer3 软件的目录.

以示例文件鉴定得到的 DSS 结果作为示范进行操作：

5.2.1 引物设计

```
hzy@hmp1 ~/Demo$ mkdir primer (1)
hzy@hmp1 ~/Demo$ IdenDSS plugin \
> --primer \
> -i plugin_input.txt \
> -d database/example_db.fasta \
> -o primer \
> --bin /home/smalltools/primer3-2.5.0/src
Start design Codonopsis_angshen primers
Design Codonopsis_angshen primers. Done
Start design Codonopsis_pilosula primers
Design Codonopsis_pilosula primers. Done (3)
Start design Codonopsis_sumodesta primers
Design Codonopsis_sumodesta primers. Done
hzy@hmp1 ~/Demo$ ls primer
Codonopsis_angshen_primer.txt Codonopsis_pilosula_primer.txt Codonopsis_sumodesta_primer.txt (4)
hzy@hmp1 ~/Demo$
```

图 5.3 使用插件模块设计引物

- (1) 创建存放设计引物结果的文件夹.
- (2) 设计引物命令.
- (3) 软件运行过程中的提示信息.
- (4) 结果文件,具体解释见下:

5.2.2 识别限制性核酸内切酶酶切位点

```
hzy@hmp1 ~/Demo$ mkdir RFLP (1)
hzy@hmp1 ~/Demo$ IdenDSS plugin \
> --rflp \
> -i plugin_input.txt \
> -d database/example_db.fasta \
> -o RFLP (2)
Begin to search Codonopsis_angshen RFLP sites
Codonopsis_angshen RFLP sites searches done
Begin to search Codonopsis_pilosula RFLP sites
Codonopsis_pilosula RFLP sites searches done (3)
Begin to search Codonopsis_sumodesta RFLP sites
Codonopsis_sumodesta RFLP sites searches done
All RFLP sites searches done
hzy@hmp1 ~/Demo$ ls RFLP
Codonopsis_angshen_rflp.txt Codonopsis_pilosula_rflp.txt Codonopsis_sumodesta_rflp.txt (4)
```

图 5.5 使用插件模块识别限制性核酸内切酶酶切位点

- (1) 创建存放识别限制性核酸内切酶酶切位点结果的文件夹.
- (2) 识别限制性核酸内切酶酶切位点命令.
- (3) 软件运行过程中的提示信息.
- (4) 结果文件,具体解释见下:

```
hzy@hmp1 ~/Demo$ csvtk pretty -T -t -H RFLP/Codonopsis_angshen_rflp.txt
group      assembly      seq      position      GC      enzyme
Codonopsis_angshen chuandangshen2scaffold_11 TAAATATCTAGAAAGTTATCCTGGGAATCGAGTGATTCCG 18868-18907 37.5 XbaI
Codonopsis_angshen chuandangshen2scaffold_11 TAGATCTCTTTCCITTTTTTTTTTTTACAATCTTTACTA 25527-25566 20.0 BglII
Codonopsis_angshen chuandangshen2scaffold_11 AAGATCGTTTAGATCTCTTTCCITTTTTTTTTTTTACAAT 25518-25557 22.5 BglII
```

图 5.6 限制性核酸内切酶酶切位点识别结果

前五列同 DSS 鉴定结果（见 4.3 节）

第六列：enzyme：能够切割该 DSS 标记的限制性核酸内切酶。

5.2.3 合并 DSS

```
hzy@hmp1 ~/Demo$ mkdir combine (1)
hzy@hmp1 ~/Demo$ IdenDSS plugin \
> --combine \
> -i plugin_input.txt \
> -d database/example_db.fasta \
> -o combine (2)
Begin to combine Codonopsis_angshen DSSs
Combining Codonopsis_angshen DSSs done
Begin to combine Codonopsis_pilosula DSSs
Combining Codonopsis_pilosula DSSs done (3)
Begin to combine Codonopsis_sumodesta DSSs
Combining Codonopsis_sumodesta DSSs done
Combining all groups DSSs done
hzy@hmp1 ~/Demo$ ls combine/
Codonopsis_angshen_combined.txt Codonopsis_pilosula_combined.txt Codonopsis_sumodesta_combined.txt (4)
hzy@hmp1 ~/Demo$
```

图 5.7 使用插件模块合并 DSS

- (1) 创建存放合并 DSS 结果的文件夹.
- (2) 合并 DSS 结果命令.

(3) 软件运行过程中的提示信息.

(4) 结果文件。具体内容见图 5.8, 格式与 DSS 结果文件完全一致(见 4.3 节)。

group→	assembly→	seq→	position→	GC↓
Codonopsis_pilosula→	G140224LY0334→	GCATTATCGCTTAGAAATGCTTTTCTAGCATTTGATTGCGTACCCCTTGAAGTCTTTGACGCACACTTGAAAAATAAC→	2393-2471→	37.9746835443038↓
Codonopsis_pilosula→	G140224LY0334→	CAGACGAAAAACAAATCAATTTTACGATAAAAGAAAAACAAAAATCAATGCCAAATGGGCCGATCATCCCTTATCT→	4362-4440→	31.645569620253166↓
Codonopsis_pilosula→	G140224LY0334→	CTCGGCCACGCCCATTTATTTTCACTCTTGATTAAATATATTATACAAGTGGCAGGCGCTTTGTCAACTGCTCCCA→	7720-7798→	44.30379746835443↓
Codonopsis_pilosula→	G140224LY0334→	AATAATTTTCGAGAAAGTAAAGACTTAAAGATCTAATTTACAAATAAA→	9261-9309→	18.367346938775512↓
Codonopsis_pilosula→	G140224LY0334→	ACTTAGTAGCTAAGTTCTCTACCTTATTTCTATAAAAGTACGGTTTCATATTAT→	9525-9578→	27.77777777777778↓
Codonopsis_pilosula→	G140224LY0334→	GAACCCTTACACAAATGCCCTTTTGTGTGATTGAAAGGTTTGGTGAGCCGTATCTACCAAGACTCTCCAGCCAAA→	9629-9707→	44.30379746835443↓
Codonopsis_pilosula→	G140224LY0334→	ATAACACATTCGAGGAAAAATATAAATTCGCGATTTTCGATTT→	10002-10045→	29.545454545454547↓
Codonopsis_pilosula→	G140224LY0334→	AATTTGAGCTGCAATCCGACCCGAGAAACCGAGAGACCAACATTAATAGCAGGTCGGATTCCAGCATTGAATAGATCG→	11372-11450→	46.835443037974684↓
Codonopsis_pilosula→	G140224LY0334→	CCCTTCTTGATCCGCAAAACCTCACTCATTAACAACCTCAACATTATGCGATTTCAAATTCAGAGCAATG→	12233-12305→	41.0958904109589↓

图 5.8 合并后的 DSS

5.2.4 统计 DSS 鉴定结果

```
hzy@hmp1 ~/Demo$ mkdir summary (1)
hzy@hmp1 ~/Demo$ IdenDSS plugin \
> --statistic \ (2)
> -i plugin_input.txt \
> -d database/example_db.fasta \
> -o summary
Begin to summary all groups results (3)
Summary all groups results done
hzy@hmp1 ~/Demo$ ls summary/summary.tsv (4)
summary/summary.tsv
hzy@hmp1 ~/Demo$
```

图 5.9 使用插件模块统计 DSS 鉴定结果

(1) 创建存放统计 DSS 鉴定结果的文件夹.

(2) 统计 DSS 鉴定结果命令.

(3) 软件运行过程中的提示信息.

(4) 结果文件为输出文件夹路径中的 **summary.tsv** 文件。内容见图 5.10。

注: 虽然该功能输出的结果为单个文件, 但是为了与其他功能输入格式统一, 其

-o/--output 选项仍为文件夹路径, 而非文件路径。结果文件为输出文件夹路径中的 **summary.tsv** 文件。

```
hzy@hmp1 ~/Demo$ csvtk pretty -t -T -H summary/summary.tsv
group          dss_num
Codonopsis_angshen    3789
Codonopsis_pilosula   3876
Codonopsis_sumodesta  46989
```

图 5.10 DSS 统计结果

第一列：group：目标分类单元名称。

第二列：dss_num：目标分类单元 DSS 数量。

5.2.5 同时运行多个插件

如前所述，各插件可以互相独立运行，因此可以同时运行。例如使用如下代码同时合并 DSS、识别限制性核酸内切酶酶切位点、统计 DSS 鉴定结果。

```
hzy@hmp1 ~/Demo$ mkdir all_plugins (1)
hzy@hmp1 ~/Demo$ IdenDSS plugin \
> --statistic --combine --rflp \
> -i plugin_input.txt \
> -d database/example_db.fasta \
> -o all_plugins (2)
Begin to search Codonopsis_angshen RFLP sites
Codonopsis_angshen RFLP sites searches done
Begin to search Codonopsis_pilosula RFLP sites
Codonopsis_pilosula RFLP sites searches done
Begin to search Codonopsis_sumodesta RFLP sites
Codonopsis_sumodesta RFLP sites searches done
All RFLP sites searches done
Begin to combine Codonopsis_angshen DSSs
Combining Codonopsis_angshen DSSs done
Begin to combine Codonopsis_pilosula DSSs
Combining Codonopsis_pilosula DSSs done
Begin to combine Codonopsis_sumodesta DSSs
Combining Codonopsis_sumodesta DSSs done
Combining all groups DSSs done
Begin to summary all groups results
Summary all groups results done (3)
hzy@hmp1 ~/Demo$ ls all_plugins/
Codonopsis_angshen_combined.txt Codonopsis_pilosula_combined.txt Codonopsis_sumodesta_combined.txt summary.tsv (4)
Codonopsis_angshen_rflp.txt Codonopsis_pilosula_rflp.txt Codonopsis_sumodesta_rflp.txt
```

图 5.11 同时运行多个插件

(1) 创建存放全部插件结果的文件夹。

(2) 同时运行多个插件命令。

(3) 软件运行过程中的提示信息。

(4) 结果文件。其中以_combined.txt 结尾的为合并 DSS 的结果；以_rflp.txt 结尾的为识别限制性核酸内切酶酶切位点的结果；summary.tsv 为统计 DSS 鉴定结果的文件。同时运行与单独运行各插件结果一致。

5.3 图形化界面操作

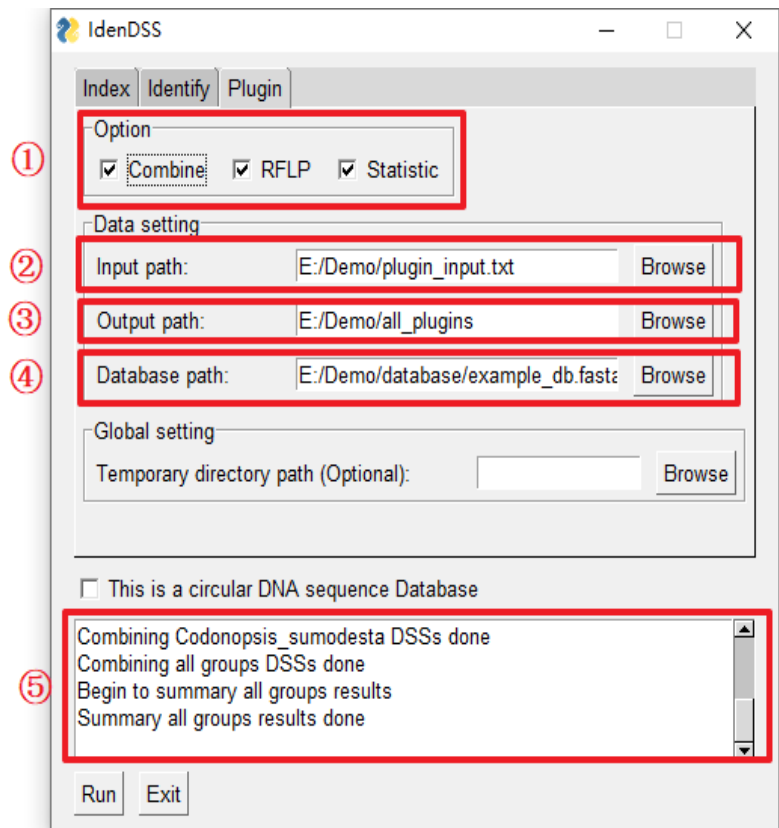


图 5.12 插件模块图形化界面

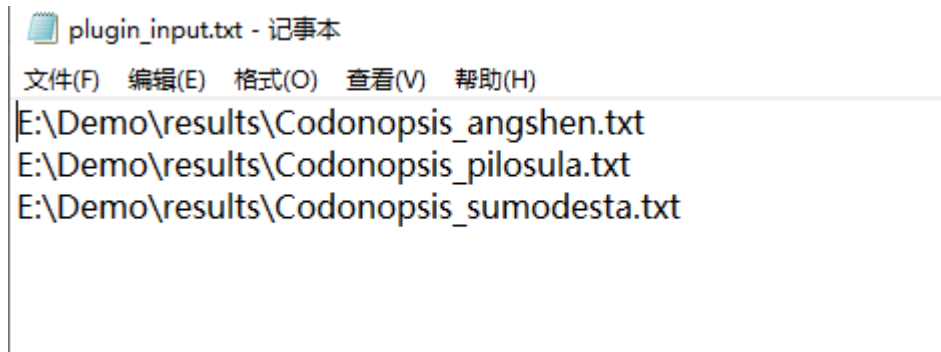


图 5.13 plugin_input.txt 内容

(1) 选取 Plugin 标签

(2) 填选相关信息，操作同 3.2，各选项说明如下：

① 对应 `--combine`，`--rflp`，`--statistic`，用户根据自身需求勾选需要执行的功能。

-
- ② 对应-i/--input，输入文件路径(输入文件内容见图 5.13).
- ③ 对应-o/--output，输出结果存储文件夹.
- ④ 对应-d/--database，3.2 中建立的数据库路径.
- (3) 点击 Run 按钮运行。软件运行过程中的提示信息见图 5.13 中的⑤。
- (4) 结果文件(图 5.14),详细格式解读见 5.2 节下各项。








名称	修改日期	类型	大小
 Codonopsis_angshen_combined.txt	2022/10/8 15:49	Text Document	61 KB
 Codonopsis_angshen_rflp.txt	2022/10/8 15:48	Text Document	12 KB
 Codonopsis_pilosula_combined.txt	2022/10/8 15:49	Text Document	11 KB
 Codonopsis_pilosula_rflp.txt	2022/10/8 15:48	Text Document	57 KB
 Codonopsis_sumodesta_combined.txt	2022/10/8 15:49	Text Document	134 KB
 Codonopsis_sumodesta_rflp.txt	2022/10/8 15:49	Text Document	405 KB
 summary.tsv	2022/10/8 15:49	TSV 文件	1 KB

图 5.14 Windows 系统下图形化界面运行后结果

6 验证

IdenDSS 软件支持使用高通量测序数据对鉴定得到的 DSS 序列可信度进行进一步验证。原理简述如下：如果一条 DSS 出现在目标分类单元的高通量测序数据中，且不出现在背景分类单元的高通量测序单元中，则认为该 DSS 序列可信，可进一步进行实验验证。

受 KMC 软件和硬件要求限制，目前该功能仅限 Linux 系统。需要注意的是，该功能对硬件有一定要求，具体如下：使用测序量为 1 Gbp 的文库(FASTQ 格式)进行验证约需要 5 GB 内存、2 GB 的硬盘空间和 1 分钟的运行时间；使用测序量 5 Gbp 的文库(FASTQ 格式)进行验证约需要 11 GB 内存、4GB 的硬盘空间和 4 分钟的运行时间。

6.1 输入文件

该功能需要如下输入文件：

- (1) DSS 鉴定结果文件(由 identity 模块得到)
- (2) 背景分类单元高通量测序文件(fastq 格式)
- (3) 目标分类单元高通量测序文件(fastq 格式)

6.2 命令行界面操作

```
hzy@hmp1 ~/Demo$ IdenDSS validate -h
usage: IdenDSS validate [-h] -i INPUT --sp SP --bg BG [--min MIN] -o OUTPUT [-t TMP] [--bin BIN]

optional arguments:
  -h, --help            show this help message and exit
  -i INPUT, --input INPUT
                        <File path>
  --sp SP                <File path> The intraspecies HTS FASTQ file (could be a list, separate by comma)
  --bg BG                <File path> The background species HTS FASTQ file (could be a list, separate by comma)
  --min MIN              <Int> The minium k-mer occurance
  -o OUTPUT, --output OUTPUT
                        <File path> validated DSS result path
  -t TMP, --tmp TMP      <Directory path> Temporary directory path <Default system temporary directory>
  --bin BIN              <Directory path> KMC3 exec directory, containing kmc and kmc dump <If your KMC3 software not in PATH>
```

图 6.1 验证模块(validate)提示信息

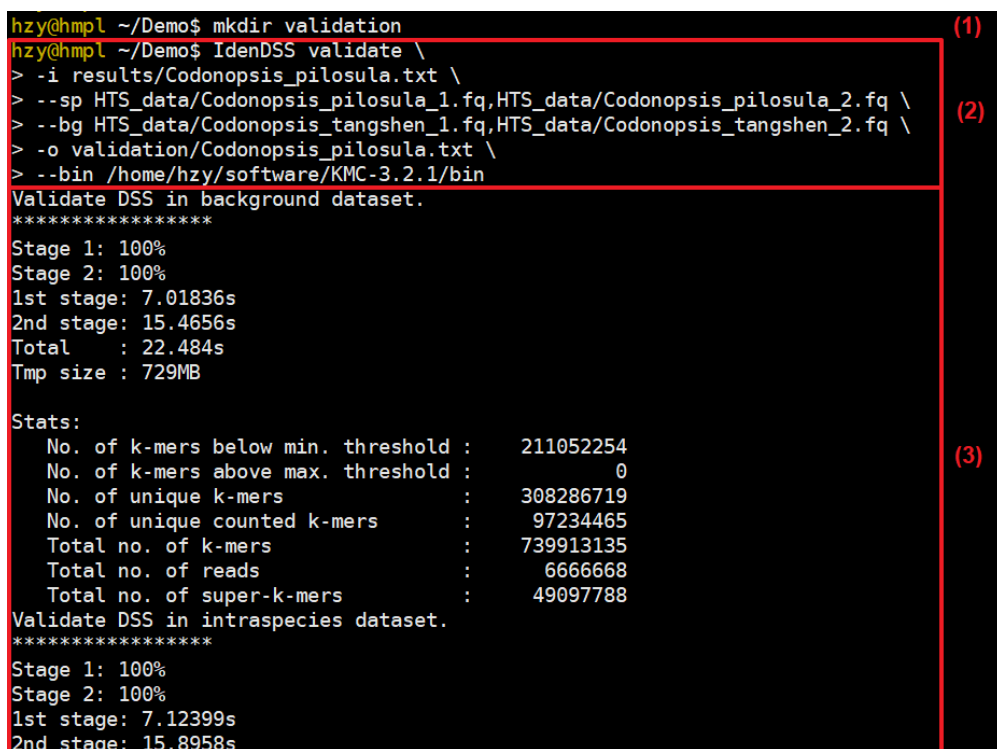
必填选项说明：

- `-i/--input` 需要验证的 DSS 鉴定结果文件路径.
- `--sp` 目标分类单元高通量测序文件(fastq 格式).
- `--bg` 背景分类单元高通量测序文件(fastq 格式).
- `-o/--output` 经验证的 DSS 鉴定结果文件路径.

其他选项说明：

- `--min` 最小有效出现次数，默认值为 2.如果 DSS 序列在高通量数据集中出现的次数小于该值，则认为其是高通量测序造成的假阳性结果.
- `-t/--tmp` 指定临时文件夹路径，默认为系统临时文件夹路径.
- `--bin` 若 KMC 软件不在环境变量中，使用该选项指定 KMC 软件的目录.

以 4.1 鉴定得到的 *Codonopsis pilosula* 目标类群的 DSS 进行示范操作：



```
hzy@hmp1 ~/Demo$ mkdir validation
hzy@hmp1 ~/Demo$ IdenDSS validate \
> -i results/Codonopsis_pilosula.txt \
> --sp HTS_data/Codonopsis_pilosula_1.fq,HTS_data/Codonopsis_pilosula_2.fq \
> --bg HTS_data/Codonopsis_tangshen_1.fq,HTS_data/Codonopsis_tangshen_2.fq \
> -o validation/Codonopsis_pilosula.txt \
> --bin /home/hzy/software/KMC-3.2.1/bin
Validate DSS in background dataset.
*****
Stage 1: 100%
Stage 2: 100%
1st stage: 7.01836s
2nd stage: 15.4656s
Total : 22.484s
Tmp size : 729MB

Stats:
  No. of k-mers below min. threshold : 211052254
  No. of k-mers above max. threshold : 0
  No. of unique k-mers : 308286719
  No. of unique counted k-mers : 97234465
  Total no. of k-mers : 739913135
  Total no. of reads : 6666668
  Total no. of super-k-mers : 49097788
Validate DSS in intraspecies dataset.
*****
Stage 1: 100%
Stage 2: 100%
1st stage: 7.12399s
2nd stage: 15.8958s
```

图 6.2 使用验证模块基于高通量测序数据验证 DSS 标记

(1) 创建文件夹用于存放验证结果.

(2) 验证 DSS 命令.其中 Codonopsis_pilosula_1.fq 和 Codonopsis_pilosula_2.fq 分别为党参样品经双端测序得到的正向测序结果和反向测序结果,作为目标分类单元高通量测序文件输入,以逗号分隔; Codonopsis_tangshen_1.fq 和 Codonopsis_tangshen_2.fq 分别为川党参样品经双端测序得到的正向测序结果和反向测序结果,作为背景分类单元高通量测序文件输入,以逗号分隔.

(3) 软件运行过程中的提示信息.

```
hzy@himpl ~/Demo$ ls validation/
Codonopsis_pilosula.txt
hzy@himpl ~/Demo$ head validation/Codonopsis_pilosula.txt
group assembly seq position GC
Codonopsis_pilosula G140224LY0334 AGGTTATGATTGATTTGTCATGGTTCTTTGGATCCAGATA 95914-95953 35.0
Codonopsis_pilosula G140224LY0334 AACTTTCTCTACCTTATTCTATAAAAGTACCGTTTCATAT 9536-9575 27.5
Codonopsis_pilosula G140224LY0334 AATGCCCTTTTGTGTGATTTGAACGGTTTGGTGAGCC 9641-9680 42.5
Codonopsis_pilosula G140224LY0334 ATGGGGCTGATTTTCTCCTAGCCCTAAAAACCAACGAG 95994-96033 45.0
Codonopsis_pilosula G140224LY0334 CGAGAAAGTAAAGACTTAAAGATCTAATTTACAAATAAA 9270-9309 22.5
Codonopsis_pilosula G140224LY0334 AGCTAACTTTCTCTACCTTATTCTATAAAAGTACCGTTTC 9532-9571 32.5
Codonopsis_pilosula G140224LY0334 AATCTAACCTATATGAATCAAAATTTGGATTTTATAGAC 146319-146358 22.5
Codonopsis_pilosula G140224LY0334 ACCAACTCATCGCTTTCATTATCTGGATCCAAGAACCA 146595-146634 42.5
Codonopsis_pilosula G140224LY0334 AATTTTACGATAAAAGAAAAACAAAAATCAATGCCAAA 4379-4418 20.0
```

图 6.3 经验证可信度高的 DSS

(4) 结果文件(图 6.3)