# EasyMiner 使用说明

2023/09/09

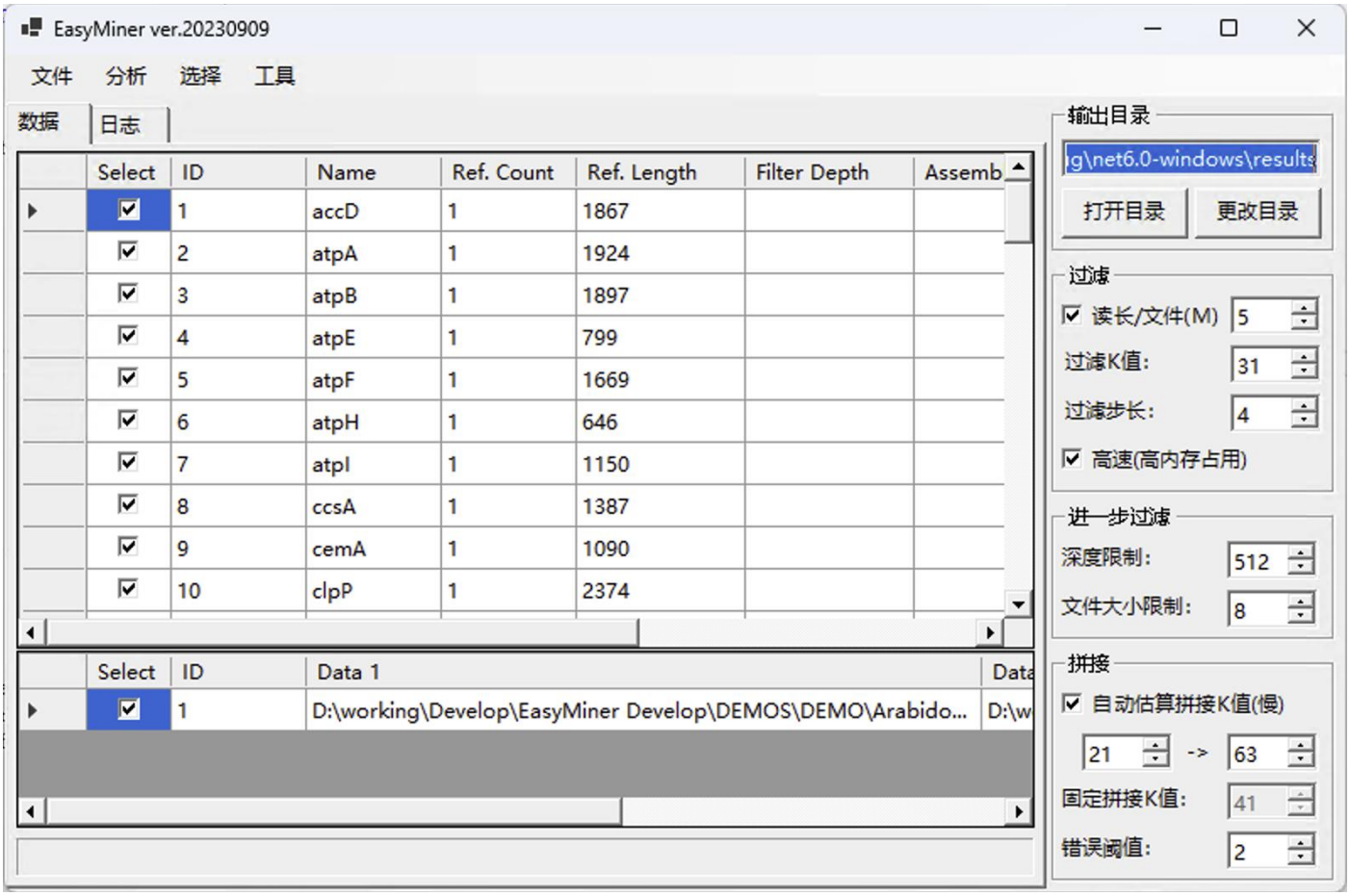
余岩 [1,*], 于欣艺 [1]

[1] *四川大学生命科学学院， 生物资源与生态环境教育重点实验室*

*yyu@scu.edu.cn*

# 目录

# 1. 介绍




EasyMiner 是基于我们前期开发的 Easy353 和 GeneMiner，主要为 Windows 用户设计的增强版本。EasyMiner 具有用户友好的 Windows 图形界面，可在个人电脑上高效准确地进行分子标记提取，无需依赖服务器。EasyMiner 通过特别设计的拼接算法，能基于近源物种的参考基因从二代测序中快速准确的提取分子标记，并同时兼具细胞器基因组组装、gnebank 文件中基因序列分解、二代测序文件导出等功能。

# 2. 安装和需求

EasyMiner 是基于.net 平台开发的，仅提供 x64 版本，需要在计算机上安装有.NET 6.0 Desktop Runtime x64。如果不满足需求，软件会在第一次运行时提醒您下载。您也可以从此处获取.NET 6.0 Desktop Runtime x64 的安装包:

https://dotnet.microsoft.com/zh-cn/download/dotnet/thank-you/runtime-desktop-6.0.21-windows-x64-installer

EasyMiner 的最新版本的编译文件和所有源代码均保存在 github 上，您可以从此处获取最新的安装包:

https://github.com/sculab/EasyMiner/releases/latest

如果您需要在 macOS 或 Linux 上使用命令行版本的基因挖掘工具，请访问:

Easy353: https://github.com/plant720/Easy353

GeneMiner: https://github.com/sculab/GeneMiner

你也可以使用 github 上 scripts 文件夹中的 python 脚本，这些脚本提供了 EasyMiner 的所有核心功能，并可以在 macOS 或 Linux 上部署。

# 3. 使用方法

## 3.1 简单实例

该实例演示了从利用来自琴叶拟南芥(*Arabidopsis lyrata*)的基因序列，从拟南芥拟南芥(*Arabidopsis thaliana*)的二代测序文件中获取对应的基因。

**数据准备:**

以下所有示例文件均保存于 github 的 DEMO 目录中:

https://github.com/sculab/EasyMiner/tree/master/DEMO

您也可以自行准备:

**（1）测序数据:** 二代测序的数据文件，文件格式为.gz 或.fq。EasyMiner 主要针对短读长的测序文件（reads 长度为 100、150、300 等）。一般而言，浅层基因组、转录组、全基因组的双端或者单端测序文件都可以使用。

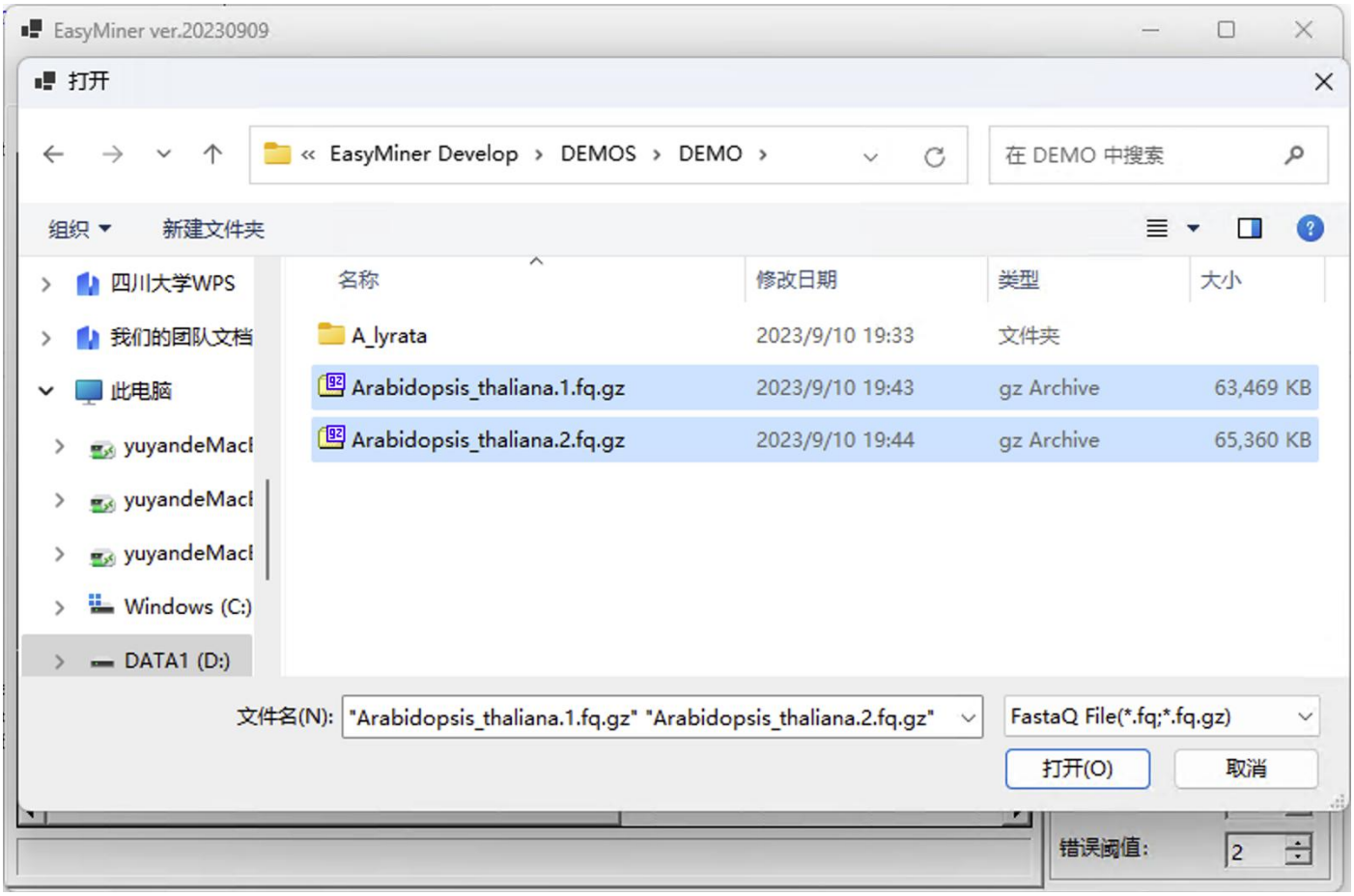
**（2）参考序列:** 近源物种的参考基因序列文件。可以使用每个文件可以 fasta 或 genbank 格式。对于 fasta 格式，文件名通常为基因名，每个文件中可以包含多个不同物种的同一个基因。对于 genbank，同一个 gb 文件中可以包含多个物种的多个基因，EasyMiner 会自动按基因名进行分解和组合。

**载入数据:**

点击 [文件>载入测序文件] 选择测序数据文件。

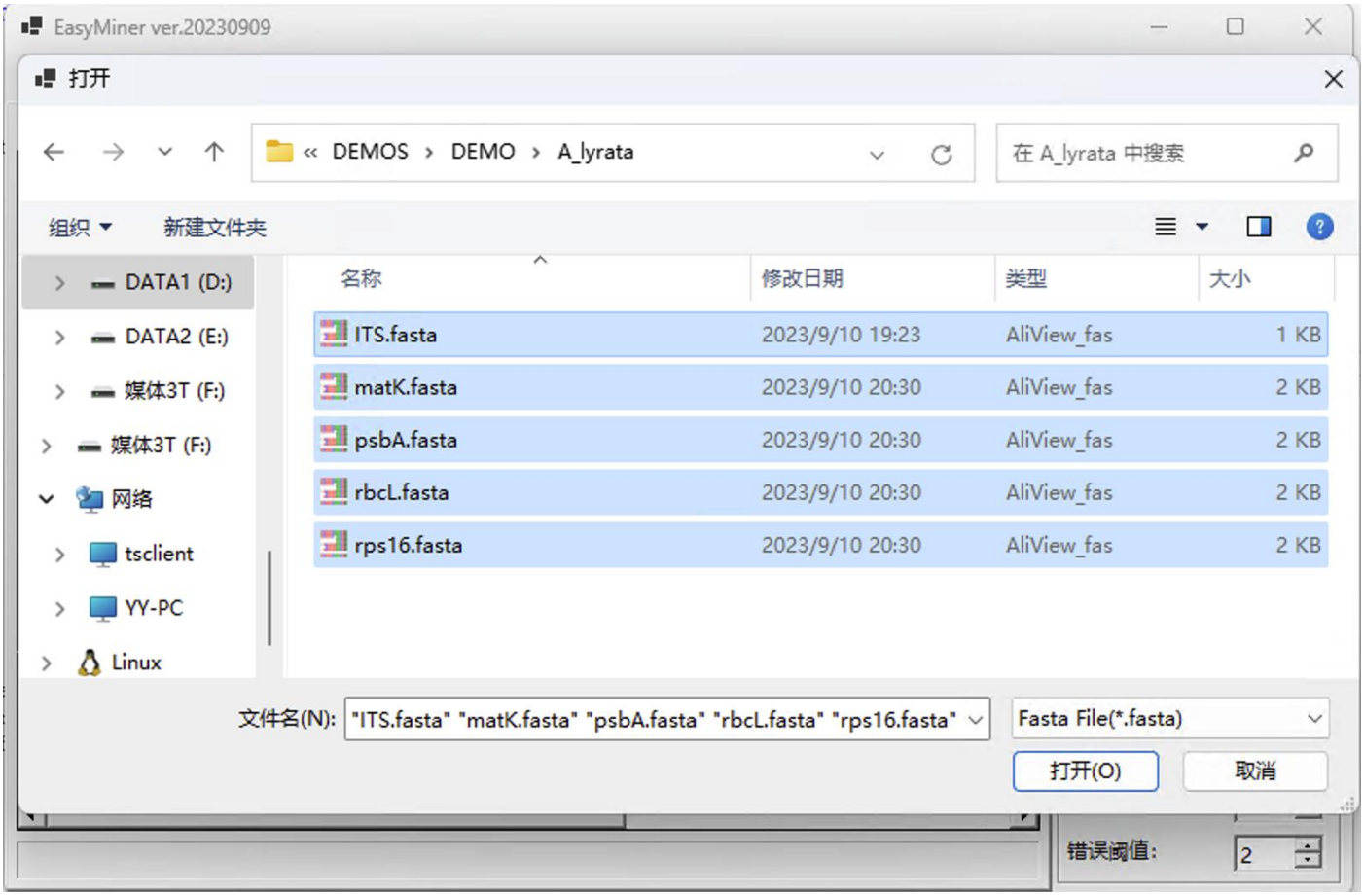
示例：打开 Arabidopsis_thaliana.1.fq.gz 和 Arabidopsis_thaliana.1.fq.gz 两个文件。这两个数据文件是来自拟南芥(*Arabidopsis thaliana*)的双端二代测序文件，测序 fang150，每个文件中保存了 1M ($2^{20}$)条 reads。

**注意:** 对于配对(paired)的序列文件，需要同时选中两个（偶数个）数据文件一起载入，如只选取一个，则会作为单端测序数据载入。




点击 [文件>载入参考序列] 选择 fasta 格式的参考序列文件，可以一次选择多个参考序列文件。

示例：载入 DEMO/A_lyrata/ 下的所有 fasta 文件（ITS、martK、psbA、rbcL、rps16），包括 1 个核基因和 4 个叶绿体基因的参考序列，所有这些序列来都自拟南芥同属的近缘种琴叶拟南芥(*A. lyrata*)。




导入文件后会显示参考序列的 ID、基因名、序列数量、序列平均长度等信息。

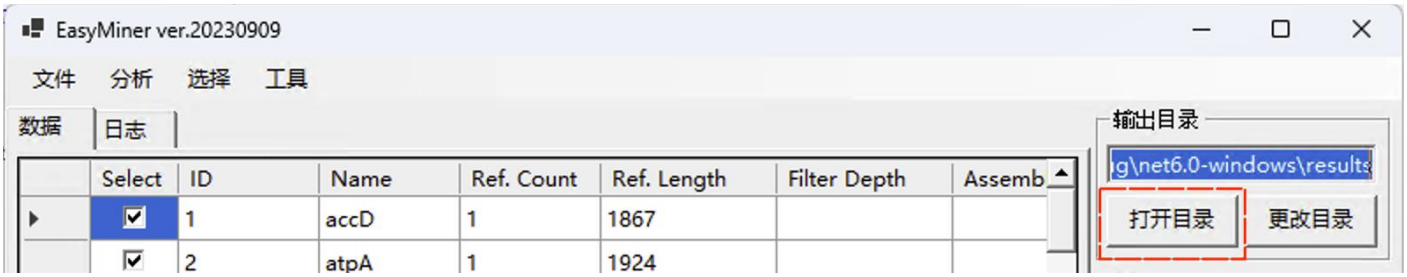
**运行程序**

点击[分析>自动] 使用默认参数运行程序，等待程序运行结束。

注意：切勿手动关闭成都弹出的命令行窗口，请耐心等待窗口自动关闭。

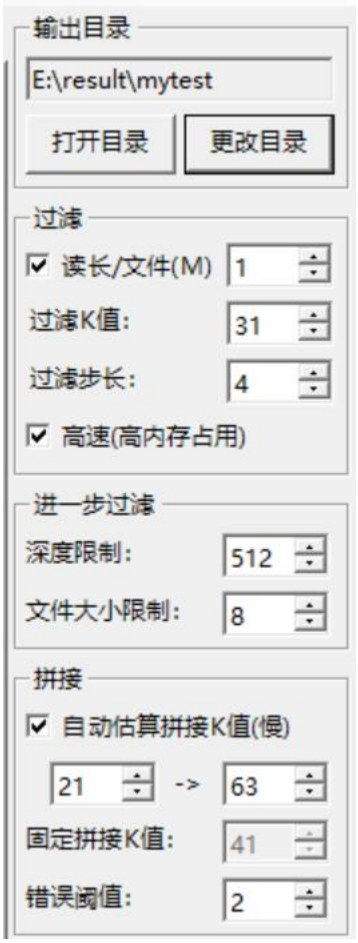
**查看结果**

点击“打开目录”按钮，查看结果文件。拼接后的文件以 fasta 格式保存于 results 目录中。

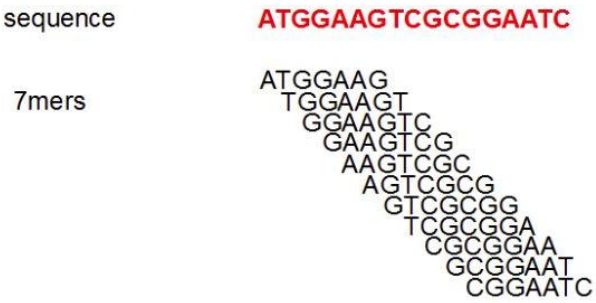



更多实例请参见 github 主页。

## 3.2 主面板参数设置

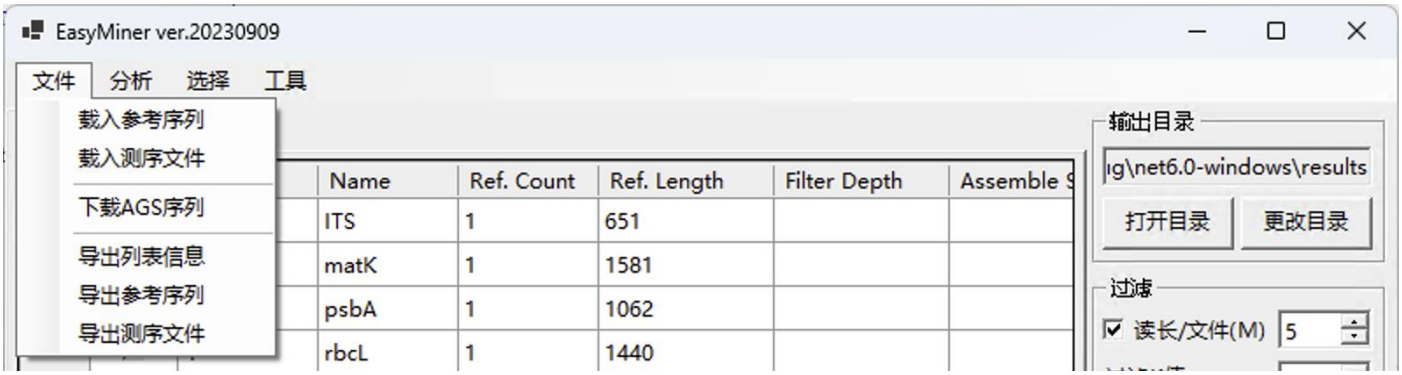



**输出目录:** 保存结果的文件夹, 默认为 EasyMiner 应用程序所在目录的 results 文件夹。
**打开目录:** 在 Windows 资源管理器中打开输出目录。
**更改目录:** 选择保存结果的文件夹。注意: 保存结果的文件夹在运行过程中可能被反复清空，请务必不要选取保存有资料的文件夹。建议每次分析都新建文件夹进行输出。如果为了延续之前的分析而选取同样输出文件夹，请在后续的分析中选择不清空文件夹。
**读长/文件(M)**: 设置在过滤过程中，每个测序文件所使用的读长的数量（或待导出的读长数量），以 M($2^{20}$)为单位。
**过滤 K 值**: 在初次过滤过程中分解参考序列和 reads 时所采用的 k-mer 值, 默认为 31。
**过滤步长**: 切取 kmer 时滑动窗口的前进的步数。
例如: 当过滤 K 值为 7，步长为 1 时，对 reads 的切割方式如下所示:

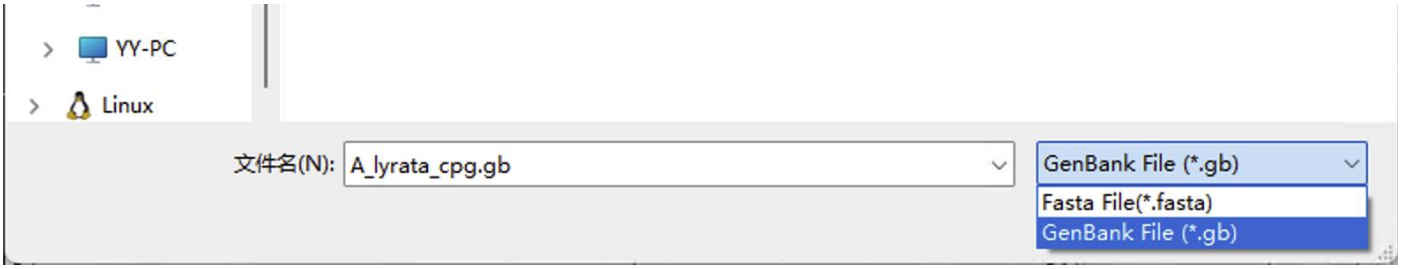



**高速(高内存占用)**: 在生成参考序列的字典时考虑反向互补的序列。选中该选项会占用更高的内存，但可以显著提高过滤速度，推荐有大内存的电脑使用。
**深度限制**: 对于过滤得到的 fq 文件，如果估算过滤深度（Filter Depth 列显示）超过了该值，则在进一步过滤中提高 K 值重新进行过滤。其中，过滤深度(Filter Depth)=reads 测序长度*过滤出的 reads 数量/参考序列的平均长度。
**进一步过滤-文件大小限制**: 对于过滤得到的 fq 文件，如果文件大小超过了该值，则在进一步过滤中提高 K 值重新进行过滤。
例如：在默认参数下，如果过滤结果文件深度超过 512，fq 文件大小大于 8MB，需要进行进一步过滤。
**自动估算拼接值(慢)**: 在拼接时对每个基因动态估算合适的 kmer 值。
**固定拼接 K 值**: 在拼接时对所有基因都使用指定的 kmer 值。
**错误阈值**: 在拼接过程中，不使用出现次数小于该值-mer。

## 3.3 文件菜单




**[文件>载入参考序列]** 选择 fasta 或者 genbank 格式的参考序列文件，可以一次选择多个参考序列文件。
在打开对话框中，可以切换待选取的数据类型:

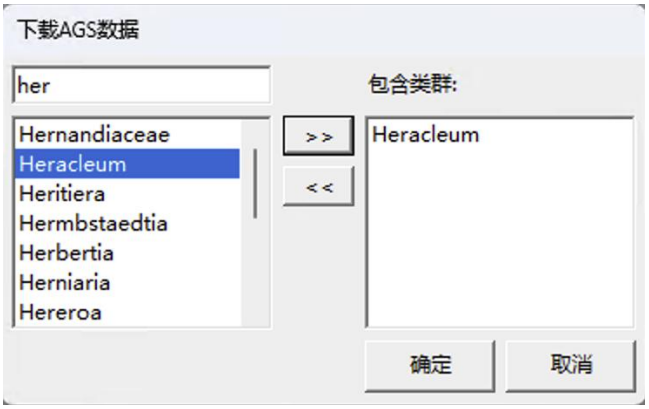



如果选择了 genbank 格式的文件，EasyMiner 会对其中的基因按照基因名自动分解，会弹出如下设置对话框:


基因最小长度：要处理的基因的最小长度
基因最大长度：要处理的基因的最大长度
扩展边界长度：在每个基因两侧延展的长度

**[文件>载入测序文件]** 载入二代测序的数据文件，文件格式为.gz 或.fq。对于配对(paired)的序列文件，需要同时选中两个（偶数个）数据文件一起载入，如只选取一个，则会作为单端测序数据载入。

**[文件>下载 AGS 序列]** 从 Kew Tree of Life Explorer (https://treeoflife.kew.org)获取 Angiosperms353 Gene Set (AGS)作为参考序列。




在输入框中输入属或以上分类阶元的拉丁学名，在下方选中类群，点击>>按钮添加到右侧列表中。

注意：如果找不到您所研究的类群，这意味着该类群在 Kew Tree of Life Explorer 没有数据，请选择更高分类阶元的类群代替。

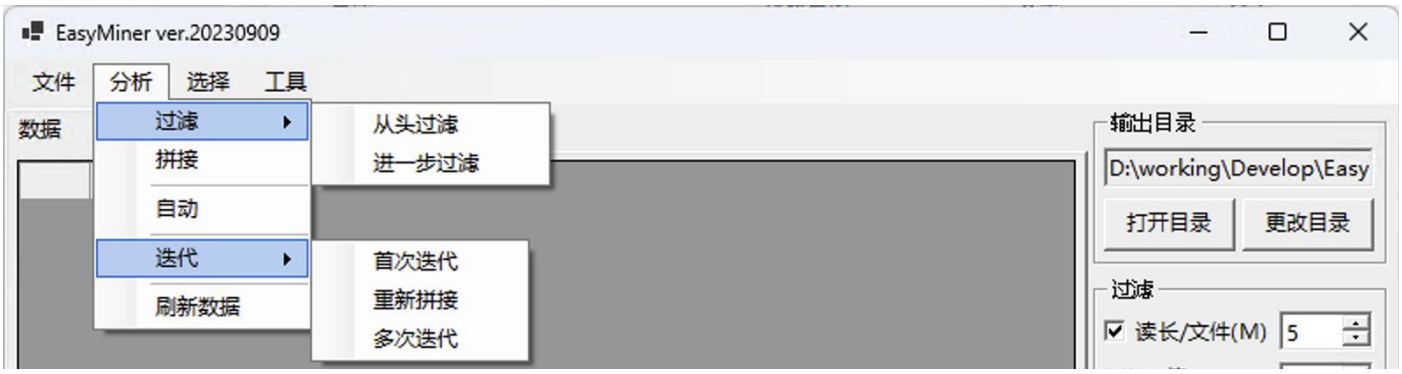
下载完成后会在参考序列列表中显示，可直接作为参考序列分析。建议使用[文件>导出参考序列]导出备用，以免重复下载。

**[文件>导出列表信息]** 指定要保存的文件名，将参考序列信息列表保存为 csv 格式。

**[文件>导出参考序列]** 选择输出文件夹，将选取的参考基因导出为 fasta 格式。

**[文件>导出测序文件]** 选择输出文件夹，在弹出的对话框中设定要跳过的读长的数量，并导出在“**读长/文件(M)**”中设置的读长的数量。对于每一对测序文件，导出的文件以 project*.1.fq 和 project*.2.fq(*为数字)命名。

## 3.4 分析菜单



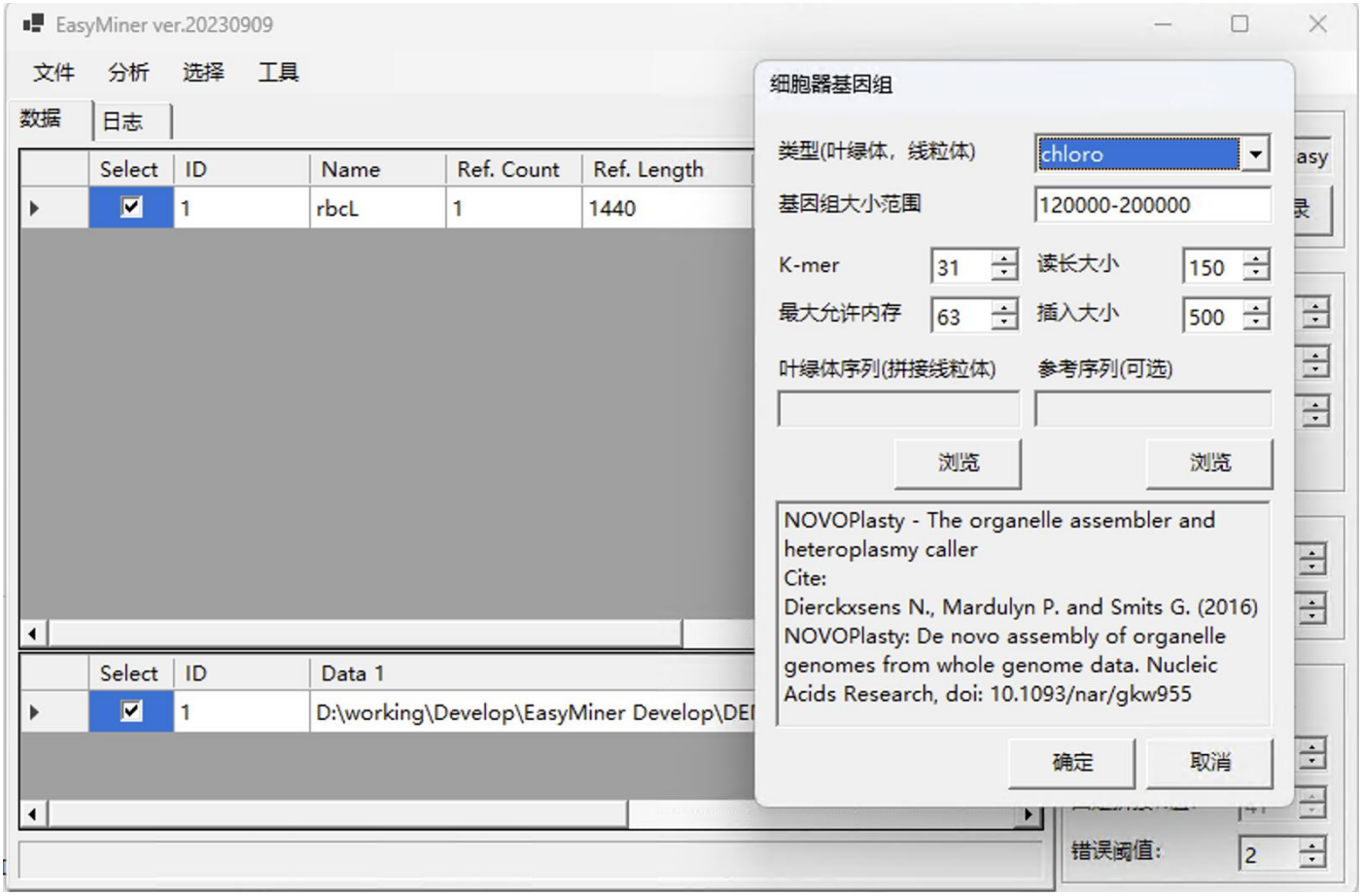

**[分析>过滤>从头过滤]**: 使用参考基因对测序数据进行批量过滤，获得与目标基因关联的 reads。过滤结果的 fq 文件保存在输出目录中的 filtered 文件夹中。如果过滤深度过高或文件过大，则建议进行进一步过滤。运行结束后，会在主界面列表中显示过滤结果的估算深度，用户可以在输出目录的 filtered 文件夹中查看每个基因过滤文件的大小。
**[分析>过滤>进一步过滤]**: 对过滤结果中过大或深度过深的数据进行进一步过滤，过大或深度过深的原始数据会储存在 large_files 文件夹中，filtered 中则保存进一步过滤之后的数据。
**[分析>拼接]**: 使用过滤后的序列进行拼接，拼接的最终结果保存在输出目录的 results 文件夹中。
**[分析>自动]**: 使用当前设定的测试自动完成过滤、（进一步过滤）、拼接的全部步骤，所有结果保存在输出目录中。
**[分析>迭代>首次迭代]**: 将输出目录中 contigs_all 中的序列作为参考序列，重新执行所有的过滤和拼接过程。结果保存在输出目录的 iteration 文件夹中。
**[分析>迭代>重新拼接]**: 使用迭代的结果进行拼接，而非使用原始参考序列过滤的结果，拼接结果直接保存在输出目录的 results 文件夹中。
**[分析>迭代>多次迭代]**: 可以自行设置迭代次数，进行多次迭代。迭代结果保存在输出目录的 iteration 文件夹中。
**[分析>刷新数据]**: 对于关闭程序后重新进行的分析，如果输出目录和之前分析的输出目录相同，在载入同样的参考序列之后，可以重新获取之前已经得到参考序列信息和拼接结果。

## 3.5 选择菜单

对参考序列进行重选，可进行全选、清空、反选。同时可以筛选选择失败、成功、过短、过深、过浅的项。

## 3.6 工具菜单


**[工具>细胞器基因组]**: EasyMiner调用NOVOPlasty进行细胞器基因组组装，选中一个参考序列作为种子序列，再载入数据文件，即可进行线粒体、叶绿体基因组的组装。通常保持默认参数即可，具体参数的含义详见NOVOPlasty的github主页https://github.com/ndierckx/NOVOPlasty。可以选择添加近源物种的叶绿体基因组序列作为参考序列，可以解决叶绿体重复区域的倒转重复问题。要进行更加细致的默认参数设定，可以手动编辑应用程序包analysis目录下的NOVO_config.txt文件，修改时请勿删除$及其之间的内容。点击确定按钮开始运行，所有结果将保存在输出目录。




**输入文件**:

Option_*_Project1.fasta: 可选组装结果。

Contigs_*_Project1.fasta: 拼接出来的contig。

Merged_contigs_Project1.txt: 把contigs拼接成环。

log_Project1.txt: 日志文件。

contigs_tmp_Project1.txt: 备用contig文件。

注意: 不要中途关闭命令行窗口。

**[工具>多序列比对]**: 使用 muscle 对参考序列和结果序列进行多序列比对。如果选择不合并，则仅仅对参考序列进行比对；选择合并，将结果序列与参考序列合并后再进行多序列比对。结果可在输出目录的 align 文件夹下查看。

# 4. 输出结果

## 4.1 输出目录

输出文件夹的目录结构和文件具体含义如下：
**align**：多序列比对的结果。
**contigs_all**：所有可能的组装结果。
**filtered**：过滤后得到的 fq 文件。
**iteration**：首次或多次迭代得到的文件，内部文件名和文件含义与上级文件夹相同。
**large_files**：进一步过滤时超过深度限制或者文件大小限制的原始 fq 文件。如果所有过滤结果都在限制以内，则不会出现该文件夹。
**log.txt**：日志文件。
**results**：拼接结果中权重最大的序列，即最终结果。

**kmer_dict_k31.dict**: kmer 字典文件，格式为：kmer 片段(十六进制)，kmer 计数（十六进制）。
**result_dict.txt**：结果文件，格式为：基因名，序列拼接状态，拼接上的 reads 数量。
**ref_reads_count_dict.txt**：每个参考基因序列拆分成 kmer 的总条数。

## 4.2 结果列表

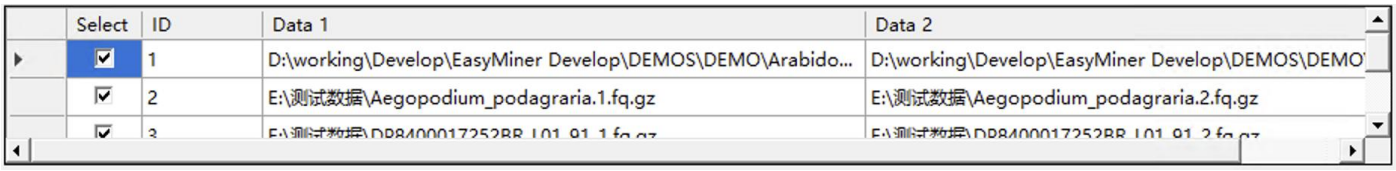
数据 | 日志

| Select | ID | Name | Ref. Count | Ref. Length | Filter Depth | Assemble State | Ass. Length | Ass. Depth |
|---|---|---|---|---|---|---|---|---|
| ☑ | 1 | ITS | 1 | 651 | 1915 | success | 1641 | 555 |
| ☑ | 2 | matK | 1 | 1581 | 367 | success | 2372 | 222 |
| ☑ | 3 | psbA | 1 | 1062 | 419 | success | 1919 | 196 |
| ☑ | 4 | rbcL | 1 | 1440 | 413 | success | 2166 | 243 |
| ☑ | 5 | rps16 | 1 | 1155 | 350 | success | 1874 | 196 |

Select：是否使用该条参考序列。
ID：参考基因的编号。
Name：参考基因的名称。
Ref. Count：参考基因的数量。
Ref. Length：参考基因的平均长度(bp)。
Filter Depth：使用参考基因过滤后的深度。过滤深度(Filter Depth)=reads 测序长度*过滤出的 reads 数量/参考序列的平均长度。
Assemble State：序列拼接的状态，包括：
- no reads：未过滤出 reads，请降低过滤 K 值或者提供更近源的参考序列
- distant references：参考序列过于远源，请提供更近源的参考序列
- insufficient reads：过滤出的 reads 太少，请减小过滤 K 值或者提供更近源的参考序列
- no seed：无法找到合适的种子，请减小拼接 K 值或者提供更近源的参考序列
- no contigs：没有拼接出结果
- low quality：结果准确度较低，reads 不足以覆盖拼接出的结果
- success：拼接成功

Ass. Length：拼接结果的长度
Ass. Depth：拼接结果的 reads 覆盖深度=reads 测序长度*用于拼接的 reads 数量/拼接结果的长度

### 4.3 数据列表




Select: 是否使用该组数据文件

Data1: 测序文件的左端(1 端)

Data2: 测序文件的右端(2 端)，如果是单端测序，则自动与 Data1 中的内容相同。

注意: 如果选取了多组数据，则会进行合并过滤，例如可以同时载入同一个目标物种的浅层测序和转录组测序，从而实现更好的挖掘效果。也可以同时载入同一个物种不同批次的测序文件。

注意 1: 如果同时选取不同 reads 长度的测序结果进行拼接，会导致无法获得正确的 depth，但依然可以使用。

注意 2: 除非有特别的理由，不建议同时使用不同样本的数据进行数据挖掘。

## 5. 常见问题

**1. 结果列表中过滤深度(filter depth)的含义?**

列表中显示的是如果将所有 reads 都用于拼接，理论上所能达到的最大深度，这个数值远大于实际拼接深度。

**2. 组装 kmer 值如何确定?**

将所有读长序列与参考序列进行比对，计算其最大共有序列的长度作为 kmer 值。

## 6. 参考文献

Dierckxsens N., Mardulyn P. and Smits G. (2016) NOVOPlasty: De novo assembly of organelle genomes from whole genome data. Nucleic Acids Research, doi: 10.1093/nar/gkw955

Dierckxsens N., Mardulyn P. and Smits G. (2019) Unraveling heteroplasmy patterns with NOVOPlasty. NAR Genomics and Bioinformatics, https://doi.org/10.1093/nargab/lqz011

Zhen Zhang, Pulin Xie, Yongling Guo, Wenbin Zhou, Enyan Liu, Yan Yu. Easy353: A tool to get Angiosperms353 genes for phylogenomic research. Molecular Biology and Evolution. msac261 (2022). https://doi.org/10.1093/molbev/msac261.

Baker W.J., Bailey P., Barber V., Barker A., Bellot S., Bishop D., Botigue L.R., Brewer G., Carruthers T., Clarkson J.J., Cook J., Cowan R.S., Dodsworth S., Epitawalage N., Francoso E., Gallego B., Johnson M., Kim J.T., Leempoel K., Maurin O., McGinnie C., Pokorny L., Roy S., Stone M., Toledo E., Wickett N.J., Zuntini A.R., Eiserhardt W.L., Kersey P.J., Leitch I.J. & Forest F. A Comprehensive Phylogenomic Platform for Exploring the Angiosperm Tree of Life. Systematic Biology. 71: 301–319. https://doi.org/10.1093/sysbio/syab035.