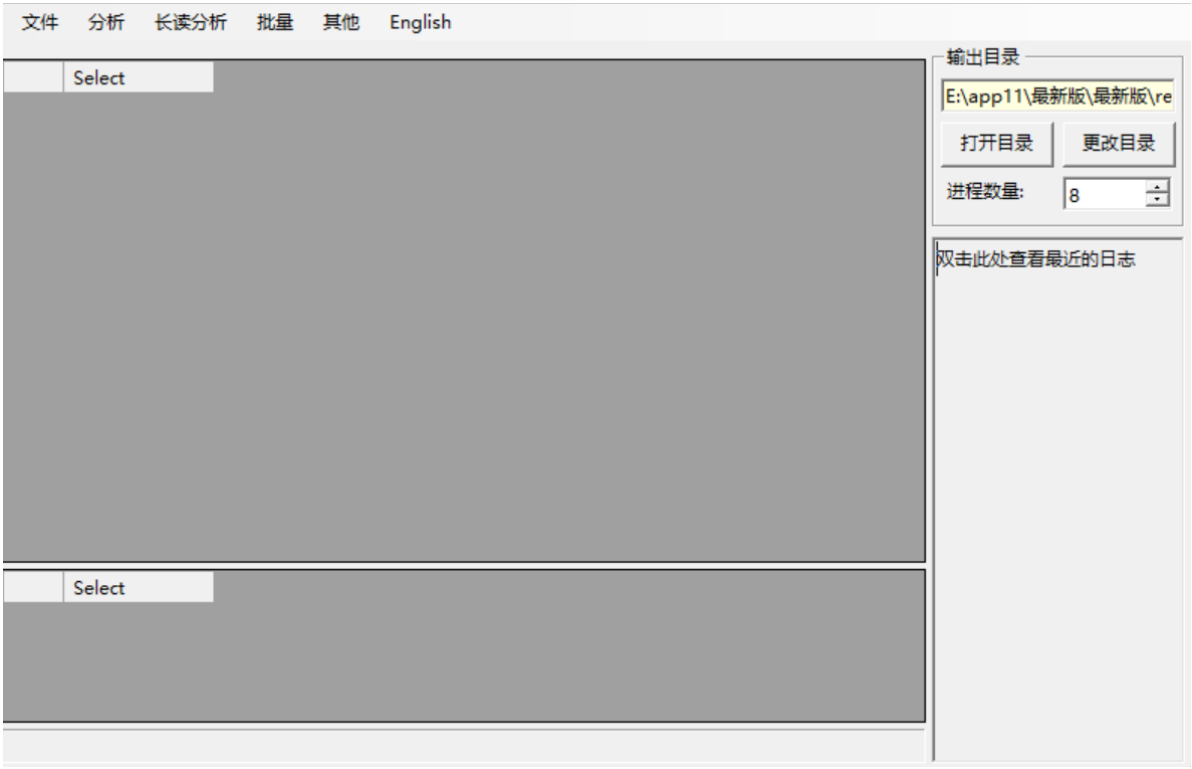
安装和需求



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的源代码均保存在github和Gitee上，您可以从此处获取最新的安装包:

Index of /database/app/EasyMiner (tpddns.cn)

或

EasyMiner download | SourceForge.net

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。

https://github.com/sculab/EasyMiner

使用方法

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(*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:

DEMO · sculab/EasyMiner - 码云 - 开源中国 (gitee.com)

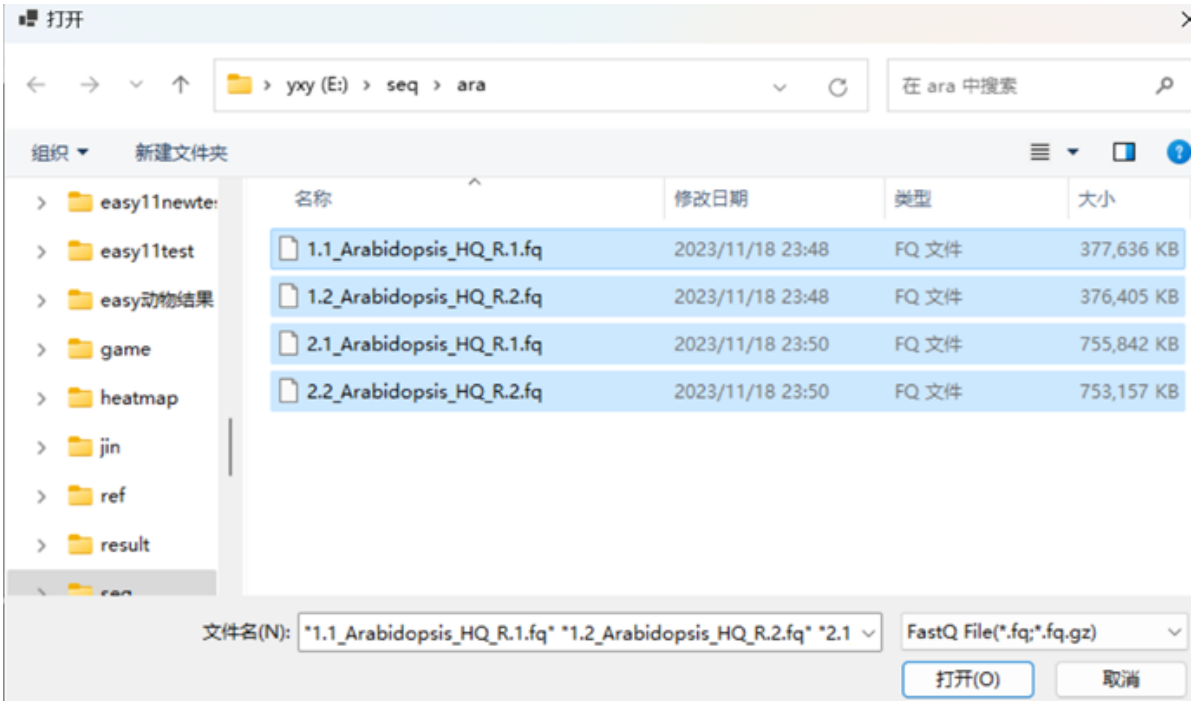
您也可以自行准备:

（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:

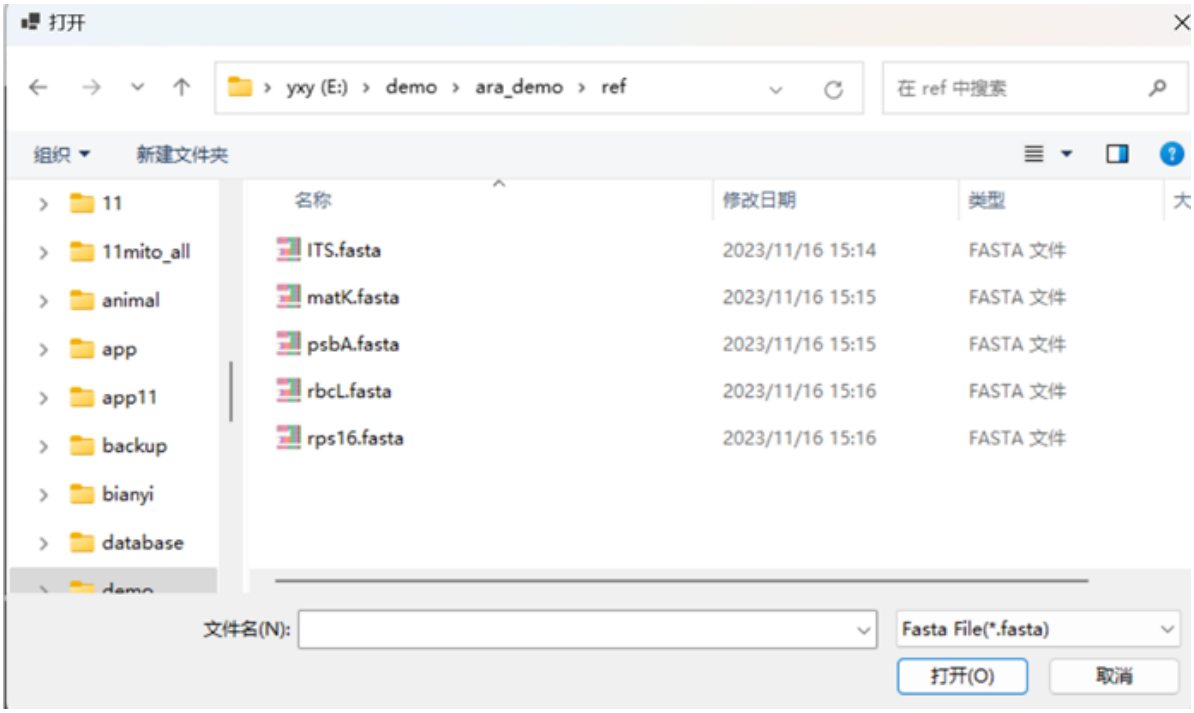
点击 **[文件>载入测序文件]** 选择测序数据文件。



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，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。

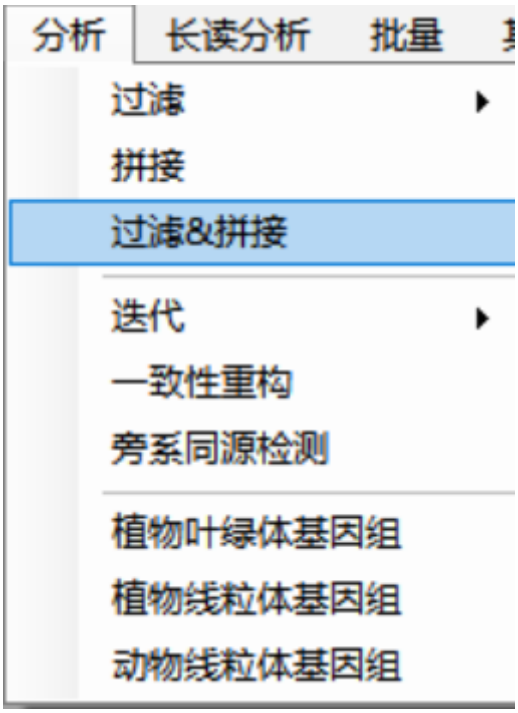


示例: 载入 DEMO/A_lyrata/ 下的所有fasta文件 (ITS、martK、psbA、rbcL、rps16)，包括1个核基因和4个叶绿体基因的参考序列，所有这些序列来都自拟南芥同属的近缘种琴叶拟南芥(*A. lyrata*)。

Select	ID	Name	Ref. Count	Ref. Length
☑	1	aly_its	1	649
☑	2	matK	1	1581
☑	3	psbA	1	1062
☑	4	rbcL	1	1440
☑	5	rps16	1	1155

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。

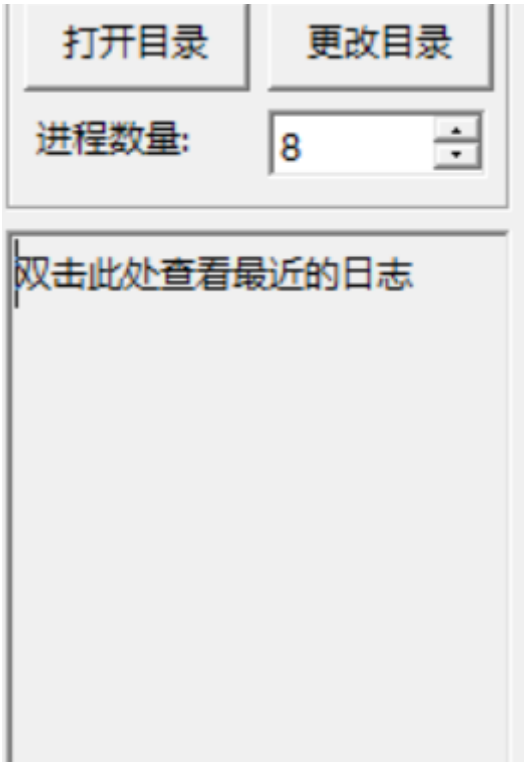
运行程序



点击[分析>过滤&拼接] 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

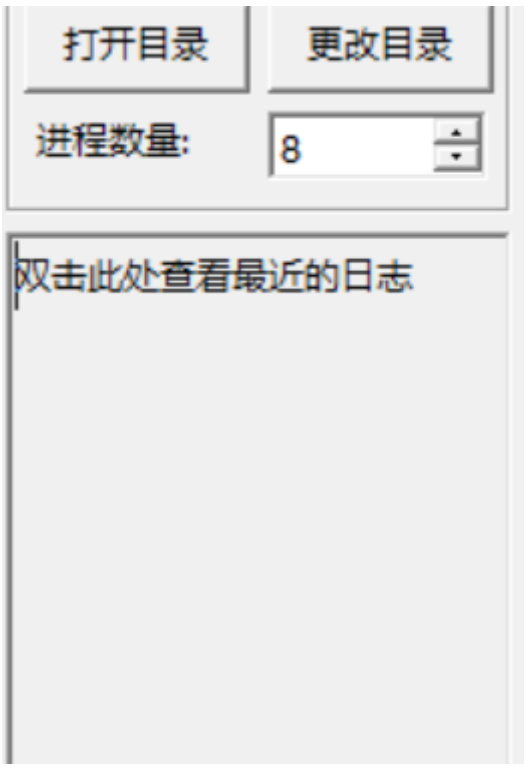


点击“打开目录”按钮，查看结果文件。拼接后的文件以fasta格式保存于results目录中。

更多实例请参见Gitee主页。

菜单

侧边栏:



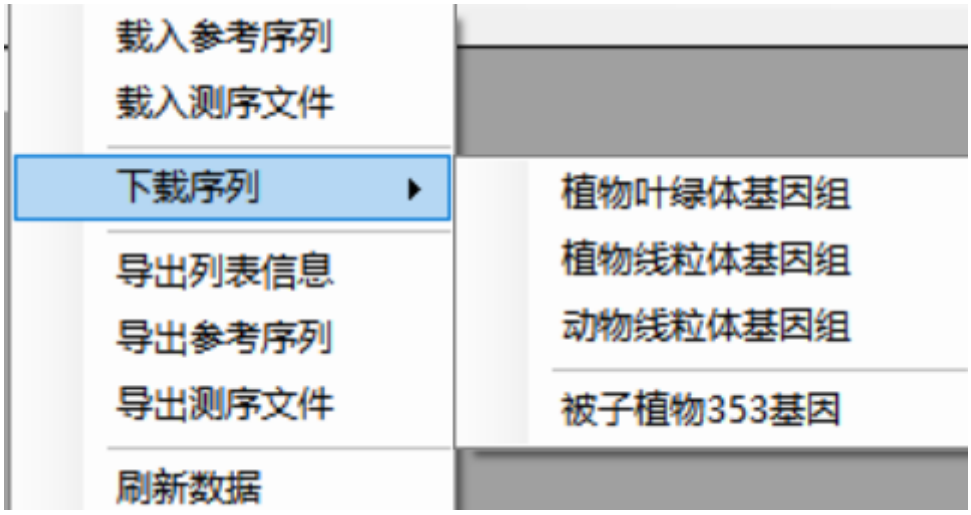
输出目录: 保存结果的文件夹，默认为EasyMiner应用程序所在目录的results文件夹。

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。

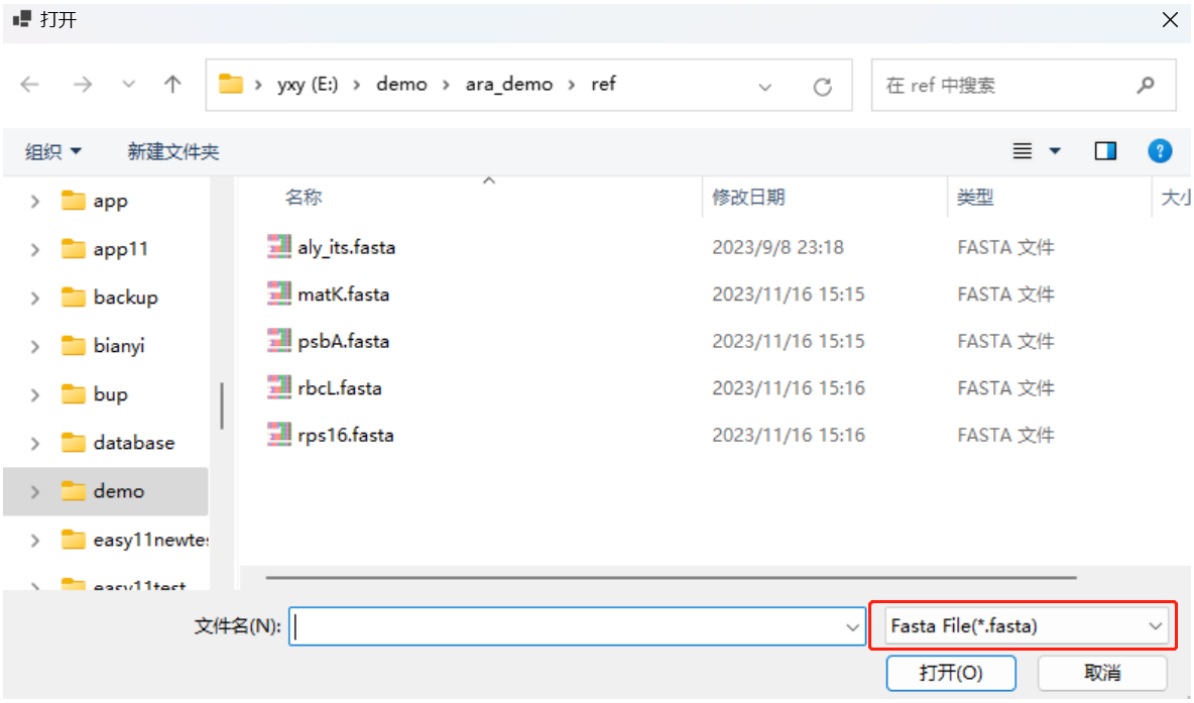
进程数量：默认为8，最大值为电脑逻辑处理器减一。可自行根据个人电脑内核决定进程数量。

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，右下角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

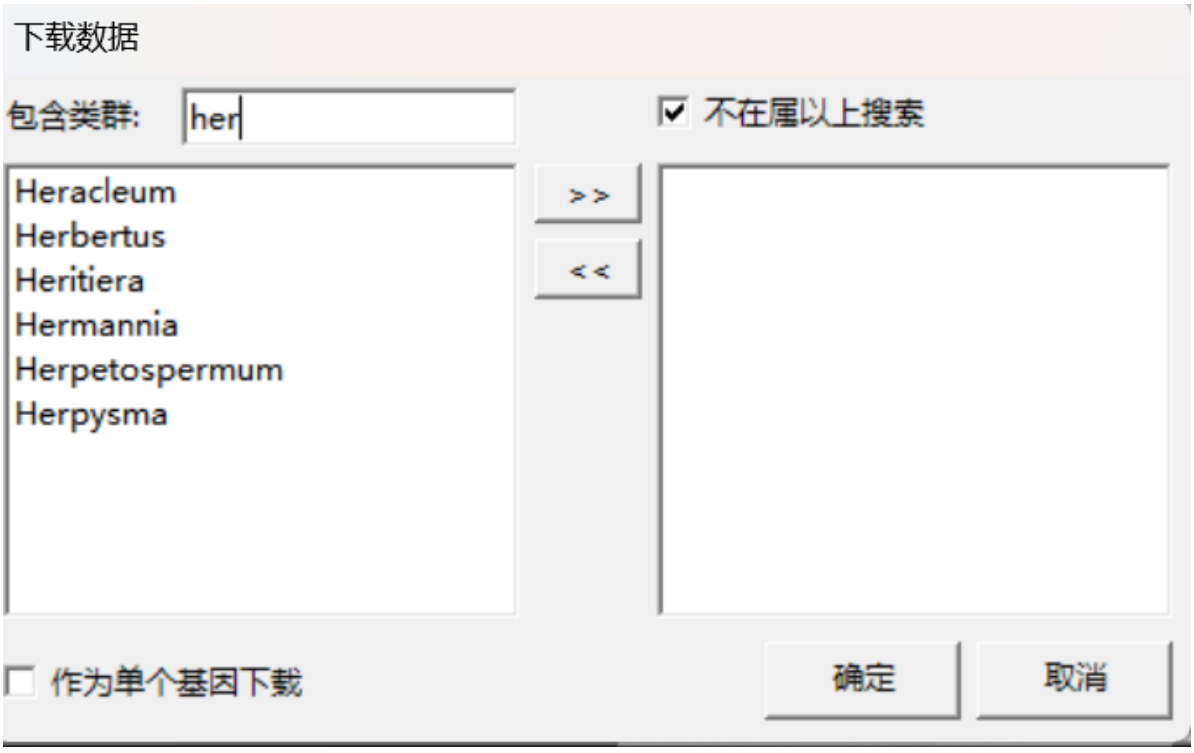
基因最大长度：要处理的基因的最大长度

扩展边界长度：在每个基因两侧延展的长度（扩展的内含子区长度）

[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>下载序列>下载植物叶绿体基因组] [文件>下载序列>下载植物线粒体基因组] [文件>下载序列>下载动物线粒体基因组]：：

从软件本地数据库获取细胞器基因组作为参考序列导入。本地数据库从NCBI（https://www.ncbi.nlm.nih.gov/）获取数据。



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，这意味着该类群在软件本地数据库没有数据，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。

不在属以上搜索：仅显示属级相关分类阶元。不勾选默认显示属级及以上分类阶元。建议勾选。

作为单个基因下载：将下载fasta格式的叶绿体全基因组序列。不勾选默认下载gb格式的叶绿体全基因组序列，将分割为多条基因参考序列导入。建议不勾选。

[文件>下载序列>下载被子植物353基因]：从软件本地数据库获取Angiosperms353 Gene Set (AGS)作为参考序列导入。本地数据库从Kew Tree of Life Explorer (https://treeoflife.kew.org)获取数据。

[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。

注意：如果在同一文件夹中导出，需要更改先前已经导出的序列名，否则将会被覆盖。

[文件>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的参考序列信息和拼接结果。

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。

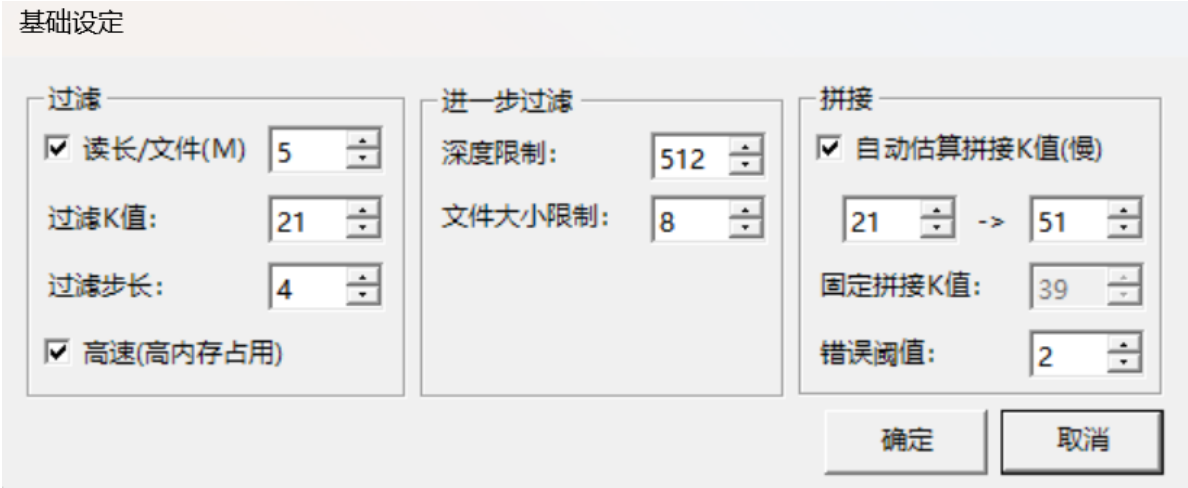
[分析>过滤>进一步过滤]: 对过滤结果中过大或深度过深的数据进行进一步过滤，过大或深度过深的原始数据会储存在large_files文件夹中，filtered中则保存进一步过滤之后的数据。

[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>过滤&拼接]: 使用当前设定的参数自动完成过滤、（进一步过滤）、拼接的全部步骤，所有结果保存在输出目录中。

当进行以上步骤会出现参数设定选项，具体含义如下：

基础设定：



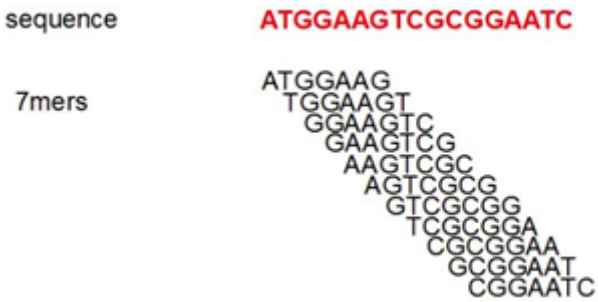
过滤：

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。

进一步过滤：

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。

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。

拼接：

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的kmer。

[分析>迭代>运行迭代]: 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。可以增强序列的长度和精度，建议运行。

[分析>迭代>用迭代覆盖]: 将results中的结果文件替换为迭代后的结果文件。

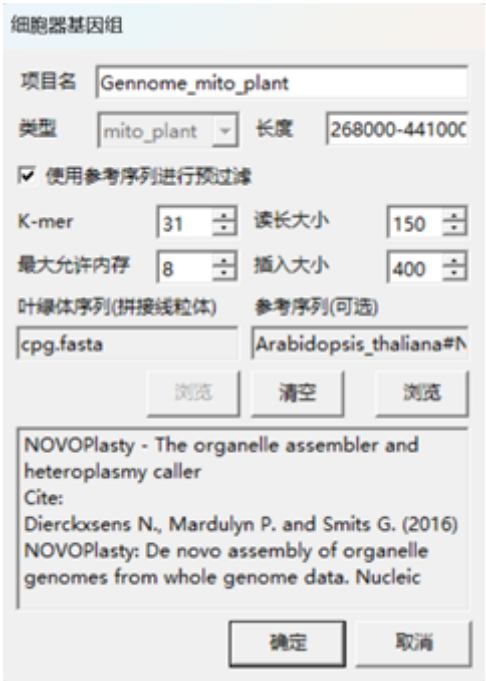
[分析>一致性重构]: 将结果序列和过滤后的fq文件进行映射。按照提示进行阈值设定，提高阈值会增加模糊碱基的数量，如果想区分混杂序列建议选择默认（0.75），如果想得到无简并碱基的结果建议选择0.25。

[分析>旁系同源检测]: 对提取的结果序列进行旁系同源基因的检测。

[分析>植物叶绿体基因组]: EasyMiner调用NOVOPlasty进行细胞器基因组组装，软件提供近源物种的叶绿体基因组序列下载作为参考序列并选用其中近源的作为种子序列，可以解决叶绿体重复区域的倒转重复问题；之后载入数据文件，即可进行叶绿体基因组的组装。通常保持默认参数即可。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的Organelle文件夹中。

[分析>植物线粒体基因组]: EasyMiner调用NOVOPlasty进行细胞器基因组组装，**需要选中先前拼接的叶绿体基因组**，再载入数据文件，即可进行线粒体基因组的组装。通常保持默认参数即可。软件提供近源物种的线粒体基因组序列下载作为参考序列，可以解决线粒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，所有结果保存在输出目录的Organelle文件夹中。

[分析>动物线粒体基因组]: EasyMiner调用NOVOPlasty进行细胞器基因组组装，软件提供近源物种的线粒体基因组序列下载作为参考序列并选用其中近源的作为种子序列；之后载入数据文件，即可进行叶绿体基因组的组装。



NOVOPlasty参数设置，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。

注意: 不要中途关闭命令行窗口。

批量菜单

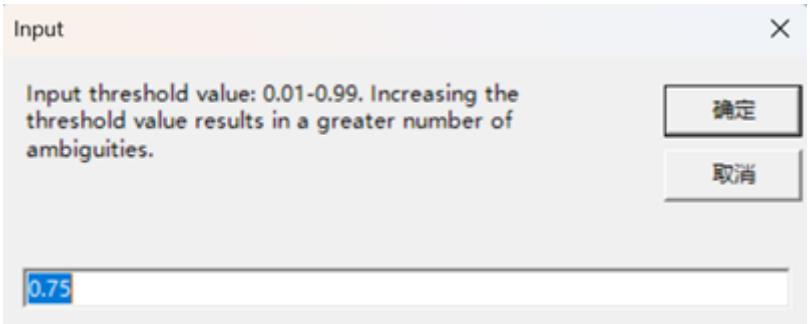
[批量>过滤&拼接]：对选中的不同物种的测序文件进行批量分析，使用当前设定的参数自动完成过滤、（进一步过滤）、拼接的全部步骤，所有结果以测序文件名命名保存在输出目录。

注意：此处弹出的**基础设定**窗口的具体含义见上述**[分析菜单]**中的介绍

[批量>合并结果]：将批量分析的结果合并，不同物种、同一基因的分析结果将合并在一个fasta文件中。

[批量>合并+切齐]：将批量分析的结果合并，并对合并后的fasta文件进行多序列比对并切齐。

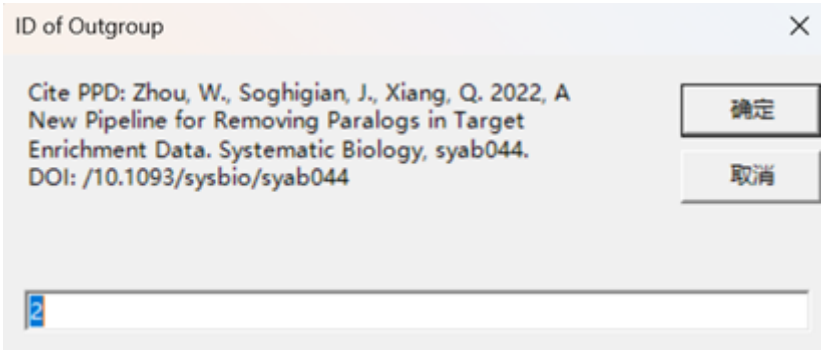
[批量>一致性重构]：将勾选物种过滤后的fq文件批量与结果序列进行映射。按照提示进行阈值设定，提高阈值会增加模糊碱基的数量。



阈值设定具体数值选择见**[分析菜单]**

[批量>旁系同源检测]: 对提取的结果序列进行旁系同源基因的检测。

[批量>PPD（仅用于353）]：**注意，一致性重构之后才可进行PPD**，即进行353提取结果的旁系同源检测。需要输入外类群的测序文件所在的ID号。



[批量>叶绿体基因组]：对选中的不同物种的测序文件进行批量叶绿体基因组拼接。

[批量>植物线粒体基因组]：对选中的不同物种的测序文件进行批量线粒体基因组拼接。

[批量>动物线粒体基因组]：对选中的动物不同物种的测序文件进行批量线粒体基因组拼接。

[批量>植物线粒体基因组]：对选中的植物不同物种的测序文件进行批量线粒体基因组拼接。

注意：有关细胞器基因组拼接的数值选择和下载规范，见上述**[分析菜单]**

[批量>统计汇总结果]：汇总过滤、拼接、一致性重构和旁系同源分析的结果。

输出结果

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。

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。

Organelle：细胞器基因组的拼接结果。

Gennome_cp.fasta：植物叶绿体基因组拼接结果。

Gennome_cp.gb：注释后的植物叶绿体基因组拼接结果。

Gennome_mito_plant.fasta：植物线粒体基因组拼接结果。

Gennome_mito_plant.gb：注释后的植物线粒体基因组拼接结果。

Gennome_mito_plant_warning.txt：报错日志文件，请检查错误原因并重试。

temp：因终端关闭终止分析，未完成的细胞器基因组的拼接过程文件。

Gennome_mito.fasta：动物线粒体基因组拼接结果。

Gennome_mito.gb：注释后的动物线粒体基因组拼接结果。

一致性重构结果：

consensus：将结果序列和过滤后的fq文件进行映射。存在设定值以上的模糊碱基数的序列将会被保留。

supercontigs：一致性重构的结果文件，使用IPUAC代码生成的退化序列，使用简并碱基标注了SNP位点。

旁系同源筛选结果：

paralogs：旁系同源基因筛选的结果文件，其中_ref.fasta文件储存旁系同源基因，csv文件记录不同位置碱基map出现的次数，.pec.csv文件记录碱基变异的频率。

summary.csv：统计汇总结果。

批量分析结果：

您的测序文件名：以测序序列名命名的文件夹，储存每个测序序列分别得到的拼接结果。

combined_results：储存合并后的结果文件。

combined_trimed：储存合并并切齐后的结果文件。

aligned: 多序列比对的结果。

PPD结果：

PPD>result>supercontig>s8 _rm_paralogs> Final_kept_genes：为最终结果文件，其他文件具体含义见PPD github（https://github.com/Bean061/putative_paralog#part2-trims-alignment-and-detectes-the-putative-paralogs）。

软件结果目录

输出结果目录

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

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:批量功能针对不同物种的测序序列进行分析。

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。

3. 测序数据是否需要去除接头和低质量reads?

建议使用测序公司提供的HQ版本的数据，使用低质量数据可能导致提取结果效果不好。如果没有HQ数据建议去除接头和低质量reads。

4. 得不到结果序列可能原因和解决办法?

选用的参考序列不够近源（手动查找更为近源的序列）

测序数据的深度太浅（可以尝试把过滤K值调低）

尝试迭代重新分析

*降低kmer得到的结果可能精确度不够，对于假阳性等错误序列，需要自己手动筛选分辨

5. 软件对电脑内存的需求?

对内存需求不大，可以对进程数量进行调节以适应电脑内存。

6. 我该如何获取内含子序列数据?
首先通过叶绿体基因组组装得到完整的gb文件，之后再将gb文件导入，勾选去除外显子区域，并在扩展边界长度选择您需要的内含子区长度。

7.有关切齐功能?

[批量>合并&切齐]是针对结果目录中的文件，进行批量比对并切齐。[其他>切齐比对]是将窗口中的参考序列对应结果目录中的序列文件，进行比对并切齐。

8.PPD没有结果?

PPD仅针对353数据应用，请保证选择了三个及以上的物种类群进行批量提取。

保证无中文目录文件夹。

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.

Wenbin Z,John S,Jenny Q X. A New Pipeline for Removing Paralogs in Target Enrichment Data.[J]. Systematic biology,2021,71(2).