

The manual of GFAP

目录

1、介绍.....	4
2、安装系统.....	4
3、安装.....	5
4、关于 GFAP 软件使用的具体信息.....	5
4.1 软件界面的介绍.....	5
4.2 软件功能介绍	6
4.2.1 功能注释	6
4.2.2 统计及绘图	9
4.2.3 其他功能	12
4.3 快速开始	13
4.3.1 GO/KEGG/Pfam	13
4.3.2 miRNA-lncRNA.....	14
4.3.3 基因家族注释	14
4.3.4 统计作图	15
4.3.5 气泡图以及网络图	16
4.3.6 翻译	17
4.3.7 RNA2DNA.....	17
4.3.8 extraction	18
4.3.9 conversion	19
4.3.10 对 Linux 系统用户.....	19
5、使用的注意事项	22

1. Introduction	24
2. Installation system.....	25
3. Installation	25
4. Specific information on the use of GFAP software.....	25
4.1 Introduction of software interface.....	25
4.2 Software function introduction	26
4.2.1 Functional annotation.....	26
4.2.2 Statistics and graphs	29
4.2.3 Other functions.....	33
4.3 Quick start	34
4.3.1 GO/KEGG/Pfam	34
4.3.2 miRNA-lncRNA.....	35
4.3.3 Gene Family Annotation.....	35
4.3.4 Statistical plotting.....	36
4.3.5 Bubble Chart and Network Diagram.....	36
4.3.6 Translation.....	37
4.3.7 RNA2DNA.....	38
4.3.8 Extraction	39
4.3.9 conversion	39
4.3.10 For Linux system users	40
5. Precautions for use	43

1、介绍

GFAP(Gene Functional Annotation for Plants)是用于植物基因功能注释的程序。该程序包括数据库和 GFAP 软件两个部分。目前, GFAP 数据库收录了来自 85 个科的 208 种植物, 这些植物涵盖了藻类, 苔藓, 地衣, 蕨类植物, 裸子植物, 双子叶以及单子叶植物在内的几乎所有植物类群以为用户提供足够的信息来对未知序列进行 GO(Gene Ontology)、KEGG(Kyoto Encyclopedia of Genes and Genomes)、蛋白结构域注释。考虑到基因家族以及 ncRNA(non-coding RNA, 非编码 RNA, GFAP 主要注释长链非编码 RNA 以及 microRNA) 信息同样在植物研究中发挥着重要作用, 因此, 在 GFAP 数据库中, 我们总结了目前主要研究的基因家族以及 ncRNAs 并收集以及构建了相关隐马尔可夫模型(HMM)以便对未知序列进行家族信息以及 ncRNA 方面的注释。我们是通过 Pfam 以及 KEGG 官网中的 HMM 模型对近缘物种库进行注释的。在注释完成后, 我们收集了注释这些植物基因所涉及的 HMM 模型构建了植物特异性的 HMM 库。因此, 除使用近缘物种信息对序列进行注释外, GFAP 同样支持使用 HMM 模型进行功能注释。除植物特有的 HMM 库, GFAP 也将提供总库信息以便用户对未知序列的注释。从这个角度讲, GFAP 适用于所有生物的基因功能注释。效率的差异是使用近缘物种信息以及使用 HMM 模型进行注释的重要区别点。测试阶段, GFAP 可以在 5s 内使用近缘物种库对 1 Mb 文件(2627 个基因)进行 GO、KEGG 以及蛋白结构域注释, 这个效率要远高于使用 HMM 模型进行注释的效率。当然, 选择以何种方式进行注释, 主要由用户自行决定。GFAP 软件则采用点击、拖拽等方式而不是采用命令行输入的方式实现功能注释。这将有助于湿生物学家们的功能注释分析。除注释功能外, GFAP 软件还提供了数据统计以及可视化功能(包括柱状图, 热图, 气泡图以及网络图等形式)以帮助用户更加便捷的分析数据。

2、安装系统

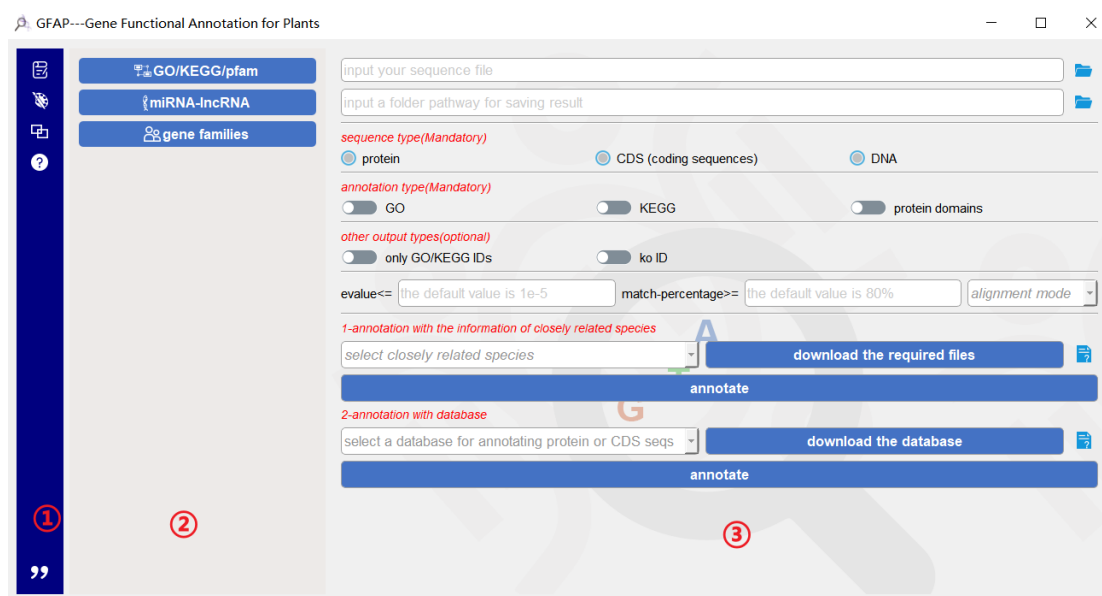
GFAP 可以在 Windows 以及 MacOS 系统上进行安装

3、安装

Windows 上直接双击 exe 文件即可安装。在 MacOS 系统中，是双击 pkg 文件进行安装的。
安装后程序默认会放在“Application”(即“应用”)文件夹下。

4、关于 GFAP 软件使用的具体信息

4.1 软件界面的介绍



第一个板块是模块板块 (①)。这里一共有“注释” (📄), “绘图” (📊), “其他功能” (🔧) 以及“帮助模块” (❓)。在每个模块下根据不同的功能类型，设定了不同的“选项”模块 (②)。第三个模块则是不同选项模块下的“功能”模块 (③)。“选项”以及“功能”的分布如下所示：




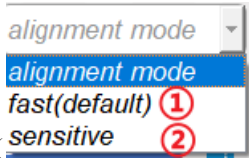
4.2 软件功能介绍


4.2.1 功能注释

GO/KEGG/Pfam

如箭头所示，用于进行 GO/KEGG/Pfam（蛋白结构域）注释。在①中输入 fasta 格式的序列文件；②填入保存路径。这里只需要填入文件夹名字即可，之后所有的注释结果将存入到

该文件夹中，可以采用拖拽的方式将文件以及文件夹直接拖入到指定的位置（①，②），当然也可以通过单击后面的文件夹图标（）来选择文件或者文件夹的路径；之后，需要在③处选择输入文件所含有的序列的类型，这个是**必选选项**；可以根据需要在④处选择注释的类型，这里是**必选**的，同时可以进行多选；考虑到 GO、KEGG 以及 ko ID 也是其他分析网站或者软件常用的输入信息，选择⑤处的选项后，会在生成注释结果的同时生成只含有 ID 的文件；GFAP 使用比对程序是 Diamond (<https://github.com/bbuchfink/diamond>)，设置了



fast 和 sensitive 两种比对的模式（），这两种比对模式在比对精度以及速度上有差异，用户可根据自身需要进行选择（这是非必须的选项，默认是 fast 模式）；⑥和⑦则分别是对比对结果进行筛选的设置，对 evalue(⑥)默认的是 1e-5，对匹配分数(⑦)默认的是 80%，如果这两个值可以满足条件，则不需要设置；如果用户想要使用近缘物种进行注释则需要在⑧处选择需要的物种，后面的⑨会根据用户选择的近缘物种以及用户所输入的序列类型，从我们的 ftp 网站上自动下载相应文件到指定位置。在我们的测试中该功能可能是有效的，但它在执行的过程中可能会受到用户网络环境的影响。为了确保用户的注释过程，我们将拟南芥的文件随软件携带（即不需要从网络下载），如果用户点击了“annotate”按钮而 GFAP 没有检测到相应文件则会使用所携带的拟南芥数据对用户序列进行注释。如果用户对该注释结果不满意，可以直接去 ftp 网址（<ftp://ftp.agis.org.cn/~panweihua/GFAP/>）自行下载数据并放入到相应文件夹下，之后点击按钮进行注释即可，需要的文件以及文件下载后需要放入 GFAP 的位置，如下表所示：

序列类型	ftp 的位置	需要放入 GFAP 的位置
蛋白 / CDS (coding sequences)	ftp://ftp.agis.org.cn/~panweihua/GFAP/protein-alignment/GFAP/protein-alignment (macos:在 GFAP 上右键，选择显示包内容，Contents，Resources，然后放入 protein-alignment，下同，不再赘述)
DNA (non-coding sequences)	ftp://ftp.agis.org.cn/~panweihua/GFAP/DNA-alignment/GFAP/DNA-alignment

如果用户想要使用 HMM 文件对序列进行注释，首先需要关注的是序列类型，因为目前含有的 HMM 文件是通过蛋白结构域进行构建的，因此，该功能只能接受蛋白或者 CDS 文件并且如果输入的是 CDS 文件，GFAP 会先将 CDS 文件进行翻译，之后注释。用户也可以选



择“其他”中的翻译功能（），在注释前先进行翻译，之后注释。在注释的时候，仍然需要先点击“download the database”对所需要的文件先进行下载，该过程是自动的，却仍有可能受到网络状况的影响，如果下载有问题(点击注释按钮后会显示没有检测到数据库)，用户可选择自行下载，需要下载的文件以及需要放置的位置如下表所示：

类型	ftp 的位置	需要放入 GFAP 的位置
KEGG	ftp://ftp.agis.org.cn/~panweihua/GFAP/database/akegg.txt.gz/GFAP/database/akegg.txt.gz
Pfam	ftp://ftp.agis.org.cn/~panweihua/GFAP/database/apg.txt.gz/GFAP/database/apg.txt.gz

miRNA-lncRNA

对 ncRNA(miRNA 以及 lncRNA)的注释包括以下几步：将 fasta 格式的序列文件放入①；通过②的选择来决定要对未知序列进行 miRNA 注释还是 lncRNA 注释；在③处放入 evaluate 值以便对结果进行筛选（默认值为 $1e-5$ ）；之后设定保存位置并命名保存文件，点击“annotate”即可进行注释。

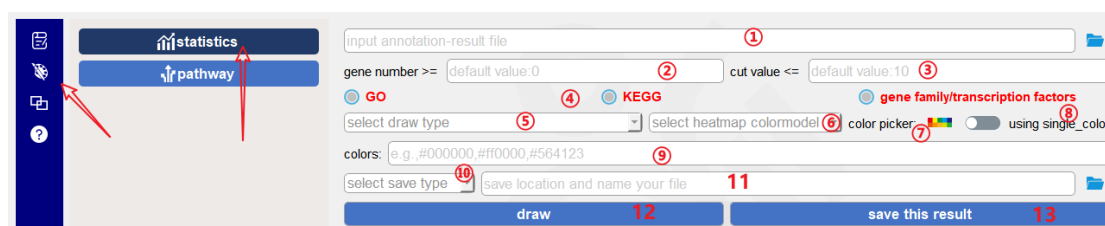
gene families

对家族成员的注释包括两部分功能。第一个部分是查找特定家族的成员并将提取序列，具体步骤：蛋白序列放入①；从②中选择需要的家族 HMM 模型；设定保存位置并命名

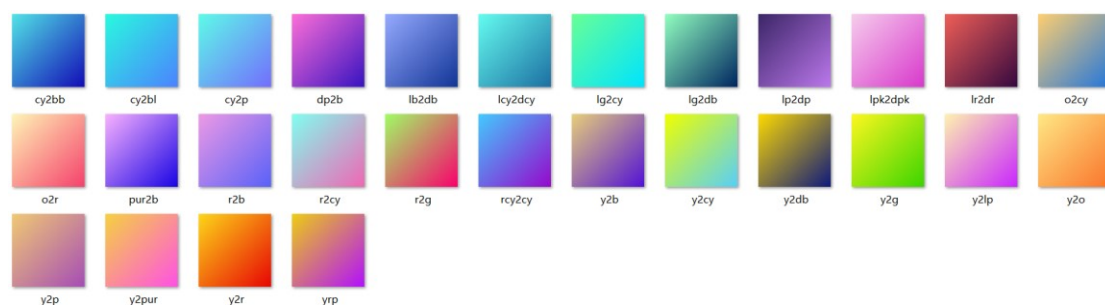
④；点击⑥；则家族成员 ID 会在⑦中显示；之后，点击⑨即可将相应序列从输入文件中提取出来。第二个部分是统计输入文件中所有序列都分布在哪些家族中，具体使用步骤：①中放入序列；④处设定保存位置并命名；⑤处选择需要注释的类型；之后点击⑧，需要的统计结果即会输入并显示在⑦处。


4.2.2 统计及绘图

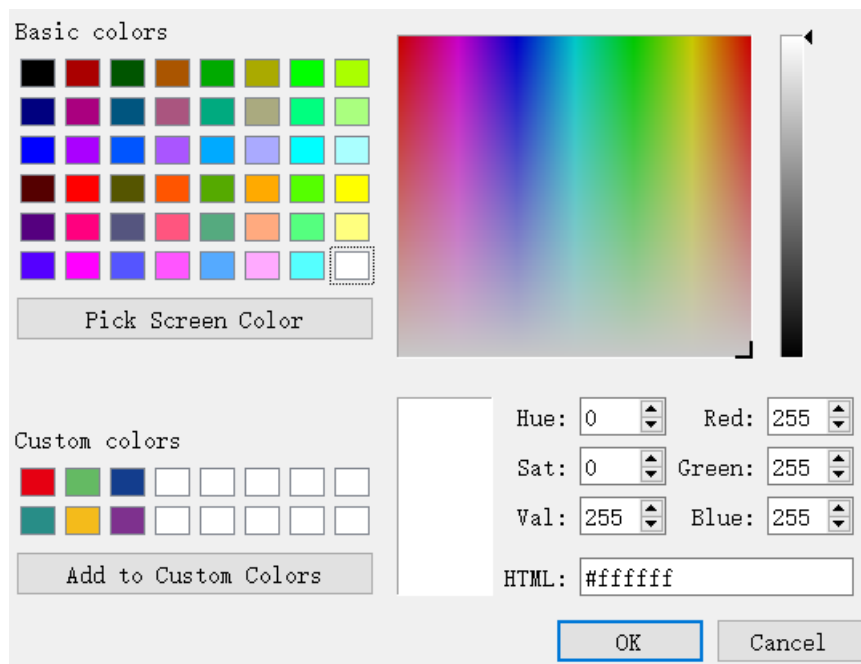
statistics



除注释功能外，GFAP 还能够对自身的注释结果进行统计和可视化。如上图所示功能主要用于绘制柱状图以及热图。将注释结果放入①，GFAP 会对注释结果中每个功能进行统计并计算具有该功能的基因个数，设置了两个参数对统计结果进行过滤；如②所示，是对基因数量进行过滤（默认值为 0，即不对基因数量过滤），例如在此处填写 10，则代表显示基因数量超过 10 的功能；③处的功能用于显示前 n%的功能（默认值为 10，即只显示前 10 的功能），例如在此处填写 10，则代表 GFAP 会显示拥有基因数量占前 10%的功能。这两处需要填入的数值为整数；接下来是必选选型（④），即需要用户告诉 GFAP 所放入的注释结果是关于什么的结果，因为不同类型的注释结果会有不同的展示方式，比如 GO 注释会以生物进程（biological process）、细胞组分（cellular component）以及分子功能（molecular function）分别进行显示；⑤处提供的是绘图类型的选择，即需要用户从柱状图或者热图中选择一个进行绘制（默认为柱状图）；⑥处给用户安排了 28 种配色方案以使用户根据自身喜好进行选择；以下是 28 种配色方案及代码：



同时，考虑这些配色方案可能仍然无法满足用户的需求，后面的 color picker () 则允许用户选择自己感兴趣的颜色，打开后的界面如下：



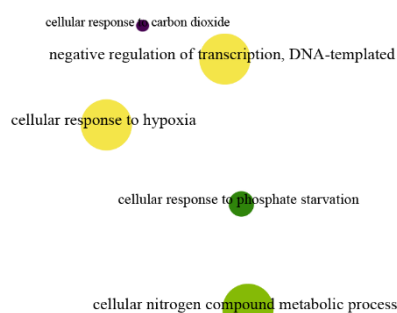
“pick screen color”可用于提取电脑屏幕中的任一颜色，“add to custom colors”可用于添加喜欢的颜色。在选定一个颜色后，单击 OK，所选择的颜色就会自动加入颜色对话框⑨中。这些是对于多色彩的设定。如果仅仅想用一种颜色表示注释结果，则可以勾选⑧，然后在上述颜色对话框中填入颜色，即可以这种颜色显示结果。在 GFAP 中我们设置的保存格式为 SVG 或者 PDF 格式（默认保存的是 SVG 格式），原因是这种矢量图能够最大程度的保证图片的清晰度。我们则推荐 Adobe Illustrator 对产生的图片做进一步的修改、整合等操作。在进行上述设定后点击⑫即可进行绘图，如果绘图结果满足要求，则可在⑪处设定保存位置，然后点击⑬保存图片。

Pathway

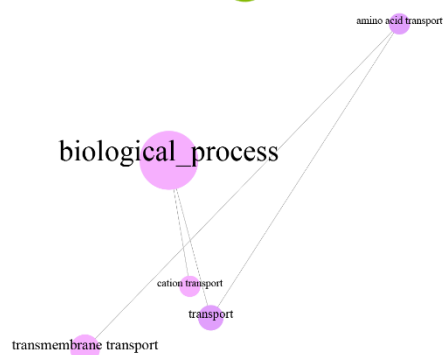


该部分主要用于绘制气泡图和网络图。

气泡图：



网络图：



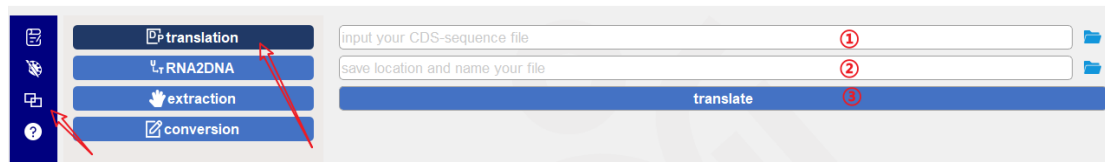
基本参数的设定及操作如前。与之不同的是这里我们通过 t 检验对 pvalue 值进行计算，计算该值的前提是需要选择近缘物种作为背景数据集⑦。在输入文件并设定相应参数后，点击⑪完成的是气泡图，同时计算的结果会显示在⑬。如下图所示：

#					
GO:0005524	ATP binding	278	0		
GO:0020037	heme binding	36	0		
GO:0005509	calcium ion binding	33	0		
GO:0003676	nucleic acid binding	68	0		
GO:0043565	sequence-specific DNA binding	64	0		
GO:0003677	DNA binding	92	0		
GO:0005515	protein binding	237	0		
GO:0008270	zinc ion binding	150	0		
GO:0003723	RNA binding	95	0		
GO:0046872	metal ion binding	55	0		
#					
GO:0005524	ATP binding	278	0		
GO:0022857	transmembrane transporter activity			61	0
GO:0003677	DNA binding	92	0		
GO:0016791	phosphatase activity	70	0		

具有相互关系的一组数据由#键进行分类。如果用户需要将气泡图中具有相互关系的功能展示出来（即想要绘制网路图），则在此基础上点击 14 完成绘图。功能 14 的数据来源即是显示在 16 中的内容，所以在绘制网络图前，用户也可以通过编辑 16 中的内容来选择需要展示的内容。

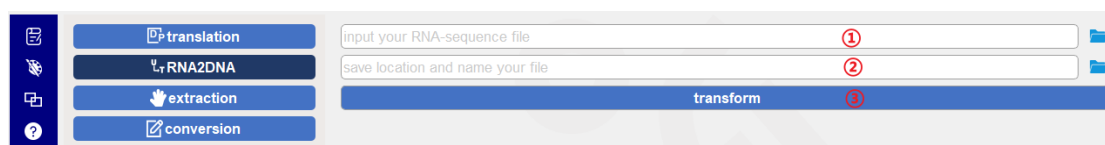
4.2.3 其他功能

translation



GFAP 支持对 DNA 以及蛋白序列的注释。尽管如此，我们仍然希望用户尽可能使用蛋白序列来完成注释过程。考虑到有些用户可能只有 CDS 序列文件。因此，我们设置了翻译功能，它能够将 CDS 文件中的基因序列批量转为蛋白文件以用于之后的注释。翻译过程：将 CDS 文件放入 1；在 2 处设置保存的位置并命名；点击 3 执行功能。

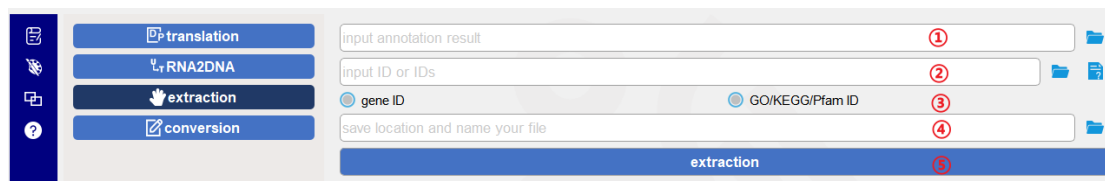
RNA2DNA



在注释 ncRNA 序列时，所能够识别的是 DNA 序列。因此，设置了可以将 RNA 序列转为 DNA 序列的功能。将 RNA 序列整理成 fasta 格式，放入 1；输入保存位置并命名文件 2；

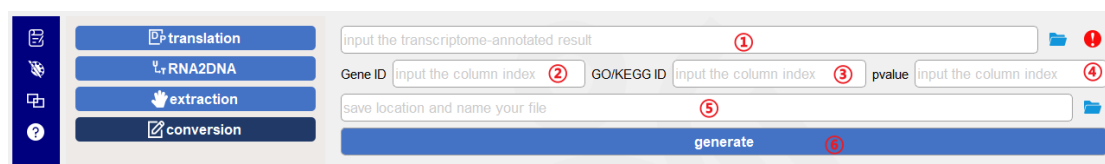
点击③完成转换。

extraction



GFAP 支持基因组水平的基因注释过程。但在实际的操作过程中，我们往往只关注部分基因，这个时候就需要将关注的基因从整体文件中提取出来。在 GFAP 中，我们设计的提取功能的输入文件是注释结果①；同时需要输入 ID②（可以是单个 ID 也可以将感兴趣的基因整理到一个文件，然后以之作为第二个输入文件）；并且这里的 ID 可以是基因的 ID 或者是 GO/KEGG/Pfam ID (③)；保存并命名 (④)；点击⑤执行功能。

conversion



设置转换功能的目的是将转录组分析的结果转为 GFAP 能够识别的格式从而使用 GFAP 进行相应分析。将转录组的结果放入①，这个转录组结果的格式为 txt 并以 tab 键作为列与列之间的分隔符。接下来需要告知 GFAP 如下信息：1、基因 ID 位于输入文件的第几列②；2、GO/KEGG ID 位于第几列③；3、pvalue 值位于第几列④。输入这些值后保存并命名，之后点击⑥执行功能即可。

4.3 快速开始

该部分叙述的内容所有具有默认值的参数都保持默认参数不变，在实际操作中用户需要根据自身的需求进行设定，一般而言，默认值能够满足大部分用户的需求。

4.3.1 GO/KEGG/Pfam

1、放入蛋白文件



2、设定工作目录

E:\try\result

3、选择输入的序列类型

sequence type(Mandatory)
☒ protein ☐ CDS (coding sequences) ☐ DNA

4、选择需要的注释类型

annotation type(Mandatory)
☒ GO ☒ KEGG ☒ protein domains

当使用近缘物种注释：

5、选择近缘物种单击下载，下载完成后单击“annotate”完成注释

1-annotation with the information of closely related species
Aquilegia_coerulea(Ranunculaceae) download the required files
annotate

当使用 HMM 文件进行注释，单击下载按钮下载文件，然后单击“annotate”等待注释完成

2-annotation with database
plant-special database download the database
annotate

4.3.2 miRNA-lncRNA

1、输入 fasta 文件

E:\example\test\miRNA.txt

2、选择序列类型

the input type: ☒ miRNA ☐ lncRNA

3、设定保存位置并命名，单击“annotate”完成注释

E:\example\test\annotate_result.txt
annotate

4.3.3 基因家族注释

1、输入蛋白文件

E:\try\Populus_trichocarpa.Pop_tri_v3.pep.all.fa

如果对单个家族进行注释：

2、选择需要的 HMM 模型

ARF (Auxin response factors)

3、保存并命名

E:\try\ARF.txt

4、单击按钮执行功能

show members of a single family

5、单击按钮自动提取序列

extract sequences

如果统计输入文件包含的所有家族:

2、保存并命名

E:\try\ARF.txt

3、选择需要分析的类型

Please select one of the following choices if you want to identify the members of all families

☒ transcription factor

☐ gene family

4、单击按钮完成功能

show genes containing domains of families

4.3.4 统计作图

1、输入注释结果文件

E:\try\result\GO_database_result.txt

2、指明是哪种注释类型

☒ GO

☐ KEGG

☐ gene family/transcription factors

3、单击“draw”绘图

draw

4、选择保存类型，输入保存位置并命名文件

svg E:\try\result\draw_statistics.svg

5、单击保存即可

save this result

4.3.5 气泡图以及网络图

1、输入注释的结果文件

E:\try\result\GO_database_result.txt

2、选择注释类型

GO KEGG gene family/transcription factors

3、选择近亲物种以计算 pvalue，如果是对 GO 注释结果进行可视化，则需要从三大类本体中选择一类进行可视化

select colormodel Abies_alba(Pinaceae) molecular_function

4、点“draw”绘图

draw

5、保存时，选择保存类型，输入保存位置并命名

svg E:\try\result\draw_statistics.svg draw save this result

6、编辑统计内容

select colormodel	Abies_alba(Pinaceae)	molecular_function
select save type	save location and name your file	
draw		save this result
GO:0004672	protein kinase activity	112 0
GO:0016887	ATPase activity	55 0
GO:0003700	DNA-binding transcription factor activity	72 0
#		
GO:0005524	ATP binding	278 0
GO:0020037	heme binding	36 0
GO:0005509	calcium ion binding	33 0
GO:0003676	nucleic acid binding	68 0
GO:0043565	sequence-specific DNA binding	64 0

7、点“draw network”绘制网络图

draw network

8、保存时，选择保存类型，输入保存位置并命名

svg E:\try\result\draw_pathway.svg draw save this result

4.3.6 翻译

- 1、输入 fasta 格式的 cds 文件

E:\example\test\CDS.fa

- 2、选择保存位置并命名

E:\example\test\dna2protein.fa

- 3、执行功能

translate

“fasta 格式的 cds 文件”，格式如下：

```
>Potri.T155100.3 pacid=37219601 polypeptide=Potri.T155100.3.p locus=Potri.T155100 ID=Potri.T155100.3.v3.1 annot-version=v3.1
ATGGCTATATCGAAGCTTTTGATTGTTTTCTTGTCGCATCTCTCTTGCTCCGCCCTTGTTGAAGCTGATCAGAAGGT
GGTGAACCTCAAATATTCAGCTGCTAGCTATCCTCTGGGAAGAATATCGCAGATTGTGGTGGCGCTTGCAATGCTAGGT
GTTCTTATCTCCAGGCCACGCTTTGCAAGAGGGCTTGCGGGACTTGCTGTGCACGATGCAAGTGTGCCCCCAGGC
ACTTCCGGCAACCATTATACCTGCCCTTGCTATGCCACCATGACTACTCGAGGTGGCCGACTCAAGTGTCTTGA
>Potri.T155100.1 pacid=37219602 polypeptide=Potri.T155100.1.p locus=Potri.T155100 ID=Potri.T155100.1.v3.1 annot-version=v3.1
ATGGCTATATCGAAGCTTTTGATTGTTTTCTTGTCGCATCTCTCTTGCTCCGCCCTTGTTGAAGCTGATCAGAAGGT
GGTGAACCTCAAATATTCAGCTGCTAGCTATCCTCTGGGAAGAATATCGATTGTGGTGGCGCTTGCAATGCTAGGTGTT
CCTTATCCTCCAGGCCACGCTTTGCAAGAGGGCTTGCGGGACTTGCTGTGCACGATGCAAGTGTGCCCCCAGGCCT
TCCGGCAACCATTATACCTGCCCTTGCTATGCCACCATGACTACTCGAGGTGGCCGACTCAAGTGTCTTGA
>Potri.T155200.1 pacid=37219603 polypeptide=Potri.T155200.1.p locus=Potri.T155200 ID=Potri.T155200.1.v3.1 annot-version=v3.1
ATGAAGGCAACGAGGAGCGGATACAGCGGTGGCGAGATGGGTAAAGGAGCAACACCAAGATGAAGGCTTTTTAGC
GGTGGTTTCAGGGTTGGCGGCTTGTTTTCTTAGAATGGTTGTTGCTGATCATGATAACCTCTTTGTTGCTGCTGAGG
TTGTTCACTTATTTGAATCTCTGTTCTTATTACAAGCTCATGAAGGAGAAGACTTGTGCTGGACTTCTACTAAAATCA
CAGGAGCTAATAGCTATATTTTAGCTGTTAGGCTCTATTGCAGTTTTGTCATGGAGTATGACATACACACTCTACTTGA
TTCAGCTACGTTGCTGACACCCCTTTGGGTAATCTATACGATCCGTTTCAACTTGAGGTCCAGTTACATGGAAGGCAAG
ATAAATTTGCAATTTACTATTTGGTGATACCATGTGCTGTGCTAGCTTTGTATATTCATCCAAGAACACATCACCACATA
GTCAACAGGATTGCTGGGCTTTTGTTTACCTTGAATCTATTTCAAGTTGCTCAGCTGCGAGTAATGCAAAACAC
AAAGATTGTTGAACCTTCCAGGCACATTATGATTTCGCTTGGAGTTGCCAGGTTCTTGGGTGTGCACATTGGATCC
TCCAGTACTGGACACTCGAGGACGCTATTGACAGCACTGGGCTATGGAATGTGGCTCTTATGCTCTACTTTCAGAA
ATTGTGCAGACATTCTTTGCCGATTTTGCTACTACTACGTCAAGAGTGTCTTGGTGGCAACTGTTCTAAGGCT
CCCCCTCAGGAGTGGTGTA
```

4.3.7 RNA2DNA

- 1、输入 fasta 格式的 RNA 序列文件

E:\example\test\miRNA.txt

- 2、选择保存位置并命名

E:\example\test\DNA2RNA.txt

- 3、点击按钮即可执行功能

transform

“fasta 格式的 RNA 序列文件” 格式：

```

1 >miR156
2 UGACAGAAGAGAGUGAGCAC
3 >miR164
4 UGGAGAAGCAGGGCACGUGCA
5 >miR166
6 UCGGACCAGGCUUCAUCCCC
7 >miR168
8 UCGCUUGGUGCAGAUCCGGAC

```

4.3.8 extraction

1、输入注释结果

E:\try\result\GO_database_result.txt

2、当放入单个 ID 时：

GO:0050660

或者可以放入含有多个 ID 的文件

E:\try\result\IDs.txt

格式为：

```

GO:0009058
GO:0008236
GO:0004842
GO:0006801
GO:0016020
GO:0050660
GO:0020037
GO:0020037
GO:0016788
GO:0055114
GO:0055114
GO:0009165
GO:0055114
GO:0009733
GO:0035145

```

3、选择输入的 ID 的类型

gene ID GO/KEGG/Pfam ID

4、设定保存位置并命名

E:\try\result\GO_extraction.txt

5、点击按钮提取信息

extraction

4.3.9 conversion

1、放入转录组结果

E:\try\transcriptome_example.txt

2、放入关键词所在的列 ID

Gene ID 1 GO/KEGG ID 28 pvalue 5

3、放入保存位置并命名

E:\try\conversion_result.txt

4、点击按钮生成相应文件

generate

“转录组结果”格式：

gene_id	HL4_readcount	HL0_readcount	log2foldchange	pval	padj	Gene Length	NR GI	NR ID	NR Score	NR Evalue	NR Description	NT GI	NT ID	NT Score
cluster-36310.157224	43.08993596	16404.86177	-8.5675	1.08E-105	3.84E-100	1406	K00799
cluster-36310.62757	2418.036399	66.53141246	5.1824	1.40E-100	2.48E-95	325	566189800	XP_006378352.1	343	1.00E-30	hypothetical protein POPTR_0010s08600g [Popul
cluster-36310.157767	4.878306409	2279.082823	-8.8766	8.15E-91	9.63E-86	1384
cluster-41605.0	3940.35515	41.90246648	6.5543	3.85E-89	3.42E-84	1436	224747075	ACN62215.1	147	2.50E-07	"omega-gliadin, partial [Triticum aestivum x Loph
cluster-36310.62762	2818.833033	60.9114402	-5.5307	7.97E-79	5.65E-74	328	566254885	XP_006387619.1	249	8.30E-20	hypothetical protein POPTR_0770s00220g [Popul

4.3.10 对 Linux 系统用户

需要安装 python3 版本并将版本加入到环境命令中。

解压后需要将路径切换到 GFAP 文件夹，使用命令：

```
cd */GFAP
```

4.3.10.1 GO/KEGG/Pfam/nr/swissprot annotation

4.3.10.1.1 使用近缘物种注释

python GFAP-linux.py -go/kegg/pfam -qp/qn (protein/CDS 序列文件) -aws (物种拉丁学名) -o (保存路径，无需命名文件，只提供路径即可)。例如：

```
python GFAP-linux.py -qp example_protein.txt -aws Rosa_chinensis -go -o ./
```

也可以设置 cpu 个数以及不同的模式 (fast/sensitive, fast 模式在保证一定正确率的前提下拥有极高的比对效率, sensitive 则对比对结果更为敏感, 但效率相对低)。结果以

GO/KEGG/PFAM 呈现。

4.3.10.1.2 使用数据库注释

有 4 种数据库可供选择：植物专用数据库（plant-special database, PSD），总数据库(total database, TD)，nr 数据库以及 swissprot。PSD 是在我们注释超 200 个物种后，搜集了与这些物种基因相关的所有 hmm 模型后构建。其余数据库是直接从 Pfam, KEGG, NCBI 等官网直接下载的最新版本，我们对其中的 nr 数据库做了针对植物的过滤操作并将 nr 以及 swissprot 库制成了比对数据库。从效果来看，PSD 与 TD 数据库的准确性高于 nr 和 swissprot，但注释的基因相对要少，因此，关于数据库的选择需要根据用户自身需求灵活选择；从注释结果看，PSD 与 TD 数据库会注释 GO/KEGG/Pfam，而另外两种则直接以比对最高的蛋白功能进行注释。

python GFAP-linux.py -go/kegg/pfam -qp/qn (protein/CDS 序列文件) -awd (数据库名称) -o (保存路径，无需命名文件，只提供路径即可)。例如：

```
python GFAP-linux.py -qp example_protein.txt -awd psd -pfam -o ./
```

4.3.10.2 筛选特定家族成员

python GFAP-linux.py -sf -qp (蛋白序列文件) -mn/mp (输入库中家族简写/如果库中没有所需要的家族，可以自行下载 hmm 文件并输入文件位置) -o (结果文件的保存路径)。例如：

```
python GFAP-linux.py -sf -qp ./example_protein.txt -mn WRKY -o gene_family.txt
```

```
python GFAP-linux.py -sf -qp ./example_protein.txt -mp ./WRKY.hmm -o gene_family.txt
```

这条命令除了生成与该家族模型相关的基因 ID 文件外，还会生成包含这些 ID 序列的 fasta 文件。

4.3.10.3 筛选输入文件中包含的、所有与家族相关的蛋白结构域

python GFAP-linux.py -mf -atf/agf (筛选转录因子/非转录因子家族) -qp (蛋白序列文件) -o (结果文件的保存路径)。例如：

```
python GFAP-linux.py -mf -agf -qp ./example_protein.txt -o gene_family.txt
```

```
python GFAP-linux.py -mf -atf -qp ./example_protein.txt -o gene_family.txt
```

4.3.10.4 识别输入的 ncRNA 序列是否为已知 ncRNA

python GFAP-linux.py -na -nt (miRNA/lncRNA) -qn (序列文件) -o (结果文件的保存路径)。例如：

```
python GFAP-linux.py -na -nt miRNA -qn ./ncRNA.txt -o ncRNA_result.txt
```

4.3.10.5 翻译

python GFAP-linux.py -t -qn (CDS 文件) -o (结果文件的保存路径)。例如：

```
python GFAP-linux.py -t -qn TAIR10_cds.fa -o protein.txt
```

4.3.10.6 将 RNA 转 DNA

python GFAP-linux.py -rd -qn (CDS 文件) -o (结果文件的保存路径)。例如：

```
python GFAP-linux.py -rd -qn RNA.txt -o DNA.txt
```

4.3.10.7 根据输入 ID 或者含有多个 ID 的文件（一个 ID 一行）来提取注释文件中的内容

python GFAP-linux.py -ex -ar (注释结果文件) -ID (ID 或者 ID 文件) -exfid/exgid (需要告知程序想要提取的是功能 ID 还是基因 ID) -o (结果文件的保存路径)。例如：

```
python GFAP-linux.py -ex -ar GFAP-Arabidopsis_thaliana\Brassicaceae\kegg_annotate.txt -ID ./gene IDs.txt -exgid -o result.txt
```

```
python GFAP-linux.py -ex -ar GFAP-Arabidopsis_thaliana\Brassicaceae\kegg_annotate.txt -ID function IDs.txt -exfid -o result.txt
```

4.3.10.8 针对转录组及与转录组呈现结果方式类似的其他组学的格式转换

python GFAP-linux.py -cf -gf (结果文件) -gid (基因 ID 在结果文件中的 index) -fid (go/pfam/kegg 在结果文件中的 index) -pvalue (pvalue 在结果文件中的 index) -o (结果文件的保存路径)。例如：

```
python GFAP-linux.py -cf -gf transcriptome_example.tab -gid 1 -fid 30 -pvalue 5 -o result.txt
```

该命令会将组学结果不利于后续分析的格式，如下：

BP Description	Gene Ontology Molecular Function
--	--
oxidation-reduction	GO:0005506//GO:0020037//GO:0016705
glycerolipid metab	GO:0004144
photosynthesis	--
response to wound	GO:0004867
--	--
gluconeogenesis	GO:0005525//GO:0004611
--	GO:0005515
blood coagulation	GO:0008191//GO:0034235
--	GO:0005524
--	--
oxidation-reduction	GO:0016491
response to wound	GO:0004867

转为基因 ID 与单个功能 ID 对应的格式以便于后续分析，如下：

```
Cluster-36310.161598 GO:0016747 transferase activity, transferring acyl group
Cluster-36310.156351 GO:0005509 calcium ion binding molecular_function 8.44E
Cluster-36310.156351 GO:0008266 poly(U) RNA binding molecular_function 8.44E
Cluster-36310.156351 GO:0010242 oxygen evolving activity molecular_function
Cluster-36310.251474 GO:0008121 ubiquinol-cytochrome-c reductase activity m
Cluster-36310.251474 GO:0016849 phosphorus-oxygen lyase activity molecular
Cluster-36310.251474 GO:0005509 calcium ion binding molecular_function 1.17E
Cluster-36310.86742 GO:0016651 oxidoreductase activity, acting on NAD(P)H molec
Cluster-36310.163845 GO:0003723 RNA binding molecular_function 2.55E-24
Cluster-36310.183781 GO:0005515 protein binding molecular_function 2.53E-24
Cluster-36310.183781 GO:0016820 ATPase-coupled transmembrane transporter acti
Cluster-36310.38952 GO:0005515 protein binding molecular_function 4.23E-24
Cluster-36310.143554 GO:0008168 methyltransferase activity molecular_function
Cluster-36310.160747 GO:0016491 oxidoreductase activity molecular_function 7
Cluster-36310.225771 GO:0004866 endopeptidase inhibitor activity molecular
Cluster-36310.161571 GO:0005509 calcium ion binding molecular_function 7.93E
Cluster-36310.163698 GO:0016747 transferase activity, transferring acyl group
Cluster-61376.0 GO:0003677 DNA binding molecular_function 2.06E-23
```

4.3.10.9 信息查询

4.3.10.9.1 查询库中所包含的物种

```
python GFAP-linux.py -as
```

4.3.10.9.2 查询库中所包含的家族

```
python GFAP-linux.py -af
```

4.3.10.10 合并注释结果

GFAP 直接产生的注释结果所采用的格式是有利于下游直接分析的（比如可视化）。但，在最终呈现时需要对结果进行进一步的归纳整理。因此，开发了合并功能。

`python GFAP-linux.py -mr -qp/qn` (序列文件，以便于为整理结果提供所有 ID) `-rp` (需要将待归纳结果放入一个空文件夹并在此输入该文件夹的位置，注意移动文件的过程中不要更改文件名字) `-o` (结果文件的保存路径)。例如，

```
python GFAP-linux.py -mr -qp example_protein.txt -rp ./1/ -o annotation_result.txt
```

5、使用的注意事项

5.1 请尽量避免路径中含有空格；

5.2 在进行功能注释时，不要以压缩文件作为输入文件；

5.3 GFAP 提供了对 DNA 序列的注释功能，这主要是考虑用户很难通过二代转录组获得蛋白序列。尽管如此，我们仍然鼓励用户使用蛋白序列进行注释。并且如果不得不对 DNA

序列进行注释，请注意，GFAP 只能使用近缘物种库对序列进行注释，而 HMM 库仅仅只适用于蛋白序列的注释。

5.4 在使用热图对统计结果进行绘制的时候，将结果保存成 pdf 的时候，有点儿技术上的难题没有克服，因此，这部分结果目前只能保存成 svg 格式。这个缺陷将在后续版本中加以改善。

5.5 GFAP 的测试文件以及使用视频可以在 ftp 中下载。

5.6 当用户使用 DNA 作为输入文件进行功能注释时，结果中“-r”表示这条序列的反向互补序列能够比对到库，具有注释的结果。

5.7 功能测试是在 Windows（64 位，win10 系统，8G 内存）和 MacOS（64 位，8G 内存，M1)机器上进行的。

5.8 Linux 版本不设可视化功能。

1 Introduction

GFAP (Gene Functional Annotation for Plants) is a program for functional annotation of plant genes. The program includes two parts: database and GFAP software. At present, the GFAP database includes 208 species of plants from 85 families, which cover almost all plant taxa including algae, mosses, lichens, ferns, gymnosperms, dicots and monocots to provide users with sufficient information to perform GO (Gene Ontology), KEGG (Kyoto Encyclopedia of Genes and Genomes), and protein domain annotation for unknown sequences. Considering that gene family and ncRNA (non-coding RNA, GFAP mainly annotates long non-coding RNA and microRNA) information also plays an important role in plant research, therefore, in the GFAP database, we summarized the current main gene families and ncRNAs. The related Hidden Markov Models (HMMs) were collected and constructed for family information and ncRNA annotation of unknown sequences. We annotated the library of closely-related species through Pfam and the HMM model on the KEGG official website. After the annotation was completed, we collected the HMM models involved in annotating these plant genes to construct a plant-specific HMM library. Therefore, in addition to the use of closely-related species information to annotate sequences, GFAP also supports the use of HMM models for functional annotation. Furthermore, GFAP will also provide general library information for user annotation of unknown sequences. From this perspective, GFAP is suitable for gene function annotation of all organisms. The difference in efficiency is an important distinguishing point between using closely-related species information and using HMM models for annotation. In the testing phase, GFAP can perform GO, KEGG and protein domain annotations on 1 Mb files (2627 genes) using the related species library within 5s, which is much more efficient than using the HMM models. Of course, the choice of how to annotate is largely up to the user. GFAP software implements function annotation by clicking, dragging, etc. instead of command line input. This will facilitate functional annotation analysis for wet biologists. In addition to the annotation function, GFAP software also provides data statistics and visualization functions (including histograms, heat maps, bubble charts, and network graphs) to help users analyze data more conveniently.

2. Installation system

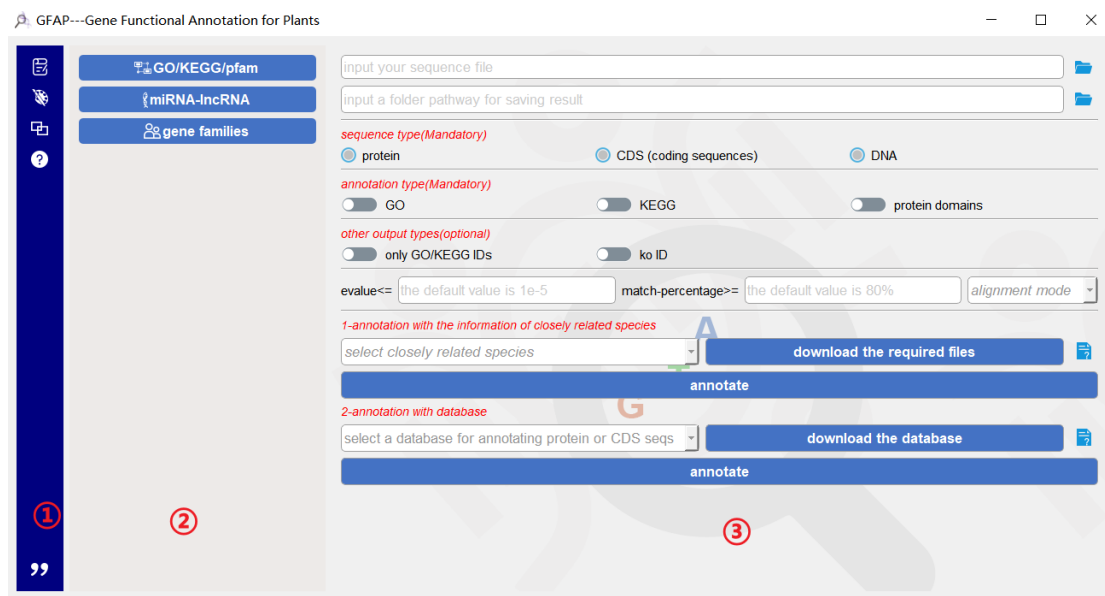
GFAP can be installed on Windows and MacOS systems

3. Installation

Double-click the exe file directly on Windows to install. In MacOS system, double-click the pkg file to install. After installation, the program will be placed in the "Application" folder by default.

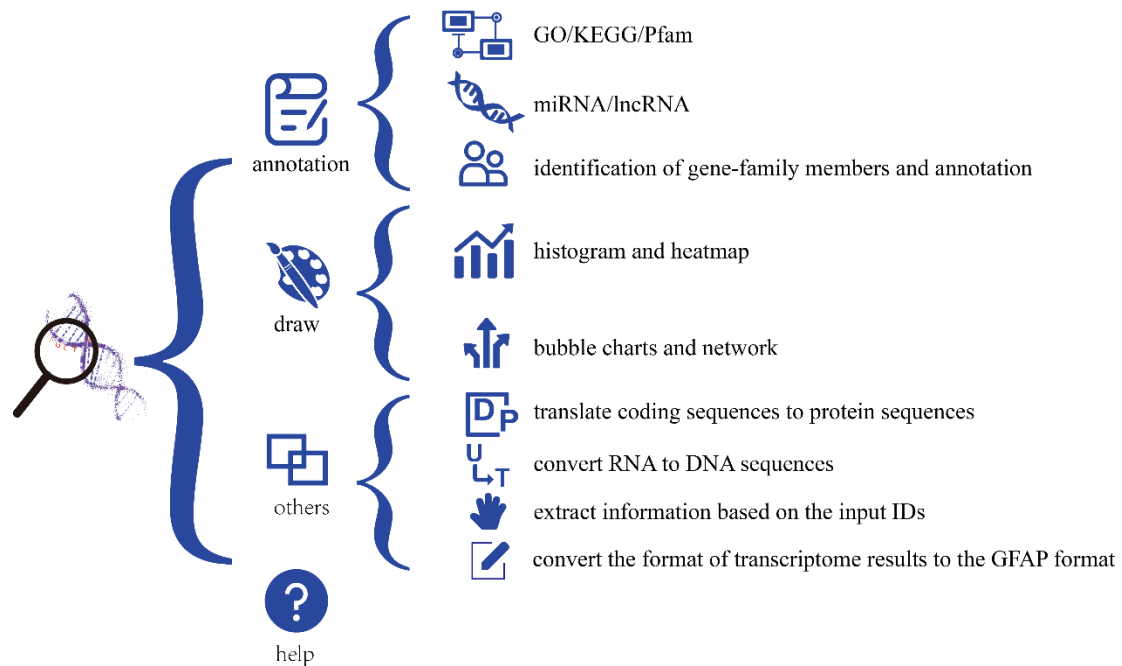
4. Specific information on the use of GFAP software

4.1 Introduction of software interface



The first block is the module block (①). There are "Annotations" (📄), "Drawings" (🔬), "Other Functions" (🔲) and "Help" (❓) modules. Under each module, different "option" modules (②) are set according to different function types. The third module is the "Function" module (③) under the different option modules. The distribution of "Options" and "Function" is

as follows:



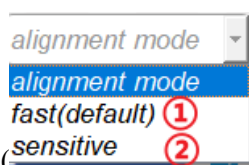
4.2 Software function introduction

4.2.1 Functional annotation

GO/KEGG/Pfam

As indicated by arrows, the 'GO/KEGG/pfam' button is used for GO/KEGG/Pfam (protein domain) annotation. Enter the sequence file (fasta format) into ①; ② fill in the save path. You only need to fill in the folder name here, and then all the annotation results will be stored in this folder. You can drag and drop files and folders directly into the specified location (①, ②). Of

course, You can also select the path of the file or folder by clicking the folder icon (📁) behind it; after that, you need to select the type of sequence contained in the input file at ③, which is a required option; you can select the type of annotation at ④ as needed, which is required here, and multiple selections are possible at the same time; considering that GO, KEGG and ko ID are also commonly used input information for other analysis websites or software, after selecting the option at ⑤, it will be generated at the same time as the annotation result. The file containing only ID; GFAP uses the alignment program Diamond (<https://github.com/bbuchfink/diamond>), and sets



two alignment modes (① and ②). There are differences of the two modes in accuracy and speed. Users can select one of them, according to their own needs (this is a non-essential option, the default is fast mode); ⑥ and ⑦ are the settings for comparing and filtering the results, and the default evalule (⑥) is 1e-5, the default matching score (⑦) is 80%, if these two values can meet your requirements, no need to set; if the user wants to use a closely-related species to annotate, it is needed to select the desired species at ⑧, followed by ⑨, it will automatically download the corresponding file from our ftp website to the specified location, according to the relative species and the sequence type selected by the user. In our testing, this function is effective, but its execution may be affected by the user's network environment. In order to ensure the user's annotation process, The *Arabidopsis* file is carried with the software (that is, it does not need to be downloaded from the network). If the user clicks the "annotate" button and GFAP does not detect the corresponding file, the carried *Arabidopsis* file will be used to finish the annotation process. If unsatisfied with the annotation result, users can directly go to the ftp website (<ftp://ftp.agis.org.cn/~panweihua/GFAP/>) to download the data and put it into the corresponding folder, and then click the button to annotate genes, the required files and the location where the files need to be placed in GFAP after downloading are shown in the following table:

Sequence type	ftp	GFAP location
Protein/ CDS (coding sequences)	ftp://ftp.agis.org.cn/~panweihua/GFAP/protein-alignment/GFAP/protein-alignment (in MacOS, GFAP->show the content of package->Contents->Resources-

		>protein-alignment. Similar below, not in further)
DNA (non-coding sequences)	ftp://ftp.agis.org.cn/~panweihua/GFAP/DNA- alignment/GFAP/DNA- alignment

If the user wants to use the HMM files to annotate the sequences, the first thing to pay attention to is the sequence type, because the HMM files currently contained is constructed by the protein domain, therefore, this function can only accept protein or CDS files and if the input is a CDS file, and GFAP will translate the CDS file first, and then annotate it. Users can also select the



translation function () in "Other" to translate before annotation. When annotating, you still need to click "download the database" to download the required files first. This process is automatic, but it may still be affected by network conditions. If there is a problem with the download (no database detected after clicking the annotate button), users can choose to download by themselves. The files to be downloaded and the locations to be placed are shown in the following table:

Annotation type	ftp	GFAP location
KEGG	ftp://ftp.agis.org.cn/~panweihua/GFAP/database/akeg g.txt.gz/GFAP/database/akegg.txt.g z

miRNA-lncRNA

The annotation of ncRNA (miRNA and lncRNA) includes the following steps: put the fasta-format sequence file into ①; decide whether to perform miRNA annotation or lncRNA annotation on the unknown sequence through the selection of ②; put the evaluate value (the default value is 1e-5) at ③ to filter the results; then set the save location, name the save file and click "annotate" for performing the annotation process.

gene families

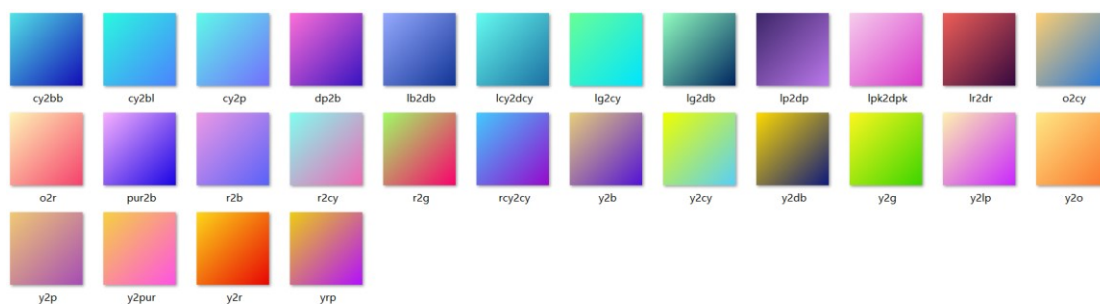
Annotation for family members consists of two-part functions. The first part is to find members of a specific family and extract the sequences. Here are the specific steps: put the protein sequences into ①; select the desired family HMM model from ②; set the save location and name it at ④; click ⑥; then the family member ID will be displayed in ⑦; after that, click ⑨ to directly extract the corresponding sequence from the input file. The second part is to count which families all the sequences in the input file are distributed in. Here are the specific steps: put the sequence in ①; set the save location and name it at ④; select the type to be annotated at ⑤; then click ⑧, the statistical result you need is entered and displayed at ⑦.


4.2.2 Statistics and graphs

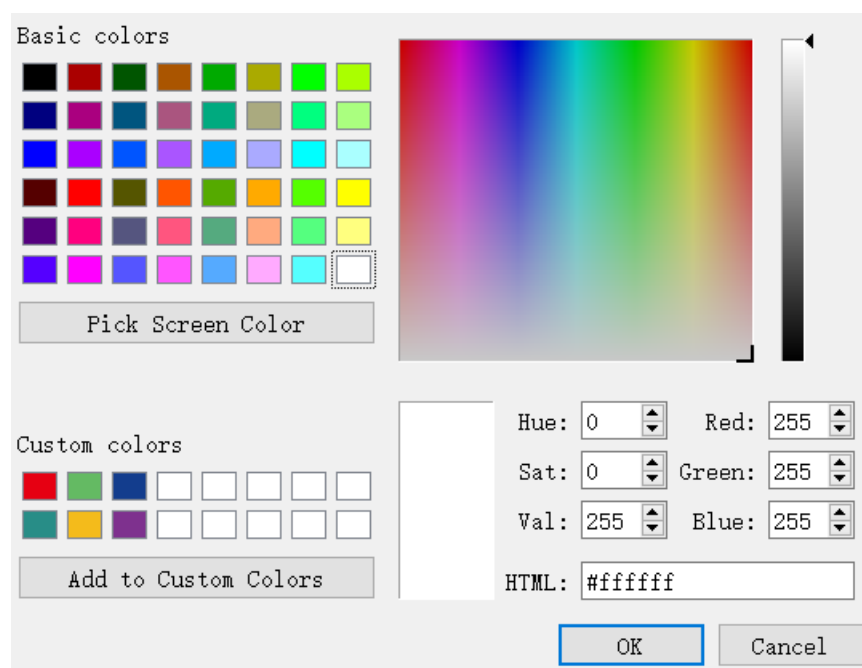
Statistics

In addition to the annotation function, GFAP can also perform statistics and visualization of its own annotation results. The functions shown above are mainly used to draw histograms and heat maps. Put the annotation results in ①, GFAP will count each function in the annotation results and calculate the number of genes with this function, and set two parameters to filter the statistical results; as shown in ②, the number of genes is filtered (The default value is 0, that is, the number of genes is not filtered), for example, if you fill in 10 here, it means to display the functions with the number of genes exceeding 10; the function at ③ is used to display the top n% functions (the

default value is 10, that is, only display the top 10 functions), for example, fill in 10 here, it means that GFAP will display the functions with the top 10% of the genes. The values that need to be filled in these two places are integers; the next is the mandatory selection (④), which requires the user to tell GFAP what the annotation results put in are about, because different types of annotation results will be displayed differently. For example, the GO annotation will be displayed in terms of biological process, cellular component and molecular function; ⑤ provides the choice of plot type, that is, the user need to select one from a histogram or a heatmap to draw (the default is a histogram) (the default is a histogram); at ⑥, 28 color schemes are arranged for users to choose according to their own preferences; the following are 28 color schemes and codes:



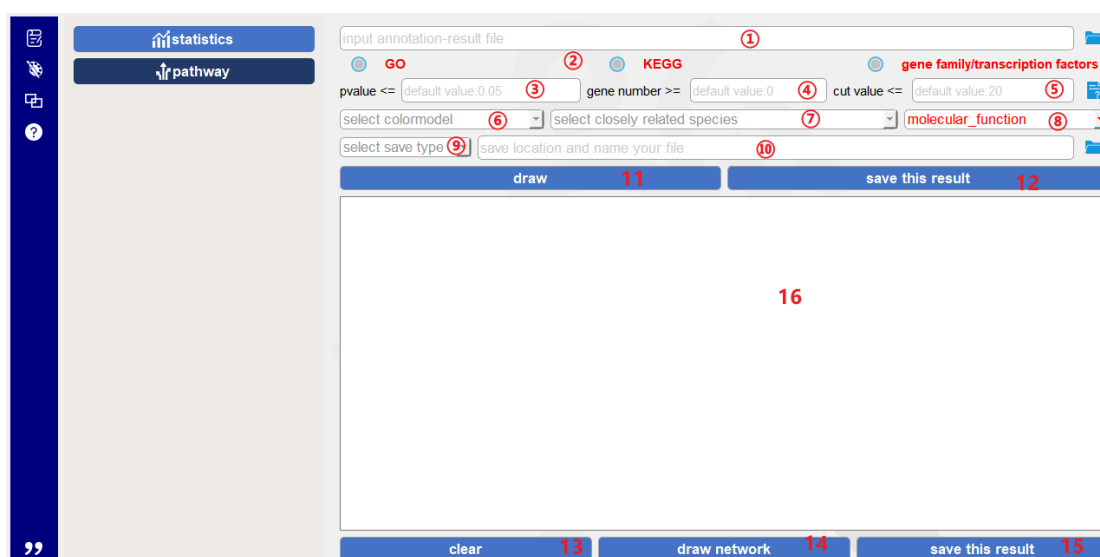
At the same time, considering that these color schemes may still not meet the needs of users, the latter color picker () allows users to select the colors they are interested in. The interface after opening is as follows:



"pick screen color" can be used to extract any colors from the computer screen, and "add to

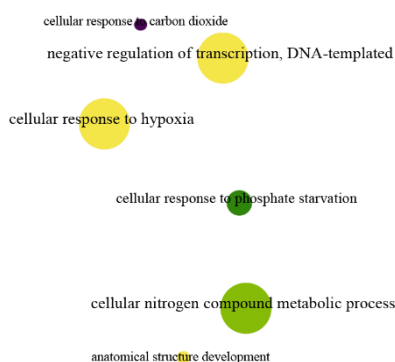
custom colors" can be used to add favorite colors. After selecting a color, click OK, and the selected color will be automatically added to the color dialog ⑨. These are settings for multicolor. If you only want to use one color to represent the annotation result, you can check ⑧, and then fill in the color in the above color dialog box, that is, the result can be displayed in this color. In GFAP, we set the save format to SVG or PDF format (the default save is SVG format), because this vector graphics can ensure the clarity of the picture to the greatest extent. We recommend Adobe Illustrator to further modify and integrate the generated images. After making the above settings, click ⑫ to draw. If the drawing result meets the requirements, you can set the save location at ⑪, and then click ⑬ to save the picture.

Pathway

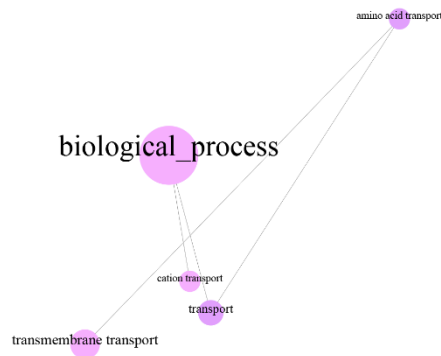


This part is mainly used for drawing bubble charts and network diagrams.

bubble charts:



network diagrams:



The basic parameters are set and operated as before. The difference is that here we use the t test to calculate the pvalue value. The premise of calculating this value is to select a closely related species as the background data set ⑦. After entering the file and setting the corresponding parameters, click ⑪ to complete the bubble chart, and the calculation result will be displayed in

⑫. As shown below:

#					
GO:0005524	ATP binding	278	0		
GO:0020037	heme binding	36	0		
GO:0005509	calcium ion binding	33	0		
GO:0003676	nucleic acid binding	68	0		
GO:0043565	sequence-specific DNA binding	64	0		
GO:0003677	DNA binding	92	0		
GO:0005515	protein binding	237	0		
GO:0008270	zinc ion binding	150	0		
GO:0003723	RNA binding	95	0		
GO:0046872	metal ion binding	55	0		
#					
GO:0005524	ATP binding	278	0		
GO:0022857	transmembrane transporter activity			61	0
GO:0003677	DNA binding	92	0		
GO:0016791	phosphatase activity	70	0		

A set of data with interrelationships is classified by the # key. If the user needs to display the functions with mutual relationships in the bubble chart (that is, want to draw a network diagram), click ⑭ to complete the drawing on this basis. The data source of the function ⑭ is the content displayed in ⑫, so before drawing the network diagram, the user can also select the content to be displayed by editing the content in ⑫.

4.2.3 Other functions

Translation

GFAP supports annotation of DNA and protein sequences. Nonetheless, we still hope that users will utilize protein sequences whenever possible to complete the annotation process. Take into account that some users may only have CDS sequence files. Therefore, we set up the translation function, which can batch convert the gene sequences in the CDS file into protein files for later annotation. The translation process is to put the CDS file in ①; set the save location and name it at ②; click ③ to execute the function.

RNA2DNA

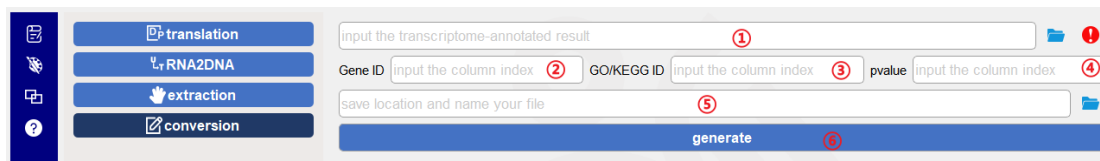
When annotating ncRNA sequences, what is identified is the DNA sequence. Therefore, a function that can convert RNA into DNA sequences is provided. Arrange the RNA sequence into fasta format and put it in ①; enter the save location and name the file ②; click ③ to complete the conversion.

extraction

GFAP supports the gene annotation process at the genome level. However, in the actual operation process, we often only pay attention to some genes. At this time, we need to extract the concerned genes from the overall file. In GFAP, the input file of the extraction function we designed is the annotation result ①; the input ID ② is required (it can be a single ID. If you want to extract information of many genes, these IDs of the genes should be input into one file, and then used as the second input file. One ID, one line.); and the ID here can be the gene ID or GO/KEGG/Pfam

ID (③); save and name (④); click ⑤ to execute the function.

conversion



The purpose of setting the conversion function is to convert the results of transcriptome analysis into a format that GFAP can recognize so that GFAP can be used for corresponding analysis. Put the results of the transcriptome into ①, the format of the transcriptome result is txt and the tab key should be used as the separator between columns. Next, GFAP needs to be informed of the following information: 1. Which column of the input file the gene ID is located in ②; 2. Which column is the GO/KEGG ID located in ③; 3. Which column is the pvalue located in ④. After entering these values, save and name them ⑤, and then click ⑥ to execute the function.

4.3 Quick start

All parameters with default values described in this section remain the same as default parameters. In actual operation, users need to set them according to their own needs. Generally speaking, the default values can meet the needs of most users.

4.3.1 GO/KEGG/Pfam

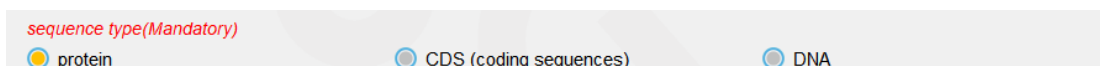
1. Insert the protein file



2. Set the working directory



3. Select the input sequence type



4. Select the desired annotation type



When using closely related species annotations:

5. Select the related species and click download. After the download is complete, click "annotate" to complete the annotation.



1-annotation with the information of closely related species

Aquilegia_coerulea(Ranunculaceae)

download the required files

annotate

When annotating with an HMM file, click the download button to download the file, then click "annotate" to wait for the annotation to complete



2-annotation with database

plant-special database

download the database

annotate

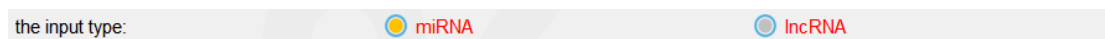
4.3.2 miRNA-IncRNA

1. Input fasta file



E:\example\test\miRNA.txt

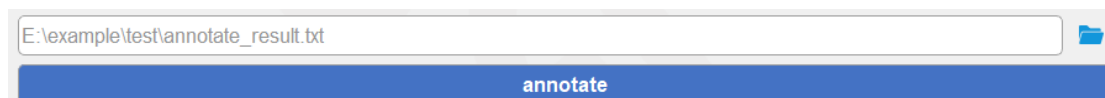
2. Select the sequence type



the input type:

miRNA IncRNA

3. Set the save location and name it, click "annotate" to complete the annotation



E:\example\test\annotate_result.txt

annotate

4.3.3 Gene Family Annotation

1. Input protein file



E:\try\Populus_trichocarpa.Pop_tri_v3.pep.all.fa

If annotating a single family:

2. Select the required HMM model



ARF (Auxin response factors)

3. Save and name



E:\try\ARF.txt

4. Click the button to execute the function

show members of a single family

5. Click the button to automatically extract the sequence

extract sequences

If the count input file contains all families:

2. Save and name

E:\try\ARF.txt

3. Select the type of analysis to be analyzed

Please select one of the following choices if you want to identify the members of all families

☒ transcription factor

☐ gene family

4. Click the button to complete the function

show genes containing domains of families

4.3.4 Statistical plotting

1. Enter the annotation result file

E:\try\result\GO_database_result.txt

2. Indicate which annotation type it is

☒ GO

☐ KEGG

☐ gene family/transcription factors

3. Click "draw"

draw

4. Select the save type, enter the save location and name the file

svg

E:\try\result\draw_statistics.svg

5. Click Save

save this result

4.3.5 Bubble Chart and Network Diagram

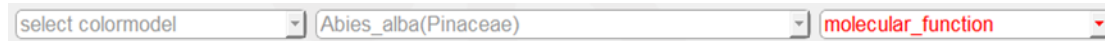
1. Enter the result file of the annotation

E:\try\result\GO_database_result.txt

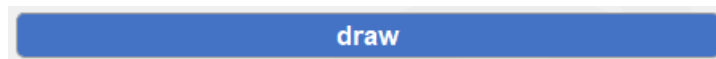
2. Select the annotation type



3. Select close relatives to calculate pvalue. If you want to visualize the GO annotation results, you need to select one of the three types of ontology for visualization



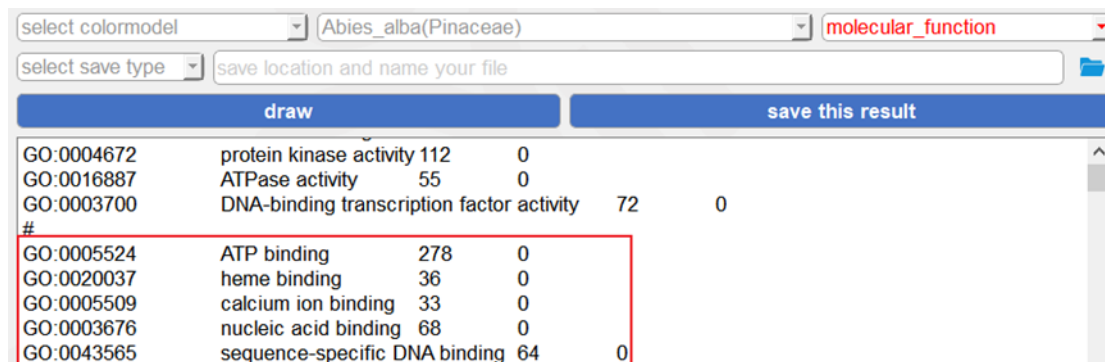
4. Click "draw"



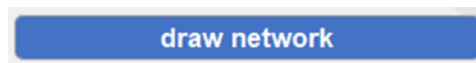
5. When saving, select the save type, enter the save location and name it



6. Editing statistics



7. Click "draw network" to draw a network diagram



8. When saving, select the save type, enter the save location and name it



4.3.6 Translation

1. Input the CDS file in fasta format



2. Select the save location and name it



3. Executive function

translate

The format of the fasta-format file is as follows:

```
>Potri.T155100.3 pacid=37219601 polypeptide=Potri.T155100.3.p locus=Potri.T155100 ID=Potri.T155100.3.v3.1 annot-version=v3.1
ATGGCTATATCGAAGCTTTTGATTGTTTTCTTGTCGCATCTCCTTGCTCCGCCCTTGTTGAAGCTGATCAGAAGGT
GGTGAACCTCAAATATTCAAGCTGCTAGCTATCCTCTGGGAAGAATATCGCAGATTGTGGTGGCGCTTGCAATGCTAGGT
GTTCTTATCTCCAGGCCACGCTCTTGAAGAGGGCTTGCGGGACTTGCTGTGCACGATGCAAGTGTGCCCCCAGGC
ACTTCGGCAACATTATACCTGCCCTTGCTATGCCACCATGACTACTCGAGGTGGCCGACTCAAGTGCTCTTGA
>Potri.T155100.1 pacid=37219602 polypeptide=Potri.T155100.1.p locus=Potri.T155100 ID=Potri.T155100.1.v3.1 annot-version=v3.1
ATGGCTATATCGAAGCTTTTGATTGTTTTCTTGTCGCATCTCCTTGCTCCGCCCTTGTTGAAGCTGATCAGAAGGT
GGTGAACCTCAAATATTCAAGCTGCTAGCTATCCTCTGGGAAGAATATCGATTGTGGTGGCGCTTGCAATGCTAGGTGT
CCTTATCTCCAGGCCACGCTCTTGAAGAGGGCTTGCGGGACTTGCTGTGCACGATGCAAGTGTGCCCCCAGGCCT
TCCGGCAACATTATACCTGCCCTTGCTATGCCACCATGACTACTCGAGGTGGCCGACTCAAGTGCTCTTGA
>Potri.T155200.1 pacid=37219603 polypeptide=Potri.T155200.1.p locus=Potri.T155200 ID=Potri.T155200.1.v3.1 annot-version=v3.1
ATGAAGGCAACGAGGAGCCGATACAGCGGTGGCGAGATGGGTAAAGGAGCAACCAAGATGAAGGCTTTTTAGC
GGTGGTTTCAGGGTTGGCGGCTTGGTTTTCTTAGAATGGTTGTCGTGATCATGATAACCTCTTGTGTGCTGAGG
TTGTTCAATCTATTGGAATCTCTGTTCTTATTACAAGCTCATGAAGGAGAAGACTTGTGCTGGACTTCACTAAATCA
CAGGAGCTAATAGCTATATTTTAGCTGTAGGCTCTATTGCAGTTTTGTGCATGGAGTATGACATACACACTACTTGA
TTCAGCTAGTTGCTGACAACCTTTGGGTAACTATACGATCCGTTCAACTTGAGGTCAGTTACATGGAAGGCAAG
ATAAATTTGCAATTTACTATTGGTGATACATGTCTGTGCTAGCTTTGTATATTCATCCAAGAACACATCACCACATA
GTCAACAGGATTGTGTTGGGCTTTTGTGTTTACCTTGAATCTATTTCAAGTGTGCTCAGCTGCGAGTAATGCAAAACAC
AAAGATTGTTGAACCTTACGGCACATTATGATTTGCGCTTGGAGTTGCCAGGTTCTGGGTTGTGCACATTGGATCC
TCCAGGTACTGGACACTCGAGGACGCTATTGACAGCACTGGGCTATGGAATGTGGCTCTTATGGTCTACTTTTCAAGAA
ATTGTGCAGACATTCATTCTGGCCGATTTTGTCTACTACTACGTCAAGAGTGTCTTGGTGGCAACTGTGTTCTAAGGT
CCCCCAGGAGTGGGTGTA
```

4.3.7 RNA2DNA

1. Input the RNA sequence file in fasta format

E:\example\test\miRNA.txt

2. Select the save location and name it

E:\example\test\DNA2RNA.txt

3. Click the button to execute the function

transform

"fasta-format RNA-seq file" format:

```
1 >miR156
2 UGACAGAAGAGAGUGAGCAC
3 >miR164
4 UGGAGAAGCAGGGCACGUGCA
5 >miR166
6 UCGGACCAGGCUUCAUUC CCC
7 >miR168
8 UCGCUUGGUGCAGAUCCGGGAC
```

4.3.8 Extraction

1. Enter the comment result

E:\try\result\GO_database_result.txt

2. When putting a single ID:

GO:0050660

Or you can put a file with multiple IDs

E:\try\result\IDs.txt

The format is:

```
GO:0009058
GO:0008236
GO:0004842
GO:0006801
GO:0016020
GO:0050660
GO:0020037
GO:0020037
GO:0016788
GO:0055114
GO:0055114
GO:0009165
GO:0055114
GO:0009733
GO:0035145
```

3. Select the type of ID to enter

☒ gene ID ☐ GO/KEGG/Pfam ID

4. Set the save location and name it

E:\try\result\GO_extraction.txt

5. Click the button to extract the information

extraction

4.3.9 conversion

1. Put the transcriptome results

E:\try\transcriptome_example.txt

2. Put the column ID where the keyword is located

Gene ID GO/KEGG ID pvalue

3. Put it in the save location and name it

E:\try\conversion_result.txt

4. Click the button to generate the corresponding file

generate

The format of transcriptome results:

gene_id	HLA_readcount	HLO_readcount	log2foldChange	pval	padj	Gene Length	NR GI	NR ID	NR Score	NR Evalue	NR Description	NT GI	NT ID	NT Score
cluster-36310.157224	43.08993596	16404.86177	-8.5675	1.08E-105	3.84E-100	1406	--	--	--	--	K00799	--	--	--
cluster-36310.62757	2418.036399	66.53141246	5.1824	1.40E-100	2.48E-95	325	566189800	XP_006378352.1	343	1.00E-30	hypothetical protein POPTR_0010s08600g [Popul	--	--	--
cluster-36310.157767	4.878306409	2279.082823	-8.8766	8.15E-91	9.63E-86	1384	--	--	--	--	--	--	--	--
cluster-41605.0	3940.35515	41.90246648	6.5543	3.85E-89	3.42E-84	1436	224747075	ACH62215.1	147	2.50E-07	*omega-gliadin, partial [Triticum aestivum x Loph	--	--	--
cluster-36310.62762	2818.833033	60.9114402	5.5307	7.97E-79	5.65E-74	328	566254885	XP_006387619.1	249	8.30E-20	hypothetical protein POPTR_0770s00220g [Popul	--	--	--

4.3.10 For Linux system users

You need to install Python 3 on your Linux system and add the version to the command environment

After extracting the files, you need to navigate to the GFAP folder using the following command:

```
cd */GFAP
```

4.3.10.1 GO/KEGG/Pfam/nr/swissprot annotation

4.3.10.1.1 annotation with the information of closely related species

python GFAP-linux.py -go/kegg/pfam -qp/qn (protein/coding-sequence fasta-format file) **-aws** (Latin scientific name of a species) **-o** (save pathway, No need to name the file, just provide the path). For example:

```
python GFAP-linux.py -qp example_protein.txt -aws Rosa_chinensis -go -o ./
```

You can also specify the number of CPUs and different modes (fast/sensitive, The “fast” mode achieves extremely high alignment efficiency while ensuring a certain level of accuracy. The “sensitive” mode is more sensitive to alignment results, but it has a relatively lower efficiency).

4.3.10.1.2 Annotation with database

python GFAP-linux.py -go/kegg/pfam -qp/qn (protein/coding-sequence fasta-format file) **-awd** (database name) **-o** (save pathway, No need to name the file, just provide the path). For example:


```
python GFAP-linux.py -qp example_protein.txt -awd psd -pfam -o ./
```

4.3.10.2 Identification of special family members

python GFAP-linux.py -sf -qp (protein fasta-format file) **-mn/mp** (Enter the abbreviation of the family in the database. If the required family is not available in the database, you can download the HMM file yourself and enter the HMM file location.) **-o** (save pathway and name your file).

For example:

```
python GFAP-linux.py -sf -qp ./example_protein.txt -mn WRKY -o gene_family.txt
```

```
python GFAP-linux.py -sf -qp ./example_protein.txt -mp ./WRKY.hmm -o gene_family.txt
```

In addition to generating a file containing the IDs of the family members, this command will also create associated FASTA files for these IDs.

4.3.10.3 Identification all genes containing protein domains in the GFAP database

python GFAP-linux.py -mf -atf/agf (transcription factor/non-transcription factor) **-qp** (protein fasta-format file) **-o** (save pathway and name your file)。例如：

```
python GFAP-linux.py -mf -agf -qp ./example_protein.txt -o gene_family.txt
```

```
python GFAP-linux.py -mf -atf -qp ./example_protein.txt -o gene_family.txt
```

4.3.10.4 Determine whether the input ncRNA sequence is a known ncRNA

python GFAP-linux.py -na -nt (miRNA/lncRNA) **-qn** (fasta-format file) **-o** (save pathway and name your file). For example:

```
python GFAP-linux.py -na -nt miRNA -qn ./ncRNA.txt -o ncRNA_result.txt
```

4.3.10.5 Translation

python GFAP-linux.py -t -qn (coding sequences, fasta format) **-o** (save pathway and name your file). For example:

```
python GFAP-linux.py -t -qn TAIR10_cds.fa -o protein.txt
```

4.3.10.6 Convert RNA to DNA

python GFAP-linux.py -rd -qn (fasta-format file) **-o** (save pathway and name your file). For example:

```
python GFAP-linux.py -rd -qn RNA.txt -o DNA.txt
```

4.3.10.7 Extract the content from the annotation file based on the input ID or a file containing

multiple IDs (one ID per line)

python GFAP-linux.py -ex -ar (annotation file) **-ID** (ID or ID file) **-exfid/exgid** (functional ID (for example, GO:2123461) or gene ID) **-o** (save pathway and name your file). For example:

```
python GFAP-linux.py -ex -ar GFAP-Arabidopsis_thaliana\Brassicaceae\kegg_annotate.txt -ID ./gene_IDs.txt -exgid -o result.txt
```

```
python GFAP-linux.py -ex -ar GFAP-Arabidopsis_thaliana\Brassicaceae\kegg_annotate.txt -ID function_IDs.txt -exfid -o result.txt
```

4.3.10.8 Format conversion for different omics files

python GFAP-linux.py -cf -gf (omics file) **-gid** (the index of gene ID) **-fid** (the index of go/pfam/kegg ID) **-pvalue** (the index of pvalue) **-o** (save pathway and name your file). For example:

```
python GFAP-linux.py -cf -gf transcriptome_example.tab -gid 1 -fid 30 -pvalue 5 -o result.txt
```

The format of omics file:

BP Description	Gene Ontology Molecular Function
--	--
oxidation-reduction	GO:0005506//GO:0020037//GO:0016705
glycerolipid metab	GO:0004144
photosynthesis	--
response to wound	GO:0004867
--	--
gluconeogenesis	GO:0005525//GO:0004611
--	GO:0005515
blood coagulation	GO:0008191//GO:0034235
--	GO:0005524
--	--
oxidation-reduction	GO:0016491
response to wound	GO:0004867

The results of the format conversion are as follows.

```
Cluster-36310.161598 GO:0016747 transferase activity, transferring acyl group
Cluster-36310.156351 GO:0005509 calcium ion binding molecular_function 8.44E
Cluster-36310.156351 GO:0008266 poly(U) RNA binding molecular_function 8.44E
Cluster-36310.156351 GO:0010242 oxygen evolving activity molecular_functio
Cluster-36310.251474 GO:0008121 ubiquinol-cytochrome-c reductase activity m
Cluster-36310.251474 GO:0016849 phosphorus-oxygen lyase activity molecular
Cluster-36310.251474 GO:0005509 calcium ion binding molecular_function 1.17E
Cluster-36310.86742 GO:0016651 oxidoreductase activity, acting on NAD(P)H molec
Cluster-36310.163845 GO:0003723 RNA binding molecular_function 2.55E-24
Cluster-36310.183781 GO:0005515 protein binding molecular_function 2.53E-24
Cluster-36310.183781 GO:0016820 ATPase-coupled transmembrane transporter acti
Cluster-36310.38952 GO:0005515 protein binding molecular_function 4.23E-24
Cluster-36310.143554 GO:0008168 methyltransferase activity molecular_functio
Cluster-36310.160747 GO:0016491 oxidoreductase activity molecular_function 7
Cluster-36310.225771 GO:0004866 endopeptidase inhibitor activity molecular
Cluster-36310.161571 GO:0005509 calcium ion binding molecular_function 7.93E
Cluster-36310.163698 GO:0016747 transferase activity, transferring acyl group
Cluster-61376.0 GO:0003677 DNA binding molecular_function 2.06E-23
```

4.3.10.9 Information query

4.3.10.9.1 Query the species included in the database

```
python GFAP-linux.py -as
```

4.3.10.9.2 Query the families included in the database

```
python GFAP-linux.py -af
```

4.3.10.10 Merge the annotation results

python GFAP-linux.py -mr -qp/qn (protein or CDS sequence file for providing gene IDs) **-rp** (folder name containing only annotation results. Note, when removing annotation results, please do not change the name of annotation results) **-o** (save pathway and name your file). For example,

```
python GFAP-linux.py -mr -qp example_protein.txt -rp ./1/ -o annotation_result.txt
```

5. Precautions for use

5.1 Please try to avoid spaces in the path.

5.2 When making functional annotations, do not use compressed files as input files.

5.3 GFAP provides an annotation function for DNA sequences, which is mainly because it is difficult for users to obtain protein sequences through the second-generation transcriptome. Nonetheless, we encourage users to annotate with protein sequences. And if you have to annotate DNA sequences, please note that GFAP can only annotate sequences using the related species library, while the HMM library is only suitable for protein sequence annotation.

5.4 When drawing the statistics results, the heatmap legend is hard to transform svg format to pdf format. Therefore, it can only be saved as a svg file. These functions will be improved in subsequent versions.

5.5 The testing data and tutorial videos can also be downloaded from our ftp website.

5.6 When a DNA-sequences file was used as the input file, “-r” in a gene ID means that the reverse complementary sequences of this ID can be aligned to the genes in database, and this alignment was with high confidence.

5.7. The functional tests were performed on Windows (64-bit system, win10, 8 Gb random access

memory) and MacOS (64-bit system, 8 Gb random access memory).