

# “实用生物信息技术”课程小组讨论总结报告

组：G3 次：R5 组长：陈伟辉 执笔：陶益可

一、时间 2026 年 5 月 13 日星期三

二、方式 面对面交流

三、主题 课后复习与组内互助

四、内容

## (1) TBtools 相关内容简介

① 相关文献搜索

## (2) TBtools 数据库使用练习

① 基因家族验证

② HMM 初筛 + SwissProt 验证

③ 系统发生树

④ 启动子顺式作用元件

⑤ 简单构建系统发生树

## (3) 问题

### (1) TBtools 相关内容简介

#### ① 相关文献搜索

其主要特点是：(1) 用户友好的操作界面和高兼容性（跨平台，兼容不同操作系统）；(2) 统一的功能使用逻辑，即 IOS(Input-Output-Start, 输入-输出-启动)；(3) 全面的多功能集成，覆盖了绝大多数日常生信数据解读需要；(4) 独立自主研发的绘图引擎；(5) 强大的数据交互功能，可以实现点击实时响应，悬停菜单，元件编辑等功能；(6) 庞大的用户和开发者互助社群。

### (2) TBtools 数据库使用练习

#### ① 基因家族验证

##### A、基于序列的筛选

已知：经初步的转录组&蛋白质组数据分析筛选，获得候选基因，UniProt 蛋白质数据库查询该蛋白质序列包含果胶裂解酶（Pectate lyase, 简称 PL）的 Domain;该序列为机器注释，下面将结合 TBtools



b.将上述第一轮蛋白序列作为 query，比对到下载到本地的 Swissprot 蛋白数据

库中，与（上述的操作大致相同；**注意输出文件的后缀名为 xml**。Outfmt 参数: BlastXML，其余参数不变，然后使用 Blast XML to Table 插件将上述 xml 输出文件转换为 Summary 格式的表格文件 xls。

打开此输出文件，进行人工筛选。主要看表格中的 B（属性）列和 C 列（功能），如果此 2 列不能确定，可以参考 D 列（注释）。所得筛选结果初步候选。

**InterPro** Classification of protein families

Home Search Browse Results Release notes Download Help

/ Browse / By Entry / InterPro / IPR002022 / Overview

### IPR002022 Pectate lyase

InterPro entry

Short name Pec\_lyase

Overlapping homologous superfamilies

- Pectin lyase fold/virulence factor (IPR011050)
- Pectin lyase fold (IPR012334)

**Description**

Pectate lyase 4.2.2.2 is an enzyme involved in the maceration and soft rotting of plant tissue. Pectate lyase is responsible for the eliminative cleavage of pectate, yielding oligosaccharides with 4-deoxy-alpha-D-mann-4-enuronosyl groups at their non-reducing ends. The protein is maximally expressed late in pollen development. It has been suggested that the pollen expression of pectate lyase genes might relate to a requirement for pectin degradation during pollen tube growth<sup>[1]</sup>.

The structure and the folding kinetics of one member of this family, pectate lyase C (pelC)1 from *Erwinia chrysanthemi* has been investigated in some detail<sup>[2,3]</sup>. PelC contains a parallel  $\beta$ -helix folding motif. The majority of the regular secondary structure is composed of parallel  $\beta$ -sheets (about 30%). The individual strands of the sheets are connected by unordered loops of varying length. The backbone is then formed by a large helix composed of  $\beta$ -sheets. There are two disulphide bonds in pelC and 12 proline residues. One of these prolines, Pro220, is involved in a cis peptide bond. The folding mechanism of pelC involves two slow phases that have been attributed to proline isomerization.

Contributing Member Database Entries

**Pfam** PF00544 **SMART** SM00656

Representative structure

## B. HMM 初筛 + SwissProt 验证

通过 TPIA 网站，下载了 LJ43 作为研究物种参考蛋白质组数据及 gff 文件；shuyunzaov2.0 参考基因组作为近缘数据 gff 文件；通过 Ensembl Plants 数据库，下载了拟南芥 11 参考蛋白质组植物传统模式生物数据 gff 文件；通过 Ensembl Plants 数据库下载多年生木本植物模式生物 *Populus trichocarpa* v4.1 蛋白质组数据 gff 文件，已知该候选蛋白已被 UniProt 蛋白序列数据列 Pfam 数据库(版本号) ”下载 HMM 模型，检索 PF (PL Domain 编号) 保存文本，作为结构域鉴定标准。

**Pfam** PF00544 Pectate lyase

Pfam entry

### Profile HMM Information

HMM build commands

Build method: hmmbuild HMM.ann SEED.ann  
Search method: hmmsearch --cpu 8 -E 1000 -Z 90746521 HMM pfamseq

Gathering threshold

Sequence: 24  
Domain: 24

Download [Download the raw HMM for this family](#)

在标准流程中，这两个步骤是环环相扣的。

a. **初筛（HMM 扫描）**：你下载的杨树全蛋白组作为“检索列表”，下载的 HMM 模型作为“检索条件”。一次 `hmmsearch` 就能初步获得一份候选基因名单。但这份名单是“根据模型特征的推测”，并不绝对可靠。

b. **精炼**：将 HMM 扫描获得的候选基因序列提交到 NCBI 的 CDD (Conserved Domain Database, 保守结构域数据库)，或使用 SMART (Simple Modular Architecture Research Tool, 蛋白质结构域分析工具) 等服务器进行在线分析。

c. **验证（SwissProt 比对）**：使用 **BLASTp** 工具，将上一步验证通过的候选蛋白序列，与 **SwissProt 数据库** 中的所有蛋白进行比对。这一步主要是为了给这些序列“贴上”功能标签，得到高质量的注释信息

隐马尔可夫模型(HMM)下载 pfam 数据库归入到了 InterPro 数据库之中，下载方法如下：进入数据库，检索目标基因家族 [Download - InterPro](#) 选择并点击目标基因家族—点击左下角 Profile HMM→Download; 3.打开检查 HMM 文件格式，必须把下载的 hmm 文件里 HMMER3/f 要换成 HMMER3/b（使筛选结果更完整可靠）

Gene Density Profile

Note

Gene density varies in different regions across genomes. Occasionally, users would like to obtain gene density profiles of genomes, which could be used for visualization to assist data analyses. Thus, here is a function implementation.

Set Input GFF3/GTF File **Gff 文件，与下一步骤保持一致**

Drag and Drop a GFF3/GTF File Over Here

(Optional) Chromosome Length Info. (ChrID|Length)

Drag and Drop an Input File Here.

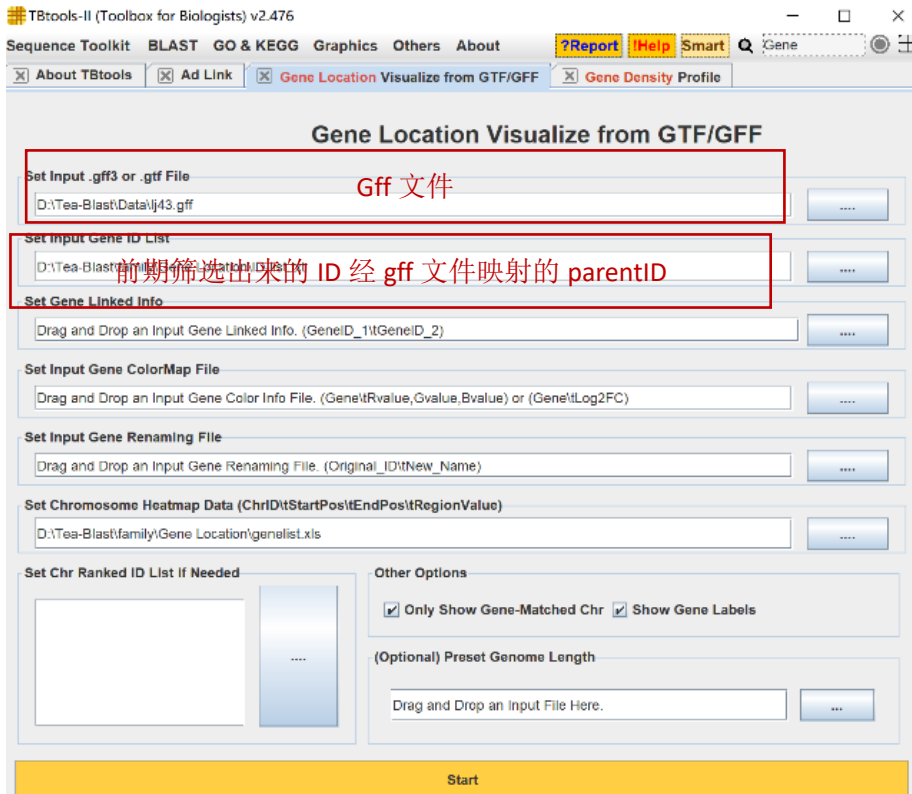
Other Parameters

Bin Size: 100000 Defined Feature Tag: Guess e.g. "exon" for non-coding RNAs

Set Output Tab-delimited File Path **输出格式为 xsl，再处理为文档**

Drag and Drop an Output Directory Here

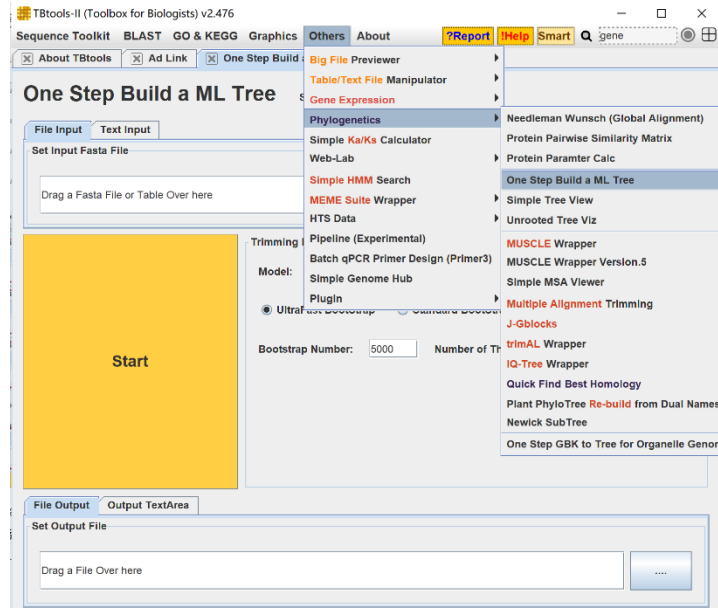
Start



将得到的所有序列汇聚在同一文件中，命名为 All Sequences.fasta，以备后用。

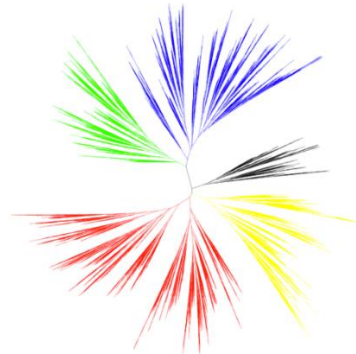
### C. 系统发生树

打开 TBtools，找到构建系统发生树的位置：Others→Phylogenetics→One Step Build a ML Tree，“Set Inport Fasta File”拖入上述 All Sequences.fasta 文件，Bootstrp Number 设置为 1000，其余参数不变。结果保留 Newick 格式，iTol 平台 ([iTOL: Interactive Tree Of Life](http://itool.ebi.ac.uk/)) 进一步美化：



## Various display modes. Support for large trees.

TOL can visualize trees with 50'000 or more leaves. With advanced search capabilities and display of unrooted, circular and regular cladograms or phylograms, exploring and navigating trees of any size is simple.



**ITOL** INTERACTIVE TREE OF LIFE [Tree of Life](#) [Upload](#) [Data sharing](#) [Help](#)

Use this page to upload and visualize a new phylogenetic tree anonymously. It should be in a plain text file and in one of supported formats (Newick, Nexus or PhyloXML). You can also use *.jplace* files generated by RaxML or pplacer, or *.qza* tree files generated by QIIME 2. Please check the [help pages](#) for detailed instructions.

Trees uploaded anonymously are deleted after 1 day. If you want to keep them private and protected, or have multiple trees to visualize, we recommend creating an [ITOL account](#). If you already have an account, please [login first](#).

Datasets and other annotation files should be dragged and dropped directly onto the interactive tree display. Please check the [help pages](#) for detailed instructions and dataset template files. Example tree and annotation files are [available for download](#).

**Upload a new tree**

Tree name:  
optional

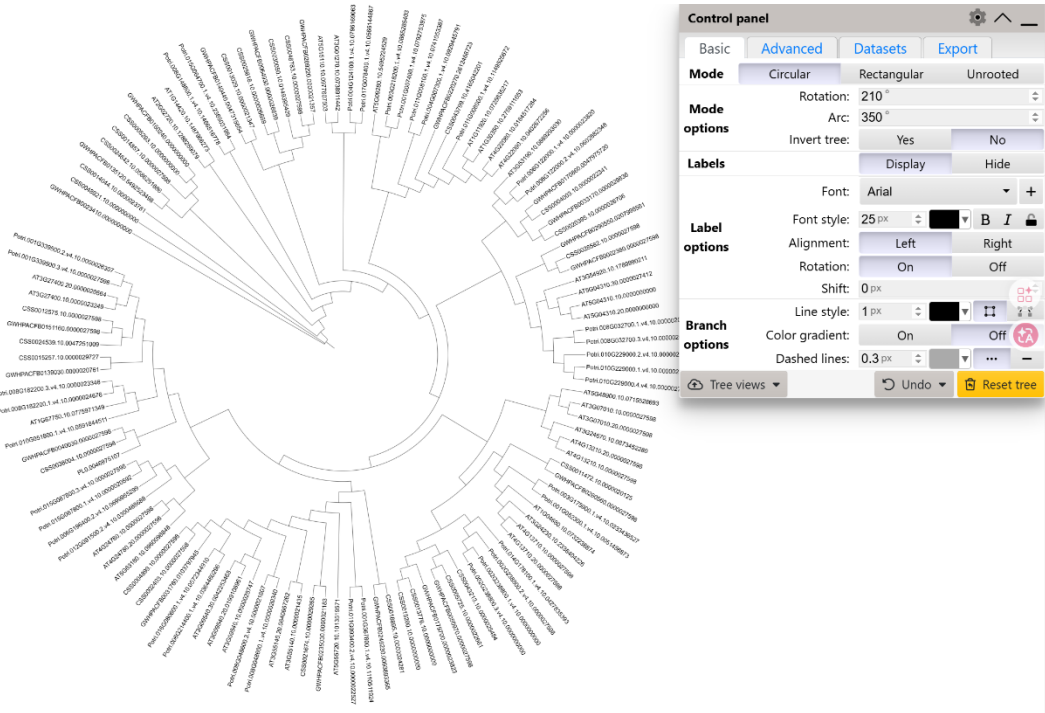
Paste your tree into the box below, or select a file using the **Tree file** selector. You can also simply drag and drop the tree file onto the page (only a regular plain text file, not QIIME QZA files).

Tree text:

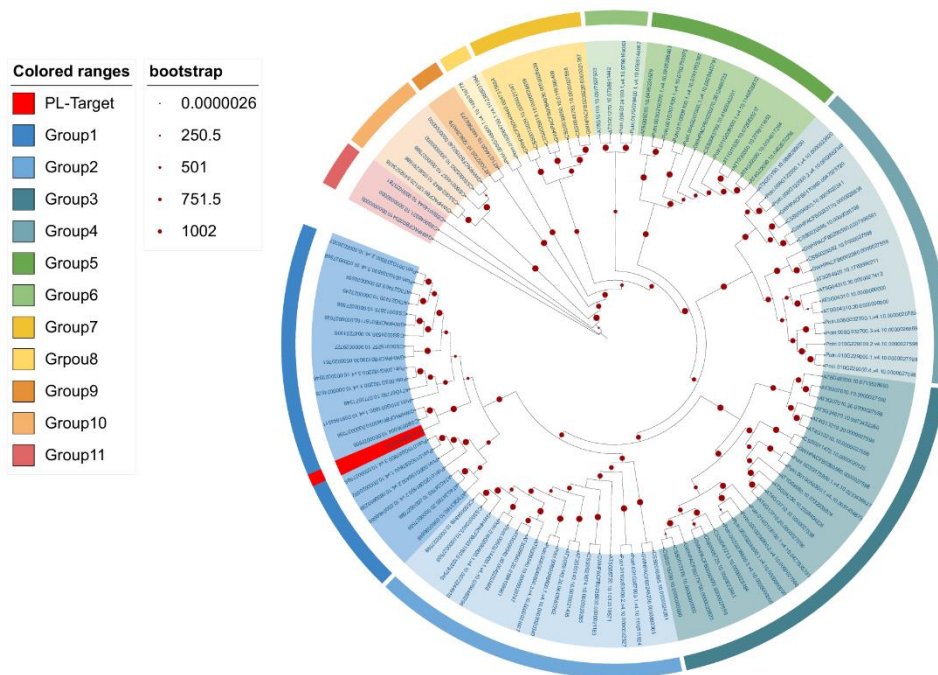
Tree file:  
 未选择文件

保存的 Newick 格式文件，或直接粘贴 TBtools 的 Newick export 复制来的代码

而后可根据个人需要进行美化，具体教程可线上搜索



美化后的结果如图所示：

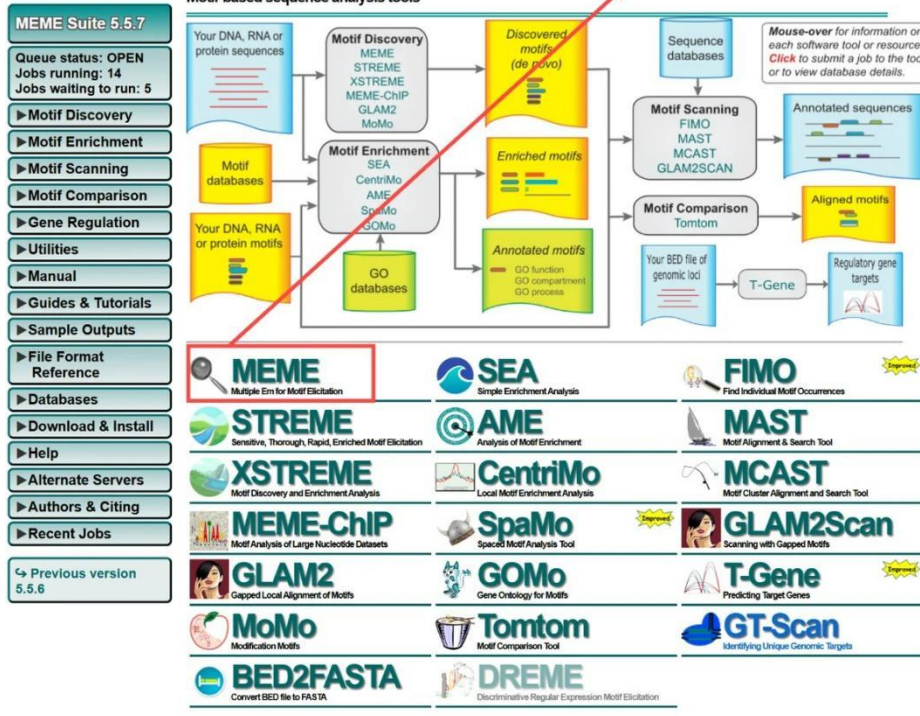


### C. Motif 分析

首先我们需要使用 The MEME Suite 在线软件(<https://meme-suite.org/meme/>)进行 motifs 分析(图 2)，点击 MEME 模块后放入已筛选出来的基因家族成员蛋白序列并设置需要的 motifs 数量，一般在 5-10 个

# The MEME Suite

点击此处，进入motif预测页面



Development of the MEME Suite was funded by grant R01 GM103544 from

**MEME Suite 5.5.7**  
Queue status: OPEN  
Jobs running: 14  
Jobs waiting to run: 9

**Data Submission Form**  
Perform motif discovery on DNA, RNA, protein or custom alphabet datasets.

**Select the motif discovery mode**  
 Classic mode    Discriminative mode    Differential Enrichment mode

**Select the sequence alphabet**  
Use sequences with a standard alphabet or specify a custom alphabet.  
 DNA, RNA or Protein    Custom

**Input the primary sequences**  
Enter sequences in which you want to find motifs.  
 **1 选择上传的文件fa格式文件**

**Select the site distribution**  
How do you expect motif sites to be distributed in sequences?

**Select the number of motifs**  
How many motifs should MEME find?  
 **2 设置需要查找的motif个数**

**Input job details**  
 **3 填写邮箱，结果会通知**

**Advanced options**

Note: if the combined form inputs exceed 80MB the job will be rejected.

**4 点击提交**

点击 start search 后需要等待一段时间，待结果出来之后我们主要使用第一个和第五个链接。第一个链接主要包含分析出来的 motifs logo 和结构域可视化信息，这里的信息都可以进行导出，

第五个链接点击右键后下载链接文件，利用 TBtools 的 MEME 可视化模块进行数据分析。

**MEME Suite 5.5.7**  
Multiple Em for Motif Elicitation

Your MEME job is complete. The results should be displayed below.

**Queue status:** OPEN  
Jobs running: 6  
Jobs waiting to run: 0

**Job Details ...**

**Results**

- MEME HTML output (1) 点击此处可查看在线的结果
- MEME XML output
- MEME text output
- MAST HTML output
- MAST XML output (2) 右键选择将链接另存为，可下载，可用于后续TBtools软件
- MAST text output
- (Primary) Sequences

**Status Messages**

- Parsing arguments
- Arguments ok
- Starting meme
- meme 23\_species\_FGP2\_refseq\_protein.fasta -protein -oc . -nostatus -time 14400 -mod zoops -motifs 10 -minw 6 -maxw 50 -objfun classic -markov\_order 0
- meme ran successfully in 9.69 seconds
- Starting mast
- mast meme.xml 23\_species\_FGP2\_refseq\_protein.fasta -oc . -nostatus

## Meme 结果:

For further information on how to interpret these results <https://meme-suite.org/meme/doc/mast.html>.  
To get a copy of the MEME software please access <https://meme-suite.org>.

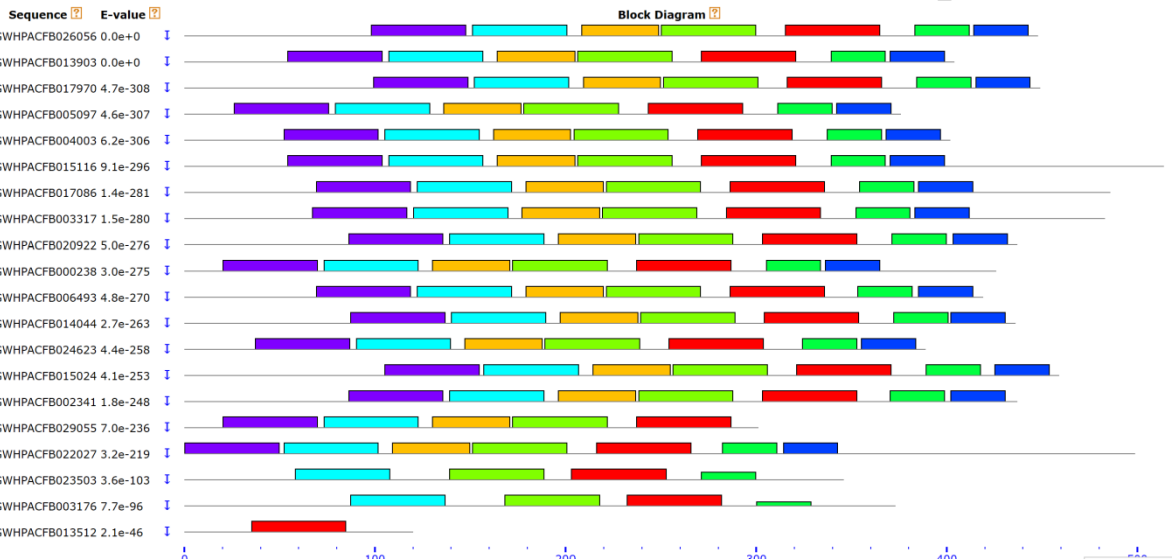
If you use MAST in your research, please cite the following paper:  
Timothy L. Bailey and Michael Gribskov, "Combining evidence using p-values: application to sequence homology searches", *Bioinformatics*, 14(1):48-54, 1998. [\[full text\]](#)

[MOTIFS](#) | [SEARCH RESULTS](#) | [INPUTS & SETTINGS](#) | [PROGRAM INFORMATION](#) | [RESULTS IN TEXT FORMAT](#) | [RESULTS IN XML FORMAT](#)

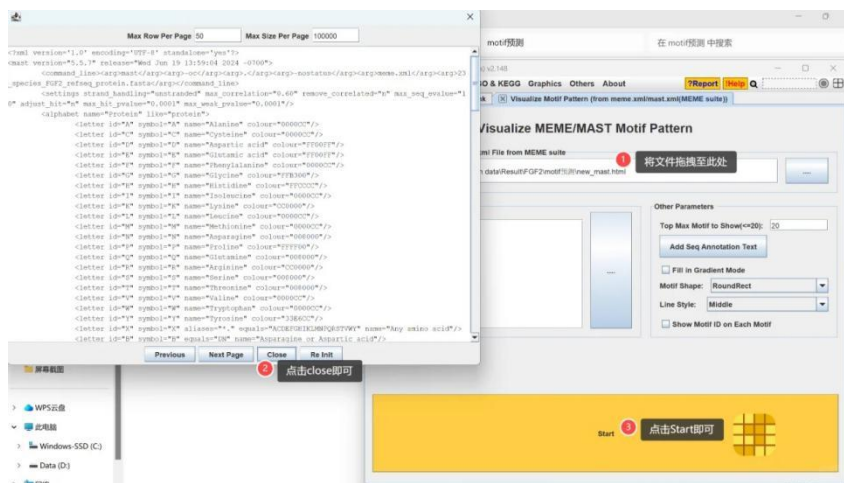
### MOTIFS

Logo	Name	Alt. Name	Width	Motif Simil		
				1.	2.	3.
	FNHFGEGLVQRMPRCRHGYFHVVNDYTHWEMYAIGGSAHPTINSQGNRF	MEME-1	50	--	0.12	0.15
	PKPGLTRHAVIQEPLWIFARDMVIKZELIMNSFKTIDGRGANVHIA	MEME-2	50	0.12	--	0.13
	SIFGSSHIWIDHCSLNCADGLIDAIMGSTAITISNYYFTHHNEVMLLGH	MEME-3	50	0.15	0.13	--
	GNPIDDCWRCDPNWEQNRKRLADCAIGFRNAIGGKDKFYVTDPSDDD	MEME-4	50	0.12	0.16	0.14
	QYVTNVIHGJHIDCKPGGNAMVRDSPHYGWRRTISDGDG	MEME-5	41	0.17	0.15	0.17
	ZSEWKNNWRSEGLMLNGAFFTPSGAGA	MEME-6	29	0.13	0.23	0.21
	SYARASSLGAKPSSLVGLTSFAGVLNCK	MEME-7	29	0.18	0.2	0.21

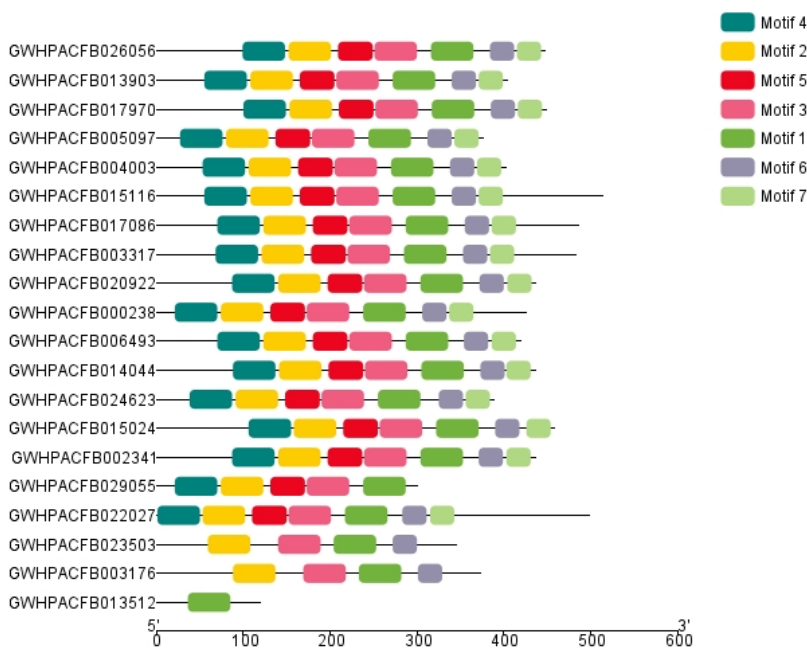
■ FNHFGEGLVQRMPRCRHGYFHVVNDYTHWEMYAIGGSAHPTINSQGNRF 
 ■ PKPGLTRHAVIQEPLWIFARDMVIKZELIMNSFKTIDGRGANVHIA 
 ■ SIFGSSHIWIDHCSLNCADGLIDAIMGSTAITISNYYFTHHNEVMLLGH  
■ GNPIDDCWRCDPNWEQNRKRLADCAIGFRNAIGGKDKFYVTDPSDDD 
 ■ QYVTNVIHGJHIDCKPGGNAMVRDSPHYGWRRTISDGDG 
 ■ ZSEWKNNWRSEGLMLNGAFFTPSGAGA 
 ■ SYARASSLGAKPSSLVGLTSFAGVLNCK



分别放入下载的 MEME 链接文件和基因家族成员原始 id



得到结果：进行美化处理



#### D. 启动子顺式作用元件分析

找到基因 ATG 翻译起始位点上游截取 2000bp 启动子序列（常用长度）进入官网：  
<http://bioinformatics.psb.ugent.be/webtools/plantcare/html/>，点击 Search a sequence，粘贴启动子 FASTA 序列，直接提交，等待结果，下拉找到所有顺式元件列表，导出数据：复制元件名称、起始位置、终止位置、功能，粘贴到 Excel 整理

**Gtf/Gff3 Sequences Extract**

Set Input GTF/GFF3 File  
genomic.gff  
Feature Tag: CDS || Feature ID (Should be Unique!): Parent Initialize

Set Sequences of Input Genome  
bi.fna

Other Parameters  
Max Feature (such as 'CDS') Counts: Up Stream Bases: 2000 Down Stream Bases:  
Min Feature Counts:  Retain Attributes in Header  Retain Only UpStream or DownStream Bases

Set an Output Fasta File  
fasta

Start

**Plant CARE** Search for CARE

email address to send the results back  
back .com

reference name or ID for the sequence (optional)  
optional header

Sequence to submit  
paste as raw DNA sequence (no headers)

please only submit fasta formatted sequences as pure/simple text files, no word documents or other binary things  
选择文件 3序列提取.fasta currently file size limited to 100Kb

Search Reset Form Demo

数据筛选：

- 第二列（顺式作用元件名称）：空白、CAAT-box、TATA-box（不包括带前后缀的行）和 Unnamed\_n 行删除
- 第八列（功能描述）中的空白行删掉
- 第二列、第三列（顺式作用元件序列）、第六列（链方向）与第七列（物种名）删除
- 基因起始位置（原第四列）与终止位置（原第五列）数值进行外理（起始位置：原第四列数值-20.终止位置：原第五列数值+20）

	A	B	C	D	E	F	G	H	I
1	PlantCARE_25904	117	137	Nicotiana glutinosa					
2	PlantCARE_25904	177	197	Nicotiana glutinosa					
3	PlantCARE_25904	195	215	Pisum sativum					
4	PlantCARE_25904	201	221	Pisum sativum					
5	PlantCARE_25904	214	234	Pisum sativum					
6	PlantCARE_25904	578	598	Nicotiana glutinosa					
7	PlantCARE_25904	604	624	Nicotiana glutinosa					
8	PlantCARE_25904	616	636	Pisum sativum					
9	PlantCARE_25904	639	659	Nicotiana glutinosa					
10	PlantCARE_25904	652	672	Nicotiana glutinosa					
11	PlantCARE_25904	693	713	Pisum sativum					
12	PlantCARE_25904	696	716	Arabidopsis thaliana					
13	PlantCARE_25904	797	817	Nicotiana glutinosa					
14	PlantCARE_25904	883	903	Nicotiana glutinosa					
15	PlantCARE_25904	983	1003	Pisum sativum					
16	PlantCARE_25904	1103	1123	Pisum sativum					
17	PlantCARE_25904	1245	1265	Nicotiana glutinosa					
18	PlantCARE_25904	1261	1281	Nicotiana glutinosa					
19	PlantCARE_25904	1335	1355	Arabidopsis thaliana					
20	PlantCARE_25904	1336	1356	Nicotiana glutinosa					
21	PlantCARE_25904	1353	1373	Nicotiana glutinosa					
22	PlantCARE_25904	1462	1482	Nicotiana glutinosa					
23	PlantCARE_25904	1473	1493	Nicotiana glutinosa					
24	PlantCARE_25904	1478	1498	Nicotiana glutinosa					
25	PlantCARE_25904	1545	1565	Nicotiana glutinosa					
26	PlantCARE_25904	1550	1570	Arabidopsis thaliana					
27	PlantCARE_25904	1551	1571	Nicotiana glutinosa					
28	PlantCARE_25904	1568	1588	Nicotiana glutinosa					
29	PlantCARE_25904	1581	1601	Pisum sativum					
30	PlantCARE_25904	1651	1671	Pisum sativum					

- e. 再准备一个序列长度文件.txt（筛到的顺式作用元件+都写 2000 的一列）。然后使用 MEGA11 软件打开基因家族蛋白序列，选择 Data-Phylogenetic Analysis，然后在软件主页面中，选择 NJ 法建树，点击 File，选择 Export Current Tree 点击 File ,save，输出文件名 Newick Export.nwk，将这一串结果粘贴进 TBtools。

**Simple BioSequence Viewer**

Set Input Gene/Protein Len in Tab-delimited Format  
基因长度.txt

Set Input ID list (four column:GeneID|Start|End|DomainName)

GeneID	Start	End	DomainName
gi0487840.1	1701	9	wound-responsive element
gi0495010.1	1303	9	wound-responsive element
gi0546000.1	1527	9	wound-responsive element

Other Parameters

Motif Shape: RoundRect

Line Style: Middle

Fill in Gradient

Null Ele.Type: Line

JJplot2 Engine

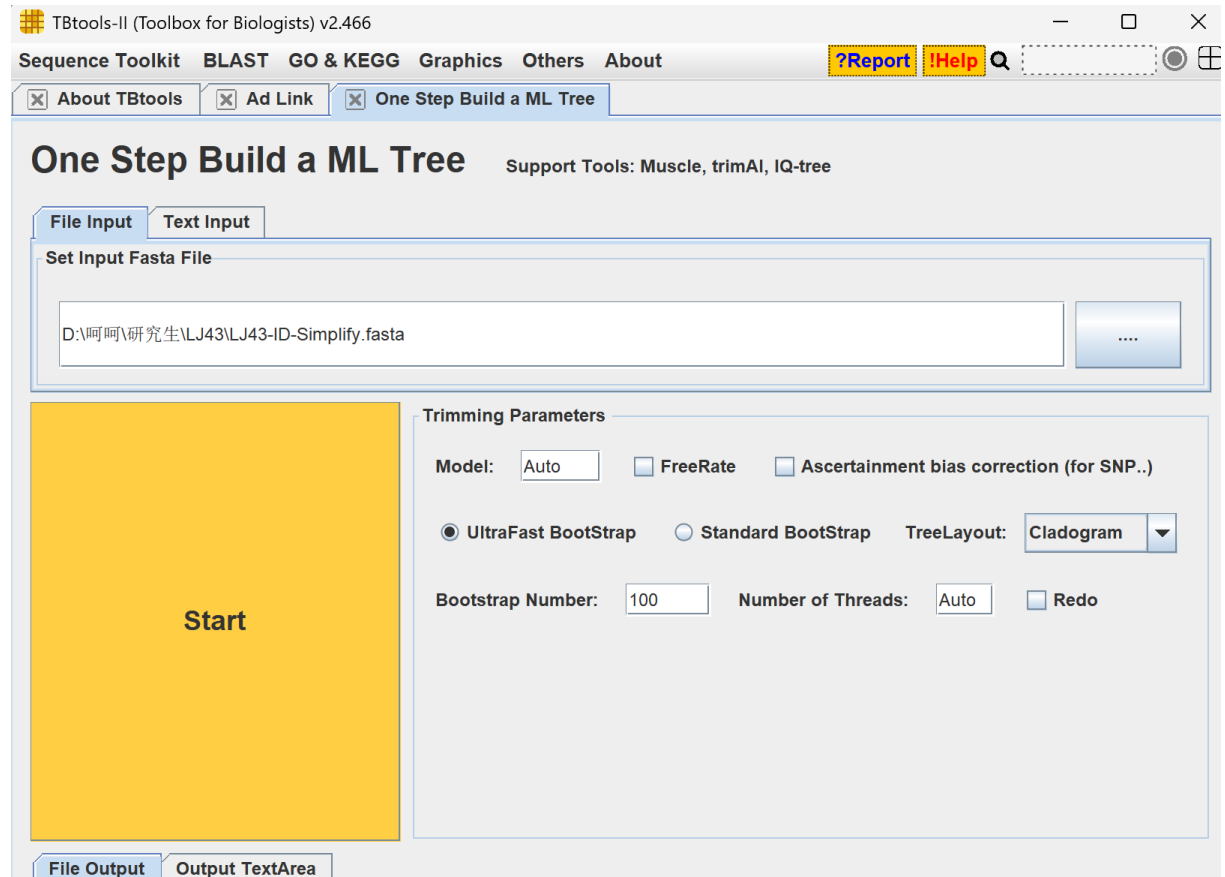
(Optional) Newick Tree String - [Not available for JJplot2 Engine]

(Optional) Set Tree Renaming File  
Drag Input Gene Renaming Info (OldID|NewName)

Start

## E. 简单构建进化树

TBtools 直接一键建树。顶部搜索：Phylogenetic，选择 Neighbor-Joining (NJ) Tree，Input Alignment File：选刚才生成的 align.fasta，Bootstrap: 1000（检验可信度，必须开），点 Run → 等待完成（1-5 分钟）



TBtools 进行进化树的构建相较 MEGA 更快，但是同时也存在缺陷就是其大多数数据处理方法是内设的方法，参考了最常用可信度最大的参数，用户对于参数的调控选项较少，遇到特殊的情况需要使用其他的建树方法还是需要回到 MEGA 进行处理。

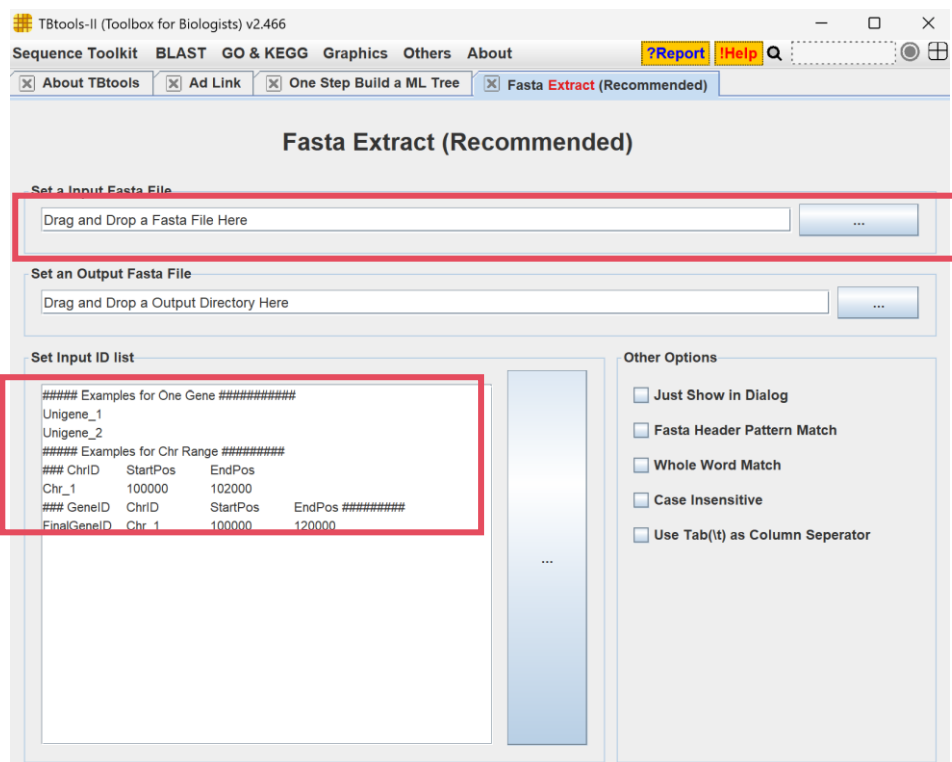
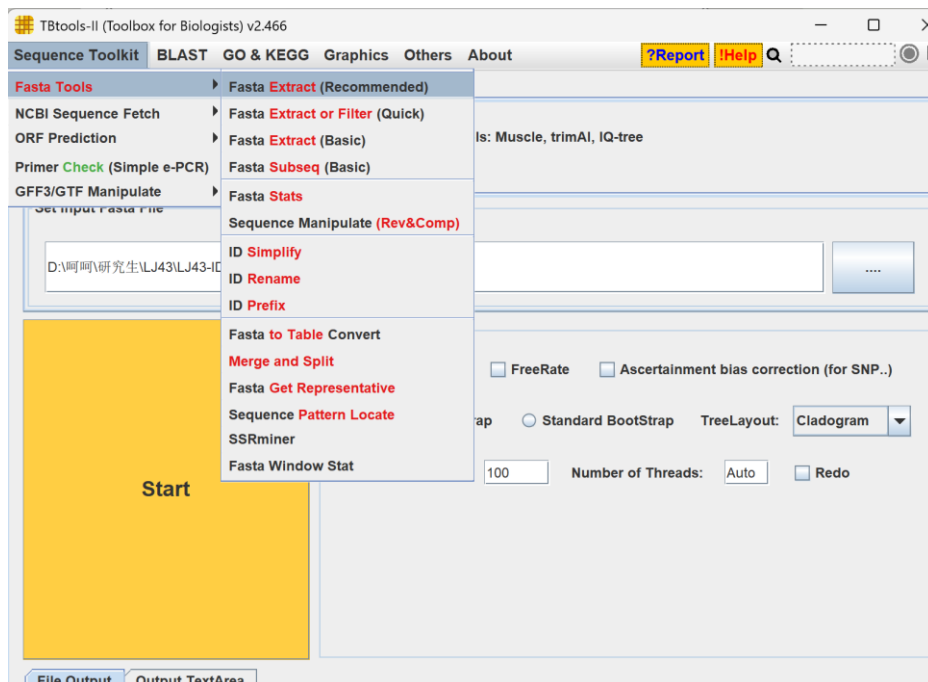
## F. 从全基因组数据中找到目标序列

1. Fasta 序列文件输入文本框，用户可以直接拖拽硬盘中的 Fasta 文件并放置到文本框中，路径会自动获取；也可以点击跟随文本框的按钮“...”，在弹出文件选择框中选取对应文件即可

2. Initialize 按钮，在设置 Fasta 序列文件后，可以看到 Start 按钮仍然不可点击。需要用户点击 Initialize 按钮，创建 Fasta 序列索引文件（如前期已有，则会软件会自动复用，节省计算时间）

3. 输出文件设置文本框，用户同样可以拖拽放置文件或者文件夹，程序会自动获取输出文

文件夹，用户需要补全一个输出文件名；当然也可以直接点击跟随文本框的摁钮，在弹出的选择框中设置对应输出文件即可



### (3) 问题

**Q1:** 在使用软件过程中发现当文件过大或是序列过多的情况下会出现报错的问题，请问 TBtools 在进行分析时，是否有处理文件大小上限？

**Q2:** 为什么使用 TBtools 的 One Step Build a ML Tree 功能模块构建系统发生树时会在跳出 Congratulations 后报错？