“实用生物信息技术”课程小组讨论总结报告

组：G3　次：R1　组长：陈伟辉　执笔：陶益可

一、时间 2026 年 4 月 8 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) BLAST 数据库使用练习

① 如何通过部分基因序列搜索完整序列以及序列来源

② 结果解读以及关键参数筛选

(2) Uniprot 使用复习、NCBI-EBI、CNCB 使用练习

① Uniprot 数据库使用

② EBI 使用

③ 从 UniProt 数据库中提取人、小鼠、大鼠血红蛋白 alpha 亚基蛋白质序列，进行双序列全局比对。

(3) TBtool 下载以及热图制作学习

① 热图制作以及优化

(4) 数据库 PHYTOZOME 使用

(5) UniProt 蛋白质数据库

① UniProt 数据库概况

② UniProt 中的帮助文档包括哪些信息？

③ 豌豆内膜蛋白注释信息

(6) 课题相关物种信息

① 拟南芥

② 大肠杆菌

③ 高粱

④ 粳稻

⑤ 茶

⑥ 美国黑杨

(7) 课题相关蛋白信息

① CKX5-ARATH

② Q9M041

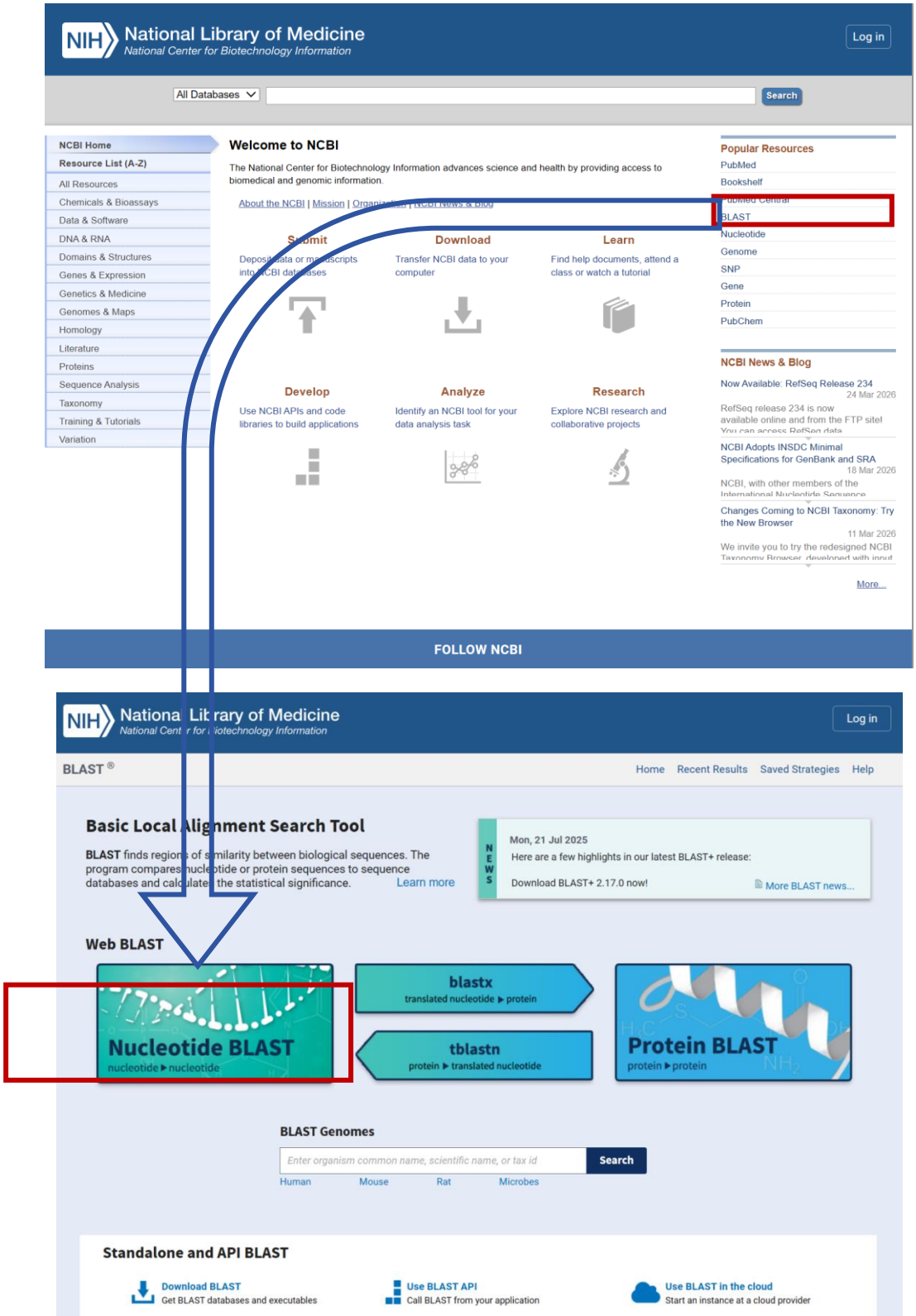
③ C79A1-SORBI

(8) 问题

（1）BLAST 数据库使用练习

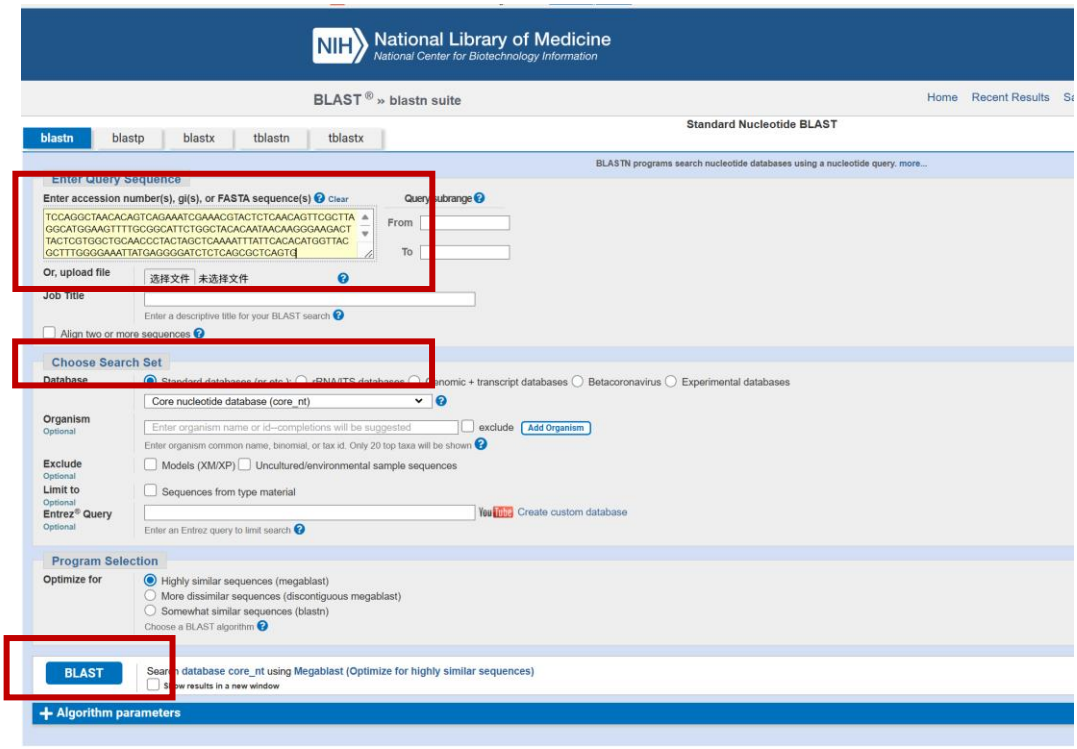
① 如何通过部分基因序列搜索完整序列以及序列来源

给到已知序列，在 blast 中找到该序列的相关信息，复制需要进行搜索的基因序列在 NCBI BLAST



在 ENTER 中输入序列（也可以选择文件）：

CTGATGAATCCCCTAATGATTTTGGTAAAAATCATTAAGTTAAGGTGGATACACATCTTGTC
ATATGATCAAATGGTTTCGCGAAAAATCAATAATCAGACAACAAGATGTGCGAACTCGATATTTT
ACACGACTCTCTTTACCAATTCTGCCCCGAATTACACTTAAAACGACTCAACAGCTTAACGTTG
GCTTGCCACGCATTACTTGACTGTAAAACTCTCACTCTTACCGAACTTGGCCGTAACCTGCCAA
CCAAAGCGAGAACAAAACATAACATCAAACGAATCGACCGATTGTTAGGTAATCGTCACCTCC
ACAAAGAGCGACTCGCTGTATACCGTTGGCATGCTAGCTTTATCTGTTCGGGCAATACGATGCCC
ATTGTACTTGTTGACTGGTCTGATATTCGTGAGCAAAAACGACTTATGGTATTGCGAGCTTCAGT
CGCACTACACGGTCGTTCTGTTACTCTTTATGAGAAAGCGTTCCCGCTTTCAGAGCAATGTTCA
AAGAAAGCTCATGACCAATTTCTAGCCGACCTTGCGAGCATTCTACCGAGTAACACCACACCGC
TCATTGTCAGTGATGCTGGCTTTAAAGTGCCATGGTATAAATCCGTTGAGAAGCTGGGTTGGTAC
TGGTTAAGTCGAGTAAGAGGAAAAGTACAATATGCAGACCTAGGAGCGGAAAACTGGAAACCT
ATCAGCAACTTACATGATATGTCATCTAGTCACTCAAAGACTTTAGGCTATAAGAGGCTGACTAA
AAGCAATCCAATCTCATGCCAAATTCTATTGTATAAATCTCGCTCTAAAGGCCGAAAAAATCAGC
GCTCGACACGGACTCATTGTCACCACCCGTCACCTAAAATCTACTCAGCGTCGGCAAAGGAGC
CATGGGTTCTAGCAACTAACTTACCTGTTGAAATTCGAACACCCAAACAACTTGTTAATATCTAT
TCGAAGCGAATGCAGATTGAAGAAACCTTCCGAGACTTGAAAAGTCCTGCCTACGGACTAGGC
CTACGCCATAGCCGAACGAGCAGCTCAGAGCGTTTTGATATCATGCTGCTAATCGCCCTGATGCT
TCAACTAACATGTTGGCTTGCGGGCGTTCATGCTCAGAAACAAGGTTGGGACAAGCACTTCCA
GGCTAACACAGTCAGAAATCGAAACGTACTCTCAACAGTTCGCTTAGGCATGGAAGTTTTGCG
GCATTCTGGCTACACAATAACAAGGGAAGACTTACTCGTGGCTGCAACCCTACTAGCTCAAAAT
TTATTCACACATGGTTACGCTTTGGGGAAATTATGAGGGGATCTCTCAGCGCTCAGTG



② 结果解读以及关键参数筛选

在 choose search set 中进行关键参数设置

A. 排除模式生物：在 nr 数据库搜索时，建议在 Entrez Query 中输入排除项。例如，如果序列包含常见的宿主菌（如 Escherichia coli）的污染片段，可输入：NOT Escherichia[orgn],这能防止结果被大量 E. coli 染色体序列淹没，让你更专注于质粒特有的匹配项。

B. 期望值（Expect threshold）：保持默认（10）即可。如果比对结果非常短且 E-value 接近 0.001，需警惕是否为随机匹配。

C. 低复杂度区域过滤：建议取消勾选“Filter low complexity regions”。质粒中的某些区域（如多克隆位点、重复序列）可能被过滤掉，导致漏掉关键比对。

注意在 BLAST 中有几个数据库，经过搜索总结(在此处使用 Blastn）：

工具	输入	数据库	翻译对象	核心用途
blastn	核酸	核酸	无	找高度相似的 DNA 序列
blastp	蛋白	蛋白	无	找同源蛋白、功能分析
blastx	核酸	蛋白	输入核酸	从新序列中预测编码基因
tblastn	蛋白	核酸	数据库核酸	在未注释序列中找特定蛋白的编码基因
tblastp	蛋白	蛋白	无	同 blastp，蛋白-蛋白比对

搜索结果筛选其中关键参数有：

A. 查询覆盖率（Query Cover）：如果覆盖率低于 70%，说明你的质粒序列中只有一部分与数据库匹配，其余部分可能是未知功能区域、不同模块的拼接，或来自其他来源的嵌合体。

B. 相似度（Percent Identity）：

a. >99%：极有可能是同一质粒，或该质粒直接来源于此。

b. 95% - 99%：可能是同源质粒，但存在突变或不同分离株的差异。

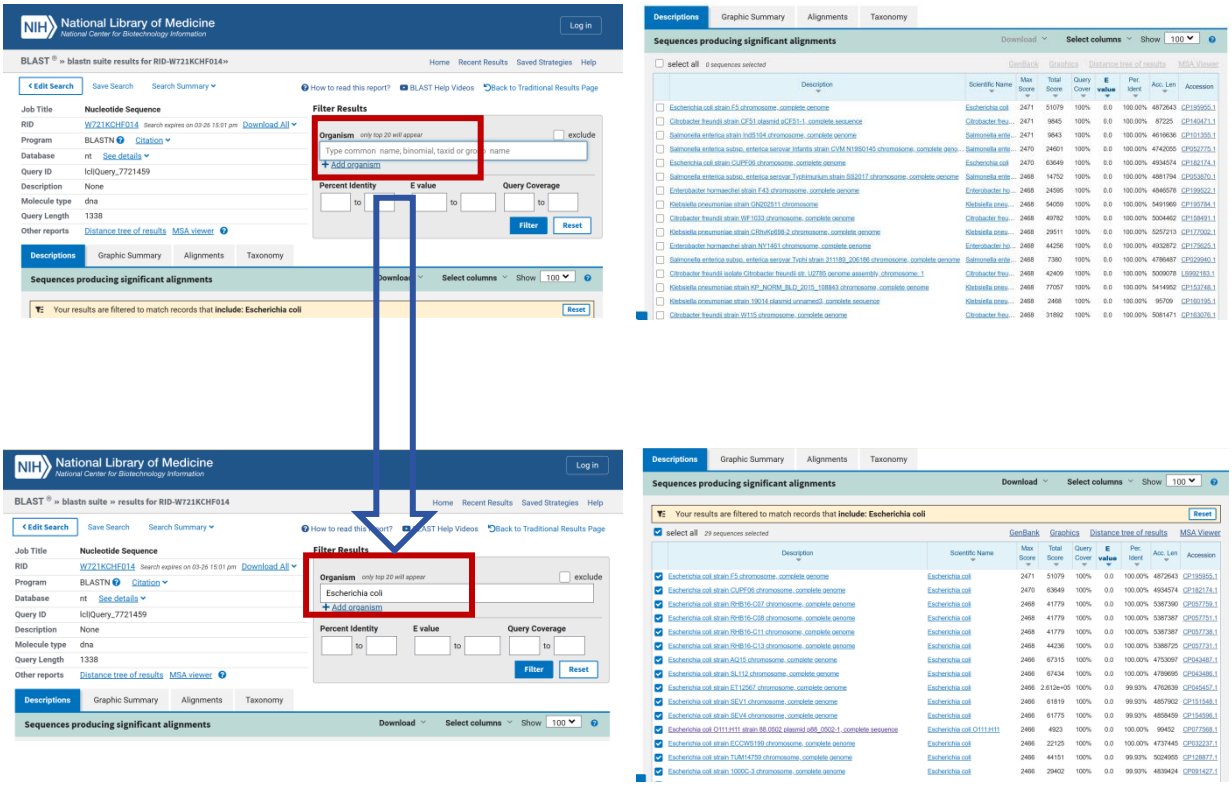
c. <90%（但 BLASTx 匹配良好）：说明质粒骨架不同，但功能基因（如抗性、接合转移）高度相似，表明功能模块的来源。

C. 匹配的物种/质粒名称：

如果多个 top hits 均指向某特定菌属（如 Lactococcus），则质粒很可能源自该菌。

如果 hits 混杂在不同属的质粒上，通常表明质粒携带了可移动元件（转座子、整合性接合元件），其来源应追踪这些核心元件的最佳匹配。

如何进行进一步筛选，可以通过右侧灰色框架内的 filter 进行筛选，organism 等进行筛选。在此处参考输入大肠杆菌后的结果对比。



（2）Uniprot 使用复习、NCBI-EBI、CNCB 使用练习

① Uniprot 数据库使用

A. 基础检索

在官网首页（https://www.uniprot.org/）的搜索框中输入关键词，点击搜索即可。

可以输入的内容：

a. 基因名，如 tp53 或 ACE2

b. 蛋白名，如 insulin receptor

c. 物种名，如 human、mouse 或拉丁名 Homo sapiens

d. UniProt 登录号，如 P04637（TP53 蛋白）

e. 关键词，如 kinase、apoptosis

B. 高级检索（Advanced Search）

当需要更精确的检索时，点击搜索框右侧的 Advanced 进入高级检索页面。

按字段检索：从下拉菜单中选择检索字段（如 Organism、Gene name、Function、Disease 等），再输入关键词

组合条件：通过“Add field”添加多个条件，并用 AND、OR、NOT 进行逻辑组合。

C. 批量检索与下载

如果需要一次性获取多个蛋白的信息，可以使用以下方式：

高级搜索 + 导出：构建检索条件得到结果列表后，点击 Download，选择 TSV 或 Excel 格式，即可下载包含 UniProt ID、基因名、物种、序列长度等字段的表格。下载前可点击 Customize columns 自定义需要导出的字段。

② EBI 使用

EBI 的入口是 Ensembl（基因组浏览器）和 EMBL-EBI 主站上的工具列表。

A. 蛋白质深度分析（UniProt）

使用方法：搜索蛋白名称或基因名 → 进入条目后，左侧导航栏可查看功能、亚细胞定位、结构域、序列等信息。自定义导出：搜索结果页面可自定义列（如基因名、物种、分子量），一键下载为 Excel。

B. 多序列比对与全局比对

a. Clustal Omega：EBI 提供的多序列比对工具支持数十至数千条序列，结果可视化好，支持输出多种格式。

b. EMBOSS Needle：双序列全局比对的首选工具，能严格从一端到另一端比对两条序列，计算精确的相似度。

C. 功能结构域分析（InterProScan）

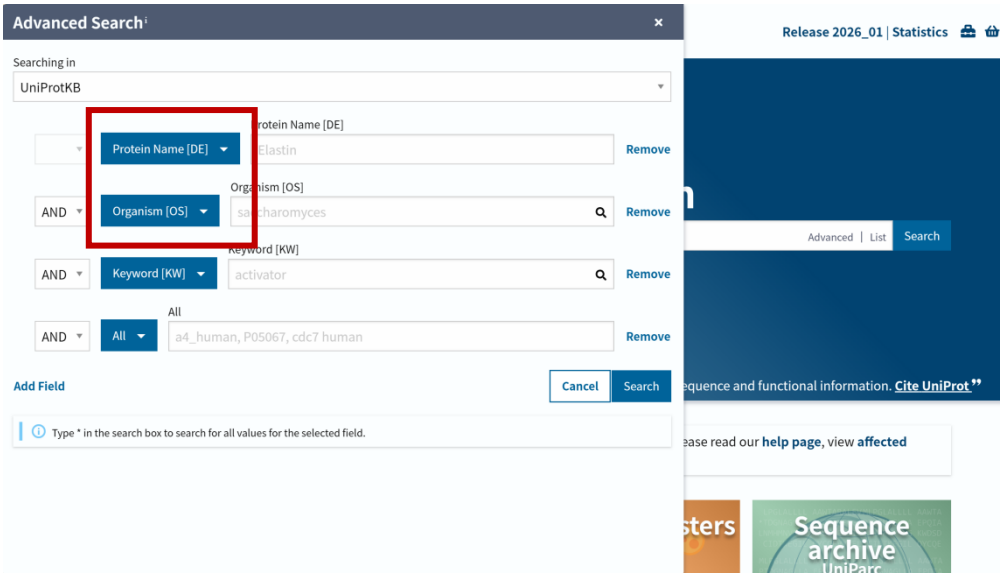
用途：输入蛋白序列，可一次性整合 Pfam、PRINTS、PROSITE 等数十个数据库的预测结果，给出蛋白的结构域、家族、功能位点等信息。这是 NCBI 的 CDD（保守结构域数据库）之外的最佳补充。

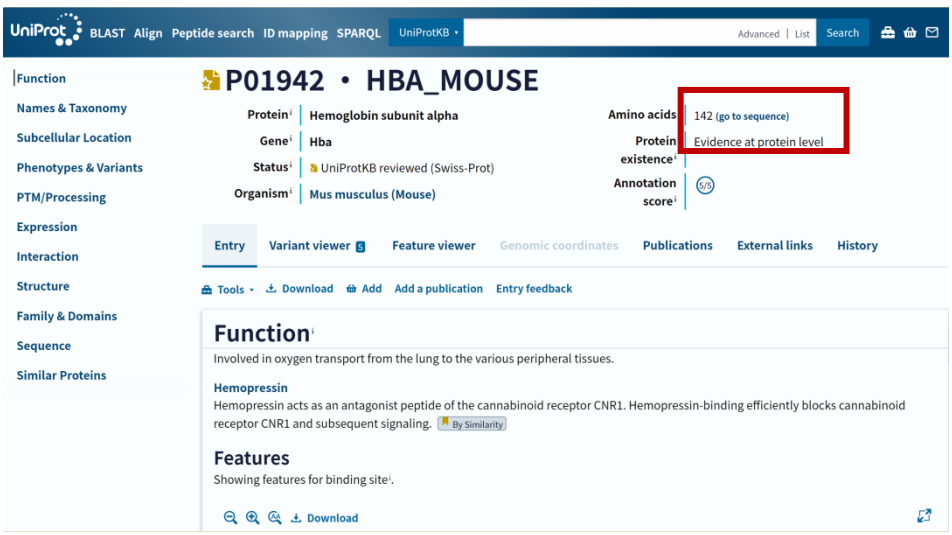
D. 蛋白结构预测（AlphaFold DB）

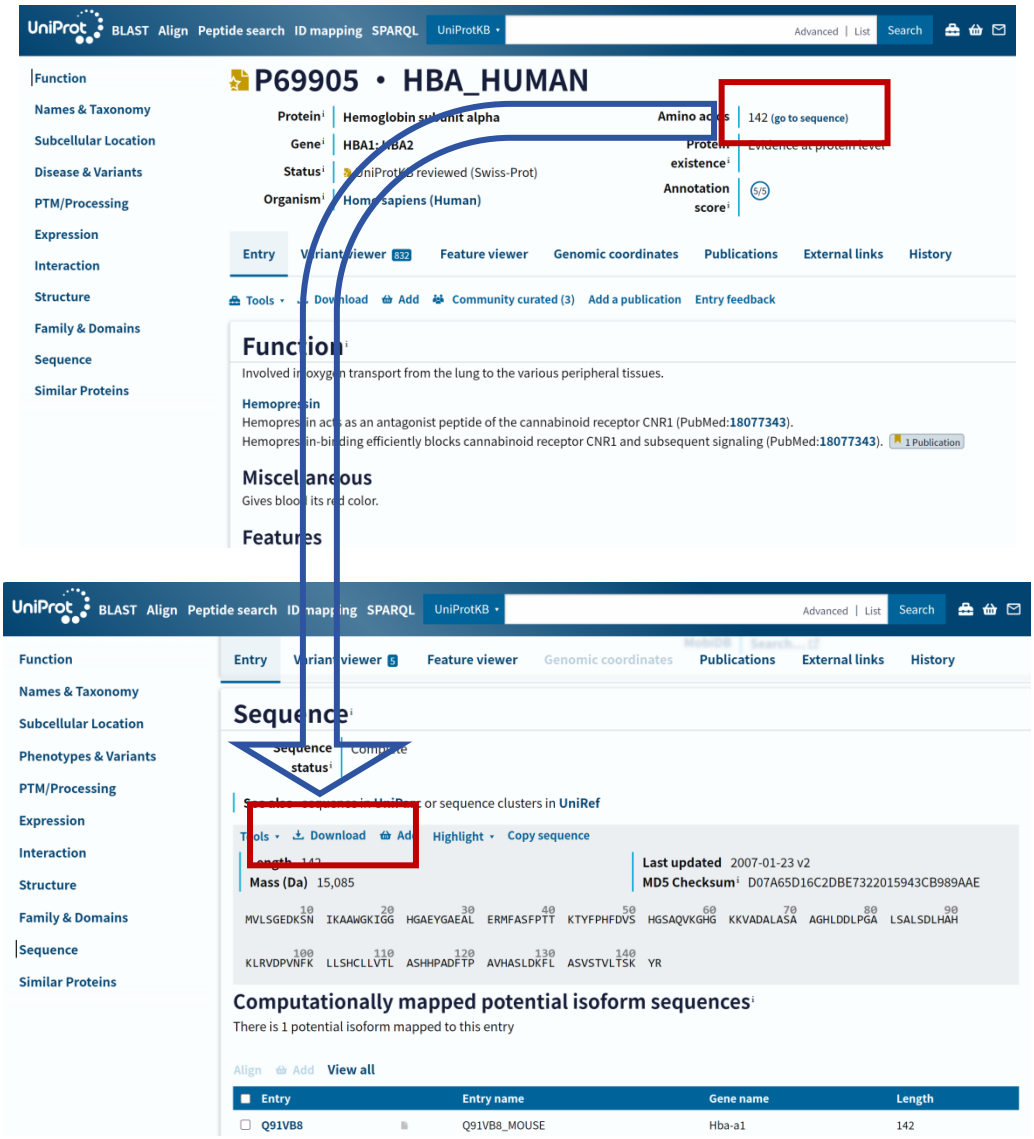
输入 UniProt ID 或基因名，可直接查看 EBI 与 DeepMind 合作的 AlphaFold 预测的蛋白质三维结构模型，支持下载 PDB 文件。

③ 从 UniProt 数据库中提取人、小鼠、大鼠血红蛋白 alpha 亚基蛋白质序列，进行双序列全局比对。

A. 使用方法：

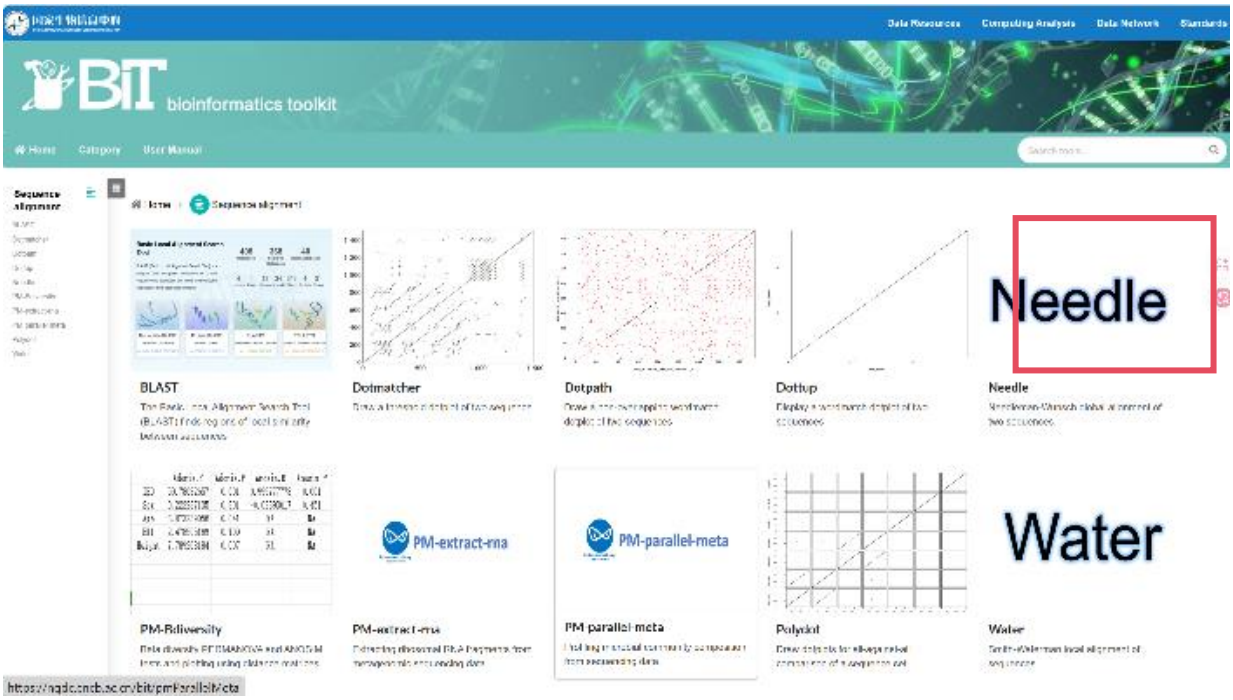






双序列全局比对平台：

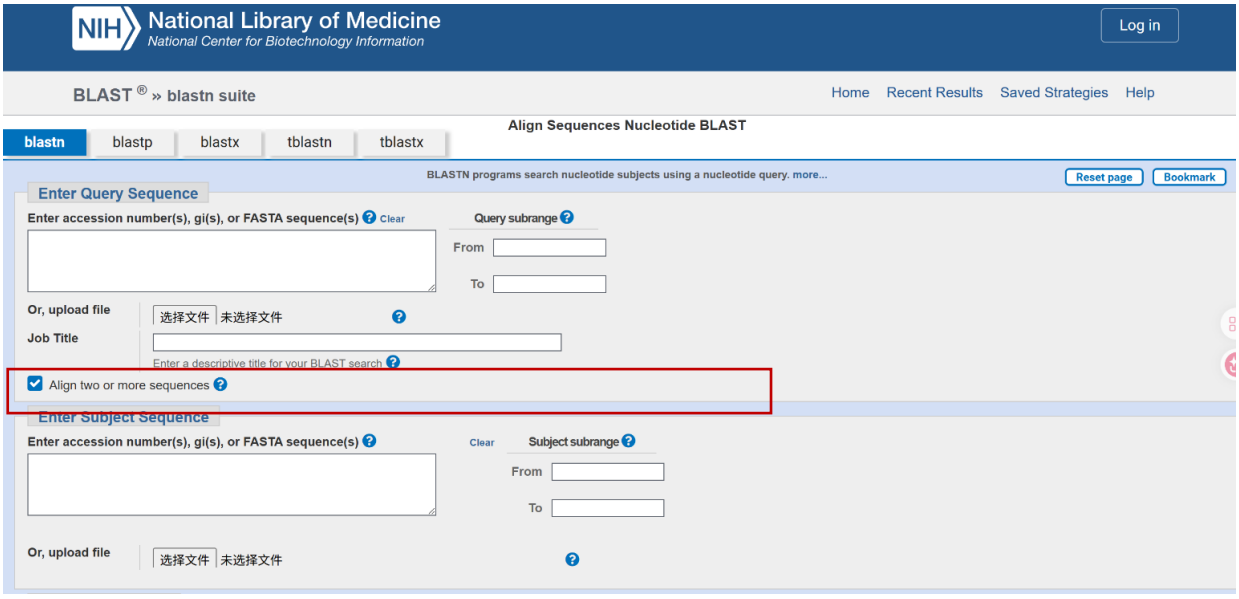
a. Needle：CNCB➡BiT➡Needle➡在窗口提交待比对序列（网址：NGDC Cloud）



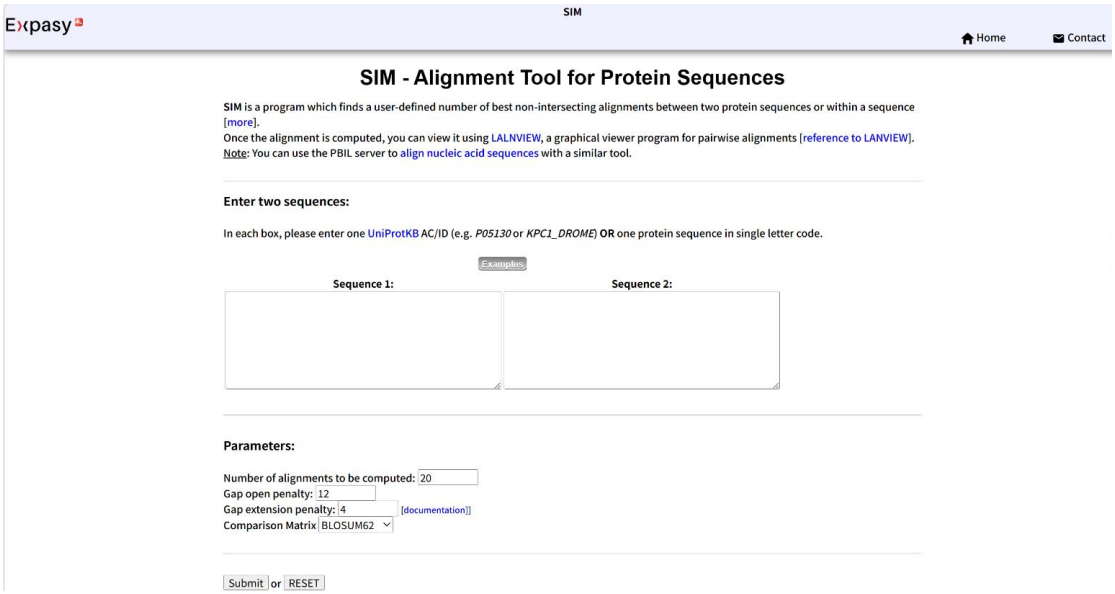
b. BLAST：NCBI➞BLAST(blastn / blastp)（最常用，局部比对）➞选择 Align two or more sequences 模式➞在窗口提交待比对序列（网址：Nucleotide BLAST: Align two or more sequences using BLAST）

附：详细教程可参考官方报告

（报告网址 C:\Users\tao\AppData\Local\Temp\1\mso6BC3.tmp）



c. Expasy SIM（经典双序列比对平台）



注意事项：进行序列比对时，应注意各平台支持的序列文件格式，如 BLAST 支持 FASTA、裸序列以及序列标识符（即直接输入 Accession 号或 GI 号）。

B.人-大鼠双序列全局对比

a. 人序列长度：142 aa（原始序列 C 端为 ...KYR---，末尾三个缺失）

b. 大鼠序列长度：147 aa（C 端有额外 AHKYH 5 个残基）

比对总长度：147 aa（含人序列中的 3 个空位和大鼠序列的 2 个空位，总空位数 5）

score：587，这些数值表明两条序列高度同源，符合哺乳动物血红蛋白 α 亚基的进化保守性。

高度保守区域

a. 血红素结合位点（如第 58 位 His，第 87 位 His 等）均完全保留。

b. 与 β 亚基相互作用的区域（如 α1β2 接触面）也高度一致。

c. N 端前 50 个残基几乎完全相同，仅少数位置为保守替换（如人 Ala 被大鼠 Asp 替换）。血红蛋白 α 亚基在哺乳动物中高度保守，人和大鼠的序列相似性 >80%，保证了基本功能（携氧、协同性）的稳定性。

生物学意义：

a. 少数差异多位于表面暴露区域或 C 端，可能与物种特异的生理调节（如 pH 敏感性、与别构效应物的结合）相关。

b. C 端延伸（大鼠多 5 个残基）可能影响亚基间接触界面，从而影响氧亲和力或四聚体稳定性，是功能研究的一个潜在关注点。

总结：

a. 该比对结果清晰展示了人和大鼠血红蛋白 α 亚基的序列相似性及细微差异。高一致性

（75.5%）和高相似性（81.6%）证实了其直系同源关系；

b. C 端长度差异和少数非保守替换是功能分化研究的重要线索。

C.小鼠 vs 大鼠 血红蛋白 α 亚基（HBA）Needle 全局比对

核心统计指标（最关键）

比对长度：142 个氨基酸（两条序列长度完全一致）

a. 一致性 (Identity)：120/142 = 84.5%,84.5% 的氨基酸完全相同

b. 相似性 (Similarity)：127/142 = 89.4%,包含性质相似的氨基酸（如疏水→疏水、带电→带电），整体功能保守性很高

c. 空位 (Gaps)：0/142 = 0%,没有插入 / 缺失（Indel），序列长度完全匹配，结构高度保守

d. 比对得分 (Score)：632.0,分值很高，说明比对结果可靠，同源性极强

D. 小鼠 vs 人 血红蛋白 α 亚基（HBA）Needle 全局比对

比对长度 142 两条序列比对后的总长度（无空位）

a. 一致性（Identity）122/142 (85.9%)完全相同的氨基酸残基占比

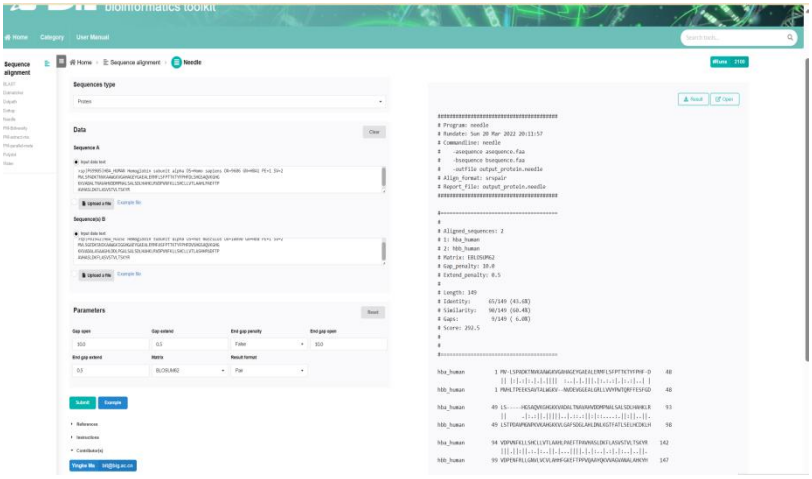
b. 相似性（Similarity）131/142 (92.3%)理化性质相似的氨基酸残基占比（含完全一致）

c. 空位（Gaps）0/142 (0.0%)比对中无任何插入 / 缺失（gap）

d. 比对得分（Score）648.0 基于 EBLOSUM62 矩阵和罚分计算的总得分，分值极高，说明比对质量非常好

从比对片段可以看到：

两条序列完全没有空位，长度完全一致，框架高度保守。绝大多数位点是完全匹配（|）或高度相似（:/.），仅在少数位点存在差异（如人序列的 SPADKTNV vs 小鼠的 SGEDKSNIK 区域）。整体高度保守，符合同源蛋白在不同物种间的功能保守性（α - 珠蛋白是携氧核心蛋白，进化压力大，序列高度保守）。



```
########################################
# Program: needle
# Rundate: Wed 25 Mar 2026 23:35:59
# Commandline: needle
#    -asequence /blastdisk2/blastdata/toolkitv3/task/8GONPDXTQVF/sequence_a.faa
#    -bsequence /blastdisk2/blastdata/toolkitv3/task/8GONPDXTQVF/sequence_b.faa
#    -gapopen 10.0
#    -gapextend 0.5
#    -endweight false
#    -endopen 10.0
#    -endextend 0.5
#    -datafile EBLOSUM62
#    -aformat pair
#    -outfile output_protein.needle
# Align_format: pair
# Report_file: output_protein.needle
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: hba_human
# 2: hba_rat
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 147
# Identity:     111/147 (75.5%)
# Similarity:   120/147 (81.6%)
# Gaps:           5/147 ( 3.4%)
# Score: 587.0
#
#
#=======================================

hba_human           1 MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLS     50
                      ||||..||||:|..|||:|.|.||||.|||:|||.:|||||||.|.|:|
hba_rat             1 MVLSADDKTNIKNCWGKIGGHGGEYGEEALQRMFAAFPTTKTYFSHIDVS     50

hba_human          51 HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFK    100
                      .||||||.|||||||||..|..||:|:|.|||.||||||||||||||||||
hba_rat            51 PGSAQVKAHGKKVADALAKAADHVEDLPGALSTLSDLHAHKLRVDPVNFK    100

hba_human         101 LLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR-----    142
                      .|||||||||||.|.|.:||||:|||||||||||||||||||||
hba_rat           101 FLSHCLLVTLACHHPGDFTPAMHASLDKFLASVSTVLTSKYRAHKYH    147


#---------------------------------------
#---------------------------------------
```

```
########################################
# Program: needle
# Rundate: Wed 25 Mar 2026 23:38:30
# Commandline: needle
#    -asequence /blastdisk2/blastdata/toolkitv3/task/WBL7T5DXKBJ/sequence_a.faa
#    -bsequence /blastdisk2/blastdata/toolkitv3/task/WBL7T5DXKBJ/sequence_b.faa
#    -gapopen 10.0
#    -gapextend 0.5
#    -endweight false
#    -endopen 10.0
#    -endextend 0.5
#    -datafile EBLOSUM62
#    -aformat pair
#    -outfile output_protein.needle
# Align_format: pair
# Report_file: output_protein.needle
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: HBA_MOUSE
# 2: HBA_RAT
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 142
# Identity:     120/142 (84.5%)
# Similarity:   127/142 (89.4%)
# Gaps:           0/142 ( 0.0%)
# Score: 632.0
#
#
#=======================================

HBA_MOUSE           1 MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVS     50
                      |||..:||:|||..||||||.|||.|||:||||:||||||||.|.|||
HBA_RAT             1 MVLSADDKTNIKNCWGKIGGHGGEYGEEALQRMFAAFPTTKTYFSHIDVS     50

HBA_MOUSE          51 HGSAQVKGHGKKVADALASAAGHLDDLPGALSALSDLHAHKLRVDPVNFK    100
                      .|||||.|||||||||.||.|::|||||.|||||||||||||||||||
HBA_RAT            51 PGSAQVKAHGKKVADALAKAADHVEDLPGALSTLSDLHAHKLRVDPVNFK    100

HBA_MOUSE         101 LLSHCLLVTLASHHPADFTPAVHASLDKFLASVSTVLTSKYR    142
                      .|||||||||.|||.||||:|||||||||||||||||||
HBA_RAT           101 FLSHCLLVTLACHHPGDFTPAMHASLDKFLASVSTVLTSKYR    142


#---------------------------------------
#---------------------------------------
```

符号含义：

|：完全匹配的氨基酸

:：理化性质相似的氨基酸

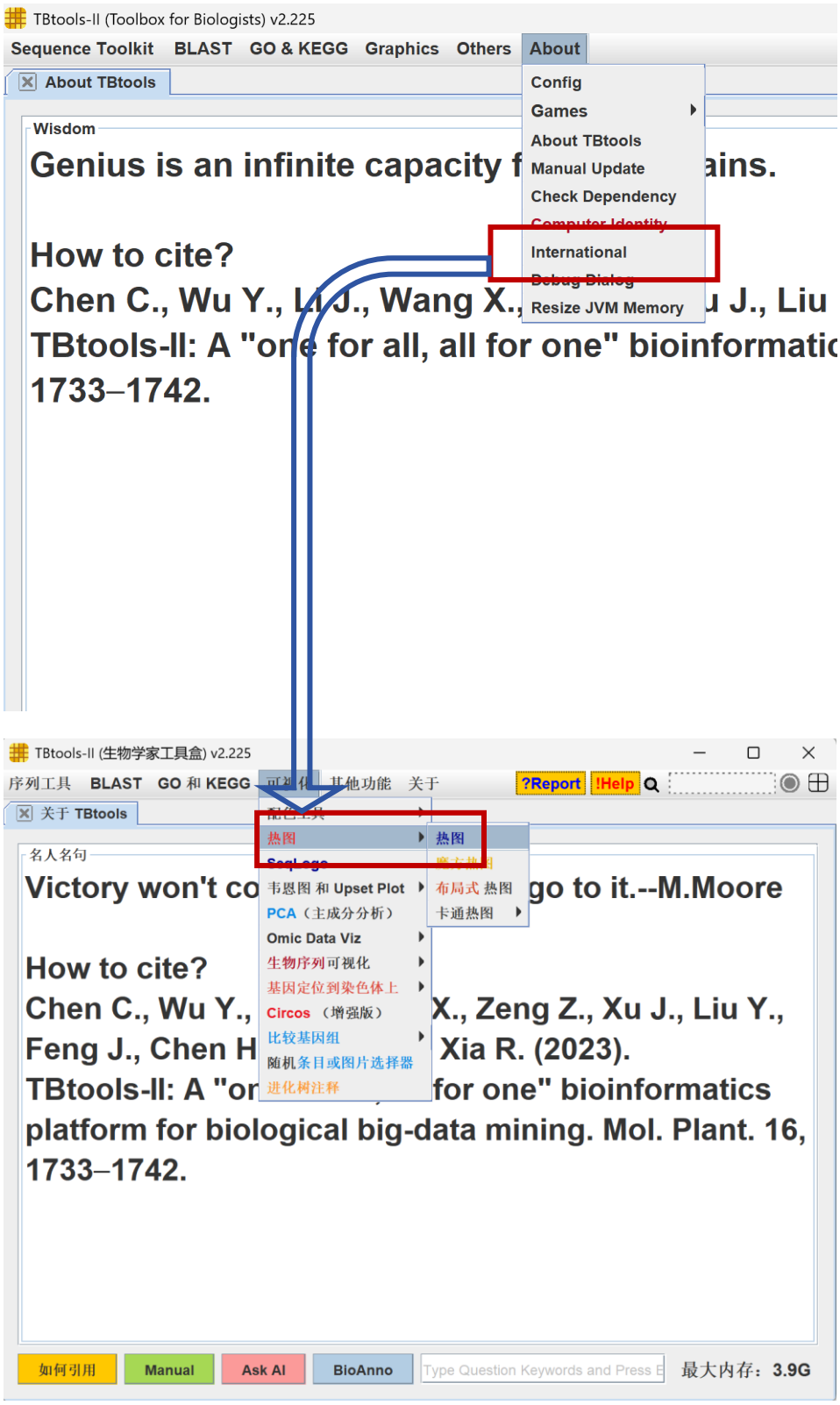
.：理化性质较弱相似的氨基酸

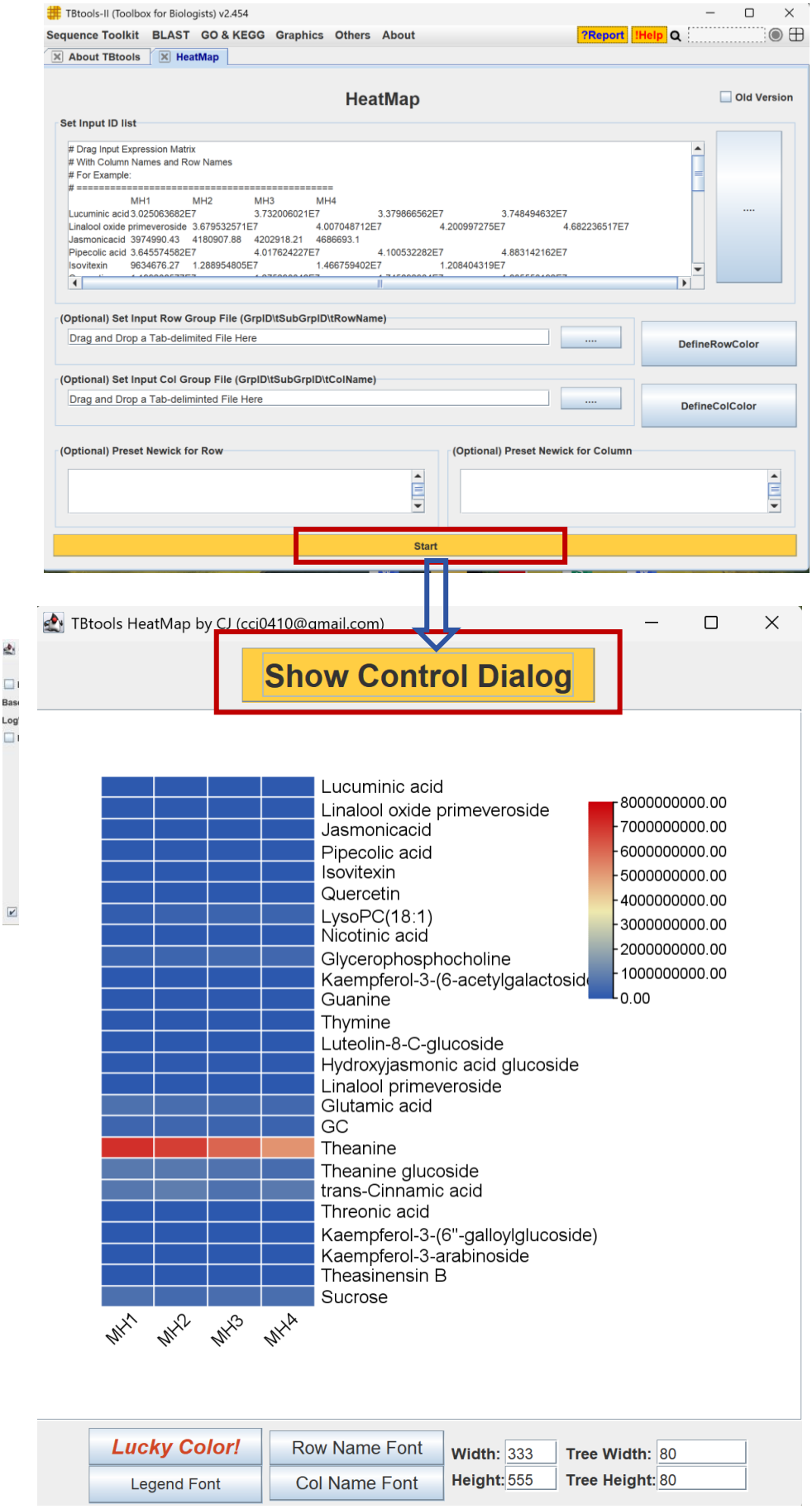
空格：不相似 / 不匹配

-：序列空位

（3）TBTOOLS 下载以及热图制作学习

④热图制作以及优化

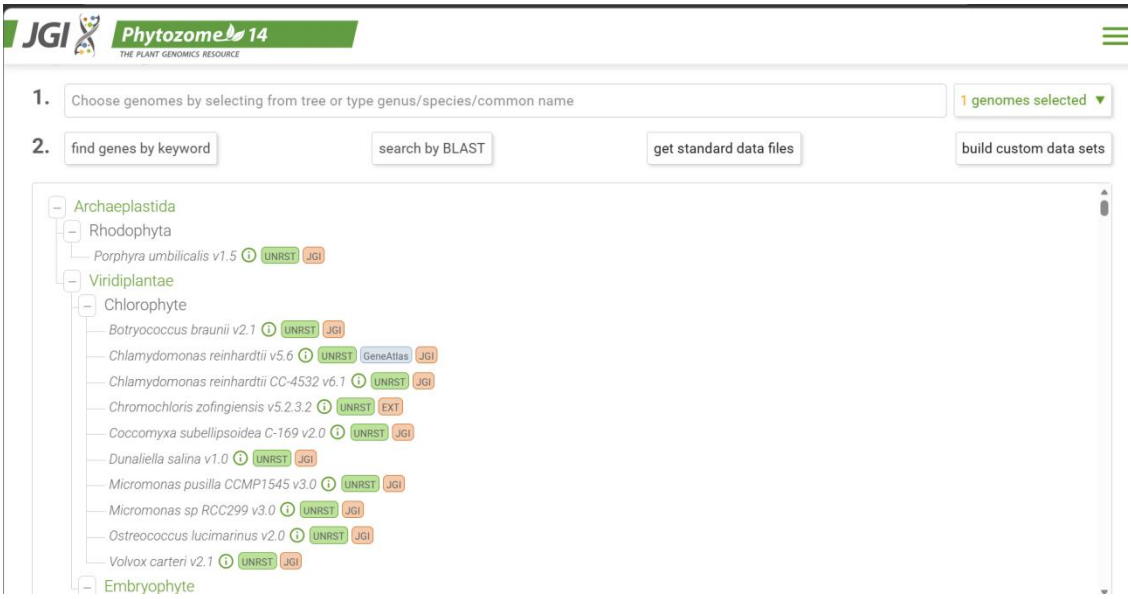



TBtools-II (Toolbox for Biologists) v2.454
Sequence Toolkit BLAST GO & KEGG Graphics Others About
?Report !Help
About TBtools
HeatMap
HeatMap
Old Version
Set Input ID list
Drag Input Expression Matrix
With Column Names and Row Names
For Example:
MH1 MH2 MH3 MH4
Lucuminic acid 3.025063682E7 3.732006021E7 3.379866562E7 3.748494632E7
Linalool oxide primeveroside 3.679532571E7 4.007048712E7 4.200997275E7 4.682236517E7
Jasmonicacid 3974990.43 4180907.88 4202918.21 4686693.1
Pipecolic acid 3.645574582E7 4.017624227E7 4.100532282E7 4.883142162E7
Isovitexin 9634676.27 1.288954805E7 1.466759402E7 1.208404319E7
(Optional) Set Input Row Group File (GrpID\tSubGrpID\tRowName)
Drag and Drop a Tab-delimited File Here
DefineRowColor
(Optional) Set Input Col Group File (GrpID\tSubGrpID\tColName)
Drag and Drop a Tab-deliminted File Here
DefineColColor
(Optional) Preset Newick for Row
(Optional) Preset Newick for Column
Start
TBtools HeatMap by CJ (ccj0410@gmail.com)
Show Control Dialog
Lucuminic acid
Linalool oxide primeveroside
Jasmonicacid
Pipecolic acid
Isovitexin
Quercetin
LysoPC(18:1)
Nicotinic acid
Glycerophosphocholine
Kaempferol-3-(6-acetylgalactosid
Guanine
Thymine
Luteolin-8-C-glucoside
Hydroxyjasmonic acid glucoside
Linalool primeveroside
Glutamic acid
GC
Theanine
Theanine glucoside
trans-Cinnamic acid
Threonic acid
Kaempferol-3-(6''-galloylglucoside)
Kaempferol-3-arabinoside
Theasinensin B
Sucrose
MH1
MH2
MH3
MH4
8000000000.00
7000000000.00
6000000000.00
5000000000.00
4000000000.00
3000000000.00
2000000000.00
1000000000.00
0.00
Lucky Color!
Row Name Font
Width: 333
Tree Width: 80
Legend Font
Col Name Font
Height: 555
Tree Height: 80

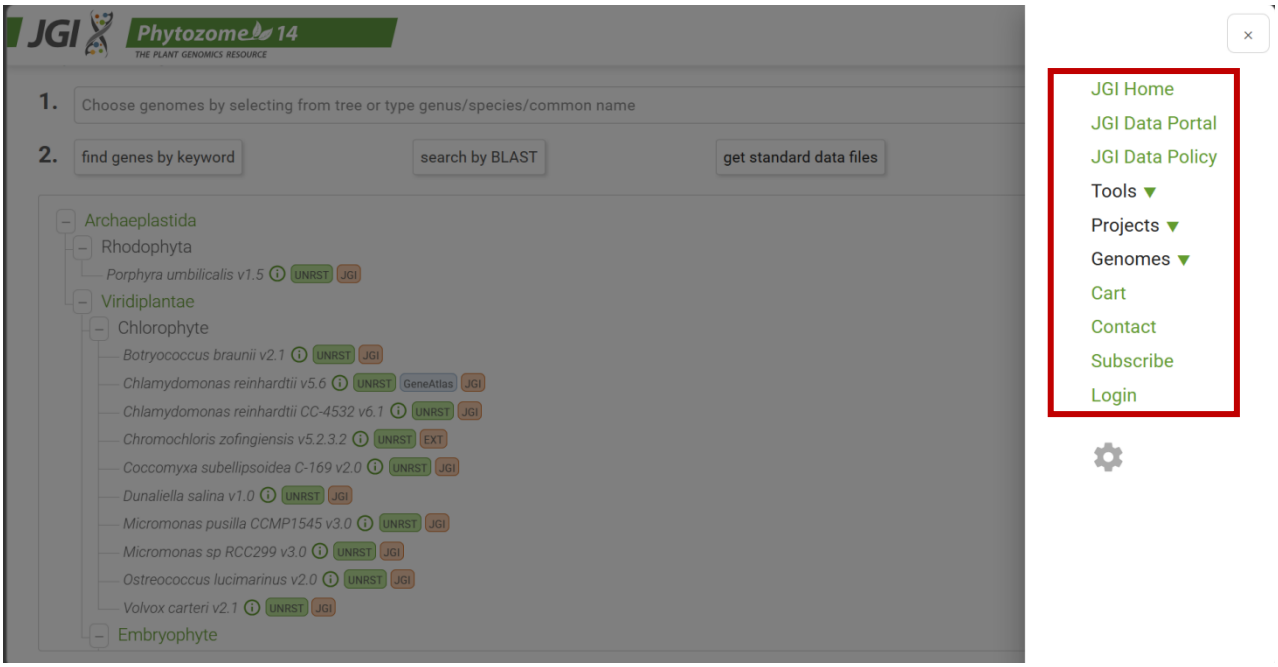
（4）数据库 PHYTOZOME 使用

https://phytozome-next.jgi.doe.gov/

PHYTOZOME——为高粱这种小众作物提供了高质量参考基因组与完善的比较基因组学工具，有效弥补了研究资源不足的问题，便于开展功能基因挖掘与进化分析。



查找特定物种：点击顶部菜单栏的 “Genomes”，可以通过分类树（如按纲目）或搜索框快速定位目标物种（如拟南芥、水稻、玉米、高粱等）。



查看基因组详情：点击物种名后，会进入基因组信息页，包含基因组统计、注释版本、文献引用信息等。

基因在 6 号染色体上的结构，包括其转录本位置以及基于 EST、BLASTx 同源比对和 RNA-Seq 覆盖度的多重证据。

该图展示了高粱基因 Sobic.006G109000 在 6 号染色体上的完整基因组序列（反向链，长度 2039 bp）。

基因与其共表达基因在相应代谢通路中协同发挥作用。

（5）UniProt 蛋白质数据库

①UniProt 数据库概况

UniProt 是国际知名蛋白质数据库，主要包括 UniProtKB 知识库、UniParc 归档库和 UniRef 参考序

列集三部分。 UniProtKB 知识库是 UniProt 的核心，除蛋白质序列数据外，还包括大量注释信息。UniProtKB 知识库分 Swiss-Prot 和 TrEMBL 两个子库。Swiss-Prot 子库中 50 多万条序列均由人工审阅和注释，而 TrEMBL 子库中 1.4 亿多条序列是 由核酸序列数据库 EMBL 中的蛋白质编码序列翻译所得，并由计算机根据一定规则进行注释。 UniParc 归档库将存放于不同 数据库中的同一个蛋白质归并到一个记录中以避免冗余，并赋予序列唯一性特定标识符。 UniRef 参考序列集按相似性程度将 UniProtKB 和 UniParc 中的序列分为 UniRef 100、UniRef 90 和 UniRef 50 三个数据集。UniProt 网站为用户提供了高效实用的 高级检索系统和大量帮助文档。UniProt 数据库每 4 周发布新版的同时也发布统计报表，用户可通过统计报表了解该数据库 的数据量及更新情况、数据类别和物种分布等基本信息，查看常规注释信息、序列特征注释信息和数据库交叉链接等统计数 据。UniProt 是目前国际上序列数据最完整、注释信息最丰富的非冗余蛋白质序列数据库，自本世纪初创建以来，为生命科学 领域提供了宝贵资源。

1) UniProt 数据库由哪三部分组成？

UniProtKB 知识库、UniParc 归档库和 UniRef 参考序列集三部分。

2) UniProtKB 知识库由哪两部分组成？

UniProtKB 知识库分 Swiss-Prot 和 TrEMBL 两个子库。

3) UniProt 统计报表（Statistics）包括哪些主要信息？

Introduction——版本简介

Taxonomic origin——分类起源

Sequence size——序列大小

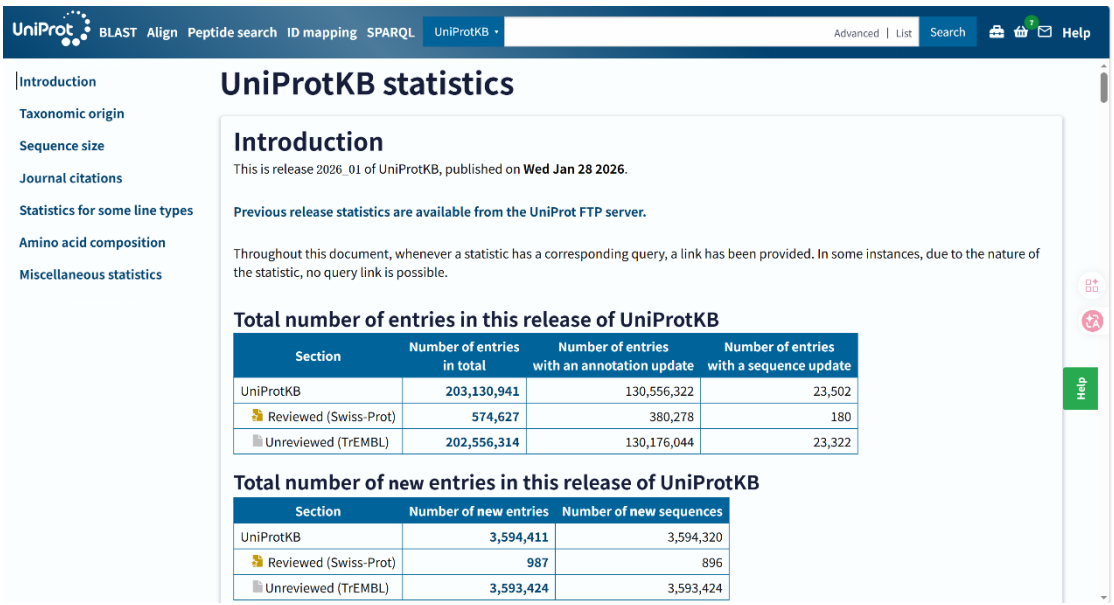
Journal citations——期刊引用

Statistics for some line types——某些通路类型的统计数据

Amino acid composition——氨基酸组成

Miscellaneous statistics——杂项数据统计

当前版本为 UniProtKB 2026-01 版本，发布于 2026 年 1 月 28 日星期三，以往数据可在 UniProt FTP 查找（网址 https://ftp.uniprot.org/pub/databases/uniprot/previous_releases/release-YYYY_NN/knowledgebase/）



4) UniProt 常规注释信息（General Annotation）包括哪些主要部分？

该类是基于整条序列的常规注释信息，如功能、表达、亚细胞定位等。

5) UniProt 序列特征注释信息（Sequence Annotation）包括哪些主要部分？

该类注释信息不是基于整条序列或整个蛋白质，而是基于序列特定区域或特定位点，因此也称序列特征注释信息。序列特征注释信息共分以下七大类：

分子加工 Molecular Processing

包含：信号肽 Signal peptide、转运肽 Transit peptide、前肽 Propeptide、N 端甲硫氨酸 N-terminal methionine 等。

序列区域 Region

包含：结构域 Domain、基序 / 模体 Motif、重复序列 Repeat、卷曲螺旋 Coiled coil、跨膜螺旋 Transmembrane helix、锌指结构 Zinc finger、DNA 结合区 DNA-binding domain、核苷酸结合区 Nucleotide-binding region、钙结合区 Calcium-binding region 等。

序列位点 Site

包含：活性位点 Active site、金属结合位点 Metal-binding site 等。

氨基酸修饰 Amino Acid Modification

包含：二硫键 Disulfide bond、糖基化 Glycosylation、脂质修饰 Lipidation、交联键 Cross-link、非标准氨基酸 Non-standard residue 等。

天然变异 Natural Variations

包含：天然突变位点、可变剪接产物 Alternative splicing 等。

实验相关信息 Experimental Information

包含：定点突变 Mutagenesis、非相邻氨基酸 Non-adjacent residues、非末端氨基酸 Non-terminal residues、序列不确定位点 Sequence uncertainty、序列冲突 Sequence conflict 等。

二级结构 Secondary Structure

包含：α 螺旋 Alpha helix、β 折叠 Beta sheet、β 转角 Beta turn。

6) UniProt 与哪几大类数据库建立了交叉链接（Cross Reference）？

- **序列数据库 Sequence Databases**

包含：NCBI 人鼠共有编码序列数据库 CCDS (Consensus Coding Sequences)、NCBI 参考序列数据库 RefSeq、EBI 核酸序列数据库 EMBL。

- **蛋白质三维结构数据库 3D Structure Databases**

包含：国际蛋白质数据库 PDB (Protein Data Bank)、EBI 蛋白质结构概览 PDBSum、蛋白质模型库 Protein Model Portal、瑞士生物信息研究所同源建模库 SMR (Swiss Model Repository)。

- **蛋白质互作数据库 Protein-protein Interaction Databases**

包含：BioGRID 互作数据库、EBI 大分子互作数据库 IntAct、复合物数据库 Complex Portal、STRING 互作网络数据库、哺乳动物复合物库 CORUM、DIP 互作数据库。

- **化学小分子数据库 Chemistry Databases**

包含：ChEMBL 生物活性小分子库、DrugBank 药物靶点库、药理学指南 Guide to Pharmacology、BindingDB 分子结合数据库。

- **特殊蛋白家族数据库 Family/Group Databases**

包含：过敏原库 Allergome、蛋白酶数据库 MEROPS、碳水化合物酶数据库 CAZy、多功能蛋白库 MoonDB、过氧化物酶库 PeroxiBase、限制性内切酶库 REBASE、转运蛋白分类库 TCDB、凝集素库 UniLectin、真菌木质纤维素蛋白库 mycoCLAP。

- **翻译后修饰数据库 PTM Databases**

包含：修饰位点整合库 iPTMnet、羰基化数据库 CarbonylDB、糖基化数据库 Glyconnec、糖生物学库 UniCarbKB、去磷酸化数据库 DEPOD。

- **多态性与突变数据库 Polymorphism Databases**

包含：NCBI 单核苷酸多态性库 dbSNP、癌症多态突变库 BioMuta。

- **双向凝胶电泳数据库 2D Gel Databases**

包含：瑞士双向电泳库 Swiss-2DPAGE、生殖相关 2D 电泳库 Reproduction-2DPAGE、都柏林大学 2D 电泳库 UCD 2D-PAGE。

- **蛋白质组数据库 Proteome Databases**

包含：EBI 蛋白组鉴定库 PRIDE、CTDB 蛋白组库、蛋白组动态百科 EPD、蛋白丰度库 PaxDB、MaxDB、肽段图谱库 PeptideAtlas、日本蛋白组库 jPOST、奥地利蛋白组库 ProMex。

- **基因组注释数据库 Genome Annotation Databases**

包含：Ensembl 基因组注释平台、UCSC 基因组浏览器、NCBI Gene 数据库、KEGG 通路基因组数据库、植物基因组库 Gramene、病原菌数据库 PATRIC、无脊椎动物媒介数据库 VectorBase。

- **物种专用数据库 Organism-specific Databases**

模式生物库：小鼠 MGI、大鼠 RGD、爪蟾 Xenbase、斑马鱼 ZFIN、果蝇 FlyBase、线虫 WormBase；拟南芥 TAIR/Araport、玉米 MaizeDB；酵母 SGD/PomBase；微生物 EcoBase、Tuberculist 等；

人类疾病库：OMIM 单基因病库、GeneCards、蛋白表达库 HPA、药理基因组 PharmGKB、疾病变异库 DisGeNET；

毒素 & 命名库：蛛毒素 ArachnoServer、芋螺毒素 ConoServer；基因命名库 HGNC/VGNC。

- **系统发育数据库 Phylogenomic Databases**

包含：Ensembl GeneTree、EBI TreeFam、直系同源库 eggNOG、OrthoDB、OMA、inParanoid。

- **酶与代谢通路数据库 Enzyme and Pathway Databases**

包含：代谢通路 Reactome、酶学数据库 BRENDA、信号网络 SIGNOR、生化动力学 SABIO-RK、KEGG 代谢通路库。

- **基因表达数据库 Gene Expression Databases**

包含：EBI ExpressionAtlas 表达图谱、组织表达库 Bgee。

- **蛋白家族与结构域数据库 Family/Domain Databases**

包含：整合注释平台 InterPro、保守域 Pfam/SMART/CDD、功能分类 PANTHER/PIRSF、结构域分类 CATH/Gene3D/SuperFamily、蛋白指纹 PRINTS、功能位点 Prosite、结构域 ProDom。

②UniProt 中的帮助文档包括哪些信息？

说明从 UniProt 数据库中检索以下序列条目的步骤和结果

1) 所有拟南芥序列

2) 已审阅拟南芥序列

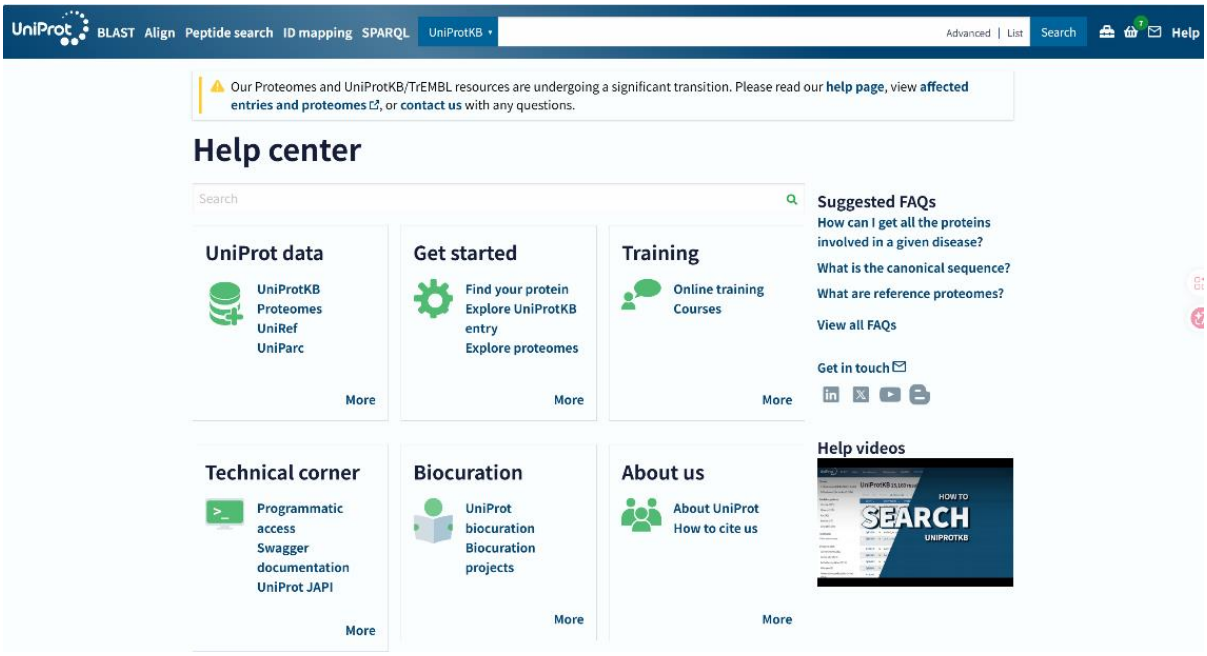
3) 已审阅拟南芥序列中具有蛋白质证据的序列

4) 已审阅拟南芥序列中具有蛋白质证据、且具有跨膜螺旋的序列

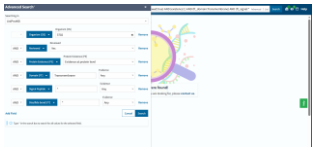
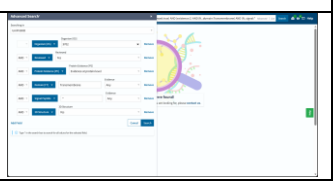
5) 已审阅拟南芥序列中具有蛋白质证据、具有跨膜螺旋和信号肽的序列

6) 已审阅拟南芥序列中具有蛋白质证据、具有跨膜螺旋和信号肽、并具有二硫键的序列

7) 已审阅拟南芥序列中具有蛋白质证据、具有跨膜螺旋、信号肽、二硫键，且已经测定三维结构的序列



步骤	检索内容	示例查询语句	网站	结果说明
1	物种（organism_id）	organism_id:3702	(organism_id:3702) in UniProtKB search (136309) | UniProt	返回数据库中所有与拟南芥（Arabidopsis thaliana，TAXID 3702）相关的序列，包括未审阅（TrEMBL）和已审阅（Swiss-Prot）部分，为136309。
2	物种 + 数据库来源（reviewed）	organism_id:3702 AND reviewed:true	organism_id:3702 AND reviewed:true in UniProtKB search (16418) | UniProt	仅返回经过人工审阅的Swiss-Prot 条目，质量更高，注释更可靠，为16418。
3	步骤 2 + 存在蛋白质水平证据（existence）	organism_id:3702 AND reviewed:true	(organism_id:3702) AND	在已审阅序列中，进一步限定为有直接实验证据

		AND existence:"evidence at protein level"	(reviewed:true) AND (existence:1) in UniProtKB search (7752) | UniProt	（如质谱、Edman 测序等）证明蛋白质存在的条目，为 7752。
4	步骤 3 + 存在跨膜螺旋区域（feature）	(organism_id:3702) AND (reviewed:true) AND (existence:1) AND (ft_domain:Transmembrane)	(organism_id:3702) AND (reviewed:true) AND (existence:1) AND (ft_domain:Transmembrane) in UniProtKB search (41) | UniProt	利用 feature 字段，筛选出注释有跨膜螺旋（Transmembrane helix）区域的序列，为 41。
5	步骤 4 + 存在信号肽（feature）	(organism_id:3702) AND (reviewed:true) AND (existence:1) AND (ft_domain:Transmembrane) AND (ft_signal:*)	(organism_id:3702) AND (reviewed:true) AND (existence:1) AND (ft_domain:Transmembrane) AND (ft_signal:*) in UniProtKB search (1) | UniProt	在具有跨膜螺旋的基础上，增加存在信号肽（Signal peptide）的限定，为 1。
6	步骤 5 + 存在二硫键（feature）			
7	步骤 6 + 存在三维结构信息			

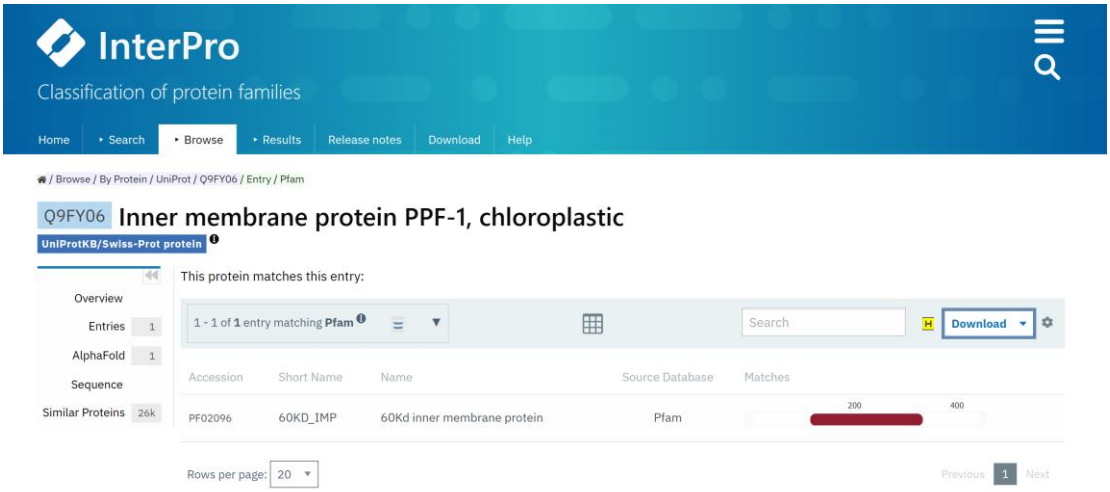
③豌豆内膜蛋白注释信息

1) 以豌豆内膜蛋白 PPF1_PEA 为例，说明该序列条目包括哪些注释信息。

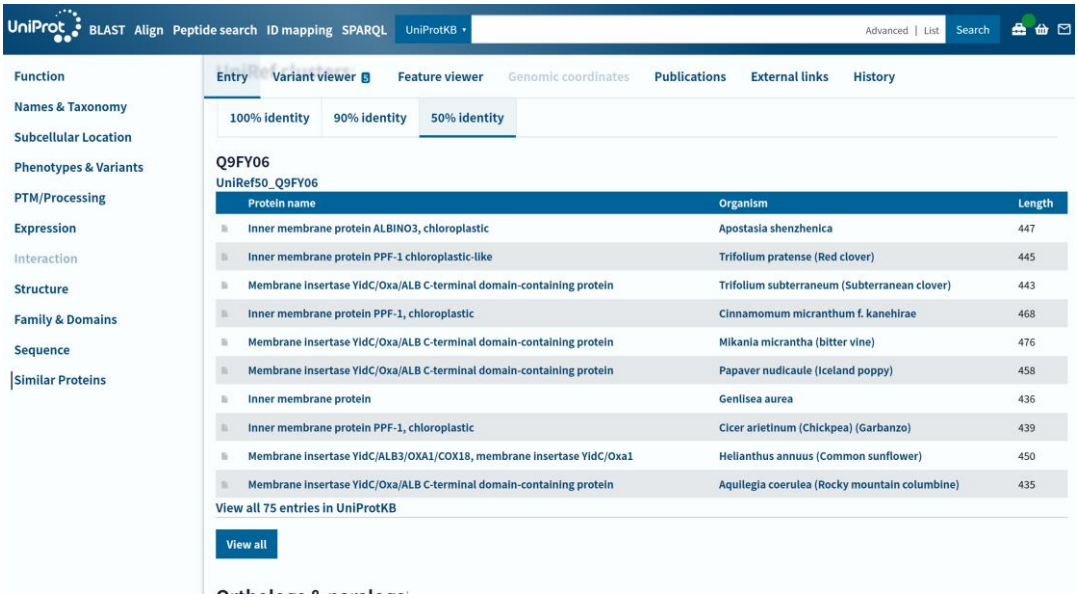
PPF-1 - Inner membrane protein PPF-1, chloroplastic - Pisum sativum (Garden pea) | Variants viewer | UniProtKB | UniProt

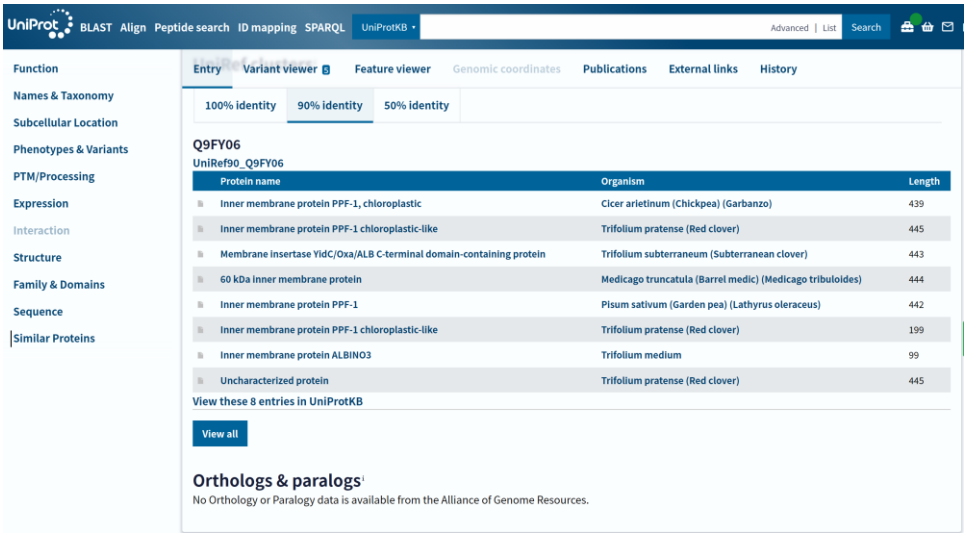
2) 通过注释信息或高级检索，查找拟南芥中与 PPF1_PEA 属于同一家族的内膜蛋白。

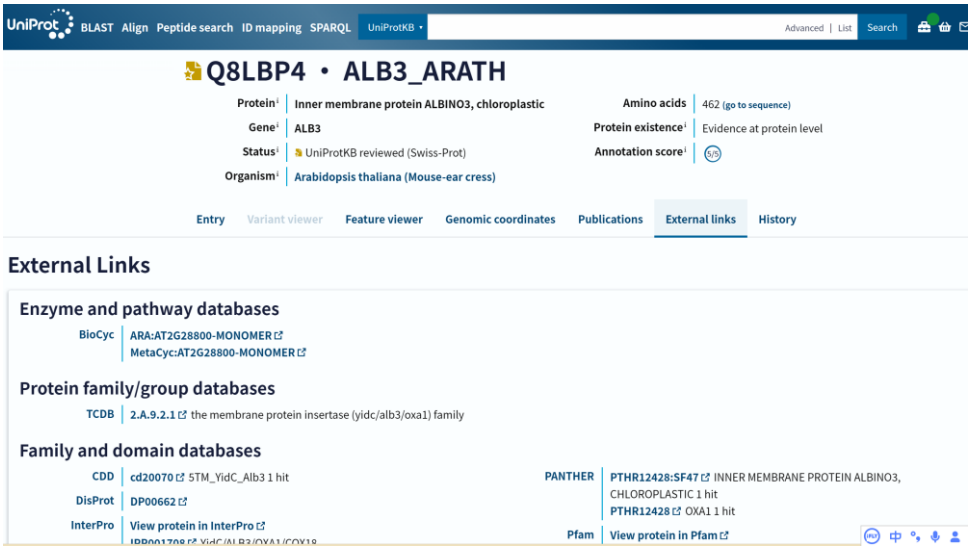
Inner membrane protein PPF-1, chloroplastic (Q9FY06) - protein - InterPro



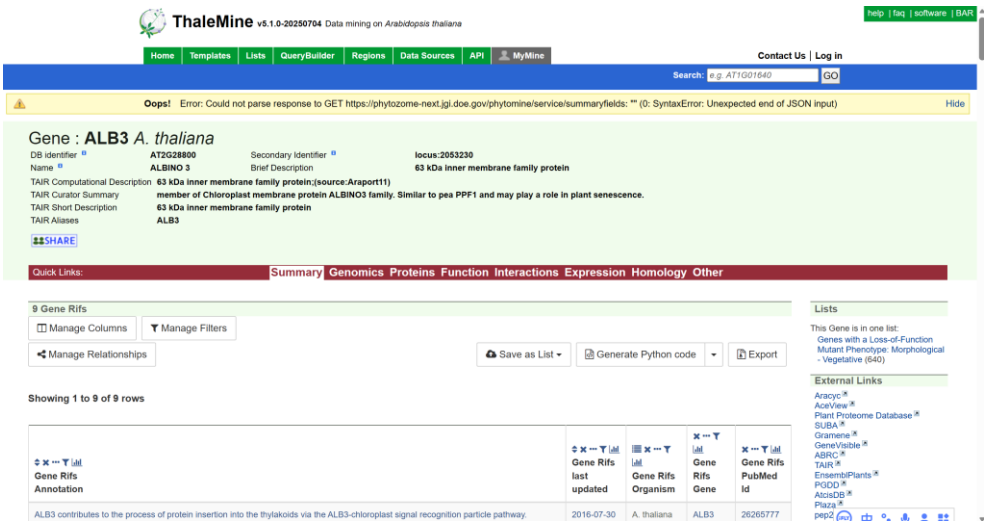
3) 通过查看注释信息和多序列比对，找出拟南芥中 PPF1_PEA 的直系同源蛋白 ALB3_ARATH。

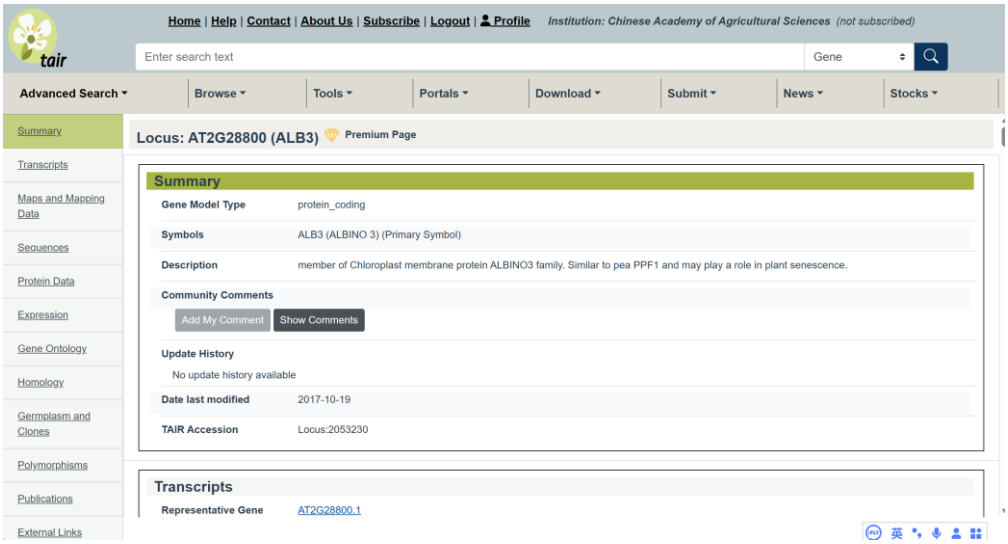






4) 查看 ALB3_ARATH 的注释信息，特别是拟南芥专门数据库 AraPort 和 TAIR，并与 PPF1_PEA 的注释信息进行比较， 说明如何将模式生物研究结果用于非模式生物。





（1）通过直系同源关系进行功能推断

方法：利用 UniProt、OrthoDB 或 TAIR 提供的直系同源映射，将拟南芥中已知的 ALB3 功能直接转移给豌豆的 PPF1。

依据：保守的结构域（Pfam）、相同的亚细胞定位、以及已验证的相互作用网络（如与 cpSRP 的互作）提供了强有力的功能同源性证据。

（2）利用模式生物的突变体表型预测非模式生物的表型

拟南芥 alb3 突变体表现为白化、幼苗致死。在豌豆中，通过化学诱变或基因沉默获得的 PPF1 功能缺失突变体也表现出类似的叶绿体发育缺陷。

应用：当非模式生物（如大豆、玉米）中发现与 ALB3 直系的基因时，可推测其功能缺失将导致叶绿体发育异常，从而指导表型观察和实验设计。

（3）借用模式生物的蛋白质相互作用网络

拟南芥中 ALB3 与 cpSRP43、cpSRP54 的互作已通过多种实验验证（Co-IP、酵母双杂交）。

在豌豆中，可通过同源预测，假设 PPF1 同样与 cpSRP 互作，并直接使用拟南芥的抗体或互作实验方案进行验证。

（4）利用模式生物的调控网络和表达模式

通过 TAIR 或 AraPort 的共表达网络，可以发现与 ALB3 协同表达的基因（如叶绿体生物发生相关基因）。对于非模式生物，可将其直系同源基因的表达数据（如 RNA-seq）与拟南芥的共表达模式进行对比，推断其参与的生物学过程。

（5）借助模式生物的结构信息

拟南芥 ALB3 虽然尚未有晶体结构，但其同源蛋白（如细菌 YidC、酵母 Oxa1）的结构已解析。通过同源建模，可以预测非模式生物中 PPF1 的三维结构，并推测其关键活性位点。

（6）课题相关物种信息

①拟南芥的物种信息

1) 该物种的中文名、英文名、拉丁文学名、分类学登录号。

拟南芥——Arabidopsis——Arabidopsis thaliana——3702

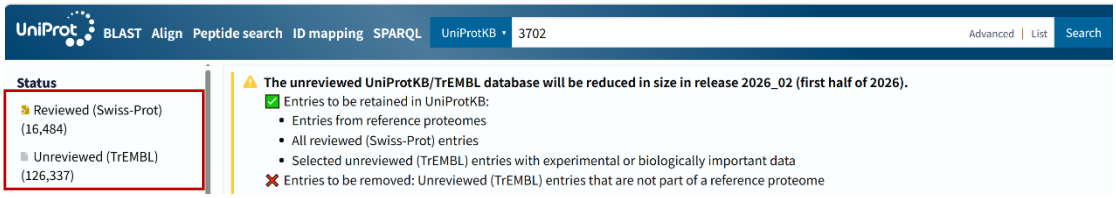
2) 该物种的分类学地位（界、门、纲、目、科、属、种）。

3) 该物种在 Swiss-Prot 和 TrEMBL 子库中序列条目数。

截至 2026.04——

该物种在 Swiss-Prot 中经人工注释序列条目 16484 条；

该物种在 TrEMBL 中计算机注释序列条目 126337 条；

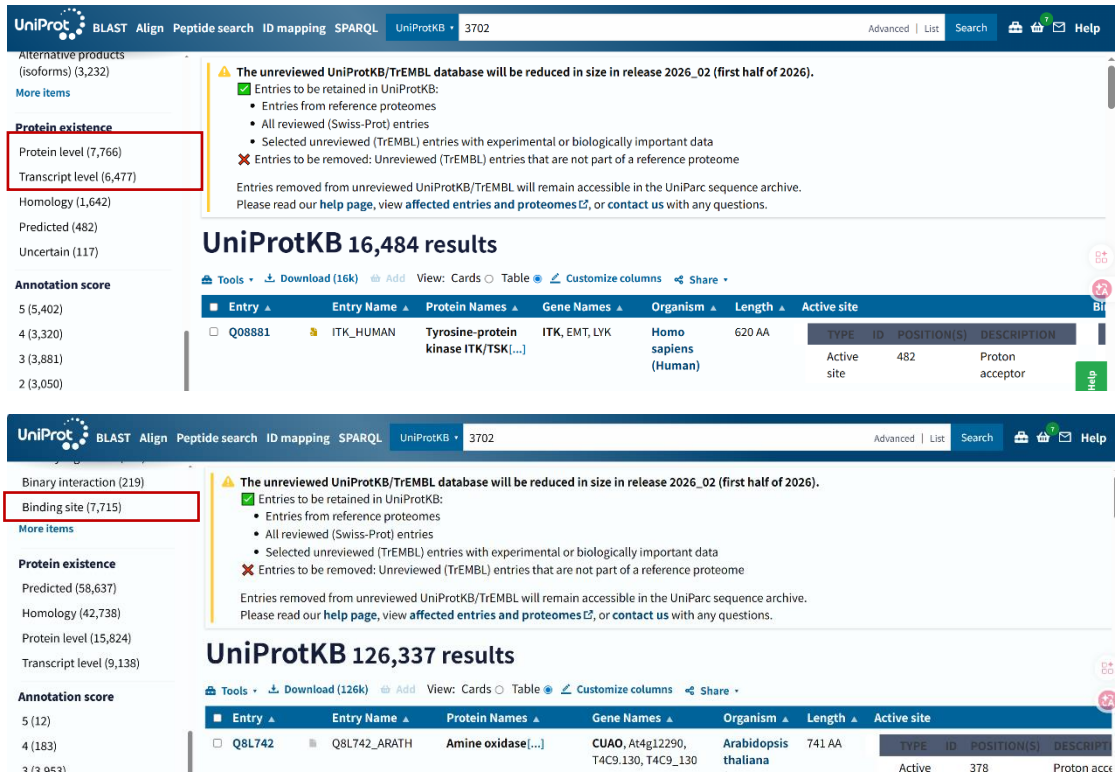


4) 该物种在 Swiss-Prot 和 TrEMBL 子库中具有蛋白质水平证据的序列条目数。

截至 2026.04——

该物种在 Swiss-Prot 子库中具有蛋白质水平证据的序列条目数 7766 条；

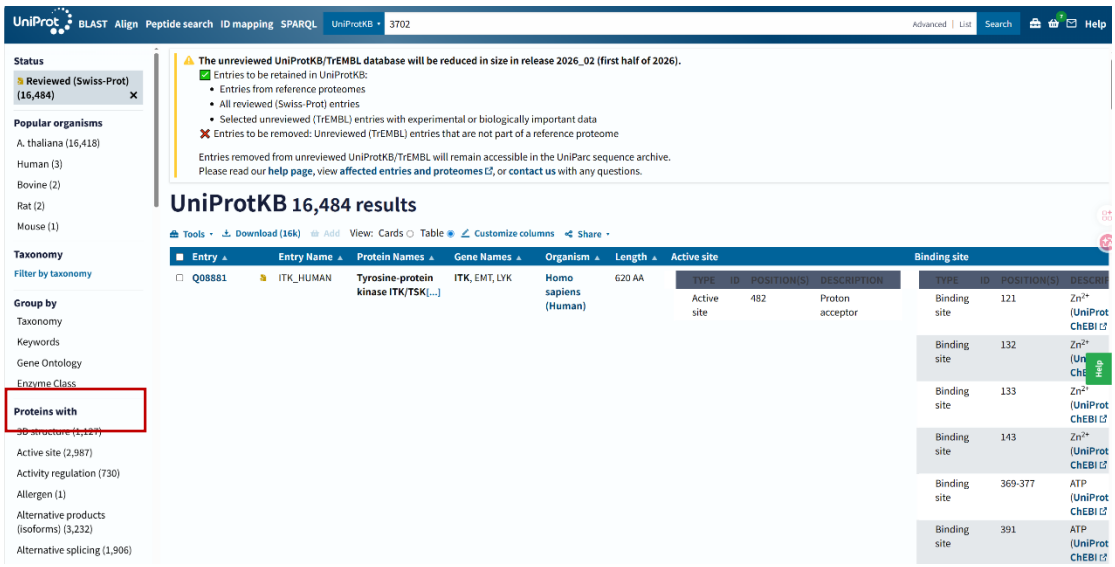
该物种在 TrEMBL 子库中具有蛋白质水平证据的序列条目数 15824 条；



5) **该物种在 Swiss-Prot 子库中具有三维空间结构的序列条目数。**

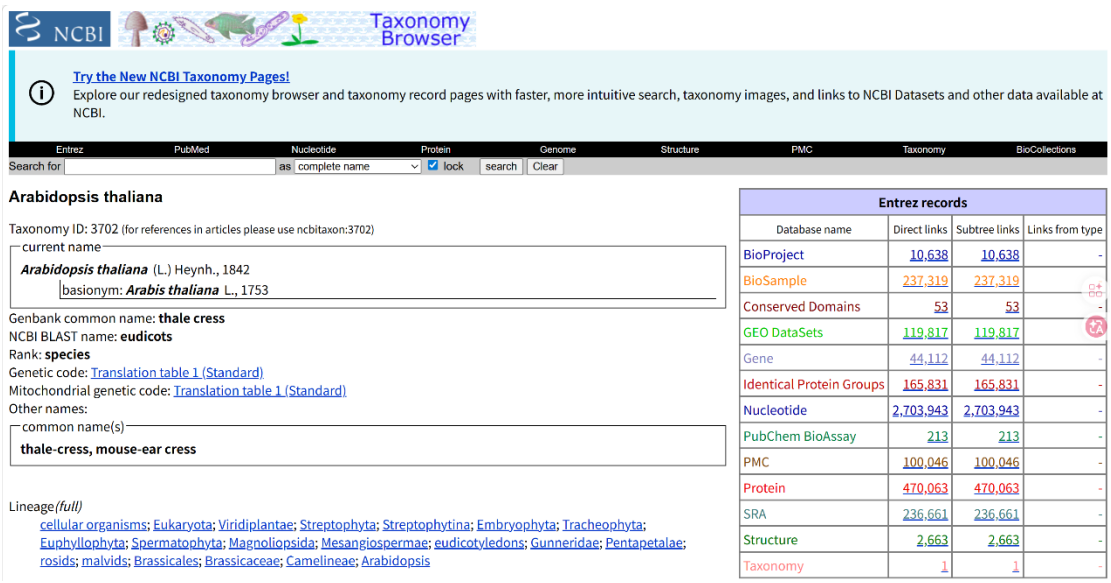
截至 2026.04——

该物种在 Swiss-Prot 子库中具有三维空间结构的序列条目数 1127 条；



6) **查阅该物种在 NCBI 分类学网站中与其它数据库的交叉链接，列表说明其基本信息。**

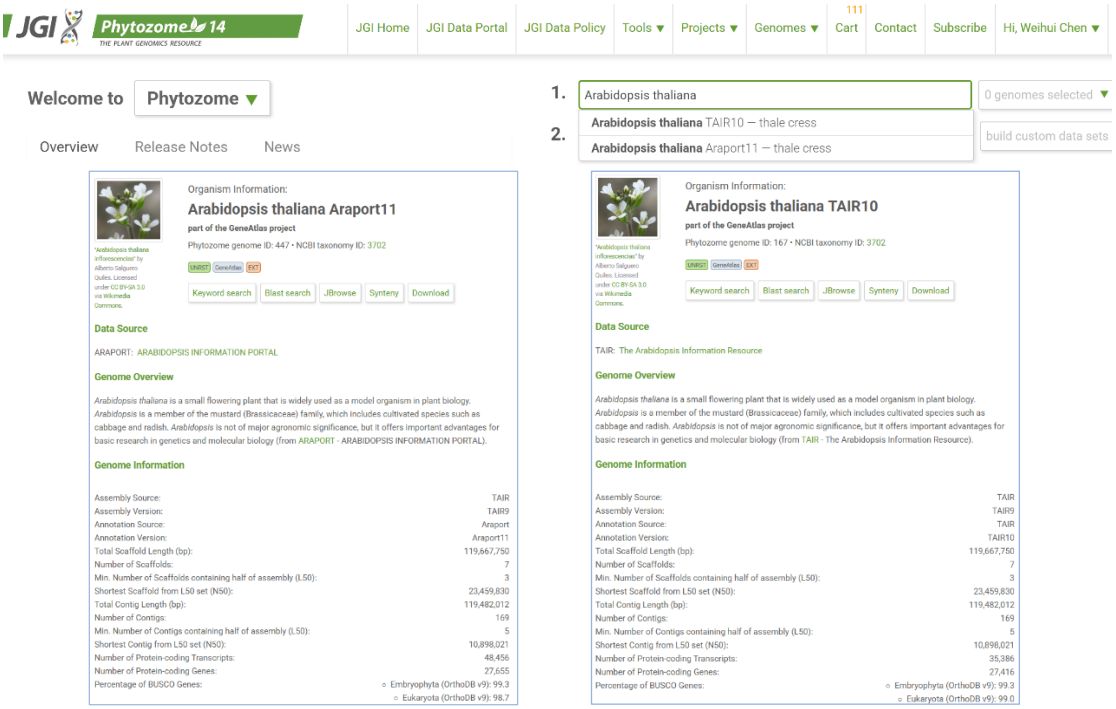
步骤：打开 NCBI → 点选 Resource list 下的 Taxonomy → 输入**学名 / 俗名** → 点**斜体学名** → 看 **Lineage** 和 **TaxID** → 一键跳序列 / 基因 / 论文



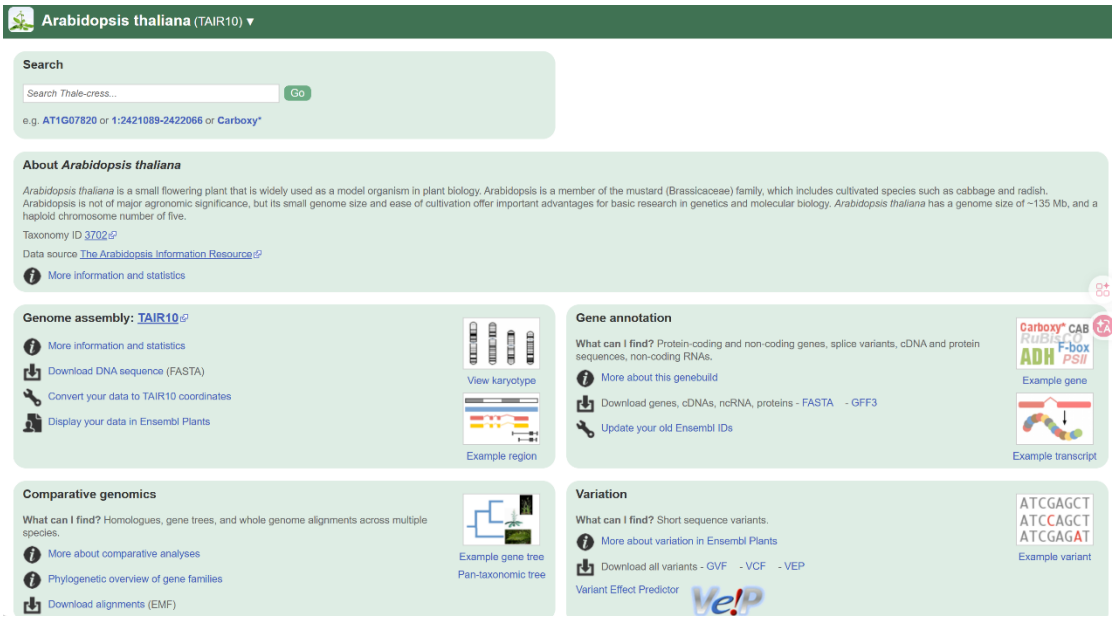
如图所示，列出了拟南芥在 13 个 NCBI 子库（Nucleotide、Protein、Gene、GEO、SRA、PMC 等）的条目数，Direct links 与 Subtree links 数值一致，说明数据均直接对应拟南芥本身，无额外亚种 / 变种数据。其中，核酸序列≈270 万条、蛋白序列≈47 万条、基因注释≈4.4 万条、表达数据≈ GEO/SRA 合计近 36 万条、相关文献≈10 万篇；

7) 查阅 Ensembl 或 Phytozome 等基因组数据库，若该物种已完成基因组测序，熟悉其基因组基本信息；若该物种尚未测序， 找出与其亲缘关系最近的物种，熟悉其基本信息。

Phytozome 中收录了拟南芥最新的两版基因组数据，Araport11 版与 TAIR10 版基本一致；



Ensembl 中仅收录了拟南芥基因组 TAIR10 版



8) 搜索 Database Common，找出该物种相关数据库，熟悉相关数据库的基本信息和已发表论文。

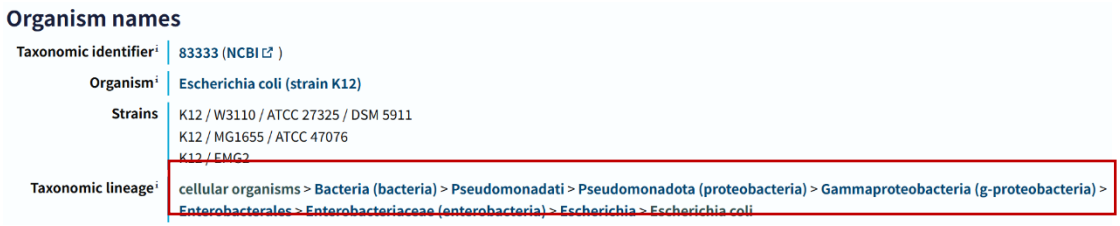
拟南芥权威数据库：TAIR、Araport。

②大肠杆菌

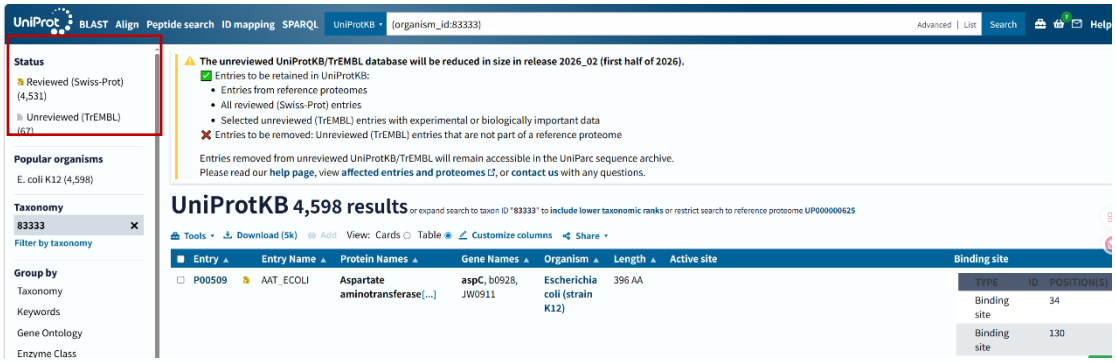
1) 该物种的中文名、英文名、拉丁文学名、分类学登录号。

大肠杆菌（通用名）、大肠埃希氏菌（学名）——Escherichia coli——**Escherichia coli（E.coli）**——83333

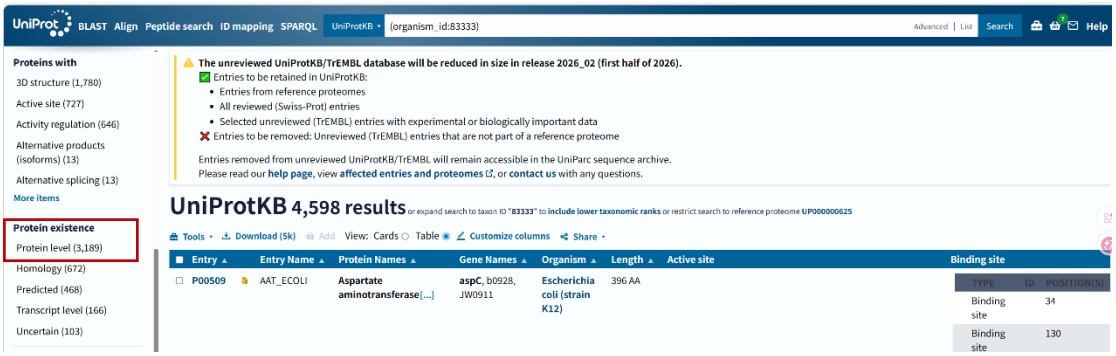
2) 该物种的分类学地位（界、门、纲、目、科、属、种）。



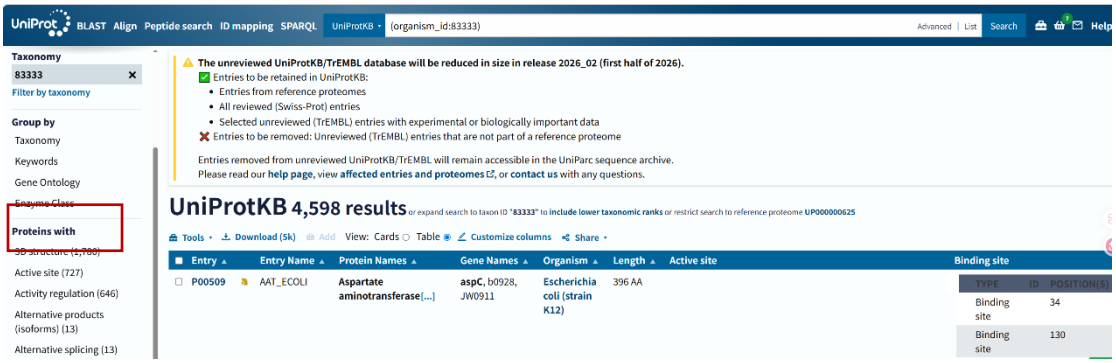
3) 该物种在 Swiss-Prot 和 TrEMBL 子库中序列条目数。



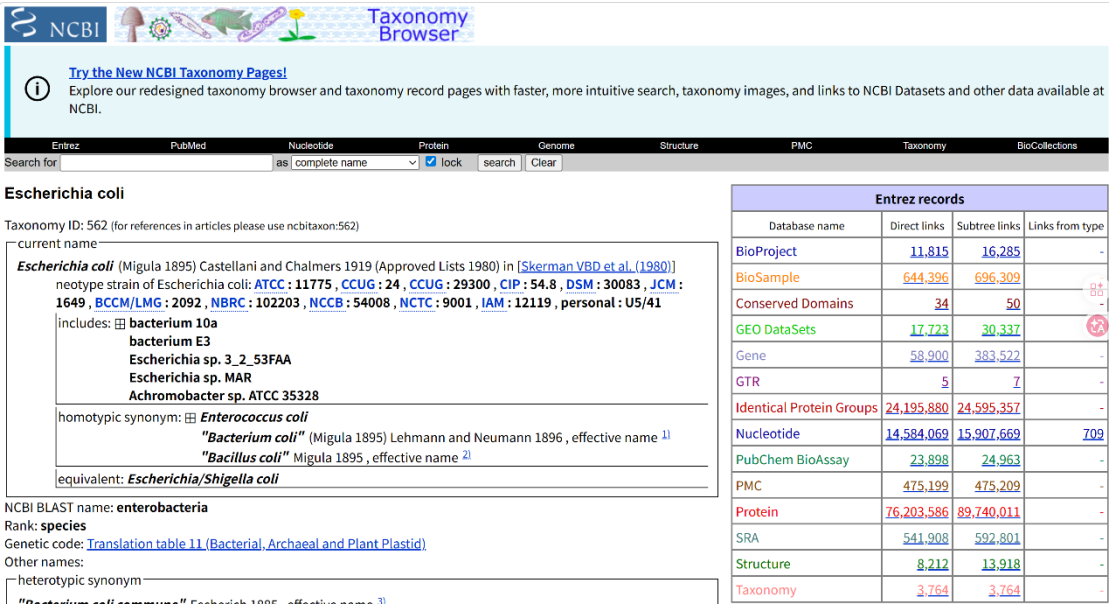
4) 该物种在 Swiss-Prot 和 TrEMBL 子库中具有蛋白质水平证据的序列条目数。



5) 该物种在 Swiss-Prot 子库中具有三维空间结构的序列条目数。



6) 查阅该物种在 NCBI 分类学网站中与其它数据库的交叉链接，列表说明其基本信息。



这张表展示了大肠杆菌在 NCBI 各数据库中的数据条目数，Dieect links 仅统计属于「大肠杆菌（*Escherichia coli*）」这个物种的直接数据条目,Subtree links 统计大肠杆菌及其所有亚种、菌株（如 K-12、BL21、O157:H7 等）的全部数据，范围更广，所以数值普遍比 Direct links 大，数据整体体现了大肠杆菌作为研究最广泛的模式生物之一，拥有海量的序列、功能、文献和组学数据资源。

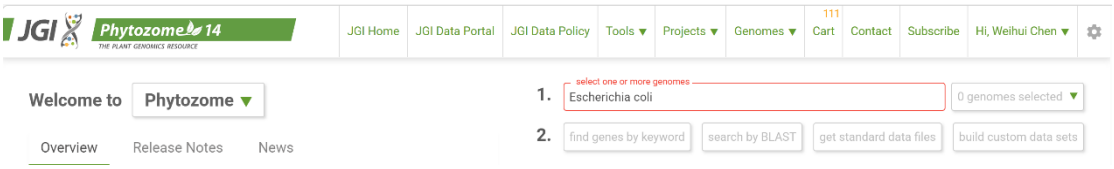
核酸序列、蛋白序列数据量极大，分别超 1458 万和 7620 万条；

相关文献（PMC）超 47.5 万篇，说明研究热度极高；

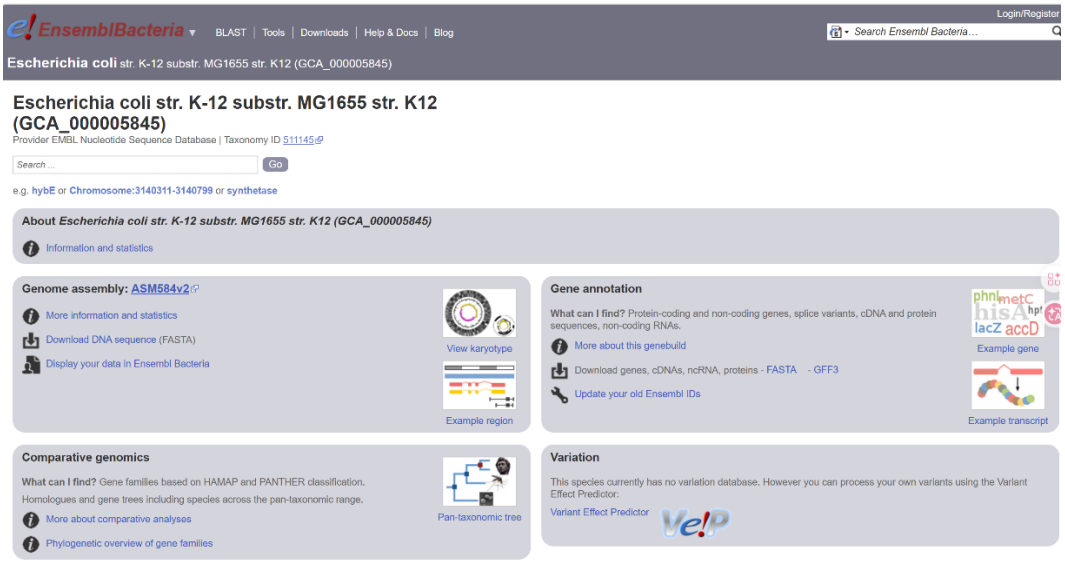
基因（Gene）、保守结构域、蛋白结构、表达数据（GEO/SRA）也覆盖全面。

7) 查阅 Ensembl 或 Phytozome 等基因组数据库，若该物种已完成基因组测序，熟悉其基因组基本信息；若该物种尚未测序， 找出与其亲缘关系最近的物种，熟悉其基本信息。

如图、Phytozome 数据库查询不到大肠杆菌的基因组信息



Ensembl Bacteria



8) 搜索 Database Common，找出该物种相关数据库，熟悉相关数据库的基本信息和已发表论文。

大肠杆菌专用注释 / 基因库——**EcoGene（K-12 精注释）**

http://bmb.med.miami.edu/EcoGene/EcoWeb/

大肠杆菌专属代谢——**EcoCyc（E. coli 百科全书）**

https://ecocyc.org/

大肠杆菌专属通路——**KEGG（eco: E. coli K-12）**

https://www.genome.jp/kegg-bin/show_organism?org=eco

大肠杆菌专属调控——**RegulonDB（转录调控、操纵子）**

https://regulondb.ccg.unam.mx/

PATRIC（病原菌基因组）

https://www.patricbrc.org/view/Genome/511145.12

EnteroBase（肠杆菌科菌株分型）

https://enterobase.warwick.ac.uk/species/ecoli

③高粱

1) 中文名：高粱

英文名：Sorghum

拉丁学名：*Sorghum bicolor* (L.) Moench

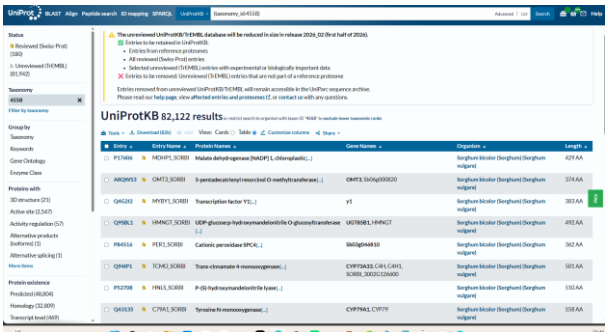
分类学登录号（NCBI Taxonomy ID）：4558

2)分类学地位：

cellular organisms > Eukaryota (eukaryotes) > Viridiplantae > Streptophyta > Streptophytina > Embryophyta (land plants) > Tracheophyta > Euphyllophyta > Spermatophyta > Magnoliopsida (flowering plants) > Mesangiospermae > Liliopsida (monocots) > Petrosaviidae > commelinids > Poales > Poaceae > PACMAD clade > Panicoideae > Andropogonodae > Andropogoneae > Sorghinae > Sorghum

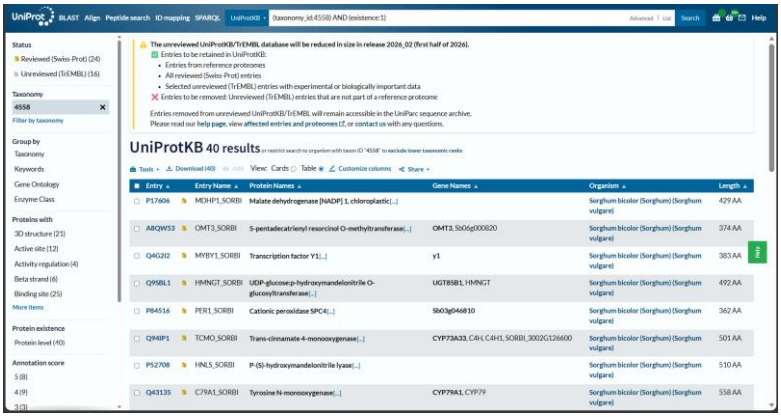
3)该物种在 Swiss-Prot 和 TrEMBL 子库中序列条目数。

(taxonomy_id:4558) in UniProtKB search (82122) | UniProt



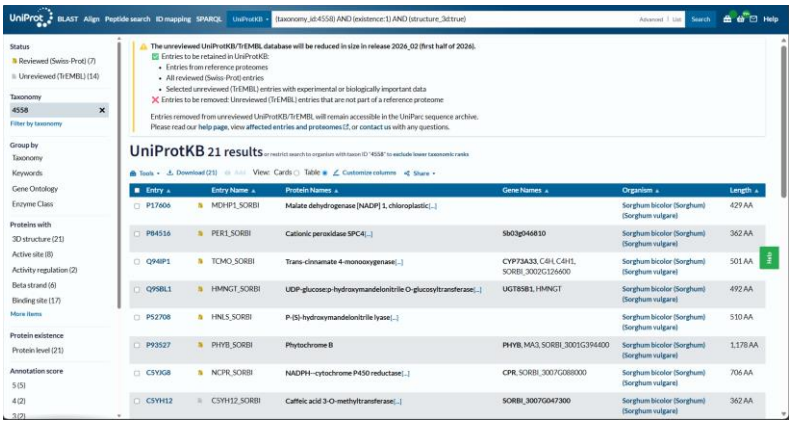
4)该物种在 Swiss-Prot 和 TrEMBL 子库中具有蛋白质水平证据的序列条目数。

(taxonomy_id:4558) AND (existence:1) in UniProtKB search (40) | UniProt



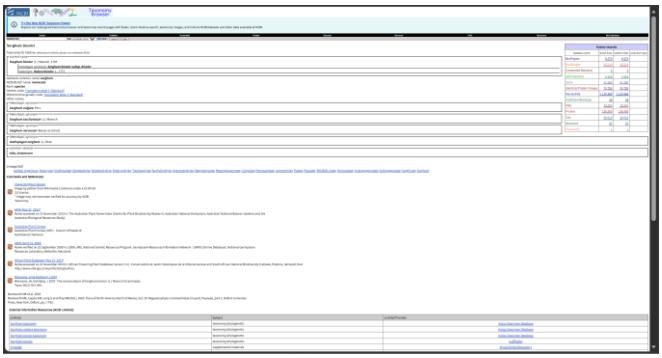
5)该物种在 Swiss-Prot 子库中具有三维空间结构的序列条目数。

(taxonomy_id:4558) AND (existence:1) AND (structure_3d:true) in UniProtKB search (21) | UniProt



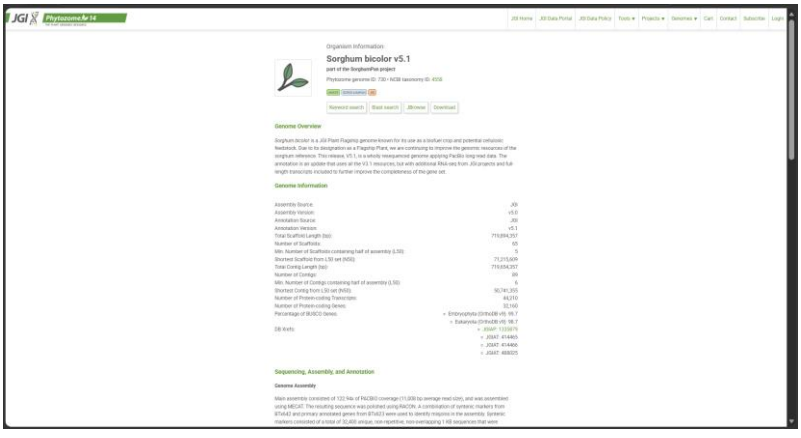
6)查阅该物种在 NCBI 分类学网站中与其它数据库的交叉链接，列表说明其基本信息。

Taxonomy browser Taxonomy Browser (Sorghum bicolor)



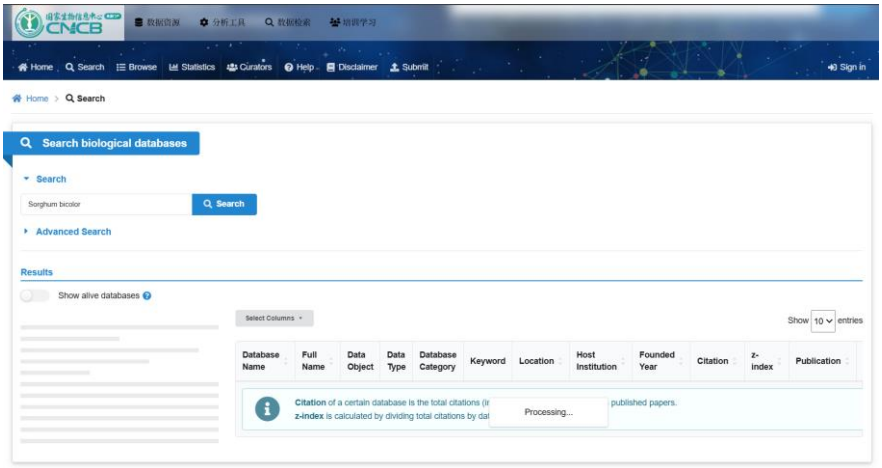
7)查阅 Ensembl 或 Phytozome 等基因组数据库，若该物种已经完成基因组测序，熟悉其基因组基本信息；若该物种尚未测序， 找出与其亲缘关系最近的物种，熟悉其基本信息。

S.bicolor v5.1: Phytozome



8) 搜索 Database Common，找出该物种相关数据库，熟悉相关数据库的基本信息和已发表论文。

Search - Database Commons



④粳稻

1) 中文名：粳稻

英文名：Japanese rice

拉丁学名：Oryza sativa subsp. japonica Kato

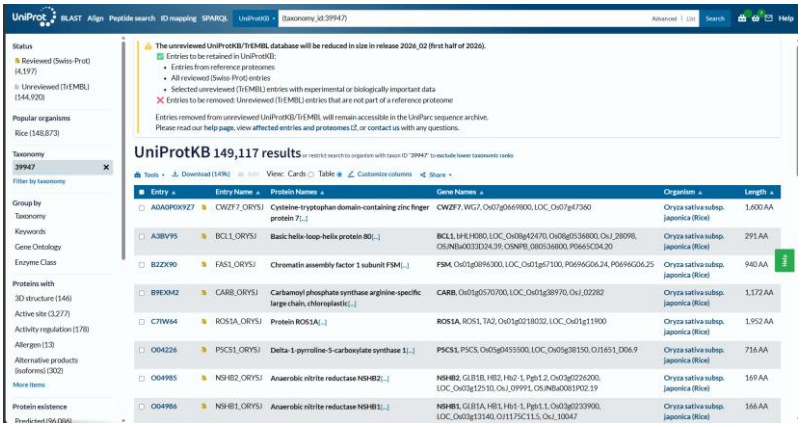
分类学登录号（NCBI Taxonomy ID）：39947

2)分类学地位：

cellular organisms > Eukaryota (eukaryotes) > Viridiplantae > Streptophyta > Streptophytina > Embryophyta (land plants) > Tracheophyta > Euphyllophyta > Spermatophyta > Magnoliopsida (flowering plants) > Mesangiospermae > Liliopsida (monocots) > Petrosaviidae > commelinids > Poales > Poaceae > BOP clade > Oryzoideae > Oryzeae > Oryzinae > Oryza

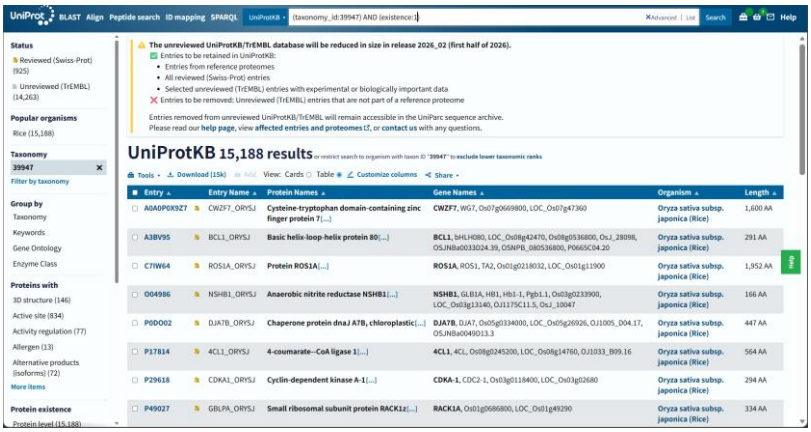
3) 该物种在 Swiss-Prot 和 TrEMBL 子库中序列条目数。

(taxonomy_id:39947) in UniProtKB search (149117) | UniProt



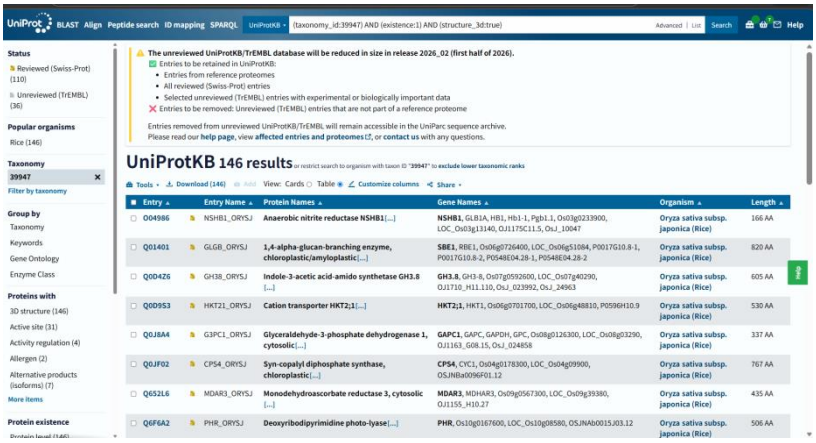
4) 该物种在 Swiss-Prot 和 TrEMBL 子库中具有蛋白质水平证据的序列条目数。

(taxonomy_id:39947) AND (existence:1) in UniProtKB search (15188) | UniProt



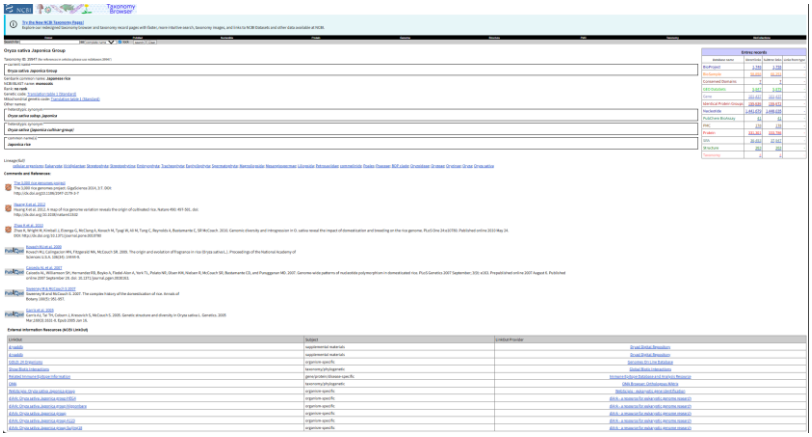
5) 该物种在 Swiss-Prot 子库中具有三维空间结构的序列条目数。

[(taxonomy_id:39947) AND (existence:1) AND (structure_3d:true) in UniProtKB search (146) | UniProt]



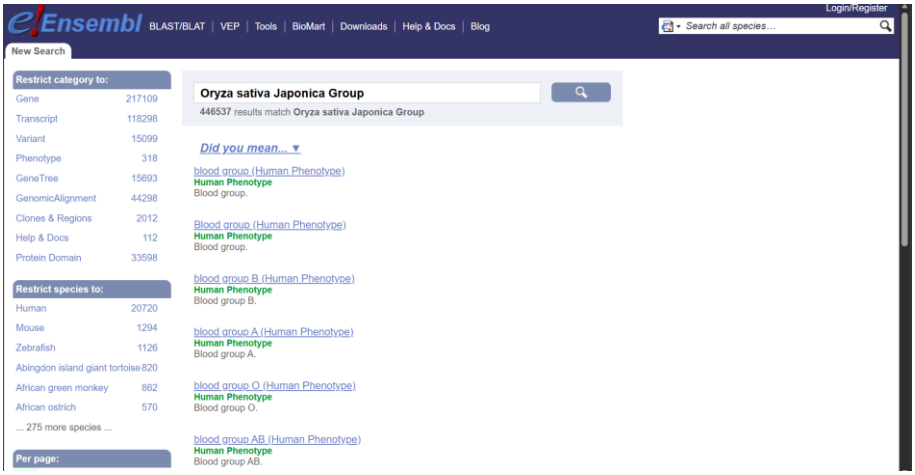
6) 查阅该物种在 NCBI 分类学网站中与其它数据库的交叉链接，列表说明其基本信息。

[Taxonomy browser Taxonomy Browser (Oryza sativa Japonica Group)]



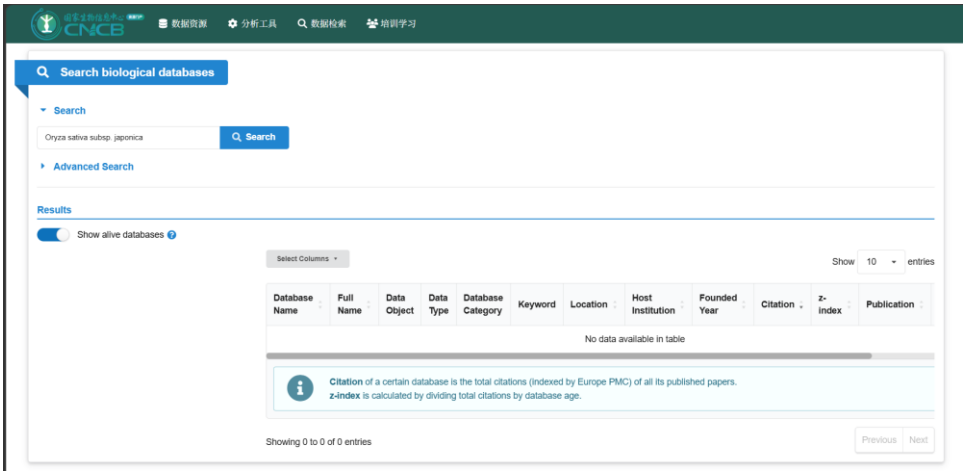
7) 查阅 Ensembl 或 Phytozome 等基因组数据库，若该物种已经完成基因组测序，熟悉其基因组基本信息；若该物种尚未测序， 找出与其亲缘关系最近的物种，熟悉其基本信息。

[Oryza sativa Japonic... - Search - Homo_sapiens - Ensembl genome browser 115]



8) 搜索 Database Common，找出该物种相关数据库，熟悉相关数据库的基本信息和已发表论文。

[Search - Database Commons]



⑤茶

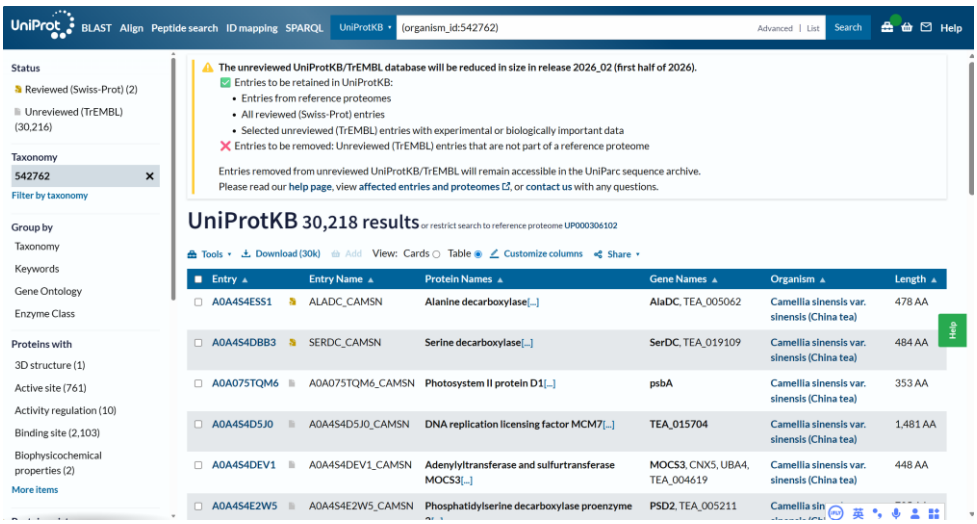
1) 该物种的中文名、英文名、拉丁文学名、分类学登录号。

茶，Chinese tea，Camellia sinensis var. sinensis ，Taxon ID：542762

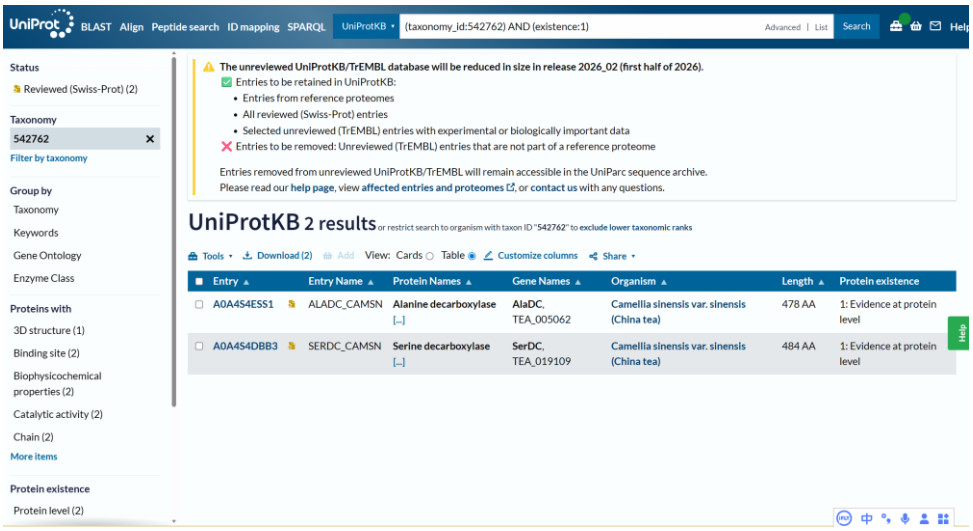
2) 该物种的分类学地位（界、门、纲、目、科、属、种）。

cellular organisms > Eukaryota (eukaryotes) > Viridiplantae > Streptophyta > Streptophytina > Embryophyta (land plants) > Tracheophyta > Euphyllophyta > Spermatophyta > Magnoliopsida (flowering plants) > Mesangiospermae > eudicotyledons > Gunneridae > Pentapetalae > asterids > Ericales > Theaceae > Camellia > Camellia sinensis (Tea plant) (Thea sinensis)

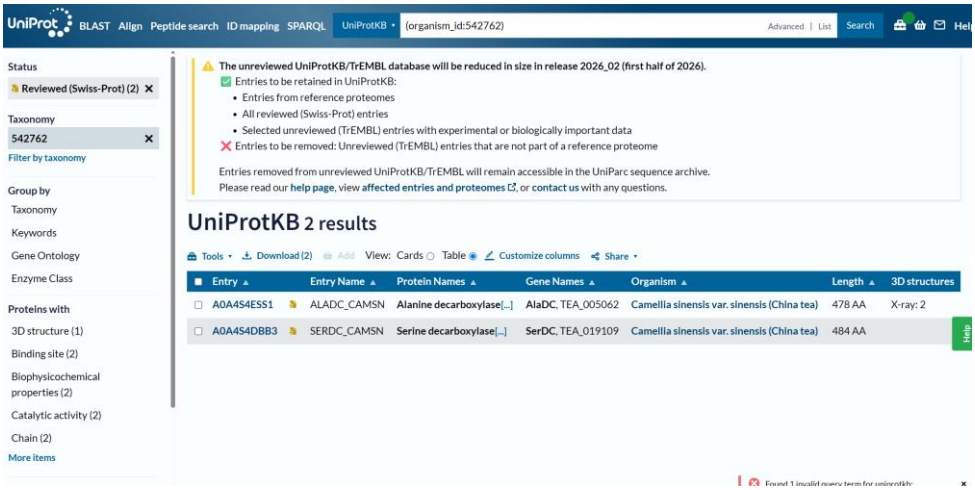
3) 该物种在 Swiss-Prot 和 TrEMBL 子库中序列条目数。



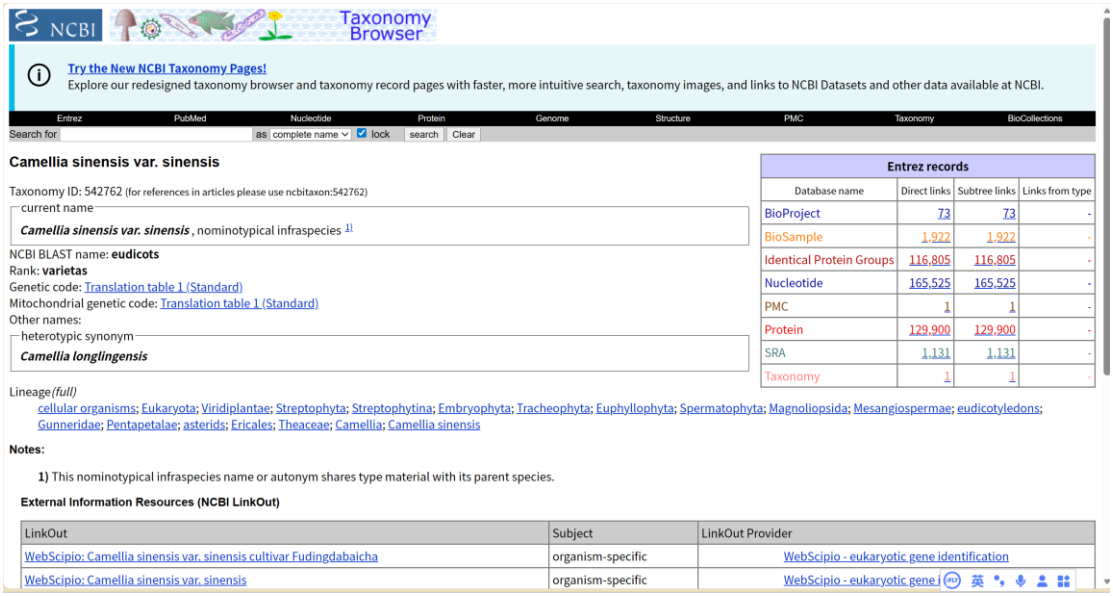
4) 该物种在 Swiss-Prot 和 TrEMBL 子库中具有蛋白质水平证据的序列条目数。



5) 该物种在 Swiss-Prot 子库中具有三维空间结构的序列条目数。



6) 查阅该物种在 NCBI 分类学网站中与其它数据库的交叉链接，列表说明其基本信息。



7) 查阅 Ensembl 或 Phytozome 等基因组数据库，若该物种已经完成基因组测序，熟悉其基因组基本信息；若该物种尚未测序， 找出与其亲缘关系最近的物种，熟悉其基本信息。

C.sativa_Acsn-226 v1.1: Phytozome

8) 搜索 Database Common，找出该物种相关数据库，熟悉相关数据库的基本信息和已发表论文。

Search - Database Commons

⑥美国黑杨

1) 该物种的中文名、英文名、拉丁文学名、分类学登录号。

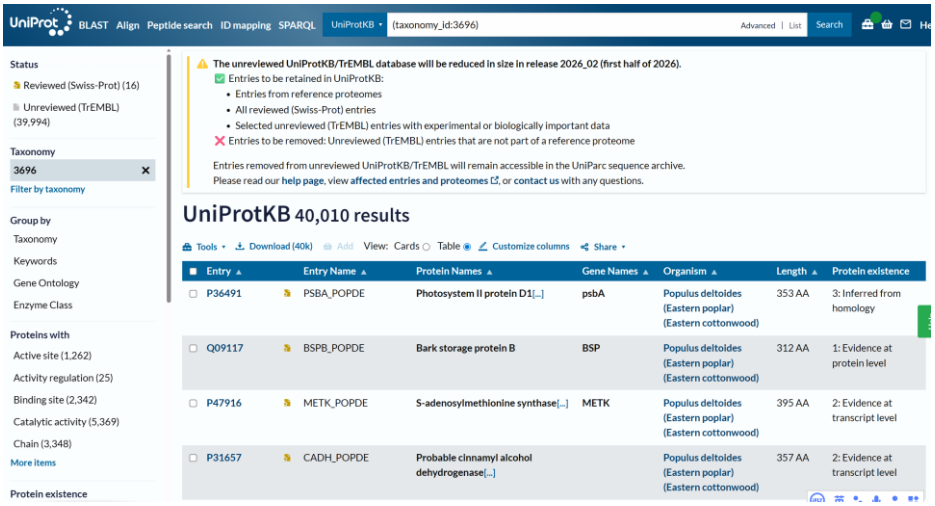
美洲黑杨，Populus，Populus deltoides，3696

2) 该物种的分类学地位（界、门、纲、目、科、属、种）。

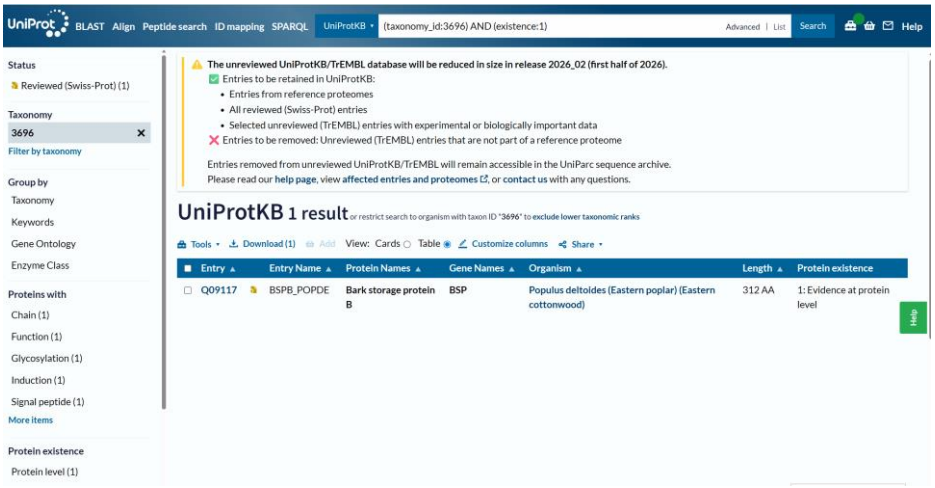
cellular organisms > Eukaryota (eukaryotes) > Viridiplantae > Streptophyta > Streptophytina > Embryophyta (land plants) > Tracheophyta > Euphyllophyta > Spermatophyta > Magnoliopsida (flowering

plants) > Mesangiospermae > eudicotyledons > Gunneridae > Pentapetalae > rosids > fabids > Malpighiales > Salicaceae > Saliceae > Populus (poplars)

3)该物种在 Swiss-Prot 和 TrEMBL 子库中序列条目数。

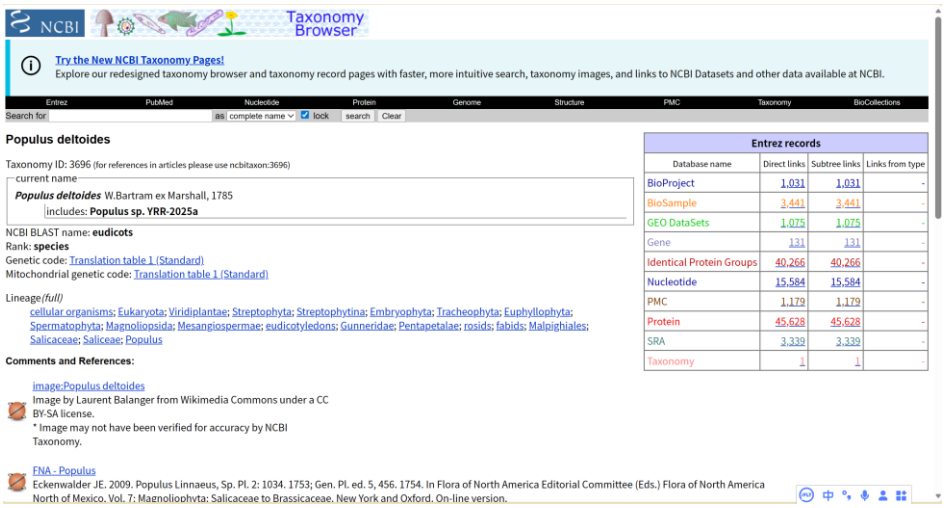


3) 该物种在 Swiss-Prot 和 TrEMBL 子库中具有蛋白质水平证据的序列条目数。



4) 该物种在 Swiss-Prot 子库中具有三维空间结构的序列条目数。

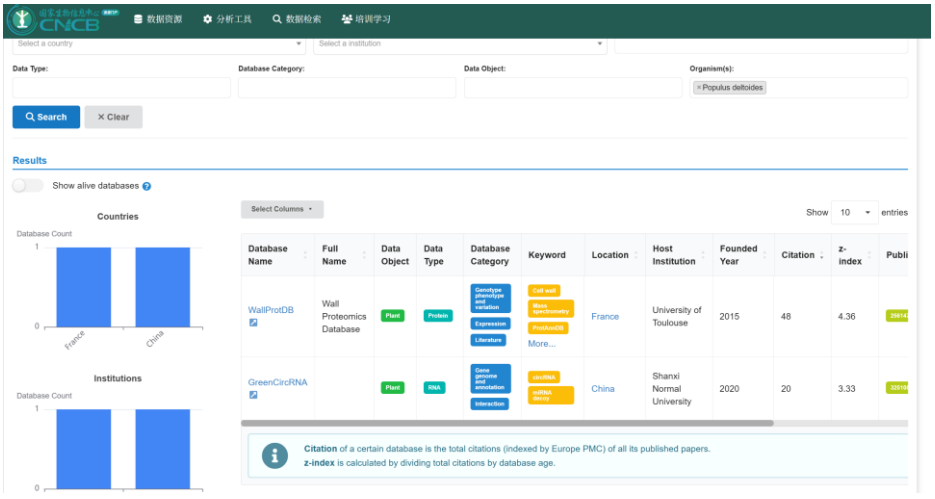
5) 查阅该物种在 NCBI 分类学网站中与其它数据库的交叉链接，列表说明其基本信息。

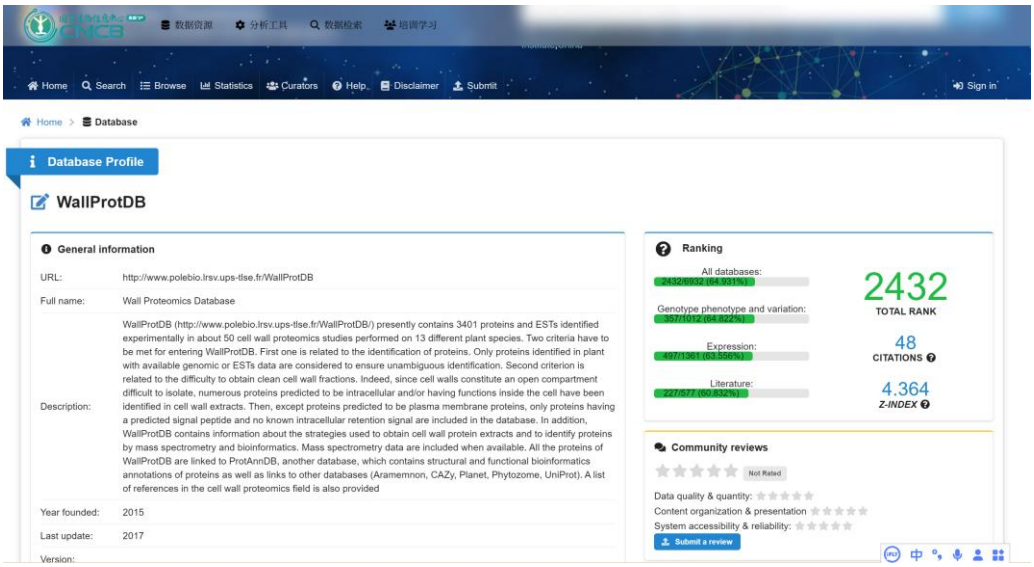


6) 查阅 Ensembl 或 Phytozome 等基因组数据库，若该物种已经完成基因组测序，熟悉其基因组基本信息；若该物种尚未测序， 找出与其亲缘关系最近的物种，熟悉其基本信息。

P.deltoidesWV94 v2.1: Phytozome

8) 搜索 Database Common，找出该物种相关数据库，熟悉相关数据库的基本信息和已发表论文。



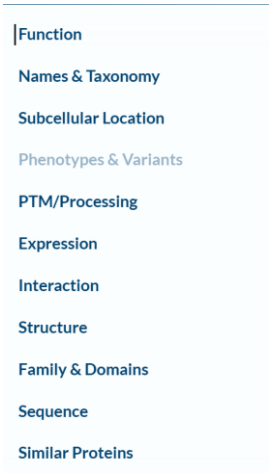


（7）课题相关蛋白信息

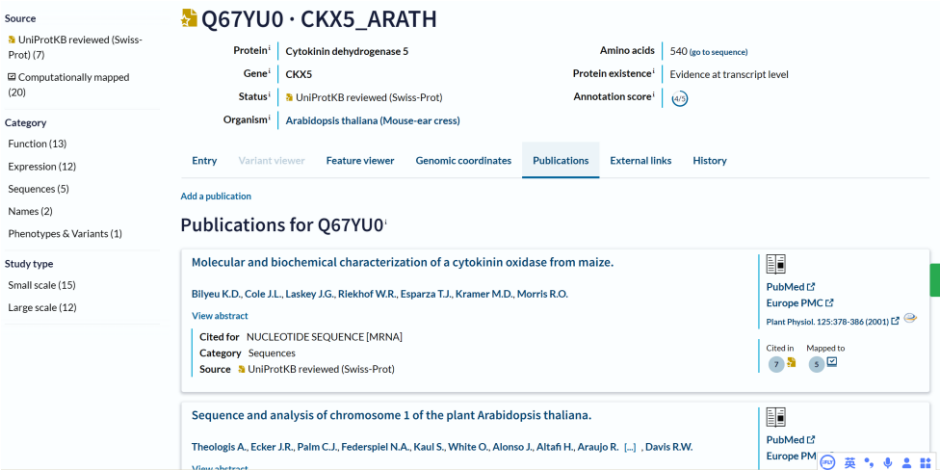
①CKX5-ARATH

1) 该序列条目的蛋白名、基因名和物种名：

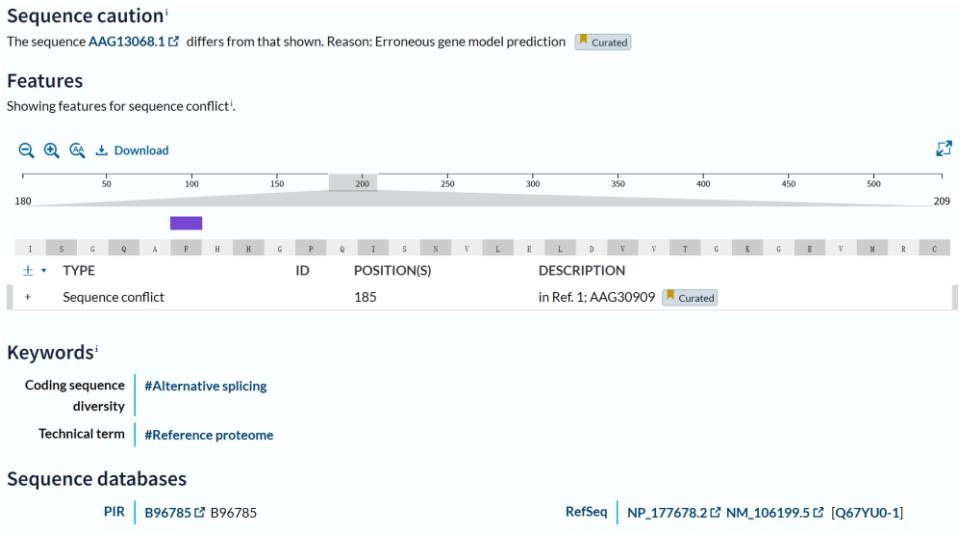
2) 该序列条目是否经过人工审阅，包括哪几大类注释信息：



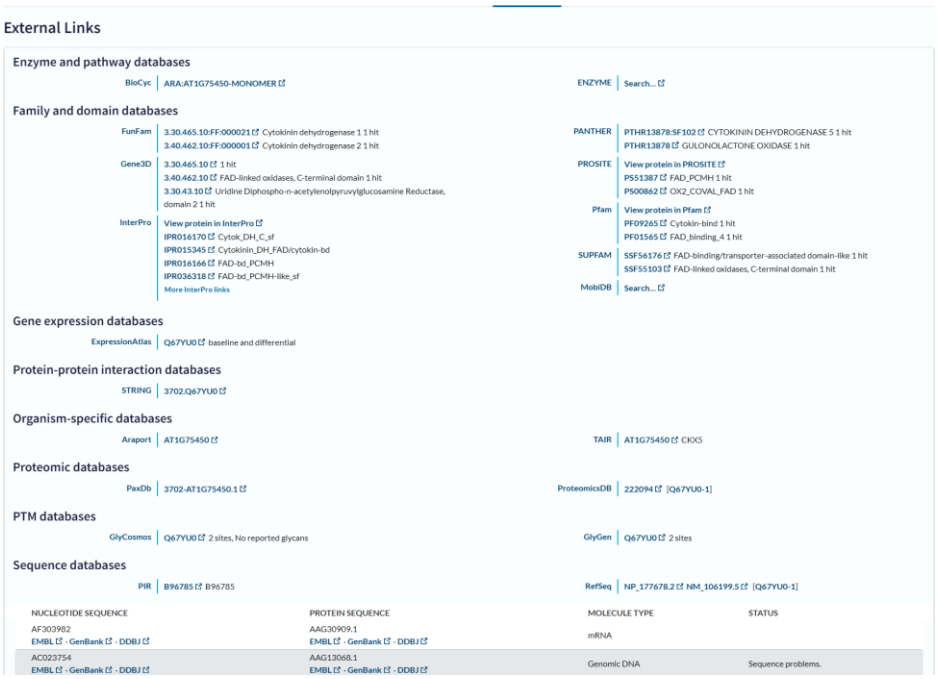
3) 该序列条目收录了多少篇相关文献，共分哪几大类：



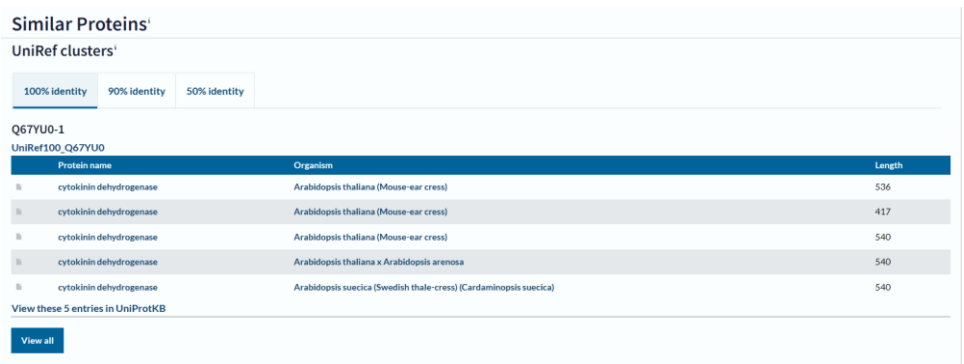
4) 该序列条目包括哪些特征位点信息？

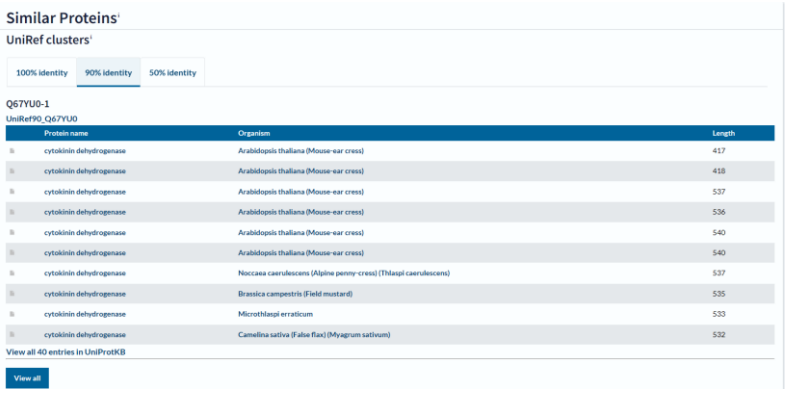


5) 该序列条目包括哪几类数据库交叉链接，其中你最感兴趣的有哪些数据库？从中可以获得哪些你感兴趣的信息？



6) Swiss-Prot 子库中与该序列条目相同位点占 100%, 90%和 50%的序列条目数：





Similar Proteins

UniRef clusters

100% identity | 90% identity | 50% identity

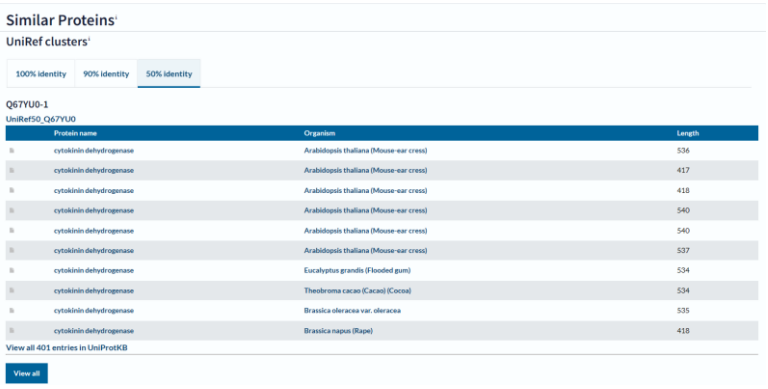
Q67YU0-1

UniRef90_Q67YU0

Protein name	Organism	Length
cytokinin dehydrogenase	Arabidopsis thaliana (Mouse-ear cress)	417
cytokinin dehydrogenase	Arabidopsis thaliana (Mouse-ear cress)	418
cytokinin dehydrogenase	Arabidopsis thaliana (Mouse-ear cress)	537
cytokinin dehydrogenase	Arabidopsis thaliana (Mouse-ear cress)	536
cytokinin dehydrogenase	Arabidopsis thaliana (Mouse-ear cress)	540
cytokinin dehydrogenase	Arabidopsis thaliana (Mouse-ear cress)	540
cytokinin dehydrogenase	Noccaea caerulescens (Alpine penny-cress) (Thlaspi caerulescens)	537
cytokinin dehydrogenase	Brassica campestris (Field mustard)	535
cytokinin dehydrogenase	Microthlaspi erraticum	533
cytokinin dehydrogenase	Camelina sativa (False flax) (Myagrum sativum)	532

View all 40 entries in UniProtKB

View all



Similar Proteins

UniRef clusters

100% identity | 90% identity | 50% identity

Q67YU0-1

UniRef50_Q67YU0

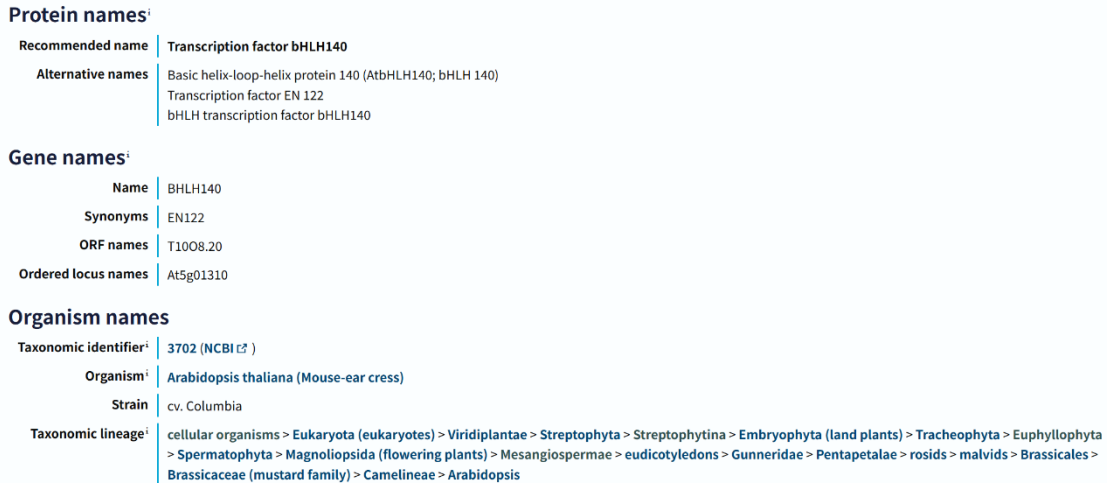
Protein name	Organism	Length
cytokinin dehydrogenase	Arabidopsis thaliana (Mouse-ear cress)	536
cytokinin dehydrogenase	Arabidopsis thaliana (Mouse-ear cress)	417
cytokinin dehydrogenase	Arabidopsis thaliana (Mouse-ear cress)	418
cytokinin dehydrogenase	Arabidopsis thaliana (Mouse-ear cress)	540
cytokinin dehydrogenase	Arabidopsis thaliana (Mouse-ear cress)	540
cytokinin dehydrogenase	Arabidopsis thaliana (Mouse-ear cress)	537
cytokinin dehydrogenase	Eucalyptus grandis (Flooded gum)	534
cytokinin dehydrogenase	Theobroma cacao (Cacao) (Cocoa)	534
cytokinin dehydrogenase	Brassica oleracea var. oleracea	535
cytokinin dehydrogenase	Brassica napus (Rape)	418

View all 401 entries in UniProtKB

View all

②Q9M041

1) 该序列条目的蛋白名、基因名和物种名：



Protein names

Recommended name	**Transcription factor bHLH140**
Alternative names	Basic helix-loop-helix protein 140 (AtbHLH140; bHLH 140) Transcription factor EN 122 bHLH transcription factor bHLH140

Gene names

Name	BHLH140
Synonyms	EN122
ORF names	T10O8.20
Ordered locus names	At5g01310

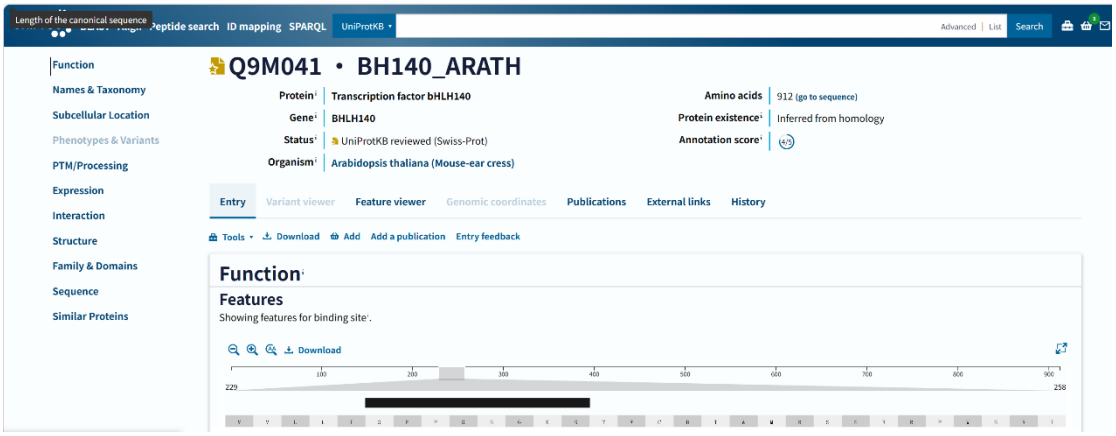
Organism names

Taxonomic identifier	3702 (NCBI)
Organism	Arabidopsis thaliana (Mouse-ear cress)
Strain	cv. Columbia
Taxonomic lineage	cellular organisms > Eukaryota (eukaryotes) > Viridiplantae > Streptophyta > Streptophytina > Embryophyta (land plants) > Tracheophyta > Euphyllophyta > Spermatophyta > Magnoliopsida (flowering plants) > Mesangiospermae > eudicotyledons > Gunneridae > Pentapetalae > rosids > malvids > Brassicales > Brassicaceae (mustard family) > Camelineae > Arabidopsis

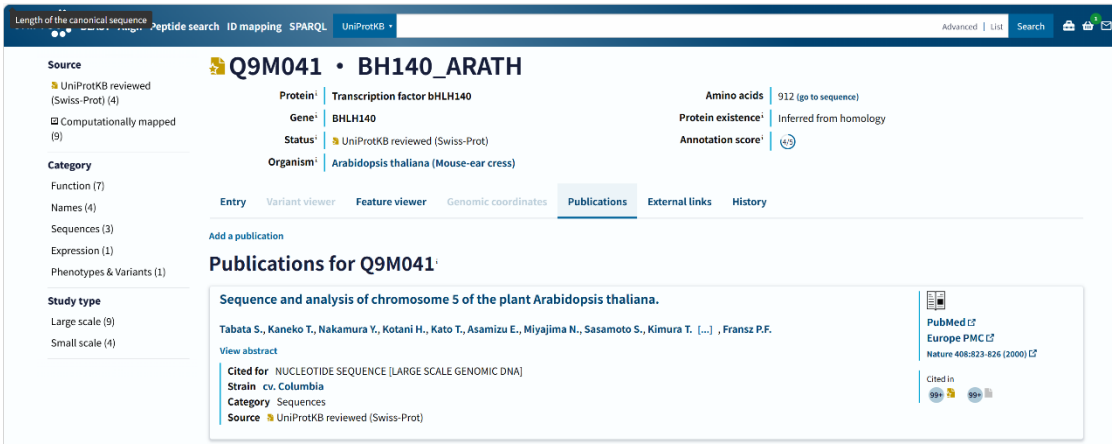
蛋白名：转录因子 bHLH140；　　基因名：BHLH140；　　物种名：拟南芥

2) 该序列条目是否经过人工审阅，包括哪几大类注释信息：

该序列已经过人工审阅，注释信息包括名称与分类学信息、亚细胞定位、表达、互作、家族、序列片段以及相似蛋白；

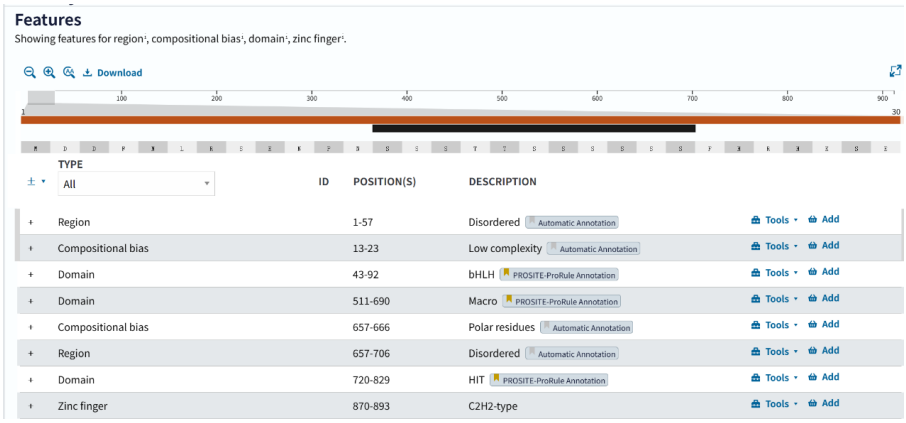


3) 该序列条目收录了多少篇相关文献，共分哪几大类：



截至 2026 年 4 月，该序列收录了 13 篇文献

4) 该序列条目包括哪些特征位点信息？

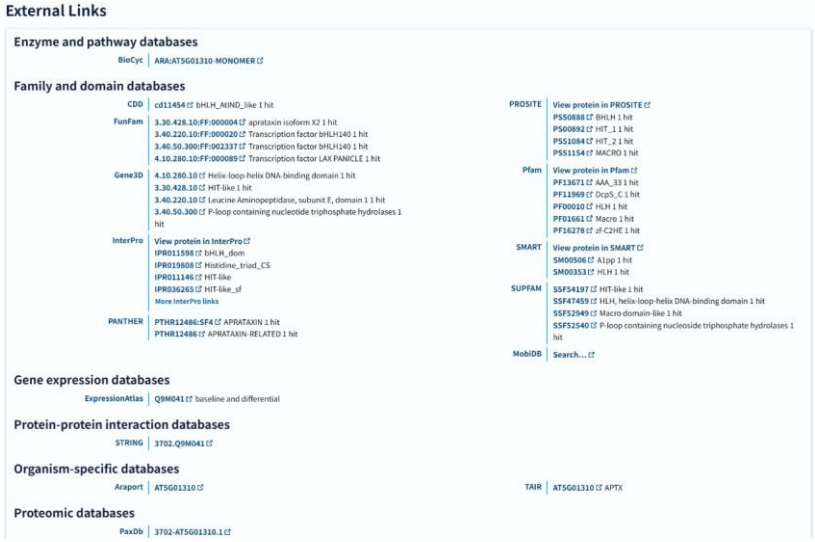


Automatic Annotation：UniProt 自动预测的结果，可信度中等，可作为参考。

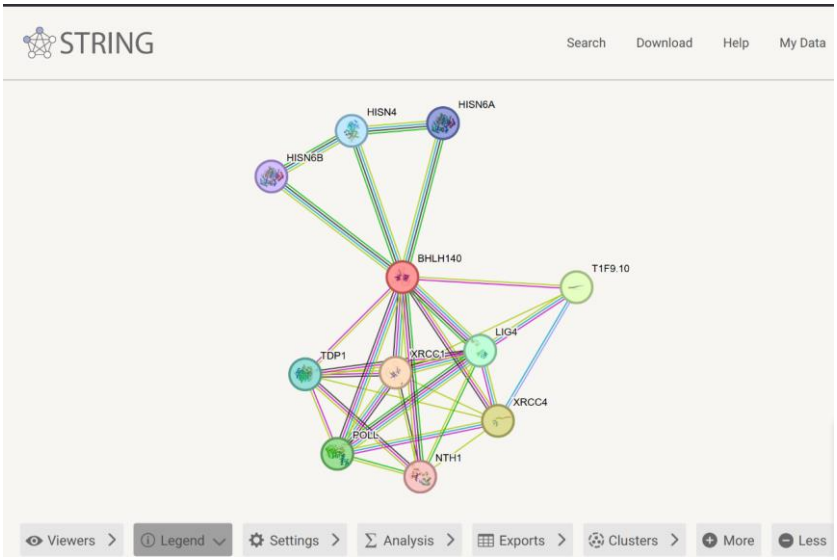
PROSITE-ProRule Annotation：来自 PROSITE 数据库的人工 / 半人工注释，可信度高，是公认的功能结构域注释。

以 bHLH transcription factor 研究为例，该序列中应重点关注 43-92 区域，如果要做定点突变验证功能，优先考虑这个区域的关键位点

5) 该序列条目包括哪几类数据库交叉链接，其中你最感兴趣的有哪些数据库？从中可以获得哪些你感兴趣的信息？



以蛋白质与蛋白质间的互作为例展开 STRING ：



所有互作蛋白可以分成两大类，功能差异非常明显：

• 组氨酸合成相关蛋白：HISN6B、HISN4、HISN6A

• DNA 修复/损伤响应相关蛋白：TDP1、POLL、LIG4、XRCC4、XRCC1、NTH1、T1F9.10

BHLH140 的互作网络揭示了它可能同时扮演两个关键角色：

①它通过与组氨酸合成通路的酶互作 / 调控，直接影响细胞内组氨酸的合成，进而调控植物的生长发育和基础代谢。作为 bHLH 家族蛋白，它的核心身份就是转录因子，这部分互作很可能是它的本职工作 —— 调控代谢相关基因的表达。

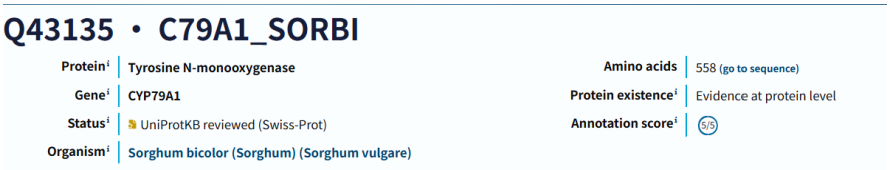
②BHLH140 与几乎所有核心 DNA 修复通路的关键蛋白都存在互作（NHEJ、BER、TDP1 通路）。这说明它可能不只是一个普通的转录因子，还直接参与 DNA 损伤的修复过程：

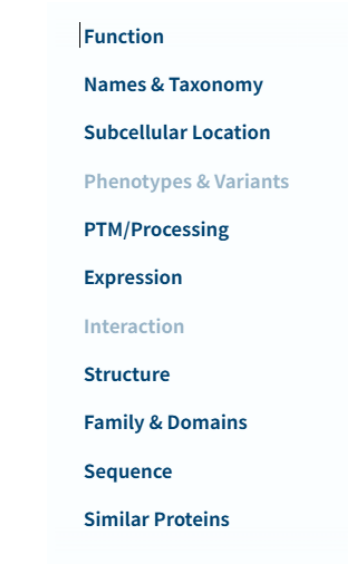
当植物受到环境胁迫（如紫外线、氧化胁迫、电离辐射）造成 DNA 损伤时，BHLH140 可能被激活，通过与这些修复蛋白互作，快速启动修复程序，维持基因组的稳定性。

它可能同时作为转录因子，上调 DNA 修复相关基因的表达，又作为蛋白互作因子，直接参与修复复合物的组装，形成“转录调控 + 直接参与”的双重调控模式。

③C79A1-SORBI

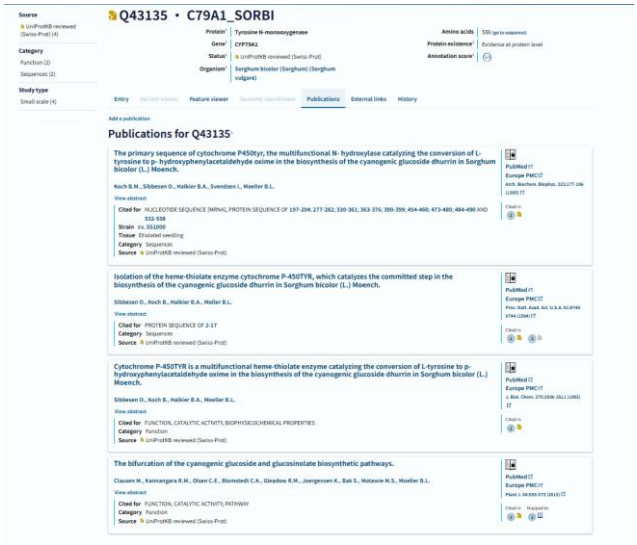
1) 该序列条目的蛋白名、基因名和物种名：





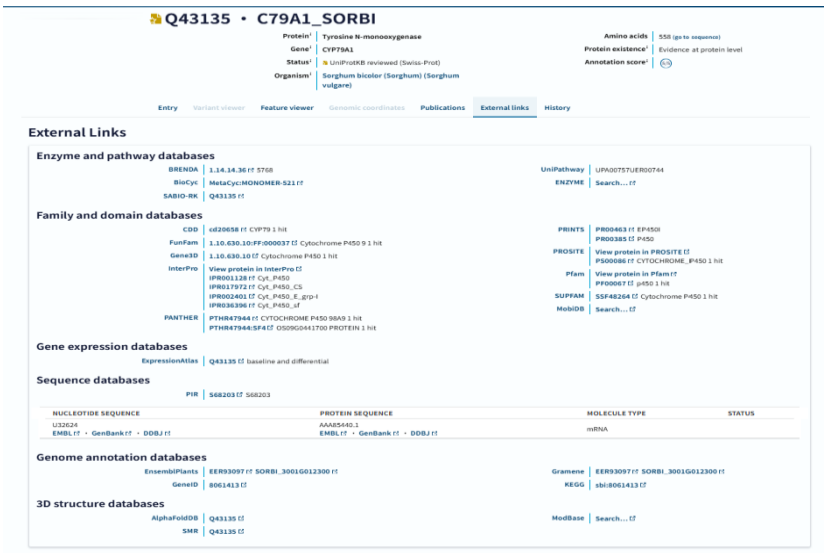
2) 该序列条目是经过人工审阅，包括哪几大类注释信息：

3) 该序列条目收录了几篇相关文献，共分哪几大类：

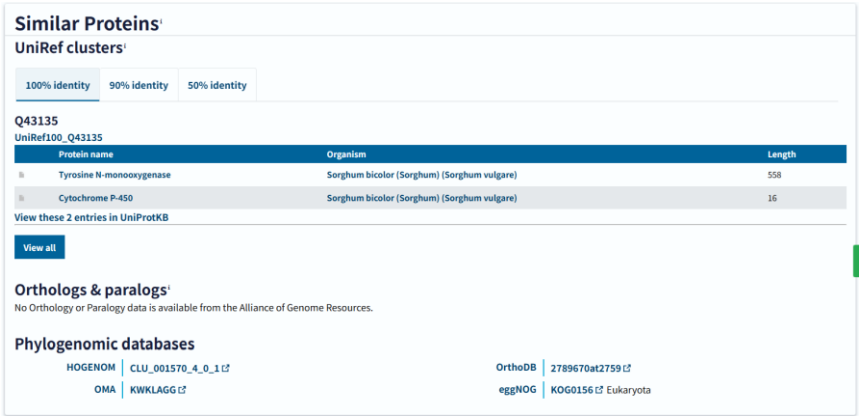


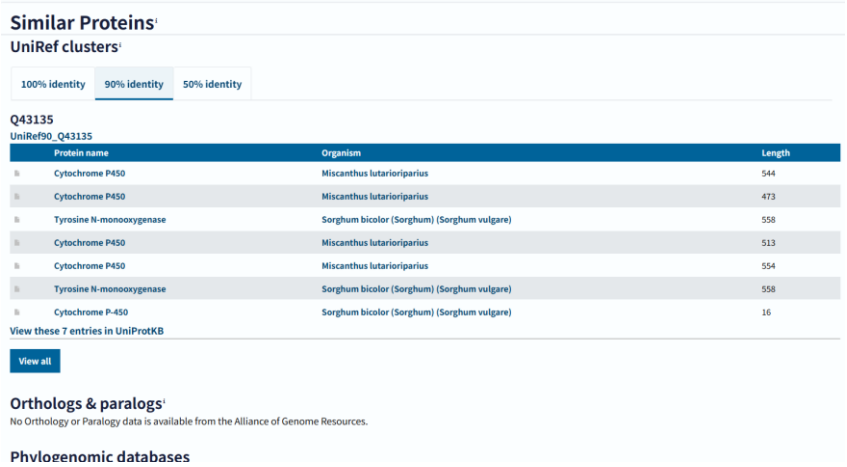
4) 该序列条目包括哪些特征位点信息？

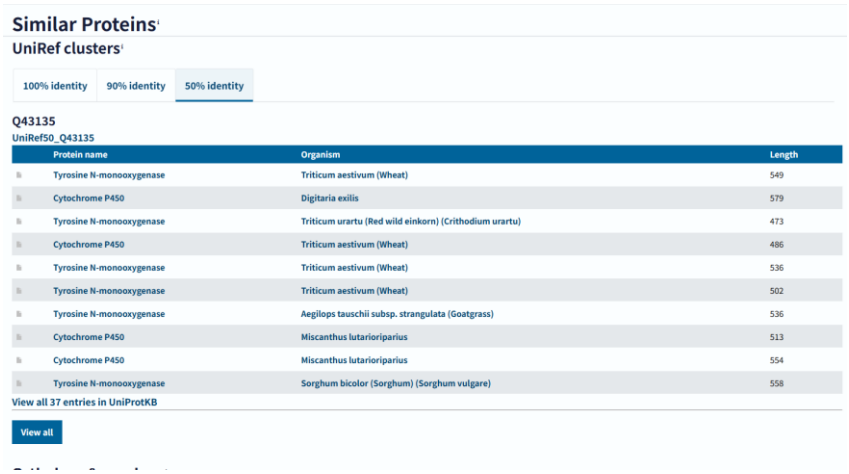
5) 该序列条目包括哪几类数据库交叉链接，其中你最感兴趣的有哪些数据库？从中可以获得哪些你感兴趣的信息？



6) Swiss-Prot 子库中与该序列条目相同位点占 100%, 90%和 50%的序列条目数：



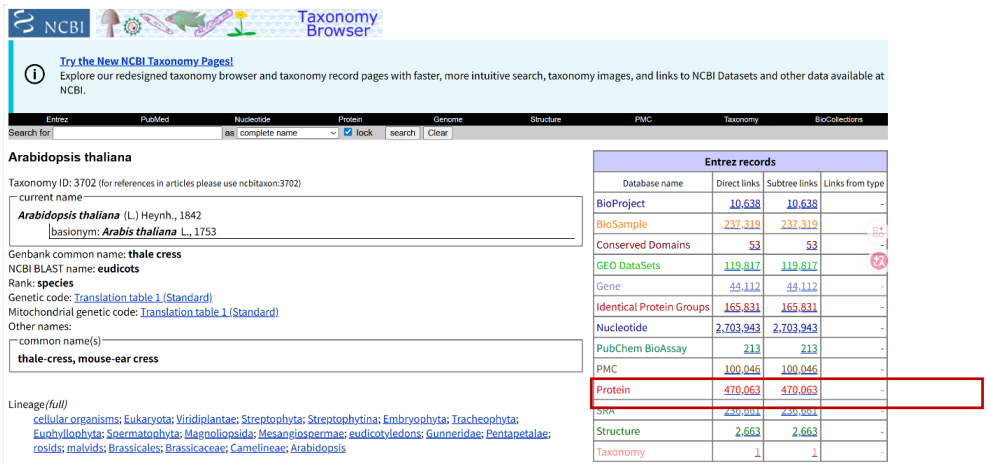


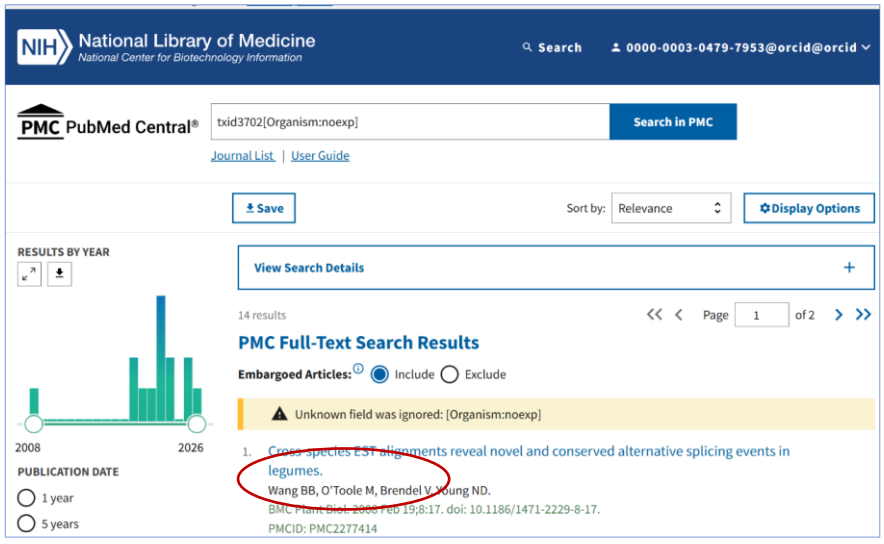


（8）问题

Q1：如何能够在数据库中一次性获得某个基因家族的全基因

Q2：为什么 Entrez records 中 PMC 显示数值不等于点击数字跳转后数值？





建议：Ensembl 收录的基因组信息以动物为主，若目标检索物种属植物/细菌大类，建议精准检索数据库 Ensembl Plants/ Bacteria（植物/细菌专用入口）

Ensembl Plants　网址 https://plants.ensembl.org/

Ensembl Bacteria　网址 https://bacteria.ensembl.org/