

---

# “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

The image shows two screenshots of the NCBI website. The top screenshot is the NCBI homepage, and the bottom screenshot is the BLAST tool page. A blue arrow traces the path from the 'BLAST' link in the 'Popular Resources' section of the homepage to the 'Nucleotide BLAST' button on the BLAST tool page. Red boxes highlight the 'BLAST' link in the top screenshot and the 'Nucleotide BLAST' button in the bottom screenshot.

**NCBI Home**  
National Library of Medicine  
National Center for Biotechnology Information

**Welcome to NCBI**  
The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

**Popular Resources**  
PubMed  
Bookshelf  
BLAST  
Nucleotide  
Genome  
SNP  
Gene  
Protein  
PubChem

**Web BLAST**

- Nucleotide BLAST**  
nucleotide → nucleotide
- blastx**  
translated nucleotide → protein
- tblastn**  
protein → translated nucleotide
- Protein BLAST**  
protein → protein

**BLAST Genomes**  
Enter organism common name, scientific name, or tax id  
Human Mouse Rat Microbes Search

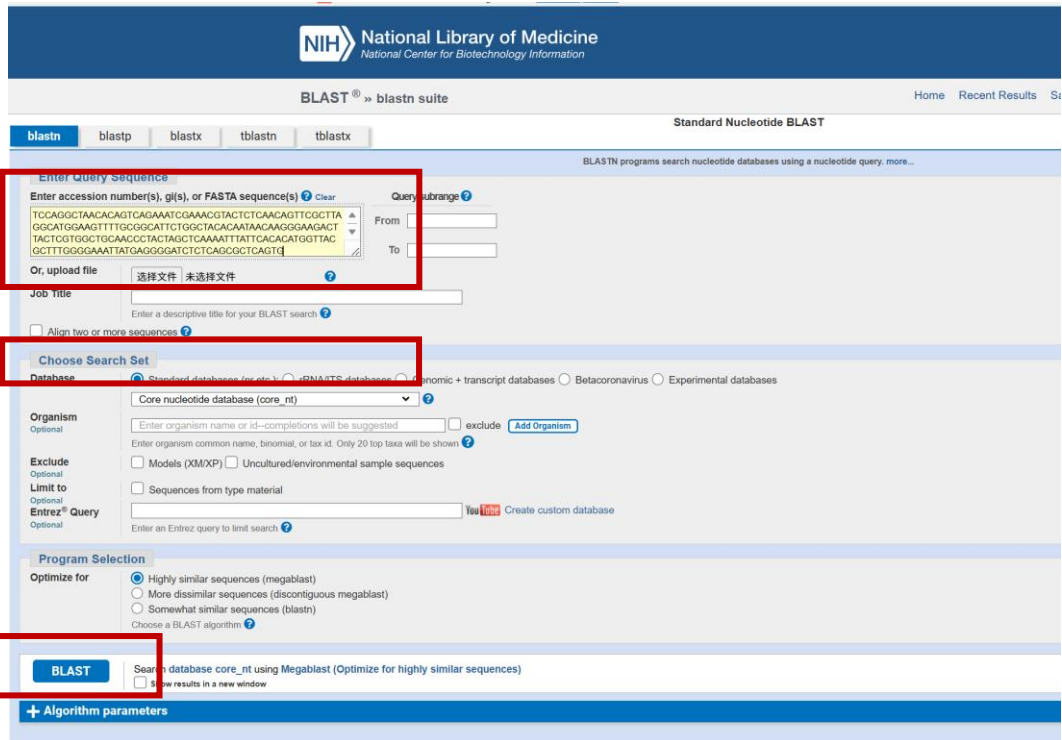
**Standalone and API BLAST**

- Download BLAST  
Get BLAST databases and executables
- Use BLAST API  
Call BLAST from your application
- Use BLAST in the cloud  
Start an instance at a cloud provider

---

在 ENTER 中输入序列（也可以选择文件）：

CTGATGAATCCCCTAATGATTTTGGTAAAAATCATTAAGTTAAGGTGGATACACATCTTGTC  
ATATGATCAAATGGTTTCGCGAAAAATCAATAATCAGACAACAAGATGTGCGAACTCGATATTTT  
ACACGACTCTCTTTACCAATTCTGCCCCGAATTACACTTAAAACGACTCAACAGCTTAACGTTG  
GCTTGCCACGCATTACTTGACTGTAAACTCTCACTCTTACCGAACTTGGCCGTAACCTGCCAA  
CCAAAGCGAGAACAAAACATAACATCAAACGAATCGACCGATTGTTAGGTAATCGTCACCTCC  
ACAAAGAGCGACTCGCTGTATAACGTTGGCATGCTAGCTTTATCTGTTCGGGCAATACGATGCC  
ATTGTACTTGTTGACTGGTCTGATATTCGTGAGCAAAAACGACTTATGGTATTGCGAGCTTCAGT  
CGCACTACACGGTCGTTCTGTTACTCTTTATGAGAAAGCGTTCCCGCTTTCAGAGCAATGTTCA  
AAGAAAGCTCATGACCAATTTCTAGCCGACCTTGCGAGCATTCTACCGAGTAACACCACACCGC  
TCATTGTCAGTGATGCTGGCTTTAAAGTGCCATGGTATAAATCCGTTGAGAAGCTGGGTTGGTAC  
TGGTTAAGTCGAGTAAGAGGAAAAGTACAATATGCAGACCTAGGAGCGGAAAACCTGGAAACCT  
ATCAGCAACTTACATGATATGTCATCTAGTCACTCAAAGACTTTAGGCTATAAGAGGCTGACTAA  
AAGCAATCCAATCTCATGCCAAATTCTATTGTATAAATCTCGCTCTAAAGGCCGAAAAAATCAGC  
GCTCGACACGGACTCATTGTCACCACCCGTCACCTAAAATCTACTCAGCGTCGGCAAAGGAGC  
CATGGGTTCTAGCAACTAACTTACCTGTTGAAATTCGAACACCCAAACAACCTTGTTAATATCTAT  
TCGAAGCGAATGCAGATTGAAGAAACCTTCCGAGACTTGAAAAGTCCTGCCTACGGACTAGGC  
CTACGCCATAGCCGAACGAGCAGCTCAGAGCGTTTTGATATCATGCTGCTAATCGCCCTGATGCT  
TCAACTAACATGTTGGCTTGCGGGCGTTCATGCTCAGAAACAAGGTTGGGACAAGCACTTCCA  
GGCTAACACAGTCAGAAATCGAAACGTA CTCTCAACAGTTCGCTTAGGCATGGAAGTTTTGCG  
GCATTCTGGCTACACAATAACAAGGGAAGACTTACTCGTGGCTGCAACCCTACTAGCTCAAAT  
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

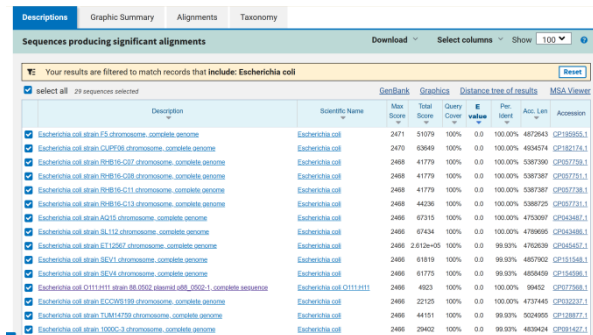
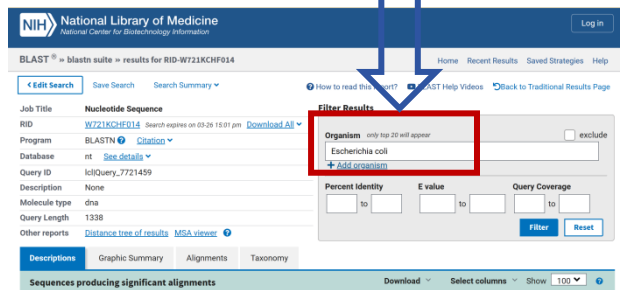
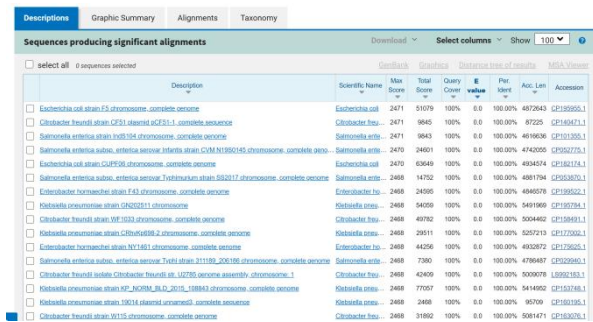
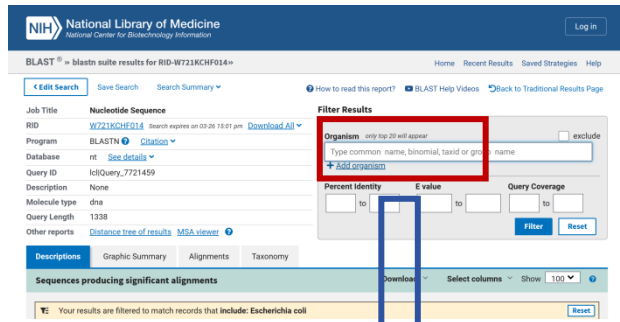
搜索结果筛选其中关键参数有：

- 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:

The image shows two screenshots from the UniProt database. The top screenshot is the 'Advanced Search' interface. It shows a search for 'Elastin' in the 'Protein Name [DE]' field, 'Saccharomyces' in the 'Organism [OS]' field, and 'activator' in the 'Keyword [KW]' field. The 'Organism [OS]' dropdown menu is highlighted with a red box. The bottom screenshot is the protein entry page for 'P01942 • HBA\_MOUSE'. It shows the protein name 'Hemoglobin subunit alpha', gene 'Hba', and organism 'Mus musculus (Mouse)'. The 'Amino acids' field shows '142 (go to sequence)' and the 'Protein existence' field shows 'Evidence at protein level', both highlighted with red boxes.

**UniProt** BLAST Align Peptide search ID mapping SPARQL UniProtKB Advanced | List Search

**P69905 • HBA\_HUMAN**

Protein: Hemoglobin subunit alpha  
 Gene: HBA1; HBA2  
 Status: UniProtKB reviewed (Swiss-Prot)  
 Organism: Homo sapiens (Human)

Amino acids: 142 (go to sequence)

Function: Involved in oxygen transport from the lung to the various peripheral tissues.

Hemopresin: Hemopresin acts as an antagonist peptide of the cannabinoid receptor CNR1 (PubMed:18077343). Hemopresin-binding efficiently blocks cannabinoid receptor CNR1 and subsequent signaling (PubMed:18077343).

Miscellaneous: Gives blood its red color.

Features

Tools: Download Add Community curated (3) Add a publication Entry feedback

**UniProt** BLAST Align Peptide search ID mapping SPARQL UniProtKB Advanced | List Search

Entry Variant viewer Feature viewer Genomic coordinates Publications External links History

Sequence

Download Add Highlight Copy sequence

Length: 142  
 Mass (Da): 15,085  
 Last updated: 2007-01-23 v2  
 MD5 Checksum: D07A65D16C2DBE7322015943CB989AAE

MVLSGEDKSN IKAAMGKIGG HGAEYGAEL ERMIFASPTT KTYFPHFDVS HGSQAQVKGHG KKVADALASA AGLHDDLPGA LLSALDLHAH  
 KLRVDPVNFK LLSHCLLVTL ASHHPADFTP AVHASLDKFL ASVSTVLTSK YR

Computationally mapped potential isoform sequences:  
 There is 1 potential isoform mapped to this entry

Entry	Entry name	Gene name	Length
<input type="checkbox"/> Q91VB8	Q91VB8_MOUSE	Hba-a1	142

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)

国家生物信息中心  
 数据资源 计算分析 标准规范 数据网络

简体中文 English

计算分析 查看所有计算分析选项

**序列比对(BLAST)**  
 基于BLAST算法对核酸或蛋白质序列, 鉴定序列间的相似性和同源性, 帮助分析序列功能和进化关系

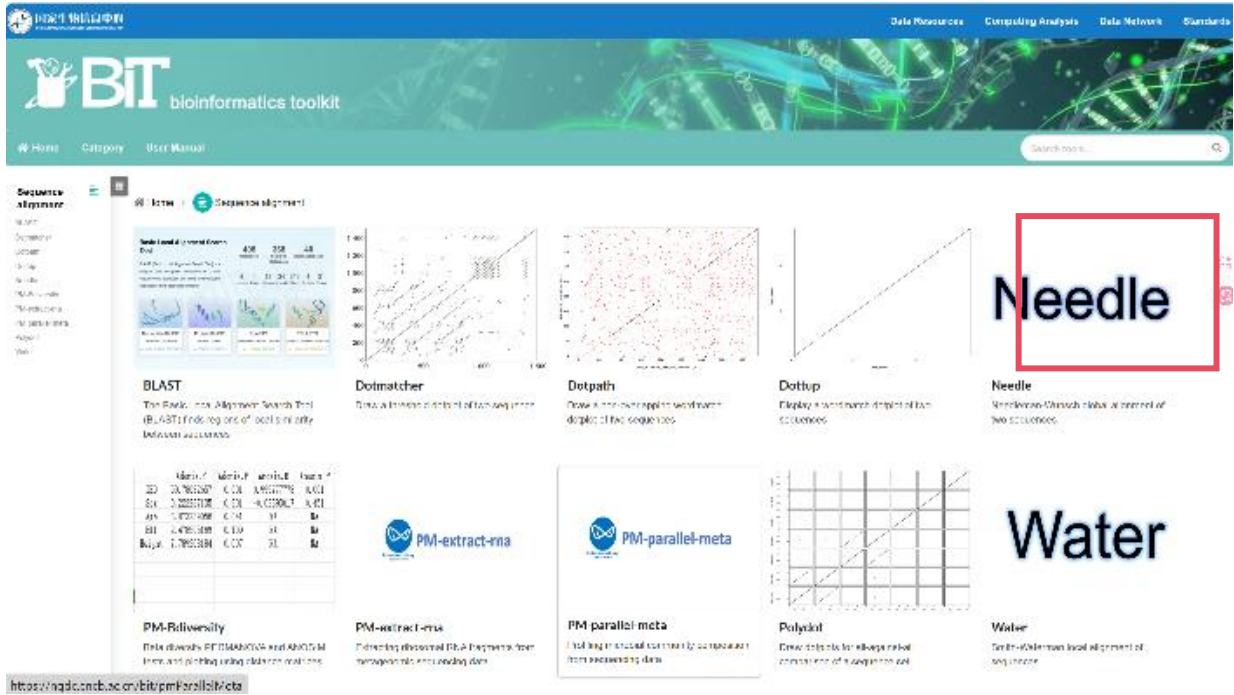
**云平台(BiT)**  
 基于云的生物信息学工具包, 集成常用生物信息方法、软件与区域网络方法, 提供免费的数据分析云服务

**基因组组装与注释**  
 将测序数据组装成连续序列及完整基因组, 并对基因组中基因的结构与功能元件进行注释和分析

**单细胞与空间组学**  
 单细胞分类、细胞的空间分布及相互作用分析, 实现高分辨率的细胞异质性和空间关系表征

**系统发生与分子进化**  
 通过比较核酸、蛋白质等分子信息, 分析物种演化与遗传变异, 揭示进化历史、遗传多样性与适应性

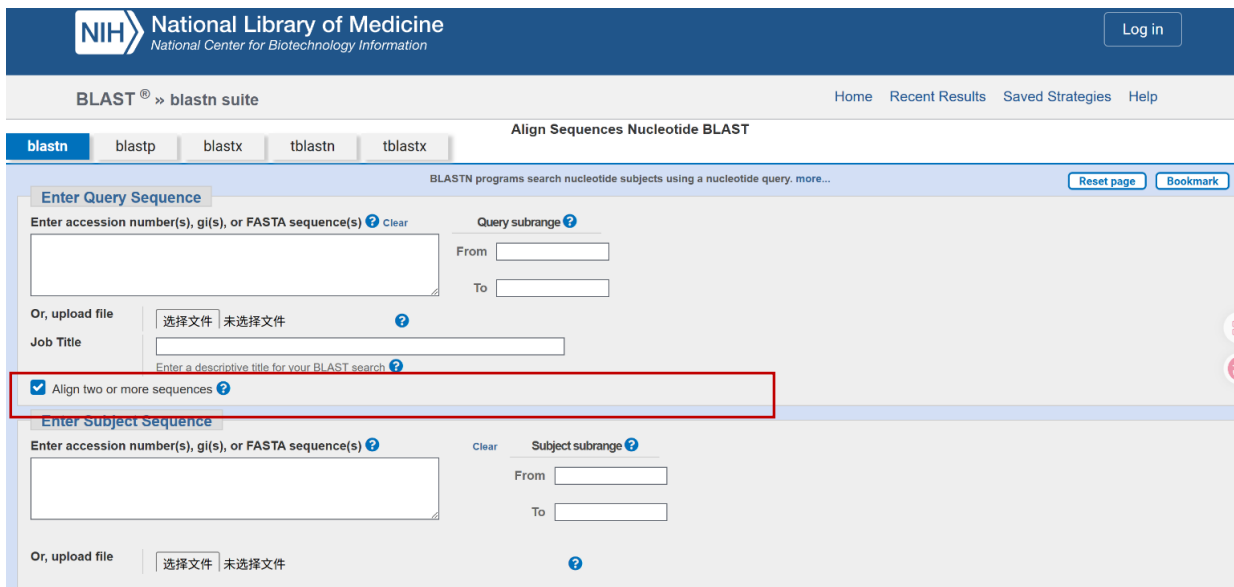
**健康与疾病检测**  
 多种定量检测分析, 以评估和检测个体健康状态, 发现疾病严重程度, 监测治疗进展等情况



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)

Expasy SIM

Home Contact

### SIM - Alignment Tool for Protein Sequences

SIM is a program which finds a user-defined number of best non-intersecting alignments between two protein sequences or within a sequence [more].  
Once the alignment is computed, you can view it using LALNVIEW, a graphical viewer program for pairwise alignments [reference to LANVIEW].  
Note: You can use the PBIL server to align nucleic acid sequences with a similar tool.

Enter two sequences:

In each box, please enter one UniProtKB AC/ID (e.g. P05130 or KPCL\_DROME) OR one protein sequence in single letter code.

Sequence 1:  Sequence 2:

Parameters:

Number of alignments to be computed:

Gap open penalty:

Gap extension penalty:  [documentation]

Comparison Matrix:

or

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

- 人序列长度：142 aa（原始序列 C 端为 ...KYR---，末尾三个缺失）
- 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  $\alpha$  亚基的进化保守性。

高度保守区域

- 血红素结合位点（如第 58 位 His，第 87 位 His 等）均完全保留。
- 与  $\beta$  亚基相互作用的区域（如  $\alpha 1 \beta 2$  接触面）也高度一致。
- 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  $\alpha$  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：

- 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。
- 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：

- 该比对结果清晰展示了人和大鼠血红蛋白  $\alpha$  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 $\alpha$ 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)

- 一致性 (Identity):  $120/142 = 84.5\%$ , 84.5% 的氨基酸完全相同
- 相似性 (Similarity):  $127/142 = 89.4\%$ , 包含性质相似的氨基酸 (如疏水  $\rightarrow$  疏水、带电  $\rightarrow$  带电), 整体功能保守性很高
- 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
- 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 $\alpha$ 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)

- 一致性 (Identity)  $122/142$  (85.9%) 完全相同的氨基酸残基占比
- 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)
- 空位 (Gaps)  $0/142$  (0.0%) 比对中无任何插入 / 缺失 (gap)
- 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

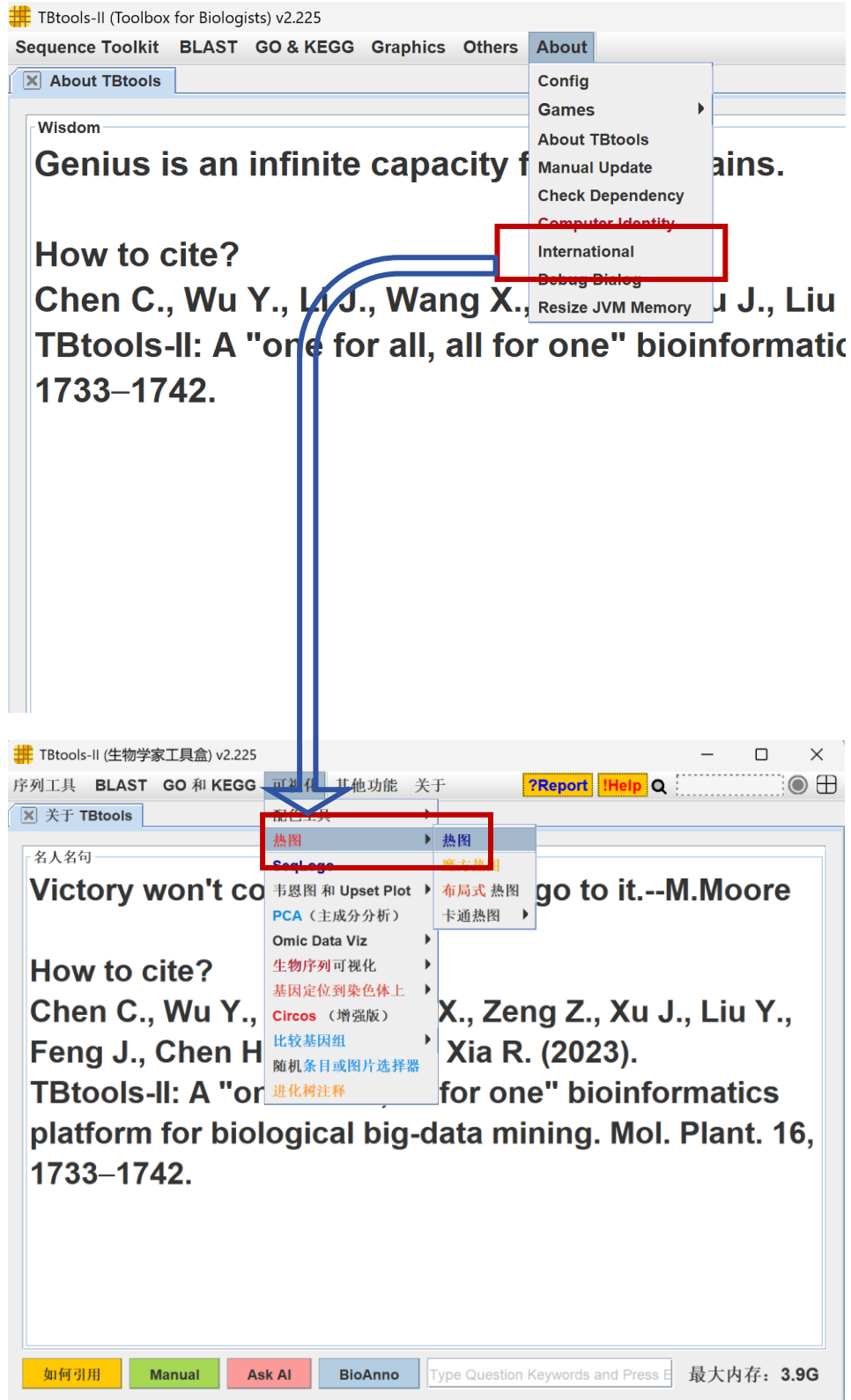
从比对片段可以看到:

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 ( $\alpha$ -珠蛋白是携氧核心蛋白, 进化压力大, 序列高度保守)。



### (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 iHelp

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 3.025963682E7 3.732006021E7 3.379866562E7 3.748494632E7
Linalool oxide primeveroside 3.679532571E7 4.007048712E7 4.200997275E7 4.682236517E7
Jasmonic acid 3974990.43 4180907.88 4202918.21 4686693.1
Pipelicolic acid 3.645574582E7 4.017624227E7 4.100532282E7 4.883142162E7
Isovitexin 9634676.27 1.288954805E7 1.466759402E7 1.208404319E7
=====

```

(Optional) Set Input Row Group File (GrpIDitSubGrpIDitRowName)  
Drag and Drop a Tab-delimited File Here

(Optional) Set Input Col Group File (GrpIDitSubGrpIDitColName)  
Drag and Drop a Tab-delimited File Here

(Optional) Preset Newick for Row

(Optional) Preset Newick for Column

TBtools HeatMap by CJ (cci0410@gmail.com)

**Show Control Dialog**

Compound	MH1	MH2	MH3	MH4
Lucuminic acid	3.025963682E7	3.732006021E7	3.379866562E7	3.748494632E7
Linalool oxide primeveroside	3.679532571E7	4.007048712E7	4.200997275E7	4.682236517E7
Jasmonic acid	3974990.43	4180907.88	4202918.21	4686693.1
Pipelicolic acid	3.645574582E7	4.017624227E7	4.100532282E7	4.883142162E7
Isovitexin	9634676.27	1.288954805E7	1.466759402E7	1.208404319E7
Quercetin				
LysoPC(18:1)				
Nicotinic acid				
Glycerophosphocholine				
Kaempferol-3-(6-acetyl)galactoside				
Guanine				
Thymine				
Luteolin-8-C-glucoside				
Hydroxyjasmonic acid glucoside				
Linalool primeveroside				
Glutamic acid				
GC				
Theanine	~8E9	~7E9	~6E9	~5E9
Theanine glucoside				
trans-Cinnamic acid				
Threonic acid				
Kaempferol-3-(6"-galloyl)glucoside				
Kaempferol-3-arabinoside				
Theasinensin B				
Sucrose				

**Lucky Color!**

Legend Font

Row Name Font  Width: 333 Tree Width: 80

Col Name Font  Height: 555 Tree Height: 80

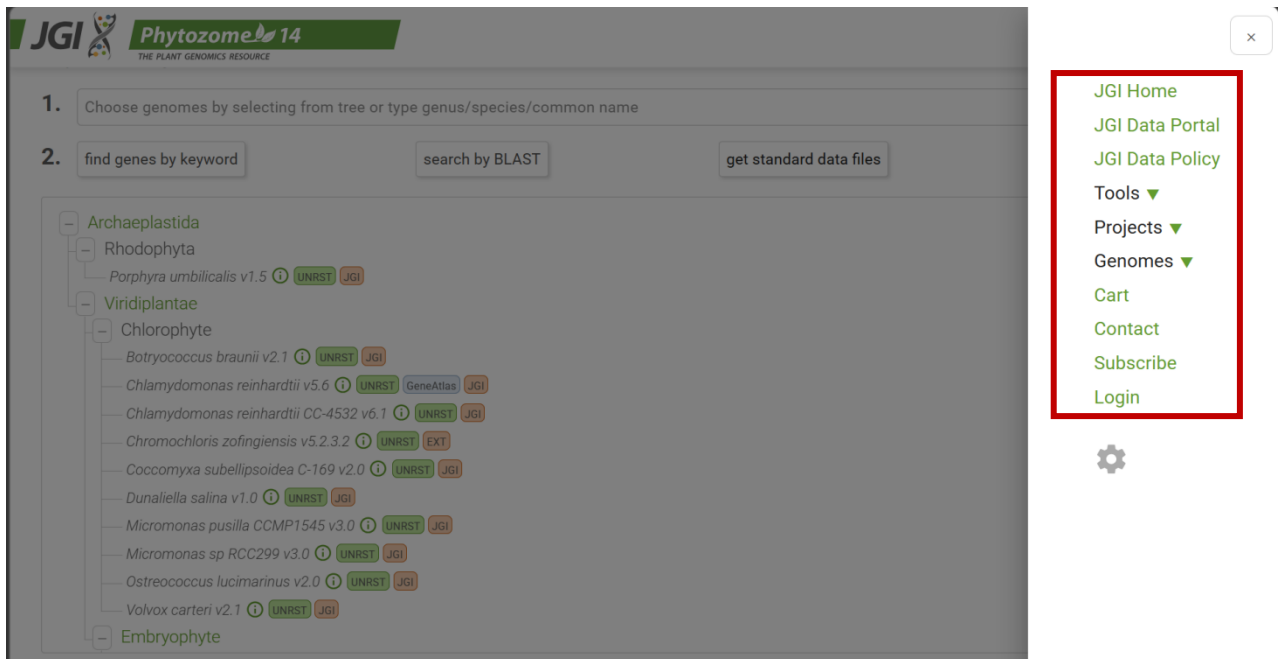
#### (4) 数据库 PHYTOZOME 使用

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

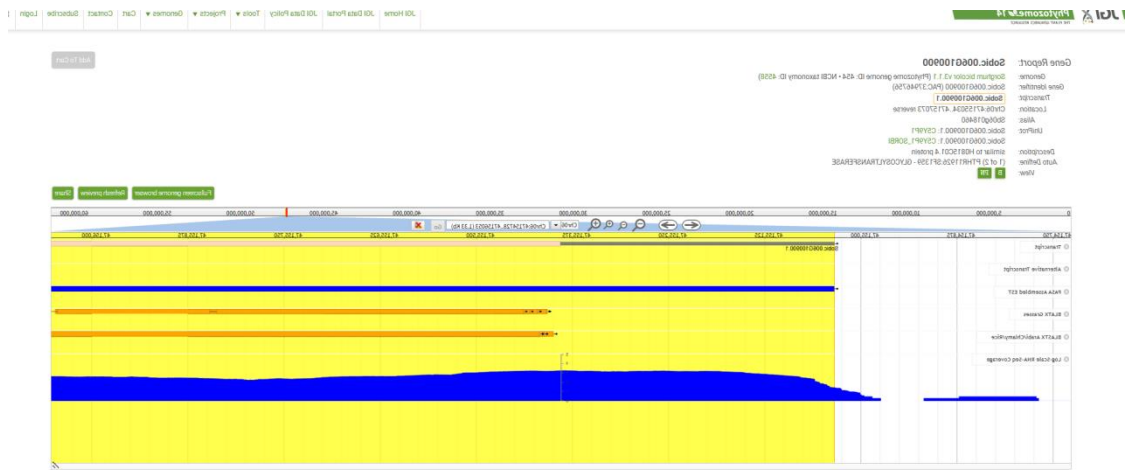
PHYTOZOME——为高粱这种小众作物提供了高质量参考基因组与完善的比较基因组学工具，有效弥补了研究资源不足的问题，便于开展功能基因挖掘与进化分析。



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

### ① UniProt 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/](https://ftp.uniprot.org/pub/databases/uniprot/previous_releases/release-YYYY_NN/knowledgebase/))

**UniProtKB statistics**

**Introduction**  
This is release 2026\_01 of UniProtKB, published on Wed Jan 28 2026.  
Previous release statistics are available from the UniProt FTP server.

Throughout this document, whenever a statistic has a corresponding query, a link has been provided. In some instances, due to the nature of the statistic, no query link is possible.

**Total number of entries in this release of UniProtKB**

Section	Number of entries in total	Number of entries with an annotation update	Number of entries with a sequence update
UniProtKB	203,130,941	130,556,322	23,502
Reviewed (Swiss-Prot)	574,627	380,278	180
Unreviewed (TrEMBL)	202,556,314	130,176,044	23,322

**Total number of new entries in this release of UniProtKB**

Section	Number of new entries	Number of new sequences
UniProtKB	3,594,411	3,594,320
Reviewed (Swiss-Prot)	987	896
Unreviewed (TrEMBL)	3,593,424	3,593,424

#### 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： $\alpha$  螺旋 Alpha helix、 $\beta$  折叠 Beta sheet、 $\beta$  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、羧基化数据库 CarbonylDB、糖基化数据库 Glyconec、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

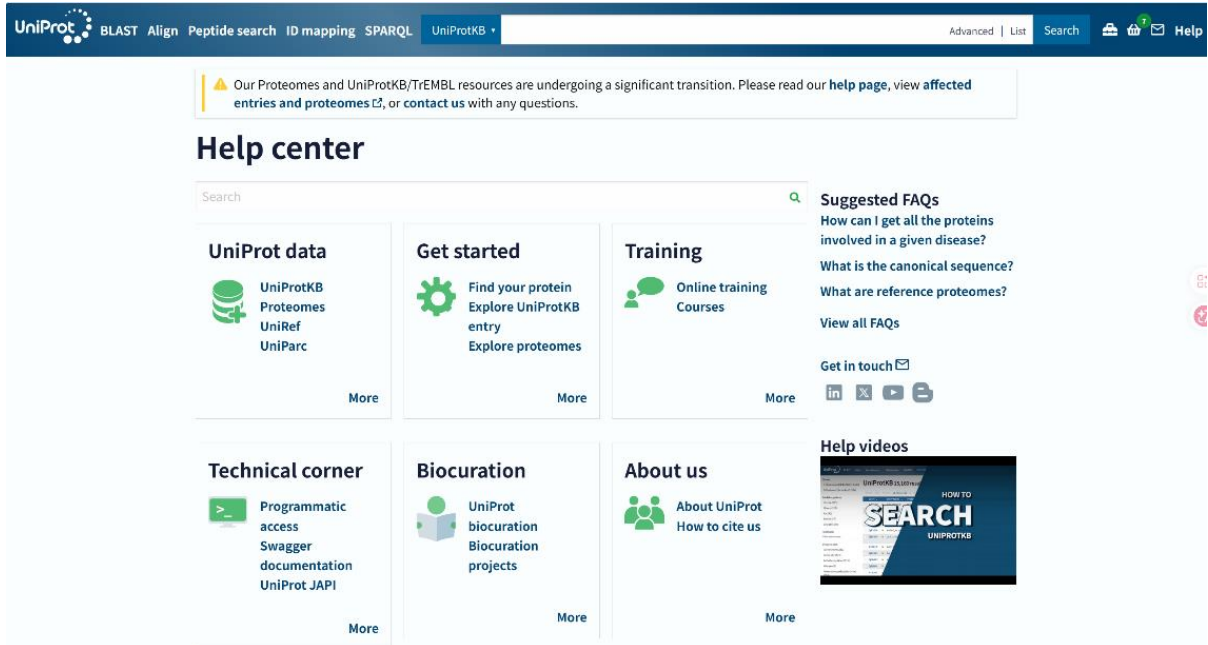
2) 已审阅拟南芥序列

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	<a href="#">(organism_id:3702) in UniProtKB search (136309)   UniProt</a>	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	<a href="#">organism_id:3702 AND reviewed:true in UniProtKB search (16418)   UniProt</a>	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	<a href="#">(organism_id:3702) AND</a>	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"	<a href="#">(reviewed:true)</a> <a href="#">AND (existence:1)</a> <a href="#">in UniProtKB</a> <a href="#">search (7752)  </a> <a href="#">UniProt</a>	(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)	<a href="#">(organism_id:3702)</a> <a href="#">)AND</a> <a href="#">(reviewed:true)</a> <a href="#">AND (existence:1)</a> <a href="#">AND</a> <a href="#">(ft_domain:Transmembrane) in</a> <a href="#">UniProtKB search</a> <a href="#">(41)   UniProt</a>	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:*)	<a href="#">(organism_id:3702)</a> <a href="#">)AND</a> <a href="#">(reviewed:true)</a> <a href="#">AND (existence:1)</a> <a href="#">AND</a> <a href="#">(ft_domain:Transmembrane) AND</a> <a href="#">(ft_signal:*) in</a> <a href="#">UniProtKB search</a> <a href="#">(1)   UniProt</a>	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](#)

- Function
- Names & Taxonomy
- Subcellular Location
- Phenotypes & Variants
- PTM/Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence
- Similar Proteins

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](#)

The screenshot shows the InterPro website interface. The main heading is "Q9FY06 Inner membrane protein PPF-1, chloroplastic". Below this, there is a search bar and a table of matches. The table has columns for Accession, Short Name, Name, Source Database, and Matches. The first row shows a match with Pfam database, accession PF02096, short name 60KD\_IMP, and name 60Kd inner membrane protein. The matches column shows a progress bar from 0 to 400, with the current match at 200. There are also navigation buttons like "Previous" and "Next".

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。

The screenshot shows the UniProt website interface with a BLAST search result. The main heading is "Q9FY06 UniRef50\_Q9FY06". Below this, there is a table of protein entries. The table has columns for Protein name, Organism, and Length. The entries are as follows:

Protein name	Organism	Length
Inner membrane protein ALBINO3, chloroplastic	Apostasia shenzhenica	447
Inner membrane protein PPF-1 chloroplastic-like	Trifolium pratense (Red clover)	445
Membrane insertase YidC/Oxa/ALB C-terminal domain-containing protein	Trifolium subterraneum (Subterranean clover)	443
Inner membrane protein PPF-1, chloroplastic	Cinnamomum micranthum f. kanehirae	468
Membrane insertase YidC/Oxa/ALB C-terminal domain-containing protein	Mikania micrantha (bitter vine)	476
Membrane insertase YidC/Oxa/ALB C-terminal domain-containing protein	Papaver nudicaule (Iceland poppy)	458
Inner membrane protein	Genlisea aurea	436
Inner membrane protein PPF-1, chloroplastic	Cicer arletinum (Chickpea) (Garbanzo)	439
Membrane insertase YidC/ALB3/OXA1/COX18, membrane insertase YidC/Oxa1	Helianthus annuus (Common sunflower)	450
Membrane insertase YidC/Oxa/ALB C-terminal domain-containing protein	Aquilegia coerulea (Rocky mountain columbine)	435

View all 75 entries in UniProtKB

UniProt BLAST Align Peptide search ID mapping SPARQL UniProtKB Advanced | List Search

Function Entry Variant viewer Feature viewer Genomic coordinates Publications External links History

Names & Taxonomy Subcellular Location Phenotypes & Variants PTM/Processing Expression Interaction Structure Family & Domains Sequence Similar Proteins

100% identity 90% identity 50% identity

**Q9FY06**  
UniRef90\_Q9FY06

Protein name	Organism	Length
Inner membrane protein PPF-1, chloroplastic	Cicer arletinum (Chickpea) (Garbanzo)	439
Inner membrane protein PPF-1 chloroplastic-like	Trifolium pratense (Red clover)	445
Membrane Insertase YidC/Oxa/ALB C-terminal domain-containing protein	Trifolium subterraneum (Subterranean clover)	443
60 kDa inner membrane protein	Medicago truncatula (Barrel medic) (Medicago tribuloides)	444
Inner membrane protein PPF-1	Pisum sativum (Garden pea) (Lathyrus oleraceus)	442
Inner membrane protein PPF-1 chloroplastic-like	Trifolium pratense (Red clover)	199
Inner membrane protein ALBINO3	Trifolium medium	99
Uncharacterized protein	Trifolium pratense (Red clover)	445

View these 8 entries in UniProtKB

View all

**Orthologs & paralogs**  
No Orthology or Paralogy data is available from the Alliance of Genome Resources.

UniProt BLAST Align Peptide search ID mapping SPARQL UniProtKB Advanced | List Search

**Q8LBP4 • ALB3\_ARATH**

Protein: Inner membrane protein ALBINO3, chloroplastic Amino acids: 462 (go to sequence)  
 Gene: ALB3 Protein existence: Evidence at protein level  
 Status: UniProtKB reviewed (Swiss-Prot) Annotation score: 0.5  
 Organism: Arabidopsis thaliana (Mouse-ear cress)

Entry Variant viewer Feature viewer Genomic coordinates Publications External links History

**External Links**

**Enzyme and pathway databases**  
BioCyc: ARA:AT2G28800-MONOMER MetaCyc:AT2G28800-MONOMER

**Protein family/group databases**  
TCDB: 2.A.9.2.1 the membrane protein insertase (yidc/alb3/oxa1) family

**Family and domain databases**  
CDD: cd20070 5TM\_YidC\_Albb3 1 hit  
DisProt: DP00662  
InterPro: View protein in InterPro  
PANTHER: PTHR12428:SF47 INNER MEMBRANE PROTEIN ALBINO3, CHLOROPLASTIC 1 hit  
PTHR12428 OXA1 1 hit  
Pfam: View protein in Pfam

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。

ThaleMine v5.1.0-20250704 Data mining on Arabidopsis thaliana

Home Templates Lists QueryBuilder Regions Data Sources API MyMine Contact Us Log in

Search: e.g. AT1G01640 GO

Oops! Error: Could not parse response to GET https://phytozone-next.igj.doe.gov/phytomine/service/summaryfields: "" (0: SyntaxError: Unexpected end of JSON input)

Gene: **ALB3** *A. thaliana*  
 DB Identifier: AT2G28800 Secondary Identifier: locus:2053230  
 Name: ALBINO3 Brief Description: 63 kDa inner membrane family protein

TAIR Computational Description: 63 kDa inner membrane family protein (source:Araport11)  
 TAIR Curator Summary: member of Chloroplast membrane protein ALBINO3 family. Similar to pea PPF1 and may play a role in plant senescence.  
 TAIR Short Description: 63 kDa inner membrane family protein  
 TAIR Aliases: ALB3

Quick Links: Summary Genomics Proteins Function Interactions Expression Homology Other

9 Gene Rifs Manage Columns Manage Filters Manage Relationships Save as List Generate Python code Export

Showing 1 to 9 of 9 rows

Gene Rifs Annotation	last updated	Gene Rifs Organism	Gene Rifs Gene	Gene Rifs PubMed Id
ALB3 contributes to the process of protein insertion into the thylakoids via the ALB3-chloroplast signal recognition particle pathway.	2016-07-30	A. thaliana	ALB3	26265777

Lists: This Gene is in one list: Genes with a Loss-of-Function Mutant Phenotype: Morphological - Vegetative (640)

External Links: Aracyc, AceView, Plant Proteome Database, SUBA, Gramene, GeneVible, ABR, TAIR, EnsemblPlants, PDD, Alzocip, Plaza, pep2

The screenshot shows the TAIR website interface. At the top, there is a navigation bar with links for Home, Help, Contact, About Us, Subscribe, Logout, and Profile. Below this is a search bar with the text "Enter search text" and a "Gene" dropdown menu. A secondary navigation bar contains "Advanced Search", "Browse", "Tools", "Portals", "Download", "Submit", "News", and "Stocks". The main content area is titled "Locus: AT2G28800 (ALB3) Premium Page". On the left, there is a sidebar with various data categories: Transcripts, Maps and Mapping Data, Sequences, Protein Data, Expression, Gene Ontology, Homology, Germplasm and Clones, Polymorphisms, Publications, and External Links. The main content area is divided into sections: "Summary" (Gene Model Type: protein\_coding; Symbols: ALB3 (ALBINO 3) (Primary Symbol); Description: member of Chloroplast membrane protein ALBINO3 family. Similar to pea PPF1 and may play a role in plant senescence.; Community Comments: Add My Comment, Show Comments; Update History: No update history available; Date last modified: 2017-10-19; TAIR Accession: Locus:2053230) and "Transcripts" (Representative Gene: AT2G28800.1).

### (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)。

Organism names	
Taxonomic identifier <sup>i</sup>	3702 (NCBI <a href="#">E</a> )
Organism <sup>i</sup>	Arabidopsis thaliana (Mouse-ear cress)
Strains	cv. Columbia cv. Bla-10 cv. Chi-1 cv. Co-1 cv. Cvi-0 <a href="#">7 more strains</a>
Taxonomic lineage <sup>i</sup>	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) > Camelinae > Arabidopsis

3) 该物种在 Swiss-Prot 和 TrEMBL

BL 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；

The screenshot shows the UniProt interface for entry 3702. A 'Status' section is highlighted with a red box, showing:

- Reviewed (Swiss-Prot): 16,484
- Unreviewed (TrEMBL): 126,337

Below this, a warning message states: 'The unreviewed UniProtKB/TrEMBL database will be reduced in size in release 2026\_02 (first half of 2026).' It lists entries to be retained (reference proteomes, all reviewed entries, and selected unreviewed entries with important data) and entries to be removed (unreviewed entries not part of a reference proteome).

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；

UniProt BLAST Align Peptide search ID mapping SPARQL UniProtKB 3702 Advanced | List Search Help

Alternative products (isoforms) (3,232) More items

**Protein existence**

- Protein level (7,766)
- Transcript level (6,477)
- Homology (1,642)
- Predicted (482)
- Uncertain (117)

**Annotation score**

- 5 (5,402)
- 4 (3,320)
- 3 (3,881)
- 2 (3,050)

The unreviewed UniProtKB/TrEMBL database will be reduced in size in release 2026\_02 (first half of 2026).

- Entries to be retained in UniProtKB:
  - Entries from reference proteomes
  - All reviewed (Swiss-Prot) entries
  - Selected unreviewed (TrEMBL) entries with experimental or biologically important data
- Entries to be removed: Unreviewed (TrEMBL) entries that are not part of a reference proteome

Entries removed from unreviewed UniProtKB/TrEMBL will remain accessible in the UniParc sequence archive. Please read our [help page](#), view [affected entries and proteomes](#), or contact us with any questions.

### UniProtKB 16,484 results

Tools Download (16k) Add View: Cards Table Customize columns Share

Entry	Entry Name	Protein Names	Gene Names	Organism	Length	Active site
Q08881	ITK_HUMAN	Tyrosine-protein kinase ITK/TSK[...]	ITK, EMT, LYK	Homo sapiens (Human)	620 AA	Active site

UniProt BLAST Align Peptide search ID mapping SPARQL UniProtKB 3702 Advanced | List Search Help

Binary interaction (219) More items

**Binding site (7,715)**

Protein existence

- Predicted (58,637)
- Homology (42,738)
- Protein level (15,824)
- Transcript level (9,138)

**Annotation score**

- 5 (12)
- 4 (183)

The unreviewed UniProtKB/TrEMBL database will be reduced in size in release 2026\_02 (first half of 2026).

- Entries to be retained in UniProtKB:
  - Entries from reference proteomes
  - All reviewed (Swiss-Prot) entries
  - Selected unreviewed (TrEMBL) entries with experimental or biologically important data
- Entries to be removed: Unreviewed (TrEMBL) entries that are not part of a reference proteome

Entries removed from unreviewed UniProtKB/TrEMBL will remain accessible in the UniParc sequence archive. Please read our [help page](#), view [affected entries and proteomes](#), or contact us with any questions.

### UniProtKB 126,337 results

Tools Download (126k) Add View: Cards Table Customize columns Share

Entry	Entry Name	Protein Names	Gene Names	Organism	Length	Active site
Q8L742	Q8L742_ARATH	Amine oxidase[...]	CUAO, At4g12290, T4C9_130, T4C9_130	Arabidopsis thaliana	741 AA	Active site

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；

UniProt BLAST Align Peptide search ID mapping SPARQL UniProtKB 3702 Advanced | List Search Help

Status

- Reviewed (Swiss-Prot) (16,484)

Popular organisms

- A. thaliana (16,418)
- Human (3)
- Bovine (2)
- Rat (2)
- Mouse (1)

Taxonomy

Filter by taxonomy

Group by

- Taxonomy
- Keywords
- Gene Ontology
- Enzyme Class

**Proteins with 3D structure (1,127)**

- Active site (2,987)
- Activity regulation (730)
- Allergen (1)
- Alternative products (isoforms) (3,232)
- Alternative splicing (1,906)

The unreviewed UniProtKB/TrEMBL database will be reduced in size in release 2026\_02 (first half of 2026).

- Entries to be retained in UniProtKB:
  - Entries from reference proteomes
  - All reviewed (Swiss-Prot) entries
  - Selected unreviewed (TrEMBL) entries with experimental or biologically important data
- Entries to be removed: Unreviewed (TrEMBL) entries that are not part of a reference proteome

Entries removed from unreviewed UniProtKB/TrEMBL will remain accessible in the UniParc sequence archive. Please read our [help page](#), view [affected entries and proteomes](#), or contact us with any questions.

### UniProtKB 16,484 results

Tools Download (16k) Add View: Cards Table Customize columns Share

Entry	Entry Name	Protein Names	Gene Names	Organism	Length	Active site	Binding site
Q08881	ITK_HUMAN	Tyrosine-protein kinase ITK/TSK[...]	ITK, EMT, LYK	Homo sapiens (Human)	620 AA	Active site	Binding site

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

Try the New NCBI Taxonomy Pages!  
Explore our redesigned taxonomy browser and taxonomy record pages with faster, more intuitive search, taxonomy images, and links to NCBI Datasets and other data available at NCBI.

Entrez records

Database name	Direct links	Subtree links	Links from type
BioProject	10,638	10,638	-
BioSample	237,319	237,319	-
Conserved Domains	53	53	-
GEO DataSets	119,817	119,817	-
Gene	44,112	44,112	-
Identical Protein Groups	165,831	165,831	-
Nucleotide	2,703,943	2,703,943	-
PubChem BioAssay	213	213	-
PMC	100,046	100,046	-
Protein	470,063	470,063	-
SRA	236,661	236,661	-
Structure	2,663	2,663	-
Taxonomy	1	1	-

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；

Phytozome 14 THE PLANT GENOMICS RESOURCE

Welcome to Phytozome

Overview Release Notes News

1. Arabidopsis thaliana  
2. Arabidopsis thaliana TAIR10 – thale cress  
Arabidopsis thaliana Araport11 – thale cress

0 genomes selected  
build custom data sets

Organism Information: Arabidopsis thaliana Araport11  
part of the GeneAtlas project  
Phytozome genome ID: 447 · NCBI taxonomy ID: 3702

Organism Information: Arabidopsis thaliana TAIR10  
part of the GeneAtlas project  
Phytozome genome ID: 167 · NCBI taxonomy ID: 3702

Data Source: ARAPORT: ARABIDOPSIS INFORMATION PORTAL

Data Source: TAIR: The Arabidopsis Information Resource

Genome Overview

Genome Information

Assembly Source:	TAIR
Assembly Version:	TAIR9
Annotation Source:	Araport
Annotation Version:	Araport11
Total Scaffold Length (bp):	119,667,750
Number of Scaffolds:	7
Min. Number of Scaffolds containing half of assembly (L50):	3
Shortest Scaffold from L50 set (NS0):	23,459,830
Total Contig Length (bp):	119,482,012
Number of Contigs:	169
Min. Number of Contigs containing half of assembly (L50):	5
Shortest Contig from L50 set (NS0):	10,898,021
Number of Protein-coding Transcripts:	48,456
Number of Protein-coding Genes:	27,655
Percentage of BUSCO Genes:	Embryophyta (OrthoDB v9): 99.3 Eukaryota (OrthoDB v9): 98.7

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。

Entrez records			
Database name	Direct links	Subtree links	Links from type
BioProject	11,815	16,285	-
BioSample	644,396	696,309	-
Conserved Domains	34	50	-
GEO DataSets	17,723	30,337	-
Gene	58,900	383,522	-
GTR	5	7	-
Identical Protein Groups	24,195,880	24,595,357	-
Nucleotide	14,584,069	15,907,669	709
PubChem BioAssay	23,898	24,963	-
PMC	475,199	475,209	-
Protein	76,203,586	89,740,011	-
SRA	541,908	592,801	-
Structure	8,212	13,918	-
Taxonomy	3,764	3,764	-

这张表展示了大肠杆菌在 NCBI 各数据库中的数据条目数，Direct 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。

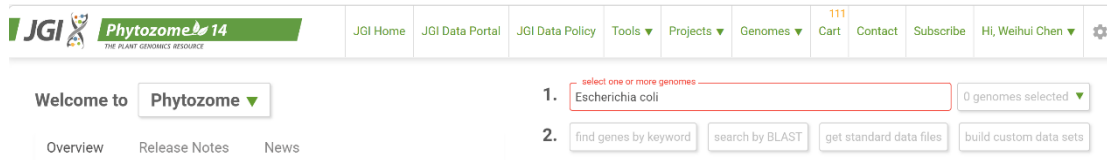
核酸序列、蛋白序列数据量极大,分别超 1458 万和 7620 万条;

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](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](#)

The screenshot shows the UniProtKB search interface with the following table of results:

Entry	Protein Name	Gene Name	Organism	Length	
P17666	MDH1L_SORBI	Mate dehydrogenase [NADP] 1, chloroplast	Sorghum bicolor (Sorghum) (Sorghum vulgare)	429 AA	
A0QW53	OMT3_SORBI	5-pentacetylcatechol O-methyltransferase	OMT3_S00Q50020	Sorghum bicolor (Sorghum) (Sorghum vulgare)	374 AA
Q4G202	MYB1L_SORBI	Transcription factor Y1L	y1	Sorghum bicolor (Sorghum) (Sorghum vulgare)	383 AA
Q98B11	HRNGT_SORBI	UDP-glucose 6-phosphatase	UGT85B1_HRNGT	Sorghum bicolor (Sorghum) (Sorghum vulgare)	492 AA
FR4516	PERL_SORBI	Cationic peroxidase SPCL	S0036046810	Sorghum bicolor (Sorghum) (Sorghum vulgare)	362 AA
Q94P91	TGMO_SORBI	Trans-cinnamate 4-monooxygenase	CYP79A33_C4H4_C4H1_SORBI_30002126400	Sorghum bicolor (Sorghum) (Sorghum vulgare)	501 AA
P52708	HNLS_SORBI	P-05-hydroxymandelonitrile lyase		Sorghum bicolor (Sorghum) (Sorghum vulgare)	510 AA
Q43135	CP9A1_SORBI	Tyrosine N-monooxygenase	CYP79A1_CYP79	Sorghum bicolor (Sorghum) (Sorghum vulgare)	558 AA

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](#)

The screenshot shows the UniProtKB search interface with the following table of results:

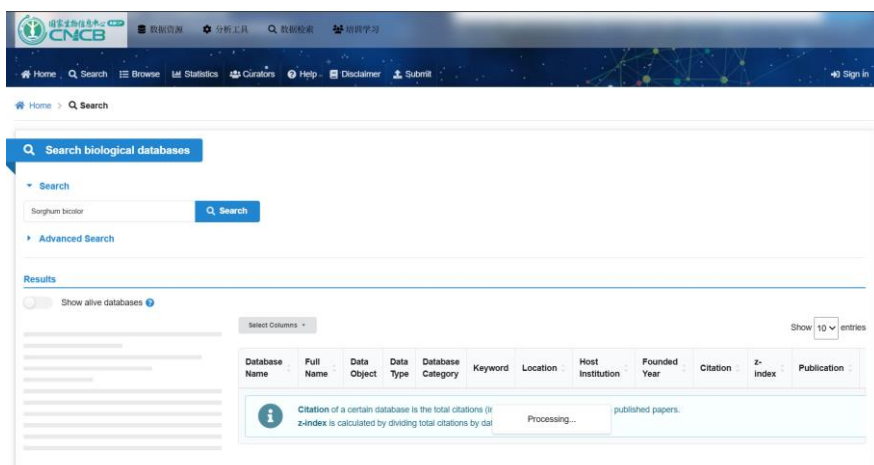
Entry	Protein Name	Gene Name	Organism	Length	
P17666	MDH1L_SORBI	Mate dehydrogenase [NADP] 1, chloroplast	Sorghum bicolor (Sorghum) (Sorghum vulgare)	429 AA	
A0QW53	OMT3_SORBI	5-pentacetylcatechol O-methyltransferase	OMT3_S00Q50020	Sorghum bicolor (Sorghum) (Sorghum vulgare)	374 AA
Q4G202	MYB1L_SORBI	Transcription factor Y1L	y1	Sorghum bicolor (Sorghum) (Sorghum vulgare)	383 AA
Q98B11	HRNGT_SORBI	UDP-glucose 6-phosphatase	UGT85B1_HRNGT	Sorghum bicolor (Sorghum) (Sorghum vulgare)	492 AA
FR4516	PERL_SORBI	Cationic peroxidase SPCL	S0036046810	Sorghum bicolor (Sorghum) (Sorghum vulgare)	362 AA
Q94P91	TGMO_SORBI	Trans-cinnamate 4-monooxygenase	CYP79A33_C4H4_C4H1_SORBI_30002126400	Sorghum bicolor (Sorghum) (Sorghum vulgare)	501 AA
P52708	HNLS_SORBI	P-05-hydroxymandelonitrile lyase		Sorghum bicolor (Sorghum) (Sorghum vulgare)	510 AA
Q43135	CP9A1_SORBI	Tyrosine N-monooxygenase	CYP79A1_CYP79	Sorghum bicolor (Sorghum) (Sorghum vulgare)	558 AA

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](#)





#### ④ 粳稻

1) 中文名: 粳稻

英文名: Japanese rice

拉丁学名: *Oryza sativa* subsp. *japonica* Kato

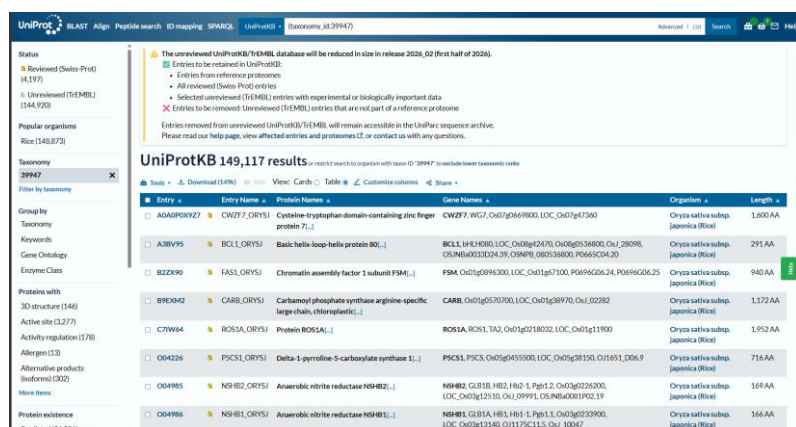
分类学登录号 (NCBI Taxonomy ID) : 39947

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](#)

Entry	Entry Name	Protein Names	Gene Names	Organism	Length
A0A0P9K3Z7	CWZ7L_ORYSJ	Cysteine-tryptophan domain-containing zinc finger protein 71	CWZ77, W57, Ory1g069800, LOC_Os1g47360	Oryza sativa subsp. japonica (Rice)	1,600 AA
A3BV95	BCL1_ORYSJ	Basic helix-loop helix protein 90	BCL1, BHLH90, LOC_Os08g42470, Os08g0536800, OsJ_28098, OS.NBA0033024.39, OS.NPB_080536800, P0665C04.20	Oryza sativa subsp. japonica (Rice)	291 AA
C7HW64	ROSLA_ORYSJ	Protein ROS1A	ROSLA, ROS1, TAD, Ory1g0218032, LOC_Os1g11900	Oryza sativa subsp. japonica (Rice)	1,952 AA
O04986	NSH1L_ORYSJ	Anaerobic nitrite reductase NSH1L	NSH1L, GLB1A, HBI, HBI-1, Pgl1-1, Ory1g0233900, LOC_Os1g13140, Q1175C11.5, OsJ_30047	Oryza sativa subsp. japonica (Rice)	166 AA
P0D002	DIAT8_ORYSJ	Chaperone protein dnaJ 1A7B, chloroplast	DIAT8, DIA7, Ory1g0193000, LOC_Os05g20026, O11005, D04.17, OS.NBA0049013.3	Oryza sativa subsp. japonica (Rice)	447 AA
P17814	4CL1_ORYSJ	4-coumarate-CoA ligase 1	4CL1, 4CL, Ory1g0245200, LOC_Os08g14760, O11033, B09.16	Oryza sativa subsp. japonica (Rice)	564 AA
P29218	CDKA1_ORYSJ	Cyclin-dependent kinase A-1	CDKA-1, CDC-1, Ory1g0118400, LOC_Os1g02680	Oryza sativa subsp. japonica (Rice)	294 AA
P49027	GBLPA_ORYSJ	Small ribosomal subunit protein RACK1Z	RACK1A, Ory1g0696000, LOC_Os1g41290	Oryza sativa subsp. japonica (Rice)	334 AA

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](#)

Entry	Entry Name	Protein Names	Gene Names	Organism	Length
O04986	NSH1L_ORYSJ	Anaerobic nitrite reductase NSH1L	NSH1L, GLB1A, HBI, HBI-1, Pgl1-1, Ory1g0233900, LOC_Os1g13140, Q1175C11.5, OsJ_30047	Oryza sativa subsp. japonica (Rice)	166 AA
Q01401	GLGB_ORYSJ	1,4-alpha-glucan branching enzyme, chloroplast	SBE1, REE1, Ory1g0726400, LOC_Os06g11084, P0017G10.8.1, P0017G10.8.2, P0548E04.28.1, P0548E04.28.2	Oryza sativa subsp. japonica (Rice)	820 AA
Q0D426	GH38_ORYSJ	Indole-3-acetic acid-amido synthetase GH3.8	GH3.8, GH3.8, Ory1g0592600, LOC_Os07g40290, O11710, H11.110, OsJ_023992, OsJ_24963	Oryza sativa subsp. japonica (Rice)	605 AA
Q0D953	HKT2L_ORYSJ	Cation transporter HKT2L	HKT2L1, HKT1, Ory1g0701700, LOC_Os06g48810, P0599H10.0	Oryza sativa subsp. japonica (Rice)	530 AA
Q0J8A4	G3PC1_ORYSJ	Glyceraldehyde 3-phosphate dehydrogenase 1, cytosolic	GAPC1, GAPC, GAPDH, GPC, Ory1g0126300, LOC_Os08g02290, O11163, G68.15, OsJ_024858	Oryza sativa subsp. japonica (Rice)	337 AA
Q0JF02	CPS4_ORYSJ	Syn-copalyl diphosphate synthase, chloroplast	CPS4, CVC1, Ory1g0178300, LOC_Os04g09000, OS.NBA0099501.12	Oryza sativa subsp. japonica (Rice)	767 AA
Q65216	MDAR3_ORYSJ	Monodehydroascorbate reductase 3, cytosolic	MDAR3, MDAR3L, Ory1g0567300, LOC_Os09g39380, O11355, H10.27	Oryza sativa subsp. japonica (Rice)	435 AA
Q6F6A2	PHR_ORYSJ	Deoxyribodipyrimidine photo-lyase	PHR, Ory1g0167600, LOC_Os11g08580, OS.NBA0013103.12	Oryza sativa subsp. japonica (Rice)	506 AA

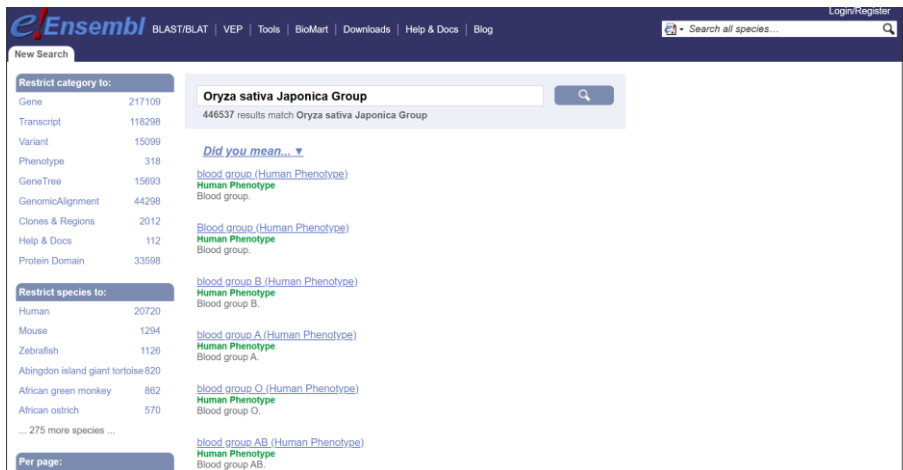
6) 查阅该物种在 NCBI 分类学网站中与其它数据库的交叉链接，列表说明其基本信息。

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

The screenshot displays the NCBI Taxonomy browser interface for the Oryza sativa Japonica Group. It shows a hierarchical tree structure with various taxonomic ranks and associated database links. The interface includes a search bar, navigation options, and a list of taxonomic entries with their corresponding database identifiers and descriptions.

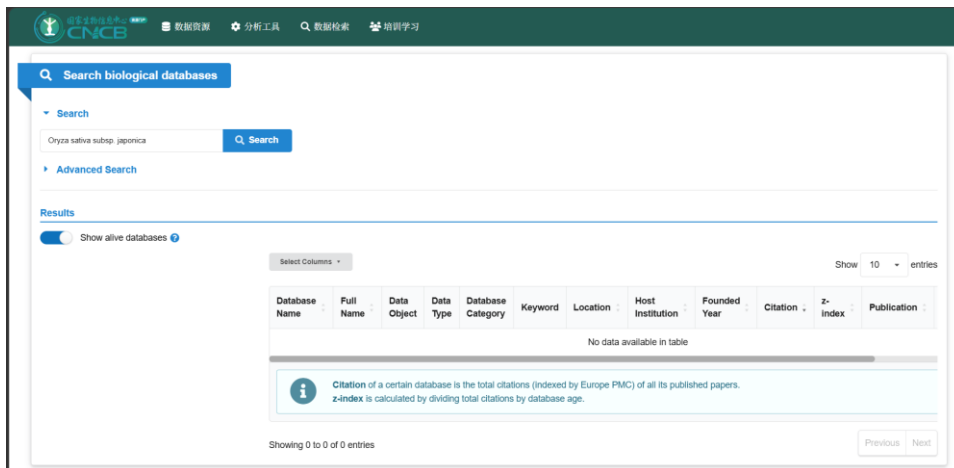
7) 查阅 Ensembl 或 Phytozome 等基因组数据库，若该物种已经完成基因组测序，熟悉其基因组基本信息；若该物种尚未测序，找出与其亲缘关系最近的物种，熟悉其基本信息。

[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。

UniProtKB (organism\_id:542762)

Status: Reviewed (Swiss-Prot) (2), Unreviewed (TrEMBL) (30,216)

Taxonomy: 542762

Group by: Taxonomy

Proteins with: 3D structure (1), Active site (761), Activity regulation (10), Binding site (2,103), Biophysicochemical properties (2)

UniProtKB 30,218 results

Entry	Entry Name	Protein Names	Gene Names	Organism	Length
A0A454ESS1	ALADC_CAMSN	Alanine decarboxylase[...]	AlaDC, TEA_005062	Camellia sinensis var. sinensis (China tea)	478 AA
A0A454DBB3	SERDC_CAMSN	Serine decarboxylase[...]	SerDC, TEA_019109	Camellia sinensis var. sinensis (China tea)	484 AA
A0A075TQM6	A0A075TQM6_CAMSN	Photosystem II protein D1[...]	psbA	Camellia sinensis var. sinensis (China tea)	353 AA
A0A454D5J0	A0A454D5J0_CAMSN	DNA replication licensing factor MCM7[...]	TEA_015704	Camellia sinensis var. sinensis (China tea)	1,481 AA
A0A454DEV1	A0A454DEV1_CAMSN	Adenylyltransferase and sulfurtransferase MOCS3[...]	MOCS3, CNX5, UBA4, TEA_004619	Camellia sinensis var. sinensis (China tea)	448 AA
A0A454E2W5	A0A454E2W5_CAMSN	Phosphatidylserine decarboxylase proenzyme	PSD2, TEA_005211	Camellia sinensis var. sinensis (China tea)	484 AA

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。

UniProtKB (taxonomy\_id:542762) AND (existence:1)

Status: Reviewed (Swiss-Prot) (2)

Taxonomy: 542762

Group by: Taxonomy

Proteins with: 3D structure (1), Binding site (2), Biophysicochemical properties (2), Catalytic activity (2), Chain (2)

UniProtKB 2 results

Entry	Entry Name	Protein Names	Gene Names	Organism	Length	Protein existence
A0A454ESS1	ALADC_CAMSN	Alanine decarboxylase [...]	AlaDC, TEA_005062	Camellia sinensis var. sinensis (China tea)	478 AA	1: Evidence at protein level
A0A454DBB3	SERDC_CAMSN	Serine decarboxylase [...]	SerDC, TEA_019109	Camellia sinensis var. sinensis (China tea)	484 AA	1: Evidence at protein level

5) 该物种在 Swiss-Prot 子库中具有三维空间结构的序列条目数。

UniProtKB (organism\_id:542762)

Status: Reviewed (Swiss-Prot) (2)

Taxonomy: 542762

Group by: Taxonomy

Proteins with: 3D structure (1), Binding site (2), Biophysicochemical properties (2), Catalytic activity (2), Chain (2)

UniProtKB 2 results

Entry	Entry Name	Protein Names	Gene Names	Organism	Length	3D structures
A0A454ESS1	ALADC_CAMSN	Alanine decarboxylase[...]	AlaDC, TEA_005062	Camellia sinensis var. sinensis (China tea)	478 AA	X-ray; 2
A0A454DBB3	SERDC_CAMSN	Serine decarboxylase[...]	SerDC, TEA_019109	Camellia sinensis var. sinensis (China tea)	484 AA	

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。

**Camellia sinensis var. sinensis**

Taxonomy ID: 542762 (for references in articles please use ncbitaxon:542762)

current name

*Camellia sinensis var. sinensis*, *nominotypical infraspecies* <sup>1</sup>

NCBI BLAST name: **eudicots**

Rank: **varietas**

Genetic code: [Translation table 1 \(Standard\)](#)

Mitochondrial genetic code: [Translation table 1 \(Standard\)](#)

Other names:

heterotypic synonym

*Camellia longlingensis*

Lineage (full)

[cellular organisms](#); [Eukaryota](#); [Viridiplantae](#); [Streptophyta](#); [Streptophytina](#); [Embryophyta](#); [Tracheophyta](#); [Euphyllophyta](#); [Spermatophyta](#); [Magnoliopsida](#); [Mesangiospermae](#); [eudicotyledons](#); [Gunneridae](#); [Pentapetalae](#); [asterids](#); [Ericales](#); [Theaceae](#); [Camellia](#); [Camellia sinensis](#)

Notes:

1) This nominotypical infraspecies name or autonym shares type material with its parent species.

External Information Resources (NCBI LinkOut)

LinkOut	Subject	LinkOut Provider
<a href="#">WebScipio: Camellia sinensis var. sinensis cultivar Fudingdabaicha</a>	organism-specific	<a href="#">WebScipio - eukaryotic gene identification</a>
<a href="#">WebScipio: Camellia sinensis var. sinensis</a>	organism-specific	<a href="#">WebScipio - eukaryotic gene i</a>

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](#)

Organism Information:

**Camelina sativa Acsn-226 v1.1**

part of the BAP project

Phytozome genome ID: 805 • NCBI taxonomy ID: 90675

[UNREF](#) [BAP](#) [JOB](#)

[Keyword search](#) [Blast search](#) [JBrowse](#) [Download](#)

**Project Overview**

The Brassicaceae Adaptation Project, BAP, a Joint Genome Institute Community Science Program project, encompasses five species: an existing crop, two new oilseed/biofuel crops and two species that balance the evolutionary distance between the other three species. Further, these five species all have extensive levels of metabolic and transcriptional diversity within both primary and specialized metabolism. Most critically, they are all relatives of the model plant *Arabidopsis thaliana* which has a large list of gold standard genes that are known to influence adaptation and metabolism. By assembling high-quality reference genomes in this collection and comparing these results to the *Arabidopsis* gold standards, it will be possible to directly test if various methods to identify key "adaptive" genes within and between species are identifying known loci.

Another key use of genomic variation is to evaluate the relative role of structural and SNP variation in the evolution of different adaptive traits. SNP variation is often linked to variation in primary metabolism and the change in gene expression patterns. In contrast, structural variants (including PAVs) are more frequently linked to biotic adaptation such as diversity in disease resistance and specialized metabolic gene families. Existing genomic variation studies largely rely on using short-read resequencing and alignment against a common reference genome to generate the necessary genomic variation information. While it is acknowledged that this approach has errors, it is fundamentally assumed that these errors are unbiased with regards to mechanism or function. However, recent work is beginning to show that this short-read resequencing approach has a potential bias whereby it underestimates genomic variation, both SNP and structural, within genes of adaptive potential.

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 > Pentapetales > rosids > fabids > Malpighiales > Salicaceae > Saliceae > Populus (poplars)

3) 该物种在 Swiss-Prot 和 TrEMBL 子库中序列条目数。

The screenshot shows the UniProtKB search interface. The search criteria are 'UniProtKB (taxonomy\_id:3696)'. The results page displays 40,010 results. A table lists the following entries:

Entry	Entry Name	Protein Names	Gene Names	Organism	Length	Protein existence
P36491	PSBA_POPDE	Photosystem II protein D1[...]	psbA	Populus deltoides (Eastern poplar) (Eastern cottonwood)	353 AA	3: Inferred from homology
Q09117	BSPB_POPDE	Bark storage protein B	BSP	Populus deltoides (Eastern poplar) (Eastern cottonwood)	312 AA	1: Evidence at protein level
P47916	METK_POPDE	S-adenosylmethionine synthase[...]	METK	Populus deltoides (Eastern poplar) (Eastern cottonwood)	395 AA	2: Evidence at transcript level
P31657	CADH_POPDE	Probable cinnamyl alcohol dehydrogenase[...]		Populus deltoides (Eastern poplar) (Eastern cottonwood)	357 AA	2: Evidence at transcript level

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。

The screenshot shows the UniProtKB search interface with the search criteria 'UniProtKB (taxonomy\_id:3696) AND (existence:1)'. The results page displays 1 result:

Entry	Entry Name	Protein Names	Gene Names	Organism	Length	Protein existence
Q09117	BSPB_POPDE	Bark storage protein B	BSP	Populus deltoides (Eastern poplar) (Eastern cottonwood)	312 AA	1: Evidence at protein level

4) 该物种在 Swiss-Prot 子库中具有三维空间结构的序列条目数。

The screenshot shows the UniProtKB search interface with the search criteria 'UniProtKB (taxonomy\_id:3696) AND (structure\_3d:true)'. The results page displays a message: 'Sorry, no results were found! If you can't find what you are looking for, please contact us.'

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。

NCBI Taxonomy Browser

Try the New NCBI Taxonomy Pages!  
Explore our redesigned taxonomy browser and taxonomy record pages with faster, more intuitive search, taxonomy images, and links to NCBI Datasets and other data available at NCBI.

Search for: [complete name] [lock] [search] [Clear]

**Populus deltoides**

Taxonomy ID: 3696 (for references in articles please use ncbitaxon:3696)

current name  
**Populus deltoides** W.Bartram ex Marshall, 1785  
(includes: **Populus sp. YRR-2025a**)

NCBI BLAST name: **euclidots**

Rank: **species**

Genetic code: [Translation table 1 \(Standard\)](#)

Mitochondrial genetic code: [Translation table 1 \(Standard\)](#)

Lineage (full)  
[cellular organisms](#); [Eukaryota](#); [Viridiplantae](#); [Streptophyta](#); [Streptophytina](#); [Embryophyta](#); [Tracheophyta](#); [Euphyllophyta](#); [Spermatophyta](#); [Magnoliopsida](#); [Mesangiospermae](#); [euclidoyledons](#); [Gunneridae](#); [Pentapetalae](#); [rosids](#); [fabids](#); [Malpighiales](#); [Salicaceae](#); [Saliceae](#); [Populus](#)

Comments and References:

[image:Populus deltoides](#)  
Image by Laurent Balanger from Wikimedia Commons under a CC BY-SA license.  
\* Image may not have been verified for accuracy by NCBI Taxonomy.

[FNA - Populus](#)  
Eckenwalder JE. 2009. Populus Linnaeus. Sp. Pl. 2: 1034. 1753; Gen. Pl. ed. 5, 456. 1754. In Flora of North America Editorial Committee (Eds.) Flora of North America North of Mexico. Vol. 7; Magnoliophyta: Salicaceae to Brassicaceae. New York and Oxford. On-line version.

Database name	Direct links	Subtree links	Links from type
BioProject	1,031	1,031	-
BioSample	3,441	3,441	-
GEO DataSets	1,075	1,075	-
Gene	131	131	-
Identical Protein Groups	40,266	40,266	-
Nucleotide	15,584	15,584	-
PMC	1,179	1,179	-
Protein	45,628	45,628	-
SRA	3,339	3,339	-
Taxonomy	1	1	-

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.deltoides WV94 v2.1: Phytozome](#)

JGI Phytozome 14 THE PLANT GENOMICS RESOURCE

JGI Home | JGI Data Portal | JGI Data Policy | Tools | Projects | Genomes | Cart | Contact | Subscribe | Login

Organism Information:  
**Populus deltoides WV94 v2.1**  
Phytozome genome ID: 445 • NCBI taxonomy ID: 3696

Keyword search | Blast search | JBrowse | Synteny | Download

**Genome Overview**

WV94 is a *Populus deltoides* (Marsh.) clone from Issaquena Co., Mississippi. It was first identified by the U.S. Forest Service for its rapid growth under field conditions [Coyle et al. 2006]. ArborGen LLC subsequently selected this clone for use in its transformation program because of WV94's ability to regenerate whole plants under in vitro conditions. The BioEnergy Science Center (BESC) at Oak Ridge National Laboratory has tested transgenic lines of WV94 as part of the U.S. Department of Energy's biofuels program.

**Genome Information**

Assembly Source:	JGI
Assembly Version:	v2.0
Annotation Source:	JGI
Annotation Version:	v2.1
Total Scaffold Length (bp):	446,783,942
Number of Scaffolds:	1,375
Min. Number of Scaffolds containing half of assembly (L50):	8
Shortest Scaffold from L50 set (N50):	21,697,863
Total Contig Length (bp):	432,523,942
Number of Contigs:	2,801
Min. Number of Contigs containing half of assembly (L50):	701

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。

NCBI Database Common

Select a country: [ ] Select an institution: [ ]

Data Type: [ ] Database Category: [ ] Data Object: [ ] Organism(s): [Populus deltoides]

Search [X Clear]

Results

Show alive databases

Countries

Institutions

Database Name	Full Name	Data Object	Data Type	Database Category	Keyword	Location	Host Institution	Founded Year	Citation	Z-index	Publi
WallProtDB	Wall Proteomics Database	Protein	Protein	Genotype and trait variation, Expression, Literature	Cell line, Gene, Microarray, Proteomics	France	University of Toulouse	2015	48	4.36	2014
GreenCircRNA		Plant	RNA	Genotype and trait variation, Literature	Cell line, Gene, Microarray, Proteomics	China	Shanxi Normal University	2020	20	3.33	2018

Citation of a certain database is the total citations (indexed by Europe PMC) of all its published papers. z-index is calculated by dividing total citations by database age.

**Database Profile**

**WallProtDB**

**General information**

URL: <http://www.polebio.irsv.upvs-tlse.fr/WallProtDB>

Full name: Wall Proteomics Database

Description: WallProtDB (<http://www.polebio.irsv.upvs-tlse.fr/WallProtDB/>) presently contains 3401 proteins and ESTs identified experimentally in about 50 cell wall proteomics studies performed on 13 different plant species. Two criteria have to be met for entering WallProtDB. First one is related to the identification of proteins. Only proteins identified in plant with available genomic or ESTs data are considered to ensure unambiguous identification. Second criterion is related to the difficulty to obtain clean cell wall fractions. Indeed, since cell walls constitute an open compartment difficult to isolate, numerous proteins predicted to be intracellular and/or having functions inside the cell have been identified in cell wall extracts. Then, except proteins predicted to be plasma membrane proteins, only proteins having a predicted signal peptide and no known intracellular retention signal are included in the database. In addition, WallProtDB contains information about the strategies used to obtain cell wall protein extracts and to identify proteins by mass spectrometry and bioinformatics. Mass spectrometry data are included when available. All the proteins of WallProtDB are linked to ProtAmdB, another database, which contains structural and functional bioinformatics annotations of proteins as well as links to other databases (Aramemnon, CAZy, Planet, Phytosome, UniProt). A list of references in the cell wall proteomics field is also provided.

Year founded: 2015

Last update: 2017

Version:

**Ranking**

All databases: 2432 (100%)

Genotype phenotype and variation: 3577 (147%)

Expression: 447 (18%)

Literature: 227 (9%)

**TOTAL RANK**

2432

48 CITATIONS

4.364 Z-INDEX

**Community reviews**

Data quality & quantity: ★★★★★

Content organization & presentation: ★★★★★

System accessibility & reliability: ★★★★★

Submit a review

## (7) 课题相关蛋白信息

### ① CKX5-ARATH

1) 该序列条目的蛋白名、基因名和物种名:

#### Q67YU0 · CKX5\_ARATH

Protein <sup>1</sup>	Cytokinin dehydrogenase 5	Amino acids	540 (go to sequence)
Gene <sup>1</sup>	CKX5	Protein existence <sup>1</sup>	Evidence at transcript level
Status <sup>1</sup>	UniProtKB reviewed (Swiss-Prot)	Annotation score <sup>1</sup>	4/5
Organism <sup>1</sup>	Arabidopsis thaliana (Mouse-ear cress)		

2) 该序列条目是否经过人工审阅, 包括哪几大类注释信息:

- Function
- Names & Taxonomy
- Subcellular Location
- Phenotypes & Variants
- PTM/Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence
- Similar Proteins

3) 该序列条目收录了多少篇相关文献, 共分哪几大类:

**Source**

UniProtKB reviewed (Swiss-Prot) (7)

Computationally mapped (20)

**Category**

Function (13)

Expression (12)

Sequences (5)

Names (2)

Phenotypes & Variants (1)

**Study type**

Small scale (15)

Large scale (12)

**Publications for Q67YU0<sup>1</sup>**

**Molecular and biochemical characterization of a cytokinin oxidase from maize.**

Bilyeu K.D., Cole J.L., Laskey J.G., Riekhof W.R., Esparza T.J., Kramer M.D., Morris R.O.

View abstract

Cited for NUCLEOTIDE SEQUENCE [MRNA]

Category Sequences

Source UniProtKB reviewed (Swiss-Prot)

PubMed Europe PMC Plant Physiol. 125:378-386 (2001)

Cited in Mapped to

**Sequence and analysis of chromosome 1 of the plant Arabidopsis thaliana.**

Theologis A., Ecker J.R., Palm C.J., Federspiel N.A., Kaul S., White O., Alonso J., Altafi H., Araujo R. [..], Davis R.W.

PubMed Europe PMC

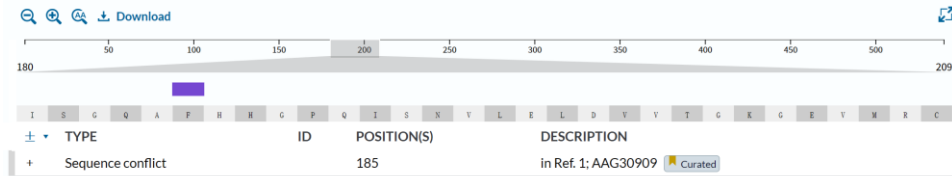
4) 该序列条目包括哪些特征位点信息?

## Sequence caution<sup>i</sup>

The sequence AAG13068.1 [differs](#) from that shown. Reason: Erroneous gene model prediction Curated

## Features

Showing features for sequence conflict<sup>1</sup>.



## Keywords<sup>i</sup>

- Coding sequence diversity [#Alternative splicing](#)
- Technical term [#Reference proteome](#)

## Sequence databases

PIR [B96785](#) [B96785](#)

RefSeq [NP\\_177678.2](#) [NM\\_106199.5](#) [\[Q67YU0-1\]](#)

5) 该序列条目包括哪几类数据库交叉链接，其中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？

## External Links

**Enzyme and pathway databases**  
BioCyc [ARA:AT1G75450-MONOMER](#) ENZYME [Search...](#)

**Family and domain databases**

Funfam	<a href="#">3.30.465.10:FF-000021</a> <a href="#">Cytokinin dehydrogenase 1</a> 1 hit	PANTHER	<a href="#">PTHR13878:SF102</a> <a href="#">CYTOKININ DEHYDROGENASE 5</a> 1 hit
	<a href="#">3.40.462.10:FF-000001</a> <a href="#">Cytokinin dehydrogenase 2</a> 1 hit		<a href="#">PTHR13878</a> <a href="#">GULONOLACTONE OXIDASE</a> 1 hit
Gene3D	<a href="#">3.30.465.10</a> <a href="#">1</a> hit	PROSITE	<a href="#">View protein in PROSITE</a>
	<a href="#">3.40.462.10</a> <a href="#">FAD-linked oxidases, C-terminal domain</a> 1 hit		<a href="#">P551387</a> <a href="#">FAD_PCMH</a> 1 hit
	<a href="#">3.30.43.10</a> <a href="#">Uridine Diphospho-n-acetylenolpyruvylglucosamine Reductase, domain 2</a> 1 hit		<a href="#">P900862</a> <a href="#">OX2_COVAL_FAD</a> 1 hit
InterPro	<a href="#">View protein in InterPro</a>	Pfam	<a href="#">View protein in Pfam</a>
	<a href="#">IPRO16170</a> <a href="#">Cytok_DH_C_sf</a>		<a href="#">PF09265</a> <a href="#">Cytokinin-blind</a> 1 hit
	<a href="#">IPRO15345</a> <a href="#">Cytokinin_DH_FAD_cytok-bd</a>		<a href="#">PF01565</a> <a href="#">FAD_binding_4</a> 1 hit
	<a href="#">IPRO16166</a> <a href="#">FAD-bd_PCMH</a>	SUPFAM	<a href="#">SSF56176</a> <a href="#">FAD-binding/transporter-associated domain-like</a> 1 hit
	<a href="#">IPRO36318</a> <a href="#">FAD-bd_PCMH-like_sf</a>		<a href="#">SSF51003</a> <a href="#">FAD-linked oxidases, C-terminal domain</a> 1 hit
	<a href="#">More InterPro links</a>	MobiDB	<a href="#">Search...</a>

**Gene expression databases**  
ExpressionAtlas [Q67YU0](#) [baseline and differential](#)

**Protein-protein interaction databases**  
STRING [3702:Q67YU0](#)

**Organism-specific databases**  
Araport [AT1G75450](#) TAIR [AT1G75450](#) [C10X5](#)

**Proteomic databases**  
PaxDb [3702:AT1G75450.1](#) ProteomicsDB [222094](#) [\[Q67YU0-1\]](#)

**PTM databases**  
GlyCosmos [Q67YU0](#) [2 sites](#), No reported glycans GlyGen [Q67YU0](#) [2 sites](#)

**Sequence databases**

NUCLEOTIDE SEQUENCE	PROTEIN SEQUENCE	MOLECULE TYPE	STATUS
AF303982	AAG30909.1	mRNA	
EMBL <a href="#">GenBank</a> <a href="#">DDBJ</a>	EMBL <a href="#">GenBank</a> <a href="#">DDBJ</a>		
AC023754	AAG13068.1	Genomic DNA	Sequence problems.
EMBL <a href="#">GenBank</a> <a href="#">DDBJ</a>	EMBL <a href="#">GenBank</a> <a href="#">DDBJ</a>		

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<sup>i</sup>

### UniRef clusters<sup>i</sup>

100% identity 90% identity 50% identity

Q67YU0-1

UniRef100\_Q67YU0

Protein name	Organism	Length
<a href="#">cytokinin dehydrogenase</a>	<a href="#">Arabidopsis thaliana (Mouse-ear cress)</a>	536
<a href="#">cytokinin dehydrogenase</a>	<a href="#">Arabidopsis thaliana (Mouse-ear cress)</a>	417
<a href="#">cytokinin dehydrogenase</a>	<a href="#">Arabidopsis thaliana (Mouse-ear cress)</a>	540
<a href="#">cytokinin dehydrogenase</a>	<a href="#">Arabidopsis thaliana x Arabidopsis arenosa</a>	540
<a href="#">cytokinin dehydrogenase</a>	<a href="#">Arabidopsis suecica (Swedish thale-cress) (Cardaminopsis suecica)</a>	540

View these 5 entries in UniProtKB

[View all](#)

## Similar Proteins<sup>i</sup>

### UniRef clusters<sup>i</sup>

100% identity 90% identity 50% identity

Q67YU0-1

UniRef90\_Q67YU0

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

View all 40 entries in UniProtKB

View all

## Similar Proteins<sup>i</sup>

### UniRef clusters<sup>i</sup>

100% identity 90% identity 50% identity

Q67YU0-1

UniRef50\_Q67YU0

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

**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<sup>i</sup>

**Name** | BHLH140

**Synonyms** | EN122

**ORF names** | T1008.20

**Ordered locus names** | At5g01310

### Organism names<sup>i</sup>

**Taxonomic identifier<sup>i</sup>** | 3702 (NCBI [↗](#))

**Organism<sup>i</sup>** | *Arabidopsis thaliana* (Mouse-ear cress)

**Strain** | cv. Columbia

**Taxonomic lineage<sup>i</sup>** | 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) > Camelinae > *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;

**Q9M041 • BHLH140**

Protein: Transcription factor bHLH140  
 Gene: BHLH140  
 Status: UniProtKB reviewed (Swiss-Prot)  
 Organism: Arabidopsis thaliana (Mouse-ear cress)

Amino acids: 912 (go to sequence)  
 Protein existence: Inferred from homology  
 Annotation score:

**Function**

**Features**  
 Showing features for binding site:

229 100 200 300 400 500 600 700 800 900 1000 238

3) 该序列条目收录了多少篇相关文献，共分哪几大类：

**Q9M041 • BHLH140**

Protein: Transcription factor bHLH140  
 Gene: BHLH140  
 Status: UniProtKB reviewed (Swiss-Prot)  
 Organism: Arabidopsis thaliana (Mouse-ear cress)

Amino acids: 912 (go to sequence)  
 Protein existence: Inferred from homology  
 Annotation score:

**Publications for Q9M041**

Sequence and analysis of chromosome 5 of the plant Arabidopsis thaliana.

Tabata S., Kaneko T., Nakamura Y., Kotani H., Kato T., Asamizu E., Miyajima N., Sasamoto S., Kimura T., Fransz P.F.

View abstract  
 Cited for NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA]  
 Strain cv. Columbia  
 Category Sequences  
 Source UniProtKB reviewed (Swiss-Prot)

PubMed  
 Europe PMC  
 Nature 408:823-826 (2000)

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？

**Features**  
 Showing features for region, compositional bias, domain, zinc finger:

100 200 300 400 500 600 700 800 900 1000 30

TYPE	ID	POSITION(S)	DESCRIPTION	Tools
Region		1-57	Disordered	
Compositional bias		13-23	Low complexity	
Domain		43-92	bHLH	
Domain		511-690	Macro	
Compositional bias		657-666	Polar residues	
Region		657-706	Disordered	
Domain		720-829	HIT	
Zinc finger		870-893	C2H2-type	

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?

**External Links**

**Enzyme and pathway databases**  
 BioCyc | ARA:AT5G61310-MONOMER

**Family and domain databases**

<b>CDD</b>	cd11454   bHLH_AIN1D_like 1 hit	<b>PROSITE</b>	View protein in PROSITE
<b>FunFam</b>	3.30.428.10.FF:000004   aprataxin isoform X2 1 hit	PS50888   BHLH 1 hit	
	3.40.220.10.FF:000020   Transcription factor bHLH140 1 hit	PS50892   HIT_1 1 hit	
	3.40.50.300.FF:002337   Transcription factor bHLH140 1 hit	PS51084   HIT_2 1 hit	
	4.10.280.10.FF:000089   Transcription factor LAX PANICLE 1 hit	PS51154   MACRO 1 hit	
<b>Gene3D</b>	4.10.280.10   Helix-loop-helix DNA-binding domain 1 hit	<b>Pfam</b>	View protein in Pfam
	3.30.428.10   HIT-like 1 hit	PF13673   AAA_33 1 hit	
	3.40.220.10   Leucine Aminopeptidase, subunit F, domain 1 1 hit	PF12868   Dcp3_C 1 hit	
	3.40.50.300   P-loop containing nucleoside triphosphate hydrolases 1 hit	PF00661   HLH 1 hit	
		PF01661   Macro 1 hit	
		PF16278   c-C2HE 1 hit	
<b>InterPro</b>	View protein in InterPro	<b>SMART</b>	View protein in SMART
	IPR011598   bHLH_dom	SM00586   Atgip 1 hit	
	IPR019608   Histidine_triad_CS	SM00353   HLH 1 hit	
	IPR011496   HIT-like	<b>SUPFAM</b>	SSF45497   HIT-like 1 hit
	IPR026265   HIT-like_sf	SSF47459   HLH_helix-loop-helix DNA-binding domain 1 hit	
	More InterPro links	SSF2949   Macro domain-like 1 hit	
		SSF2540   P-loop containing nucleoside triphosphate hydrolases 1 hit	
<b>PANTHER</b>	PTHR12486:SF4   APRATAXIN 1 hit	<b>MobiDB</b>	Search...
	PTHR12488   APRATAXIN-RELATED 1 hit		

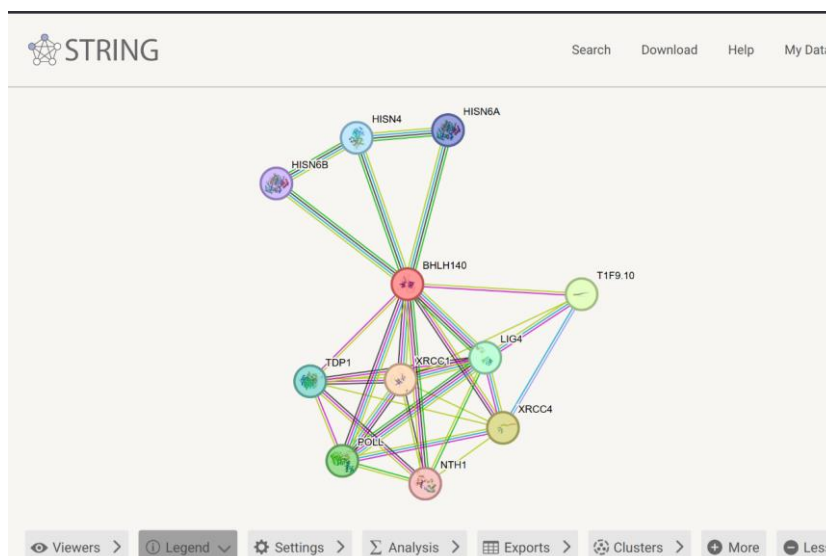
**Gene expression databases**  
 ExpressionAtlas | Q9M041 | baseline and differential

**Protein-protein interaction databases**  
 STRING | 3702.Q9M041

**Organism-specific databases**  
 Araport | AT5G61310 | TAIR | AT5G61310 | APTX

**Proteomic databases**  
 PaxDb | 3702-AT5G61310.1

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、TIF9.10

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, 这部分互动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：

#### Q43135 • C79A1\_SORBI

Protein <sup>i</sup>	Tyrosine N-monooxygenase	Amino acids	558 (go to sequence)
Gene <sup>i</sup>	CYP79A1	Protein existence <sup>i</sup>	Evidence at protein level
Status <sup>i</sup>	UniProtKB reviewed (Swiss-Prot)	Annotation score <sup>i</sup>	
Organism <sup>i</sup>	<i>Sorghum bicolor</i> (Sorghum) ( <i>Sorghum vulgare</i> )		

#### Function

#### Names & Taxonomy

#### Subcellular Location

#### Phenotypes & Variants

#### PTM/Processing

#### Expression

#### Interaction

#### Structure

#### Family & Domains

#### Sequence

#### Similar Proteins

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？

**Q43135 · C79A1\_SORBI**

Tyrosine N-monoxygenase  
 Gene: CYP79A1  
 Status: UniProtKB reviewed (Swiss-Prot)  
 Organism: Sorghum bicolor (Sorghum) (Sorghum vulgare)

Amino acids: 558 (go to sequence)  
 Protein existence: Evidence at protein level  
 Annotation score:

Entry Variant viewer Feature viewer Genomic coordinates Publications External links History

### External Links

**Enzyme and pathway databases**

BRENDA: 3.1.4.14.36 (†) ST06 UniPathway: UP00757UER00744  
 BioCyc: MetaCyc:MON0488-921 (†) ENZYME: Search... (†)  
 SABIO-RK: Q43135 (†)

**Family and domain databases**

CDD: s482658 (†) CYP79\_1 hit PRINTS: PR00463 (†) EP450  
 FunFam: 3.1.0.630.10.FPF000037 (†) Cytochrome P450 9.1 hit PROSITE: PR00385 (†) P450  
 Gene3D: 3.1.0.630.10 (†) Cytochrome P450 1 hit PROSITE: PR00088 (†) CYTOCHROME\_P450\_1 hit  
 InterPro: View protein in InterPro (†) IPRO01328 (†) Cyt\_P450 Pfam: View protein in Pfam (†)  
 IPRO01792 (†) Cyt\_P450\_C5 PROSITE: PR00807 (†) p450\_1 hit  
 IPRO02462 (†) Cyt\_P450\_E\_gcp1 SUPFAM: SSF48264 (†) Cytochrome P450 1 hit  
 IPRO38398 (†) Cyt\_P450\_LF MobiDB: Search... (†)  
 PANTHER: PTHR47944 (†) CYTOCHROME P450 38A9\_2 hit  
 PTHR47948 (†) CYP79A1 PROTEIN\_1 hit

**Gene expression databases**

ExpressionAtlas: Q43135 (†) baseline and differential

**Sequence databases**

PIR: S48202 (†) S48202

NUCLEOTIDE SEQUENCE	PROTEIN SEQUENCE	MOLECULE TYPE	STATUS
GenBank (†) · DDBJ (†)	EMBL (†) · GenBank (†) · DDBJ (†)	mRNA	

**Genome annotation databases**

EnsemblPlants: EER93097 (†) SORBI\_3001G012300 (†) Gramene: EER93097 (†) SORBI\_3001G012300 (†)  
 GeneID: 8061413 (†) KEGG: sbi8061413 (†)

**3D structure databases**

AlphaFoldDB: Q43135 (†) ModBase: Search... (†)  
 SMR: Q43135 (†)

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

**Q43135**  
 UniRef100\_Q43135

Protein name	Organism	Length
Tyrosine N-monoxygenase	Sorghum bicolor (Sorghum) (Sorghum vulgare)	558
Cytochrome P-450	Sorghum bicolor (Sorghum) (Sorghum vulgare)	16

View these 2 entries in UniProtKB

[View all](#)

**Orthologs & paralogs**  
 No Orthology or Paralogy data is available from the Alliance of Genome Resources.

**Phylogenomic databases**

HOGENOM: CLU\_001570\_4\_0\_1 (†) OrthoDB: 2789670at2759 (†)  
 OMA: KWKLAGG (†) eggNOG: KOG0156 (†) Eukaryota

### Similar Proteins

UniRef clusters:

100% identity 90% identity 50% identity

**Q43135**  
 UniRef90\_Q43135

Protein name	Organism	Length
Cytochrome P450	Miscanthus lutarioriparius	544
Cytochrome P450	Miscanthus lutarioriparius	473
Tyrosine N-monoxygenase	Sorghum bicolor (Sorghum) (Sorghum vulgare)	558
Cytochrome P450	Miscanthus lutarioriparius	513
Cytochrome P450	Miscanthus lutarioriparius	554
Tyrosine N-monoxygenase	Sorghum bicolor (Sorghum) (Sorghum vulgare)	558
Cytochrome P-450	Sorghum bicolor (Sorghum) (Sorghum vulgare)	16

View these 7 entries in UniProtKB

[View all](#)

**Orthologs & paralogs**  
 No Orthology or Paralogy data is available from the Alliance of Genome Resources.

**Phylogenomic databases**

## Similar Proteins

### UniRef clusters

100% identity 90% identity 50% identity

Q43135

UniRef50\_Q43135

Protein name	Organism	Length
Tyrosine N-monoxygenase	Triticum aestivum (Wheat)	549
Cytochrome P450	Digitaria exilis	579
Tyrosine N-monoxygenase	Triticum urartu (Red wild einkorn) (Crotodium urartu)	473
Cytochrome P450	Triticum aestivum (Wheat)	486
Tyrosine N-monoxygenase	Triticum aestivum (Wheat)	536
Tyrosine N-monoxygenase	Triticum aestivum (Wheat)	502
Tyrosine N-monoxygenase	Aegilops tauschii subsp. strangulata (Goatgrass)	536
Cytochrome P450	Miscanthus lutarioriparius	513
Cytochrome P450	Miscanthus lutarioriparius	554
Tyrosine N-monoxygenase	Sorghum bicolor (Sorghum) (Sorghum vulgare)	558

View all 37 entries in UniProtKB

View all

## (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?

Try the New NCBI Taxonomy Pages! Explore our redesigned taxonomy browser and taxonomy record pages with faster, more intuitive search, taxonomy images, and links to NCBI Datasets and other data available at NCBI.

Search for: [complete name] lock search Clear

**Arabidopsis thaliana**  
Taxonomy ID: 3702 (for references in articles please use ncbitaxon:3702)  
current name: *Arabidopsis thaliana* (L.) Heynh., 1842  
[basonym: *Arabis thaliana* L., 1753]

Genbank common name: **thale cress**  
NCBI BLAST name: **euclids**  
Rank: **species**  
Genetic code: [Translation table 1 \(Standard\)](#)  
Mitochondrial genetic code: [Translation table 1 \(Standard\)](#)  
Other names:  
common name(s): **thale-cress, mouse-ear cress**

Lineage (full)  
cellular organisms; Eukaryota; Viridiplantae; Streptophyta; Streptophytina; Embryophyta; Tracheophyta; Euphyllophyta; Spermatophyta; Magnoliopsida; Mesangiospermae; eudicotyledons; Gunneridae; Pentapetalae; rosids; malvids; Brassicales; Brassicaceae; Camelinae; Arabidopsis

Database name	Direct links	Subtree links	Links from type
BioProject	10,638	10,638	-
BioSample	237,319	237,319	-
Conserved Domains	53	53	-
GEO DataSets	119,817	119,817	-
Gene	44,112	44,112	-
Identical Protein Groups	165,831	165,831	-
Nucleotide	2,703,943	2,703,943	-
PubChem BioAssay	213	213	-
PMC	100,046	100,046	-
Protein	470,063	470,063	-
SRH	430,004	430,004	-
Structure	2,663	2,663	-
Taxonomy	1	1	-

National Library of Medicine  
National Center for Biotechnology Information

Search 0000-0003-0479-7953@orcid@orcid

PMC PubMed Central® txid3702[Organism:noexp] Search in PMC

Journal List | User Guide

Save Sort by: Relevance Display Options

RESULTS BY YEAR

View Search Details

14 results Page 1 of 2

PMc Full-Text Search Results

Embargoed Articles:  Include  Exclude

Unknown field was ignored: [Organism:noexp]

- Cross-species EST alignments reveal novel and conserved alternative splicing events in legumes.  
Wang BB, O'Toole M, Brendel V, Young ND.  
BMC Plant Biol. 2006 Feb 19;8:17. doi: 10.1186/1471-2229-8-17.  
PMCID: PMC2277414

---

**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/>