中国科学院北京基因组研究所（国家生物信息中心）
BEIJING INSTITUTE OF GENOMICS CHINESE ACADEMY OF SCIENCES / CHINA NATIONAL CENTER FOR BIOINFORMATION

基因组学 - Genomics

# 第6章 基因组注释

章张 博士

中国科学院大学
University of Chinese Academy of Sciences

国家基因组科学数据中心
National Genomics Data Center

# 内容提纲

- 基因组注释的定义
- 基因组注释的复杂性：原核生物 vs 真核生物
- 基因组结构注释
  - 原核生物
  - 真核生物
- 基因组功能注释
  - 序列比对、蛋白结构域
  - 基因本体论和代谢途径
  - 基因功能富集
- 基因组注释实例

# 基因组注释的定义

# 基因组测序完成 → 基因组注释开始

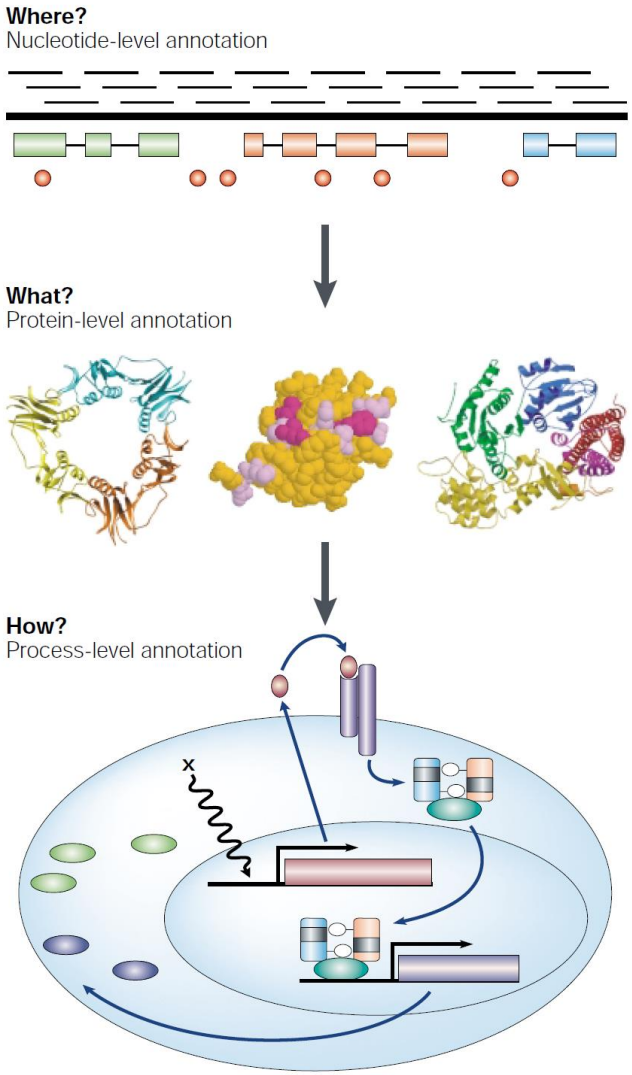
ACACTCGCTTCTGGAACGTCTGAGGTTATCAATAAGCTCCTAGTCCAGACGCCATGGGTCATTTCACAGAGGAGGACAAGGCT
ACTATCACAAGCCTGTGGGGCAAGGTGAATGTGGAAGATGCTGGAGGAGAAACCCTGGGAAGGCTCCTGGTTGTCTACCCATG
GACCCAGAGGTTCTTTGACAGCTTTGGCAACCTGTCCTCTGCCTCTGCCATCATGGGCAACCCCAAAGTCAAGGCACATGGCA
AGAAGGTGCTGACTTCCTTGGGAGATGCCACAAAGCACCTGGATGATCTCAAGGGCACCTTTGCCCAGCTGAGTGAACTGCAC
TGTGACAAGCTGCATGTGGATCCTGAGAACTTCAAGCTCCTGGGAAATGTGCTGGTGACCGTTTTGGCAATCCATTTCGGCAA
AGAATTCACCCCTGAGGTGCAGGCTTCCTGGCAGAAGATGGTGACTGCAGTGGCCAGTGCCCTGTCCTCCAGATACCACTGAG
CTCACTGCCCATGATTCAGAGCTTTCAAGGATAGGCTTTATTCTGCAAGCAATACAAATAATAAATCTATTCTGCTGAGAGAT
CACACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATCTGACTCCTGAGGAGAAGTCT
GCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTG
GACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCA
AGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCAC
TGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAA
AGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCACTAAG
CTCGCTTTCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTAAACTGGGGGATATTATGAAGGG
CCTTGAGCATCTGGATTCTGCCTAATAAAAAACATTTATTTTCATTGCACACTCGCTTCTGGAACGTCTGAGGTTATCAATAA
GCTCCTAGTCCAGACGCCATGGGTCATTTCACAGAGGAGGACAAGGCTACTATCACAAGCCTGTGGGGCAAGGTGAATGTGGA
AGATGCTGGAGGAGAAACCCTGGGAAGGCTCCTGGTTGTCTACCCATGGACCCAGAGGTTCTTTGACAGCTTTGGCAACCTGT
CCTCTGCCTCTGCCATCATGGGCAACCCCAAAGTCAAGGCACATGGCAAGAAGGTGCTGACTTCCTTGGGAGATGCCACAAAG
CACCTGGATGATCTCAAGGGCACCTTTGCCCAGCTGAGTGAACTGCACTGTGACAAGCTGCATGTGGATCCTGAGAACTTCAA
GCTCCTGGGAAATGTGCTGGTGACCGTTTTGGCAATCCATTTCGGCAAAGAATTCACCCCTGAGGTGCAGGCTTCCTGGCAGA
AGATGGTGACTGCAGTGGCCAGTGCCCTGTCCTCCAGATACCACTGAGCTCACTGCCCATGATTCAGAGCTTTCAAGGATAGG
CTTTATTCTGCAAGCAATACAAATAATAAATCTATTCTGCTGAGAGATCACACATTTGCTTCTGACACAACTGTGTTCACTAG
CAACCTCAAACAGACACCATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGA
TGAAGTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGT

# 基因组注释（Genome annotation）

- 定义：利用生物信息学方法和工具，对基因组所有基因的生物学功能进行高通量注释，是当前功能基因组学研究的一个热点。
  - 结构注释（Structural annotation）：基因位置及其结构等
  - 功能注释（Functional annotation）：基因功能及其调控等
- 目的：识别基因组序列中存在的基因和其他多种功能元件（包括编码基因、非编码RNA、转座子等重复序列、调控元件等），并推测其生物学功能（例：ENCODE）。
- 意义：基因组注释是生物学研究的基础，一个基因组的价值取决于该基因组注释的质量，基因组注释建立了从未知功能的基因组序列到该物种生物学研究的桥梁。

**Genome annotation: from sequence to biology.** *Nature Reviews Genetics* 2001

# 基因组注释计划：ENCODE




## ENCODE Encyclopedia Version 2: Genomic and Transcriptomic Annotations


The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

**The ENCODE has two levels of annotations:**

- **Integrative-level annotations** integrate multiple types of experimental data and ground level annotations.
- **Ground-level annotations** are derived directly from the experimental data, typically produced by uniform processing pipelines.

**Genome annotation: from sequence to biology.** *Nature Reviews Genetics* 2001

# Genome Annotation

NIH National Library of Medicine

Home Write to the Help Desk

NLM Support Center > Knowledge Base > KA-03574

## What is genome annotation?

Print

Views: 3746

**Genome annotation** is the process of finding and designating locations of individual genes and other features on raw DNA sequences, called **assemblies**. Annotation gives meaning to a given sequence and makes it much easier for researchers to view and analyze its contents. To visualize what annotation adds to our understanding of the sequence, you can compare the raw sequence (in FASTA format) with the GenBank or Graphics formats, both of which contains annotations. In both instances note the placement of individual genes and other features on the sequence.

When a group of researchers assemble a genome, they may also — with processes they establish themselves — annotate it at the same time. In the past, an assembly with annotation was known as a **build**. These days, the term build is rarely used, as the genome assembly process and its annotation process are often completely uncoupled. They can be conducted at different times by different parties. For example, the Genome Reference Consortium (GRC) is maintaining and updating the human reference *assembly*. GRC releases assembly (sequence) updates and deposits these to the International Nucleotide Sequence Database Collaboration (INSDC) without annotation. GRC prepared the latest major assembly update (major release designated as GRCh38) in December 2013 and it has since followed with several minor updates (patches). In further processing of an assembly update, the NCBI staff creates a RefSeq version of the submitted INSDC assembly. Following that, NCBI annotates the RefSeq version of the assembly. Each annotation release has its own designation and time stamp. For example, the latest (as of May 2021) NCBI annotation release is designated as Release 109.20210514.

In addition of the human reference genome, NCBI staff annotate numerous eukaryotic genomes via the powerful Eukaryotic Genome Annotation Pipeline. Visit the Eukaryotic Genome Annotation at NCBI page to start exploring extensive documentation on the annotation process, and to follow the progress of individual genome annotation.

NCBI staff have also developed the Prokaryotic Genome Annotation Pipeline that is available as a service to GenBank submitters and also as a stand-alone software package.

https://support.nlm.nih.gov/knowledgebase/article/KA-03574/

# 基因组注释

## 识别基因组中编码基因、非编码基因及调控区域

# 基因组注释：结构注释和功能注释

**Like parenthood, genome annotation responsibilities do not end with birth. — 基因组注释不能“管生不管养”**

## Genome Annotation

### Structural Annotation

ORFs, genes

Regulatory elements

Repeat elements

### Functional Annotation

Biological function

Biochemical function

Regulatory function

**A beginner's guide to eukaryotic genome annotation**. *Nature Reviews Genetics* 2012

# Curation: value-added annotation

- Curation (审编) involves the **annotation, translation, presentation and publication** of the data into a database that enables integration of the **scientific literature as well as large data sets**.
- Primary goals of curation:
  - accurate and comprehensive representation of biological knowledge;
  - easy access to this data for working scientists and a basis for computational analysis.

**Bioinformatics: Curation generation**. *Nature* 2011
http://www.biocurator.org

Genomics, Proteomics & Bioinformatics
Volume 12, Issue 4, August 2014, Pages 153-155

Perspective

## Bringing Biocuration to China

Zhang Zhang [1], Weimin Zhu [2 3], Jingchu Luo [4]

Show more

ISB International Society for Biocuration

+ Add to Mendeley  Share  Cite

https://doi.org/10.1016/j.gpb.2014.07.001  Get rights and content

Under a Creative Commons license  open access

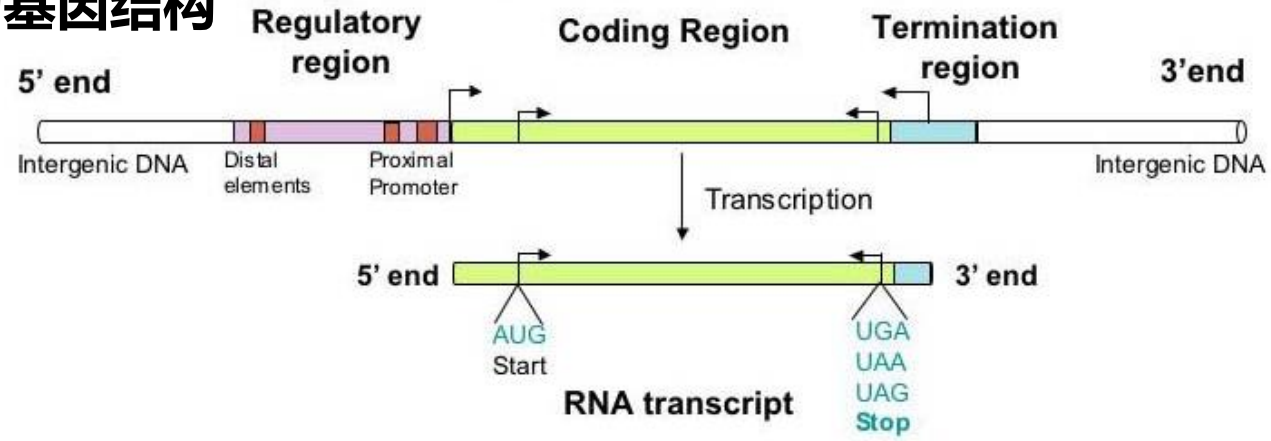
### Abstract

Biocuration involves adding value to biomedical data by the processes of standardization, quality control and information transferring (also known as data annotation). It enhances data interoperability and consistency, and is critical in translating biomedical data into scientific discovery. Although China is becoming a leading scientific data producer, biocuration is still very new to the Chinese biomedical data community. In fact, there currently lacks an equivalent acknowledged word in Chinese for the word "curation". Here we propose its Chinese translation as "审编" (Pinyin: *shěn biān*) based on its implied meanings taken by biomedical data community. The 8th International Biocuration Conference to be held in China (http://biocuration2015.tilsi.org) next year bears the potential to raise the general awareness in China of the significant role of biocuration in scientific discovery. However, challenges are ahead in its implementation.

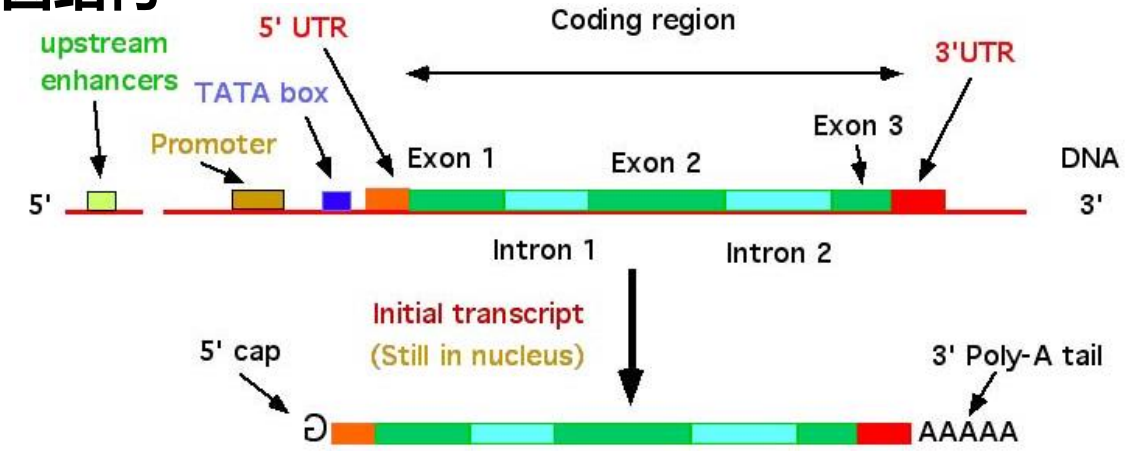
# 基因组注释的复杂性
# 原核生物 vs 真核生物

# 基因结构

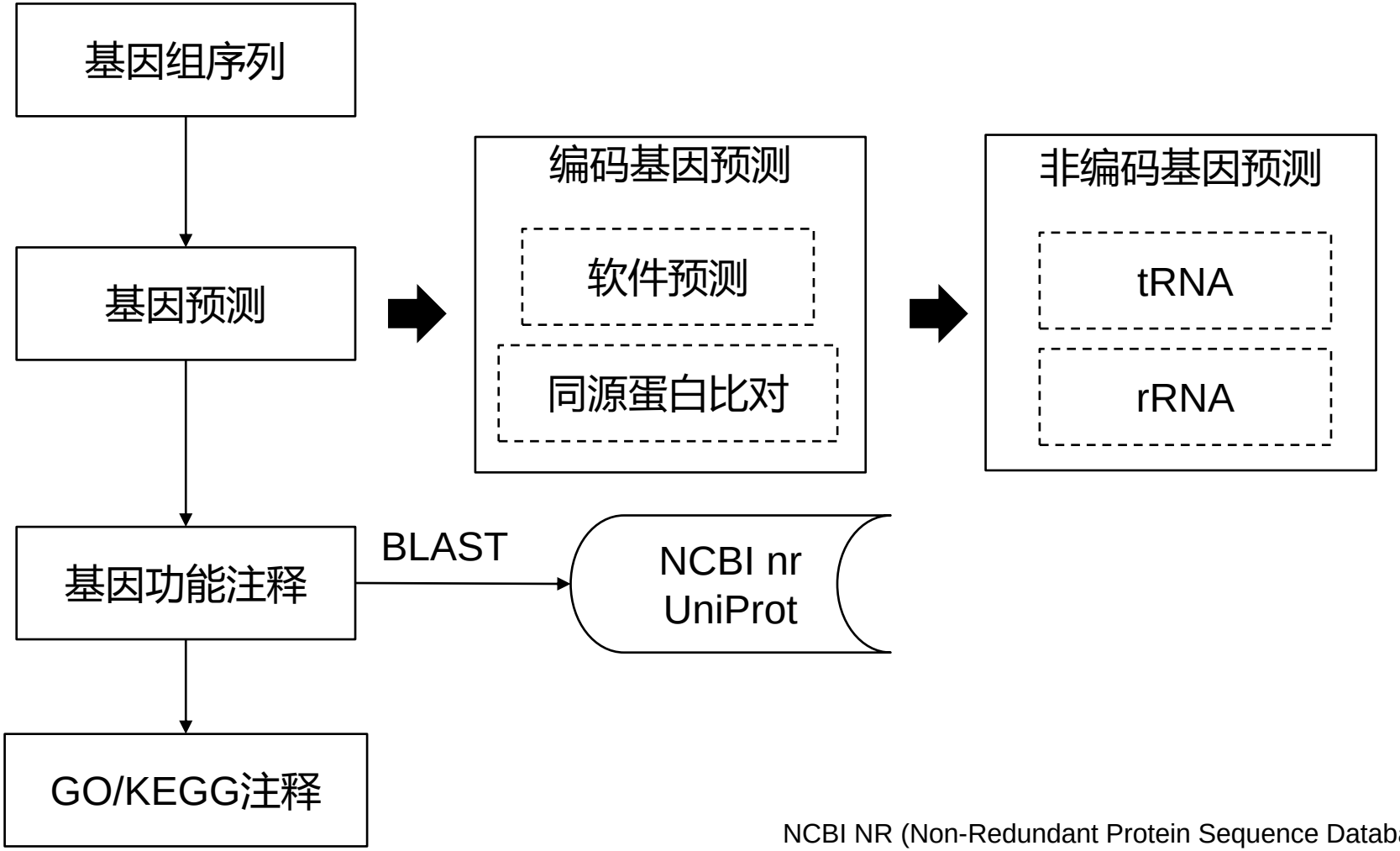
## 原核生物基因结构




## 真核生物基因结构

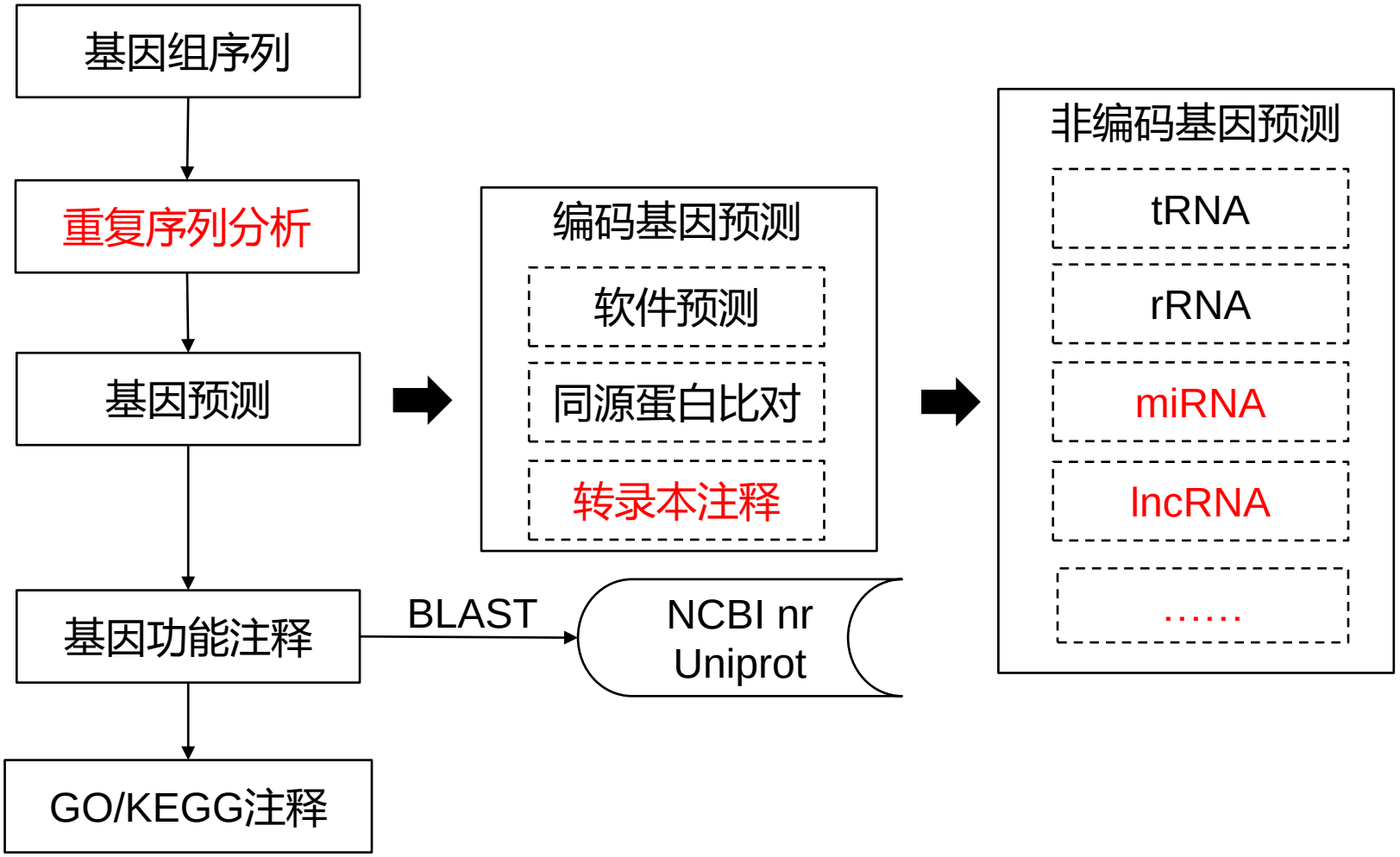


# 原核生物基因组注释流程




NCBI NR (Non-Redundant Protein Sequence Database)

# 真核生物基因组注释流程



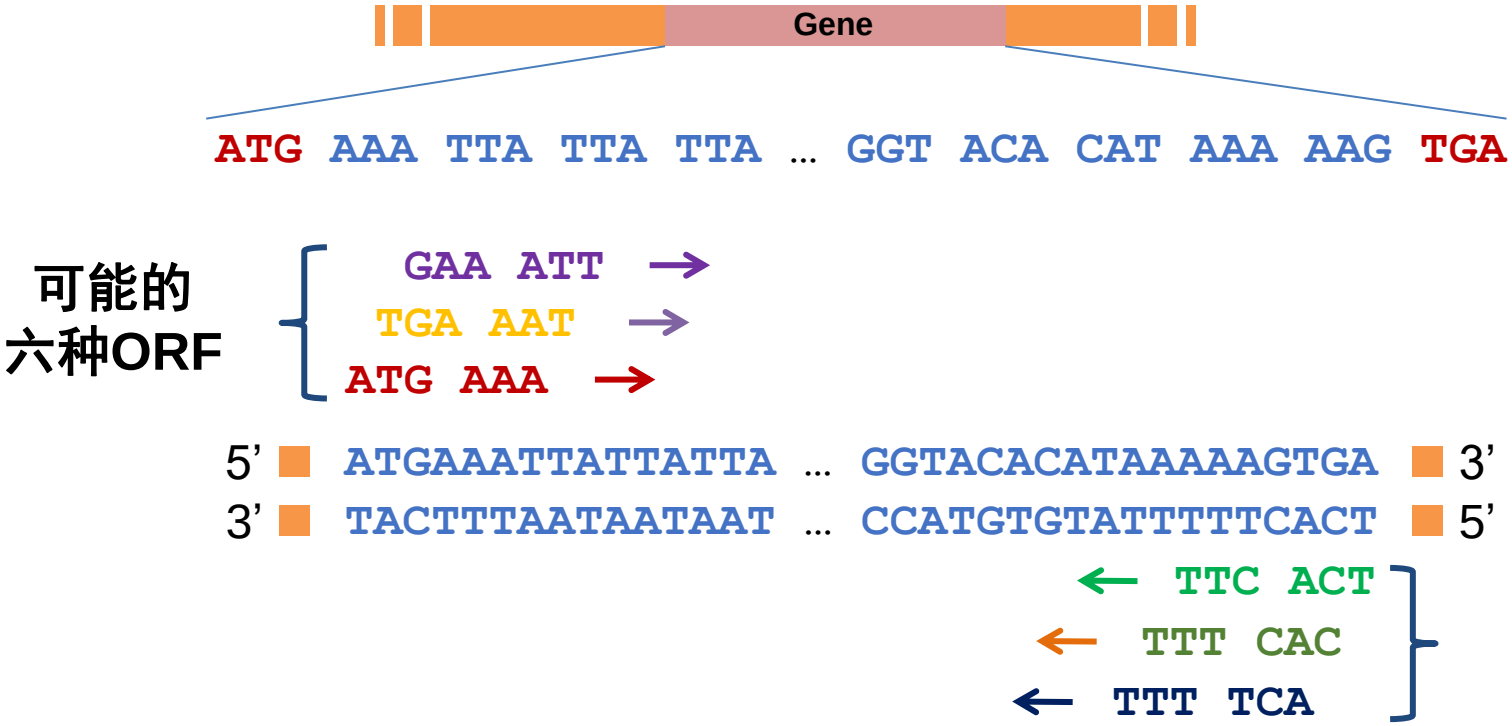
# 原核生物基因组结构注释

# 基因预测的方法

- 湿性实验手段：通过实验分析，看其是否能表达基因产物；通量低、成本高
- 干性实验预测：通过计算机对DNA序列进行特征搜寻，分析寻找基因；通量高、成本低（生物信息学）
  - 基于基因结构特征搜寻
  - 基于同源基因搜索

# 方法一：基于基因结构特征搜寻

基因的核苷酸序列并非随机排列，而是具有明显特征，可基于开放读码框（Open reading frame, ORF）预测基因



# 方法二：基于同源基因搜索

通过将数据库中的基因序列与待查的基因组序列进行比较，从中查找可与之匹配的碱基序列及其比例，用于界定基因的方法称为**同源搜索**。

同源基因有以下几种情况：

- DNA序列某些片段完全相同
- ORF排列类似
- ORF翻译成的氨基酸序列相同
- 模拟多肽高级结构相似

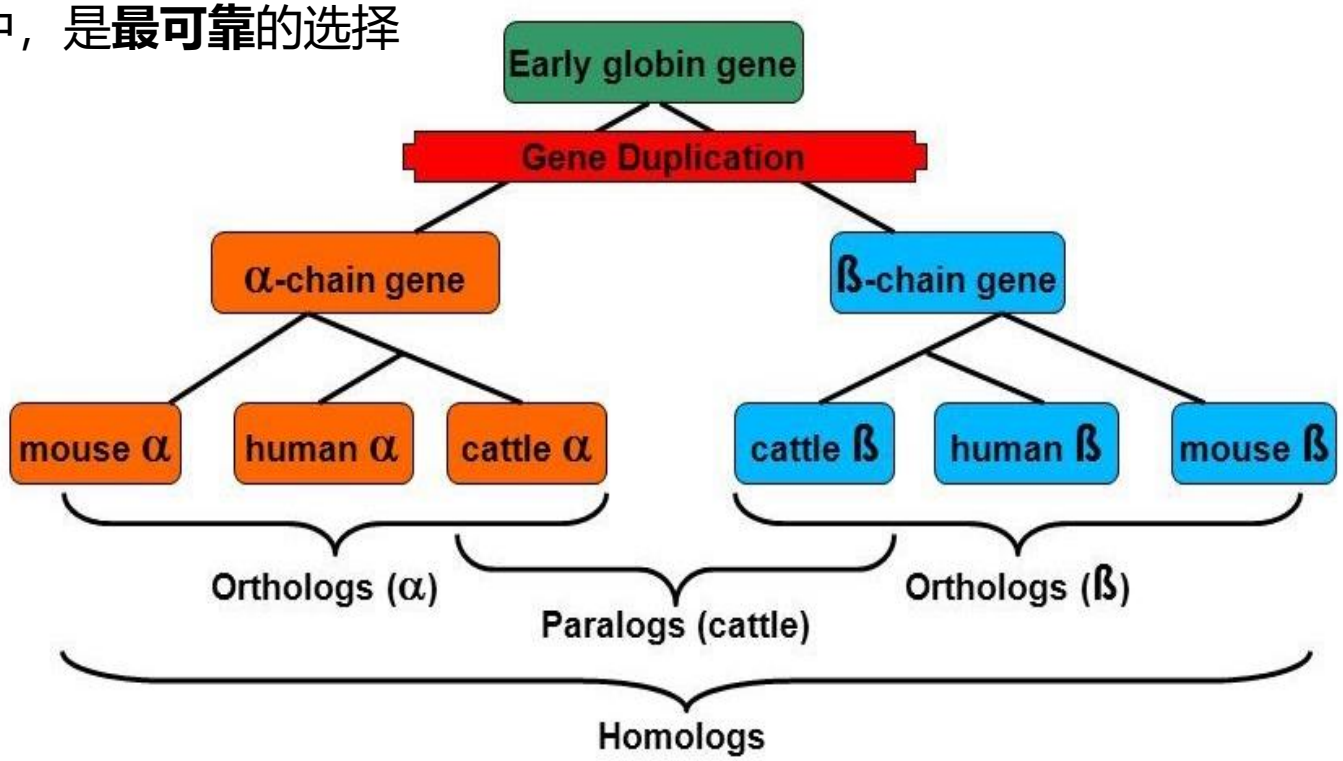
# 同源基因的类型

- 直系同源基因（Orthologous gene）
  指从**同一祖先**垂直进化而来的基因

  大多数直系同源基因**功能相同或相近**，调控途径也相似， 因此在基因组序列的注释中，是**最可靠**的选择

- 旁系同源基因（Paralogous gene）
  指由于**基因复制**而产生的同源基因

  旁系同源基因随着时间的推移，它们在**序列组成和功能上**可能会变得**不同**



# 同源性、相似性和一致性

- **同源性**（**Homology**）：进化过程中源于同一祖先的不同分支，用来描述物种之间的进化关系，所以在同源性的表达中只能用“有”或者“无”，属于**定性描述**。
- **相似性**（**Similarity**）：序列之间**相似**位点占整个序列的比例，属于**定量描述**。
- **一致性**（**Identity**）：序列之间**相同**位点占整个序列的比例，相对精确度更高的一个描述，属于**定量描述**。

# 相似性 vs 一致性

The percentage of identical residues (**identity**) or the percentage of residues conserved with **similar physicochemical properties** (**similarity**), is usually used to "quantify the **homology**."

**hemoglobin subunit beta-like [Canis lupus familiaris]**

**Sequence ID:** NP_001257812.1 **Length:** 147 **Number of Matches:** 1

See 1 more title(s)

**Range 1: 1 to 147** GenPept Graphics ▼ Next Match ▲ Previous Match

| Score | Expect | Method | Identities | Positives | Gaps |
|---|---|---|---|---|---|
| 275 bits(702) | 2e-96 | Compositional matrix adjust | 132/147(90%) | 138/147(93%) | 0/147(0%) |

```
Query   1     MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK   60
              MVHLT EEKS V+ LWGKVNVDEVGGEALGRLL+VYPWTQRFF+SFGDLSTPDAVM N K
Sbjct   1     MVHLTAEEKSLVSGLWGKVNVDEVGGEALGRLLIVYPWTQRFFDSFGDLSTPDAVMSNAK   60

Query   61    VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG   120
              VKAHGKKVL +FSDGL +LDNLKGTFA LSELHCDKLHVDPENF+LLGNVLVCVLAHHFG
Sbjct   61    VKAHGKKVLNSFSDGLKNLDNLKGTFAKLSELHCDKLHVDPENFKLLGNVLVCVLAHHFG   120

Query   121   KEFTPPVQAAYQKVVAGVANALAHKYH   147
              KEFTP VQAAYQKVVAGVANALAHKYH
Sbjct   121   KEFTPQVQAAYQKVVAGVANALAHKYH   147
```

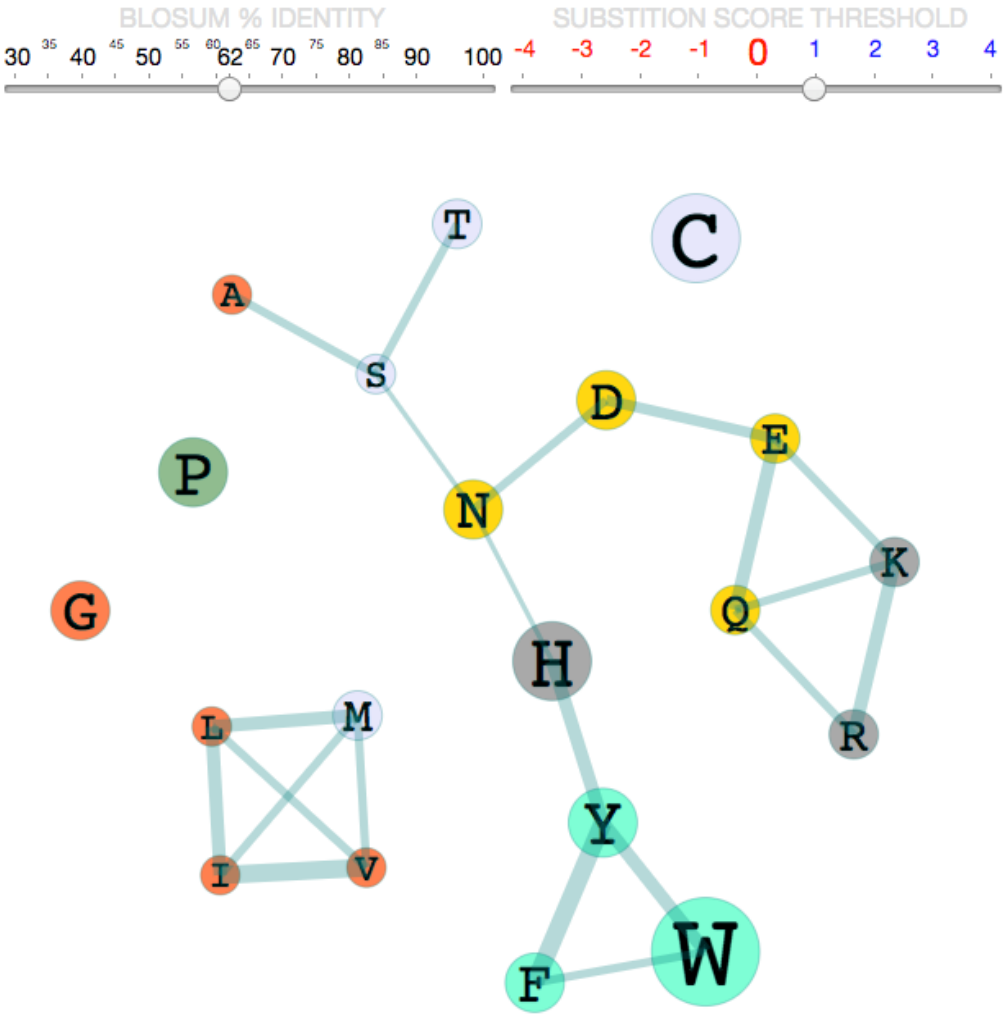
**In BLAST, similarity refers to a positive matrix score.**

# 同源性、一致性和相似性：三者的关系

- “两条序列的同源性”只能描述为“有”或者“无”，**同源性不能量化。**
- 而相似性和一致性可以看做是从不同的角度对同源性的量化指标，一般地，**相似性 > 一致性。**
- 一般来说，序列之间的**相似性和一致性越高**，序列之间**同源的可能性越大。**
- 虽然同源基因在序列上是相似的，但**相似的序列不一定是同源。**

下面这段描述是三者正确的描述：
人类NP_000509.1序列（hemoglobin subunit beta，147个氨基酸）与狗NP_001257812.1序列（hemoglobin subunit beta-like，147个氨基酸）同源，相似性高达93%，一致性达90%。

# Amino Acid Similarity Network




Node Color

- Chemical1
- Chemical2
- Polarity
- Charge
- Hydrophobicity
- Acid or Base
- Codon count
- Essentiality

Node Size

- Matrix score
- Hydropathy
- pI
- pK1
- pK2
- van der Waals
- Occurrence
- Mass(Da)

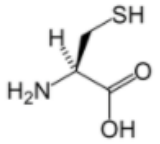
| | | |
|---|---|---|
| Acidic | Aliphatic | Aromatic |
| Basic | Cyclic | Hydroxyl/Sulfur |

## Cysteine

Cys , C

103.1388 Da | 1.9%

UGU, UGC




The sulfur atom bonds readily to heavy metal ions. Under oxidizing conditions, two cysteines can join together in a disulfide bond to form the amino acid cystine. When cystines are part of a protein, insulin for example, the tertiary structure is stabilized, which makes the protein more resistant to denaturation; therefore, disulfide bonds are common in proteins that have to function in harsh environments including digestive enzymes (e.g., pepsin and chymotrypsin) and structural proteins (e.g., keratin). Disulfides are also found in peptides too small to hold a stable shape on their own (e.g. insulin).

http://ahmetrasit.com/blosum/

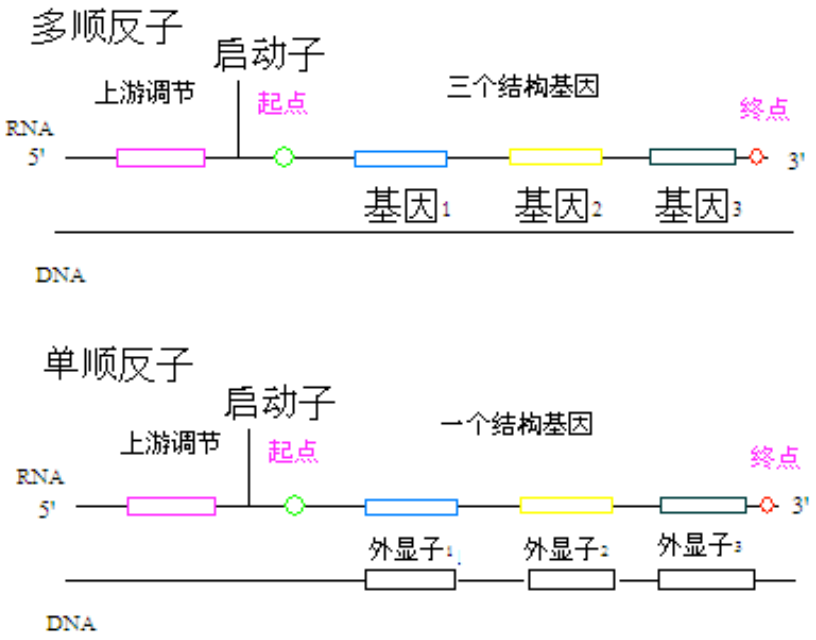
# 原核基因组结构注释的主要问题

原核基因组中80%-90%的序列参与编码基因

- 如果有两个或多个contained的ORF，哪一个是基因（假定只可能有一个）
- 在无已知同源性信息的情况下如何寻找基因

# 原核生物编码基因预测

- 原核生物基因的各种信号位点（如启动子和终止子信号位点）特异性较强且容易识别，因此相应的基因预测方法已经基本成熟。Prodigal和Glimmer是应用最为广泛的原核生物基因结构预测软件，准确度高。
- 原理：通过扫描基因组，找到从ATG（少数GTG、TTG)开始的位置，直到TGA、TAG、TAA。
- **多顺反子**：在原核细胞中，通常是几种不同的mRNA连在一起，**位于同一转录单位**内，相互之间由一段短的不编码蛋白质的间隔序列所隔开，**享有同一对起点和终点**。



# 原核生物编码基因预测软件

- Glimmer：采用内插**马尔科夫模型**（IMM）来识别编码区域和非编码区域。一般使用三种方法创建训练模型（Delcher et al. 2007. PMID: 17237039）：
  - 用亲缘关系很近的物种基因
  - 用自身序列创建的ORF数据
  - 用基因组本身的已知信息（常采用自身数据作为训练数据）
- Prodigal：原核的**动态编程**基因查找算法，针对原核生物的基因预测工具，尤其是高GC的基因组，同时也适用于宏基因组（Hyatt et al. 2010. PMID: 20211023）
- GeneMark：原理是使用**统计学模型**的从头预测(ab initio)方法，不依赖任何先验知识和经验参数，通过描述DNA序列中核苷酸的离散模型，利用编码区和非编码区的核苷酸分布概率不同来进行基因预测（Besemer, et al. 2001. PMID: 11410670）

# 非编码RNA序列的预测

非编码RNA（non-coding RNA，ncRNA），指的是不被翻译成蛋白质的RNA，如tRNA、rRNA等，这些RNA不被翻译成蛋白质，但是具有重要的生物学功能。

- miRNA（小RNA，microRNA）：通过与目标信使RNA（mRNA）结合，进而抑制转录后的基因表达，在调控基因表达、细胞周期、生物体发育时序等方面起重要作用。
- tRNA（转运RNA，transfer RNA）：携带氨基酸进入核糖体，使之在mRNA指导下合成蛋白质。
- rRNA（核糖体RNA，ribosomal RNA）：与蛋白质结合形成核糖体，其功能是作为mRNA的支架，提供mRNA翻译成蛋白质的场所。
- snRNA（小核RNA，small nuclear RNA）：主要参与mRNA前体的加工过程，使之成为成熟的mRNA，是RNA剪切体的主要成分。

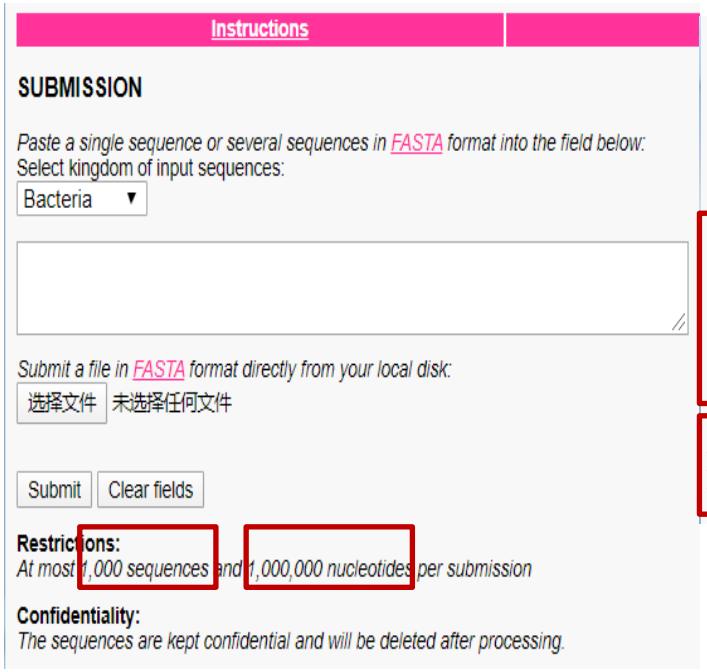
# 基因组rRNA注释

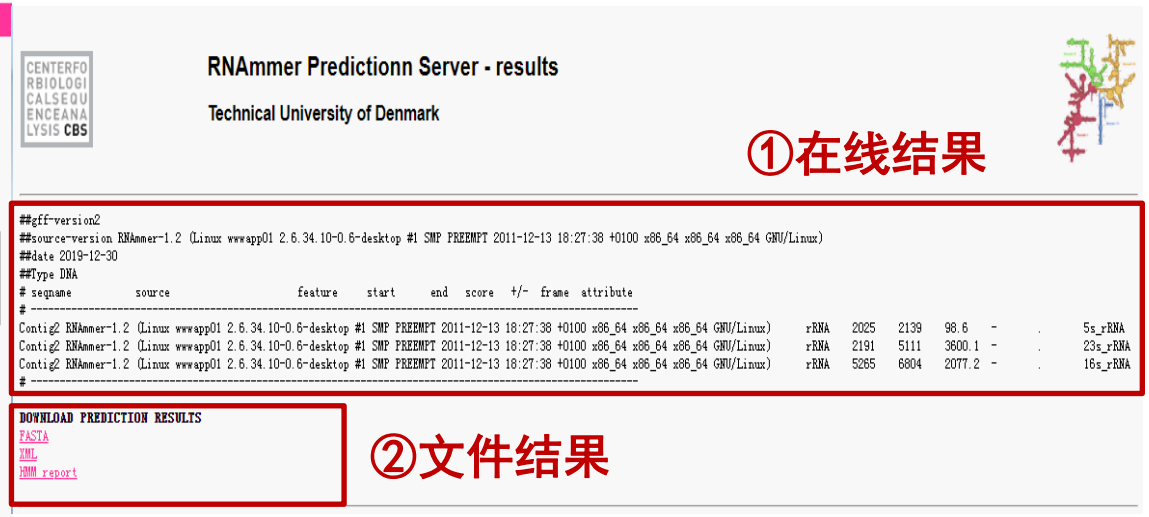
- rRNA预测
  - RNAmmer：不基于参考序列的从头预测
  - 在原核生物中，包含以下3种类型的rRNA
    - 5S rRNA
    - 16S rRNA
    - 23S rRNA

Web server: http://www.cbs.dtu.dk/services/RNAmmer/

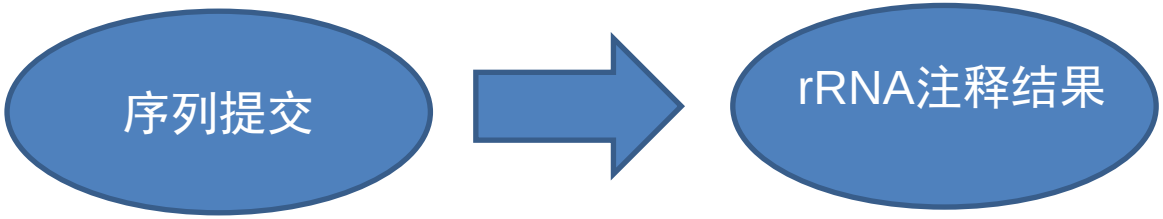
Lagesen K, et al. 2007. PMID: 17452365

# RNAmmer的使用




RNAmmer在线注释



# tRNA预测注释

tRNAscan-SE：综合多个识别和分析程序，通过分析启动子元件的保守序列模式、tRNA二级结构的分析、转录控制元件分析和除去绝大多数假阳性的筛选过程，据称能识别99%的真tRNA基因。

Web server: http://lowelab.ucsc.edu/tRNAscan-SE/

Chan PP, et al. 2019. PMID: 31020551

# tRNAscan-SE在线注释

Search options

| | |
|---|---|
| Sequence source | Bacterial |
| Search mode | Default |
| Query sequence | ◉ Formatted (FASTA) ○ Raw Sequence |
| | Sequence name (optional): (no spaces) |
| | (Queries are limited to a total of less than 5 million nucleotides at any one time) |
| | or submit a file: |
| | 选择文件 Filter5M.fasta |
| | Clear Sequence |
| Output | Output BED format |

序列提交

Run tRNAscan-SE | Reset Form

Extended options

- Disable pseudo gene checking
- Show origin of first-pass hits
- Show primary and secondary structure components to scores

| | |
|---|---|
| Genetic Code for tRNA Isotype Prediction: | Universal |
| Score cutoff: | Default cut-off value should only be changed for execptional conditions |

**Results**

Download as text

| Sequence Name | tRNA # | Predicted tRNA Structure | Similar tRNAs in GtRNAdb | tRNA Begin | tRNA End | tRNA Type | Anticodon | Intron Begin | Intron End | Infernal Score | Isotype Model | Isotype Score | Note |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MySeq1 | 1 | View | View | 13 | 85 | Thr | TGT | 0 | 0 | 78.0 | Thr | 93.1 | None |
| MySeq2 | 1 | View | View | 6 | 79 | Arg | TCT | 0 | 0 | 75.1 | Arg | 89.3 | None |
| MySeq3 | 1 | View | View | 14 | 114 | Ser | CGA | 51 | 69 | 71.8 | Ser | 118.3 | None |
| MySeq4 | 1 | View | View | 6 | 88 | Leu | AAG | 0 | 0 | 65.0 | Leu | 92.2 | None |
| MySeq5 | 1 | View | View | 3 | 89 | SeC | TCA | 0 | 0 | 146.9 | SeC | 146.9 | None |
| MySeq6 | 1 | View | View | 7 | 92 | Lys | CTT | 0 | 0 | 72.1 | Leu | 75.7 | IPD:-73.90 |

**Isotype-Specific Model Scores:**

Download as text

| tRNAscan ID | Anticodon predicted isotype | Isotype Prediction (Anticodon v. Isotype Model) | Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | SeC | Ser | Thr | Trp | Tyr | Val | iMet |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MySeq1.trna1 | Thr | Consistent | 62.6 | 63.2 | 48.1 | 34.4 | 77.1 | 62.4 | 21.9 | 36.3 | 60.9 | 66.0 | 28.5 | 46.4 | 65.2 | 39.8 | 47.1 | No Hit | 32.5 | 93.1 | 59.9 | 58.6 | 61.3 | -7.1 |
| MySeq2.trna1 | Arg | Consistent | 42.5 | 89.3 | 67.8 | 29.2 | 48.4 | 45.1 | 30.0 | 38.1 | 47.4 | 54.0 | 11.1 | 51.0 | 62.4 | 24.8 | 22.6 | No Hit | 26.2 | 65.8 | 61.6 | 44.4 | 49.2 | 25.7 |
| MySeq3.trna1 | Ser | Consistent | 24.2 | 45.4 | 8.9 | 0.1 | 58.6 | 43.5 | 8.7 | 31.9 | 47.1 | 43.5 | 74.6 | 7.8 | 0.3 | -2.7 | -11.5 | No Hit | 118.3 | 45.3 | 28.2 | 51.0 | 19.0 | No Hit |
| MySeq4.trna1 | Leu | Consistent | 40.8 | 32.4 | -12.6 | -5.6 | 34.6 | 23.7 | No Hit | 26.8 | 50.6 | 30.5 | 92.2 | -11.2 | 3.2 | -11.9 | 12.8 | No Hit | 67.7 | 44.5 | 37.5 | 29.8 | 17.6 | No Hit |
| MySeq5.trna1 | SeC | Consistent | No Hit | No Hit | 5.4 | -2.1 | -3.4 | -10.0 | No Hit | -2.0 | -2.5 | -5.8 | -5.7 | -2.0 | No Hit | No Hit | -12.7 | 146.9 | -14.6 | -0.5 | No Hit | -4.4 | No Hit | No Hit |
| MySeq6.trna1 | Lys | Inconsistent | 15.0 | 50.3 | 21.2 | 6.0 | 40.4 | 34.1 | -9.2 | 23.3 | 69.9 | 55.8 | 75.7 | 1.8 | 14.6 | -2.1 | 2.5 | No Hit | 72.4 | 51.9 | 36.6 | 54.8 | 19.2 | No Hit |

Top score | 2nd highest score | 3rd highest score

**Run Statistics:**

| Search options | |
|---|---|
| Search Mode | Eukaryotic |
| Searching with | Infernal First Pass->Infernal |
| Isotype-specific model scan | Yes |
| Covariance model | models/TRNAinf-euk.cm |
| | models/TRNAinf-euk-SeC.cm |
| Infernal first pass cutoff score | 10 |

tRNA注释结果

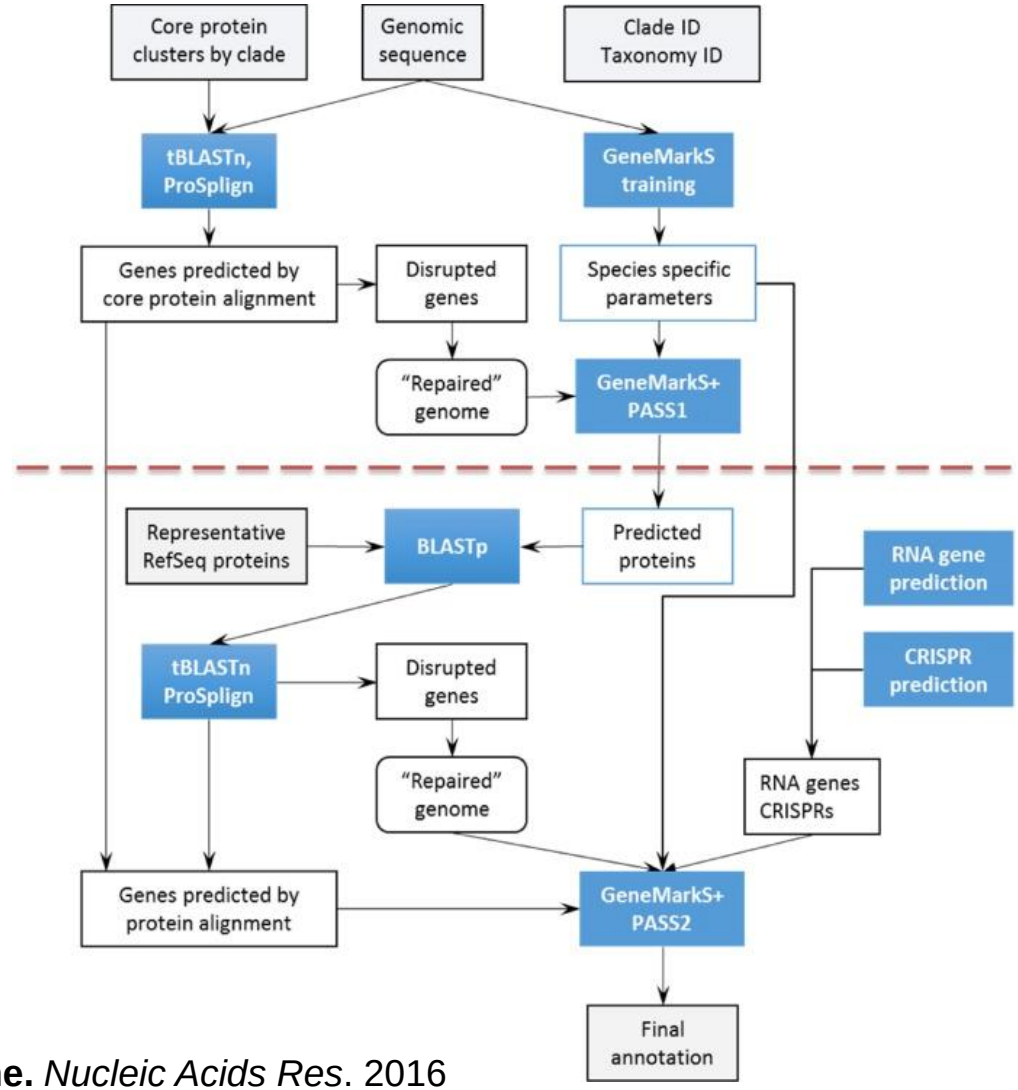
# ncRNA预测

- Infernal：建立了3,016多个RNA家族，并对每个家族建立了一致性二级结构和协方差模型，形成了**Rfam数据库**。
- 利用Rfam家族的协方差模型，采用Rfam自带的Infernal软件预测miRNA、snRNA序列等ncRNA。
- Rfam/Infernal方法应用广泛，可以预测各种RNA家族成员，但是特异性较差。如果有更好的专门预测某一类非编码RNA的软件，建议采用特异性RNA软件进行预测。

Kalvari et al. 2018. PMID: 29927072

# 原核基因组注释流程—PGAP

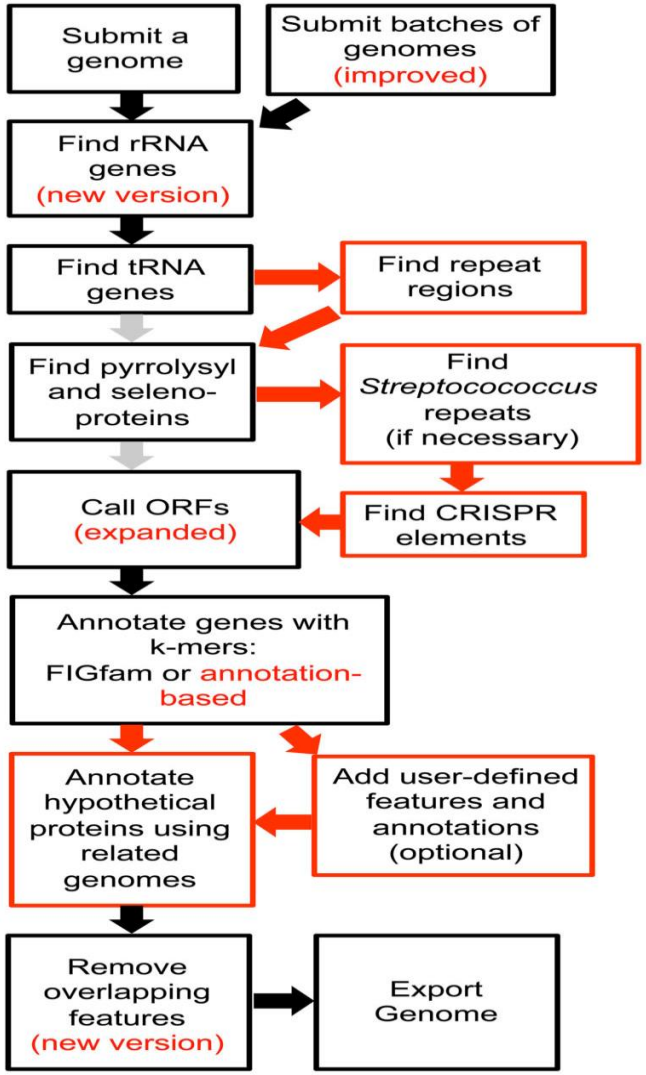
PGAP（Prokaryotic Genome Annotation Pipeline）：通过挖掘现有的蛋白质信息而进行调节，以建立新的核心蛋白簇，并根据来自于所提交细菌基因组的不断增加的大量可用证据，反复改善其注释功能。可产生高品质的注释，旨在满足NCBI对于细菌基因组序列提交的INSDC标准，并遵守UniProt命名原则。




**NCBI prokaryotic genome annotation pipeline.** *Nucleic Acids Res*. 2016

# 原核基因组注释流程—RAST

RAST（Rapid Annotation using Subsystem Technology）：针对完整的或将近完整（draft genome or complete genome）的细菌和古菌基因组注释工具。可用来预测ORF、rRNA、tRNA以及相应的功能基因，并可以利用这些信息构建代谢网络。




Brettin T, et al. 2015. PMID: 25666585

# 原核基因组注释流程—Prokka

Prokka：原核基因组注释的分析流程，包括基因鉴定、功能注释和基因组配套注释文件生成

Feature prediction tools used by Prokka

| Tool (reference) | Features predicted |
|---|---|
| Prodigal ( Hyatt 2010 ) | Coding sequence (CDS) |
| RNAmmer ( Lagesen *et al.* , 2007 ) | Ribosomal RNA genes (rRNA) |
| Aragorn ( Laslett and Canback, 2004 ) | Transfer RNA genes |
| SignalP ( Petersen *et al.* , 2011 ) | Signal leader peptides |
| Infernal ( Kolbe and Eddy, 2011 ) | Non-coding RNA |

**Prokka: rapid prokaryotic genome annotation.**
*Bioinformatics* 2014

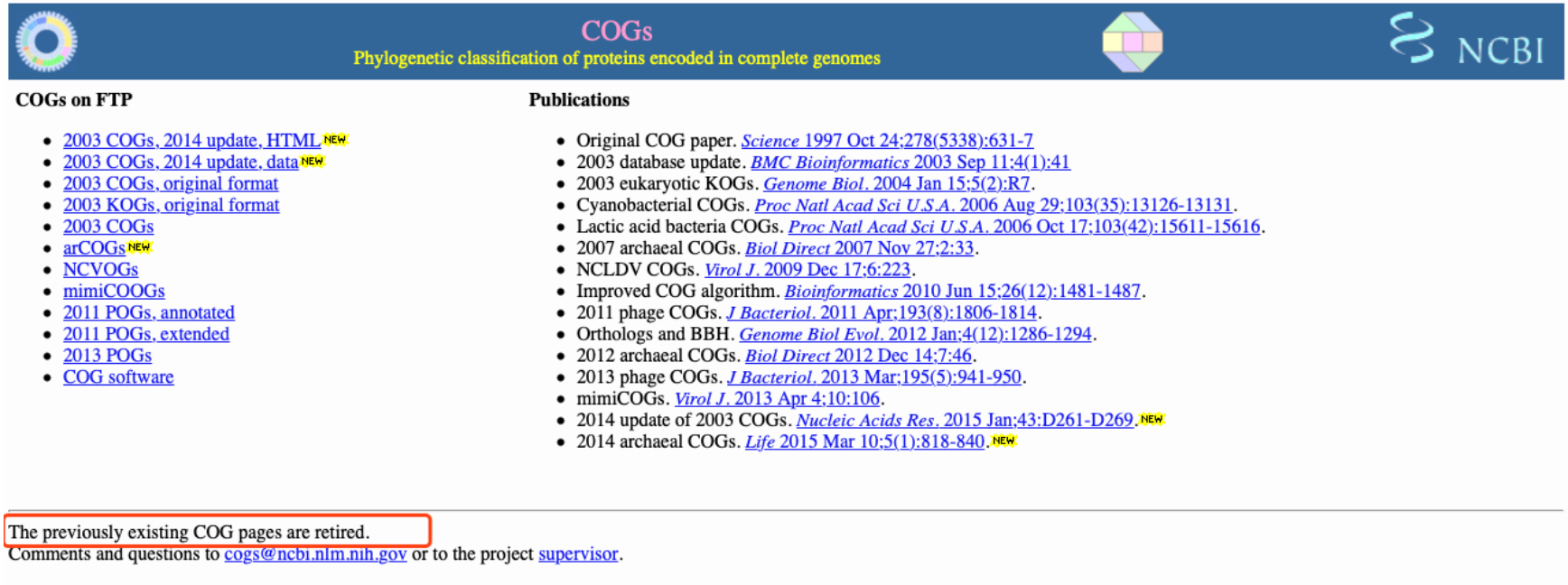
Comparison of annotation of *E.coli K-12* accession U00096.2

| Feature | Reference | Prokka | RAST | xBase2 |
|---|---|---|---|---|
| Total CDS | 4321 | **4305** | 4512 | 4444 |
| *Matching start* | – | **3828** | 3571 | 3025 |
| *Different start* | – | **318** | 533 | 1052 |
| *Missing CDS* | – | **172** | 214 | 241 |
| *Extra CDS* | – | **159** | 405 | 367 |
| *Hypothetical protein* | 18 | 276 | 638 | **156** |
| *With EC number* | 1114 | 1050 | **1118** | 0 |
| Total tRNA | 89 | **88** | 86 | **88** |
| Total rRNA | 22 | **22** | **22** | **22** |

# COG注释

**COG（Clusters of Orthologous Groups of proteins，同源蛋白簇）**

COG分为两类，一类是原核生物的（COG），另一类是真核生物（KOG），是由NCBI创建并维护的蛋白数据库，COG数据库按照功能一共可以分为二十六类。COG曾是原核生物功能注释的常用数据库，目前已停止更新。



COGs

Phylogenetic classification of proteins encoded in complete genomes

NCBI

**COGs on FTP**

- 2003 COGs, 2014 update, HTML NEW
- 2003 COGs, 2014 update, data NEW
- 2003 COGs, original format
- 2003 KOGs, original format
- 2003 COGs
- arCOGs NEW
- NCVOGs
- mimiCOOGs
- 2011 POGs, annotated
- 2011 POGs, extended
- 2013 POGs
- COG software

**Publications**

- Original COG paper. *Science* 1997 Oct 24;278(5338):631-7
- 2003 database update. *BMC Bioinformatics* 2003 Sep 11;4(1):41
- 2003 eukaryotic KOGs. *Genome Biol.* 2004 Jan 15;5(2):R7.
- Cyanobacterial COGs. *Proc Natl Acad Sci U.S.A.* 2006 Aug 29;103(35):13126-13131.
- Lactic acid bacteria COGs. *Proc Natl Acad Sci U.S.A.* 2006 Oct 17;103(42):15611-15616.
- 2007 archaeal COGs. *Biol Direct* 2007 Nov 27;2:33.
- NCLDV COGs. *Virol J.* 2009 Dec 17;6:223.
- Improved COG algorithm. *Bioinformatics* 2010 Jun 15;26(12):1481-1487.
- 2011 phage COGs. *J Bacteriol.* 2011 Apr;193(8):1806-1814.
- Orthologs and BBH. *Genome Biol Evol.* 2012 Jan;4(12):1286-1294.
- 2012 archaeal COGs. *Biol Direct* 2012 Dec 14;7:46.
- 2013 phage COGs. *J Bacteriol.* 2013 Mar;195(5):941-950.
- mimiCOGs. *Virol J.* 2013 Apr 4;10:106.
- 2014 update of 2003 COGs. *Nucleic Acids Res.* 2015 Jan;43:D261-D269. NEW
- 2014 archaeal COGs. *Life* 2015 Mar 10;5(1):818-840. NEW

The previously existing COG pages are retired.

Comments and questions to cogs@ncbi.nlm.nih.gov or to the project supervisor.

# 基因组注释信息

## Escherichia coli str. K-12 substr. MG1655, complete genome

NCBI Reference Sequence: NC_000913.3

FASTA Graphics

Go to:

```
LOCUS       NC_000913             4641652 bp    DNA     circular CON 09-MAR-2022
DEFINITION  Escherichia coli str. K-12 substr. MG1655, complete genome.
ACCESSION   NC_000913
VERSION     NC_000913.3
DBLINK      BioProject: PRJNA57779
            BioSample: SAMN02604091
KEYWORDS    RefSeq.
SOURCE      Escherichia coli str. K-12 substr. MG1655
  ORGANISM  Escherichia coli str. K-12 substr. MG1655
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales;
            Enterobacteriaceae; Escherichia.
REFERENCE   1  (bases 1 to 4641652)
  AUTHORS   Riley,M., Abe,T., Arnaud,M.B., Berlyn,M.K., Blattner,F.R.,
            Chaudhuri,R.R., Glasner,J.D., Horiuchi,T., Keseler,I.M., Kosuge,T.,
            Mori,H., Perna,N.T., Plunkett,G. III, Rudd,K.E., Serres,M.H.,
            Thomas,G.H., Thomson,N.R., Wishart,D. and Wanner,B.L.
  TITLE     Escherichia coli K-12: a cooperatively developed annotation
            snapshot--2005
  JOURNAL   Nucleic Acids Res. 34 (1), 1-9 (2006)
   PUBMED   16397293
  REMARK    Publication Status: Online-Only
REFERENCE   2  (bases 1 to 4641652)
  AUTHORS   Hayashi,K., Morooka,N., Yamamoto,Y., Fujita,K., Isono,K., Choi,S.,
            Ohtsubo,E., Baba,T., Wanner,B.L., Mori,H. and Horiuchi,T.
  TITLE     Highly accurate genome sequences of Escherichia coli K-12 strains
            MG1655 and W3110
  JOURNAL   Mol. Syst. Biol. 2, 2006 (2006)
   PUBMED   16738553
```

```
COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
            NCBI review. The reference sequence is identical to U00096.
            On Nov 3, 2013 this sequence version replaced NC_000913.2.
            Changes to proteins and annotation made on March 7, 2022.  Current
            U00096 annotation updates are derived from EcoCyc
            https://ecocyc.org/.  Suggestions for updates can be sent to
            biocyc-support@ai.sri.com. These updates are being generated from a
            collaboration  that includes EcoCyc, the University of Wisconsin,
            UniProtKB/Swiss-Prot, and the National Center for Biotechnology
            Information (NCBI).
            COMPLETENESS: full length.
FEATURES             Location/Qualifiers
     source          1..4641652
                     /organism="Escherichia coli str. K-12 substr. MG1655"
                     /mol_type="genomic DNA"
                     /strain="K-12"
                     /sub_strain="MG1655"
                     /db_xref="taxon:511145"
     gene            190..255
                     /gene="thrL"
                     /locus_tag="b0001"
                     /gene_synonym="ECK0001"
                     /db_xref="ASAP:ABE-0000006"
                     /db_xref="ECOCYC:EG11277"
                     /db_xref="GeneID:944742"
     CDS             190..255
                     /gene="thrL"
                     /locus_tag="b0001"
                     /gene_synonym="ECK0001"
                     /codon_start=1
                     /transl_table=11
                     /product="thr operon leader peptide"
                     /protein_id="NP_414542.1"
                     /db_xref="UniProtKB/Swiss-Prot:P0AD86"
                     /db_xref="ASAP:ABE-0000006"
                     /db_xref="ECOCYC:EG11277"
                     /db_xref="GeneID:944742"
                     /translation="MKRISTTITTTITITTGNGAG"
```

https://www.ncbi.nlm.nih.gov/nuccore/NC_000913.3/

```
gene            189874..190599
                /gene="rpsB"
                /locus_tag="b0169"
                /gene_synonym="ECK0168"
                /db_xref="ASAP:ABE-0000575"
                /db_xref="ECOCYC:EG10901"
CDS             189874..190599
                /gene="rpsB"
                /locus_tag="b0169"
                /gene_synonym="ECK0168"
                /codon_start=1
                /transl_table=11
                /product="30S ribosomal subunit protein S2"
                /protein_id="AAC73280.1"
                /db_xref="UniProtKB/Swiss-Prot:P0A7V0"
                /db_xref="ASAP:ABE-0000575"
                /db_xref="ECOCYC:EG10901"
                /translation="MATVSMRDMLKAGVHFGHQTRYWNPKMKPFIFGARNKVHIINLE
                KTVPMFNEALAELNKIASRKGKILFVGTKRAASEAVKDAALSCDQFFVNHRWLGGMLT
                NWKTVRQSIKRLKDLETQSQDGTFDKLTKKEALMRTRELEKLENSLGGIKDMGGLPDA
                LFVIDADHEHIAIKEANNLGIPVFAIVDTNSDPDGVDFVIPGNDDAIRAVTLYLGAVA
                ATVREGRSQDLASQAEESFVEAE"
gene            190857..191708
                /gene="tsf"
                /locus_tag="b0170"
                /gene_synonym="ECK0169"
                /db_xref="ASAP:ABE-0000579"
                /db_xref="ECOCYC:EG11033"
CDS             190857..191708
                /gene="tsf"
                /locus_tag="b0170"
                /gene_synonym="ECK0169"
                /codon_start=1
                /transl_table=11
                /product="protein chain elongation factor EF-Ts"
                /protein_id="AAC73281.1"
                /db_xref="UniProtKB/Swiss-Prot:P0A6P1"
                /db_xref="ASAP:ABE-0000579"
                /db_xref="ECOCYC:EG11033"
                /translation="MAEITASLVKELRERTGAGMMDCKKALTEANGDIELAIENMRKS
                GAIKAAKKAGNVAADGVIKTKIDGNYGIILEVNCQTDFVAKDAGFQAFADKVLDAAVA
                GKITDVEVLKAQFEEERVALVAKIGENINIRRVAALEGDVLGSYQHGARIGVLVAAKG
                ADEELVKHIAMHVAASKPEFIKPEDVSAEVVEKEYQVQLDIAMQSGKPKEIAEKMVEG
                RMKKFTGEVSLTGQPFVMEPSKTVGQLLKEHNAEVTGFIRFEVGEGIEKVETDFAAEV
                AAMSKQS"
```

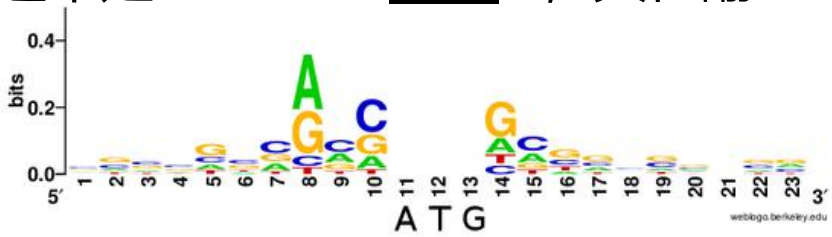
```
gene            228756..228875
                /gene="rrfH"
                /locus_tag="b0205"
                /gene_synonym="ECK0205"
                /db_xref="ASAP:ABE-0000686"
                /db_xref="ECOCYC:EG30076"
                /db_xref="GeneID:944898"
rRNA            228756..228875
                /gene="rrfH"
                /locus_tag="b0205"
                /gene_synonym="ECK0205"
                /product="5S ribosomal RNA"
                /db_xref="ASAP:ABE-0000686"
                /db_xref="ECOCYC:EG30076"
                /db_xref="GeneID:944898"
gene            228928..229004
                /gene="aspU"
                /locus_tag="b0206"
                /gene_synonym="ECK0206"
                /db_xref="ASAP:ABE-0000688"
                /db_xref="ECOCYC:EG30024"
                /db_xref="GeneID:944880"
tRNA            228928..229004
                /gene="aspU"
                /locus_tag="b0206"
                /gene_synonym="ECK0206"
                /product="tRNA-Asp"
                /note="tRNA-Asp(GUC)"
                /db_xref="ASAP:ABE-0000688"
                /db_xref="ECOCYC:EG30024"
                /db_xref="GeneID:944880"
```

https://www.ncbi.nlm.nih.gov/nuccore/NC_000913.3/

# 真核生物基因组结构注释

# Kozak序列规则

- Kozak序列是存在于真核生物mRNA的一段序列，位于mRNA 5’端帽子结构后面的一段核酸序列，通常是GCCRCCATGG，其在翻译的起始中有重要作用。




- 起始密码子：ATG/GTG

  第一个ATG的确定可依据Kozak规则，即第一个ATG侧翼序列的碱基分布所满足的统计规律，若将第一个ATG中的碱基A，T，G分别标为1，2，3位，则Kozak规则可描述如下：
  (1)第4位的偏好碱基为G
  (2)ATG的5’端约15bp范围的侧翼序列内不含碱基T
  (3)在-3，-6和-9位置，G是偏好碱基
  (4)除-3，-6和-9位，在整个侧翼序列区，C是偏好碱基

- 终止密码子：TAA，TAG，TGA
  - GC%=50% 终止密码子每64bp出现一次
  - GC%>50% 终止密码子每100-200bp出现一次

https://en.wikipedia.org/wiki/Kozak_consensus_sequence

The Kozak sequence is a nucleic acid motif that functions as the protein translation initiation site in most eukaryotic mRNA transcripts, which was determined by sequencing of 699 vertebrate mRNAs and named after the scientist who discovered it, Marilyn Kozak.

# 重复序列

重复序列可分为串联重复序列（Tandem repeat）和散在重复序列（Interspersed repeat）两大类：

- 串联重复序列：微卫星序列、小卫星序列等
- 散在重复序列：又称转座子元件TE
  - DNA转座子
  - 反转录转座子（retrotransposon)
    - LTR
    - LINE
    - SINE

# 重复序列注释方法

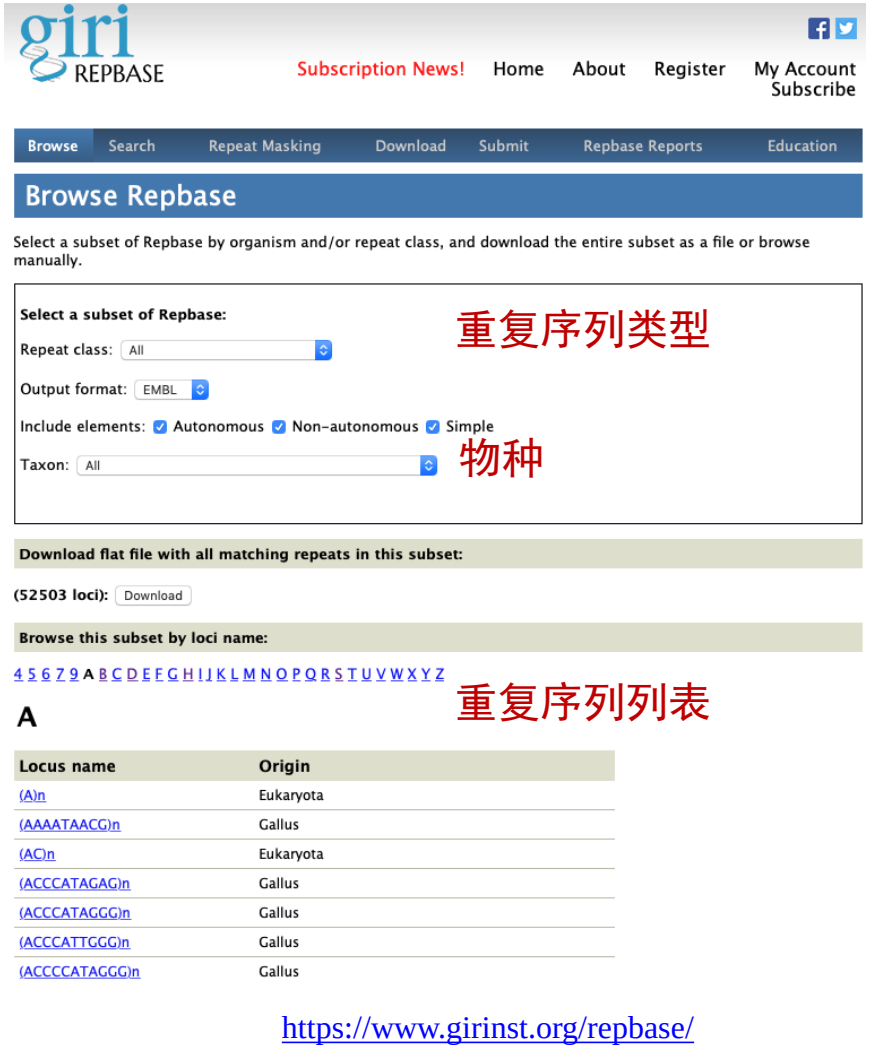
识别重复序列的方法分为两类：

- **序列比对方法**：一般采用Repeatmasker软件，**识别与已知重复序列相似的序列**，并对其进行分类。常用Repbase重复序列数据库。
- **从头预测方法**：利用重复序列或转座子自身的序列或结构特征构建从头预测算法或软件对序列进行识别。从头预测方法的优点在于能够根据转座子元件自身的结构特征进行预测，**不依赖于已有的转座子数据库，能够发现未知的转座子元件**。常见的从头预测方法有Recon、RepeatModeler等。

# 重复序列数据库Repbase

Repbase是真核生物基因组重复序列数据库，是多种注释工具（如Repeatmasker）的参考数据来源。

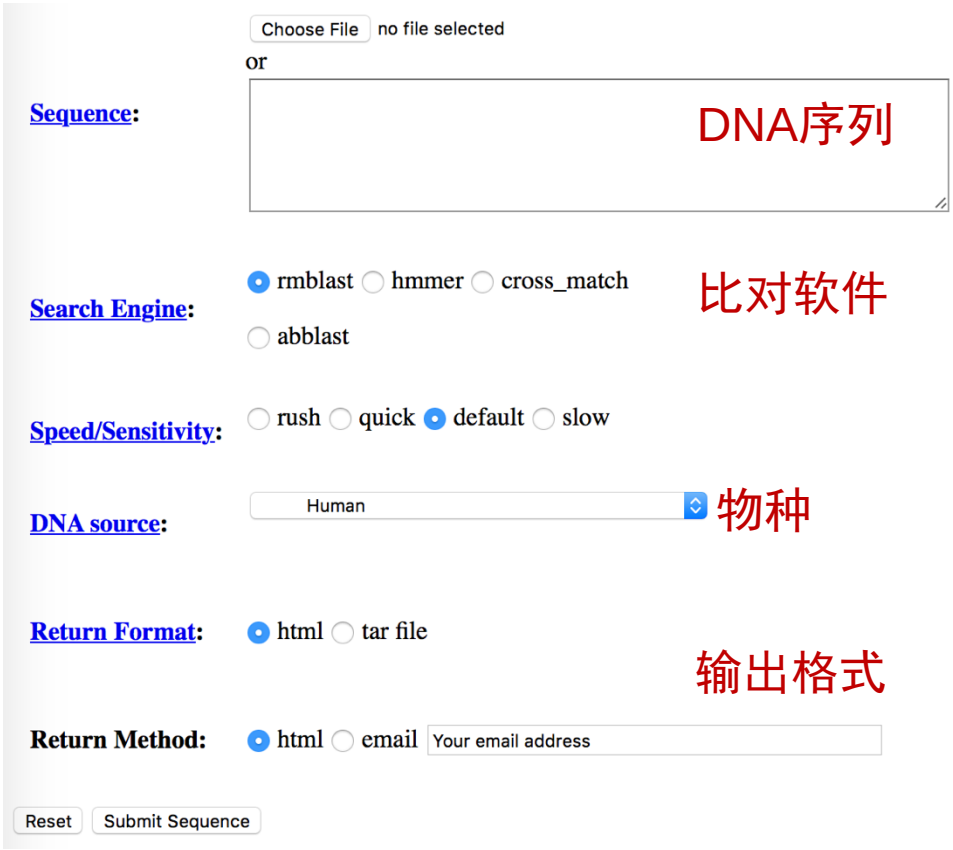
- Transposable Element
  - DNA transposon
    - Mariner/Tc1
    - hAT
    - MuDR
    - EnSpm/CACTA
    - piggyBac
    - P
    - Merlin
    - Harbinger
    - Transib
    - Novosib
    - Helitron
    - Polinton
    - Kolobok
    - ISL2EU
    - Crypton
      - CryptonA
  - LTR Retrotransposon
    - Gypsy
    - Copia
    - BEL
    - DIRS
  - Endogenous Retrovirus
    - ERV1
    - ERV2
    - ERV3
    - Lentivirus
    - ERV4
    - Lokiretrovirus
    - Spumaretrovirus
  - Non-LTR Retrotransposon
    - SINE
      - SINE1/7SL
      - SINE2/tRNA
      - SINE3/5S
      - SINE4




https://www.girinst.org/repbase/

# 重复序列注释

## RepeatMasker Web Server

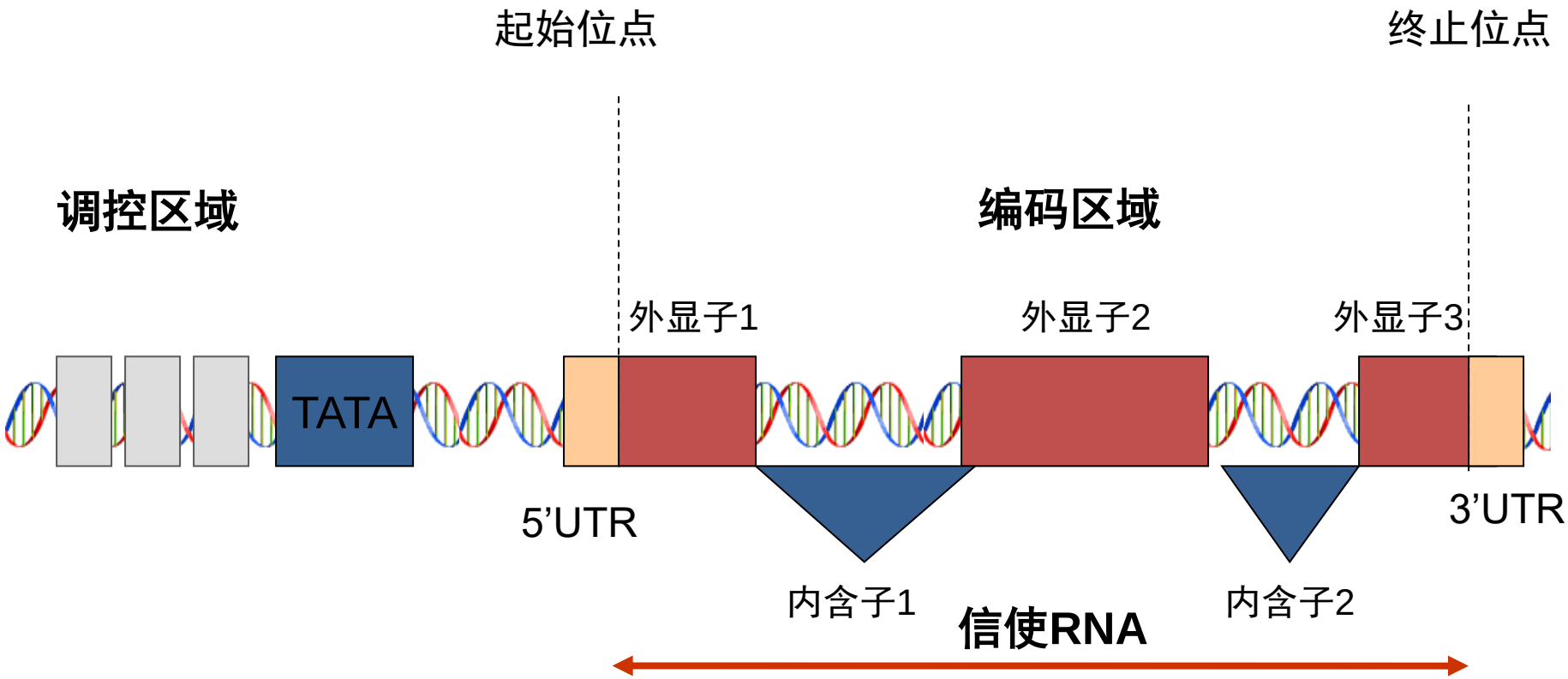



## 注释的重复序列类型

| Repeat Family | Size (bp) | Ratio (%) |
|---|---|---|
| Low_complexity | 2,011,687 | 0.51 |
| Microsatellite | 2,520,388 | 0.63 |
| Minisatellite | 9,815,590 | 2.47 |
| Satellite | 16,830,062 | 4.23 |
| SINE | | |
| *Sj-alpha1* | 4,293,863 | 1.08 |
| *Sj-alpha2* | 3,011,537 | 0.76 |
| LTR Retrotransposon | | |
| *Gulliver* | 711,318 | 0.18 |
| 18 Novel Sj-Retrotransposons | 23,803,907 | 5.98 |
| Non-LTR Retrotransposon | | |
| *Sj-pido* | 1,540,514 | 0.39 |
| *SjR1* | 3,358,700 | 0.84 |
| *SjR2* | 17,602,429 | 4.43 |
| 4 Novel Sj-Retrotransposons | 27,743,061 | 6.97 |
| 3 *Penelope*-like Retrotransposons | 3,938,326 | 0.99 |
| Transposon | 856,671 | 0.22 |
| undefined repeat | 41,582,296 | 10.45 |
| Total repeat | 159,620,349 | 40.13 |
| Genome Size* | 397,751,405 | 100 |

http://repeatmasker.org/cgi-bin/WEBRepeatMasker

# 真核生物蛋白质编码基因结构




From Case Western Reserve University

# 真核生物基因结构预测

- 通过基因结构预测，能够获得基因组详细的基因分布和结构信息，为功能注释和进化分析工作提供重要原料。
- 基因结构预测：包括预测基因组中的基因位点、开放性阅读框架（ORF）、翻译起始位点和终止位点、内含子和外显子区域、启动子、可变剪切位点以及蛋白质编码序列等。
- 基因结构预测主要采用三种方法：
  - 软件预测
  - 同源预测
  - 转录本预测

# 软件预测

- 软件预测方面主要是从基因的结构出发去预测序列中某一个部分是否存在基因。
- 软件预测
  - 优点：不需要该物种的其他信息，使用比较广泛。
  - 缺点：没有特异性，预测的准确性过低。
- 主要的代表软件有：
  - GlimmerHMM
  - Genscan
  - Augustus

# 真核生物编码基因预测

- Augustus：一个真核生物基因预测软件，有网页服务端和本地版，它基于Hidden-Markov Model（隐马尔科夫链模型HMM）的预测方法，**需要相近的物种编码序列进行训练之后再对新物种进行预测**，用于训练的物种应该具有较近的亲缘关系。

  http://augustus.gobics.de/

BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution* 2021

- GlimmerHMM：把一个基因看做几种特征序列（状态）的有序切换，这些特征序列包括内含子、基因间隔区、四种外显子（第一个外显子、中间的外显子、最后一个外显子、唯一的外显子），切换的过程形成马尔科夫链。

  http://ccb.jhu.edu/software/glimmerhmm/

# Augustus在线基因预测

You must **either** upload a *.tar.gz archive with AUGUSTUS species parameters from your computer **or** specify a project identifier: Help

**AUGUSTUS species parameters** *

选择物种

Upload an archive file (max. 100 MB): Help Choose File no file selected

or

specify a project identifier: Help

or

select an organism: Select One... Help

You must **either** upload a genome file from your computer **or** specify a web link to a genome file: Help

**Genome file** * (max. 250000 scaffolds) Help

Upload a file (max. 100 MB): Choose File no file selected 基因组序列

or

specify web link to genome file (max. 1 GB):

You may (optionally) also specify one or several of the following files that contain external evidence for prot

**cDNA file** ***Non-commercial users only*** Help

可选：cDNA序列

Upload a file (max. 100 MB): Choose File no file selected

or

specify web link to cDNA file (max. 1 GB):

**Hints file** Help

Upload a file (max. 200 MB): Choose File no file selected

The following checkboxes allow you to modify the gene prediction behavior of AUGUSTUS:

**UTR prediction** Help

predict UTRs (requires species-specific UTR parameters)

是否预测UTR

**Report genes on**

预测单/双链基因

both strands forward strand only reverse strand only

**Alternative transcripts:**

是否预测可变剪接

none few medium many

**Allowed gene structure:** Help

基因结构要求

predict any number of (possibly partial) genes
only predict complete genes
only predict complete genes - at least one
predict exactly one gene

ignore conflicts with other strand

## 基因预测结果

基因结构

```
# Predicted genes for sequence              rands
# start gene g1
HSACKI10        AUGUSTUS        gene    182     3060    0.63    +       .       g1
HSACKI10        AUGUSTUS        transcript      182     3060    0.63    +       .       g1.t1
HSACKI10        AUGUSTUS        start_codon     182     184     .       +       0       transcript_id "g1.t1";
HSACKI10        AUGUSTUS        initial 182     225     1       +       0       transcript_id "g1.t1"; gene_id
HSACKI10        AUGUSTUS        internal        1691    2300    0.86    +       1       transcript_id "g1.t1";
HSACKI10        AUGUSTUS        terminal        3049    3060    0.74    +       0       transcript_id "g1.t1";
HSACKI10        AUGUSTUS        CDS     182     225     1       +       0       transcript_id "g1.t1"; gene_id
HSACKI10        AUGUSTUS        CDS     1691    2300    0.86    +       1       transcript_id "g1.t1"; gene_id
HSACKI10        AUGUSTUS        CDS     3049    3060    0.74    +       0       transcript_id "g1.t1"; gene_id
HSACKI10        AUGUSTUS        stop_codon      3058    3060    .       +       0       transcript_id "g1.t1";
# coding sequence = [atgatgaaaccctgtctctaccaaaaagacaaaaaattagccagctcaagcaagcactactcttcctcccgcagtggag
# gaggaggaggaggaggaggatgtggaggaggaggaggagtgtcatccctaagaatttctagcagcaaaggctcccttggtggaggatttagctcaggg
# gggttcagtggtggctcttttagccgtgggagctctggtgggggatgctttgggggctcatcaggtggctatggaggattaggaggttttggtggagg
# tagctttcatggaagc                                tgggagttatggaggcagctttggagggggcaatttcggaggtggcagctttggtgggggca
# gctttggtggaggcgg                                gaggaggctttggtggtggatttggaggagatggtggcctttctctctggaaatgaaaaagta
# accatgcagaatctga                                ttggacaaagttcgggctctggaagaatcaaactatgagctggaaggcaaaatcaaggagtg
# gtatgaaaagcatggc                                tcgtgactacagcaaatactacaaaaccatcgatgaccttaaaaatcagagaacaacataa]
# protein sequence = [MMKPCLYQKDKKLASSSKHYSSSRSGGGGGGGGCGGGGGVSSLRISSSKGSLGGGFSSGGFSGGSFSRGSSGGGCFGG
# SSGGYGGLGGFGGGSFHGSYGSSSFGGSYGGSFGGGNFGGGSFGGGSFGGGGFGGGGFGGGFGGDGGLLSGNEKVTMQNLNDRLASYLDKVRAL
# EESNYELEGKIKEWYEKHGNSHQGEPRDYSKYYKTIDDLKNQRTT]
# Evidence for and against this transcript:
# % of transcript supported by hints (any source): 20
# CDS exons: 1/3
#       E:   1
# CDS introns: 0/2
# 5'UTR exons and introns: 0/0
# 3'UTR exons and introns: 0/0
# hint groups fully obeyed: 0
# incompatible hint groups: 5
#       E:   3 (gi|154723761,gi|13907559,gi|154736078)
#       M:   2
# end gene g1
###
```

CDS序列

预测证据

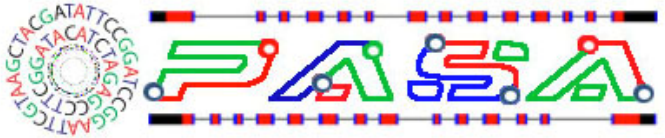
# 基于转录本注释基因结构

PASA是一种真核生物基因组注释工具，它利用表达的转录序列的剪接排列来自动建模基因结构，并保持基因结构注释与测序数据一致。PASA还识别并分类了转录本比对支持的所有剪接变体。

PASA的注释功能还包括：

- 注释 UTR 区域
- 外显子的添加、删除、边界调整
- 增加可变剪接变体
- 注释基因的ploy A位点
- 识别反义转录本
- 识别并分类所有发现的剪接变异

## Gene Structure Annotation and Analysis Using PASA

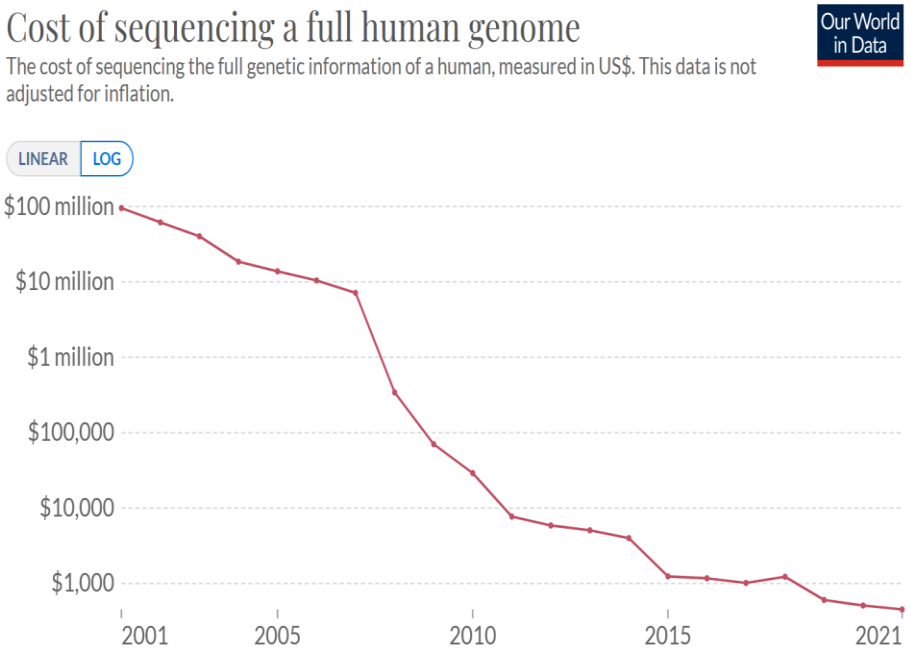


PASA, acronym for Program to Assemble Spliced Alignments (and pronounced 'pass-uh'), is a eukaryotic genome annotation tool that exploits spliced alignments of expressed transcript sequences to automatically model gene structures, and to maintain gene structure annotation consistent with the most recently available experimental sequence data. PASA also identifies and classifies all splicing variations supported by the transcript alignments.

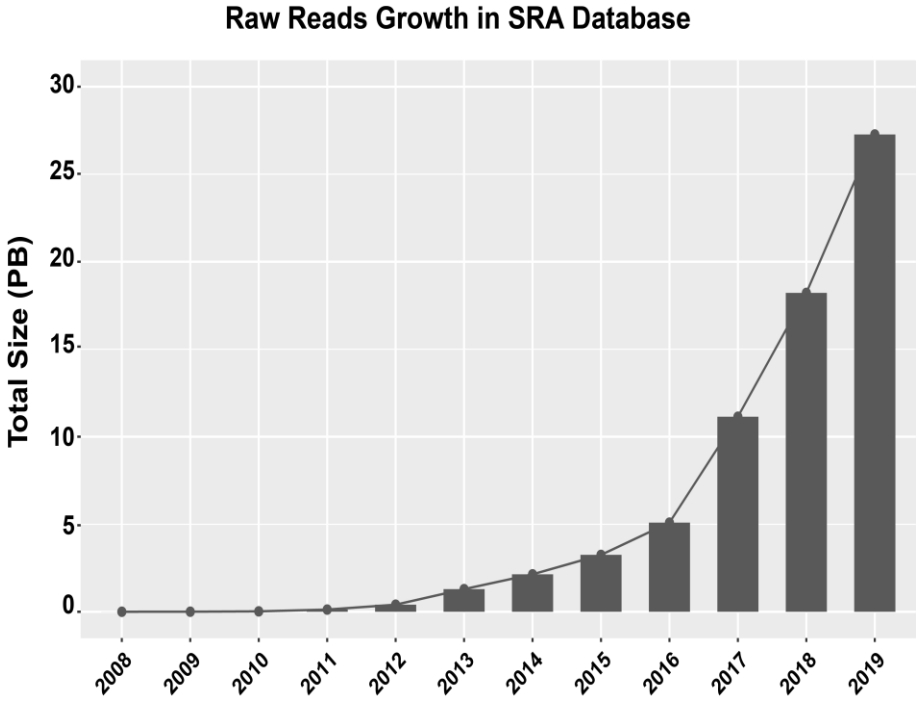
http://pasapipeline.github.io/

# 转录组数据

- 测序成本显著下降，已超越摩尔定律
- 测序数据的积累呈指数级增长
- 多组学数据（转录组、蛋白质组、表观组学等）迅速积累



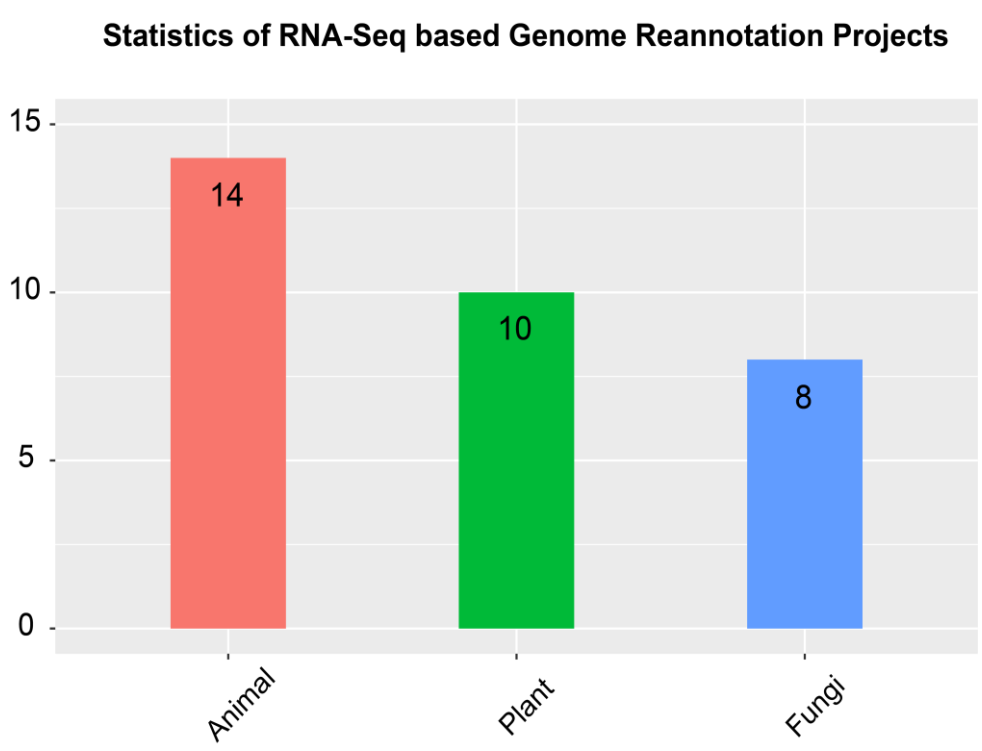

https://ourworldindata.org/grapher/cost-of-sequencing-a-full-human-genome

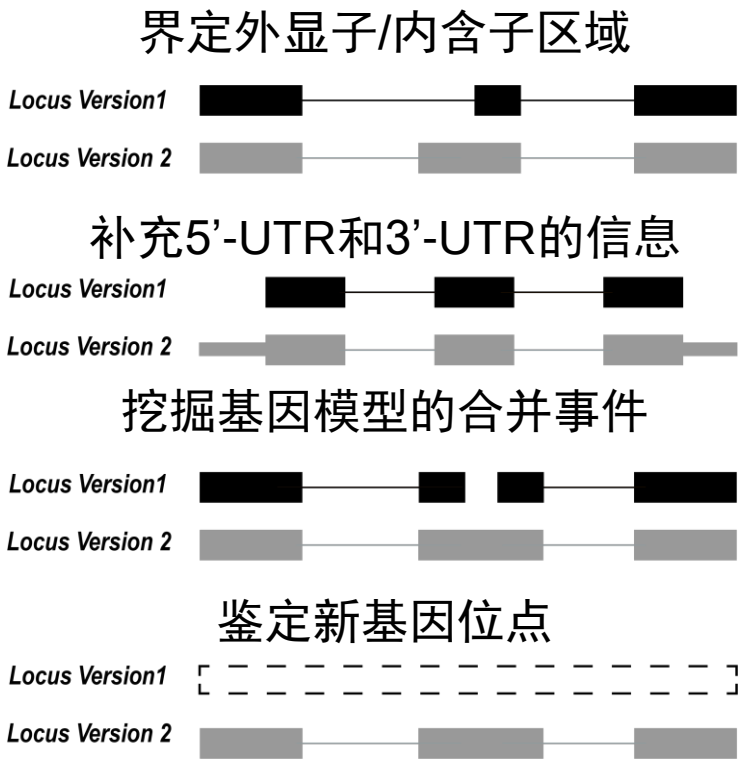



https://www.ncbi.nlm.nih.gov/sra (2019-03)

# 基于转录组数据的基因组重注释

RNA-Seq数据在基因组重注释研究中有巨大的应用潜力




Cerqueira GC, et al. *Nucleic Acids Research*, 2014

# 基于RNA-Seq数据的基因组重注释

- 基于RNA-Seq的植物 & 藻类基因组重注释（Genome Reannotation）
- 相关研究论文的发表数量呈逐年上升趋势

**2011**

黄瓜

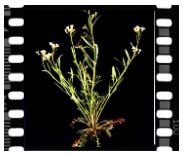
*Li et al.*, BMC genomics

**2014**

油菜

*Devisetty et al.*, G3

**2015**



琴叶拟南芥

*Rawat et al.*, PLoS One

野草莓

*Darwish et al.*, BMC genomics

**2017**

褐藻

*Cormier et al.*, NEW PHYTOL

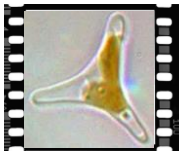
葡萄

*Harris et al.*，BMC genomics

**2018**

甜瓜

*Ruggieri et al.*, SCI REP-UK



三角褐指藻

*Rastogi et al.*, SCI REP-UK

拟南芥

*Cheng et al.*, PLANT J

猕猴桃

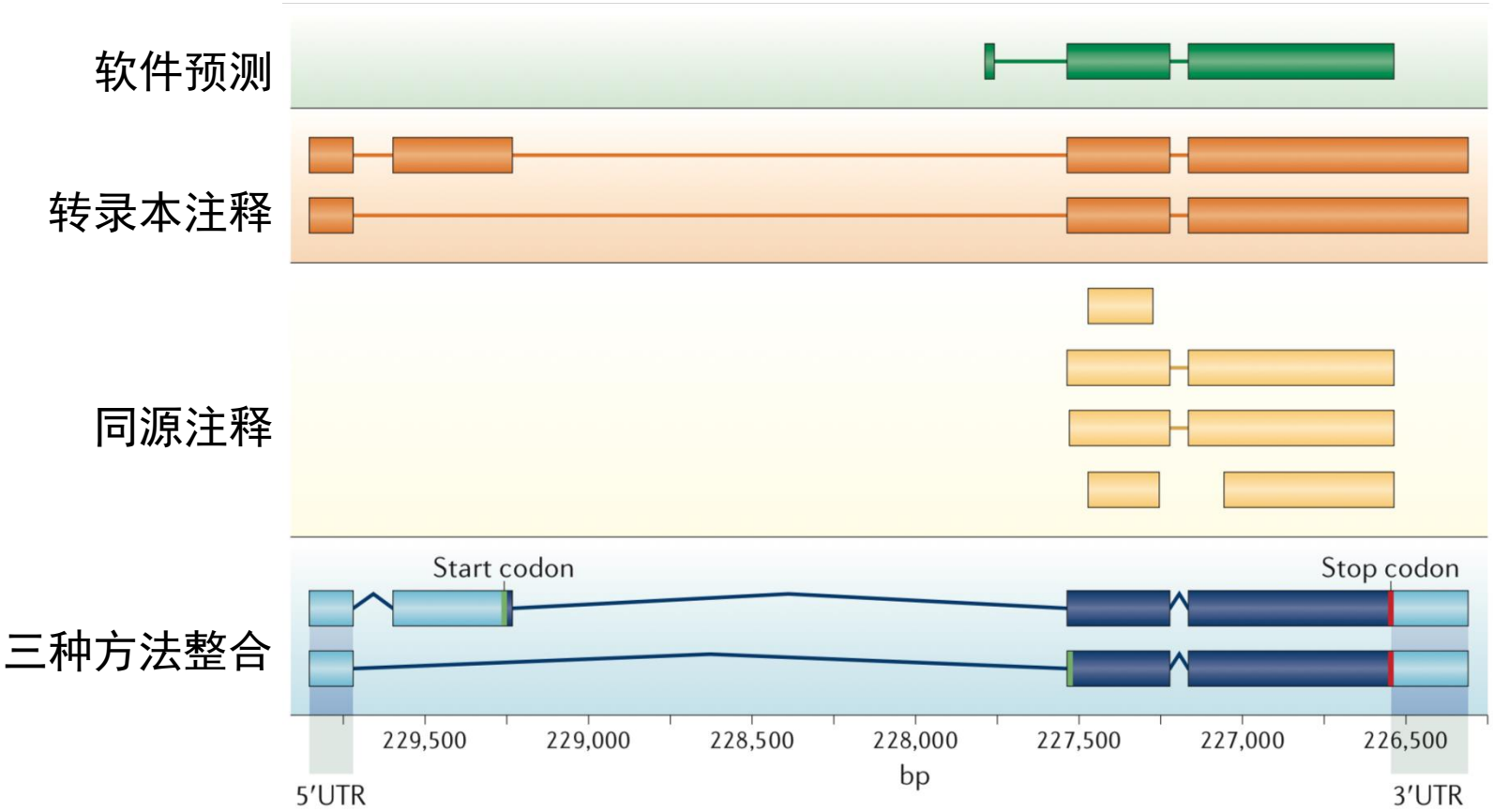
*Sarah et al.*, BMC genomics

牛角瓜

*Genevieve et al.*, G3

# 三种方法的整合

通过整合软件预测、基于同源注释和转录本预测三种方法的预测结果，得到最终的基因结构模型




**A beginner's guide to eukaryotic genome annotation**. *Nature Reviews Genetics* 2012

# 基因组功能注释

# 基因功能注释

**基因功能注释：**通过比对的方法根据已知功能的蛋白质编码基因序列预测未知蛋白质编码基因的功能。

- 功能序列（Functional Motif Detection）
- 基因本体（Gene Ontology Annotation）
- 代谢通路（KEGG Enrichment Analysis）



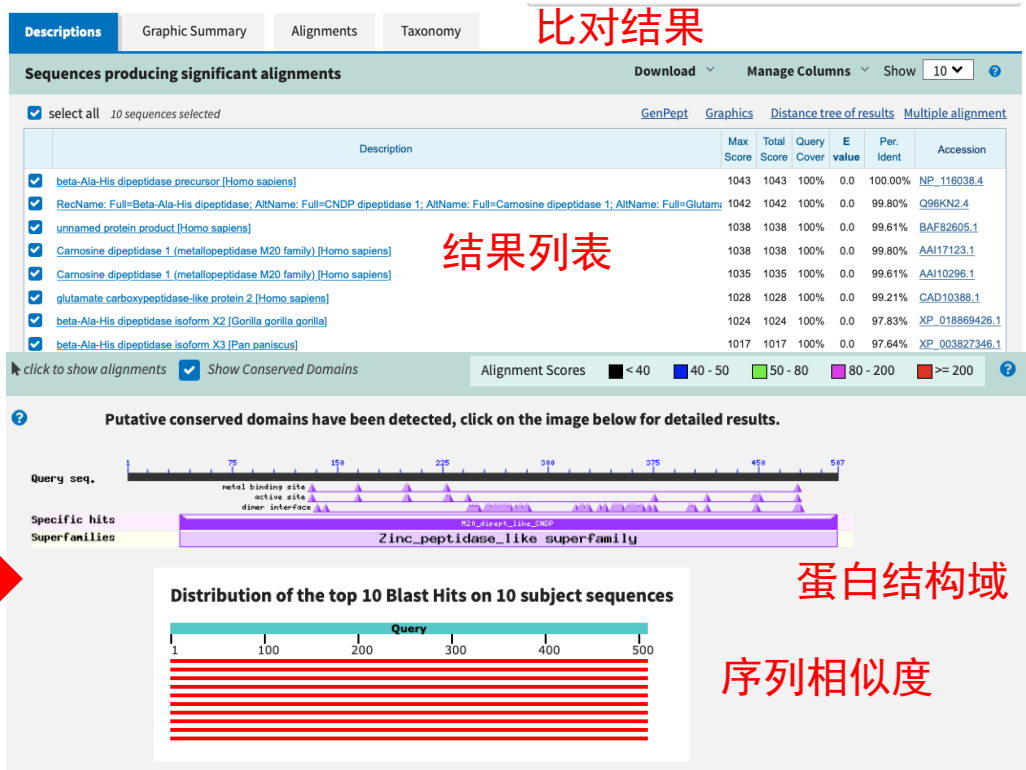


# 基因功能注释

普遍采用BLAST比对方法对预测出来的编码基因进行功能注释，通过与各种功能数据库（NCBI nr、Swiss-Prot等）进行蛋白质比对，获取该基因的功能信息。

- Swiss-Prot是经过注释的蛋白质序列数据库，由欧洲生物信息学研究所（EBI）维护。数据库由蛋白质序列条目构成，每个条目包含蛋白质序列、引用文献信息、分类学信息、注释等。
- NCBI nr 是非冗余的蛋白序列数据库，主要数据来源为GenBank CDS translations+PDB+SwissProt+PIR+PRF。

# 在线BLAST比对


```
Score          Expect  Method                    Identities      Positives       Gaps
1024 bits(2647) 0.0    Compositional matrix adjust. 496/507(98%) 501/507(98%) 0/507(0%)

Query  1    MDPKLGRMAASLLAVLLLLLERGMFSSPSPPPALLEKVFQYIDLHQDEFVQTLKEWVAIE  60
            MDPKLGRMAASLLAVLLLLLE GMFSSPSPPPALLEKVFQYIDLHQDEFVQTLKEWVAIE
Sbjct  1    MDPKLGRMAASLLAVLLLLLECGMFSSPSPPPALLEKVFQYIDLHQDEFVQTLKEWVAIE  60

Query  61   SDSVQPVPRFRQELFRMMAVAADTLQRLGARVASVDMGPQQLPDGQSLPIPPVILAELGS  120
            SDSVQPVPRFRQELFRMMA+AAD LQRLGARVA VDMGPQQLPDGQ+LPIPPVILAELGS
Sbjct  61   SDSVQPVPRFRQELFRMMAMAADMLQRLGARVALVDMGPQQLPDGQTLPIPPVILAELGS  120

Query  121  DPTKGTVCFYGHLDVQPADRGDGWLTDPYVLTEVDGKLYGRGATDNKGPVLAWINAVSAF  180
            DPTKGTVCFYGHLDVQPADRGDGWLTDPYVLTEVDGKLYGRGATDNKGPVLAWINAVSAF
Sbjct  121  DPTKGTVCFYGHLDVQPADRGDGWLTDPYVLTEVDGKLYGRGATDNKGPVLAWINAVSAF  180

Query  181  RALEQDLPVNIKFIIEGMEEAGSVALEELVEKEKDRFFSGVDYIVISDNLWISQRKPAIT  240
            RAL+QDLPVNIKFIIEGMEEAGSVALEELVEKEKDRFFSGVDYIVISDNLWISQRKPAIT
Sbjct  181  RALKQDLPVNIKFIIEGMEEAGSVALEELVEKEKDRFFSGVDYIVISDNLWISQRKPAIT  240

Query  241  YGTRGNSYFMVEVKCRDQDFHSGTFGGILHEPMADLVALLGSLVDSSGHILVPGIYDEVV  300
            YGTRGNSYFMVEVKCRDQDFHSGTFGGILHEPMADLVALLGSLVDSSGHILVPGIYDEVV
Sbjct  241  YGTRGNSYFMVEVKCRDQDFHSGTFGGILHEPMADLVALLGSLVDSSGHILVPGIYDEVV  300
```

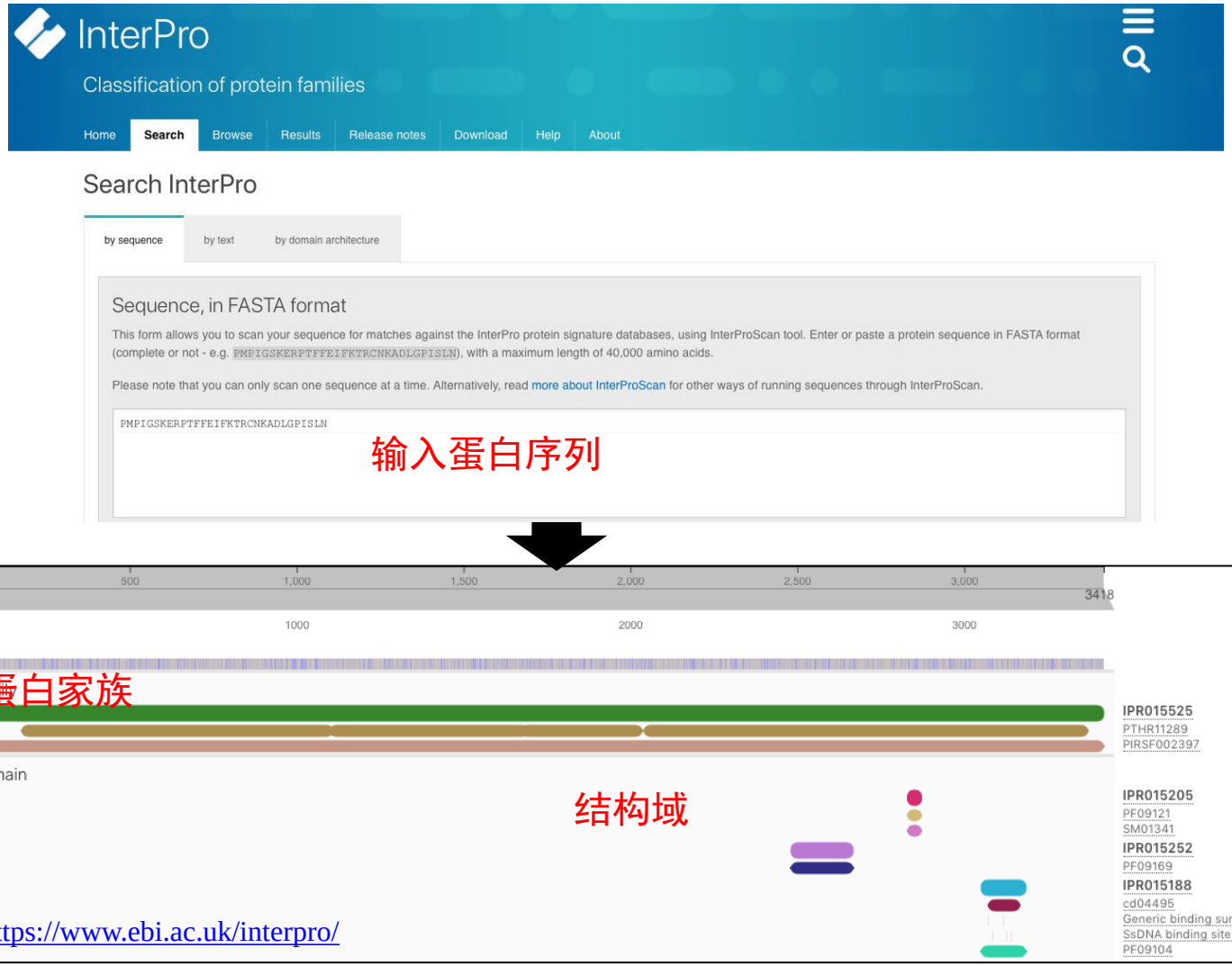
序列联配

# 蛋白结构域

蛋白质的功能与其结构密切相关，一个蛋白质的保守结构域在一定程度上体现了该蛋白质的功能

- InterPro：**集成了蛋白质家族、结构域和功能位点的非冗余蛋白质特征序列数据库**，采用interproscan软件可以对新蛋白质序列通过序列比对或者HMM算法等搜索与interpro蛋白质特征序列匹配预测蛋白质各种结构功能域、信号肽、跨膜特征、蛋白质螺旋结构等。(http://www.ebi.ac.uk/interpro/)
- CDD（Conserved Domain Database，蛋白质保守结构域数据库）：**收集了大量保守结构域序列信息和蛋白质序列信息，**通过搜索可获得蛋白质序列中所含的保守结构域信息，从而分析、预测该蛋白质的功能。(https://www.ncbi.nlm.nih.gov/cdd/)

# 在线分析蛋白结构域（InterPro）



# 在线分析蛋白结构域（CDD）

Search for Conserved Domains within a protein or coding nucleotide sequence

Enter **protein** or **nucleotide** query as accession, gi, or sequence in FASTA format. For multiple protein queries, use Batch CD-Search.

输入蛋白序列

**OPTIONS**

Search against database: CDD v3.17 - 52910 PSSMs

Expect Value threshold: 0.010000

Apply low-complexity filter

Composition based statistics adjustment

Force live search

Rescue borderline hits  Suppress weak overlapping hits

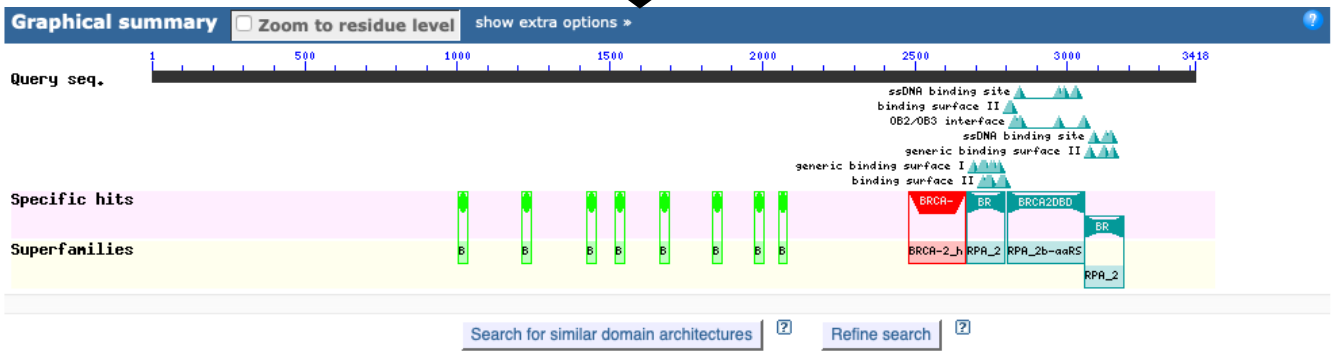
Maximum number of hits: 500

Result mode: Concise  Standard  Full

Submit  Reset

https://www.ncbi.nlm.nih.gov/cdd/

结构域




**List of domain hits**

| Name | Accession | Description | Interval | E-value |
|---|---|---|---|---|
| [+] BRCA-2_helical | pfam09169 | BRCA2, helical; Members of this family adopt a helical structure, consisting of a four-helix ... | 2481-2667 | 9.29e-125 |
| [+] BRCA2DBD_OB2 | cd04494 | BRCA2DBD_OB2: A subfamily of OB folds corresponding to the second OB fold (OB2) of the ... | 2804-3054 | 4.26e-107 |
| [+] BRCA-2_OB3 | pfam09104 | BRCA2, oligonucleotide/oligosaccharide-binding, domain 3; Members of this family assume an OB ... | 3054-3186 | 2.11e-68 |
| [+] BRCA-2_OB1 | pfam09103 | BRCA2, oligonucleotide/oligosaccharide-binding, domain 1; Members of this family assume an OB ... | 2670-2793 | 1.09e-61 |
| [+] BRCA2 | pfam00634 | BRCA2 repeat; The alignment covers only the most conserved region of the repeat. | 1213-1244 | 8.29e-09 |
| [+] BRCA2 | pfam00634 | BRCA2 repeat; The alignment covers only the most conserved region of the repeat. | 1518-1549 | 3.08e-08 |
| [+] BRCA2 | pfam00634 | BRCA2 repeat; The alignment covers only the most conserved region of the repeat. | 1003-1034 | 8.76e-08 |
| [+] BRCA2 | pfam00634 | BRCA2 repeat; The alignment covers only the most conserved region of the repeat. | 1973-2003 | 1.05e-07 |
| [+] BRCA2 | pfam00634 | BRCA2 repeat; The alignment covers only the most conserved region of the repeat. | 1665-1696 | 1.90e-07 |
| [+] BRCA2 | pfam00634 | BRCA2 repeat; The alignment covers only the most conserved region of the repeat. | 1422-1453 | 3.77e-06 |
| [+] BRCA2 | pfam00634 | BRCA2 repeat; The alignment covers only the most conserved region of the repeat. | 2052-2082 | 1.08e-04 |
| [+] BRCA2 | pfam00634 | BRCA2 repeat; The alignment covers only the most conserved region of the repeat. | 1838-1869 | 4.89e-04 |

# 基因本体论和代谢途径

# 本体论（Ontology）

- Ontology 是特定领域信息组织的一种形式，是**领域知识规范的抽象和描述**，是表达、共享、重用知识的方法。
- Ontology 是**知识体系构建的关键技术**，知识图谱是一种人工智能技术，它的关键在于让计算机能够处理人类的知识。然而，人类脑海中的知识通常是直觉性的，我们无法将这种直觉性的知识直接输入给计算机，Ontology 就是一种对知识建模，使计算机能够识别人类知识的方法。
- 本体（Ontology）通过对于**概念（Concept）、术语（Terminology）及其相互关系（Relation, Property）的规范化（Conceptualization）描述**，勾画出某一领域的基本知识体系和描述语言。

# 基因本体（Gene Ontology）

**基因本体（Gene Ontology，简称GO）是一种系统地对物种基因及其产物属性进行注释的方法和过程。**基因本体知识库是世界上最大的基因功能信息资源。这些知识既是**人类可读**的，又是**机器可读**的，是生物医学研究中大规模分子生物学和遗传学实验的计算分析的基础。其目标是：

- 维护和发展基因及其产物属性描述的词汇；
- 注释基因及其产物，传播注释数据；
- 提供方便的工具访问数据；
- 实现在实验数据的基础上，使用GO进行程式解析，例如基因富集组分分析。

http://geneontology.org/


A resource with **controlled vocabulary** to describe the function of genes and gene products

Founded in 1998 when researchers studying the genome of three model organisms—*Drosophila melanogaster* (fruit fly), *Mus musculus* (mouse), and *Saccharomyces cerevisiae* (brewer's or baker's yeast)

# GO的组成

- GO中最基本的概念是term
- GO里的每条记录(entry) 都有一个唯一的编号GO:NNNNNNN，对应一个term，比如“cell”、“DNA binding”、“signal transduction”
- 每个term属于一个ontology，总共有3个ontology：

**1. 分子功能（Molecular Function)**

由基因产物进行的分子水平的活动。分子功能术语描述在分子水平上发生的活动，如“催化”或“运输”。

**2. 细胞组件（Cellular Component)**

细胞的组成（如线粒体）或稳定的大分子复合物（如核糖体）。

**3. 生物过程（Biological Process)**

由多个分子活动完成的过程，例如：DNA修复、信号转导。

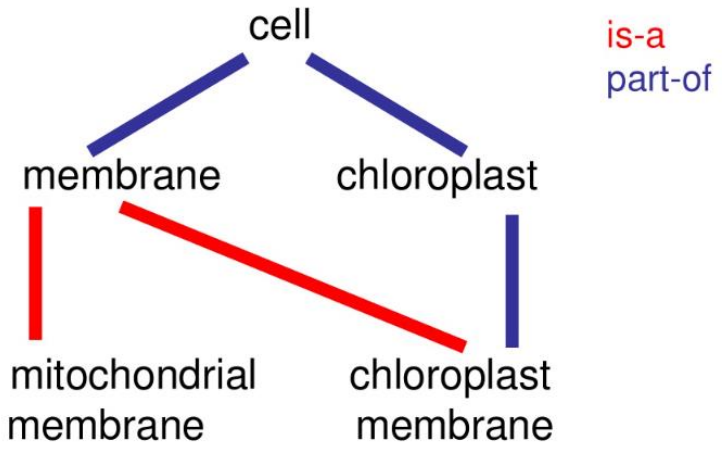
注意，一个生物过程并不等同于一个途径。

# GO功能分类

- 一个基因可能会同时出现在多个ontology里，例如：cytochrome（细胞色素）的molecular function是oxidoreductase activity（氧化还原酶活性），而biological process是oxidative phosphorylation（氧化磷酸化），cellular component是mitochondrial inner membrane（线粒体内膜）
- Term之间存在两种关系：

  is_a表示简单的包含关系

  part_of表示部分关系
- GO的结构是有向无环图（DAG），一个term可以有多个parent



# GO数据结构

| | |
|---|---|
| **Accession** | GO:0051718 |
| **Name** | DNA (cytosine-5-)-methyltransferase activity, acting on CpG substrates |
| **Ontology** | molecular_function |
| **Synonyms** | HpaII methylase, HpaII' methylase, M.BsuRIa, M.BsuRIb |
| **Alternate IDs** | None |
| **Definition** | Catalysis of the reaction: S-adenosyl-L-methionine + CpG (in DNA) = S-adenosyl-L-homocysteine + 5-MeCpG (in DNA). *Source:* EC:2.1.1.37, PMID:15689527 |
| **Comment** | None |
| **History** | See term history for GO:0051718 at QuickGO |

Data health ♥

- GO:0008168 methyltransferase activity
  - GO:0090116 C-5 methylation of cytosine
  - GO:0009008 DNA-methyltransferase activity
  - GO:0008757 S-adenosylmethionine-dependent methyltransferase activity
    - **GO:0003886 DNA (cytosine-5-)-methyltransferase activity**
      - GO:0051718 DNA (cytosine-5-)-methyltransferase activity, acting on CpG substrates
      - GO:0051719 DNA (cytosine-5-)-methyltransferase activity, acting on CpN substrates
      - GO:0051720 DNA (cytosine-5-)-methyltransferase activity, acting on CpNpG substrates

http://amigo.geneontology.org/amigo/term/GO:0051718

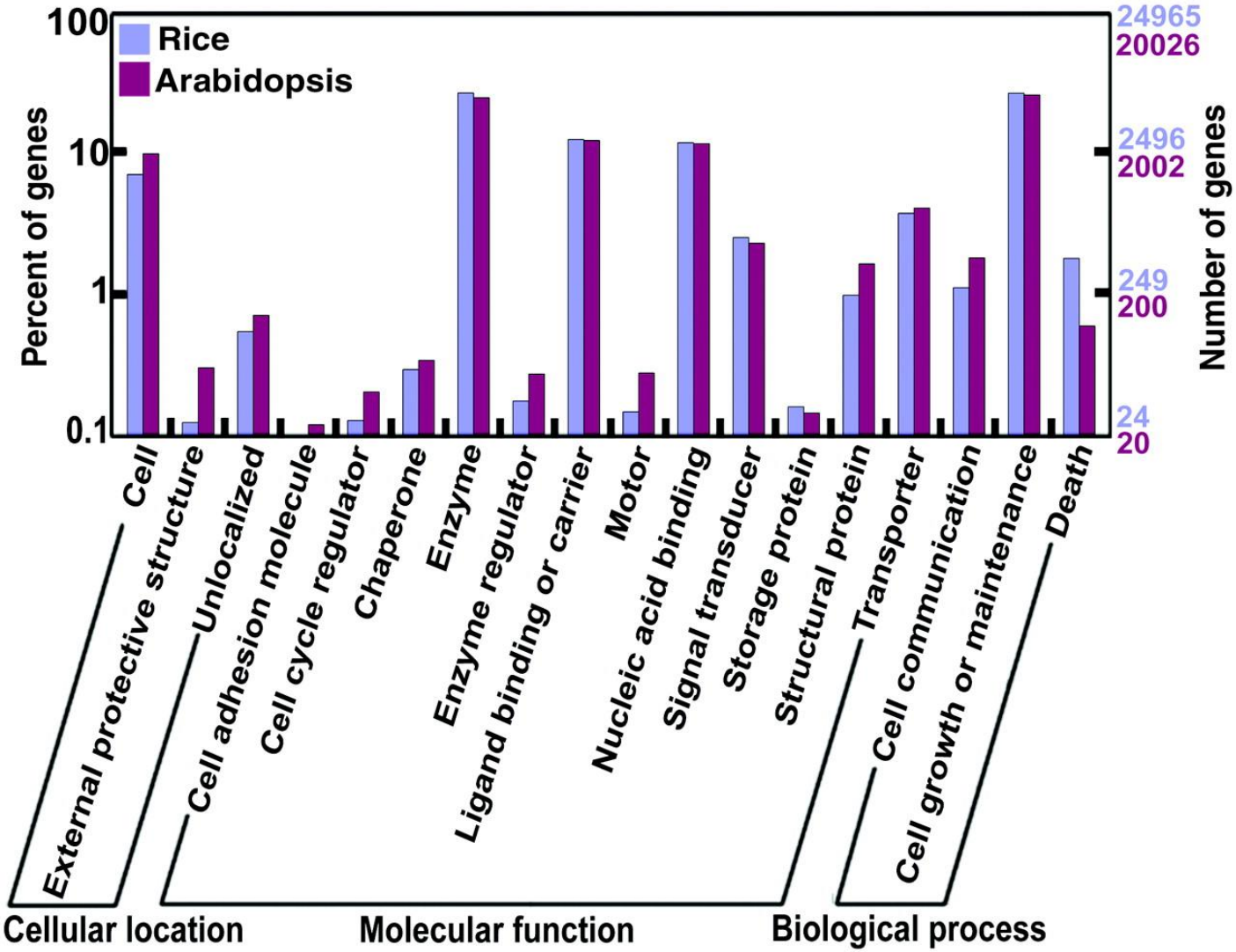
# Principles of GO Annotation

A GO annotation is a statement about the function of a particular gene. GO annotations are created by associating a gene or gene product with a GO term.

- Annotations represent the **normal functions** of gene products.
- Each annotation is supported by **GO Evidence Codes** from the Evidence and Conclusions Ontology and a reference.
- Gene products are annotated to the **most granular term** in the ontology that is supported by the available evidence.
- GO annotations are meant to reflect the **most up-to-date vi**ew of a gene product's role in biology.
- Because biological knowledge changes, annotations for a given gene product may change to **reflect changes** in knowledge and/or changes in the ontology.
- There is an **open-world assumption**, that is, if a gene product is unannotated then its role is still unknown.

http://geneontology.org/docs/go-annotations/

# 基因集的GO分类统计


Rice
Arabidopsis
Percent of genes
100
10
1
0.1
Number of genes
24965
20026
2496
2002
249
200
24
20
Cell
External protective structure
Unlocalized
Cell adhesion molecule
Cell cycle regulator
Chaperone
Enzyme
Enzyme regulator
Ligand binding or carrier
Motor
Nucleic acid binding
Signal transducer
Storage protein
Structural protein
Transporter
Cell communication
Cell growth or maintenance
Death
Cellular location
Molecular function
Biological process


http://amigo.geneontology.org/

## DNA (cytosine-5-)-methyltransferase activity, acting on CpG substrates

Term Information

GO详细信息

**Accession** GO:0051718
**Name** DNA (cytosine-5-)-methyltransferase activity, acting on CpG substrates
**Ontology** molecular_function
**Synonyms** HpaII methylase, HpaII' methylase, M.BsuRIa, M.BsuRIb
**Alternate IDs** None
**Definition** Catalysis of the reaction: S-adenosyl-L-methionine + CpG (in DNA) = S-adenosyl-L-homocysteine + 5-MeCpG (in DNA). *Source:* EC:2.1.1.37, PMID:15689527
**Comment** None
**History** See term history for GO:0051718 at QuickGO
**Subset** None
**Related** Link to all **genes and gene products** annotated to DNA (cytosine-5-)-methyltransferase activity, acting on CpG substrates.
Link to all direct and indirect **annotations** to DNA (cytosine-5-)-methyltransferase activity, acting on CpG substrates.
Link to all direct and indirect **annotations download** (limited to first 10,000) for DNA (cytosine-5-)-methyltransferase activity, acting on CpG substrates.

Annotations | Graph Views | Inferred Tree View | Neighborhood | Mappings

Total annotations: 4; showing: 1-4
Results count 10
«First <Prev Next> Last» Download

| Gene/product | Gene/product name | Annotation qualifier | GO class (direct) | Annotation extension | Contributor | Organism | Evidence | Evidence with | PANTHER family | Type | Isoform | Reference | Date |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dnmt3b | DNA methyltransferase 3B | | DNA (cytosine-5-)-methyltransferase activity, acting on CpG substrates | | MGI | Mus musculus | IDA | | dna cytosine-5- -methyltransferase 3-related pthr23068 | protein | | MGI:MGI:5811689 PMID:27856912 | 20171129 |
| Dnmt3c | DNA methyltransferase 3C | | DNA (cytosine-5-)-methyltransferase activity, acting on CpG substrates | | UniProt | Mus musculus | IMP | | dna cytosine-5- -methyltransferase 3-related pthr23068 | protein | | MGI:MGI:5913574 PMID:28854222 | 20181018 |
| Dnmt3c | DNA methyltransferase 3C | | DNA (cytosine-5-)-methyltransferase | | MGI | Mus musculus | IDA | | dna cytosine-5- -methyltransferase 3-related | protein | | MGI:MGI:5811689 PMID:27856912 | 20171129 |

## GO树状列表

- biological_process 1125537
  - behavior 6390
  - biological adhesion 15859
  - biological phase 334
  - biological regulation 389885
  - biomineralization 1306
  - carbohydrate utilization 49
  - carbon utilization 575
  - cell aggregation 542
  - cell killing 1497
  - cell population proliferation 15442
  - cellular component organization or biogenesis 213702
  - cellular process 803938
  - detoxification 9934
  - developmental process 99685
  - growth 19563
  - immune system process 26377
  - localization 229720
  - locomotion 21527
  - metabolic process 689544
  - multi-organism process 42573
  - multicellular organismal process 100800
  - negative regulation of biological process 89358
  - nitrogen utilization 1068
  - organ 3070
  - phosphorus utilization 0
  - pigmentation 1401
  - positive regulation of biological process 107029
  - regulation of biological process 340061
  - reproduction 30667
  - reproductive process 30015
  - response to stimulus 249698
  - rhythmic process 3393
  - signaling 138764
  - sulfur utilization 49
- cellular_component 1054742
- molecular_function 1051592

# GO浏览和查询

## DNA (cytosine-5-)-methyltransferase activity, acting on CpG substrates

Term Information

**Accession** GO:0051718 — Data health

**Name** DNA (cytosine-5-)-methyltransferase activity, acting on CpG substrates

**Ontology** molecular_function

**Synonyms** HpaII methylase, HpaII' methylase, M.BsuRIa, M.BsuRIb

**Alternate IDs** None

**Definition** Catalysis of the reaction: S-adenosyl-L-methionine + CpG (in DNA) = S-adenosyl-L-homocysteine + 5-MeCpG (in DNA). *Source:* EC:2.1.1.37, PMID:15689527

**Comment** None

**History** See term history for GO:0051718 at QuickGO

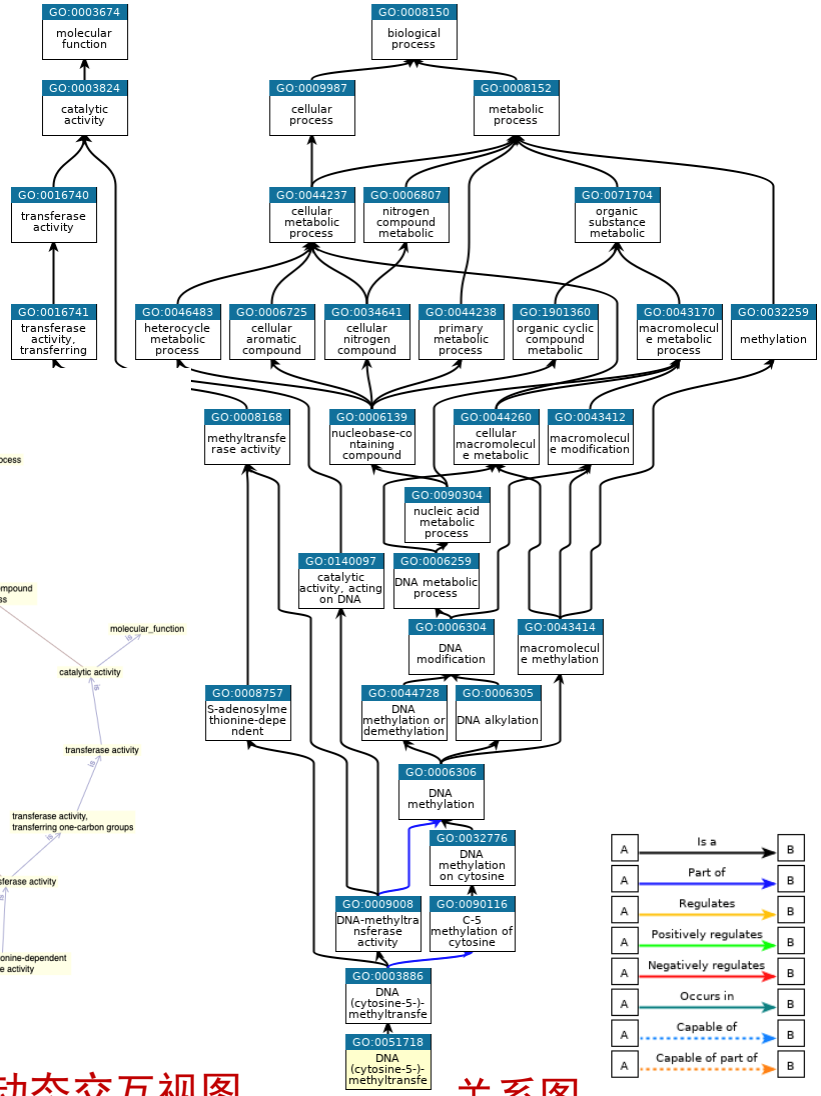
**Subset** None

**Related**
- Link to all **genes and gene products** annotated to DNA (cytosine-5-)-methyltransferase activity, acting on CpG substrates.
- Link to all direct and indirect **annotations** to DNA (cytosine-5-)-methyltransferase activity, acting on CpG substrates.
- Link to all direct and indirect **annotations download** (limited to first 10,000) for DNA (cytosine-5-)-methyltransferase activity, acting on CpG substrates.

Annotations | Graph Views | Inferred Tree View | Neighborhood | Mappings


树状图

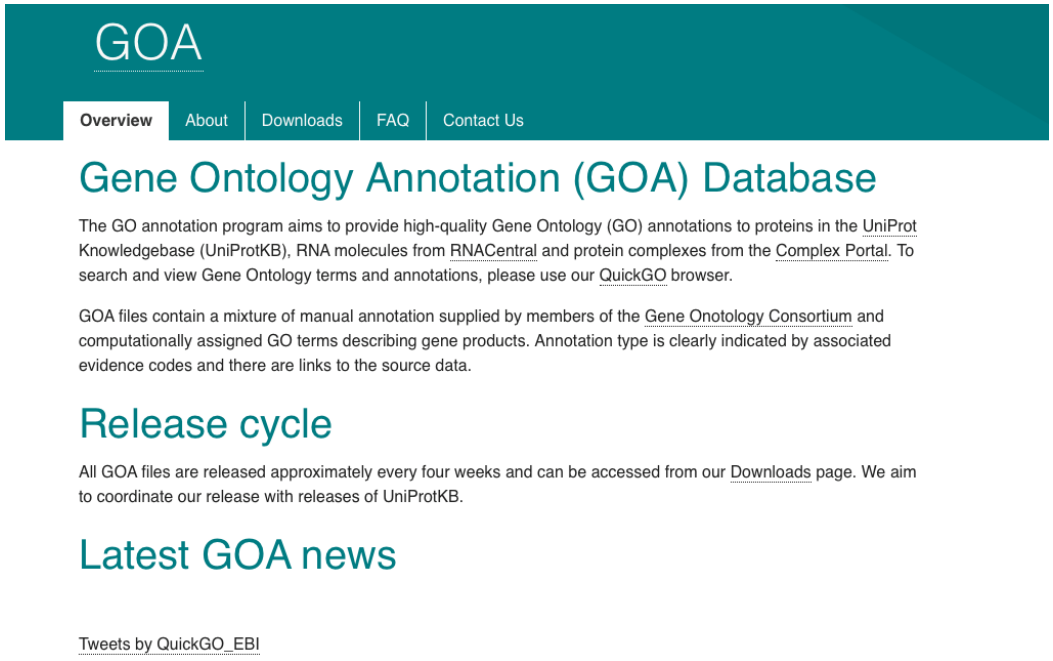

动态交互视图




关系图

# GO注释数据库（GOA）

GO注释程序旨在为UniProt知识库中的蛋白（UniProtKB）、RNACentral的RNA分子和其他蛋白质复合体提供高质量的基因本体（GO）注释




https://www.ebi.ac.uk/GOA

| Annotation Sets | | |
|---|---|---|
| UniProtKB | Current Files | Archive Files |
| Human | Current Files | Archive Files |
| Mouse | Current Files | Archive Files |
| Rat | Current Files | Archive Files |
| Arabidopsis | Current Files | Archive Files |
| Zebrafish | Current Files | Archive Files |
| Chicken | Current Files | Archive Files |
| Cow | Current Files | Archive Files |
| Dicty | Current Files | Archive Files |
| Dog | Current Files | Archive Files |
| Pig | Current Files | Archive Files |
| Fly | Current Files | Archive Files |
| Worm | Current Files | Archive Files |
| Yeast | Current Files | Archive Files |
| PDB | Current Files | Archive Files |
| Proteomes | Current Files | *See note below |

已注释的物种/数据库

# KEGG Pathway代谢途径

KEGG (Kyoto Encyclopedia of Genes and Genomes)代谢路径数据库是一组手工绘制的图形图，称为KEGG路径图，表示代谢、遗传信息处理、环境信息处理、细胞过程、组织系统、人类疾病和药物开发的分子路径。KEGG中存在三大类代谢图，每个数据路的pathway都有相应的唯一编号。

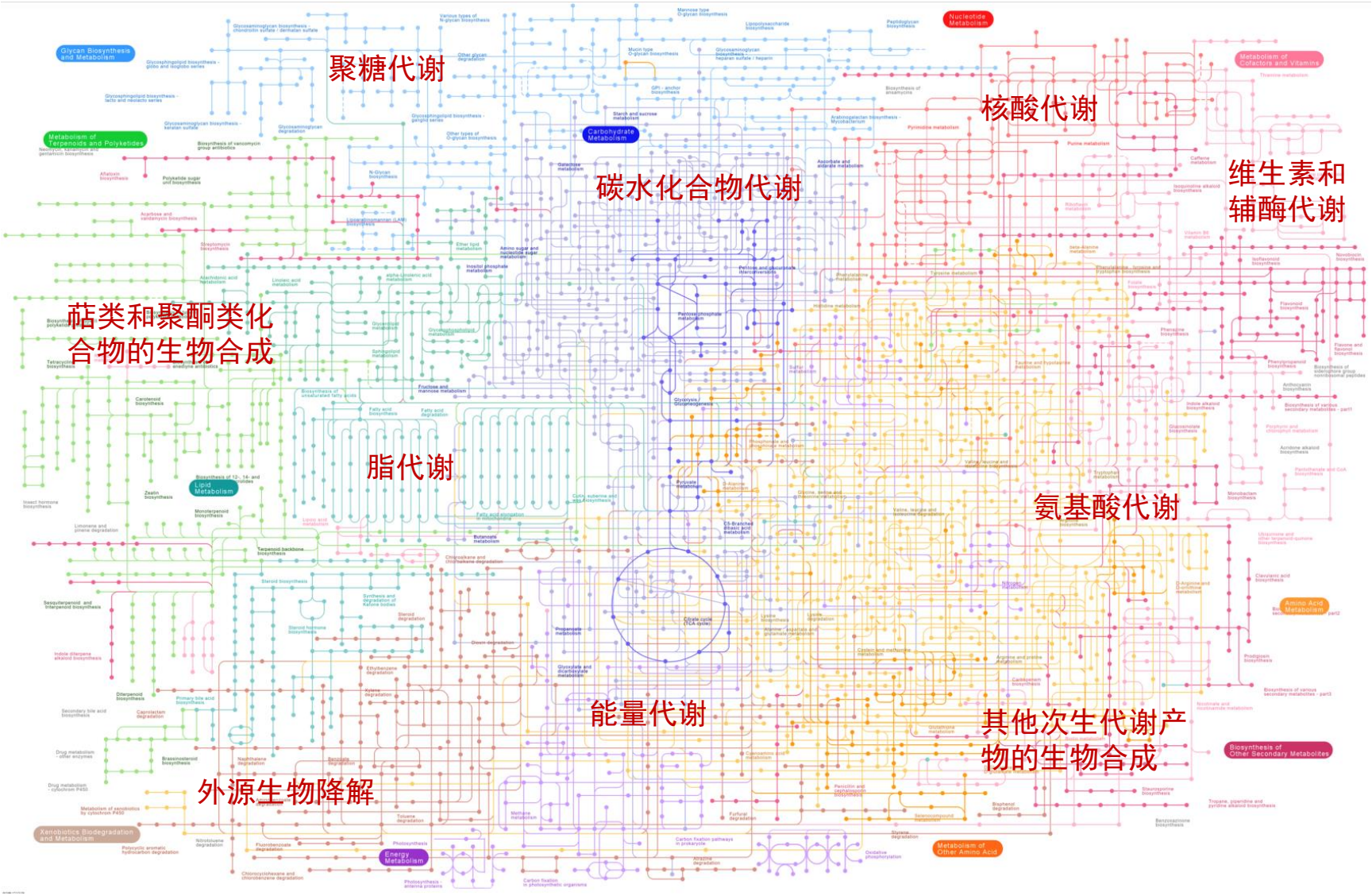
- reference pathway：根据已有的知识绘制的的具有一般参考意义的代谢图。通路图中的小框都是白色，在KEGG中名字以map开头，比map00010。
- species-specific pathway：物种特有代谢通路图。绿色小框为该物种特有的基因或酶。名字为特定物种种属英文缩写，比如人的糖酵解通路图hsa00010。
- 以ko/ec/rn开头的Reference pathway：ko通路中的节点只代表基因；ec通路中的节点只代表相关的酶；rn通路中的节点只表示该点参与的某个反应、反应物及反应类型。底色以蓝色表示。

https://www.genome.jp/kegg/

金久实 (Minoru Kanehisa)
日本京都大学

- 1948 年出生于日本长崎
- 1976 年在东京大学获物理系博士学位，之后在霍普金斯医学院从事博士后研究
- 1981年成为阿拉莫斯国家实验室的研究科学家，在此期间参与GenBank数据库开发
- 1995年至今，研发KEGG，发表20多篇数据库文章
- 2018年，被美国科睿维安列为可能获诺贝尔奖人选之一

# 主要代谢途径


聚糖代谢
核酸代谢
碳水化合物代谢
维生素和
辅酶代谢
萜类和聚酮类化
合物的生物合成
脂代谢
氨基酸代谢
能量代谢
其他次生代谢产
物的生物合成
外源生物降解

# KEGG PATHWAY代谢途径

**1. Metabolism** —一级

**1.0 Global and overview maps** 二级

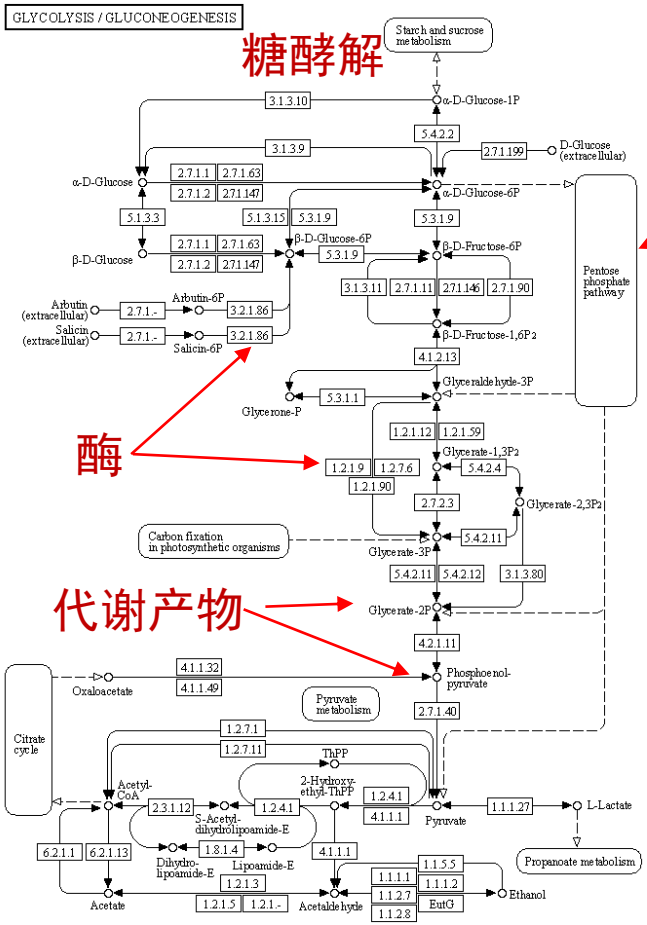
01100 Metabolic pathways
01110 Biosynthesis of secondary metabolites *Major update!*
01120 Microbial metabolism in diverse environments
01130 Biosynthesis of antibiotics
01200 Carbon metabolism
01210 2-Oxocarboxylic acid metabolism
01212 Fatty acid metabolism
01230 Biosynthesis of amino acids
01220 Degradation of aromatic compounds

**1.1 Carbohydrate metabolism**

00010 Glycolysis / Gluconeogenesis
00020 Citrate cycle (TCA cycle)
00030 Pentose phosphate pathway
00040 Pentose and glucuronate interconversions
00051 Fructose and mannose metabolism
00052 Galactose metabolism
00053 Ascorbate and aldarate metabolism
00500 Starch and sucrose metabolism
00520 Amino sugar and nucleotide sugar metabolism
00620 Pyruvate metabolism
00630 Glyoxylate and dicarboxylate metabolism
00640 Propanoate metabolism
00650 Butanoate metabolism
00660 C5-Branched dibasic acid metabolism
00562 Inositol phosphate metabolism

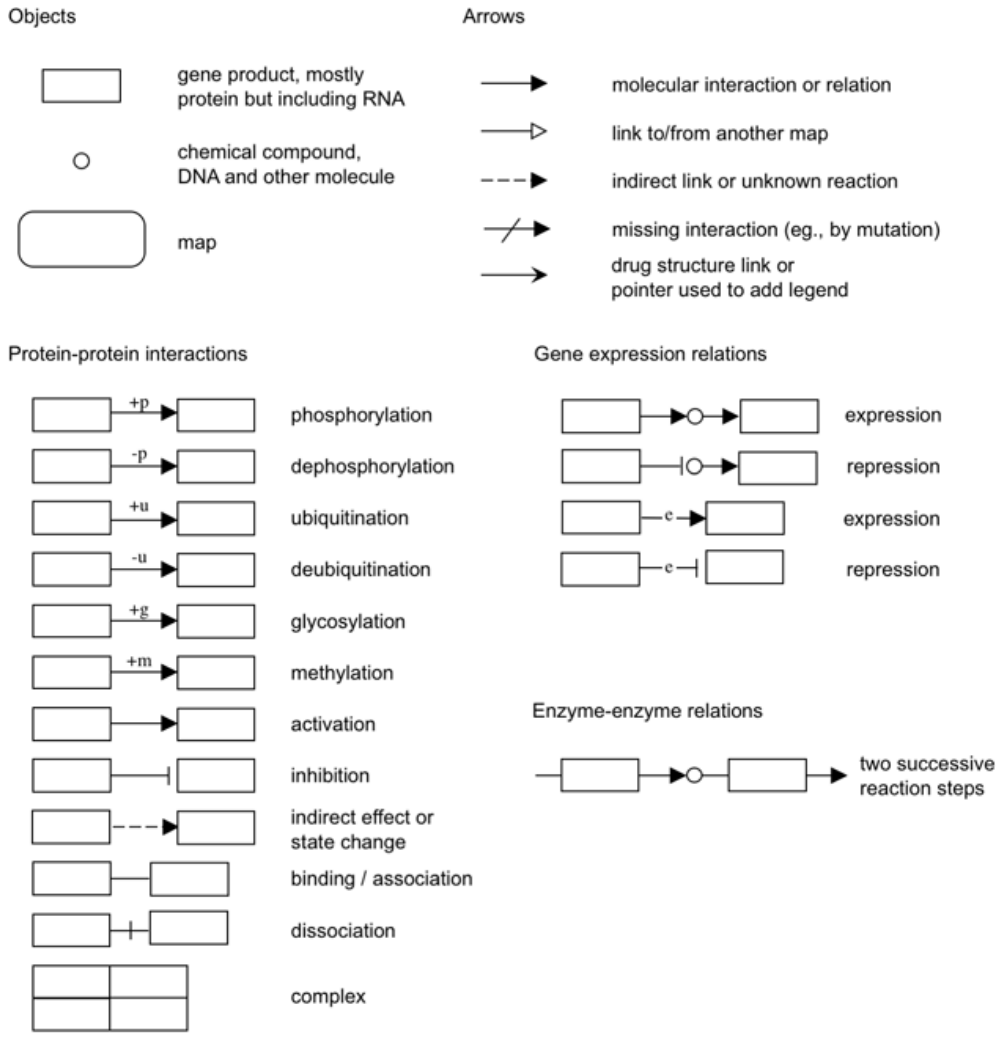
**1.2 Energy metabolism**

00190 Oxidative phosphorylation
00195 Photosynthesis
00196 Photosynthesis - antenna proteins
00710 Carbon fixation in photosynthetic organisms
00720 Carbon fixation pathways in prokaryotes
00680 Methane metabolism
00910 Nitrogen metabolism
00920 Sulfur metabolism




https://www.kegg.jp/kegg/pathway.html

# Pathway图中符号的含义



# 基因注释和功能富集

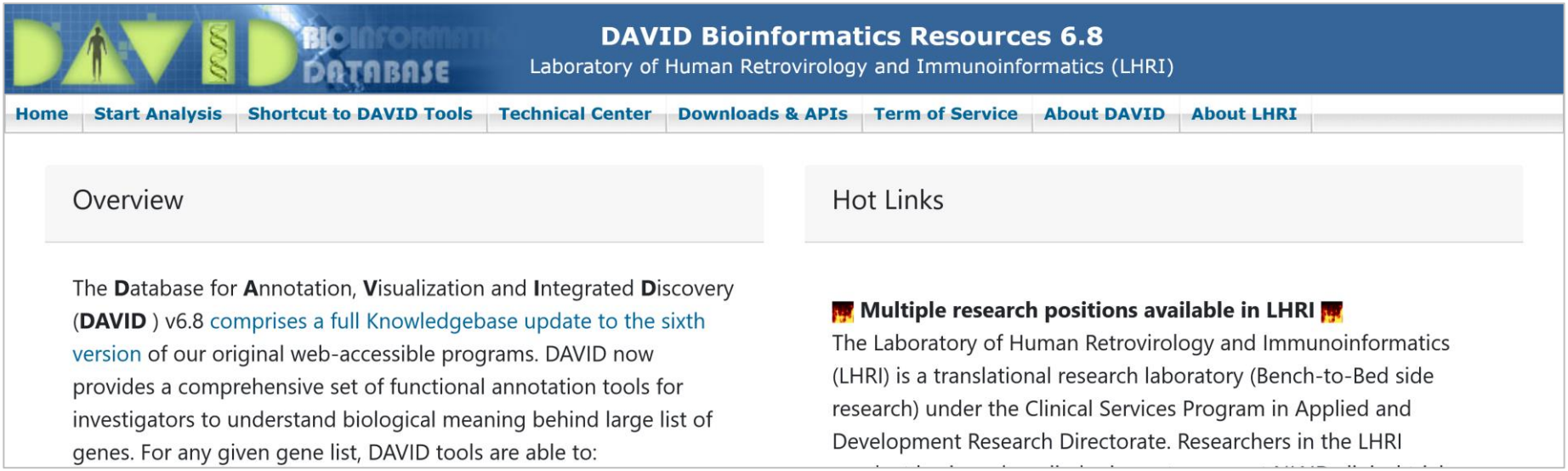
# GO/KEGG富集分析

由于GO/KEGG中**每种功能类别/代谢途径中的背景基因数量不同**，因此基因的绝对数量不能用来衡量基因在某种类别中的富集程度，需要通过统计算法来进行富集分析。

问题：在哪个类别中更富集?

| | 某物种所有基因 | 候选基因集 |
|---|---|---|
| GO类别1 | 1000 | 50 |
| GO类别2 | 10 | 5 |

# DAVID Bioinformatics Resources




Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources

BT Sherman, RA Lempicki - Nature protocols, 2009 - nature.com

DAVID bioinformatics resources consists of an integrated biological knowledgebase and analytic tools aimed at systematically extracting biological meaning from large gene/protein lists. This protocol explains how to use DAVID, a high-throughput and mining environment, to analyze gene lists derived from high-throughp experiments. The procedure first requires uploading a gene list conta common gene identifiers followed by analysis using one or more text

★ Cited by 26197 Related articles All 12 versions

Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists

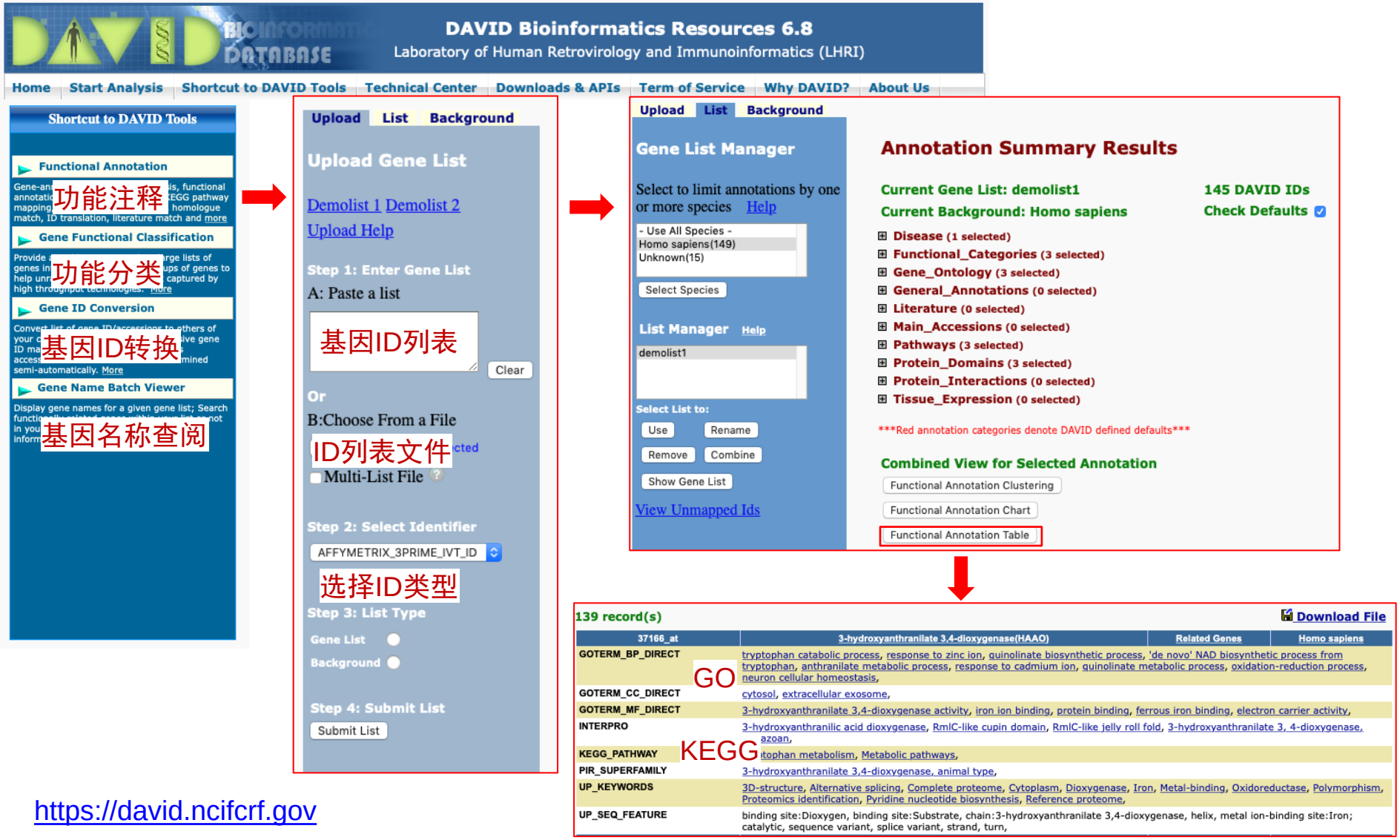
DW Huang, BT Sherman, RA Lempicki - Nucleic acids research, 2009 - academic.oup.com

Functional analysis of large gene lists, derived in most cases from emerging high-throughput genomic, proteomic and bioinformatics scanning approaches, is still a challenging and daunting task. The gene-annotation enrichment analysis is a promising high-throughput strategy that increases the likelihood for investigators to identify biological processes most pertinent to their study. Approximately 68 bioinformatics enrichment tools that are currently available in the community are collected in this survey. Tools are uniquely categorized into …

☆ Cited by 10862 Related articles All 27 versions

https://david.ncifcrf.gov

# DAVID Bioinformatics Resources




https://david.ncifcrf.gov

# DAVID Bioinformatics Resources

## Functional Annotation Clustering

Help and Manual

Current Gene List: demolist1
Current Background: Homo sapiens
145 DAVID IDs

Options  Classification Stringency: Medium

Rerun using options  Create Sublist

基因数量 富集程度

33 Cluster(s)

功能类别

Download File

| | Annotation Cluster 1 | Enrichment Score: 5.37 | G | | Count | P_Value | Benjamini |
|---|---|---|---|---|---|---|---|
| ☐ | GOTERM_CC_DIRECT | extracellular space | RT | | 32 | 7.1E-9 | 1.2E-6 |
| ☐ | GOTERM_CC_DIRECT | extracellular region | RT | | 32 | 4.1E-7 | 3.5E-5 |
| ☐ | UP_KEYWORDS | Disulfide bond | RT | | 47 | 1.3E-6 | 3.5E-4 |
| ☐ | UP_SEQ_FEATURE | signal peptide | RT | | 46 | 1.6E-6 | 1.0E-3 |
| ☐ | UP_KEYWORDS | Secreted | RT | | 33 | 1.7E-6 | 2.3E-4 |
| ☐ | UP_SEQ_FEATURE | disulfide bond | RT | | 42 | 1.9E-6 | 5.9E-4 |
| ☐ | UP_KEYWORDS | Signal | RT | | 49 | 4.5E-5 | 4.1E-3 |
| ☐ | UP_KEYWORDS | Glycoprotein | RT | | 50 | 2.2E-4 | 1.5E-2 |
| ☐ | UP_SEQ_FEATURE | glycosylation site:N-linked (GlcNAc...) | RT | | 44 | 2.3E-3 | 3.7E-1 |

## Functional Annotation Chart

Help and Manual

Current Gene List: demolist1
Current Background: Homo sapiens
145 DAVID IDs

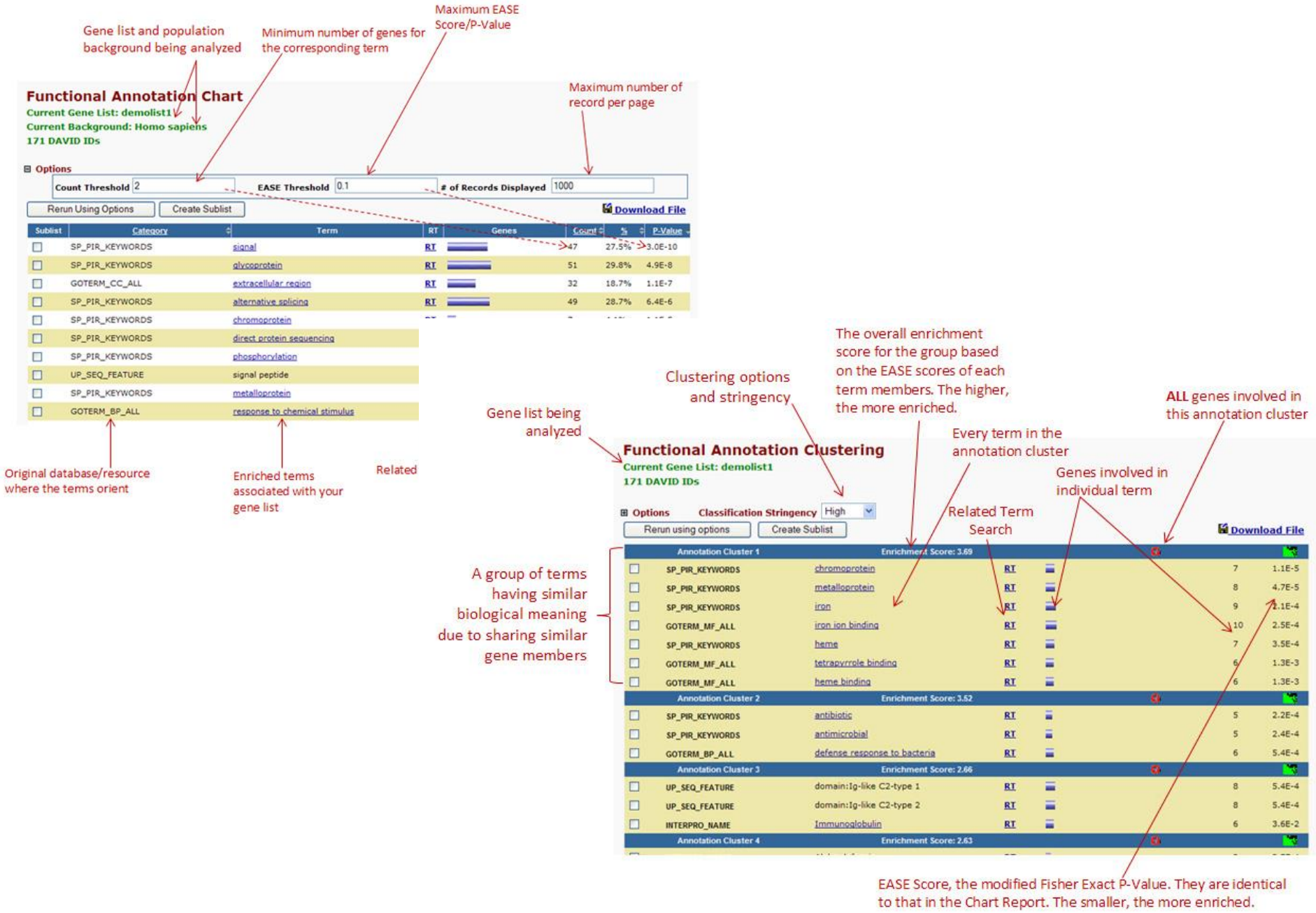
Options

Rerun Using Options  Create Sublist

220 chart records

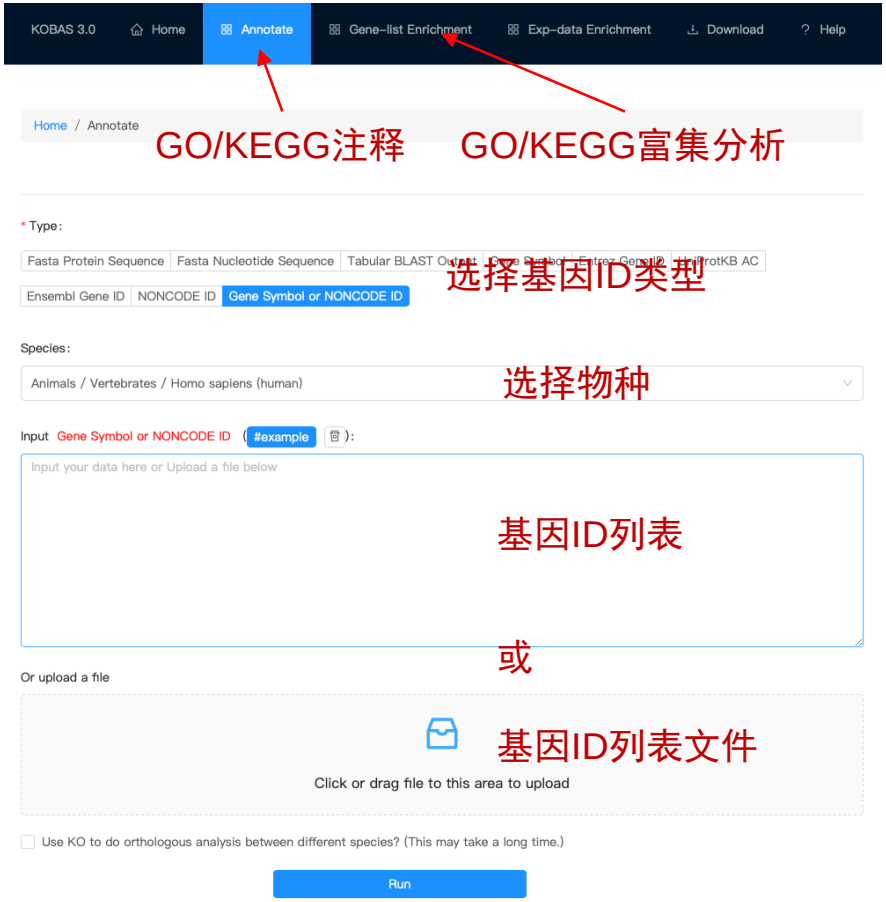
Download File

| Sublist | Category | Term | RT | Genes | Count | % | P-Value | Benjamini |
|---|---|---|---|---|---|---|---|---|
| ☐ | GOTERM_CC_DIRECT | extracellular space | RT | | 32 | 22.1 | 7.1E-9 | 1.2E-6 |
| ☐ | GOTERM_CC_DIRECT | extracellular region | RT | | 32 | 22.1 | 4.1E-7 | 3.5E-5 |
| ☐ | UP_KEYWORDS | Disulfide bond | RT | | 47 | 32.4 | 1.3E-6 | 3.5E-4 |
| ☐ | UP_SEQ_FEATURE | signal peptide | RT | | 46 | 31.7 | 1.6E-6 | 1.0E-3 |
| ☐ | UP_KEYWORDS | Secreted | RT | | 33 | 22.8 | 1.7E-6 | 2.3E-4 |
| ☐ | UP_SEQ_FEATURE | disulfide bond | RT | | 42 | 29.0 | 1.9E-6 | 5.9E-4 |
| ☐ | UP_KEYWORDS | Signal | RT | | 49 | 33.8 | 4.5E-5 | 4.1E-3 |
| ☐ | GOTERM_BP_DIRECT | defense response to fungus | RT | | 5 | 3.4 | 5.6E-5 | 5.6E-2 |


Gene list and population background being analyzed
Minimum number of genes for the corresponding term
Maximum EASE Score/P-Value
Maximum number of record per page
Functional Annotation Chart
Current Gene List: demolist1
Current Background: Homo sapiens
171 DAVID IDs
Options
Count Threshold 2
EASE Threshold 0.1
# of Records Displayed 1000
Rerun Using Options
Create Sublist
Download File
Sublist
Category
Term
RT
Genes
Count
%
P-Value
SP_PIR_KEYWORDS signal RT 47 27.5% 3.0E-10
SP_PIR_KEYWORDS glycoprotein RT 51 29.8% 4.9E-8
GOTERM_CC_ALL extracellular region RT 32 18.7% 1.1E-7
SP_PIR_KEYWORDS alternative splicing RT 49 28.7% 6.4E-6
SP_PIR_KEYWORDS chromoprotein
SP_PIR_KEYWORDS direct protein sequencing
SP_PIR_KEYWORDS phosphorylation
UP_SEQ_FEATURE signal peptide
SP_PIR_KEYWORDS metalloprotein
GOTERM_BP_ALL response to chemical stimulus
Original database/resource where the terms orient
Enriched terms associated with your gene list
Related
Gene list being analyzed
Clustering options and stringency
The overall enrichment score for the group based on the EASE scores of each term members. The higher, the more enriched.
ALL genes involved in this annotation cluster
Every term in the annotation cluster
Genes involved in individual term
Related Term Search
Functional Annotation Clustering
Current Gene List: demolist1
171 DAVID IDs
Options
Classification Stringency High
Rerun using options
Create Sublist
Download File
Annotation Cluster 1 Enrichment Score: 3.69
SP_PIR_KEYWORDS chromoprotein RT 7 1.1E-5
SP_PIR_KEYWORDS metalloprotein RT 8 4.7E-5
SP_PIR_KEYWORDS iron RT 9 1.1E-4
GOTERM_MF_ALL iron ion binding RT 10 2.5E-4
SP_PIR_KEYWORDS heme RT 7 3.5E-4
GOTERM_MF_ALL tetrapyrrole binding RT 6 1.3E-3
GOTERM_MF_ALL heme binding RT 6 1.3E-3
Annotation Cluster 2 Enrichment Score: 3.52
SP_PIR_KEYWORDS antibiotic RT 5 2.2E-4
SP_PIR_KEYWORDS antimicrobial RT 5 2.4E-4
GOTERM_BP_ALL defense response to bacteria RT 6 5.4E-4
Annotation Cluster 3 Enrichment Score: 2.66
UP_SEQ_FEATURE domain:Ig-like C2-type 1 RT 8 5.4E-4
UP_SEQ_FEATURE domain:Ig-like C2-type 2 RT 8 5.4E-4
INTERPRO_NAME Immunoglobulin RT 6 3.6E-2
Annotation Cluster 4 Enrichment Score: 2.63
A group of terms having similar biological meaning due to sharing similar gene members
EASE Score, the modified Fisher Exact P-Value. They are identical to that in the Chart Report. The smaller, the more enriched.

# KOBAS

## KOBAS网站使用方法




## 功能注释结果

| | | | |
|---|---|---|---|
| Query: | ASAH1 | | |
| Gene: | hsa:427 | ASAH1, AC, ACDase, ASAH, PHP, PHP32, SMAPME | |
| Entrez Gen | 427 | | |
| Pathway: | Metabolic pathways | KEGG PATHWAY | hsa01100 |
| | Sphingolipid metabolism | KEGG PATHWAY | hsa00600 |
| | Sphingolipid signaling pathway | KEGG PATHWAY | hsa04071 |
| | phospholipids as signalling intermediaries | BioCarta | 100183 |
| | Lysosome | KEGG PATHWAY | hsa04142 |
| | sphingosine and sphingosine-1-phosphate metabo | BioCyc | PWY3DJ-11470 |
| | Ceramide signaling pathway | PID | ceramide_pathwa |
| | Immune System | Reactome | R-HSA-168256 |
| | Metabolism of lipids | Reactome | R-HSA-556833 |
| | Metabolism | Reactome | R-HSA-1430728 |
| | Sphingolipid metabolism | Reactome | R-HSA-428157 |
| | Glycosphingolipid metabolism | Reactome | R-HSA-1660662 |
| | Innate Immune System | Reactome | R-HSA-168249 |
| | Neutrophil degranulation | Reactome | R-HSA-6798695 |
| GO: | extracellular region | Gene Ontology | GO:0005576 |
| | nitrogen compound metabolic process | Gene Ontology | GO:0006807 |
| | hydrolase activity | Gene Ontology | GO:0016787 |
| | lysosomal lumen | Gene Ontology | GO:0043202 |
| | cellular metabolic process | Gene Ontology | GO:0044237 |
| | extracellular exosome | Gene Ontology | GO:0070062 |

http://kobas.cbi.pku.edu.cn/kobas3

# KOBAS富集分析

功能类别　数据库　输入基因数量 背景基因数量　富集程度

| #Term | Database | ID | Input number | Background number | P-Value | Corrected P-Value | Input |
|---|---|---|---|---|---|---|---|
| Human cyto... | KEGG PATHW... | hsa05163 | 34 | 225 | 7.53693461... | 4.87134486... | PTK2\|IL10R... |
| Pathways i... | KEGG PATHW... | hsa05200 | 46 | 530 | 8.21562461... | 5.10128552... | BCL2L1\|PTK... |
| Circadian ... | KEGG PATHW... | hsa04713 | 23 | 97 | 4.40278876... | 2.54272886... | GRIA2\|GNAO... |
| Dopaminerg... | KEGG PATHW... | hsa04728 | 25 | 131 | 8.58136119... | 4.90324371... | SLC6A3\|CLO... |
| MAPK signa... | KEGG PATHW... | hsa04010 | 33 | 295 | 1.74062764... | 9.84095901... | JUN\|HSPA8\|... |
| Retrograde... | KEGG PATHW... | hsa04723 | 24 | 148 | 1.44598506... | 7.54750803... | GRIA2\|GNAO... |

Human cytomegalovirus infection

# 基因组注释实例

# 基于RNA-Seq数据的水稻基因组重注释

- 1503套高质量RNA-seq数据集，文件大小5.32 TB
- 含可变剪接的基因所占的比例
- 每个基因位点所对应的平均转录本数量

**Table 1 Statistics of three different annotation systems for rice genome (Overview)**

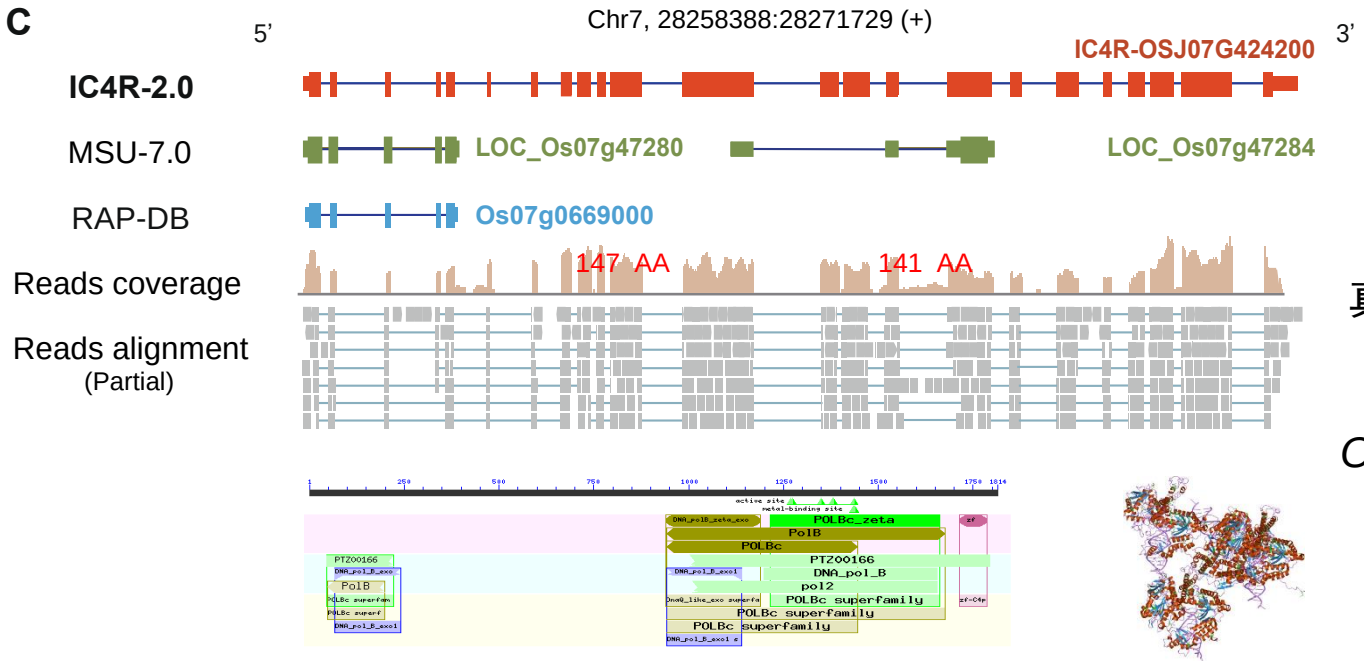
| Data Items | IC4R-2.0 | RAP-DB[a] | MSU-7.0[b] |
|---|---|---|---|
| **Protein-coding genes** | | | |
| Number of gene loci | 56,221 | 45,966 | 55,986 |
| Number of mRNAs | 80,039 | 52,733 | 66,338 |
| Percentage of spliced genes | 16.36% | 11.56% | 11.54% |
| Average spliced isoforms per gene | 1.42 | 1.15 | 1.18 |

*Note*: Data obtained from [a] MSU (http://rice.plantbiology.msu.edu) and [b] RAP-DB (https://rapdb.dna.affrc.go.jp/) on April 12, 2018.

**IC4R-2.0: Rice Genome Reannotation Using Massive RNA-seq Data**. *Genomics Proteomics Bioinformatics* 2020

# 水稻基因组重注释：基因合并

IC4R-2.0中218个独立的基因，在原注释系统中被分割注释成了不同的基因




1814 AA

真核生物DNA聚合酶

Poly zeta

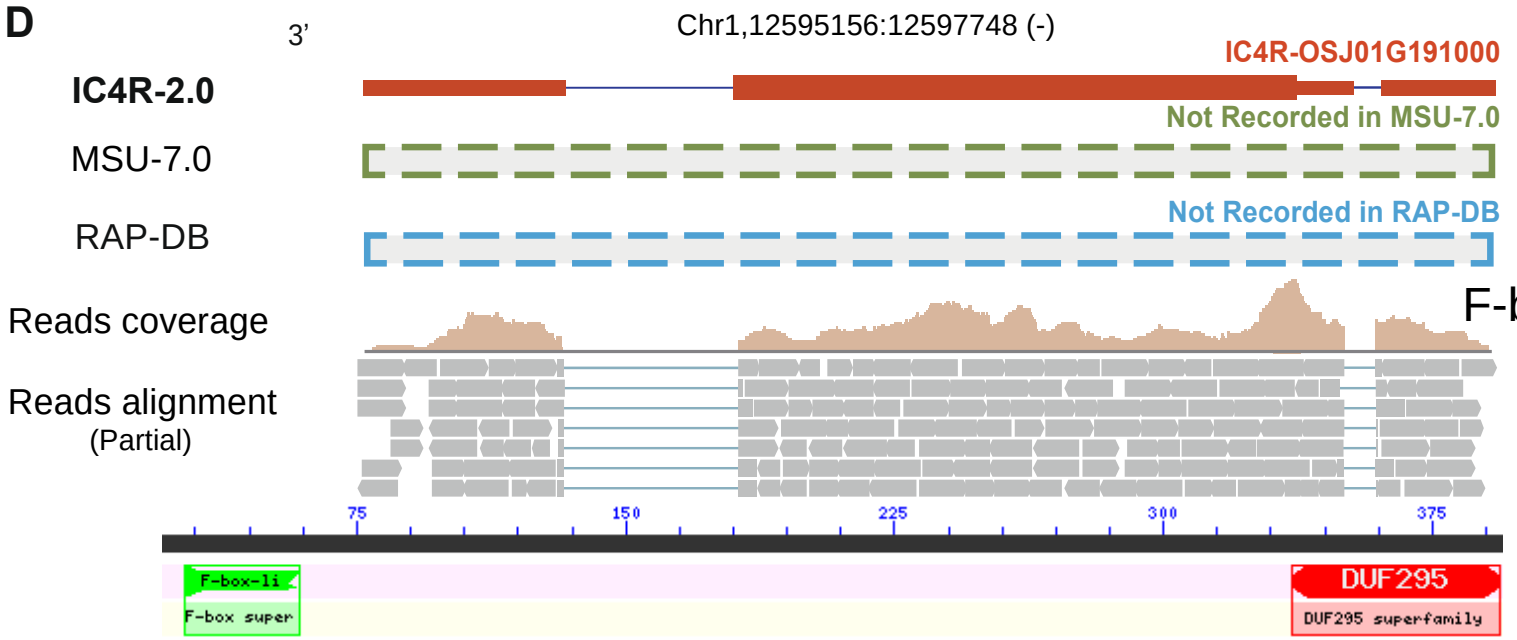
*OsPolζ* / *OsPoly-zeta*

**Lawrence et al, 2008**

- Polymerases α, β, λ, σ, μ, δ, ε, η, ι, **ζ**, κ and Rev1; 保守性高
- *Polζ*：在真核生物DNA自发性损伤修复、DNA复制过程中发挥着重要作用

# 水稻基因组重注释：新基因

IC4R-2.0注释系统中存在456个在MSU-7.0与RAP-DB未报道过的基因




F-box & Duf 蛋白

*Osfbduf3*

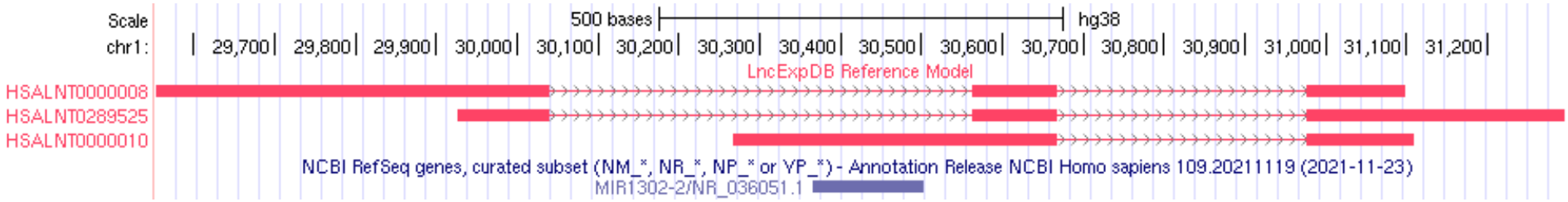
- *F-Box* 在植物中是一类重要的转录因子基因家族 (*LRRs, Kelch, DUF and PP2*)
- *Fbduf3* 基因与在其他植物中，与干旱、寒冷等非生物胁迫应答密切相关 (Song et al, 2015)

# 人类长非编码RNA注释集

| 注释集 | 基因数目 | 发表单位 |
| --- | --- | --- |
| LncBook | 95,243 | 北京基因组研究所 |
| NONCODE | 94,411 | 生物物理研究所 |
| RefLnc | 59,489 | 北京大学 |
| MiTranscriptome | 58,648 | Michigan Center for Translational Pathology |
| CHESS Comprehensive Human Expressed SequenceS | 18,887 | Johns Hopkins University |
| GENCODE | 18,811 | National Human Genome Research Institute |

# LncBook鉴定新型长非编码RNA基因

***MIR1302-2HG*基因在LncBook与GENCODE中的注释情况**

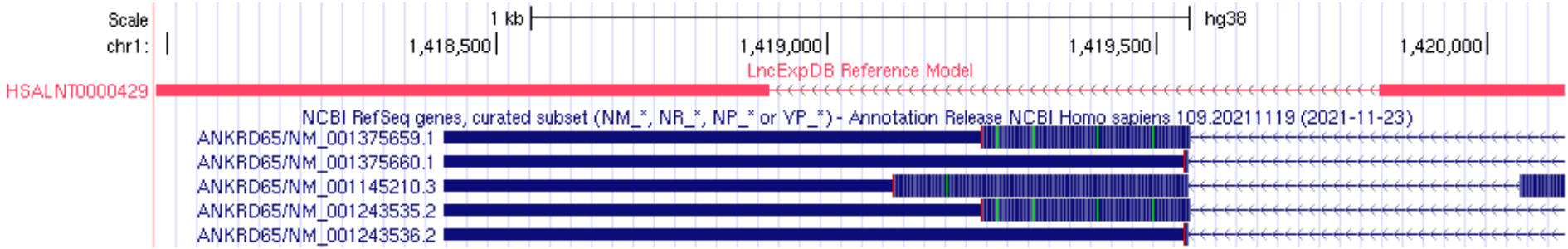



GENCODE中只注释出一条miRNA，没有鉴定出该基因

**LncBook**

**GENCODE**

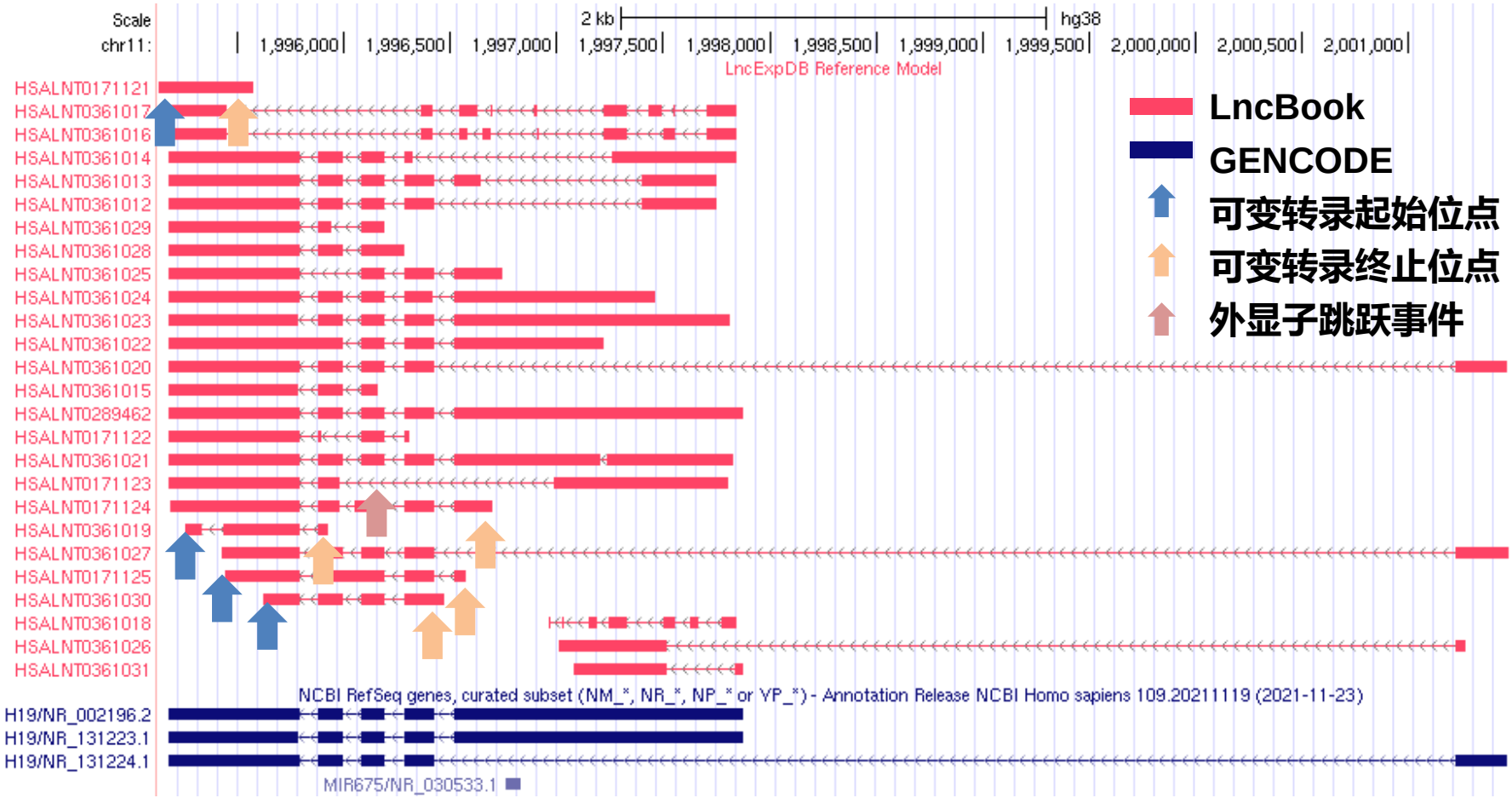
**HSALNG0000159在LncBook与GENCODE中的注释情况**




GENCODE中只注释出一个蛋白编码基因，没有鉴定出该长非编码RNA基因

# 长非编码RNA基因：*H19*

***H19*: a ~2.3-kb imprinted maternally expressed transcript, is closely associated with Beckwith-Wiedemann Syndrome and also involved in tumorigenesis.**



# LncBook包含丰富的多组学注释信息

## 表达谱数据

- HPA中32种正常组织长非编码RNA表达谱数据
- GTEx中53种正常组织长非编码RNA表达谱数据


## 甲基化数据

- 9种癌症启动子区与基因body区甲基化注释信息

## 变异数据

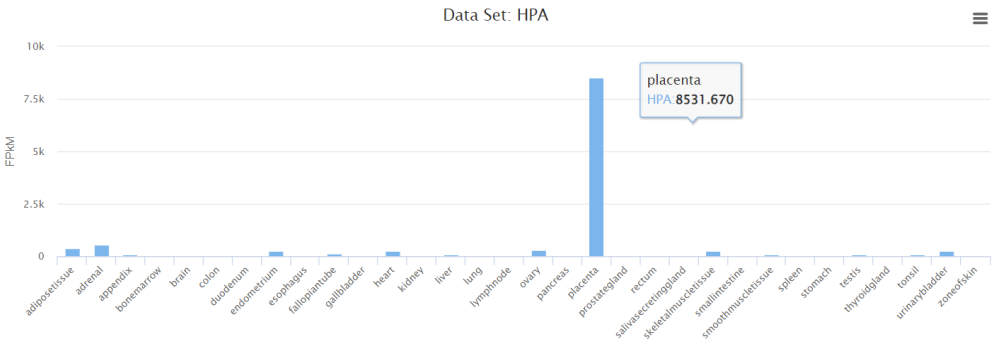
- COSMIC, Clinvar, 1000 Genomes Project变异数据注释

## 互作数据

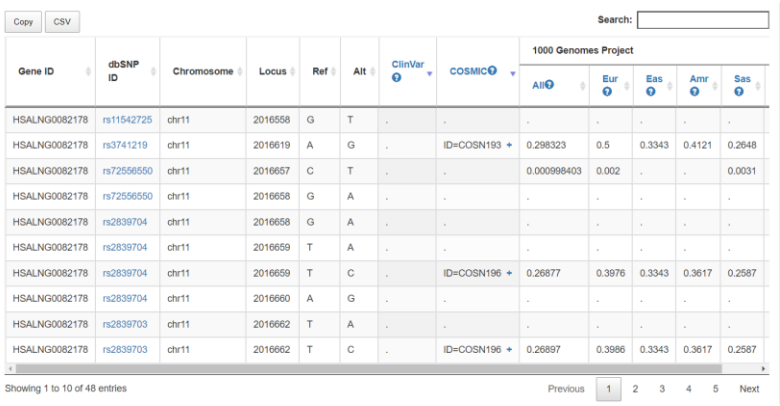
- 基于TargetScan与miRanda预测lncRNA-miRNA互作关系对
- 基于starBase搜集有实验证据支持的lncRNA-miRNA关系对

# LncBook数据库多组学图谱展示

## 表达




## 变异

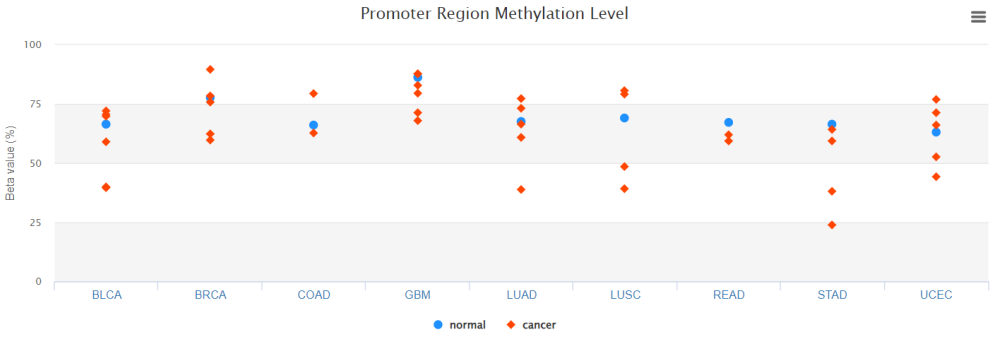


Copy CSV Search:

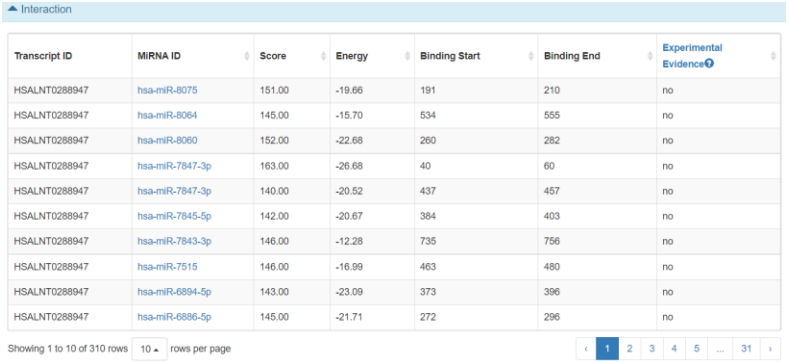
| Gene ID | dbSNP ID | Chromosome | Locus | Ref | Alt | ClinVar | COSMIC | 1000 Genomes Project | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | All | Eur | Eas | Amr | Sas |
| HSALNG0082178 | rs11542725 | chr11 | 2016558 | G | T | . | . | . | . | . | . | . |
| HSALNG0082178 | rs3741219 | chr11 | 2016619 | A | G | . | ID=COSN193 + | 0.298323 | 0.5 | 0.3343 | 0.4121 | 0.2648 |
| HSALNG0082178 | rs72556550 | chr11 | 2016657 | C | T | . | . | 0.000998403 | 0.002 | . | . | 0.0031 |
| HSALNG0082178 | rs72556550 | chr11 | 2016658 | G | A | . | . | . | . | . | . | . |
| HSALNG0082178 | rs2839704 | chr11 | 2016658 | G | A | . | . | . | . | . | . | . |
| HSALNG0082178 | rs2839704 | chr11 | 2016659 | T | A | . | . | . | . | . | . | . |
| HSALNG0082178 | rs2839704 | chr11 | 2016659 | T | C | . | ID=COSN196 + | 0.26877 | 0.3976 | 0.3343 | 0.3617 | 0.2587 |
| HSALNG0082178 | rs2839704 | chr11 | 2016660 | A | G | . | . | . | . | . | . | . |
| HSALNG0082178 | rs2839703 | chr11 | 2016662 | T | A | . | . | . | . | . | . | . |
| HSALNG0082178 | rs2839703 | chr11 | 2016662 | T | C | . | ID=COSN196 + | 0.26897 | 0.3986 | 0.3343 | 0.3617 | 0.2587 |

Showing 1 to 10 of 48 entries Previous 1 2 3 4 5 Next

## 甲基化




## 互作



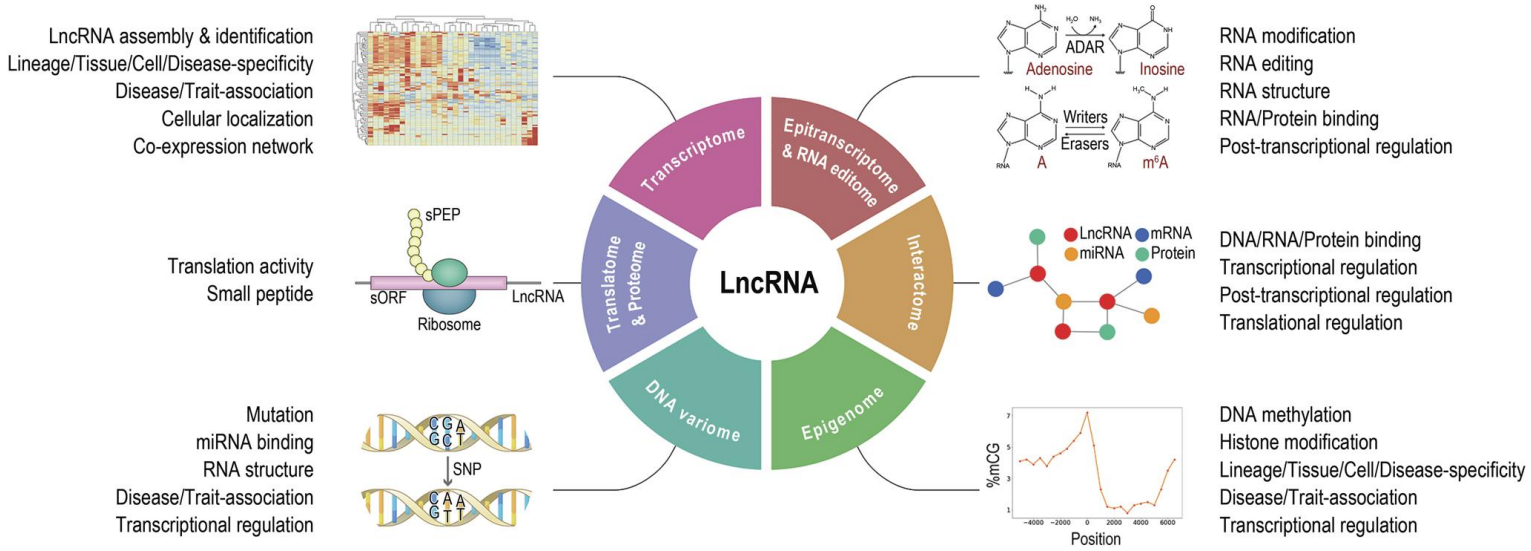
Interaction

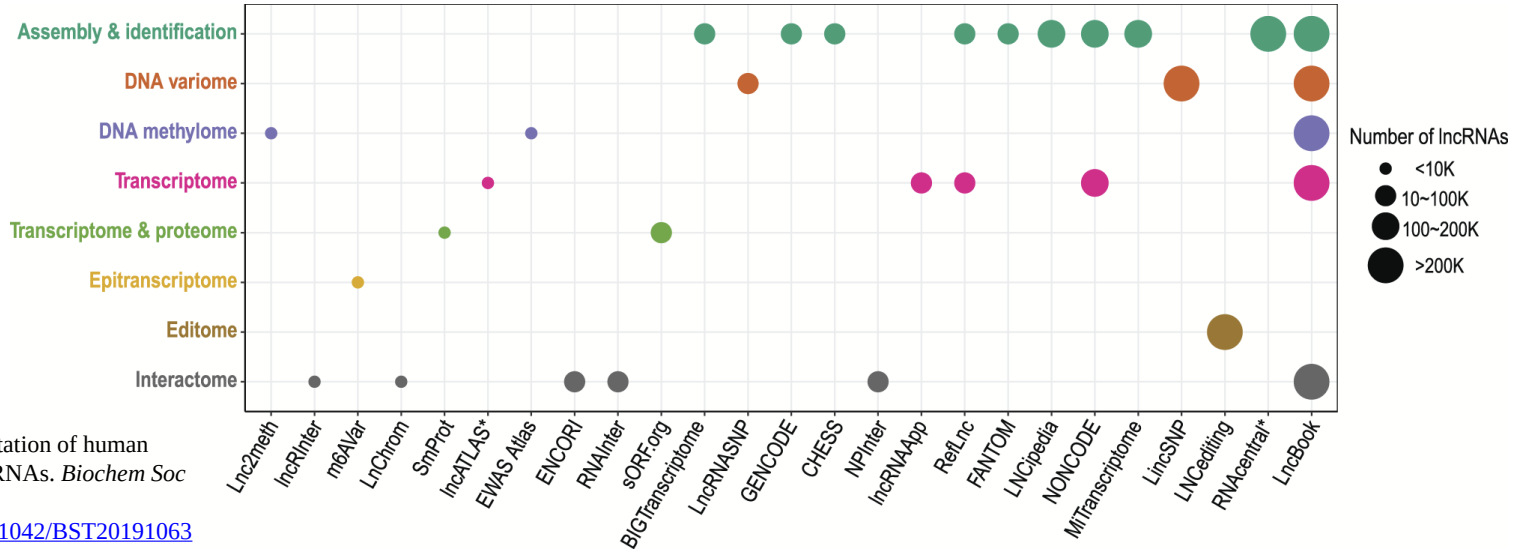
| Transcript ID | MiRNA ID | Score | Energy | Binding Start | Binding End | Experimental Evidence |
|---|---|---|---|---|---|---|
| HSALNT0288947 | hsa-miR-8075 | 151.00 | -19.66 | 191 | 210 | no |
| HSALNT0288947 | hsa-miR-8064 | 145.00 | -15.70 | 534 | 555 | no |
| HSALNT0288947 | hsa-miR-8060 | 152.00 | -22.68 | 260 | 282 | no |
| HSALNT0288947 | hsa-miR-7847-3p | 163.00 | -26.68 | 40 | 60 | no |
| HSALNT0288947 | hsa-miR-7847-3p | 140.00 | -20.52 | 437 | 457 | no |
| HSALNT0288947 | hsa-miR-7845-5p | 142.00 | -20.67 | 384 | 403 | no |
| HSALNT0288947 | hsa-miR-7843-3p | 146.00 | -12.28 | 735 | 756 | no |
| HSALNT0288947 | hsa-miR-7515 | 146.00 | -16.99 | 463 | 480 | no |
| HSALNT0288947 | hsa-miR-6894-5p | 143.00 | -23.09 | 373 | 396 | no |
| HSALNT0288947 | hsa-miR-6886-5p | 145.00 | -21.71 | 272 | 296 | no |

Showing 1 to 10 of 310 rows 10 rows per page ‹ 1 2 3 4 5 ... 31 ›

https://ngdc.cncb.ac.cn/lncbook/

# 长非编码RNA多组学注释




Multi-omics annotation of human long non-coding RNAs. *Biochem Soc Trans* (2020) https://doi.org/10.1042/BST20191063

# 基因组注释的挑战

Opinion | Published: 01 February 2007

## Genome re-annotation: a wiki solution?

Steven L Salzberg

*Genome Biology* **8**, Article number: 102 (2007) | Cite this article

**22k** Accesses | **91** Citations | **23** Altmetric | Metrics

信息分散
人人参与

Genome Biol. 2019; 20: 92.

Published online 2019 May 16. doi: 10.1186/s13059-019-1715-2

## Next-generation genome annotation: we still struggle to get it right

Steven L. Salzberg

数据暴增
自动智能

Method | Open Access | Published: 28 March 2023

## HiCAT: a tool for automatic annotation of centromere structure

Shenghan Gao, Xiaofei Yang, Hongtao Guo, Xixi Zhao, Bo Wang & Kai Ye

*Genome Biology* **24**, Article number: 58 (2023) | Cite this article

结构复杂
工具研发

# Summary

- 基因组注释：结构注释、功能注释
- 审编（curation）：value-added annotation
- 基因组注释：原核生物、真核生物
- 同源性、一致性、相似性
- 结构预测：软件预测、同源预测、转录本预测
- 功能预测：序列比对、蛋白结构域、基因本体GO、代谢KEGG、功能富集分析、DAVID
- 实例：基于RNA-seq的水稻基因组重注释、人类长非编码RNA注释

# Further Reading

- **Genome annotation: from sequence to biology**. *Nature Reviews Genetics* 2001
- **Towards multidimensional genome annotation.** *Nature Reviews Genetics* 2006
- **A beginner's guide to eukaryotic genome annotation**. *Nature Reviews Genetics* 2012
- **NCBI prokaryotic genome annotation pipeline.** *Nucleic Acids Res*. 2016
- **Genome annotation for clinical genomic diagnostics: strengths and weaknesses.** *Genome Medicine* 2017
- **Next-generation genome annotation: we still struggle to get it right.** *Genome Biology* 2019
- **Long-Read Annotation: Automated Eukaryotic Genome Annotation Based on Long-Read cDNA Sequencing**. *Plant Physiology* 2019