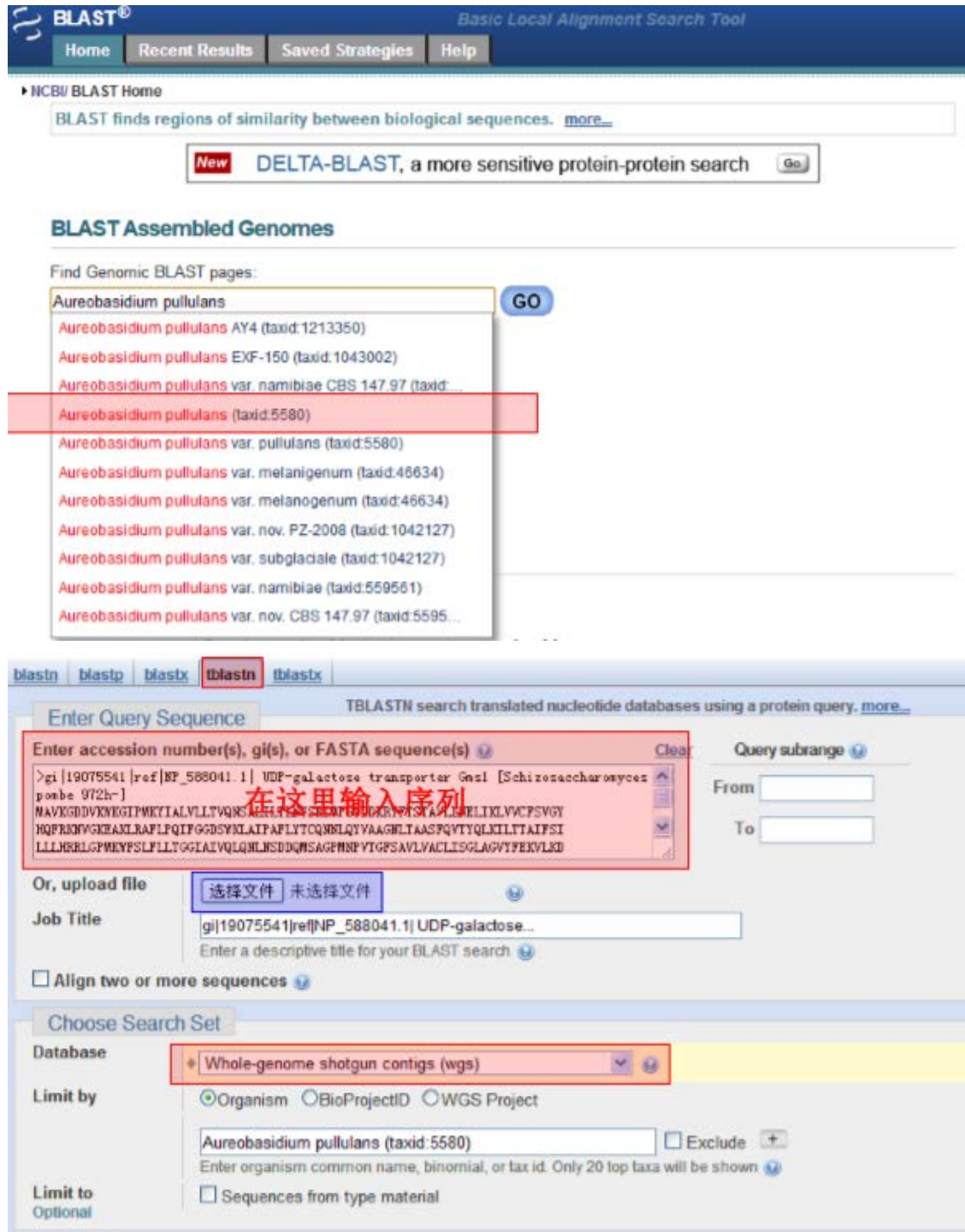


## 新基因发现与基因结构建模

### 1. 新基因范畴:

- 1.1. 该物种没有报道 => 同源基因搜索
- 1.2. 所有物种都没有报道 => EST/cDNA 序列文库、RNA-seq、从头计算鉴别新基因

### 2. AP 为例 blast: s



BLAST<sup>®</sup> Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI BLAST Home

BLAST finds regions of similarity between biological sequences. [more...](#)

New DELTA-BLAST, a more sensitive protein-protein search [Go](#)

### BLAST Assembled Genomes

Find Genomic BLAST pages:

Aureobasidium pullulans [GO](#)

- Aureobasidium pullulans AY4 (taxid:1213350)
- Aureobasidium pullulans EXF-150 (taxid:1043002)
- Aureobasidium pullulans var. namibiae CBS 147.97 (taxid:...
- Aureobasidium pullulans (taxid:5580)**
- Aureobasidium pullulans var. pullulans (taxid:5580)
- Aureobasidium pullulans var. melanigenum (taxid:46634)
- Aureobasidium pullulans var. melanigenum (taxid:46634)
- Aureobasidium pullulans var. nov. PZ-2008 (taxid:1042127)
- Aureobasidium pullulans var. subglaciale (taxid:1042127)
- Aureobasidium pullulans var. namibiae (taxid:559561)
- Aureobasidium pullulans var. nov. CBS 147.97 (taxid:5595...

blastn blastp blastx **tblastn** tblastx

Enter Query Sequence

TBLASTN search translated nucleotide databases using a protein query. [more...](#)

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)

Query subrange

From

To

Or, upload file [选择文件](#) [未选择文件](#)

Job Title

Enter a descriptive title for your BLAST search [+](#)

☐ Align two or more sequences [+](#)

### Choose Search Set

Database [+](#) Whole-genome shotgun contigs (wgs) [-](#)

Limit by ☒ Organism ☐ BioProjectID ☐ WGS Project

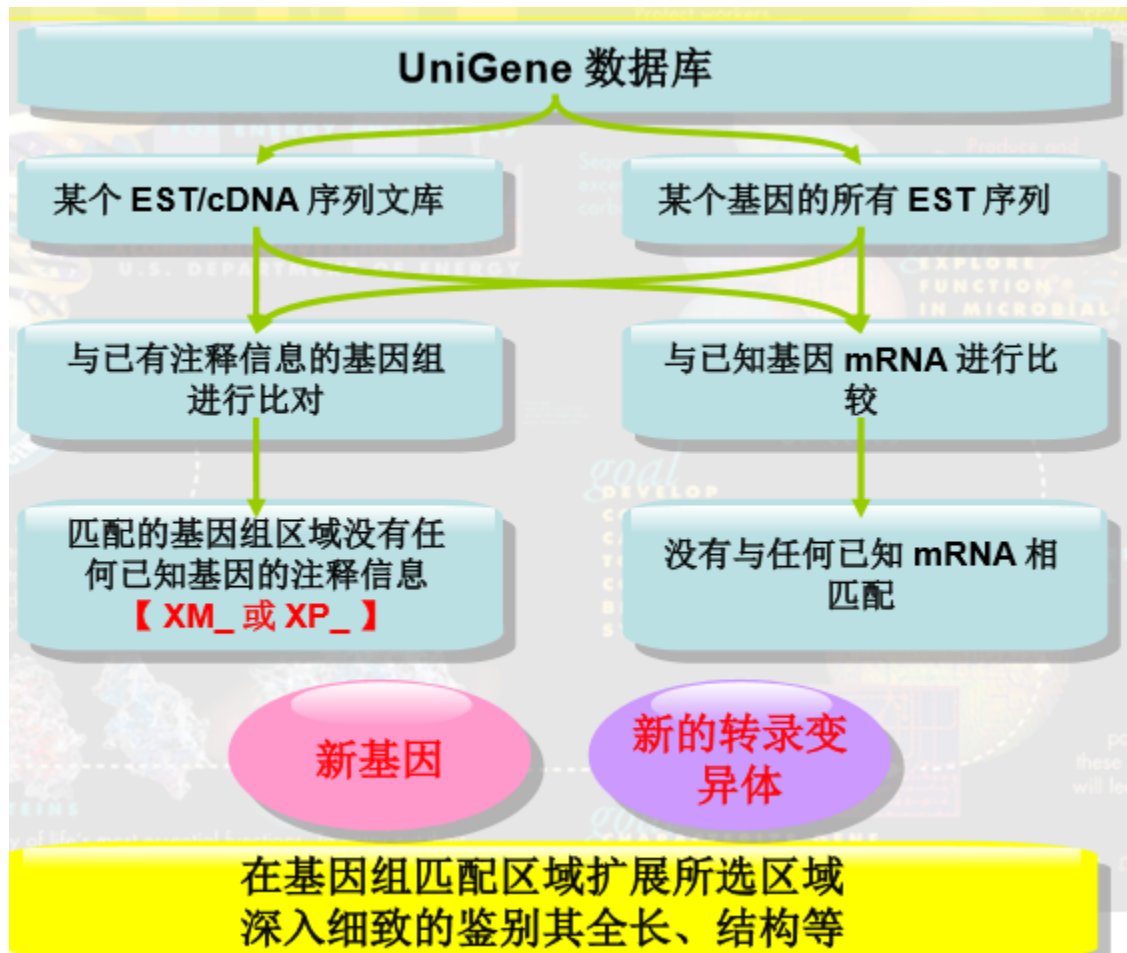
[Exclude](#) [+](#)

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown [+](#)

Limit to Optional ☐ Sequences from type material

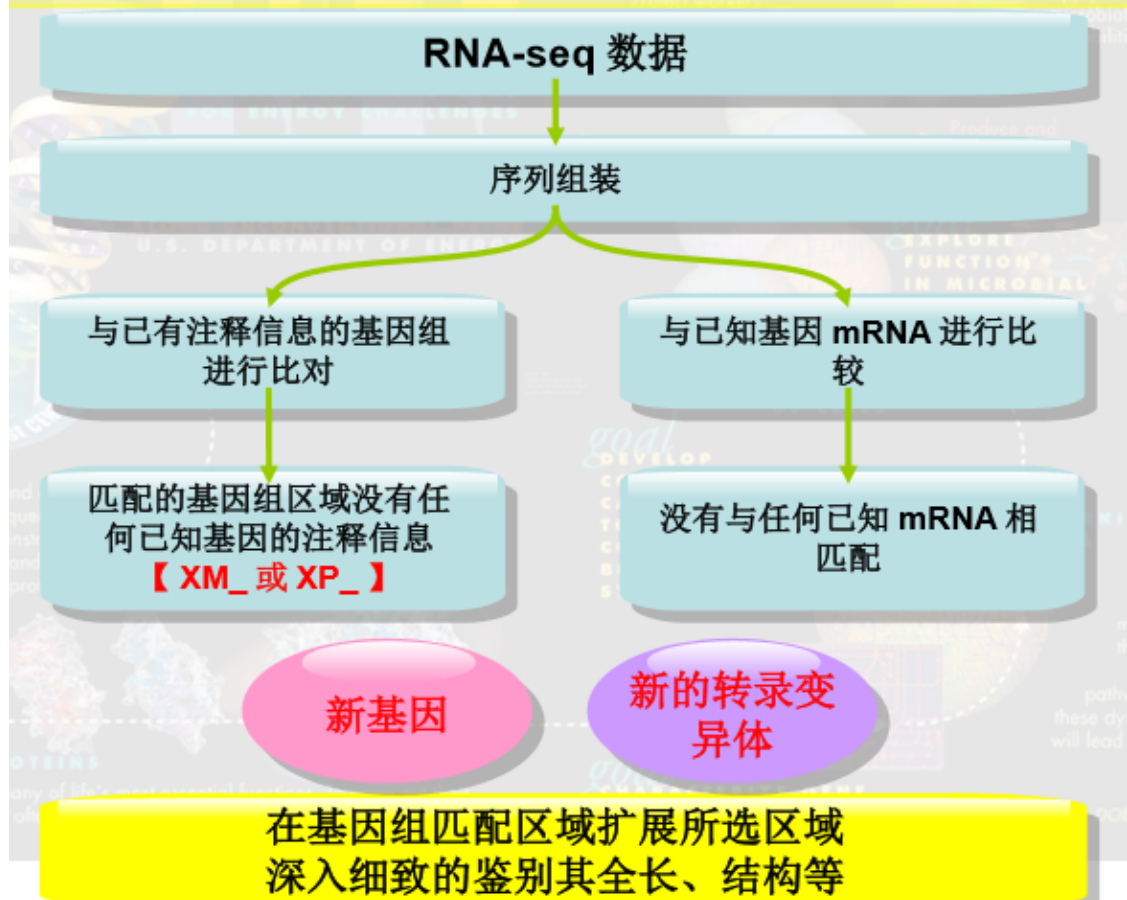
在结果中选择高相似区域，扩展上下游，各延伸 1kb

### 3. 基于 EST/cDNA 序列文库的新基因发现:



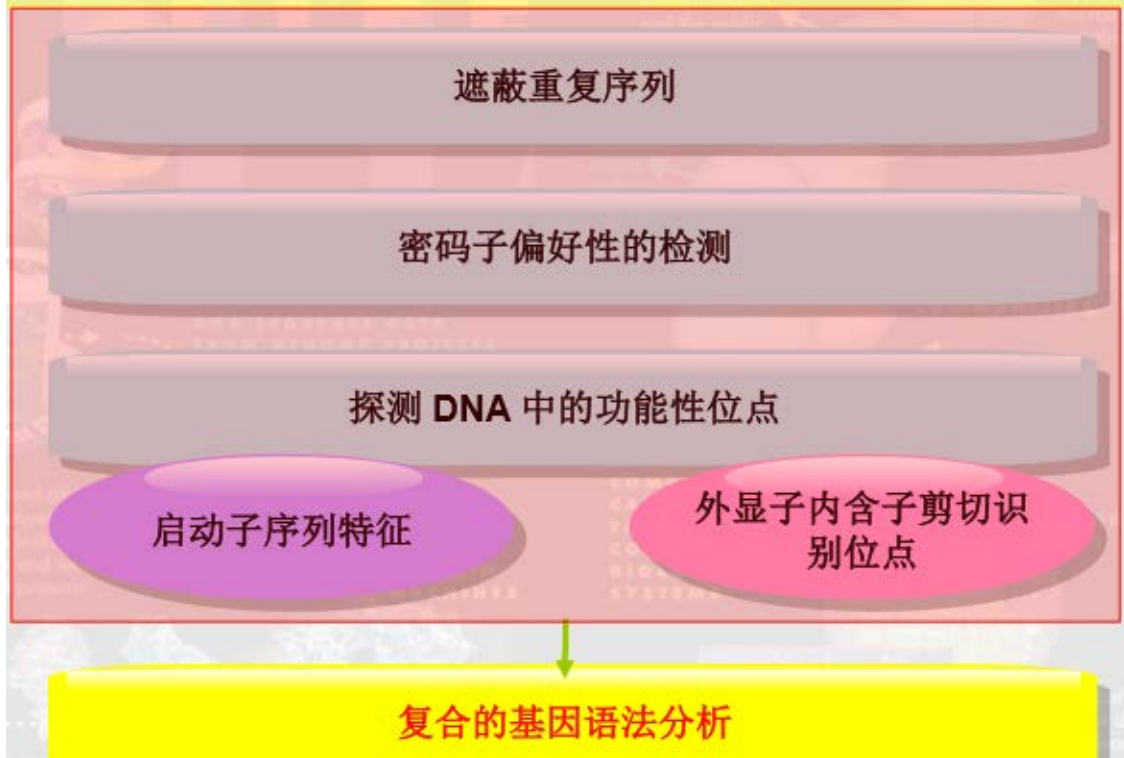
4. 基于 RNA-seq 数据的新基因发现：

## 基于 RNA-seq 数据的新基因发现



5. 从头计算鉴别新基因:

## 从头计算鉴别新基因



## 从头计算鉴别新基因 >> 相关软件

Software / WebServer	applicability	link
<b>RepeatMasker</b>	various familiar/model species	NHGRI
<b>GENSCAN</b>	Vertebrate/Arabidopsis/Maize	MIT
<b>tRNAscan-SE</b>	any	lowelab Eddy lab bioweb

6. 基因结构建模:
- NCBI Splign
  - UCSC BLAT
  - EMBL-EBI GeneWise

## NCBI >> Sequence Analysis >> Splign >> Results



## BLAT Search Results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
<a href="#">browser</a> <a href="#">details</a>	NP_035770.2	592	77	390	390	84.3%	17	+-	7572930	7579449	6520
<a href="#">browser</a> <a href="#">details</a>	NP_035770.2	99	86	255	390	77.8%	1	++	3639926	3643780	3855

一致残基的比例

相似序列得分



GeneWise

http://www.ebi.ac.uk/Tools/psa/genewise/

Input form

Web services

Help & Documentation

Share

Feedback

Tools > Pairwise Sequence Alignment > GeneWise

Pairwise Sequence Alignment

GeneWise compares a protein sequence to a genomic DNA sequence, allowing for introns and frameshifting errors.

STEP 1 - Enter your sequences

Enter or paste your protein sequence in any supported format:

```
>gi|19075541|ref|NP_588041.1| UDP-galactose transporter Gms1 [Schizosaccharomyces pombe 972h-]
MAVKGDDVVKWGIIPKRYIALVLLTQNSALILTLNYSRIMPGYDDKRYFTSTAVLLNELIKLVCFSWGY
HQRKINFGKEAKLRFLPQIFGGDSVKLAIPAFLYTCQNNLQYVAAGNLTAASFQYITQLKILITAFISI
LLHRRRLGPMKWFSLFLITGGIATVQLQNLNSDDQMSAGPHNPVITGSAVLVACLISGLAGVTFEKLVD
TNPVLWYKRVNQLSFFSLFPCLFIIIMKDYHNTAENGFFFGYNSIYWLAILLQAGGGIIVALCVAFADNIM
KNFSTISIISSLASLYLMDPKLSLITPLIGVMLVIAATPLYTKPKSKPSPSRGTYIPWTQDAADQVQ
HHH
```

Schizosaccharomyces pombe (SP) 的Gms1蛋白序列

Or, upload a file:

选择文件

未选择文件

AND

Enter or paste your DNA sequence in any supported format:

```
>gb|AMCU01000006.1|:212545-215551 Aureobasidium pullulans AY4 contig6, whole genome shotgun sequence
GGCTTGATAGGCTGGACTCGAGGGCAGAGCGTTCATGGGCGATGAGCTGCTACCTTGTGTCAGGGGATTGGGTTGCAATGATTGATCCGAGGACTCGAGGGGCTGCGG
GTGCTTCTTCGGGGGATCTATCATGTAATGCTTGGGTACTGGTCAGGGACCGGATATACATGGTCGTATGTCGGTGGTCAATTGGCATTAAAGAAATGAGCGCGATGG
GCTCGCCAACTTCAGCCTGGCTTGGCGGGTTGCTGTAAGTTGTCGGGCTGGAGCTGACGGTCTTAAGATGGTCAGATGTTAGGATTCAGTATCAATCTTGAATCATG
AGTCGGTCGTTCTGTAGCTTGCAGTCGCAATATGGTCGTCACGTTGCTTCCGTATATGCTCGTACGGTCTGGAGTGGTCTGATGGATTGTGTCATACATGATGAGTG
ACTTGTGAGTTGTCGAAGAGGCTGAGAGAGAGAGGTCGGAACACTGCTCTGGAAITAAITTCAGAAAGAACATATCTTCAAGTCATGGGAGTCAGCTACCGAGCGC
TACAGAGCGGCTGTTC
```

SP中Gms1蛋白在AP基因组中高相似区域【上下游各延伸1kb】

PageDown

7. 对比各个基因结构建模软件:

软件	目标序列	参照序列	物种来源	新基因发现	已知基因的新转录变异体鉴别	基因结构建模效果
NCBI Blast系列	基因组核酸序列	蛋白质、EST/cDNA、mRNA	亲缘关系越近越好	√	√	较差
UCSC BLAT				√	√	较差
NCBI Splign		EST/cDNA、mRNA				一般
EMBL-EBI GeneWise		蛋白质				较好
GENSCAN		从头计算				较好

只能输入一个目标信息和一个基因组信息，只可以基因结构建模

亲缘关系越近，才易得出结果

8. 原核生物从头预测基因:

8.1. 原核生物的启动子区域信号:

Pribnow box  
转录因子结合位点

8.2. ORF 区

有成百上千的碱基对长  
终止密码子

9. 真核生物从头预测基因

9.1. 真核生物的启动子区域信号:

CpG 岛  
poly(A) 尾巴的结合位点

9.2. 外显子和内含子的剪接机制:

剪接位点  
外显子

10. 密码子偏好:

RNA 二级结构、转录/基因表达、翻译延伸速度、蛋白质折叠  
JAVA Codon Adaptation Tool (JCat) <http://www.jcat.de/>

Codon Optimization Tool <http://sg.idtdna.com/CodonOpt>

11. 原核生物从头预测基因:

GLIMMER

GeneMark GeneMarkS GeneMarkS+

12. 真核生物从头预测基因:

GENSCAN

Geneid **【HMM】**

Augustus **【HMM】**

GeneMark-ES GeneMark-ET

GlimmerHMM **【HMM】**

mSplicer

CONTRAST

mGen

FGENESH **【HMM】**