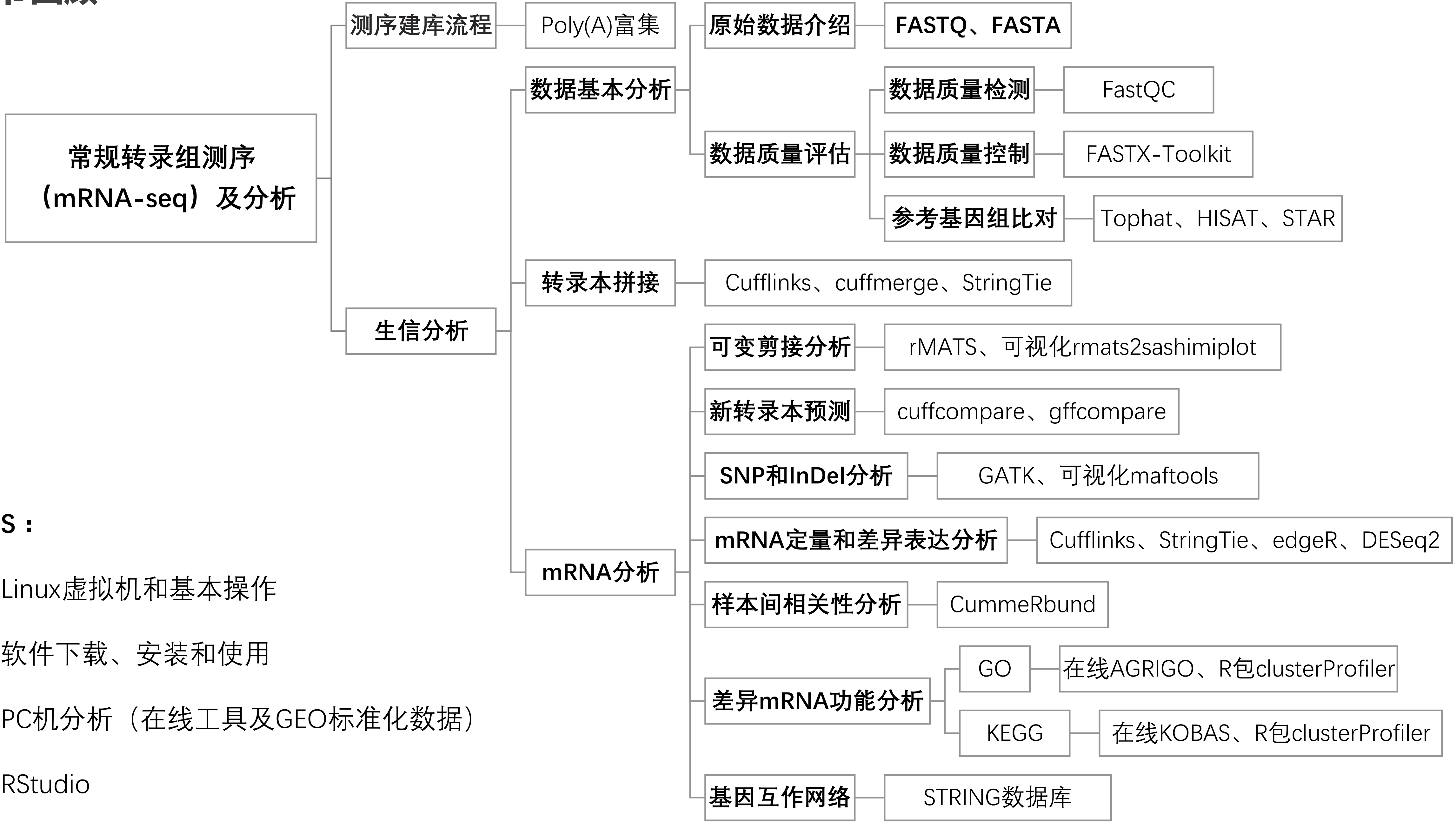
转录组数据分析

（ncRNA）

上节回顾



- **常规转录组测序（mRNA-seq）及分析**
  - 测序建库流程
    - Poly(A)富集
  - **生信分析**
    - **数据基本分析**
      - **原始数据介绍**
        - **FASTQ、FASTA**
      - **数据质量评估**
        - **数据质量检测**
          - FastQC
        - **数据质量控制**
          - FASTX-Toolkit
        - **参考基因组比对**
          - Tophat、HISAT、STAR
    - **转录本拼接**
      - Cufflinks、cuffmerge、StringTie
    - **mRNA分析**
      - **可变剪接分析**
        - rMATS、可视化rmats2sashimiplot
      - **新转录本预测**
        - cuffcompare、gffcompare
      - **SNP和InDel分析**
        - GATK、可视化maftools
      - **mRNA定量和差异表达分析**
        - Cufflinks、StringTie、edgeR、DESeq2
      - **样本间相关性分析**
        - CummeRbund
      - **差异mRNA功能分析**
        - GO
          - 在线AGRIGO、R包clusterProfiler
        - KEGG
          - 在线KOBAS、R包clusterProfiler
      - **基因互作网络**
        - STRING数据库

TIPS：

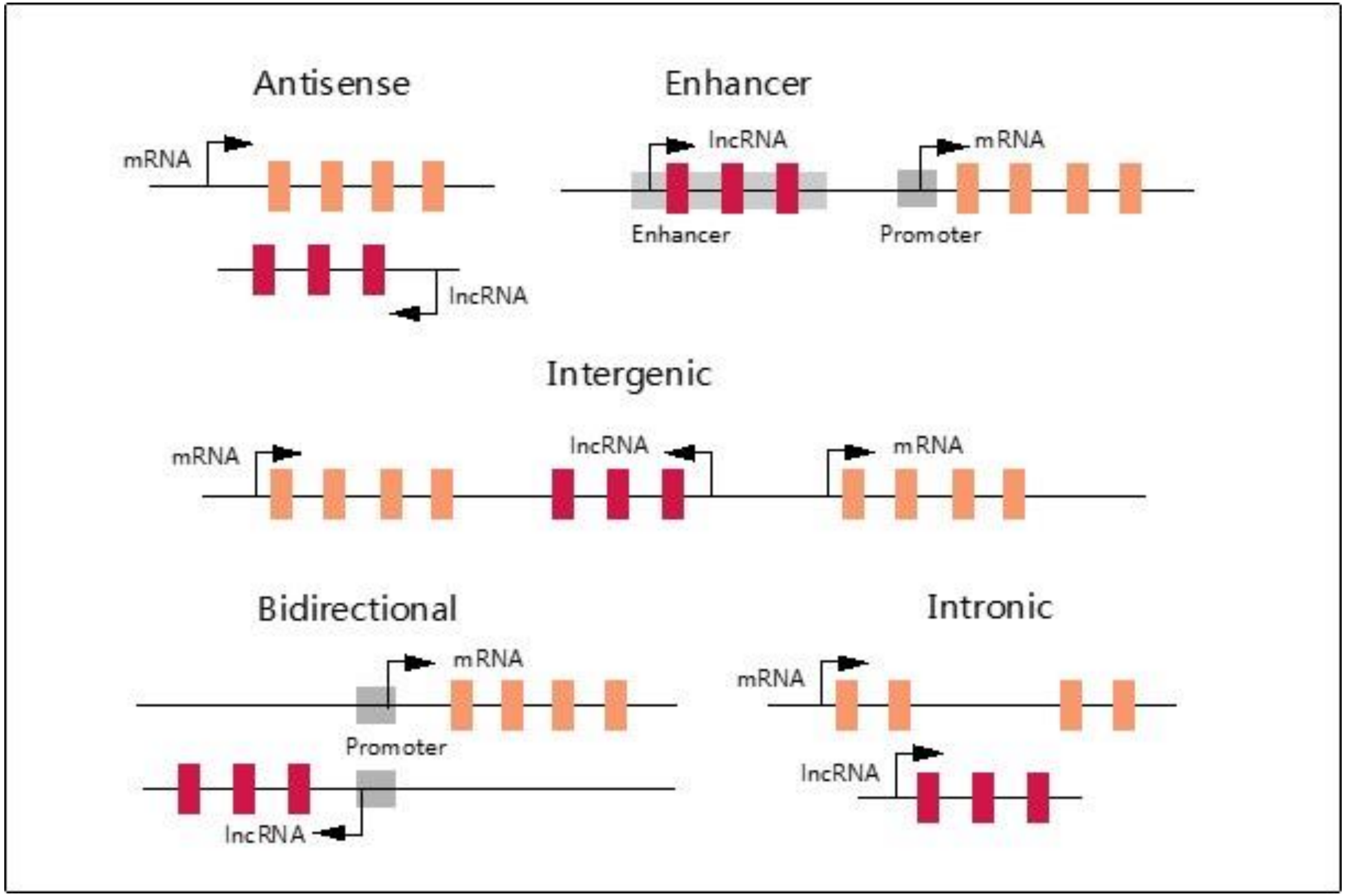
- Linux虚拟机和基本操作
- 软件下载、安装和使用
- PC机分析（在线工具及GEO标准化数据）
- RStudio

本节目录

- 非编码RNA转录组测序（ncRNA-seq）及分析
  - 长非编码RNA（lncRNA）
    - 定义
    - 测序建库流程
    - 生信分析
      - lncRNA的鉴定
        - 基本筛选
        - 编码潜能筛选
          - CPC
          - CPAT
          - PLEK
      - lncRNA分析
        - 特征分析
        - 定量分析
        - 靶基因预测
        - 差异表达分析
        - 靶基因功能分析
      - 数据库及工具
        - NONCODE
  - 小RNA（sRNA）
    - 定义
    - 测序建库流程
    - 生信分析
      - miRNA鉴定
        - miRDeep2
        - miRDeep-P
      - miRNA靶基因预测
        - Miranda
        - psRNATarget
        - miRTarBase
      - 数据库及工具
        - miRBase

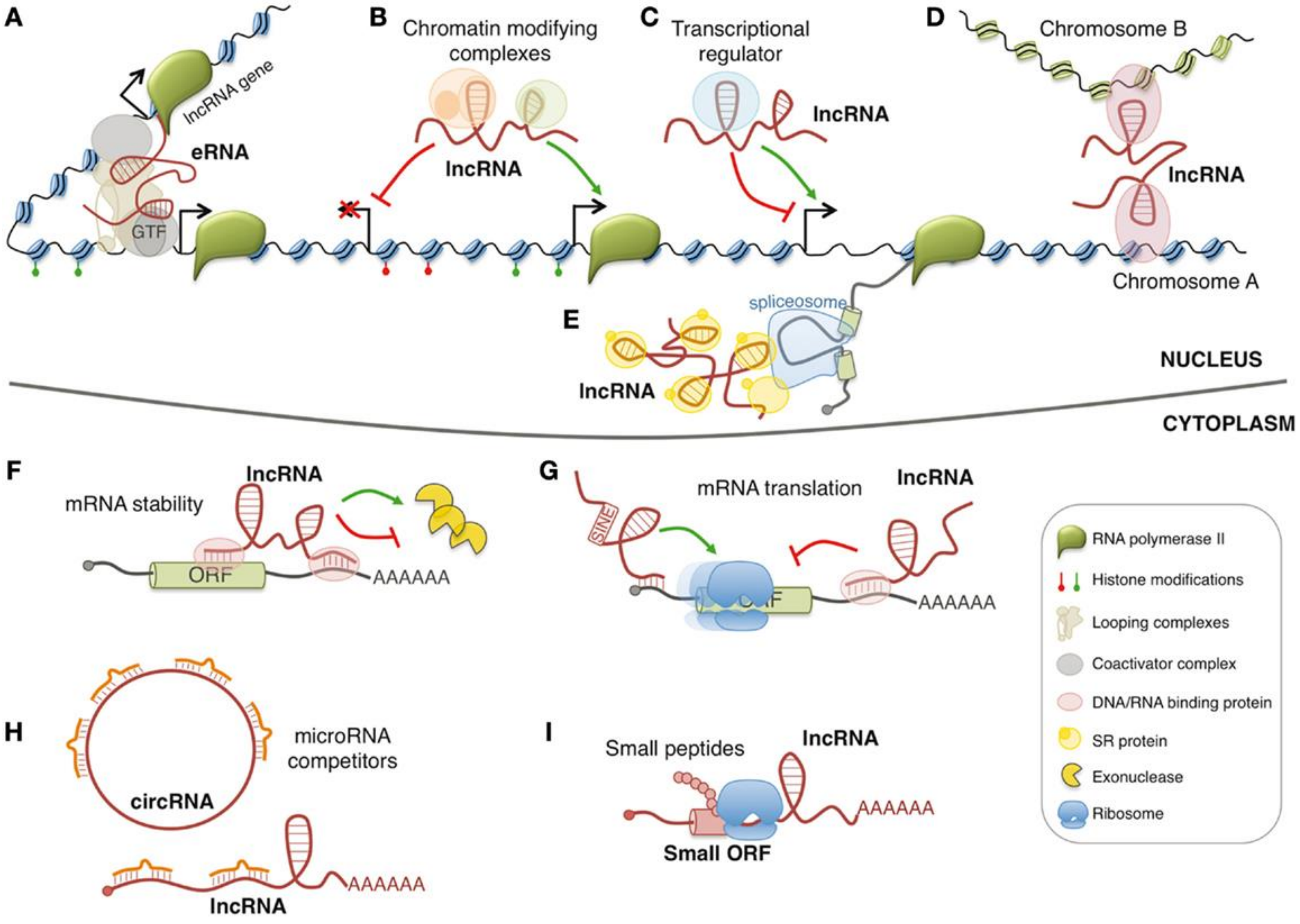
什么是长非编码RNA（lncRNA）

lncRNAs are transcripts with lengths exceeding 200 nucleotides that are not translated into protein



- **Antisense lncRNAs** are transcribed from the antisense strand and overlap with the protein-coding gene sequence.
- **Enhancer lncRNAs** are located in enhancer regions.
- **Intergenic lncRNAs** are transcripts located from protein-coding genes at distances of more than 1~5 kb.
- **Bidirectional lncRNAs** are transcribed within 1 kb of promoters of the protein-coding transcript in the opposite direction.
- **Intronic lncRNAs** are derived from the intron of a coding gene.

什么是长非编码RNA（lncRNA）



(A) regulate transcription by acting as enhancer RNA (eRNA)

(B) recruiting chromatin modifying complexes

(C) regulating transcription factors activity

(D) regulate gene expression by acting on the spatial conformation of chromosomes

(E) regulate gene expression by influencing pre-mRNA splicing.

(F) regulate mRNA expression by regulating mRNA stability,

(G) regulate mRNA expression by regulating mRNA translation

(H) regulate mRNA expression by competing for microRNA binding. In addition

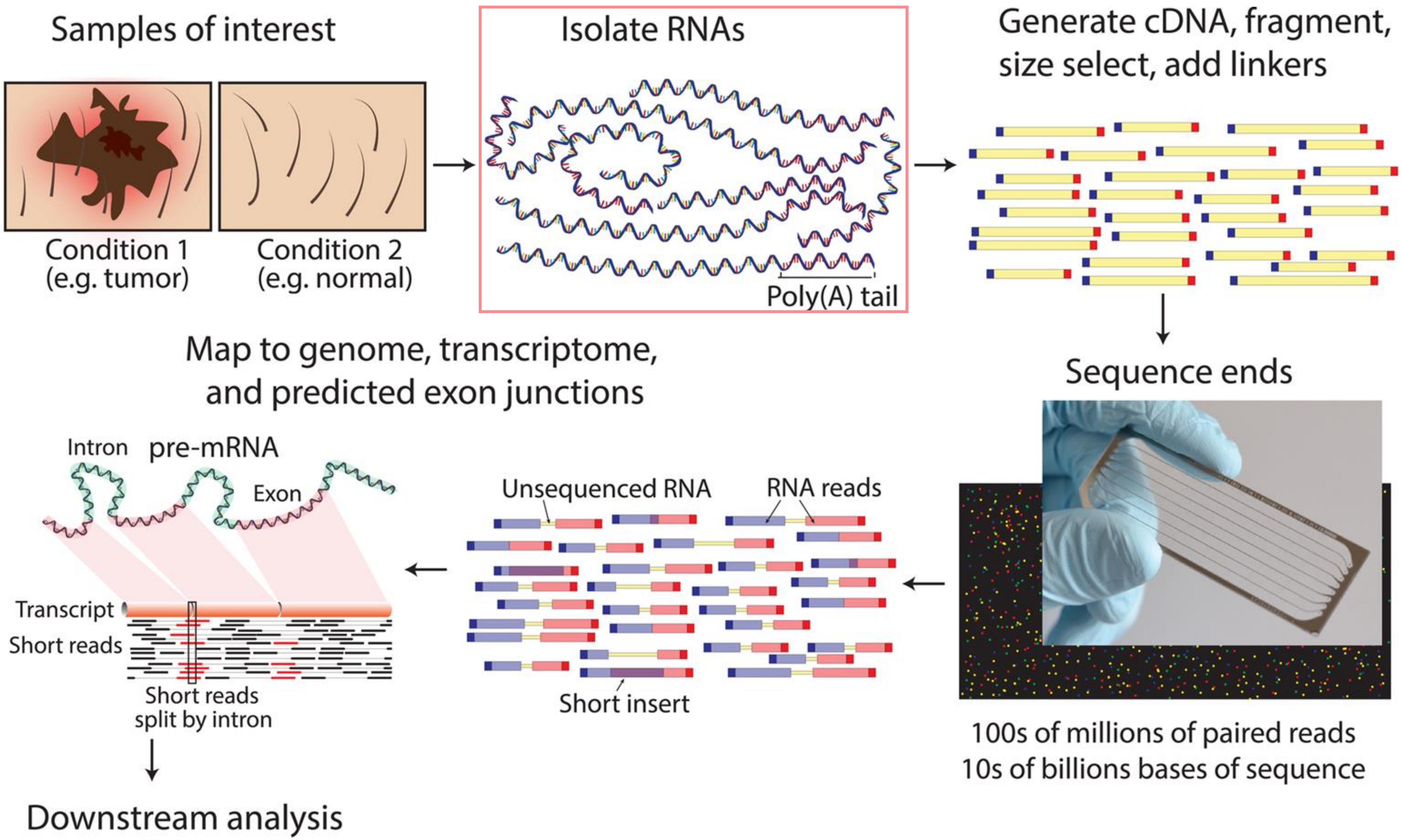
(I) few lncRNAs contain small open reading frames (ORFs) that can be translated in biological active small peptides

什么是长非编码RNA（lncRNA）

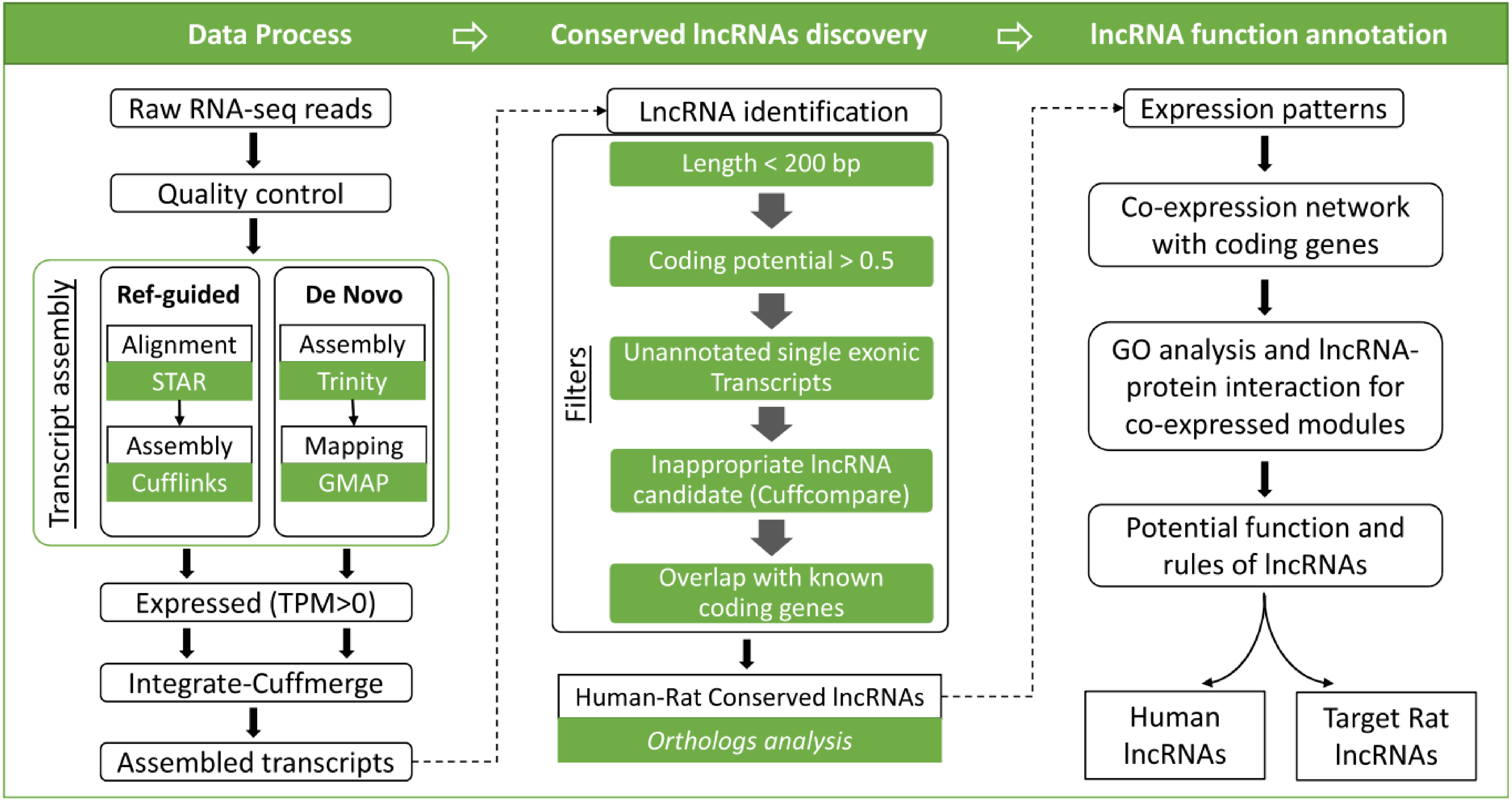
mRNA	**lncRNA**
Tissue-specific expression	Tissue-specific expression
Form secondary structure	Form secondary structure
Undergo post-transcriptional processing, i.e. 5' cap, polyadenylation, splicing	Undergo post-transcriptional processing, i.e. 5' cap, polyadenylation, splicing
Important roles in dieases and development	Important roles in diseases and development
Protein coding transcript	**Non-protein coding, regulatory functions**
Well conserved between species	**Poorly conserved between species**
Present in both nucleus and cytoplasm	**Many predominantly nuclear, others nuclear and/or cytoplasmic**
Total 20-24,000 mRNAs	**Currently ~30,000 lncRNA transcripts, predicted 3-100 fold of mRNA in number**
Expression level: low to high	**Expression level: very low to moderate**

lncRNA-seq测序建库流程

- mRNA：OligoDT磁珠富集mRNA
- lncRNA：Ribo-zero kit去除mRNA



lncRNA鉴定和分析流程



Data Process ⇨ Conserved lncRNAs discovery ⇨ lncRNA function annotation

Raw RNA-seq reads

Quality control

Transcript assembly

Ref-guided

Alignment

STAR

Assembly

Cufflinks

De Novo

Assembly

Trinity

Mapping

GMAP

Expressed (TPM>0)

Integrate-Cuffmerge

Assembled transcripts

LncRNA identification

Filters

Length < 200 bp

Coding potential > 0.5

Unannotated single exonic Transcripts

Inappropriate lncRNA candidate (Cuffcompare)

Overlap with known coding genes

Human-Rat Conserved lncRNAs

Orthologs analysis

Expression patterns

Co-expression network with coding genes

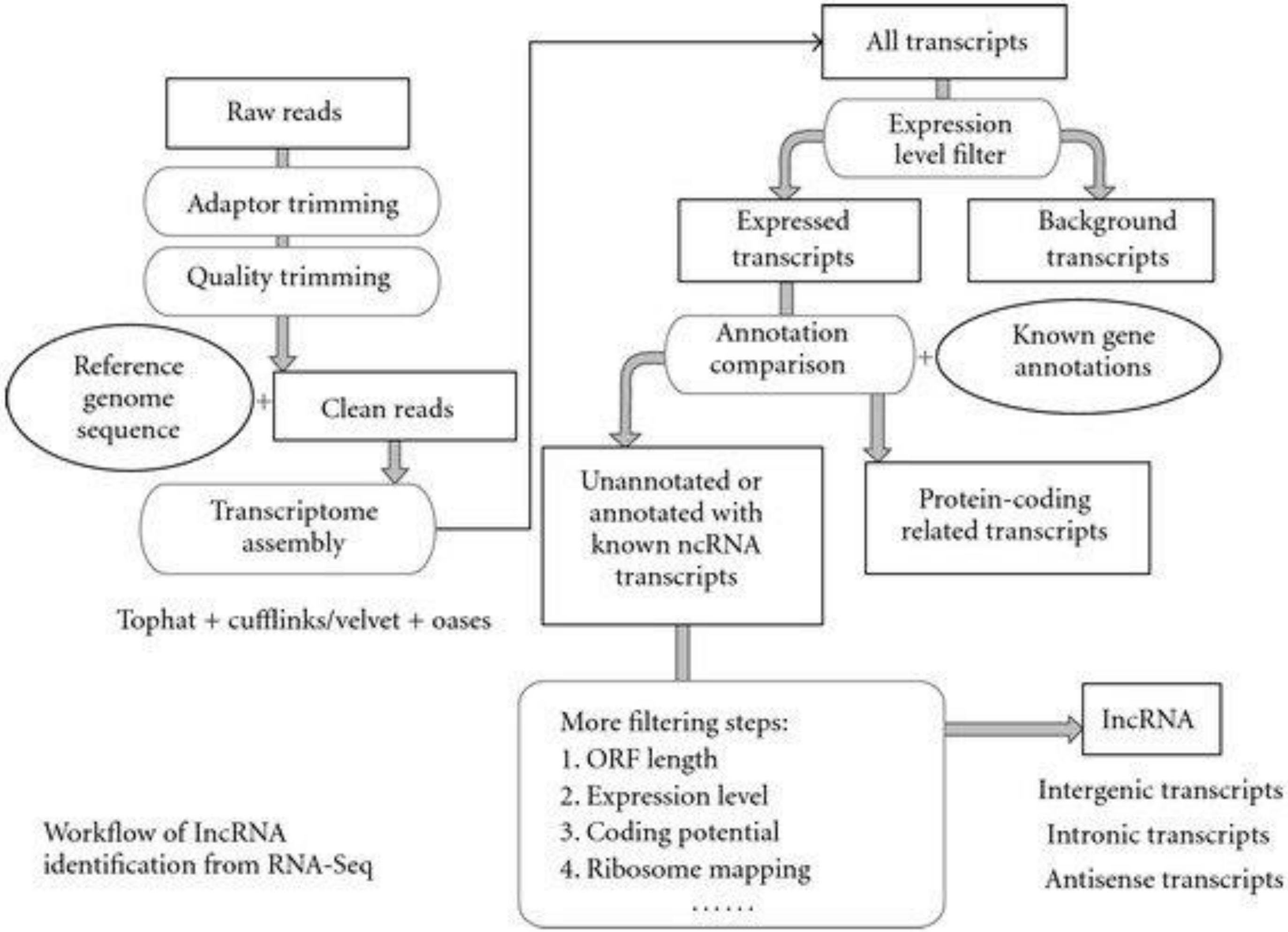
GO analysis and lncRNA-protein interaction for co-expressed modules

Potential function and rules of lncRNAs

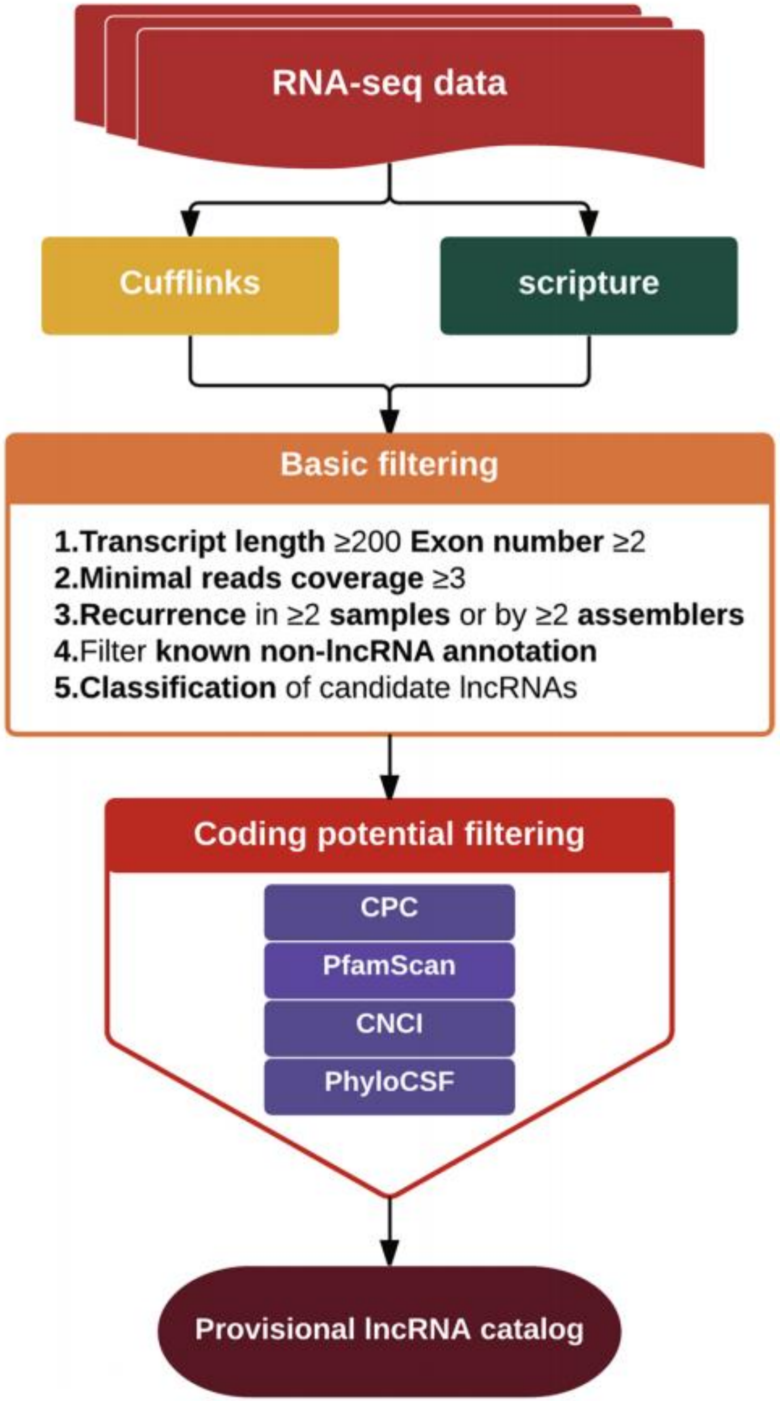
Human lncRNAs

Target Rat lncRNAs

lncRNA鉴定流程



Workflow of lncRNA identification from RNA-Seq



1.Transcript length ≥200 **Exon number** ≥2
2.Minimal reads coverage ≥3
3.Recurrence in ≥2 **samples** or by ≥2 **assemblers**
4.Filter **known non-lncRNA annotation**
5.Classification of candidate lncRNAs

lncRNA鉴定软件

Name	Species	Web Server	Input File	Main Model/Algorithm	Training-set	Year Published
LGC	Plant/Animal	http://bigd.big.ac.cn/lgc/calculator	FASTA/BED/GTF	Relationship between ORF length and GC content	NO	2019
CPAT	**Human/Fly/Mouse/Zebrafish**	**http://lilab.research.bcm.edu/cpat/index.php**	**FASTA/BED**	**Logistic Regression**	**YES**	**2013**
CPC	**All**	**http://cpc.cbi.pku.edu.cn/**	**FASTA**	**Support Vector Machine**	**NO**	**2007**
COME	Plant/Human/Mouse/Fly/Worm	https://www.ncrnalab.org/RNAfinder/CP.html	GTF	Balanced Random Forest	YES	2017
lncRScan-SVM	Human	NA	FASTA/BED/GTF/GFF	Support Vector Machine	YES	2015
CNCI	Plant/Animal	NA	FASTA/GTF	Support Vector Machine	NO	2013
PLEK	**Vertebrate**	**NA**	**FASTA**	**Support Vector Machine**	**NO**	**2014**
FEELnc	All	https://github.com/tderrien/FEELnc	FASTA/GTF	Random Forest	YES	2017
PhyloCSF	Vertebrate/Fly/Mosquito/Yeast/Worm	NA	FASTA	Phylogenetic Codon Model	YES	2011
PORTRAIT	Fungi	http://bioinformatics.cenargen.embrapa.br/portrait	FASTA	Support Vector Machine	YES	2009
PLIT	Plant	NA	FASTA	LASSO/Random Forest	YES	2018
RNAplonc	Plant	NA	FASTA	REPTree	YES	2018
PLncPRO	Plant/Animal	NA	FASTA	Random Forest	YES	2017
CREMA	Plant/Animal	NA	FASTA	Ensemble approach	YES	2018

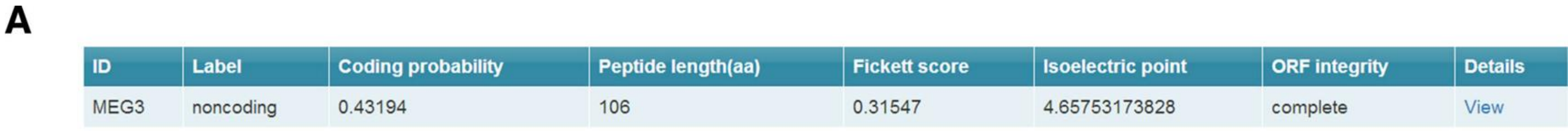
lncRNA鉴定软件——CPC

- CPC（Coding Potential Calculator）是一种蛋白质编码潜能计算工具，将转录本与已知蛋白数据库做blastx比对，依据转录本各个编码框的生物学序列特征，通过支持向量机的分类器来评估转录本的编码潜能。这里的数据库使用的是NCBI真核蛋白数据库（NRDB）。
- 2017年更新第二版。

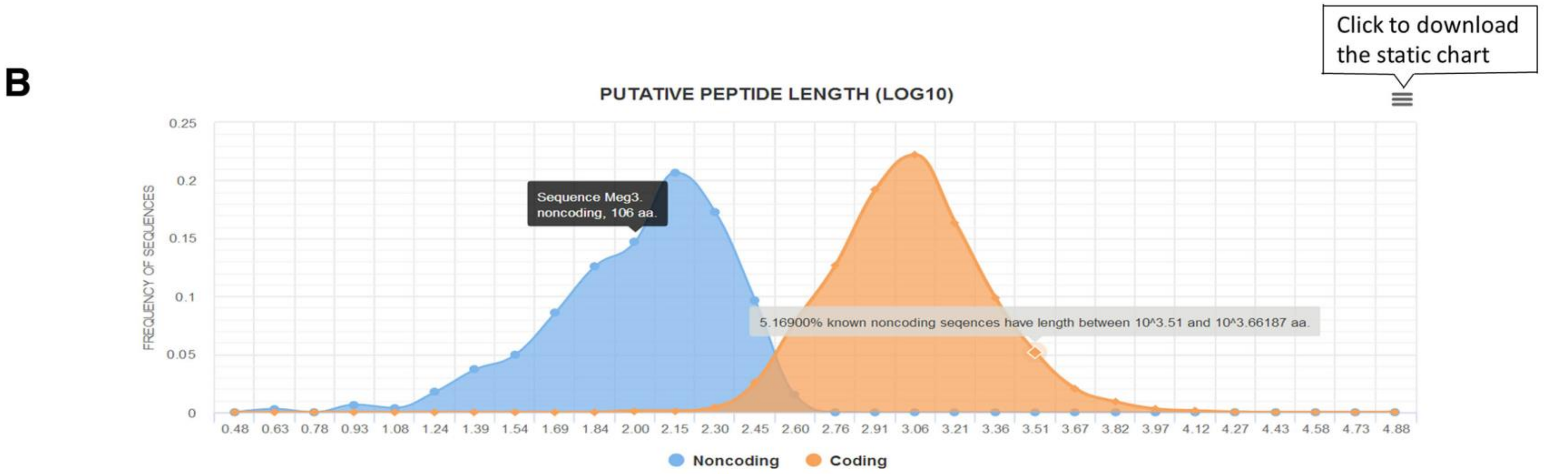
Linux运行

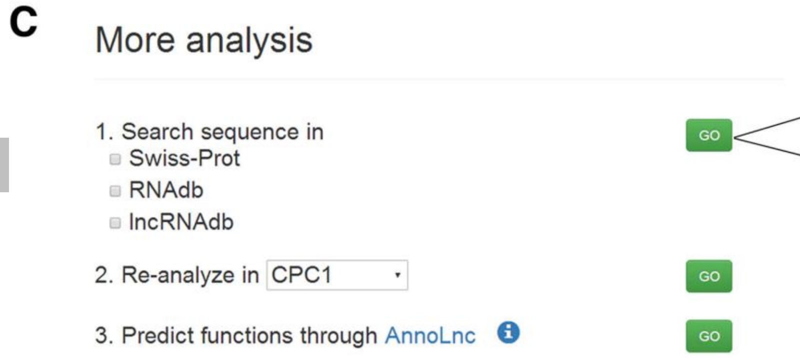
```
$ bin/CPC2.py -i (input_seq)
-o (result_in_table)
```

在线版CPC2



ID	Label	Coding probability	Peptide length(aa)	Fickett score	Isoelectric point	ORF integrity	Details
MEG3	noncoding	0.43194	106	0.31547	4.65753173828	complete	View

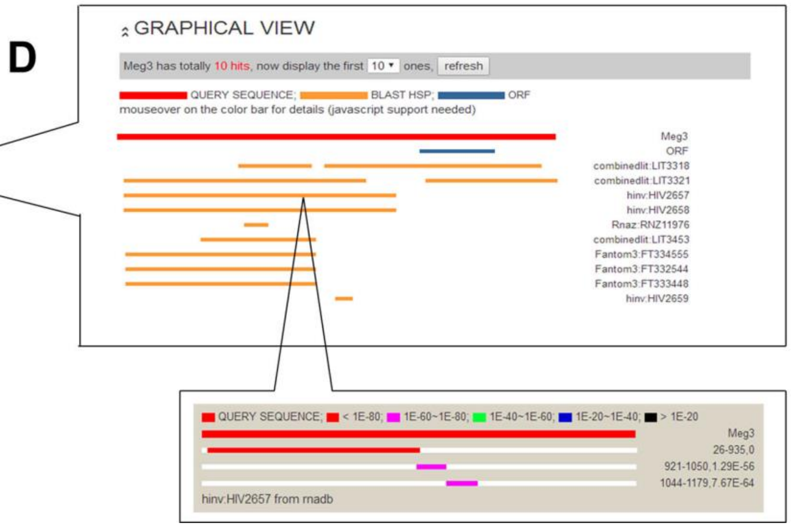




More analysis

1. Search sequence in
   - ☐ Swiss-Prot
   - ☐ RNAdb
   - ☐ lncRNAdb

   GO
2. Re-analyze in CPC1 ▾ GO
3. Predict functions through AnnoLnc ℹ GO

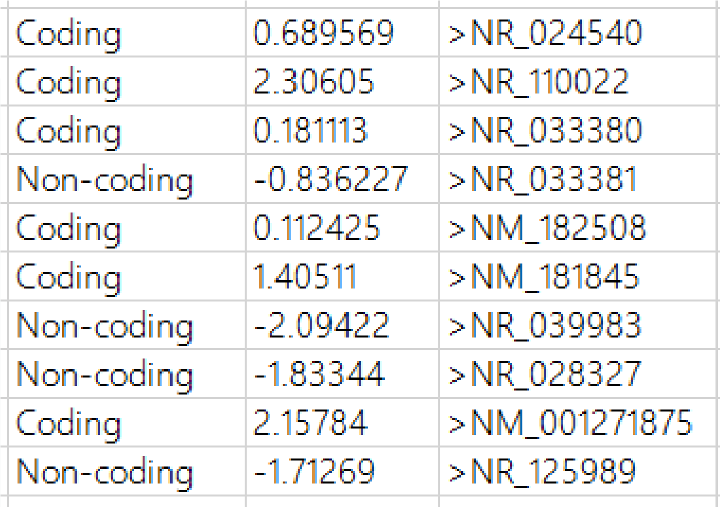


Meg3 has totally 10 hits, now display the first 10 ▾ ones, refresh

mouseover on the color bar for details (javascript support needed)

lncRNA鉴定软件——PLEK

PLEK，是使用基于改进的k-mer方案和支持向量机（SVM）算法，在缺乏基因组序列或注释的情况下将lncRNA与信使RNA（mRNA）区分开。PLEK 可在注释完好的mRNA和lncRNA转录物上进行评估。人RefSeq mRNAs和GENCODE lncRNAs的10倍交叉验证试验表明PLEK可以达到高达95.6%的准确性。



Coding	0.689569	>NR_024540
Coding	2.30605	>NR_110022
Coding	0.181113	>NR_033380
Non-coding	-0.836227	>NR_033381
Coding	0.112425	>NM_182508
Coding	1.40511	>NM_181845
Non-coding	-2.09422	>NR_039983
Non-coding	-1.83344	>NR_028327
Coding	2.15784	>NM_001271875
Non-coding	-1.71269	>NR_125989

```
#install
$ wget https://sourceforge.net/projects/plek/files/PLEK.1.2.tar.gz
$ tar xzvf PLEK.1.2.tar.gz
$ cd PLEK.1.2
$ python PLEK_setup.py

#run (默认情况下会调用内置的svm模型)
$ python PLEK.py -fasta transcript.fa -out output -thread 10

#物种已知的mRNA和lncRNA转录本序列, 也可以构建自己的模型
$ python PLEKModelling.py mRNA mRNAs.fa -lncRNA lncRNAs.fa -prefix 20190129
$ python PLEK.py -fasta transcript.fa -out output -model 20190129.mode -range 20190129.range
-thread 10
```

lncRNA鉴定软件——CPAT（在线）

CPAT
Coding Potential Assessment Tool

大部分lncRNA编码潜能软件分析的缺陷是：

- 大多数lncRNAs保守性较差，倾向于谱系特异性，这大大限制了基于比对的方法的区分能力。
- 蛋白质编码基因的一个重要部分可能有一个交替处理的同种型或一个从另一个启动子转录的，这部分ncRNA不能通过同源搜索正确分类，因为它们会与蛋白质编码基因有显着的匹配。
- 基于对齐的方法是非常缓慢的。
- 可靠性取决于对齐质量。大多数对齐工具使用启发式搜索，并不保证给出最佳对齐。

CPAT则克服了以上的缺陷。

Test region

○ BED or FASTA file, regular or compressed: (<10 MB allowed)

● BED or FASTA data:

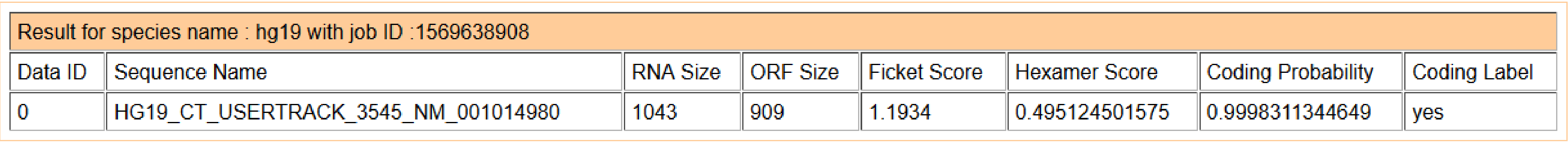
```
>hg19_ct_UserTrack_3545_NM_001014980 range=chr1:1177826-1182102
5'pad=0 3'pad=0 strand=- repeatMasking=none
CTCGCCGCGCTGAGCCGCCTCGGGACGGAGCCATGCGGCGCTGGGCCTGG
GCCGCGGTCGTGGTCCTCCTCGGGCCGCAGCTCGTGCTCCTCGGGGGCGT
CGGGGCCCGGCGGGAGGCACAGAGGACGCAGCAGCCTGGCCAGCGCGCAG
ATCCCCCCAACGCCACCGCCAGCGCGTCCTCCCGCGAGGGGCTGCCCGAG
GCCCCCAAGCCATCCCAGGCCTCAGGACCTGAGTTCTCCGACGCCCACAT
GACATGGCTGAACTTTGTCCGGCGGCCGGACGACGGCGCCTTAAGGAAGC
GGTGCGGAAGCAGGGACAAGAAGCCGCGGGATCTCTTCGGTCCCCCAGGA
CCTCCAGGTGCAGAAGTGACCGCGGAGACTCTGCTTCACGAGTTTCAGGA
GCTGCTGAAAGAGGCCACGGAGCGCCGGTTCTCAGGGCTTCTGGACCCGC
```

○ BED or FASTA url:

Select Species assembly

- ● Human (hg19, GRCh37)
- ○ Mouse (NCBI Build 37/mm9)
- ○ Mouse (GRCm38/mm10)
- ○ Fly (dm3, BDGP Release 5)
- ○ Zebrafish (Zv9/danRer7)

Can I use a different species or assembly?



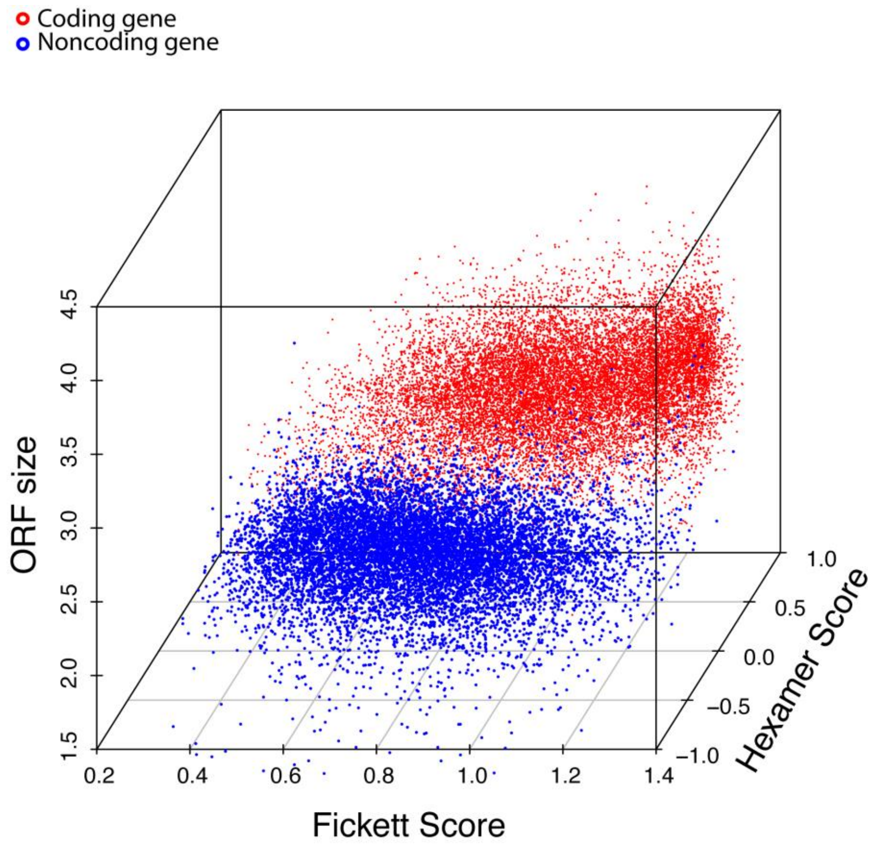
Result for species name : hg19 with job ID :1569638908

Data ID	Sequence Name	RNA Size	ORF Size	Ficket Score	Hexamer Score	Coding Probability	Coding Label
0	HG19_CT_USERTRACK_3545_NM_001014980	1043	909	1.1934	0.495124501575	0.9998311344649	yes

lncRNA鉴定软件——CPAT（linux命令行）

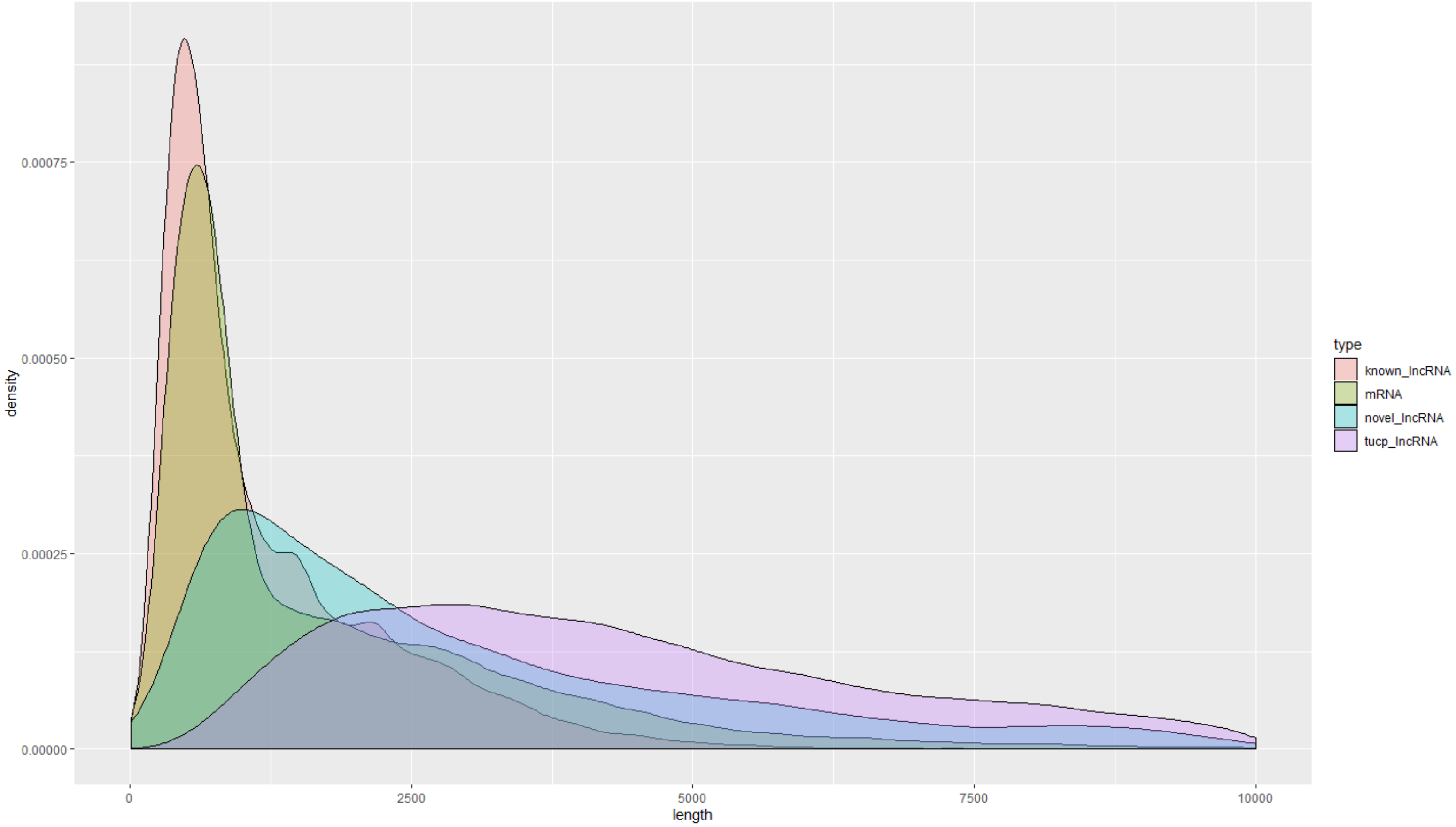
```
# use BED file as input. '-r' is required
$ cpat.py -r /database/hg19.fa -g
Human_test_coding_mRNA_hg19.bed -d
../dat/Human_logitModel.RData -x
../dat/Human_Hexamer.tsv -o output1

# use FASTA file as input. '-r' is not required
$ cpat.py -g Human_test_coding_mRNA.fa  -d
../dat/Human_logitModel.RData -x
../dat/Human_Hexamer.tsv -o output2

# output
Name
mRNA_size
ORF_size
Fickett_score
Hexamer_score
coding_prob
```

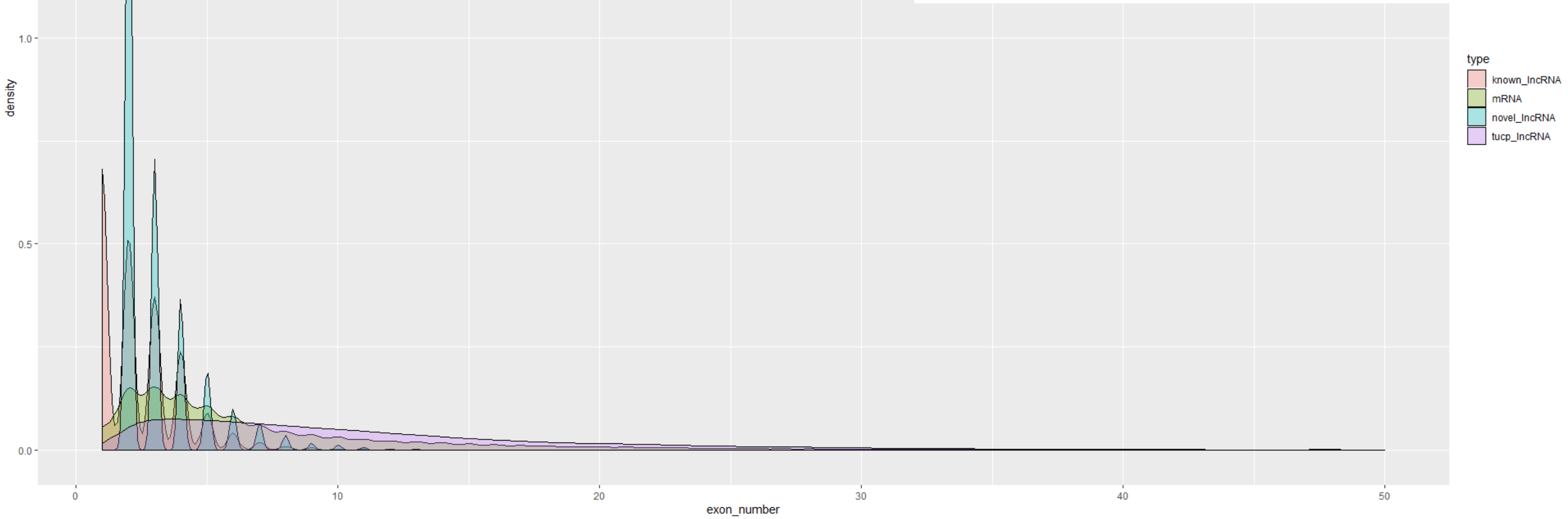
- Human coding probability (CP) cutoff: 0.364 (CP >=0.364 indicates coding sequence, CP < 0.364 indicates noncoding sequence)
- Mouse coding probability (CP) cutoff: 0.44
- Fly coding probability (CP) cutoff: 0.39
- Zebrafish coding probability (CP) cutoff: 0.38



lncRNA分析——特征分析

- 与mRNA的长度比较分析

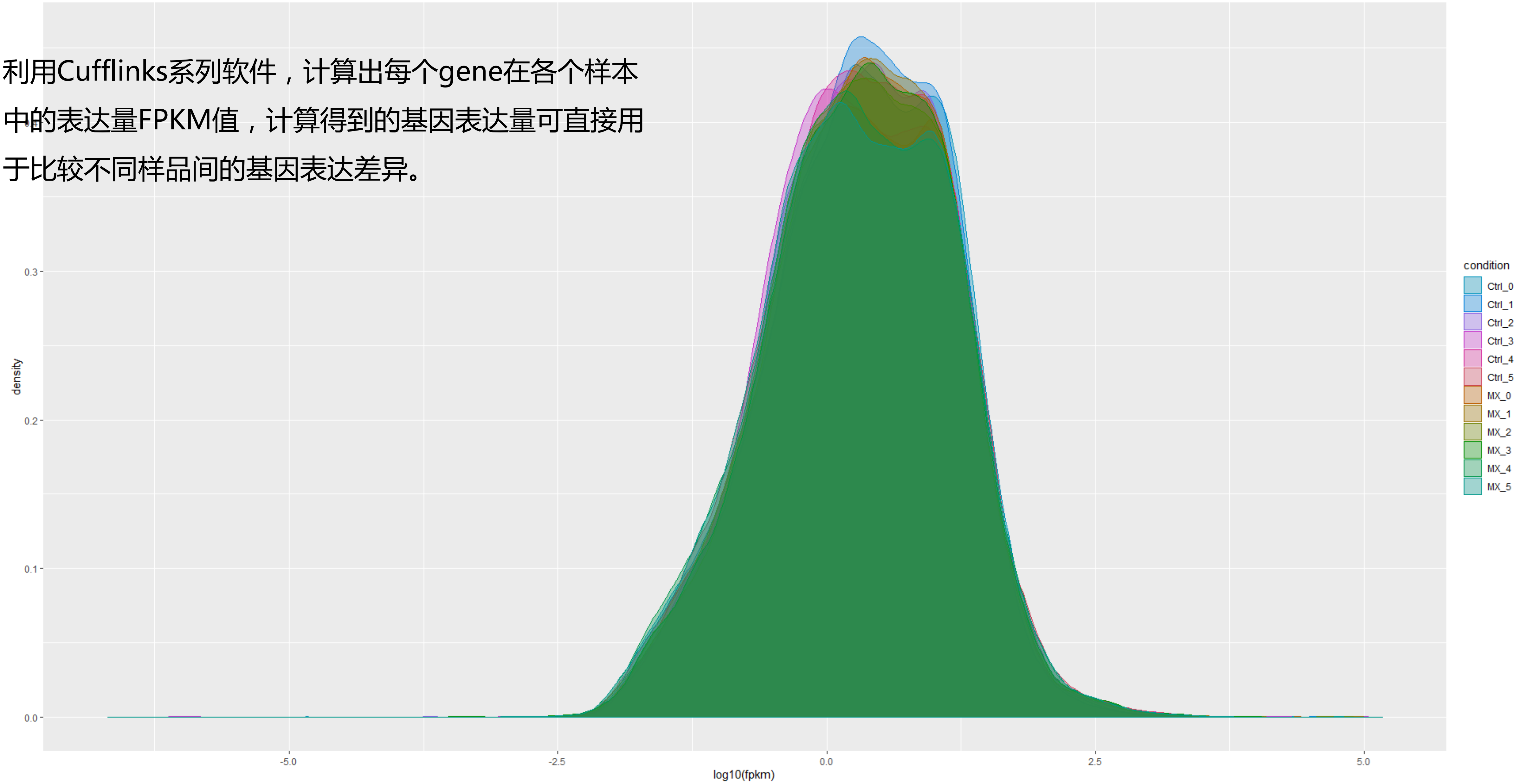




- 与mRNA外显子数目比较

lncRNA分析——定量分析

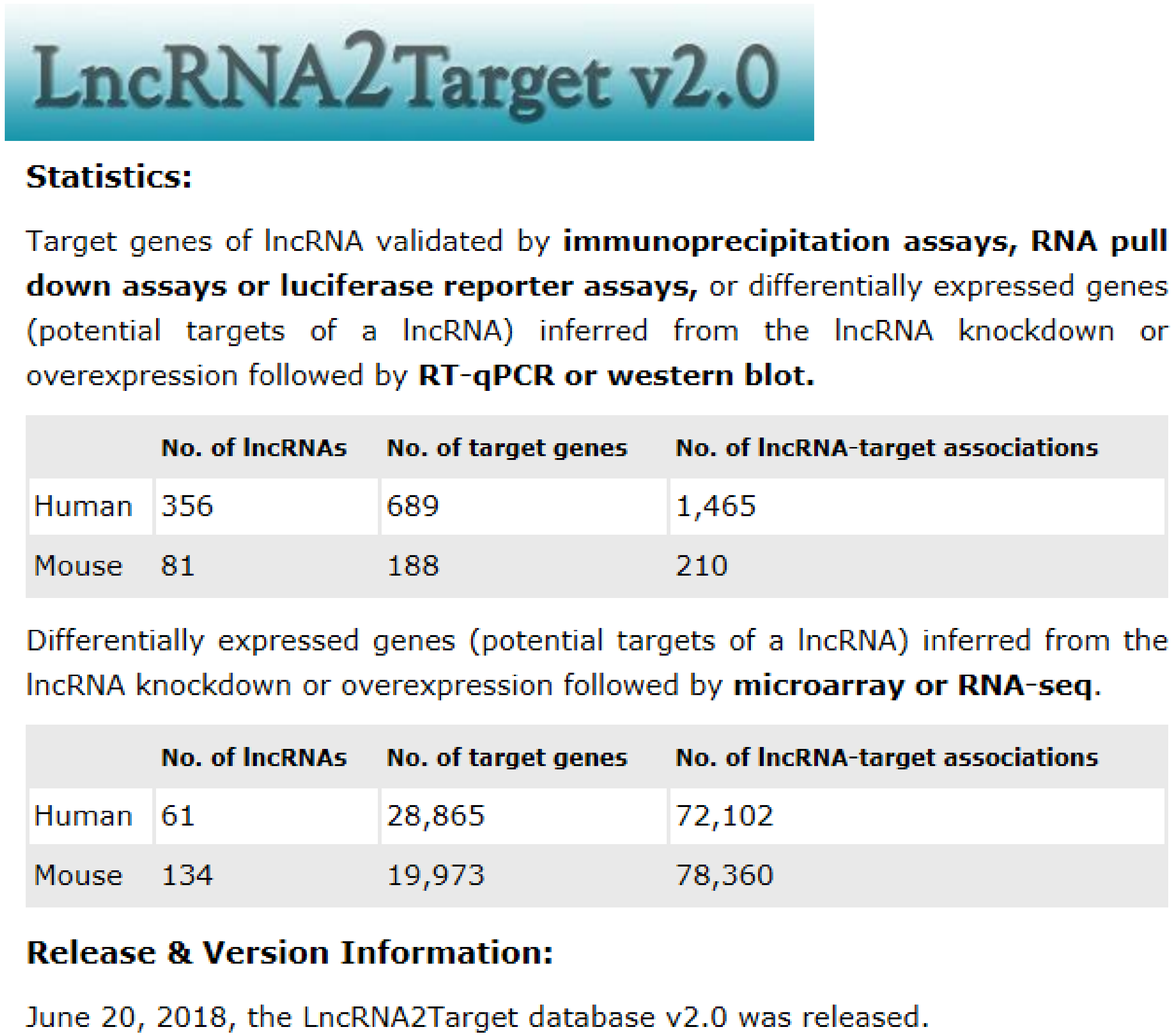
- 利用Cufflinks系列软件，计算出每个gene在各个样本中的表达量FPKM值，计算得到的基因表达量可直接用于比较不同样品间的基因表达差异。



lncRNA分析——靶基因预测

- 通过lncRNA在染色体上相对与mRNA的位置、以及lncRNA与mRNA表达的相关性来进行靶向mRNA的预测分析。

- 通过样本间lncRNA与mRNA的表达量相关性分析或共表达分析方法来预测lncRNA靶基因。此方法需要一定样品数量保证其预测的准确性，因此当样本量小于5时不进行此项分析。当样品数小于25时，采用Pearson相关系数法分析样本间lncRNA与mRNA的相关性，取相关性绝对值大于0.99的mRNA基因进行功能富集分析预测lncRNA的主要功能；当样本数大于或等于25时可采用WGCNA将不同的组织、处理或者时间点间表达模式相似的基因聚类，得到不同的共表达模块，根据模块内已知的mRNA 基因功能进一步探索研究lncRNA的功能。



LncRNA2Target v2.0

Statistics:

Target genes of lncRNA validated by **immunoprecipitation assays, RNA pull down assays or luciferase reporter assays,** or differentially expressed genes (potential targets of a lncRNA) inferred from the lncRNA knockdown or overexpression followed by **RT-qPCR or western blot.**

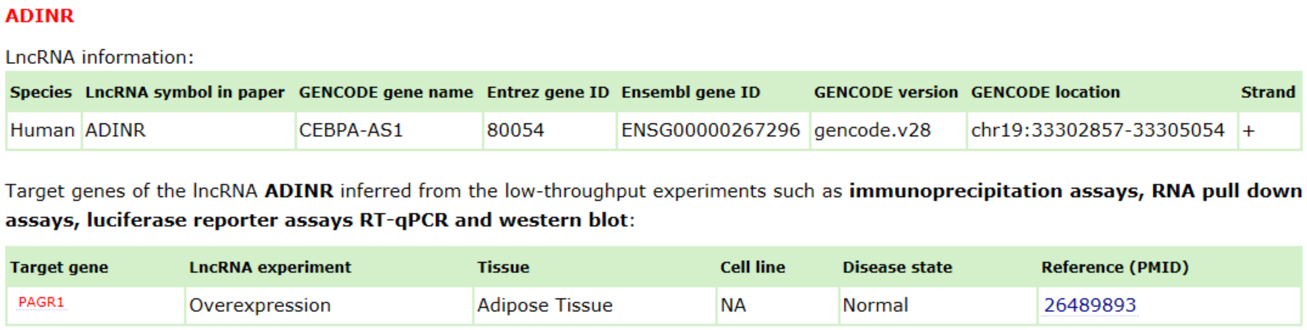
	No. of lncRNAs	No. of target genes	No. of lncRNA-target associations
Human	356	689	1,465
Mouse	81	188	210

Differentially expressed genes (potential targets of a lncRNA) inferred from the lncRNA knockdown or overexpression followed by **microarray or RNA-seq**.

	No. of lncRNAs	No. of target genes	No. of lncRNA-target associations
Human	61	28,865	72,102
Mouse	134	19,973	78,360

Release & Version Information:

June 20, 2018, the LncRNA2Target database v2.0 was released.



ADINR

LncRNA information:

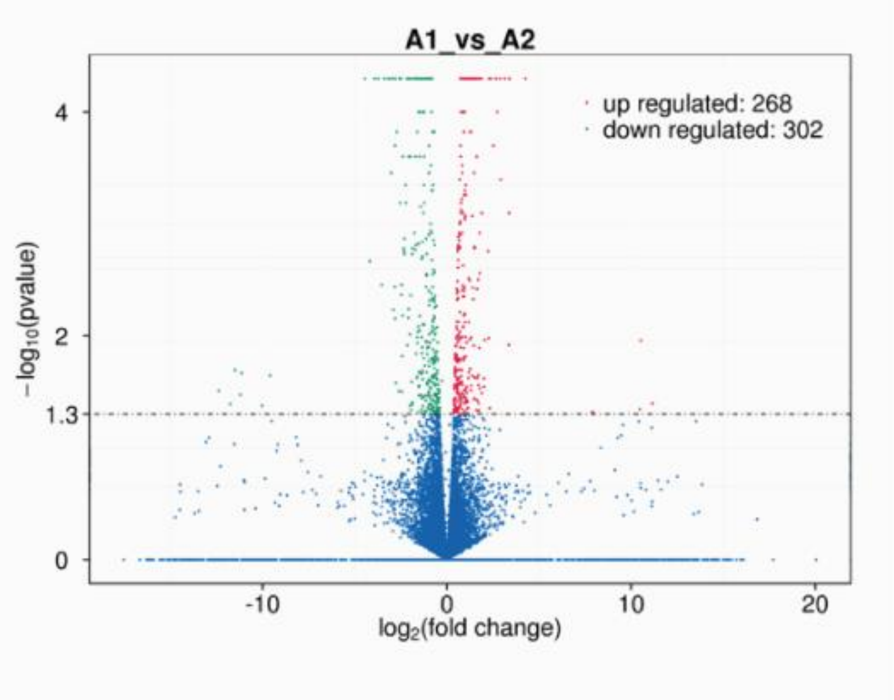
Species	LncRNA symbol in paper	GENCODE gene name	Entrez gene ID	Ensembl gene ID	GENCODE version	GENCODE location	Strand
Human	ADINR	CEBPA-AS1	80054	ENSG00000267296	gencode.v28	chr19:33302857-33305054	+

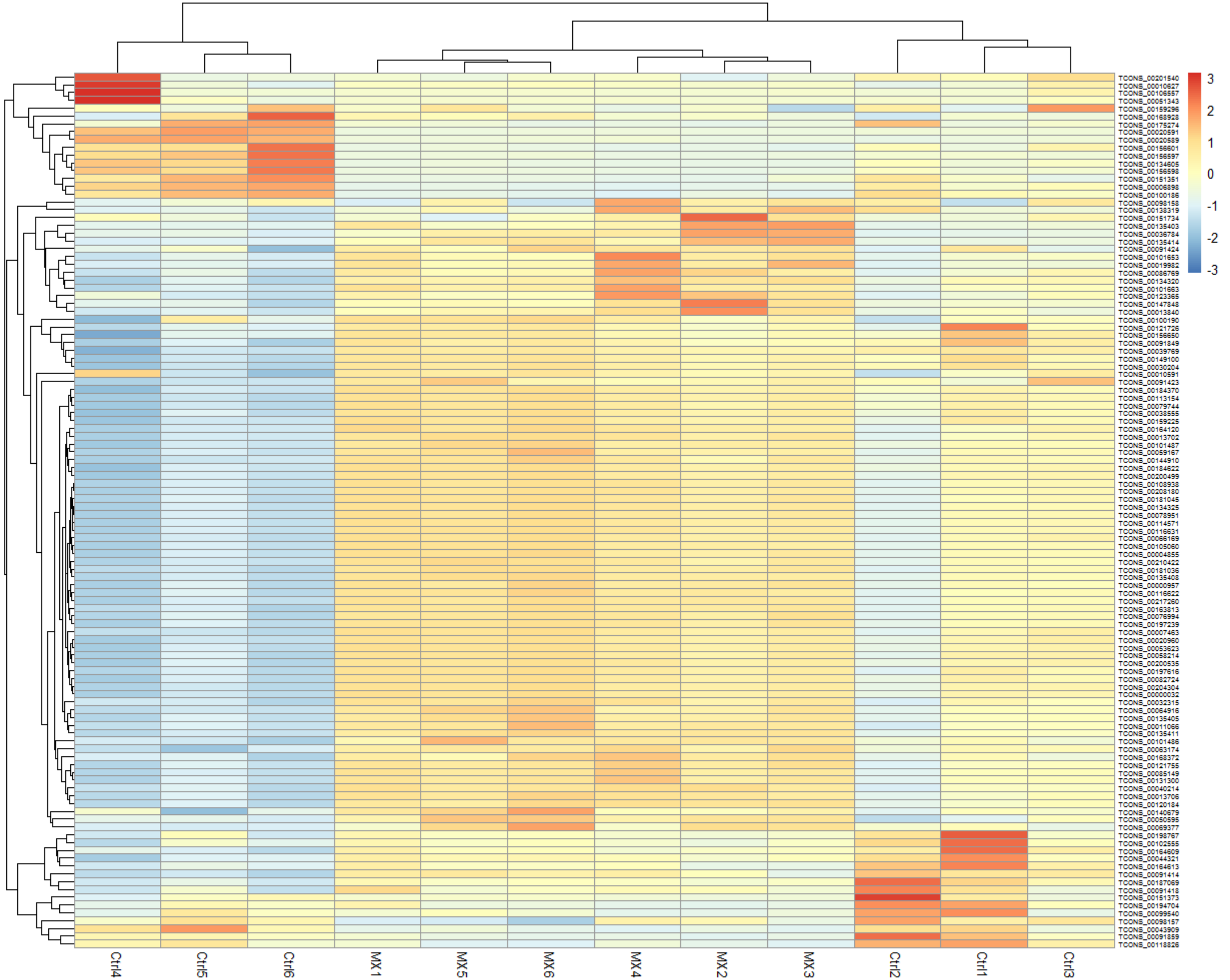
Target genes of the lncRNA **ADINR** inferred from the low-throughput experiments such as **immunoprecipitation assays, RNA pull down assays, luciferase reporter assays RT-qPCR and western blot**:

Target gene	LncRNA experiment	Tissue	Cell line	Disease state	Reference (PMID)
PAGR1	Overexpression	Adipose Tissue	NA	Normal	26489893

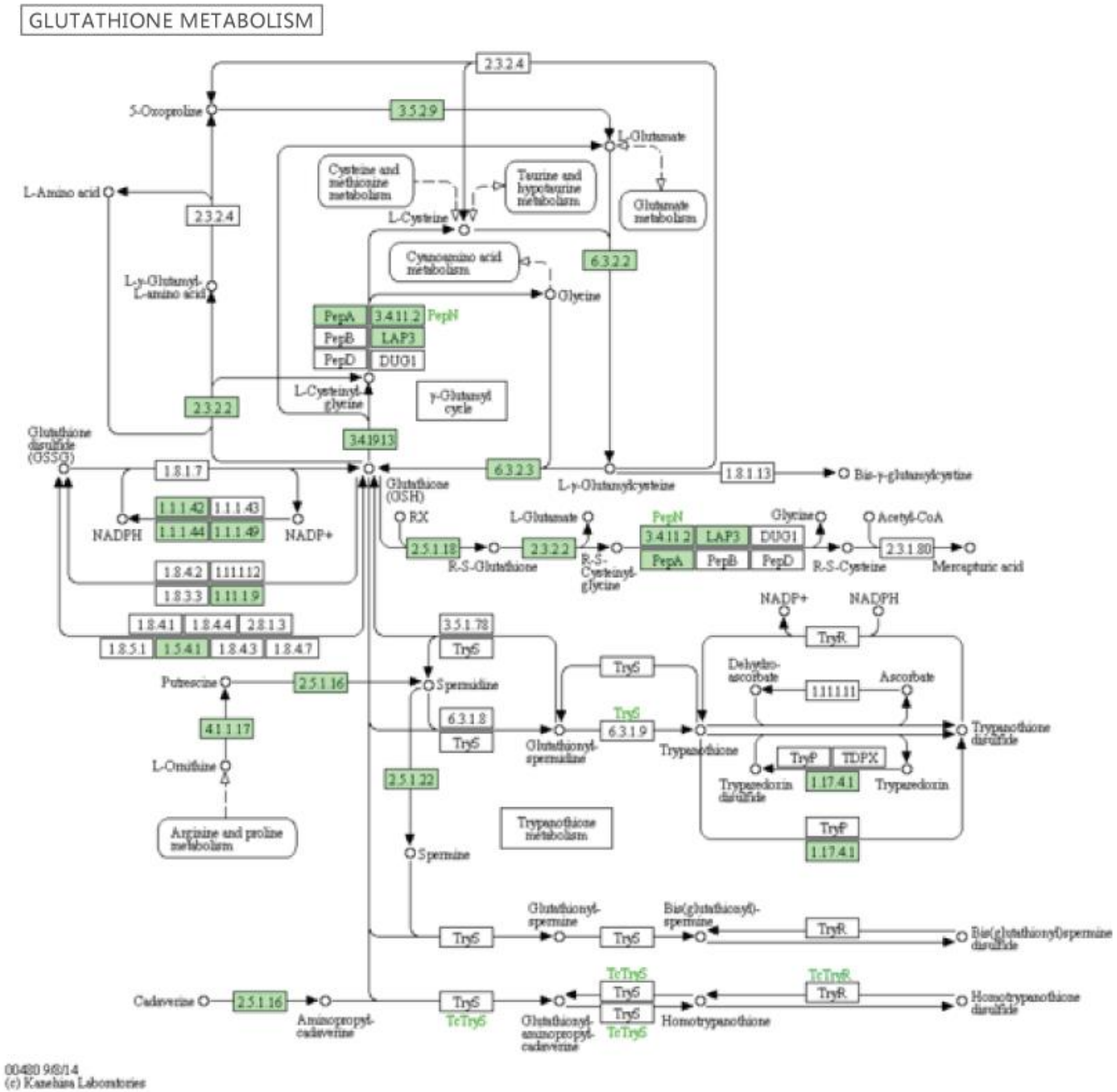
lncRNA分析——表达差异分析

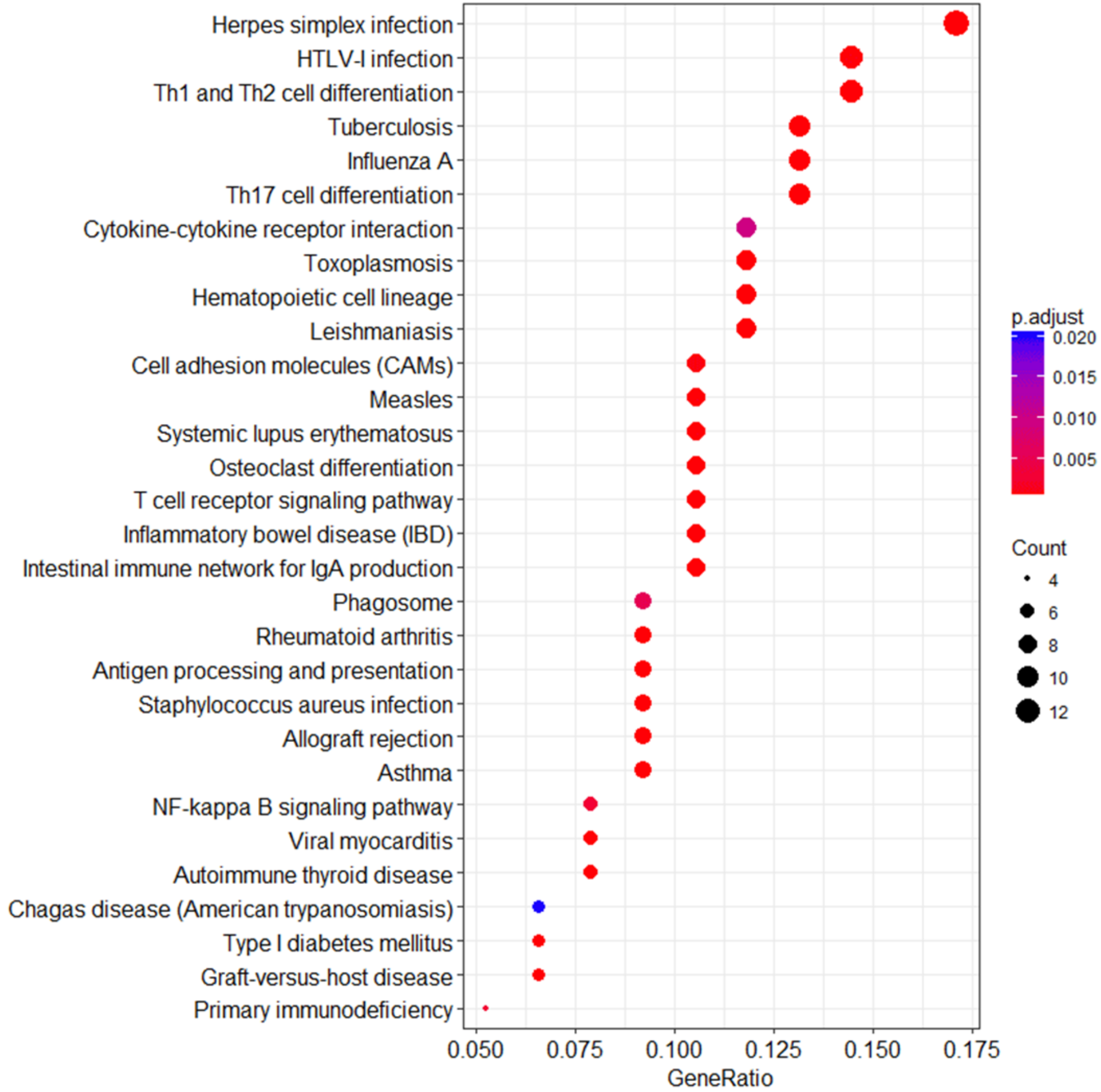
- Cuffdiff
- edgeR等





lncRNA分析——靶基因功能富集分析

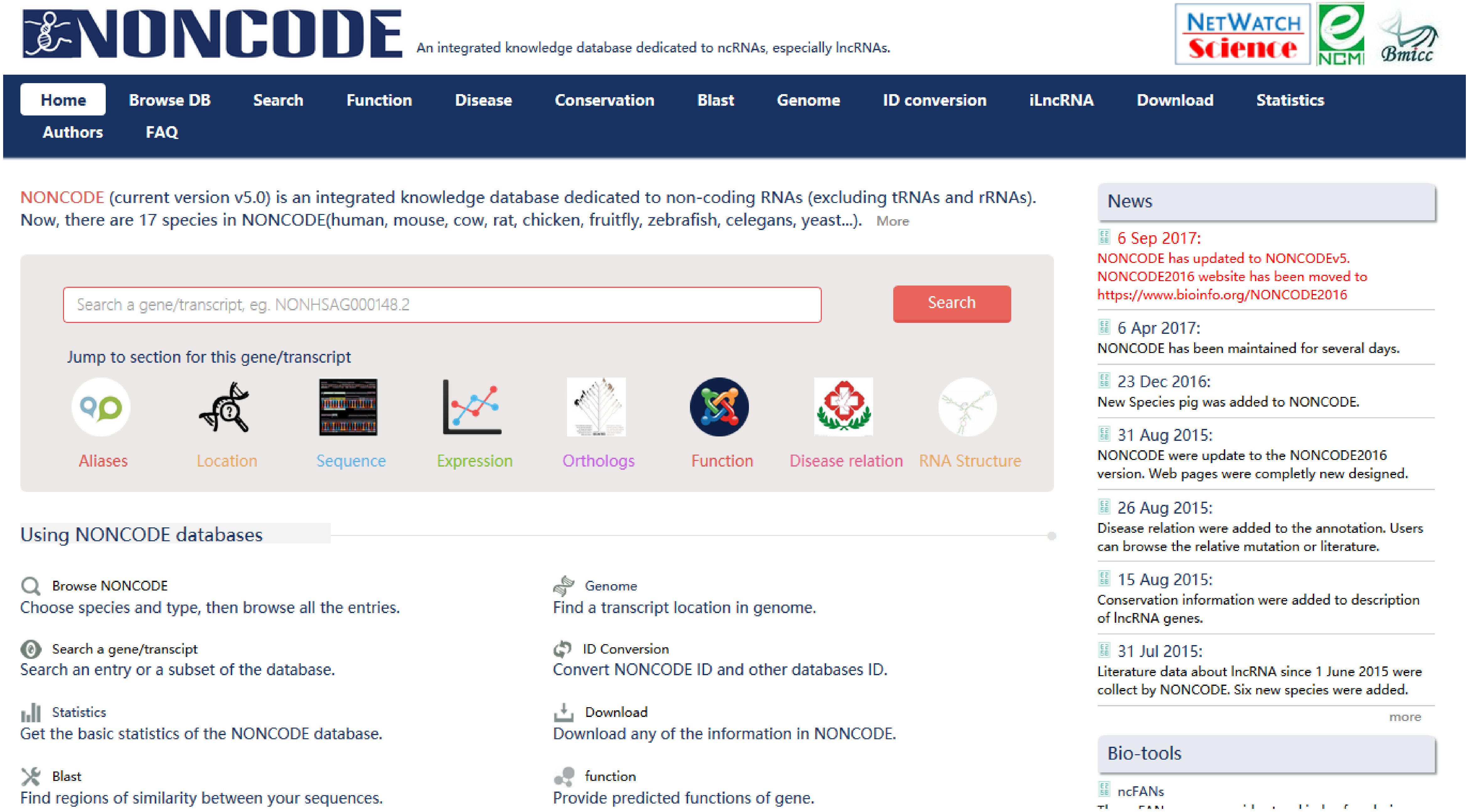




lncRNA-related databases

Name	Description	Link
deepBase	Identification, expression, evolution and function of long non-coding RNAs (LncRNAs), small RNAs and circular RNAs from deep-sequencing data	http://rna.sysu.edu.cn/deepBase/
LNCipedia	A comprehensive compendium of human long non-coding RNAs.	http://www.lncipedia.org/
lncRNAdb	The Reference Database For Functional Long Noncoding RNAs.	http://lncrnadb.com/
LncRNAWiki	A wiki-based, publicly editable and open-content platform for community curation of human long non-coding RNAs (lncRNAs)	http://lncrna.big.ac.cn/
LncBook	A comprehensive collection of 270,044 human lncRNAs and systematic curation of lncRNAs' annotation by multi-omics data integration, function annotation and disease association	http://bigd.big.ac.cn/lncbook
MONOCLdb	The MOuse NOnCode Lung database provides the annotations and expression profiles of mouse long non-coding RNAs (lncRNAs) involved in influenza and SARS-CoVinfections.	http://www.monocldb.org/
NONCODE	**An integrated knowledge database dedicated to ncRNAs, especially lncRNAs**	**http://www.noncode.org/**
lncRNome	A comprehensive searchable biologically oriented knowledgebase for long noncoding RNAs in Humans.	http://genome.igib.res.in/lncRNome/
NRED	A database of long noncoding RNA expression.	http://nred.matticklab.com/cgi-bin/ncrnadb.pl
C-It-Loci	A tool to explore and to compare the expression profiles of conserved loci among various tissues in three organisms	http://c-it-loci.uni-frankfurt.de/
MiTranscriptome	A catalog of human long poly-adenylated RNA transcripts derived from computational analysis of high-throughput RNA-Seq data from over 6,500 samples, spanning diverse cancer and tissue types	http://mitranscriptome.org/
slncky Evolution Browser	This site contains alignments and evolutionary metrics of conserved lncRNAs.	https://scripts.mit.edu/~jjenny/
Cancer LncRNA Census (CLC)	Database of long-noncoding RNAs causally implicated in cancer through in vivo, in vitro and other evidence.	https://www.gold-lab.org/clc

NONCODE



NONCODE (current version v5.0) is an integrated knowledge database dedicated to non-coding RNAs (excluding tRNAs and rRNAs). Now, there are 17 species in NONCODE(human, mouse, cow, rat, chicken, fruitfly, zebrafish, celegans, yeast...). More

Jump to section for this gene/transcript

Using NONCODE databases

Browse NONCODE
Choose species and type, then browse all the entries.

Search a gene/transcipt
Search an entry or a subset of the database.

Statistics
Get the basic statistics of the NONCODE database.

Blast
Find regions of similarity between your sequences.

Genome
Find a transcript location in genome.

ID Conversion
Convert NONCODE ID and other databases ID.

Download
Download any of the information in NONCODE.

function
Provide predicted functions of gene.

News

6 Sep 2017:
NONCODE has updated to NONCODEv5. NONCODE2016 website has been moved to https://www.bioinfo.org/NONCODE2016

6 Apr 2017:
NONCODE has been maintained for several days.

23 Dec 2016:
New Species pig was added to NONCODE.

31 Aug 2015:
NONCODE were update to the NONCODE2016 version. Web pages were completly new designed.

26 Aug 2015:
Disease relation were added to the annotation. Users can browse the relative mutation or literature.

15 Aug 2015:
Conservation information were added to description of lncRNA genes.

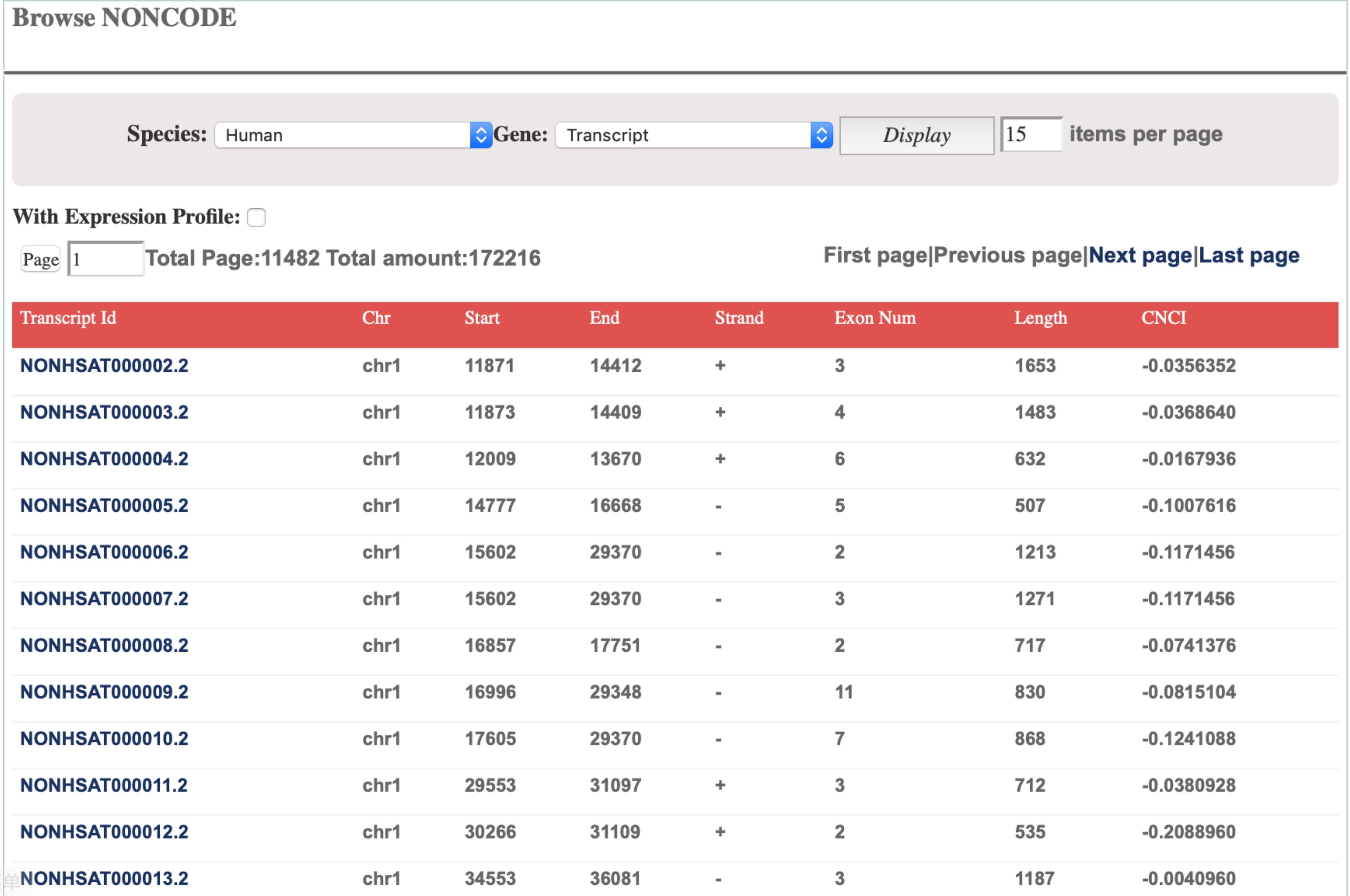
31 Jul 2015:
Literature data about lncRNA since 1 June 2015 were collect by NONCODE. Six new species were added.

more

Bio-tools

ncFANs

NONCODE



Browse NONCODE

Species: Human **Gene:** Transcript Display 15 items per page

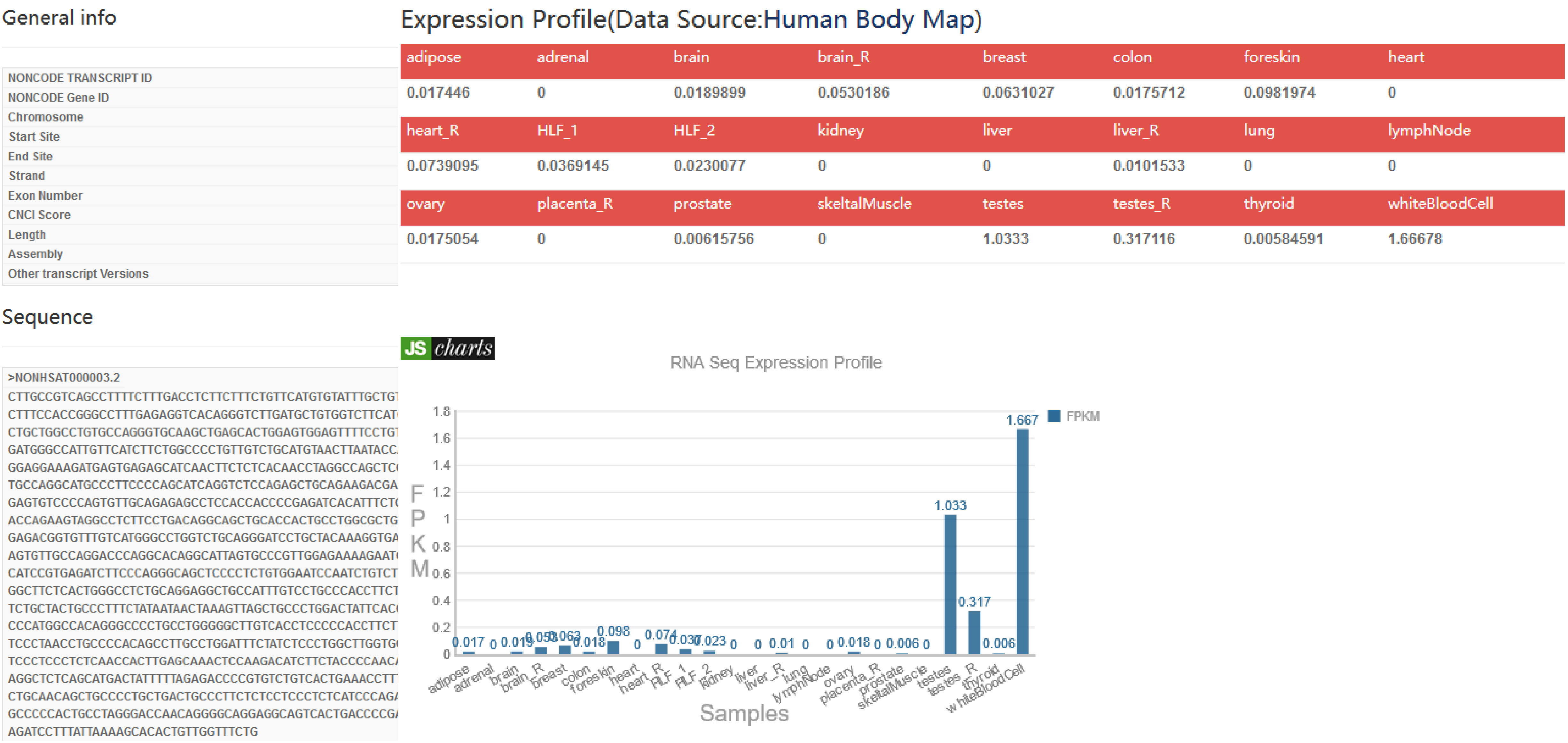
With Expression Profile: ☐

Page 1 Total Page:11482 Total amount:172216

First page|Previous page|Next page|Last page

Transcript Id	Chr	Start	End	Strand	Exon Num	Length	CNCI
NONHSAT000002.2	chr1	11871	14412	+	3	1653	-0.0356352
NONHSAT000003.2	chr1	11873	14409	+	4	1483	-0.0368640
NONHSAT000004.2	chr1	12009	13670	+	6	632	-0.0167936
NONHSAT000005.2	chr1	14777	16668	-	5	507	-0.1007616
NONHSAT000006.2	chr1	15602	29370	-	2	1213	-0.1171456
NONHSAT000007.2	chr1	15602	29370	-	3	1271	-0.1171456
NONHSAT000008.2	chr1	16857	17751	-	2	717	-0.0741376
NONHSAT000009.2	chr1	16996	29348	-	11	830	-0.0815104
NONHSAT000010.2	chr1	17605	29370	-	7	868	-0.1241088
NONHSAT000011.2	chr1	29553	31097	+	3	712	-0.0380928
NONHSAT000012.2	chr1	30266	31109	+	2	535	-0.2088960
NONHSAT000013.2	chr1	34553	36081	-	3	1187	-0.0040960

NONCODE



General info

NONCODE TRANSCRIPT ID
NONCODE Gene ID
Chromosome
Start Site
End Site
Strand
Exon Number
CNCI Score
Length
Assembly
Other transcript Versions

Sequence

```
>NONHSAT000003.2
CTTGCCGTCAGCCTTTTCTTTGACCTCTTCTTTCTGTTCATGTGTATTTGCTG
CTTTCCACCGGGCCTTTGAGAGGTCACAGGGTCTTGATGCTGTGGTCTTCAT
CTGCTGGCCTGTGCCAGGGTGCAAGCTGAGCACTGGAGTGGAGTTTTCCTG
GATGGGCCATTGTTCATCTTCTGGCCCCTGTTGTCTGCATGTAACTTAATACC
GGAGGAAAGATGAGTGAGAGCATCAACTTCTCTCACAACCTAGGCCAGCTC
TGCCAGGCATGCCCTTCCCCAGCATCAGGTCTCCAGAGCTGCAGAAGACGA
GAGTGTCCCCAGTGTTGCAGAGAGCCTCCACCACCCCGAGATCACATTTCT
ACCAGAAGTAGGCCTCTTCCTGACAGGCAGCTGCACCACTGCCTGGCGCTG
GAGACGGTGTTTGTCATGGGCCTGGTCTGCAGGGATCCTGCTACAAAGGTGA
AGTGTTGCCAGGACCCAGGCACAGGCATTAGTGCCCGTTGGAGAAAAGAAT
CATCCGTGAGATCTTCCCAGGGCAGCTCCCCTCTGTGGAATCCAATCTGTCT
GGCTTCTCACTGGGCCTCTGCAGGAGGCTGCCATTTGTCCTGCCCACCTTCT
TCTGCTACTGCCCTTTCTATAATAACTAAAGTTAGCTGCCCTGGACTATTCAC
CCCATGGCCACAGGGCCCCTGCCTGGGGGCTTGTCACCTCCCCCACCTTCT
TCCCTAACCTGCCCCACAGCCTTGCCTGGATTTCTATCTCCCTGGCTTGGTG
TCCCTCCCTCTCAACCACTTGAGCAAACTCCAAGACATCTTCTACCCCAAC
AGGCTCTCAGCATGACTATTTTTAGAGACCCCGTGTCTGTCACTGAAACCTT
CTGCAACAGCTGCCCCTGCTGACTGCCCTTCTCTCCTCCCTCTCATCCCAG
GCCCCCACTGCCTAGGGACCAACAGGGGCAGGAGGCAGTCACTGACCCCG
AGATCCTTTATTAAAAGCACACTGTTGGTTTCTG
```

Expression Profile(Data Source:Human Body Map)

adipose	adrenal	brain	brain_R	breast	colon	foreskin	heart
0.017446	0	0.0189899	0.0530186	0.0631027	0.0175712	0.0981974	0
heart_R	**HLF_1**	**HLF_2**	**kidney**	**liver**	**liver_R**	**lung**	**lymphNode**
0.0739095	0.0369145	0.0230077	0	0	0.0101533	0	0
ovary	**placenta_R**	**prostate**	**skeltalMuscle**	**testes**	**testes_R**	**thyroid**	**whiteBloodCell**
0.0175054	0	0.00615756	0	1.0333	0.317116	0.00584591	1.66678

NONCODE

iLncRNA(Identification of LncRNA)

Name:

Institute:

ACCESSIONs:

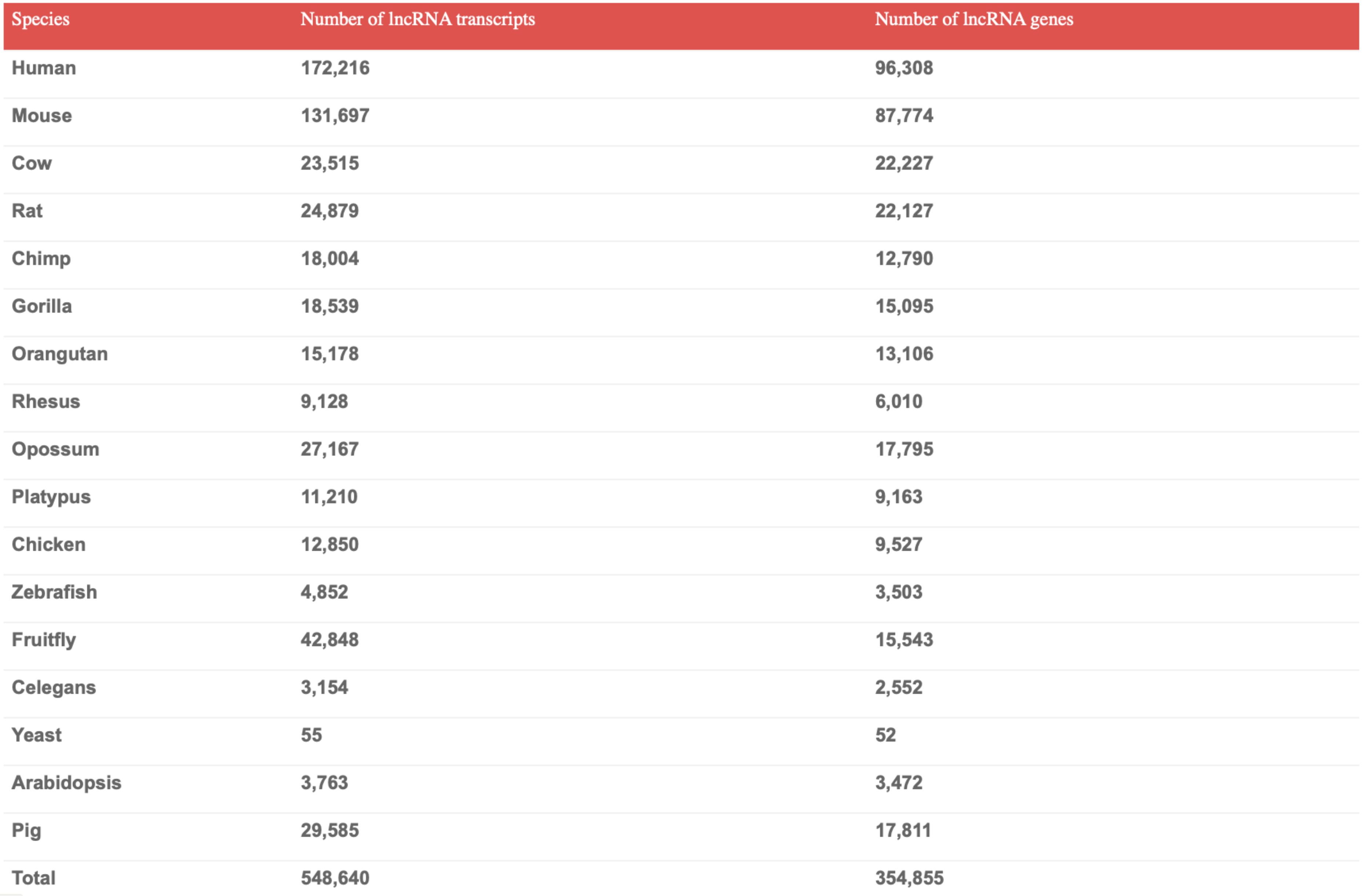
Organsim: Mouse(mm10)

Select your file:

○ **BED Format** 选取文件 未选择文件

○ **GTF Format** 选取文件 未选择文件

*E-mail:

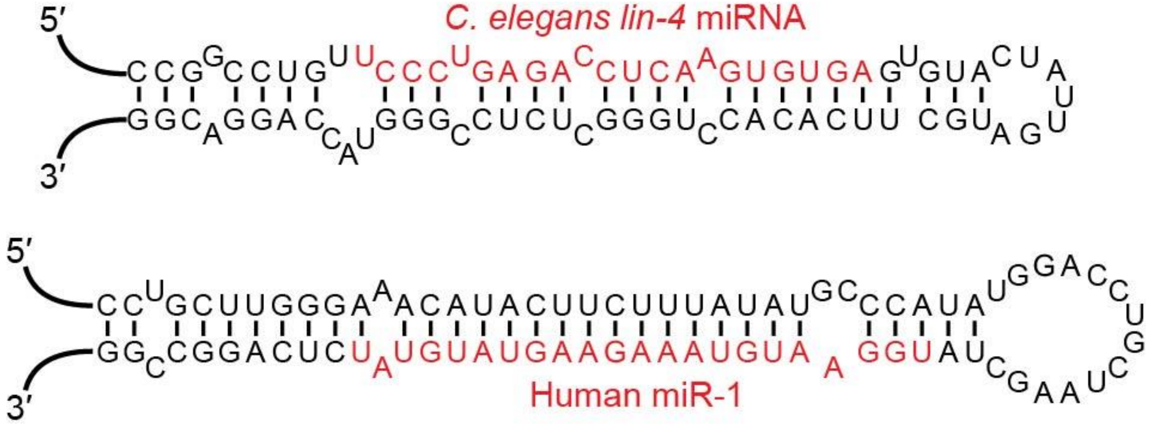


Species	Number of lncRNA transcripts	Number of lncRNA genes
Human	172,216	96,308
Mouse	131,697	87,774
Cow	23,515	22,227
Rat	24,879	22,127
Chimp	18,004	12,790
Gorilla	18,539	15,095
Orangutan	15,178	13,106
Rhesus	9,128	6,010
Opossum	27,167	17,795
Platypus	11,210	9,163
Chicken	12,850	9,527
Zebrafish	4,852	3,503
Fruitfly	42,848	15,543
Celegans	3,154	2,552
Yeast	55	52
Arabidopsis	3,763	3,472
Pig	29,585	17,811
Total	548,640	354,855

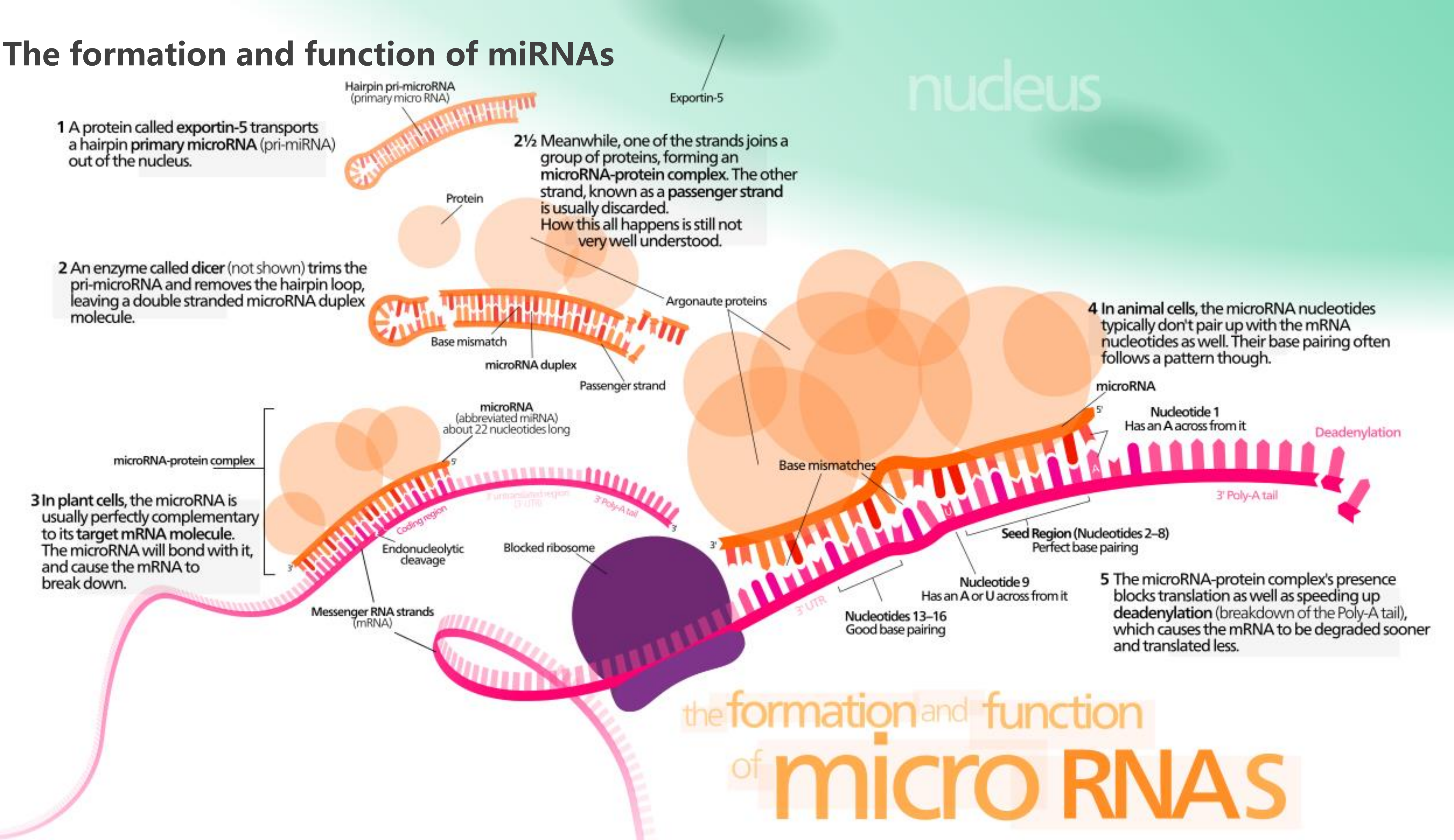
什么是小RNA（sRNA）

Small RNA are <200 nt (nucleotide) in length, and are usually non-coding RNA molecules.

- Small RNA
  - **microRNA (miRNA)**
  - Piwi-interacting RNA (piRNA)
  - small interfering RNA (siRNA)
  - small nucleolar RNA (snoRNAs)
  - tRNA-derived small RNA (tsRNA)
  - small rDNA-derived RNA (srRNA)
  - small nuclear RNA



The formation and function of miRNAs



1 A protein called **exportin-5** transports a hairpin **primary microRNA** (pri-miRNA) out of the nucleus.

2 An enzyme called **dicer** (not shown) trims the pri-microRNA and removes the hairpin loop, leaving a double stranded microRNA duplex molecule.

2½ Meanwhile, one of the strands joins a group of proteins, forming an **microRNA-protein complex**. The other strand, known as a **passenger strand** is usually discarded.
How this all happens is still not very well understood.

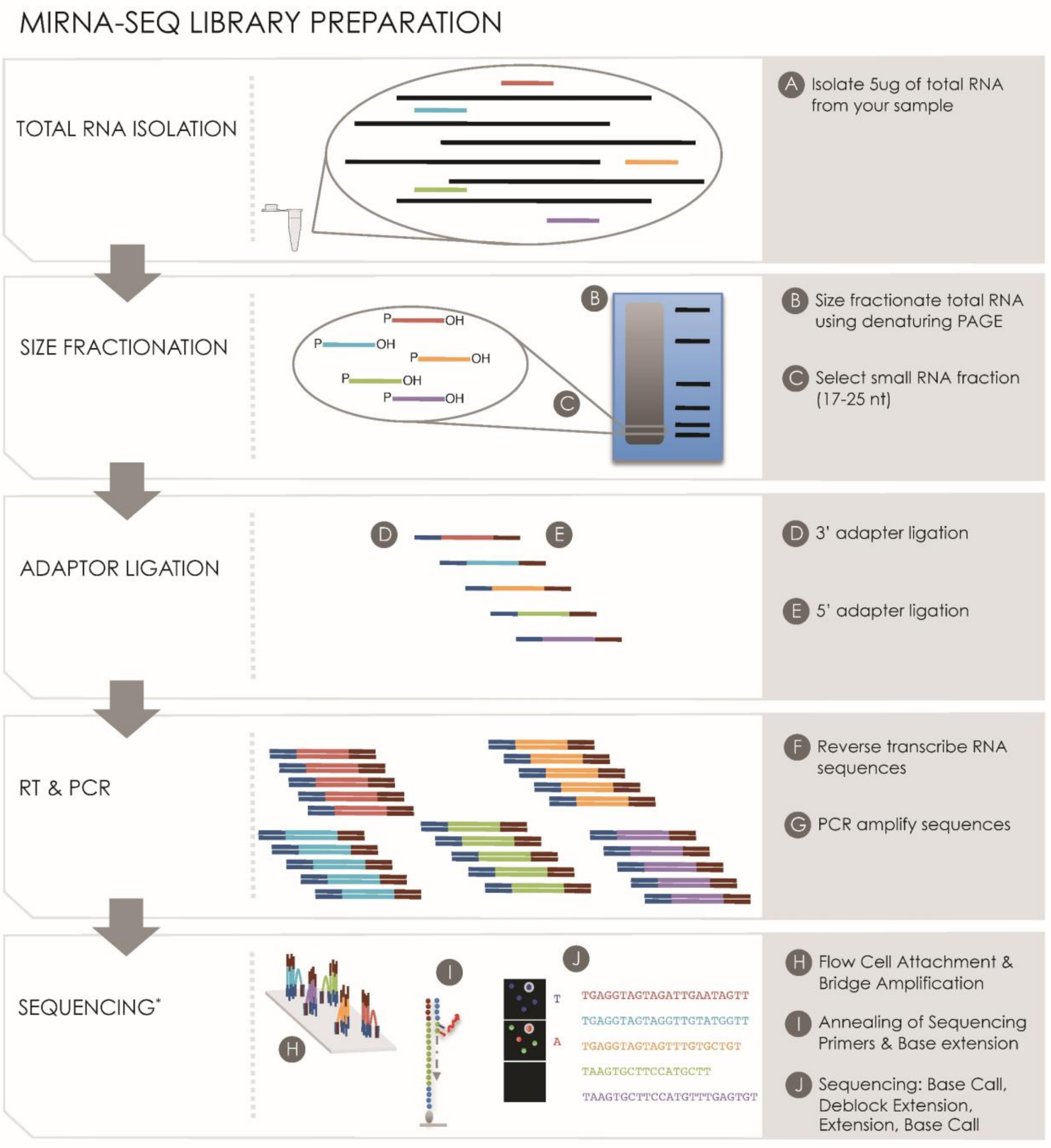
3 **In plant cells**, the microRNA is usually perfectly complementary to its **target mRNA molecule**. The microRNA will bond with it, and cause the mRNA to break down.

4 **In animal cells**, the microRNA nucleotides typically don't pair up with the mRNA nucleotides as well. Their base pairing often follows a pattern though.

5 The microRNA-protein complex's presence blocks translation as well as speeding up **deadenylation** (breakdown of the Poly-A tail), which causes the mRNA to be degraded sooner and translated less.

miRNA-seq建库测序流程

- Total RNA isolation
- Size fractionation
- Adapter ligation
- RT & PCR
- Sequencing



A Isolate 5ug of total RNA from your sample

B Size fractionate total RNA using denaturing PAGE

C Select small RNA fraction (17-25 nt)

D 3' adapter ligation

E 5' adapter ligation

F Reverse transcribe RNA sequences

G PCR amplify sequences

H Flow Cell Attachment & Bridge Amplification

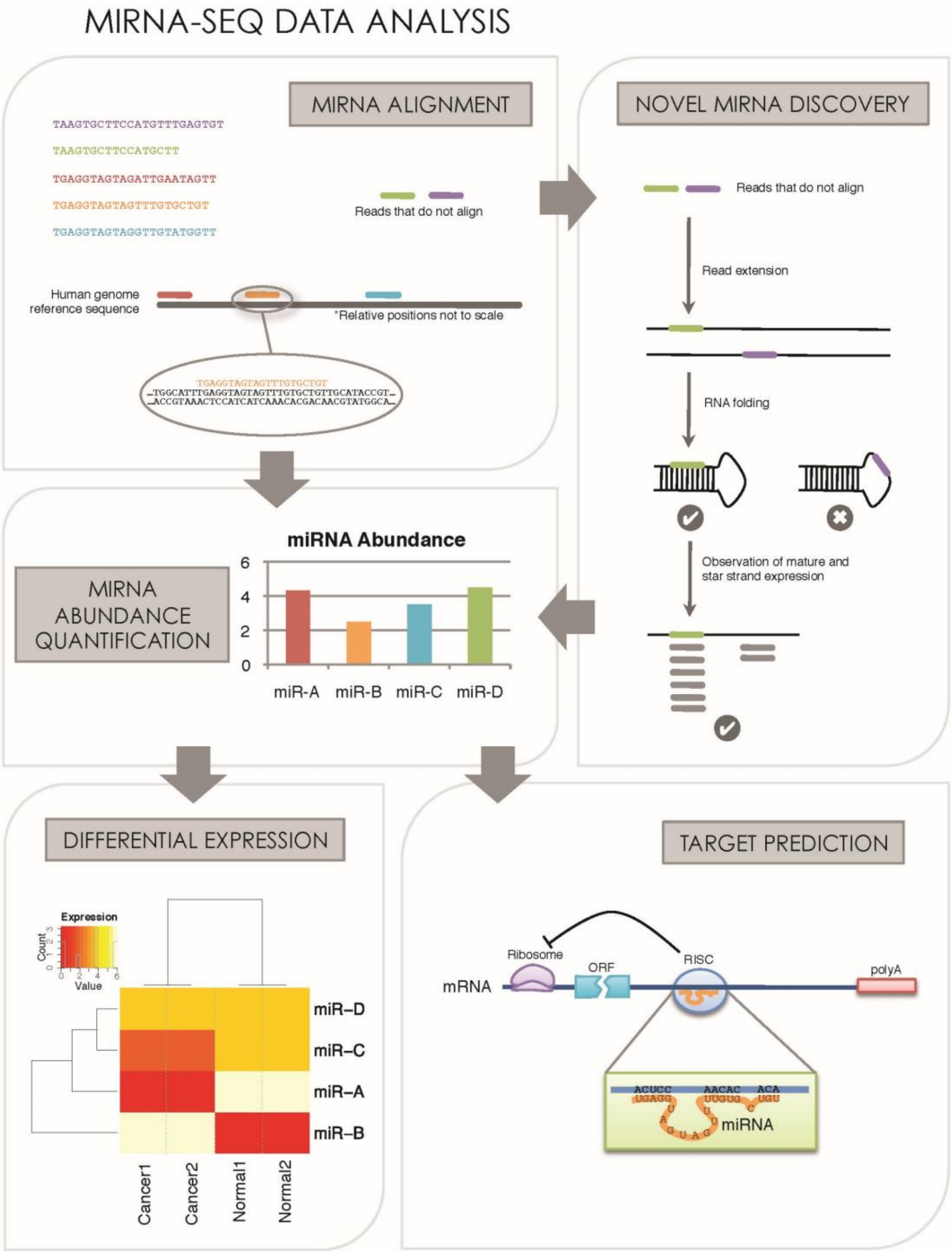
I Annealing of Sequencing Primers & Base extension

J Sequencing: Base Call, Deblock Extension, Extension, Base Call

*Illumina sequencing method depicted however other sequencing platforms can also be used.

miRNA分析流程

- Raw data quality control
- Reads that matched rRNA, tRNA, snRNA, snoRNA, repeat sequences, and other ncRNAs deposited in Rfam (http://www.sanger.ac.uk/software/Rfam) and the GenBank noncoding RNA database (http://www.ncbi.nlm.nih.gov/) were discarded
- Map to the known miRNA precursor sequences in miRBase Version 22
- Unmapped reads map to the genome to identify novel miRNAs
- Downstream analysis

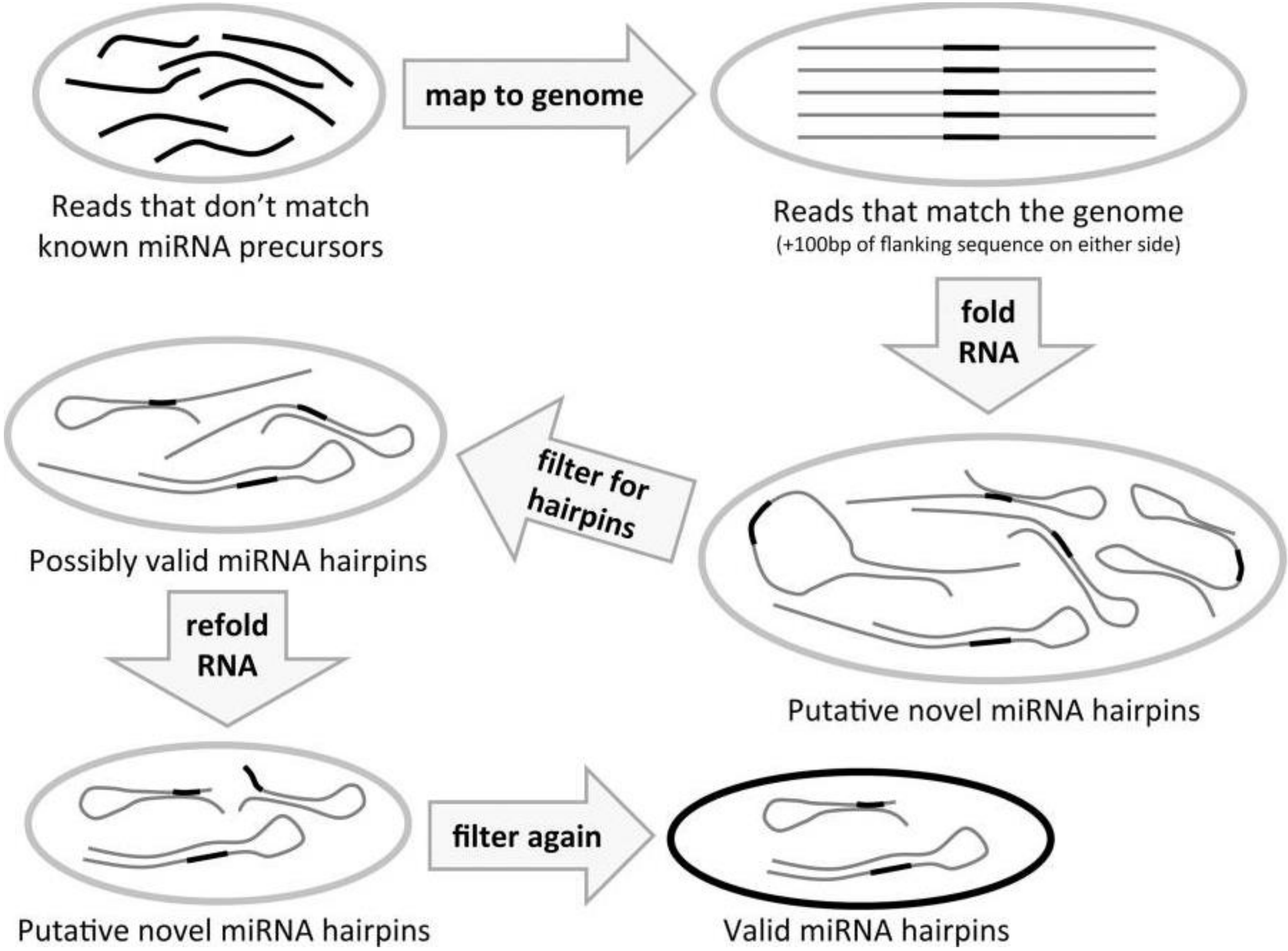


Novel miRNA discovery pipeline

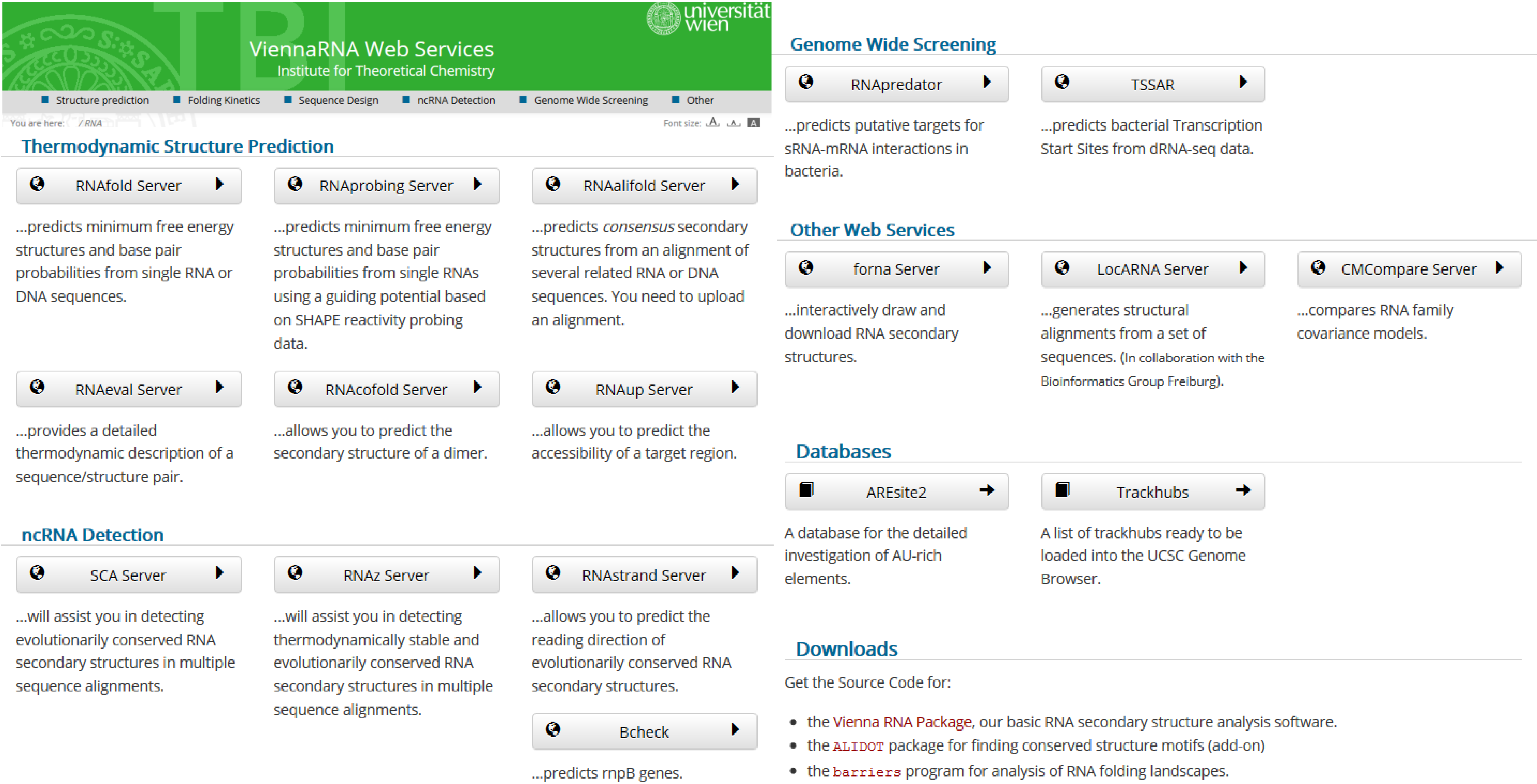
1、Obtain reads that did not align to known miRNA sequences, and map them to the genome.

2、RNA Folding Method

- Vienna package
- minimum free energy less than ~25kcal/mol



ViennaRNA Web Services



Thermodynamic Structure Prediction

RNAfold Server

...predicts minimum free energy structures and base pair probabilities from single RNA or DNA sequences.

RNAprobing Server

...predicts minimum free energy structures and base pair probabilities from single RNAs using a guiding potential based on SHAPE reactivity probing data.

RNAalifold Server

...predicts *consensus* secondary structures from an alignment of several related RNA or DNA sequences. You need to upload an alignment.

RNAeval Server

...provides a detailed thermodynamic description of a sequence/structure pair.

RNAcofold Server

...allows you to predict the secondary structure of a dimer.

RNAup Server

...allows you to predict the accessibility of a target region.

ncRNA Detection

SCA Server

...will assist you in detecting evolutionarily conserved RNA secondary structures in multiple sequence alignments.

RNAz Server

...will assist you in detecting thermodynamically stable and evolutionarily conserved RNA secondary structures in multiple sequence alignments.

RNAstrand Server

...allows you to predict the reading direction of evolutionarily conserved RNA secondary structures.

Bcheck

...predicts rnpB genes.

Genome Wide Screening

RNApredator

...predicts putative targets for sRNA-mRNA interactions in bacteria.

TSSAR

...predicts bacterial Transcription Start Sites from dRNA-seq data.

Other Web Services

forna Server

...interactively draw and download RNA secondary structures.

LocARNA Server

...generates structural alignments from a set of sequences. (In collaboration with the Bioinformatics Group Freiburg).

CMCompare Server

...compares RNA family covariance models.

Databases

AREsite2

A database for the detailed investigation of AU-rich elements.

Trackhubs

A list of trackhubs ready to be loaded into the UCSC Genome Browser.

Downloads

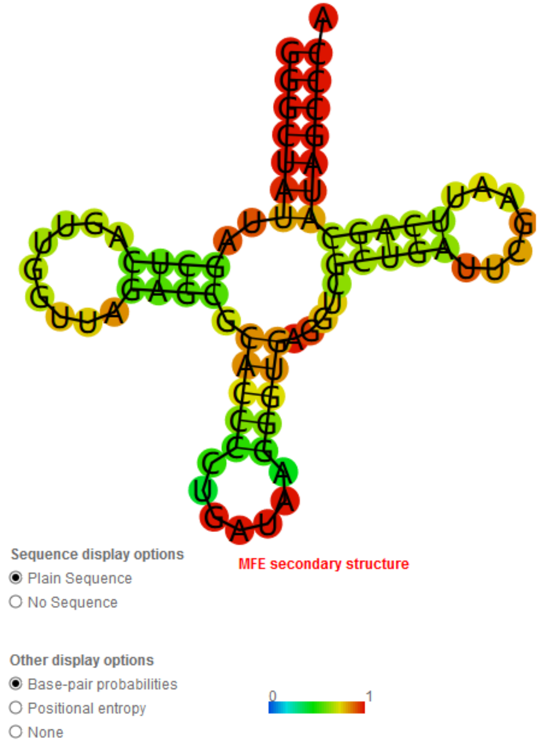
Get the Source Code for:

- the Vienna RNA Package, our basic RNA secondary structure analysis software.
- the `ALIDOT` package for finding conserved structure motifs (add-on)
- the `barriers` program for analysis of RNA folding landscapes.

ViennaRNA Web Services——RNAfold

The RNAfold web server will predict secondary structures of single stranded RNA or DNA sequences.

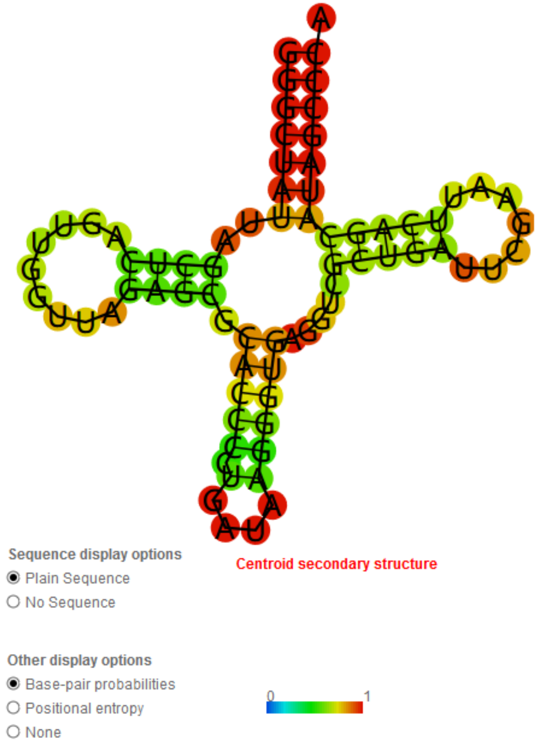
http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi



MFE secondary structure

Sequence display options
● Plain Sequence
○ No Sequence

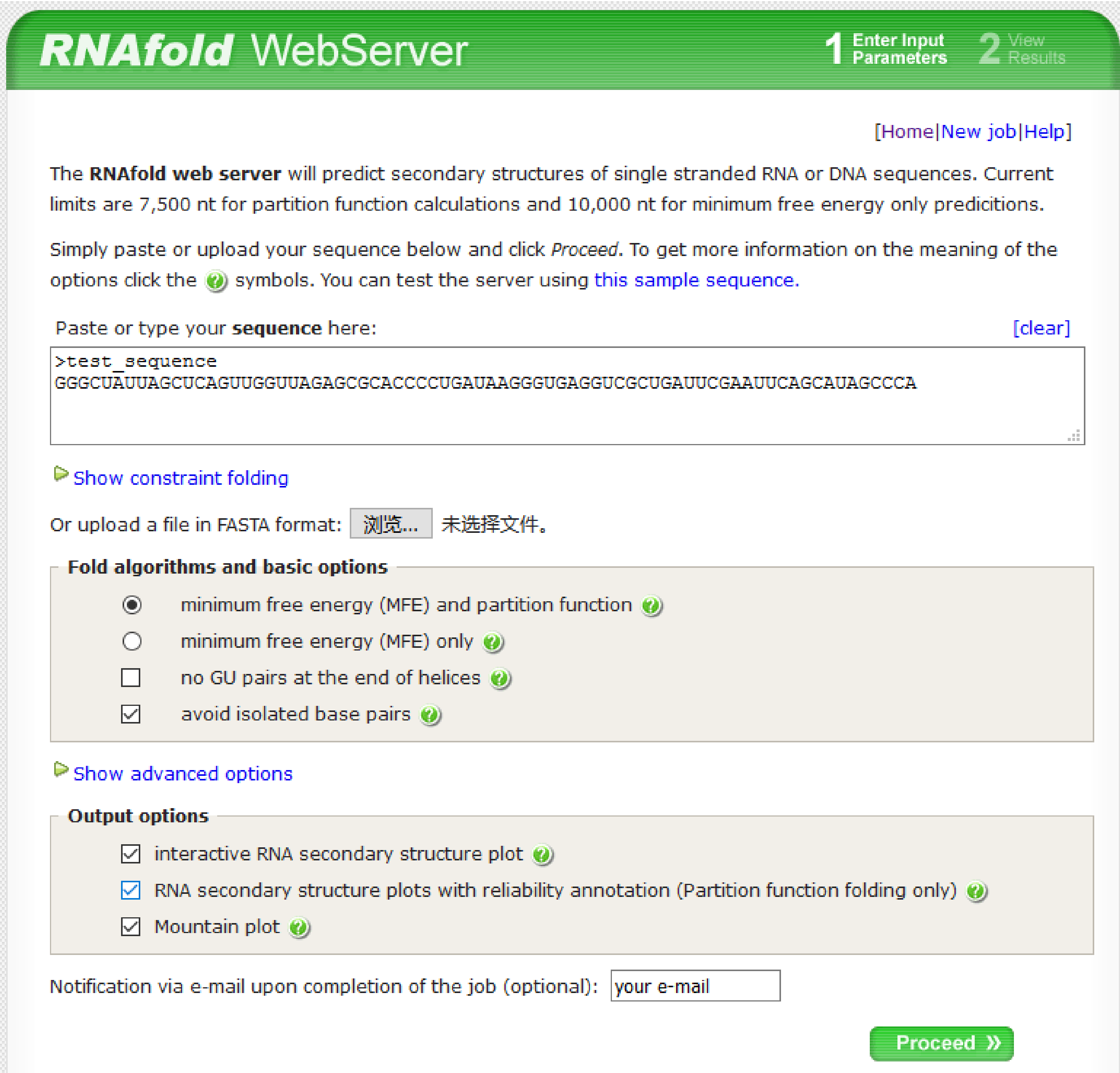
Other display options
● Base-pair probabilities
○ Positional entropy
○ None



Centroid secondary structure

Sequence display options
● Plain Sequence
○ No Sequence

Other display options
● Base-pair probabilities
○ Positional entropy
○ None



The **RNAfold web server** will predict secondary structures of single stranded RNA or DNA sequences. Current limits are 7,500 nt for partition function calculations and 10,000 nt for minimum free energy only predicitions.

Simply paste or upload your sequence below and click *Proceed*. To get more information on the meaning of the options click the symbols. You can test the server using this sample sequence.

Paste or type your **sequence** here: [clear]

```
>test_sequence
GGGCUAUUAGCUCAGUUGGUUAGAGCGCACCCCUGAUAAGGGUGAGGUCGCUGAUUCGAAUUCAGCAUAGCCCA
```

Show constraint folding

Or upload a file in FASTA format: 浏览... 未选择文件。

Fold algorithms and basic options

● minimum free energy (MFE) and partition function
○ minimum free energy (MFE) only
☐ no GU pairs at the end of helices
☑ avoid isolated base pairs

Show advanced options

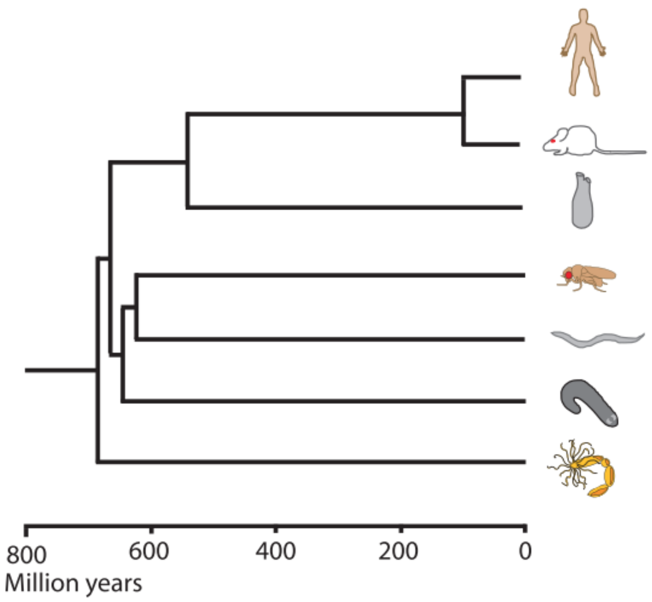
Output options

☑ interactive RNA secondary structure plot
☑ RNA secondary structure plots with reliability annotation (Partition function folding only)
☑ Mountain plot

Notification via e-mail upon completion of the job (optional): your e-mail

miRNA identification tool——miRDeep2

Identify known and novel microRNAs



organism	known miRNAs in data	sensitivity (miRNA loci correctly classified / miRNA loci)	specificity (non-miRNA loci correctly classified / non-miRNA loci)	prevalence (miRNA loci / (miRNA + non-miRNA loci))	accuracy (sensitivity • prevalence + specificity • (1 - prevalence))	novel miRNAs reported	novel miRNAs, estimated true positives	novel miRNAs, estimated true positive rate	novel miRNAs, both mature and star strands detected in a least two samples	total miRNAs detected (known and novel estimated true positives)
Human (liver)	455	79%	99.3%	2.3%	98.9%	124	75 ± 7	60 ± 6%	15	434
Human (cell lines)	631	81%	99.2%	3.2%	98.6%	147	126 ± 5	86 ± 3%	27	634
Mouse	586	82%	99.7%	1.5%	99.4%	104	85 ± 5	82 ± 5%	23	563
Sea squirt	441	71%	99.6%	0.9%	99.4%	124	55 ± 6	44 ± 5%	24	370
Fruitfly	152	78%	99.9%	0.4%	99.9%	8	1 ± 1	15 ± 15%	2	120
Nematode	154	82%	99.9%	0.4%	99.8%	32	20 ± 3	62 ± 11%	9	147
Planaria	156	88%	99.9%	0.5%	99.9%	12	11 ± 1	89 ± 4%	0	148
Anemone	40	90%	99.8%	0.2%	99.8%	28	25 ± 1	90 ± 4%	-	61

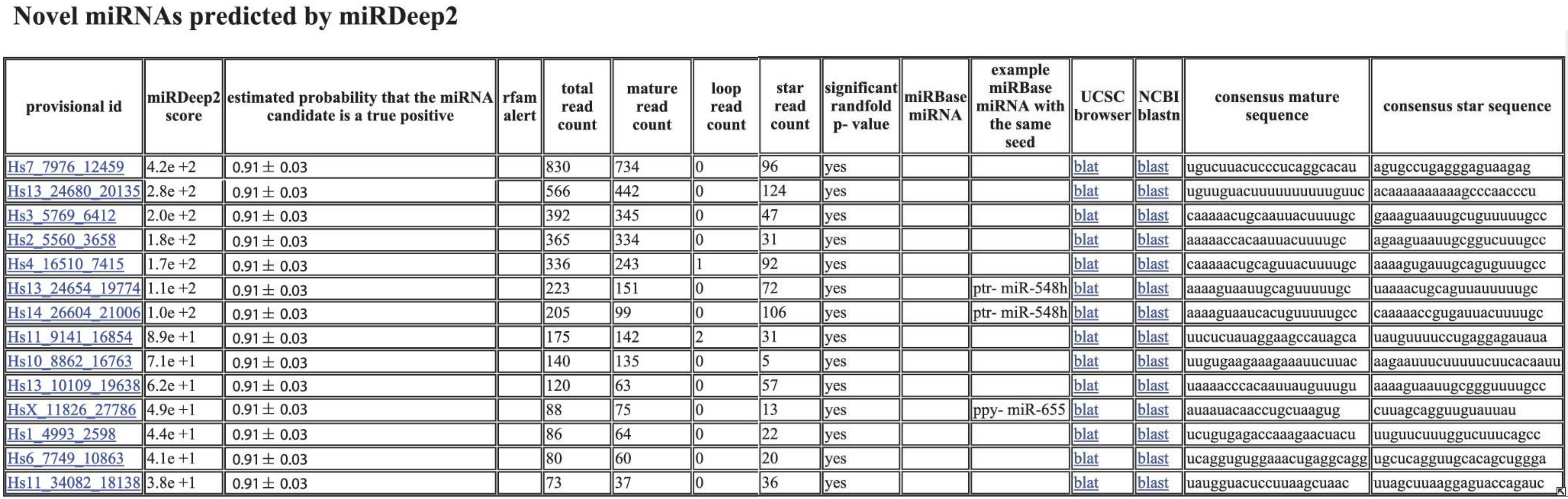
miRNA identification tool——miRDeep2 Usage

```
# 建立搜索索引，cel_cluster.fa是基因组文件，cel_cluster是index文件的索引
bowtie-build cel_cluster.fa cel_cluster

# Reads比对
perl mapper.pl reads.fa -c -j -k TCGTATGCCGTCTTCTGCTTGT -l 18 -m -p cel_cluster -s
reads_collapsed.fa -t reads_collapsed_vs_genome.arf -v

-c 指出输入文件是fasta格式，同类的参数还有-a(seq.txt format),-b(qseq.txt format),-e(fastq
format),-d (contig file)
-j 删除不规范的字母（不规范的字母是指除a,c,g,t,u,n,A,C,G,T,U,N之外的字母）
-k 剪切接头，后跟接头序列，例子中的TCGTATGCCGTCTTCTGCTTGT就是接头
-l 忽视小于某长度的序列，例子中忽视18nt长度的reads
-p 将处理过的reads map到之前建立过索引的基因组上，例子中的cel_cluster
-s 指出将处理过的reads输出到某个文件，例子中将处理过的reads输出到reads_collapsed.fa
-t 指出将mapping的结果输出到某个文件，例子中将mapping后的结果输出到reads_collapsed_vs_genome.arf文件
中
-v 在屏幕上显示处理的动作
```

miRNA identification tool——miRDeep2 Usage

```
# 比对到miRBase数据库中的前体序列进行定量分析
quantifier.pl -p precursors_ref_this_species.fa -m mature_ref_this_species.fa -r
reads_collapsed.fa  -t cel

-p miRNA前体文件，miRBase可以下载
-m 成熟miRNA序列文件，miRBase可以下载
-r reads文件
-t 物种，可以指定某个物种，这样分析的时候只考虑某个物种的数据。也可以不指定,分析所有的

# 在deep sequencing data中鉴定已知和未知的miRNA
miRDeep2.pl reads_collapsed.fa cel_cluster.fa reads_collapsed_vs_genome.arf
mature_ref_this_species.fa mature_ref_other_species.fa precursors_ref_this_species.fa -t
C.elegans 2> report.log

# reads_collapsed.fa是经过mapper.pl处理的reads，cel_cluster.fa是基因组文件，
reads_collapsed_vs_genome.arf mapping的结果，mature_ref_this_species.fa物种的成熟miRNA文件，
mature_ref_other_species.fa其他物种相关的成熟miRNA文件，precursors_ref_this_species.fa物种miRNA
前体的文件；如果你只有reads，arf文件，genome文件，其他文件没有，需要这样表示miRNAs_ref/none
miRNAs_other/none、precursors/none。
```

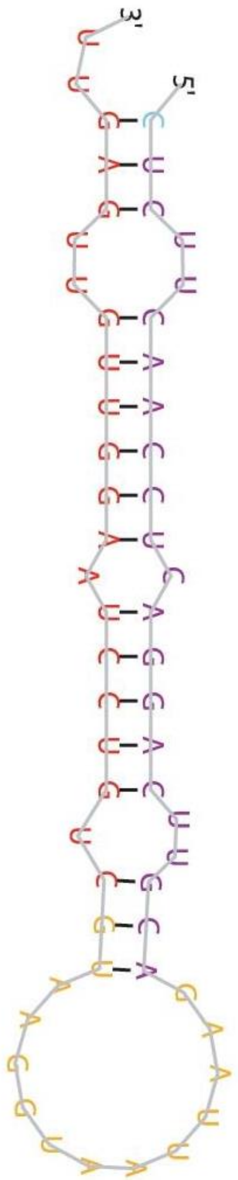
miRNA identification tool——miRDeep2 result

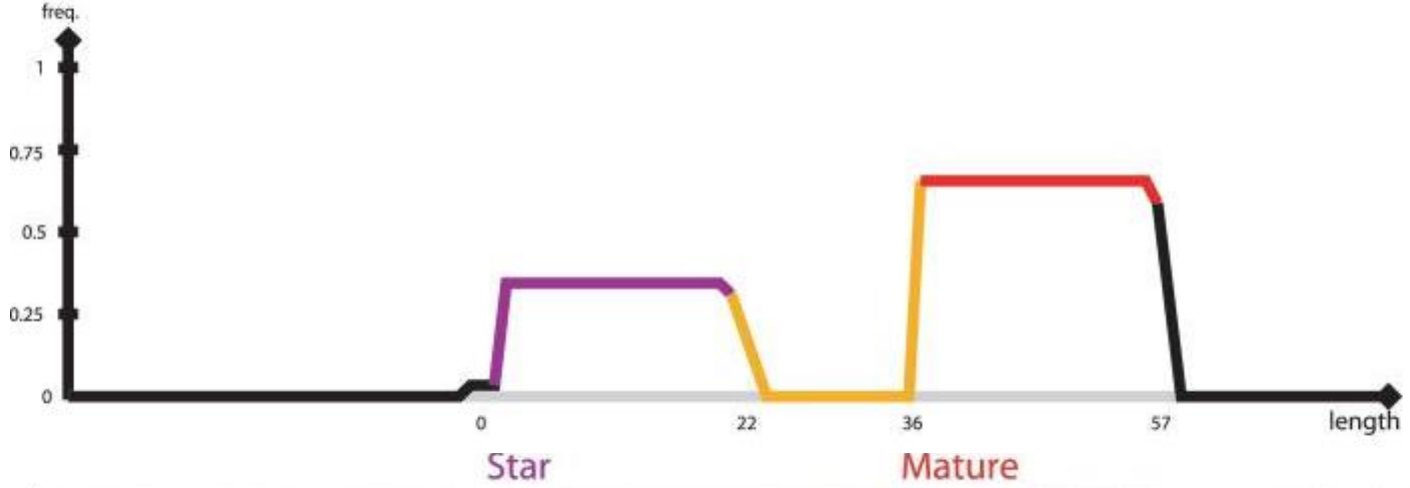


Novel miRNAs predicted by miRDeep2

provisional id	miRDeep2 score	estimated probability that the miRNA candidate is a true positive	rfam alert	total read count	mature read count	loop read count	star read count	significant randfold p- value	miRBase miRNA	example miRBase miRNA with the same seed	UCSC browser	NCBI blastn	consensus mature sequence	consensus star sequence
Hs7_7976_12459	4.2e +2	0.91 ± 0.03		830	734	0	96	yes			blat	blast	ugucuuacucccucaggcacau	agugccugagggaguaagag
Hs13_24680_20135	2.8e +2	0.91 ± 0.03		566	442	0	124	yes			blat	blast	uguuguacuuuuuuuuuuuguuc	acaaaaaaaaaagcccaacccu
Hs3_5769_6412	2.0e +2	0.91 ± 0.03		392	345	0	47	yes			blat	blast	caaaaacugcaauuacuuuugc	gaaaguaauugcuguuuuugcc
Hs2_5560_3658	1.8e +2	0.91 ± 0.03		365	334	0	31	yes			blat	blast	aaaaaccacaauuacuuuugc	agaaguaauugcggucuuugcc
Hs4_16510_7415	1.7e +2	0.91 ± 0.03		336	243	1	92	yes			blat	blast	caaaaacugcaguuacuuuugc	aaaagugauugcaguguuugcc
Hs13_24654_19774	1.1e +2	0.91 ± 0.03		223	151	0	72	yes		ptr- miR-548h	blat	blast	aaaaguaauugcaguuuuugc	uaaaacugcaguuauuuuugc
Hs14_26604_21006	1.0e +2	0.91 ± 0.03		205	99	0	106	yes		ptr- miR-548h	blat	blast	aaaaguaaucacuguuuuugcc	caaaaaccgugauuacuuuugc
Hs11_9141_16854	8.9e +1	0.91 ± 0.03		175	142	2	31	yes			blat	blast	uucucuauaggaagccauagca	uauguuuuccugaggagauaua
Hs10_8862_16763	7.1e +1	0.91 ± 0.03		140	135	0	5	yes			blat	blast	uugugaagaaagaaauucuuac	aagaauuucuuuuucuucacaauu
Hs13_10109_19638	6.2e +1	0.91 ± 0.03		120	63	0	57	yes			blat	blast	uaaaacccacaauuauguuugu	aaaaguaauugcggguuuugcc
HsX_11826_27786	4.9e +1	0.91 ± 0.03		88	75	0	13	yes		ppy- miR-655	blat	blast	auaauacaaccugcuaagug	cuuagcagguuguauuau
Hs1_4993_2598	4.4e +1	0.91 ± 0.03		86	64	0	22	yes			blat	blast	ucugugagaccaaagaacuacu	uuguucuuuggucuuucagcc
Hs6_7749_10863	4.1e +1	0.91 ± 0.03		80	60	0	20	yes			blat	blast	ucagguguggaaacugaggcagg	ugcucagguugcacagcuggga
Hs11_34082_18138	3.8e +1	0.91 ± 0.03		73	37	0	36	yes			blat	blast	uaugguacuccuuaagcuaac	uuagcuuaaggaguaccagauc

miRNA identification tool——miRDeep2 result





```
5'- agacuucauaucagauucucaccugaacgcaugacucuucaaccucaggacuugcagaauuaauggaaugcuguccuaagguuguugaguugugcauuucugggcauuuc  -3'   obs
    agacuucauaucagauucucaccugaacgcaugacucuucaaccucaggacuugcagaauuaauggaaugcuguccuaagguuguugaguugugcauuucugggcauuuc         exp
    ..........(((((.((......))..(((..((((..((((((.(((((..(((............))).))))).))))))..))))..)))...)))))......   reads   mm   sample
    ...................................ucuucaaccucaggacuug........................................................   1       0    NL2
    ...................................ucuucaaccucaggacuugc.......................................................   1       0    NL2
    ...................................ucuucaaccucaggacuugca......................................................   2       0    NL2
    ......................................................................cuguccuaagguuguugagu....................   2       0    NL2
    ......................................................................cuguccuaagguuguugaguu...................   4       0    NL2
    ......................................................................cuguccuaagguuguugaguuA..................   1       1    NL2
    ......................................................................cuguccuaagguuguugaguuU..................   1       1    NL2
    ......................................................................cuguccuaagguuguugaguug..................   1       0    NL2

    ................................gacucuucaaccucaggacuugc.......................................................   1       0    NL3
    ...................................ucuucaaccucaggacuugca......................................................   1       0    NL3
    ...................................ucuucaaccucaggacuugcaA.....................................................   1       1    NL3
    ......................................................................cuguccuaagguuguugaguu...................   2       0    NL3
    ......................................................................cuguccuaagguuguugaguug..................   1       0    NL3
    ......................................................................cuguccuaagguuguugaguuU..................   1       1    NL3
    ......................................................................cuguccuaagguuguugaguuA..................   2       1    NL3

    ...................................ucuucaaccucaggacuugc.......................................................   1       0    NL1
    ...................................ucuucaaccucaggacuugca......................................................   1       0    NL1
    ...................................ucuucaaccucaggacuugcaA.....................................................   1       1    NL1
    ......................................................................cuguccuaagguuguugaguu...................   2       0    NL1
    ......................................................................cuguccuaagguuguugaguug..................   2       0    NL1
```

miRNA identification tool——miRDeep-P

accurate and fast analysis of the microRNA transcriptome in plants

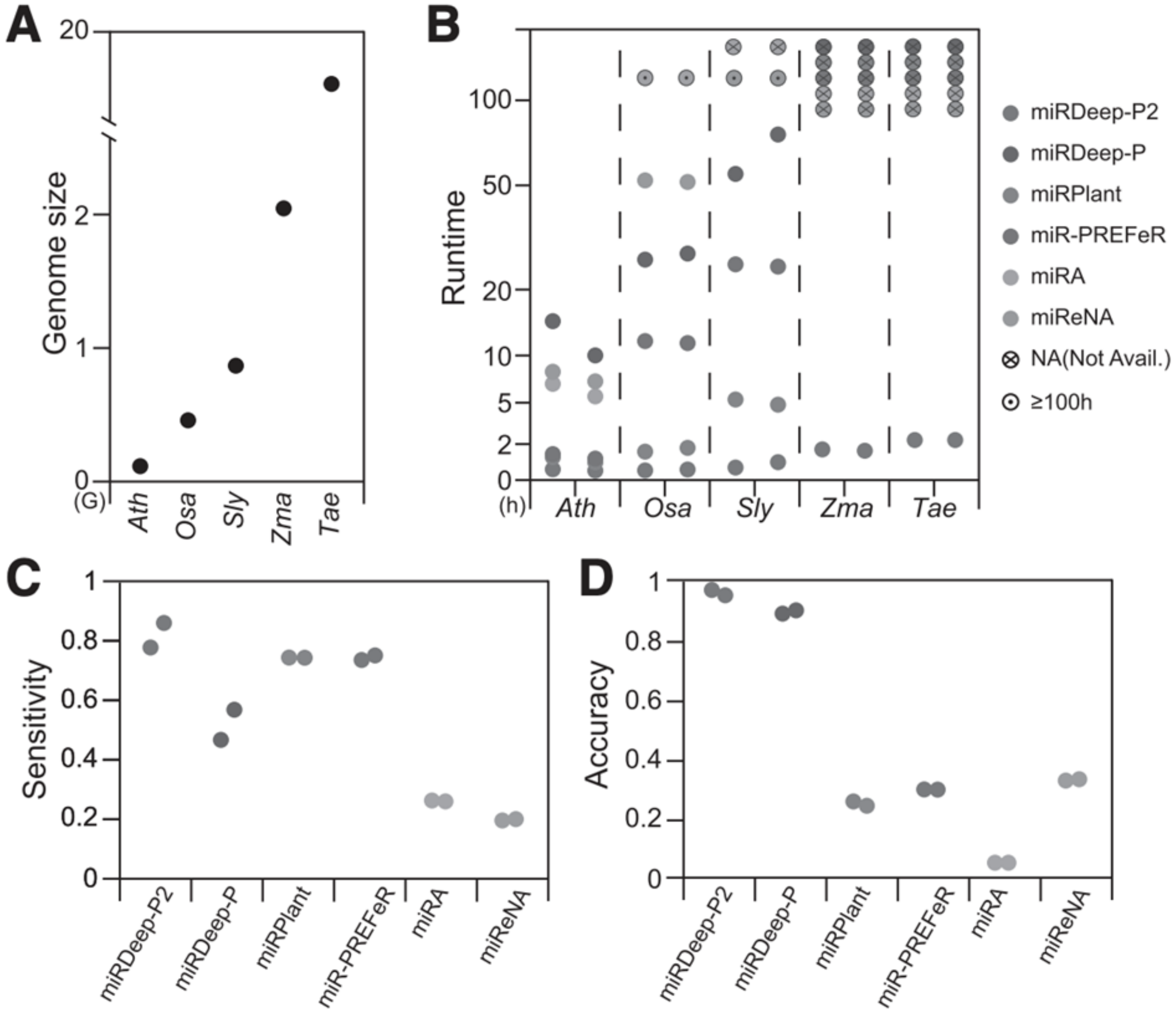


Fig. 1. Performance of miRDP2. **(A)** Genome size (in Gb) of *Arabidopsis thaliana*

```
# 分别为参考基因组和和非编码RNA的Rfam建立索引
bowtie-build --threads 8 -f ./Data/TAIR10_genome.fa ./Data/TAIR10_genome
# 为Rfam建立索引，一定得在流程的script/index 目录下
bowtie-build --threads 20 -f ./miRDP2-v1.1.2/scripts/rfam_index

# 运行分析
miRDP2-v1.1.2_pipeline.bash -g ./Data/TAIR10_genome.fa -i ./Data/TAIR10_genome -f ./Data/GSM2094927.fa -o .
# -g: 基因组序列
# -i: bowtie索引
# -f sRNA-seq 文件
# -o 输出文件夹

# 结果
“染色体编号”, “所在链”, “代表性的短读编号”, “前体编号”, “成熟miRNA位置”, “前体位置”, “成熟序列”, “前体序列 ”
```

miRNA target prediction——miranda

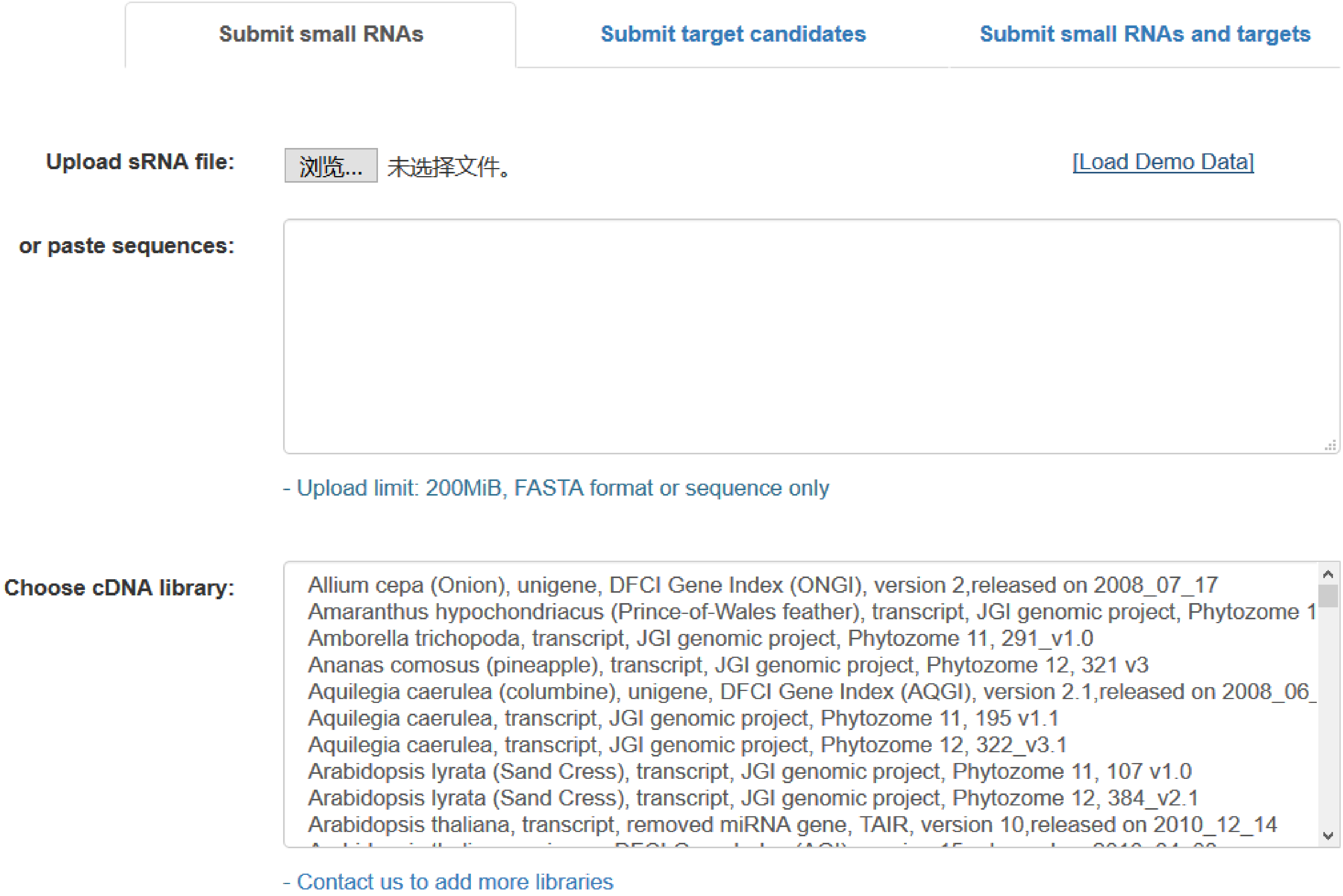
microRNA Target Scanning Algorithm

```
Usage:  miranda file1 file2 [options..]
'file1' is a FASTA file with a microRNA query
'file2' is a FASTA file containing the sequence(s) to be scanned.
```

```
Read Sequence:rno-miR-26b-3p MIMAT0004714(22 nt)
Read Sequence:lphn3 (418 nt)
=============================================================
Performing Scan: rno-miR-26b-3p vs lphn3
=============================================================
   Forward:      Score: 160.000000  Q:2 to 21  R:313 to 334 Align Len (19) (78.95%) (84.21%)
   Query:    3' cuCGGUUCAUUACCUCUUGUCc 5'
                  ||||| ||  ||||:||||
   Ref:      5' gcGCCAACTACGGGAGGACAGa 3'

   Energy:  -22.370001 kCal/Mol
Scores for this hit:
>rno-miR-26b-3p  lphn3    160.00  -22.37   2 21    313 334  19      78.95%   84.21%
Score for this Scan:
Seq1,Seq2,Tot Score,Tot Energy,Max Score,Max Energy,Strand,Len1,Len2,Positions
>>rno-miR-26b-3p lphn3    160.00  -22.37   160.00  -22.37  41      22      418      313
Complete
```

miRNA target prediction—psRNATarget (online)

https://plantgrn.noble.org/psRNATarget/analysis



Upload sRNA file: 未选择文件。

or paste sequences:

- Upload limit: 200MiB, FASTA format or sequence only

Choose cDNA library:

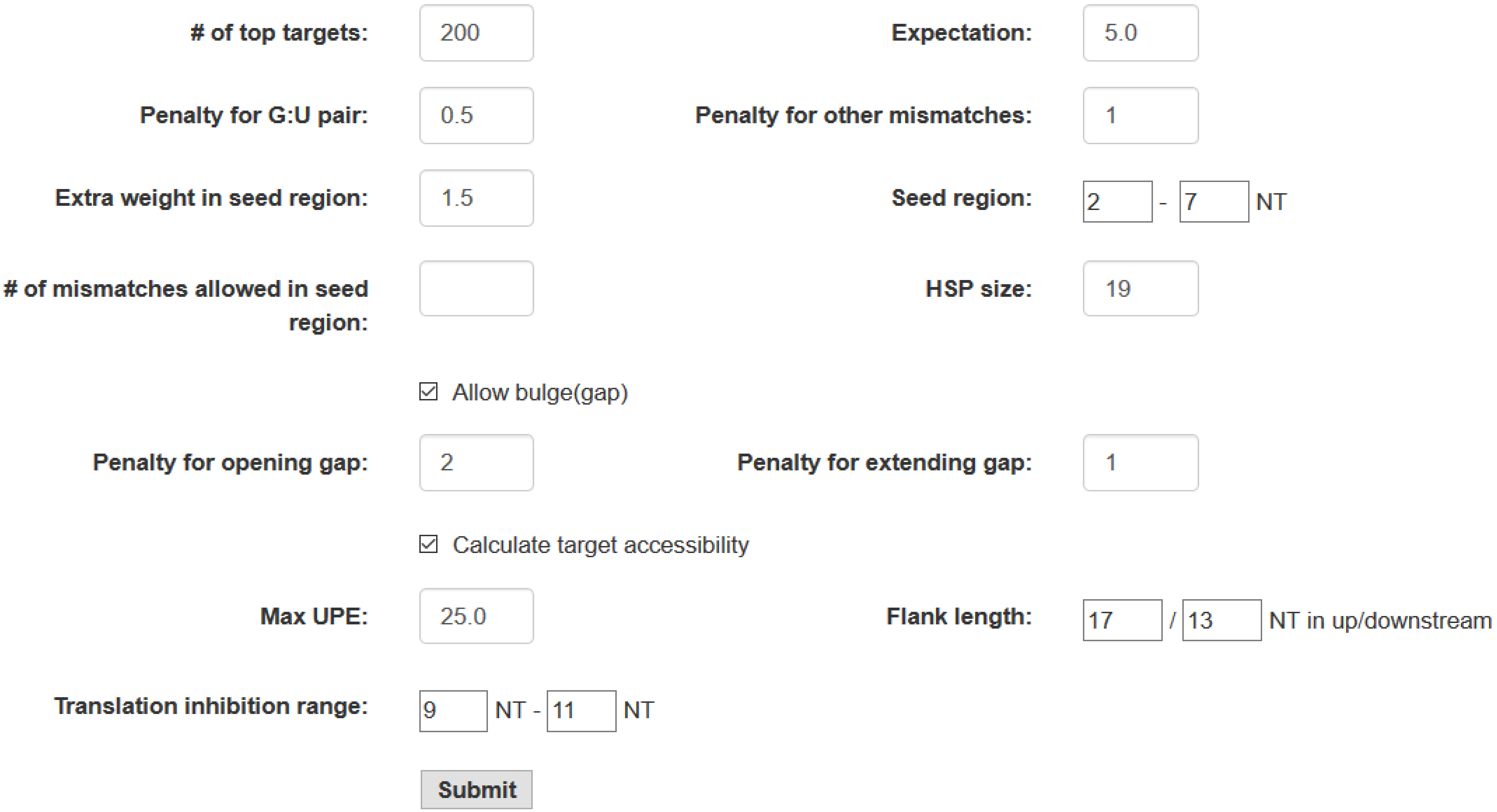
Allium cepa (Onion), unigene, DFCI Gene Index (ONGI), version 2,released on 2008_07_17
Amaranthus hypochondriacus (Prince-of-Wales feather), transcript, JGI genomic project, Phytozome 1
Amborella trichopoda, transcript, JGI genomic project, Phytozome 11, 291_v1.0
Ananas comosus (pineapple), transcript, JGI genomic project, Phytozome 12, 321 v3
Aquilegia caerulea (columbine), unigene, DFCI Gene Index (AQGI), version 2.1,released on 2008_06_
Aquilegia caerulea, transcript, JGI genomic project, Phytozome 11, 195 v1.1
Aquilegia caerulea, transcript, JGI genomic project, Phytozome 12, 322_v3.1
Arabidopsis lyrata (Sand Cress), transcript, JGI genomic project, Phytozome 11, 107 v1.0
Arabidopsis lyrata (Sand Cress), transcript, JGI genomic project, Phytozome 12, 384_v2.1
Arabidopsis thaliana, transcript, removed miRNA gene, TAIR, version 10,released on 2010_12_14

- Contact us to add more libraries

Three input types:

- Input small RNAs
- Input target candidates
- Input small RNAs and targets

miRNA target prediction—psRNATarget (online)



# of top targets: 200

Expectation: 5.0

Penalty for G:U pair: 0.5

Penalty for other mismatches: 1

Extra weight in seed region: 1.5

Seed region: 2 - 7 NT

# of mismatches allowed in seed region:

HSP size: 19

☑ Allow bulge(gap)

Penalty for opening gap: 2

Penalty for extending gap: 1

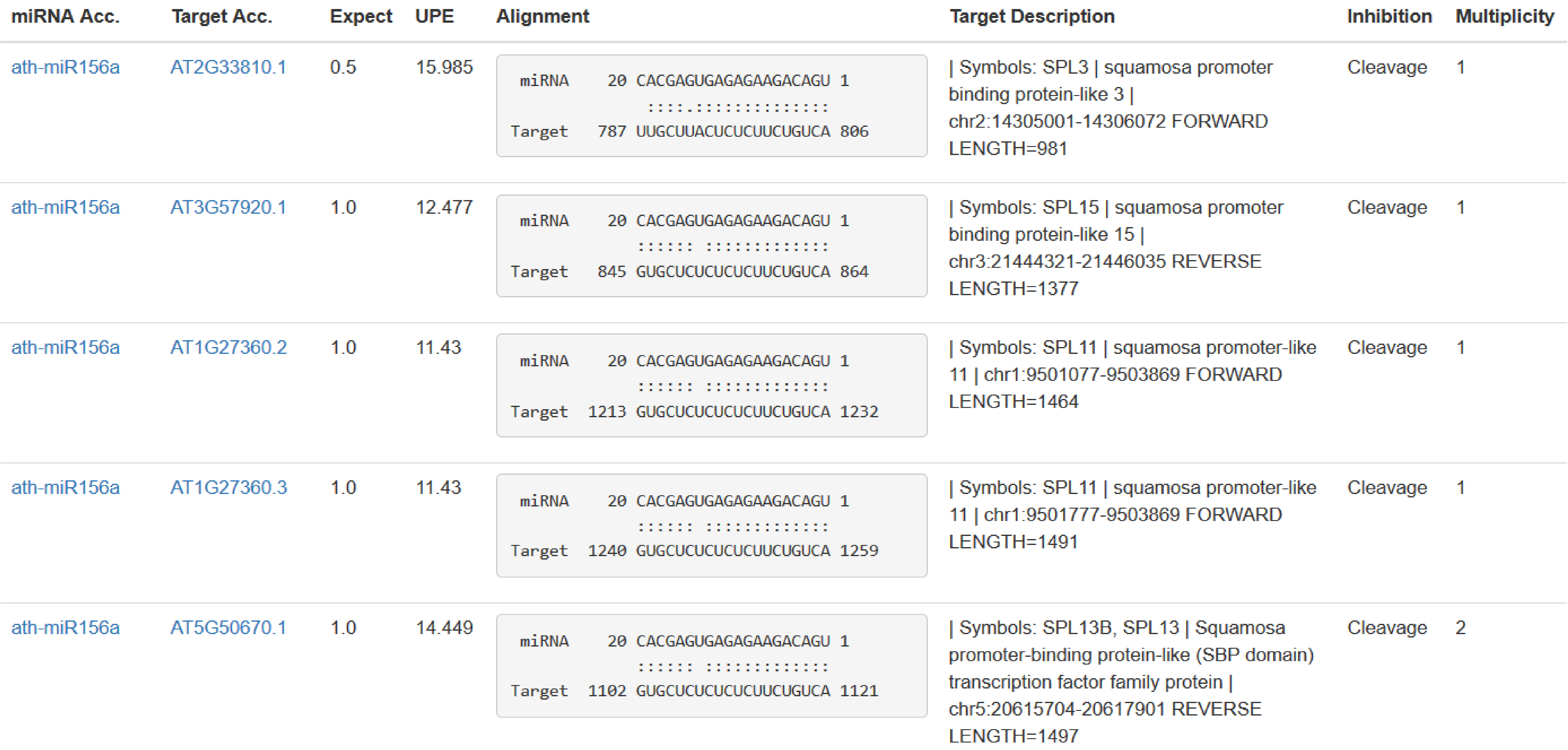
☑ Calculate target accessibility

Max UPE: 25.0

Flank length: 17 / 13 NT in up/downstream

Translation inhibition range: 9 NT - 11 NT

miRNA target prediction—psRNATarget (online)



miRNA Acc.	Target Acc.	Expect	UPE	Alignment	Target Description	Inhibition	Multiplicity

ath-miR156a	AT2G33810.1	0.5	15.985	`miRNA    20 CACGAGUGAGAGAAGACAGU 1`
				`            ::::.::::::::::::::`
				`Target   787 UUGCUUACUCUCUUCUGUCA 806`
					| Symbols: SPL3 | squamosa promoter binding protein-like 3 | chr2:14305001-14306072 FORWARD LENGTH=981	Cleavage	1

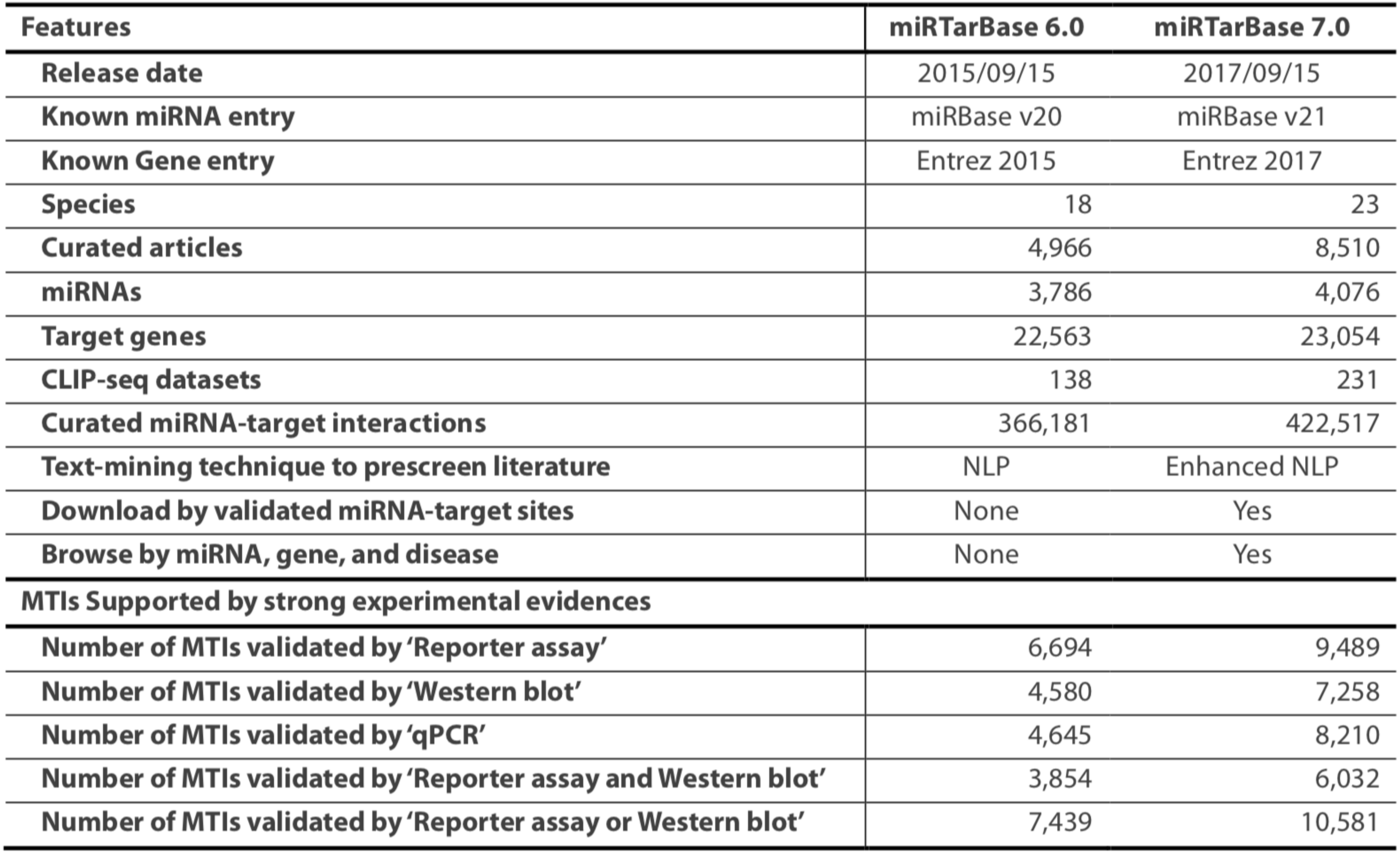
ath-miR156a	AT3G57920.1	1.0	12.477	`miRNA    20 CACGAGUGAGAGAAGACAGU 1`
				`            :::::: :::::::::::::`
				`Target   845 GUGCUCUCUCUCUUCUGUCA 864`
					| Symbols: SPL15 | squamosa promoter binding protein-like 15 | chr3:21444321-21446035 REVERSE LENGTH=1377	Cleavage	1

ath-miR156a	AT1G27360.2	1.0	11.43	`miRNA    20 CACGAGUGAGAGAAGACAGU 1`
				`            :::::: :::::::::::::`
				`Target  1213 GUGCUCUCUCUCUUCUGUCA 1232`
					| Symbols: SPL11 | squamosa promoter-like 11 | chr1:9501077-9503869 FORWARD LENGTH=1464	Cleavage	1

ath-miR156a	AT1G27360.3	1.0	11.43	`miRNA    20 CACGAGUGAGAGAAGACAGU 1`
				`            :::::: :::::::::::::`
				`Target  1240 GUGCUCUCUCUCUUCUGUCA 1259`
					| Symbols: SPL11 | squamosa promoter-like 11 | chr1:9501777-9503869 FORWARD LENGTH=1491	Cleavage	1

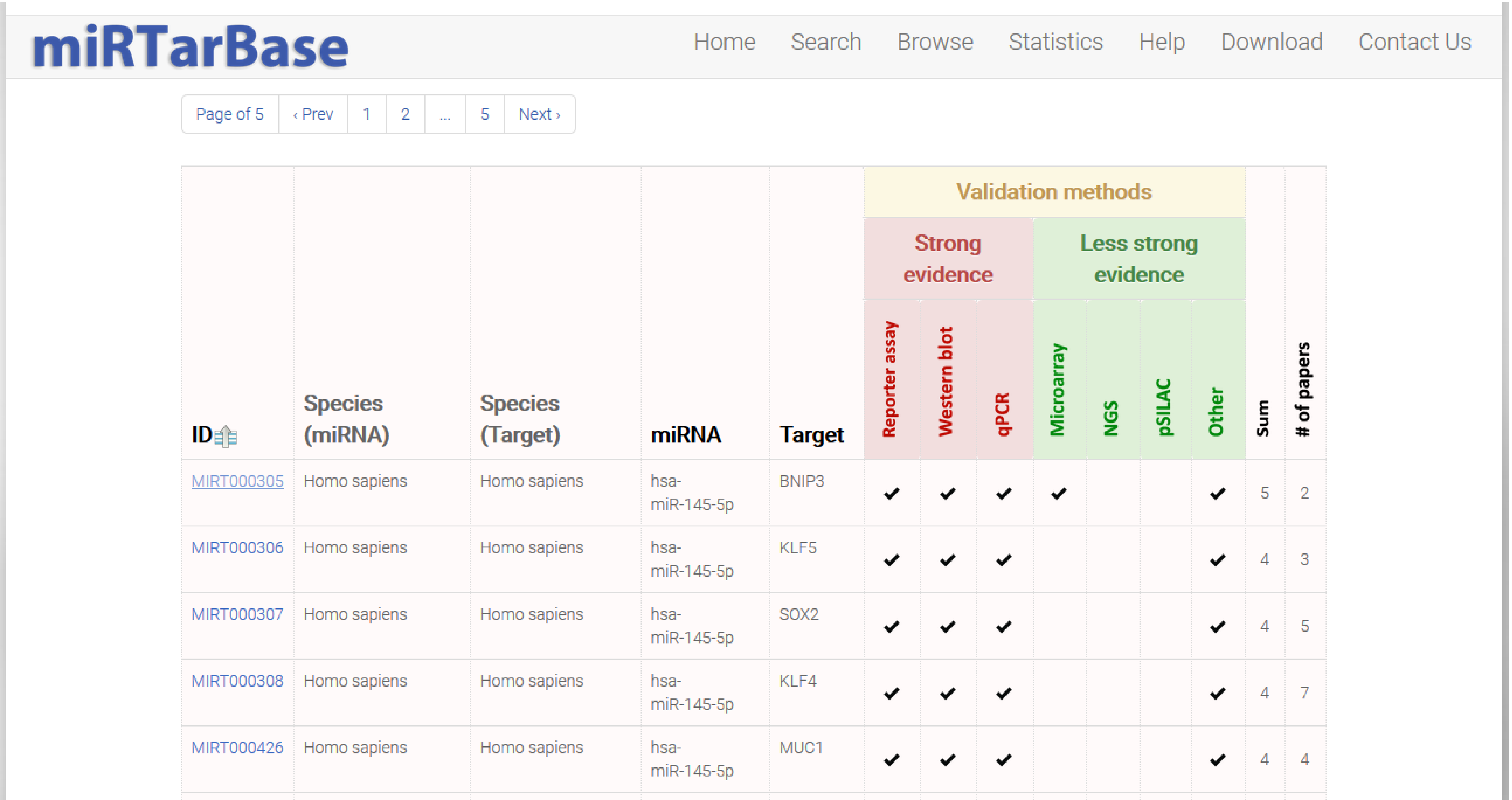
ath-miR156a	AT5G50670.1	1.0	14.449	`miRNA    20 CACGAGUGAGAGAAGACAGU 1`
				`            :::::: :::::::::::::`
				`Target  1102 GUGCUCUCUCUCUUCUGUCA 1121`
					| Symbols: SPL13B, SPL13 | Squamosa promoter-binding protein-like (SBP domain) transcription factor family protein | chr5:20615704-20617901 REVERSE LENGTH=1497	Cleavage	2

miRTarBase: the experimentally validated microRNA-target interactions database



Features	miRTarBase 6.0	miRTarBase 7.0
Release date	2015/09/15	2017/09/15
Known miRNA entry	miRBase v20	miRBase v21
Known Gene entry	Entrez 2015	Entrez 2017
Species	18	23
Curated articles	4,966	8,510
miRNAs	3,786	4,076
Target genes	22,563	23,054
CLIP-seq datasets	138	231
Curated miRNA-target interactions	366,181	422,517
Text-mining technique to prescreen literature	NLP	Enhanced NLP
Download by validated miRNA-target sites	None	Yes
Browse by miRNA, gene, and disease	None	Yes
MTIs Supported by strong experimental evidences		
Number of MTIs validated by 'Reporter assay'	6,694	9,489
Number of MTIs validated by 'Western blot'	4,580	7,258
Number of MTIs validated by 'qPCR'	4,645	8,210
Number of MTIs validated by 'Reporter assay and Western blot'	3,854	6,032
Number of MTIs validated by 'Reporter assay or Western blot'	7,439	10,581

miRTarBase: the experimentally validated microRNA-target interactions database



ID	Species (miRNA)	Species (Target)	miRNA	Target	Validation methods							Sum	# of papers
					Strong evidence			Less strong evidence					
					Reporter assay	Western blot	qPCR	Microarray	NGS	pSILAC	Other		
MIRT000305	Homo sapiens	Homo sapiens	hsa-miR-145-5p	BNIP3	✓	✓	✓	✓			✓	5	2
MIRT000306	Homo sapiens	Homo sapiens	hsa-miR-145-5p	KLF5	✓	✓	✓				✓	4	3
MIRT000307	Homo sapiens	Homo sapiens	hsa-miR-145-5p	SOX2	✓	✓	✓				✓	4	5
MIRT000308	Homo sapiens	Homo sapiens	hsa-miR-145-5p	KLF4	✓	✓	✓				✓	4	7
MIRT000426	Homo sapiens	Homo sapiens	hsa-miR-145-5p	MUC1	✓	✓	✓				✓	4	4

miRNA-related databases

Version	Date	Entries
1.0	12/02	218
1.1	01/03	262
1.2	04/03	295
1.3	05/03	332
1.4	07/03	345
2.0	07/03	506
2.1	09/03	558
2.2	11/03	593
3.0	01/04	719
3.1	04/04	899
4.0	07/04	1185
5.0	09/04	1345
5.1	12/04	1420
6.0	04/05	1650
7.0	06/05	2909
7.1	10/05	3424
8.0	02/06	3518
8.1	05/06	3963
8.2	07/06	4039
9.0	10/06	4361
9.1	02/07	4449
9.2	05/07	4584
10.0	08/07	5071
10.1	12/07	5395
11.0	04/08	6396
12.0	09/08	8619
13.0	03/09	9539
14	09/09	10883
15	04/10	14197
16	08/10	15172
17	04/11	16772
18	11/11	18226
19	08/12	21264
20	06/13	24521
21	06/14	28645
22	03/18	38589

miRBase

http://www.mirbase.org/

Latest miRBase blog posts

MicroRNA Gene Ontology annotations By sam (June 7, 2018)

You might have noticed some additional information on the mature miRNA pages in the last few weeks. See for example: http://mirbase.org/cgi-bin/mature.pl?mature_acc=MIMAT0000123 http://mirbase.org/cgi-bin/mature.pl?mature_acc=MIMAT0000069 The new section "QuickGO function" contains a set of high quality manual annotations of Gene Ontology terms for mature miRNAs, the vast majority of which come from the work of Rachael Huntley et […]

miRBase 22 release By sam (March 12, 2018)

After repeated and unreasonable delay, miRBase 22 is finally released. As you might expect with such a long gap, the number of sequences in the database has jumped significantly — by over a third. The vast majority of the increase comes from new microRNA annotations in species not previous represented in the database. Indeed, there […]

miRNA count: 38589 entries

Release 22.1: October 2018

Search by miRNA name or keyword

Download published miRNA data

Download page | FTP site

miRBase: the microRNA database

miRBase provides the following services:

- The miRBase database is a searchable database of published miRNA sequences and annotation. Each entry in the miRBase Sequence database represents a predicted hairpin portion of a miRNA transcript (termed mir in the database), with information on the location and sequence of the mature miRNA sequence (termed miR). Both hairpin and mature sequences are available for searching and browsing, and entries can also be retrieved by name, keyword, references and annotation. All sequence and annotation data are also available for download.
- The miRBase Registry provides miRNA gene hunters with unique names for novel miRNA genes prior to publication of results. Visit the help pages for more information about the naming service.

To receive email notification of data updates and feature changes please subscribe to the miRBase announcements mailing list. Any queries about the website or naming service should be directed at mirbase@manchester.ac.uk.

miRBase is managed by the Griffiths-Jones lab at the Faculty of Biology, Medicine and Health, University of Manchester with funding from the BBSRC. miRBase was previously hosted and supported by the Wellcome Trust Sanger Institute.

References

If you make use of the data presented here, please cite the following articles in addition to the primary data sources:

miRBase: from microRNA sequences to function.
Kozomara A, Birgaoanu M, Griffiths-Jones S.
Nucleic Acids Res 2019 47:D155-D162

miRBase: annotating high confidence microRNAs using deep sequencing data.
Kozomara A, Griffiths-Jones S.
Nucleic Acids Res 2014 42:D68-D73

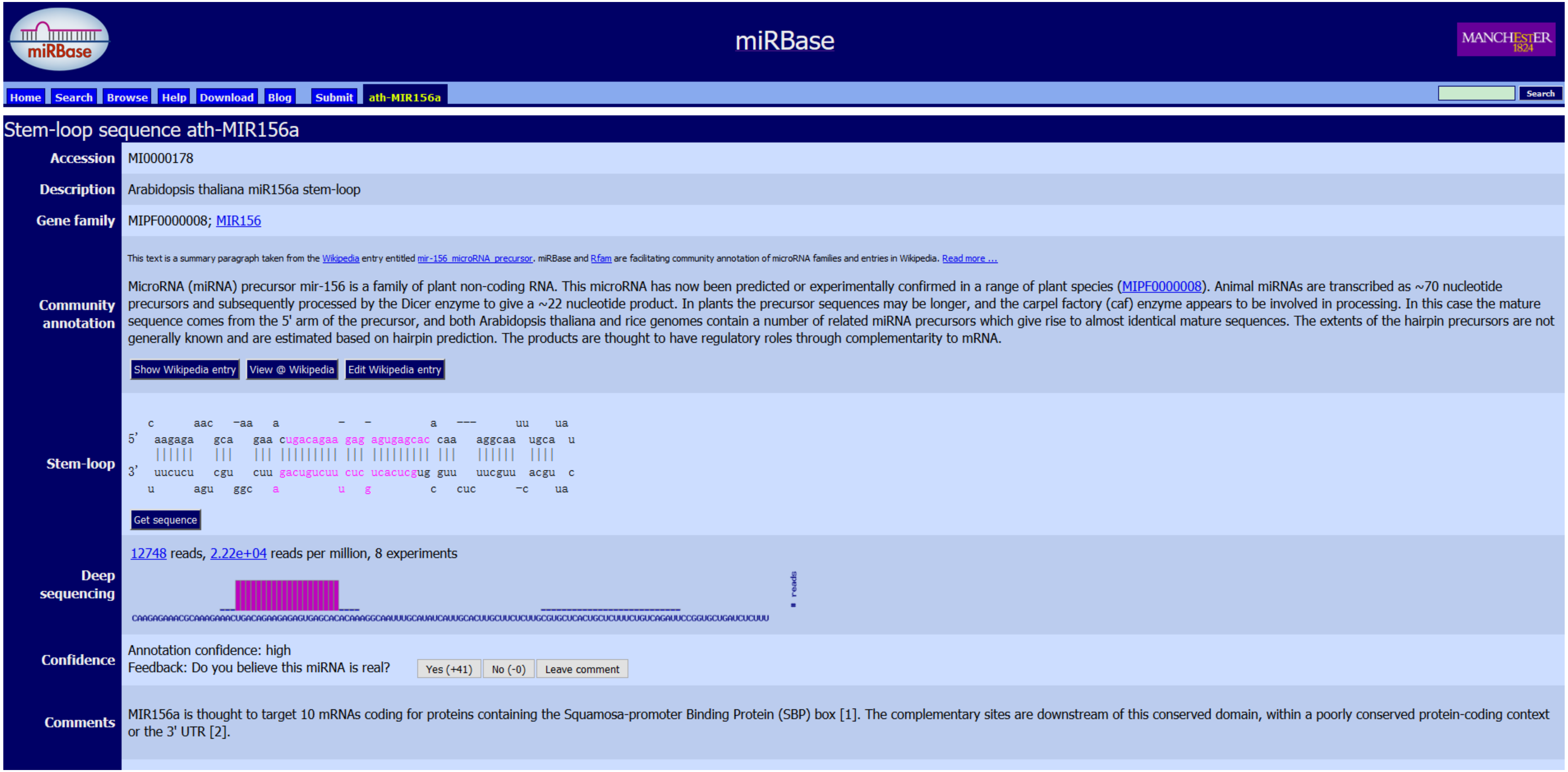
miRBase: integrating microRNA annotation and deep-sequencing data.
Kozomara A, Griffiths-Jones S.
Nucleic Acids Res 2011 39:D152-D157

miRBase: tools for microRNA genomics.
Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ.
Nucleic Acids Res 2008 36:D154-D158

miRBase: microRNA sequences, targets and gene nomenclature.
Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ.
Nucleic Acids Res 2006 34:D140-D144

miRBase

Browse -> Arabidopsis thaliana -> ath-MIR156a



Stem-loop sequence ath-MIR156a

Accession	MI0000178
Description	Arabidopsis thaliana miR156a stem-loop
Gene family	MIPF0000008; MIR156

Community annotation

This text is a summary paragraph taken from the Wikipedia entry entitled mir-156 microRNA precursor. miRBase and Rfam are facilitating community annotation of microRNA families and entries in Wikipedia. Read more ...

MicroRNA (miRNA) precursor mir-156 is a family of plant non-coding RNA. This microRNA has now been predicted or experimentally confirmed in a range of plant species (MIPF0000008). Animal miRNAs are transcribed as ~70 nucleotide precursors and subsequently processed by the Dicer enzyme to give a ~22 nucleotide product. In plants the precursor sequences may be longer, and the carpel factory (caf) enzyme appears to be involved in processing. In this case the mature sequence comes from the 5' arm of the precursor, and both Arabidopsis thaliana and rice genomes contain a number of related miRNA precursors which give rise to almost identical mature sequences. The extents of the hairpin precursors are not generally known and are estimated based on hairpin prediction. The products are thought to have regulatory roles through complementarity to mRNA.

Stem-loop

```
   c      aac   -aa   a         -   -         a   ---      uu    ua
5'  aagaga   gca   gaa cugacagaa gag agugagcac caa   aggcaa  ugca  u
    ||||||   |||   ||| ||||||||| ||| ||||||||| |||   ||||||  ||||
3'  uucucu   cgu   cuu gacugucuu cuc ucacucgug guu   uucguu  acgu  c
   u      agu   ggc   a         u   g         c   cuc      -c    ua
```

Deep sequencing

12748 reads, 2.22e+04 reads per million, 8 experiments

CAAGAGAAACGCAAAGAAACUGACAGAAGAGAGUGAGCACACAAAGGCAAUUUGCAUAUCAUUGCACUUGCUUCUCUUGCGUGCUCACUGCUCUUUCUGUCAGAUUCCGGUGCUGAUCUCUUU

Confidence

Annotation confidence: high
Feedback: Do you believe this miRNA is real?

Comments

MIR156a is thought to target 10 mRNAs coding for proteins containing the Squamosa-promoter Binding Protein (SBP) box [1]. The complementary sites are downstream of this conserved domain, within a poorly conserved protein-coding context or the 3' UTR [2].

miRBase——Search

http://www.mirbase.org/search.shtml

Search miRBase

By miRNA identifier or keyword

Enter a miRNA accession, name or keyword:

By genomic location

Select organism, chromosome and start and end coordinates. Leave the start/end boxes blank to retrieve all miRNAs on the selected chromosome.

Choose species: Chr: Start: End:

For clusters

Select organism and the desired inter-miRNA distance.

Choose species: Inter-miRNA distance: 10000

By tissue expression

Select organism and tissue.

Arabidopsis thaliana All

By sequence

Single sequence searches:

Paste a sequence here to search for similarity with miRBase miRNA sequences **(max size 1000 nts)**. You can choose to search against hairpin precursor sequences or mature miRNAs. This search may take a few minutes. Please note: this facility is designed to search for homologs of microRNA sequences, **not to predict their target sites**. For target site prediction, please use [the available bespoke tools](#).

Search sequences: Mature miRNAs

Search method: BLASTN

Choose BLASTN to search for a miRNA homolog in a longer sequence. SSEARCH is useful for finding a short sequence within the library of miRNAs (for instance, find a short motif in a miRNA or precursor stem-loop, or find mature sequences that are related to your query).

E-value cutoff: 10

Maximum no. of hits: 100

Show results only from specific organisms: ☐human ☐mouse ☐worm ☐fly ☐Arabidopsis

or choose a taxonomic classification: No species filter

Or: Select the sequence file you wish to use

根据名字搜索

根据基因组位点搜索

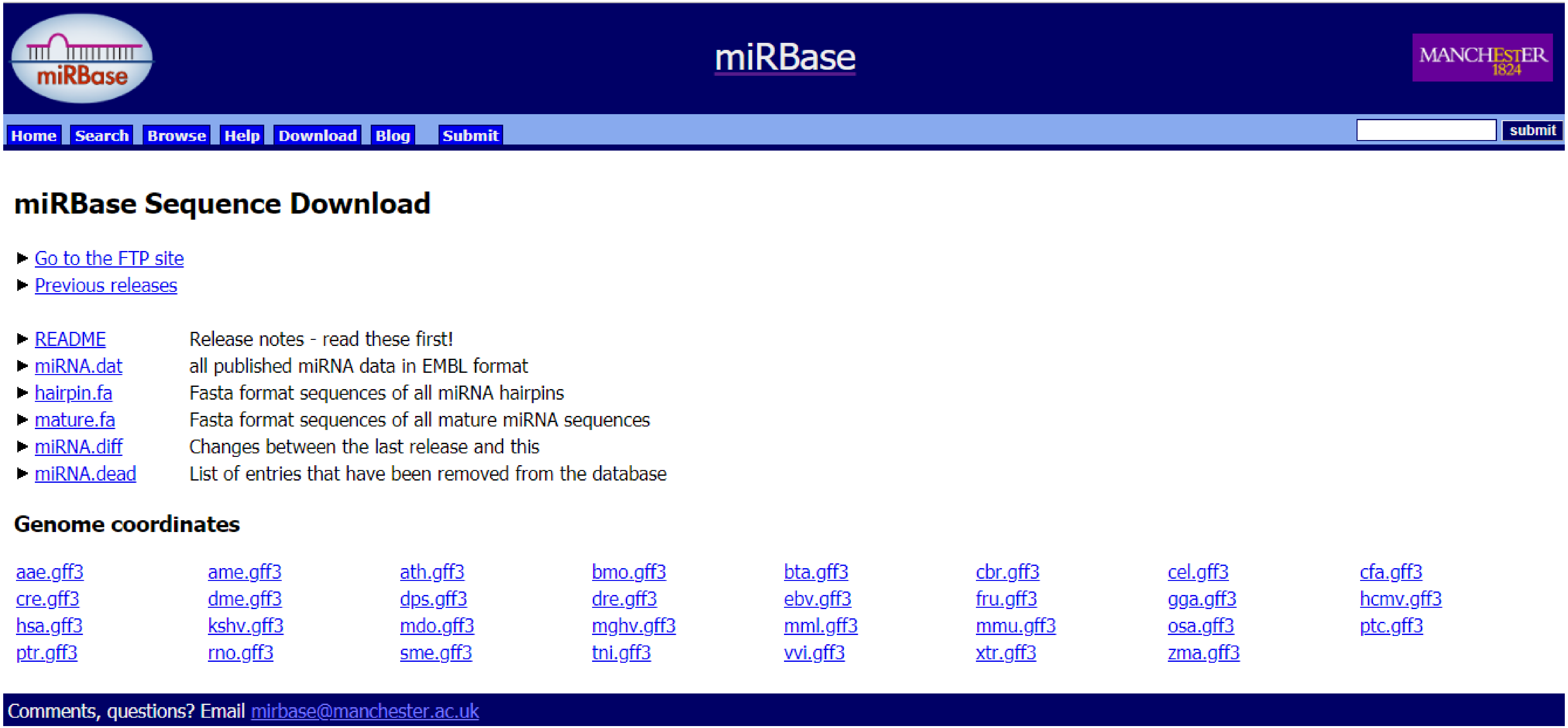
根据基因组区间搜索

根据组织搜索

根据序列搜索

miRBase——Download

http://www.mirbase.org/ftp.shtml



miRBase Sequence Download

- Go to the FTP site
- Previous releases

README	Release notes - read these first!
miRNA.dat	all published miRNA data in EMBL format
hairpin.fa	Fasta format sequences of all miRNA hairpins
mature.fa	Fasta format sequences of all mature miRNA sequences
miRNA.diff	Changes between the last release and this
miRNA.dead	List of entries that have been removed from the database

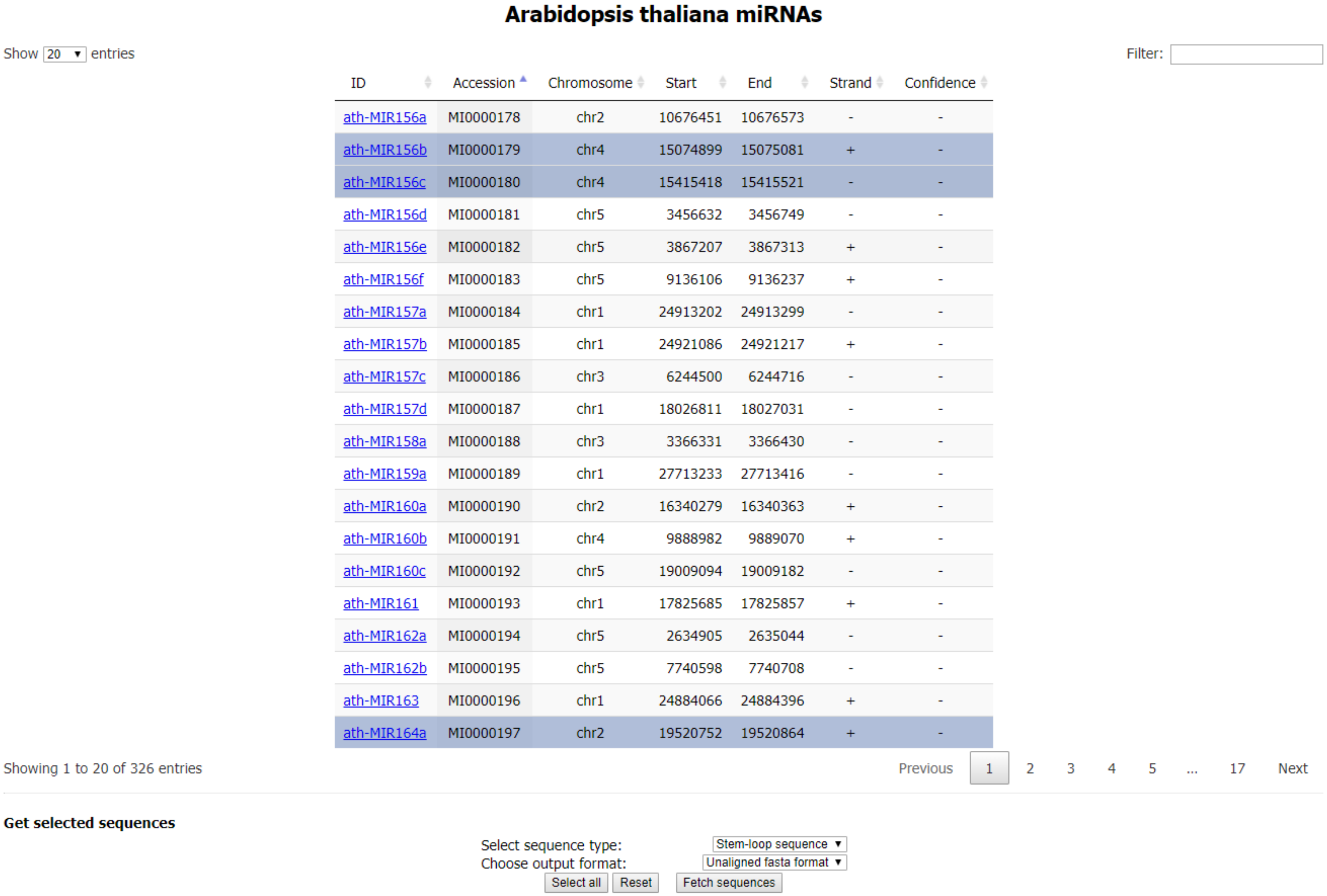
Genome coordinates

aae.gff3	ame.gff3	ath.gff3	bmo.gff3	bta.gff3	cbr.gff3	cel.gff3	cfa.gff3
cre.gff3	dme.gff3	dps.gff3	dre.gff3	ebv.gff3	fru.gff3	gga.gff3	hcmv.gff3
hsa.gff3	kshv.gff3	mdo.gff3	mghv.gff3	mml.gff3	mmu.gff3	osa.gff3	ptc.gff3
ptr.gff3	rno.gff3	sme.gff3	tni.gff3	vvi.gff3	xtr.gff3	zma.gff3	

Comments, questions? Email mirbase@manchester.ac.uk

miRBase——Browse

http://www.mirbase.org/cgi-bin/browse.pl



Arabidopsis thaliana miRNAs

Show 20 entries

ID	Accession	Chromosome	Start	End	Strand	Confidence
ath-MIR156a	MI0000178	chr2	10676451	10676573	-	-
ath-MIR156b	MI0000179	chr4	15074899	15075081	+	-
ath-MIR156c	MI0000180	chr4	15415418	15415521	-	-
ath-MIR156d	MI0000181	chr5	3456632	3456749	-	-
ath-MIR156e	MI0000182	chr5	3867207	3867313	+	-
ath-MIR156f	MI0000183	chr5	9136106	9136237	+	-
ath-MIR157a	MI0000184	chr1	24913202	24913299	-	-
ath-MIR157b	MI0000185	chr1	24921086	24921217	+	-
ath-MIR157c	MI0000186	chr3	6244500	6244716	-	-
ath-MIR157d	MI0000187	chr1	18026811	18027031	-	-
ath-MIR158a	MI0000188	chr3	3366331	3366430	-	-
ath-MIR159a	MI0000189	chr1	27713233	27713416	-	-
ath-MIR160a	MI0000190	chr2	16340279	16340363	+	-
ath-MIR160b	MI0000191	chr4	9888982	9889070	+	-
ath-MIR160c	MI0000192	chr5	19009094	19009182	-	-
ath-MIR161	MI0000193	chr1	17825685	17825857	+	-
ath-MIR162a	MI0000194	chr5	2634905	2635044	-	-
ath-MIR162b	MI0000195	chr5	7740598	7740708	-	-
ath-MIR163	MI0000196	chr1	24884066	24884396	+	-
ath-MIR164a	MI0000197	chr2	19520752	19520864	+	-

Showing 1 to 20 of 326 entries

Get selected sequences

Select sequence type: Stem-loop sequence

Choose output format: Unaligned fasta format

谢谢！