

Top 20 Useful Bioinformatic Tools

Promoter Scan

功能：启动子预测

网址：<https://www-bimas.cit.nih.gov/molbio/proscan/>



ORF Finder

功能：ORF预测

网址：<https://www.ncbi.nlm.nih.gov/orffinder/>



NCBI-BLAST

功能：序列比对


网址：<https://blast.ncbi.nlm.nih.gov/Blast.cgi>

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)


QuickBLASTP
 Try QuickBLASTP for a fast protein search of nr.
 Tue, 23 May 2017 13:00:00 EST [More BLAST news...](#)

Web BLAST



Nucleotide BLAST
nucleotide & nucleotide

blastx
translated nucleotide & protein



Protein BLAST
protein & protein

MUSCLE

功能：运行速度比较快的多序列比对

网址：<http://www.ebi.ac.uk/Tools/msa/muscle/#>

MUSCLE

Input form
Web services
Help & Documentation
Feedback
Share

Tools > Multiple Sequence Alignment > MUSCLE

Multiple Sequence Alignment

MUSCLE stands for **M**ultiple **S**equence **C**omparison by **L**og-**E**xpectation. MUSCLE is claimed to achieve both better average accuracy and better speed than [ClustalW2](#) or [T-Coffee](#), depending on the chosen options.

Important note: This tool can align up to 500 sequences or a maximum file size of 1 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of sequences in any supported format:

Clustal Omega

功能：DNA、RNA、蛋白的多序列比对

网址：<http://www.ebi.ac.uk/Tools/msa/clustalo/>

Clustal Omega

Input form
Web services
Help & Documentation
Feedback
Share

Tools > Multiple Sequence Alignment > Clustal Omega

Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three or more** sequences. For the alignment of two sequences please instead use our [pairwise sequence alignment tools](#).

Important note: This tool can align up to 4000 sequences or a maximum file size of 4 MB.

ClustalW2

功能：应用较广泛的多序列比对

网址：<http://www.ebi.ac.uk/Tools/msa/clustalw2/>

ClustalW2

[Input form](#) | [Web services](#) | [Help & Documentation](#)

[Feedback](#) | [Share](#)

Tools > Multiple Sequence Alignment > ClustalW2

ClustalW2 is a general purpose DNA or protein multiple sequence alignment program for **three or more** sequences. For the alignment of two sequences please instead use our [pairwise sequence alignment tools](#).

Please Note

The ClustalW2 services have been retired. To access similar services, please visit the [Multiple Sequence Alignment tools](#) page. For protein alignments we recommend [Clustal Omega](#). For DNA alignments we recommend trying [MUSCLE](#) or [MAFFT](#). If you have any questions/concerns please contact us via the [feedback link](#) above.

T-Coffee

功能：准确度高,速度慢的多序列比对

网址：<http://www.ebi.ac.uk/Tools/msa/tcoffee/>

T-Coffee

[Input form](#) | [Web services](#) | [Help & Documentation](#)

[Feedback](#) | [Share](#)

Tools > Multiple Sequence Alignment > T-Coffee

Multiple Sequence Alignment

T-Coffee is a multiple sequence alignment program. Its main characteristic is that it will allow you to combine results obtained with several alignment methods.

Important note: This tool can align up to 500 sequences or a maximum file size of 1 MB.

SimiTriX-SimiTetra

功能：多序列比对相似性展示

网址：<http://cotton.hzau.edu.cn/EN/tools/BioERCP/simitrix.php>

G C G I *SimiTriX-SimiTetra*
Group of Cotton Genetic Improvement

[Home](#) | [About](#) | [Blast](#) | [SimiTriX](#) | [SimiTetra](#) | [Help](#)

Control Panel

Input file
Input the value file directly
选择文件 | 未选择任何文件
提交

Dataset Name

Value Cutoff

Canvas Surface

Color Range

Show Information of selected points +

Show Result of Stats +

Venn图

功能：绘制Venn图

网址：<http://www.biovenn.nl/index.php>

Set Image Parameters

title	<input type="text" value="BioVenn"/>	Courier New bold	24	Black
subtitle	<input type="text" value="(C) 2007 - 2017 Tim Hulsen"/>	Courier New bold	18	Black
x title	<input type="text" value="ID Set X"/>	Courier New bold	12	Black
y title	<input type="text" value="ID Set Y"/>	Courier New bold	12	Black
z title	<input type="text" value="ID Set Z"/>	Courier New bold	12	Black

print numbers
 absolute nrs
 percentages

ID Set X	<input type="text"/>	<- Copy and paste your IDs .. Or input a file with IDs: <input type="button" value="选择文件"/> 未选择任何文件	Red
ID Set Y	<input type="text"/>	<- Copy and paste your IDs .. Or input a file with IDs: <input type="button" value="选择文件"/> 未选择任何文件	Lime
ID Set Z	<input type="text"/>	<- Copy and paste your IDs .. Or input a file with IDs: <input type="button" value="选择文件"/> 未选择任何文件	Blue

ID Type:
 map Affymetrix/EntrezGene IDs to Ensembl Gene IDs

background transparency
 background color

image width
 image height

(Click [here](#) to see an example)

Venn图

功能：绘制Venn图

网址：http://bioinformatics.psb.ugent.be/cgi-bin/liste/Venn/calculate_venn.html



Venn图

功能：绘制Venn图

网址：<http://bioinfogp.cnb.csic.es/tools/venny/index.html>

1. Paste up to four lists. One element per row ([example](#)),
2. Click the numbers to see the results,
3. Right-click the figure to view and save it
(actual size in pixels: 1280x1280)

UPPERCASE lowercase ←cannot be undone!

List 1 0

[clear](#)

List 2 0

[clear](#)

List 3 0

[clear](#)

List 4 0

[clear](#)

WEGO

功能：绘制GO注释结果图

网址：<http://wego.genomics.org.cn/cgi-bin/wego/index.pl>

BGI WEGO Web Gene Ontology Annotation Plotting

Introduction:

The GO (Gene Ontology) project began as the collaboration of Flybase, Saccharomyces Genome Database (SGD) and Mouse Genome Base. And now it has gone beyond what it used to be. There are so many GO resources and tools that help biologists explore the depth of gene analysis, from several genes to large-scale.

WEGO (Web Gene Ontology Annotation Plot) is a useful tool for plotting GO annotation results. It has been widely used in many important biological research projects, such as the rice genome project [Yu, J. et al. Science 296, 79-92 (2002); Yu, J. et al. PLoS Biol 3, e38 (2005)] and the silkworm genome project [Xu, Q. et al. Science 306, 1937-40 (2004)]. It has become one of the daily tools for downstream gene annotation analysis, especially when performing comparative genomics tasks. WEGO along with two other tools, namely External to GO Query and GO Archive Query, are freely available for all users. Any suggestions are welcome at wego@genomics.org.cn. Here is a sample output generated by WEGO (Fig. 1).

There are three steps to work with WEGO. The first is to upload annotation results. The input file(s) can be in WEGO native format, or if you are using InterProScan as the annotation tool, the result(s) could be used directly. We support InterProScan text, raw and XML output formats as the input format of WEGO. Then, you will be redirected to a webpage with hierarchical GO tree in which all the GO terms contained in the files uploaded are included. You could choose any GO terms interested at this page to display in the output histogram. The last step is figure setting, such as the figure caption, histogram color(s) and legend description. Currently, WEGO support SVG, PNG, PostScript, EPS and GIF as output graph format. You can also get the results by our feedback Email.

Fig 1. The sample figure of WEGO output, from the rice genome paper published on science.

CIRCOS

功能：绘制圈图

网址：<http://mkweb.bcgsc.ca/tableviewer/visualize/>



20 imperatives of information design — BioVis 2012

visualize settings samples archive about

→ 0 . READ SLOGAN BADGES

→ 1 . CHECK DATA FORMAT

Before uploading a data file, check the [samples gallery](#) to make sure that your data format is compatible.

→ 6 . WHAT IS THIS?

The Circos table viewer uses the [Circos](#) application to turn data tables into chord diagrams.

	A	B	C	D	E	F	G
A	105	450	92	94	5	301	195
B	20	46	78	33	53	28	83
C	118	553	94	317	25	89	287
D	100	18	108	104	105	25	173
H	23	83	123	342	98	40	205
I	173	428	103	325	82	215	23
J	305	173	138	49	81	258	207

into circularly composited visualizations like this

CIRCexplorer

功能：进行circRNA分析

网址：<http://circexplorer2.readthedocs.io/en/latest/>

iPath

功能：进行可视化通路图在线分析

网址：<http://pathways.embl.de/>

Select the desired version by clicking the map icons below:

iPath v2: the main interface

The default version of iPath comes with 3 overview pathways maps based on KEGG data:

- Metabolic pathways
- Regulatory pathways
- Biosynthesis of secondary metabolites

iPath v1: original version

This is the initial interactive Pathways Explorer version, which contains only the original version of the metabolic pathways overview map.

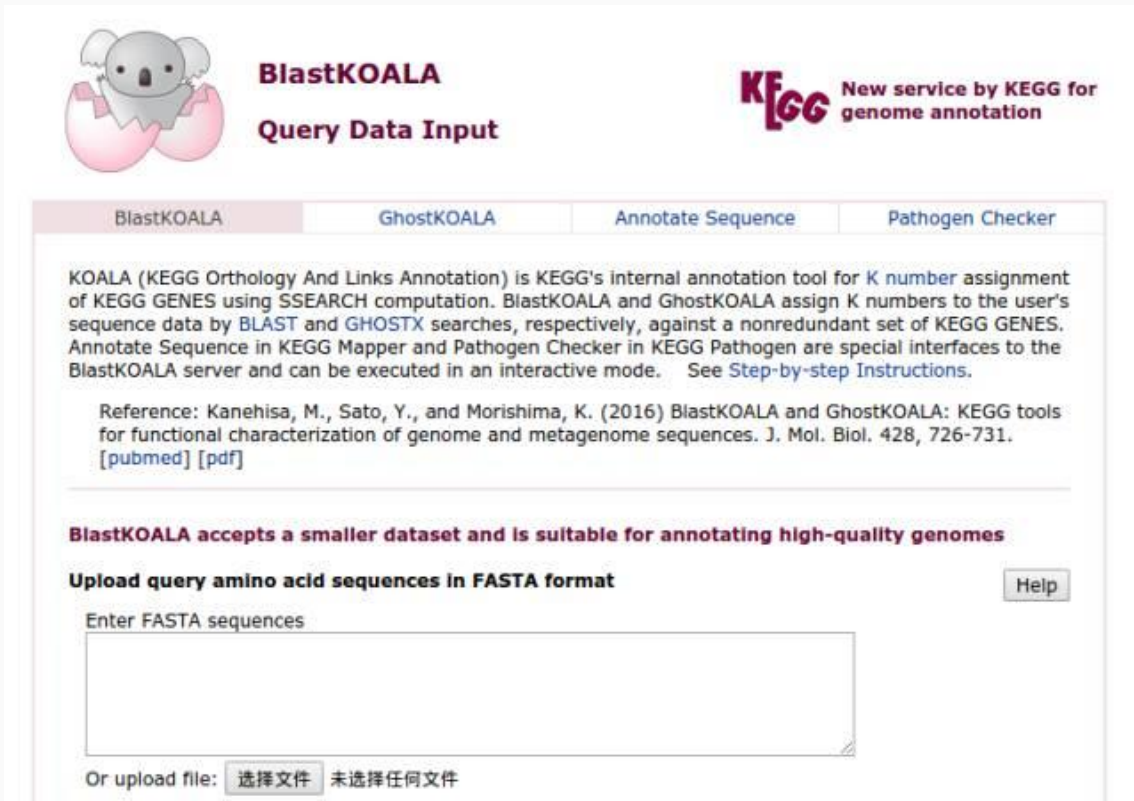
iMyc: Mycoplasma pneumoniae map

Hand curated map showing an overview of *M. pneumoniae* metabolism. Currently still using the iPath version 1 interactive interface.

KEGG

功能：进行基因的代谢通路注释

网址：<http://www.kegg.jp/blastkoala/>



BlastKOALA
Query Data Input

KEGG New service by KEGG for genome annotation

BlastKOALA GhostKOALA Annotate Sequence Pathogen Checker

KOALA (KEGG Orthology And Links Annotation) is KEGG's internal annotation tool for K number assignment of KEGG GENES using SSEARCH computation. BlastKOALA and GhostKOALA assign K numbers to the user's sequence data by BLAST and GHOSTX searches, respectively, against a nonredundant set of KEGG GENES. Annotate Sequence in KEGG Mapper and Pathogen Checker in KEGG Pathogen are special interfaces to the BlastKOALA server and can be executed in an interactive mode. See [Step-by-step Instructions](#).

Reference: Kanehisa, M., Sato, Y., and Morishima, K. (2016) BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.* 428, 726-731. [[pubmed](#)] [[pdf](#)]

BlastKOALA accepts a smaller dataset and is suitable for annotating high-quality genomes

Upload query amino acid sequences in FASTA format Help

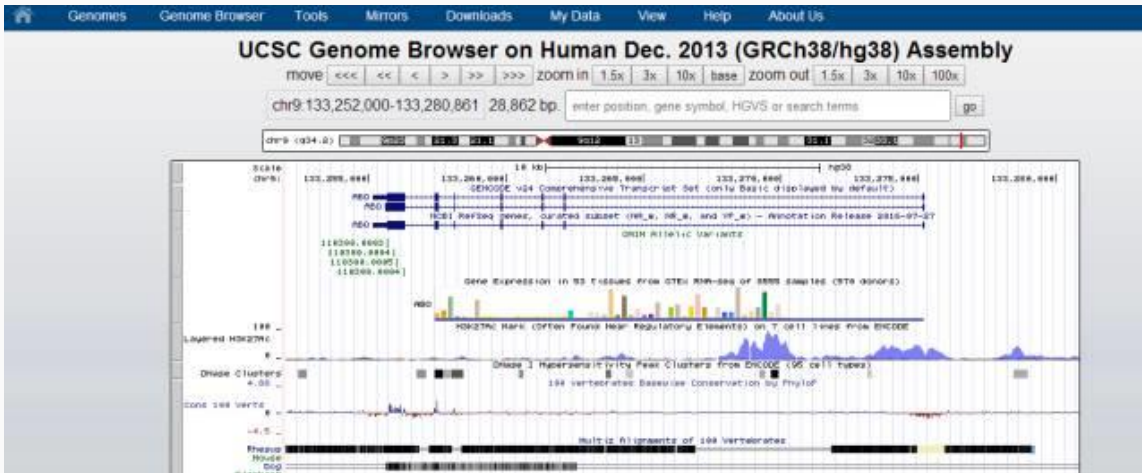
Enter FASTA sequences

Or upload file: 选择文件 未选择任何文件

UCSC

功能：进行基因组可视化

网址：<http://genome.ucsc.edu/index.html>



IBS

功能：进行序列结构示意图绘制

网址：<http://ibs.biocuckoo.org/online.php>

RAP

功能：在线分析RNA-seq

网址：<https://bioinformatics.cineca.it/rap/>

RAP

RAP

RNA-SEQ ANALYSIS PIPELINE

RAP: RNA-Seq Analysis Pipeline

RNA-Seq technology is becoming widely used in various transcriptomics studies; however, analyzing and interpreting the RNA-Seq data face serious challenges due to transcriptome complexity. A complete RNA-seq analysis involves several steps and the data can be investigated under many points of view (gene and transcript expression, differential expression, alternative splicing, polyA signals, fusion transcripts, etc.)

RAP is a web tool that performs a quite complete and customizable RNA-Seq pipeline and provides an easy and intuitive access through a web interface to intermediate and final results. The main aim of RAP is to provide to users a RNA-Seq pipeline without any installation and IT requirements. The web interface provides an easy and intuitive access for data submission and a user-friendly browsing facility of results. Users can access through RAP to several RNA-Seq algorithms, each integrated with other to maximize the overall quality and quantity of results.



AUGUSTUS

功能：基因外显子内含子，UTR,注释

网址：<http://bioinf.uni-greifswald.de/webaugustus/prediction/create>



Gene Prediction with AUGUSTUS

Navigation for: [Submit Prediction](#)

AUGUSTUS Web Server Navigation

- [Introduction](#)
- [About AUGUSTUS](#)
- [Accuracy](#)
- [Training Tutorial](#)
- [Submit Training](#)
- [Prediction Tutorial](#)
- [Submit Prediction](#)
- [Help](#)
- [Datasets for Download](#)
- [Predictions for Download](#)
- [Links & References](#)
- [Impressum](#)
- [Other AUGUSTUS Resources](#)
- [AUGUSTUS Wiki](#)
- [AUGUSTUS Forum](#)

Data Input for Running AUGUSTUS

Use this form to submit your data for running AUGUSTUS on new genomic data with already available pre-trained parameters.

Please read the [prediction tutorial](#) before submitting a job for the first time. Example data for this form is available [here](#). You may also use the button below to insert sample data. Please note that you will always need to enter the verification string at the bottom of the page, yourself, in order to submit a job!

Current problem: Regrettably, our server is currently connected to the internet via a rather unreliable connection. This may cause connection timeouts (caused by server side) when uploading big files. Please use the web link upload option, instead, if you experience such problems. We apologize for the inconvenience!

[Fill in Sample Data](#)

We recommend that you specify an **E-mail address**.

E-mail [Help](#)

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

AUGUSTUS species parameters *

Upload an archive file (max. 100 MB):

[选择文件](#) 未选择任何文件

[Help](#)

GSDS

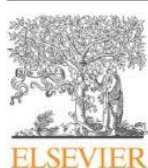
功能：基因外显子内含子，UTR,domain等区域特征展示

网址：<http://gsds.cbi.pku.edu.cn/>

The screenshot shows the GSDS 2.0 Gene Structure Display Server interface. At the top, there is a header with the logo and navigation links: Home | Help | About || Links: PlantRegMap. Below the header, there is a section for "Gene Features" with a dropdown menu set to "BED". A text area is provided for "Input features in BED format" with an "Example" button. Below this, there is a section for "Other Features to Display" with a dropdown menu set to "SVG". At the bottom, there are "Reset" and "Submit" buttons.

TCGA中miRNA神操作

John Bee文献无限好，Low Bee文献快乐多。但是Low Bee文献的快乐的到底在哪里呢？我们继续来讲讲看上次那篇1.543分的蚊帐哈（**不知道上次是哪次的就点这里哈**）。



Contents lists available at ScienceDirect

Pathology - Research and Practice

journal homepage: www.elsevier.com/locate/prp



Role of miR-452-5p in the tumorigenesis of prostate cancer: A study based on the Cancer Genome Atl(TCGA), Gene Expression Omnibus (GEO), and bioinformatics analysis

没错，就是这篇蚊帐，实在是太有内涵了。因为全程没有做实验，所以全程都用到了各种工具，看完这个，你会更长见识的。比如这段：

2.1. The clinical significance of miR-452-5p in prostate cancer from TCGA

TCGA serves as a huge repository of high throughput data on DNA, RNA, and protein in diverse human cancers, helping facilitate the comprehensive analysis of the expression of these components in various cancer types [20,21]. In the current study, we obtained the miR-452-5p expression profile of various types of human cancers and adjacent normal tissues from a TCGA data online analysis tool (http://bioinfo.life.hust.edu.cn/miR_path/index.html).

在这里，蚊帐里说到，他们分析了TCGA的数据。嗯，是TCGA中miRNA的表达数据，没错。其实他们用到的是这样一款网页神器：

Tumor-miRNA-Pathway

Browse miRNA: <ul style="list-style-type: none">let-7a-5plet-7b-5plet-7c-5plet-7d-5plet-7e-5plet-7f-5plet-7g-5pmiR-100-5pmiR-101-3pmiR-103a-3p Tumor type: <ul style="list-style-type: none">LUADPCPG <input type="button" value="Submit"/>	Search Tumor-miRNA-pathway miRNA = <input type="text" value="let-7a-5p"/> Cancer = <input type="text" value="LUAD"/> <input type="button" value="Submit"/> miRNA-target miRNA = <input type="text" value="let-7a-5p"/> <input type="button" value="Submit"/> TCGA expression profile <input type="text" value="miRNA"/> <input type="text" value="miRNA or gene symbol"/> <input type="button" value="Submit"/>
---	---

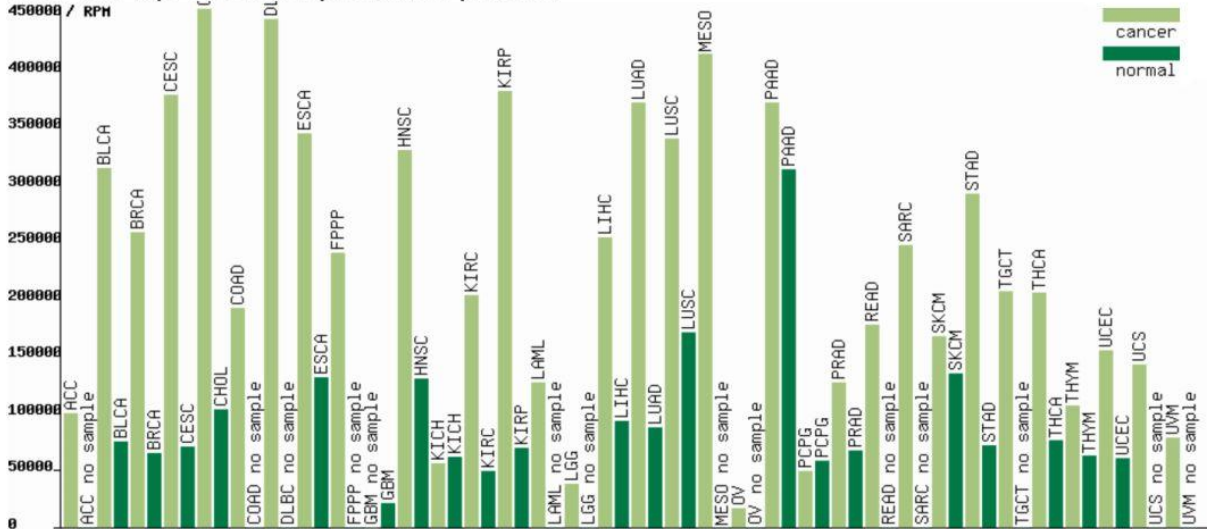
This is a database with user-friendly web interface to display the miRNA pathway regulation and their expressions in the 20 tumor types. The database provides search, browse and download function for the the miR-pathway data. Users can search miRNA regulating pathways, miRNA target genes and the expression profiles of miRNAs and genes in TCGA.

Tumor-miRNA-Pathway。没错，这个题目听了，你就知道，是有多简单粗暴了。就是研究肿瘤中 miRNA和miRNA相关的信号通路的。比如这篇蚊帐里涉及到的miRNA的表达，就是在这个界面的右侧下面这一栏：

Search Tumor-miRNA-pathway miRNA = <input type="text" value="let-7a-5p"/> Cancer = <input type="text" value="PCPG"/> <input type="button" value="Submit"/> miRNA-target miRNA = <input type="text" value="mir-21-5p"/> <input type="button" value="Submit"/> TCGA expression profile <input type="text" value="miRNA"/> <input type="text" value="miR-21-5p"/> <input type="button" value="Submit"/>
--

直接搜索miRNA就行了，我就搜了一个MiR-21-5p：

miR-21-5p TCGA expression profile



几乎和蚊帐里的图是一毛一样了：

Please cite this article as: Gao, L., Pathology - Research and Practice (2018), <https://doi.org/10.1016/j.prp.2018.03.002>

ARTICLE IN PRESS

Pathology - Research and Practice xxx (xxxx) xxx-xxx

miR-452-5p TCGA expression profile

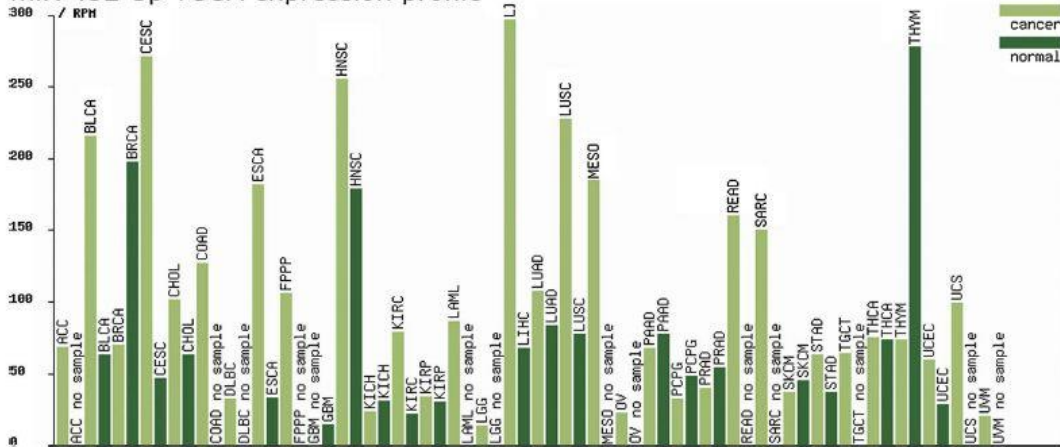


Fig. 1. Expression profile of miR-452-5p from TCGA. MiR-452-5p was down-regulated in prostate adenocarcinoma tissues compared with normal tissues.

好吧，其实这个神器还有别的功能，就是分析miRNA的信号通路：

miRNA: Tumor type:

miR-206

miR-20a-5p

miR-20b-5p

miR-210-3p

miR-211-5p

miR-215-5p

miR-21-5p

miR-217

miR-218-5p

miR-221-3p

BLCA

BRCA

CESC

CHOL

ESCA

HNSC

KIRC

KIRP

LIHC

LUAD

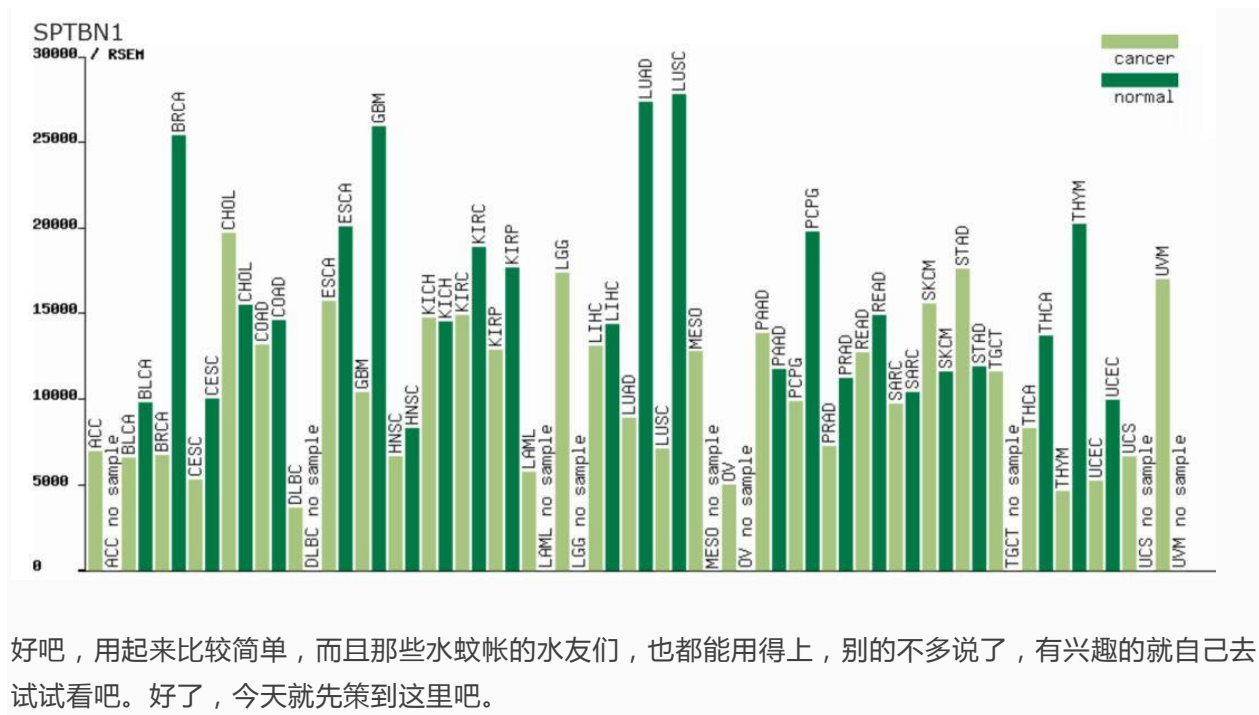
就随便浏览一下关于miR-21-5p在乳腺癌中的信号通路，当然分析的数据也都是基于TCGA的，然后就得到：

Tumor-miRNA-Pathway

miR-21-5p significantly regulates 28 pathways in BRCA(Breast invasive carcinoma)

miRNA id	Pathway alias	pathway name	P value	Targets exp.
miR-21-5p	PDZS	Synaptic Proteins at the Synaptic Junction	0.015729	Targets
miR-21-5p	TEL	Telomeres, Telomerase, Cellular Aging, and Immortality	0.015729	Targets
miR-21-5p	PPAR	Basic mechanism of action of PPARa, PPARb(d) and PPARg and effects on gene expression	0.0448604	Targets
miR-21-5p	LONGEVI	The IGF-1 Receptor and Longevity	0.0107628	Targets
miR-21-5p	TFF	Trefoil Factors Initiate Mucosal Healing	0.0303036	Targets
miR-21-5p	GCR	Corticosteroids and cardioprotection	0.015729	Targets
miR-21-5p	MAPK	MAPKinase Signaling Pathway	0.00302191	Targets

这就是跟miR-21-5p相关的信号通路了，当然也会有靶基因的表达数据，哲学数据也是基于TCGA的：



好吧，用起来比较简单，而且那些水蚊帐的水友们，也都能用得上，别的不多说了，有兴趣的就自己去试试看吧。好了，今天就先策到这里吧。

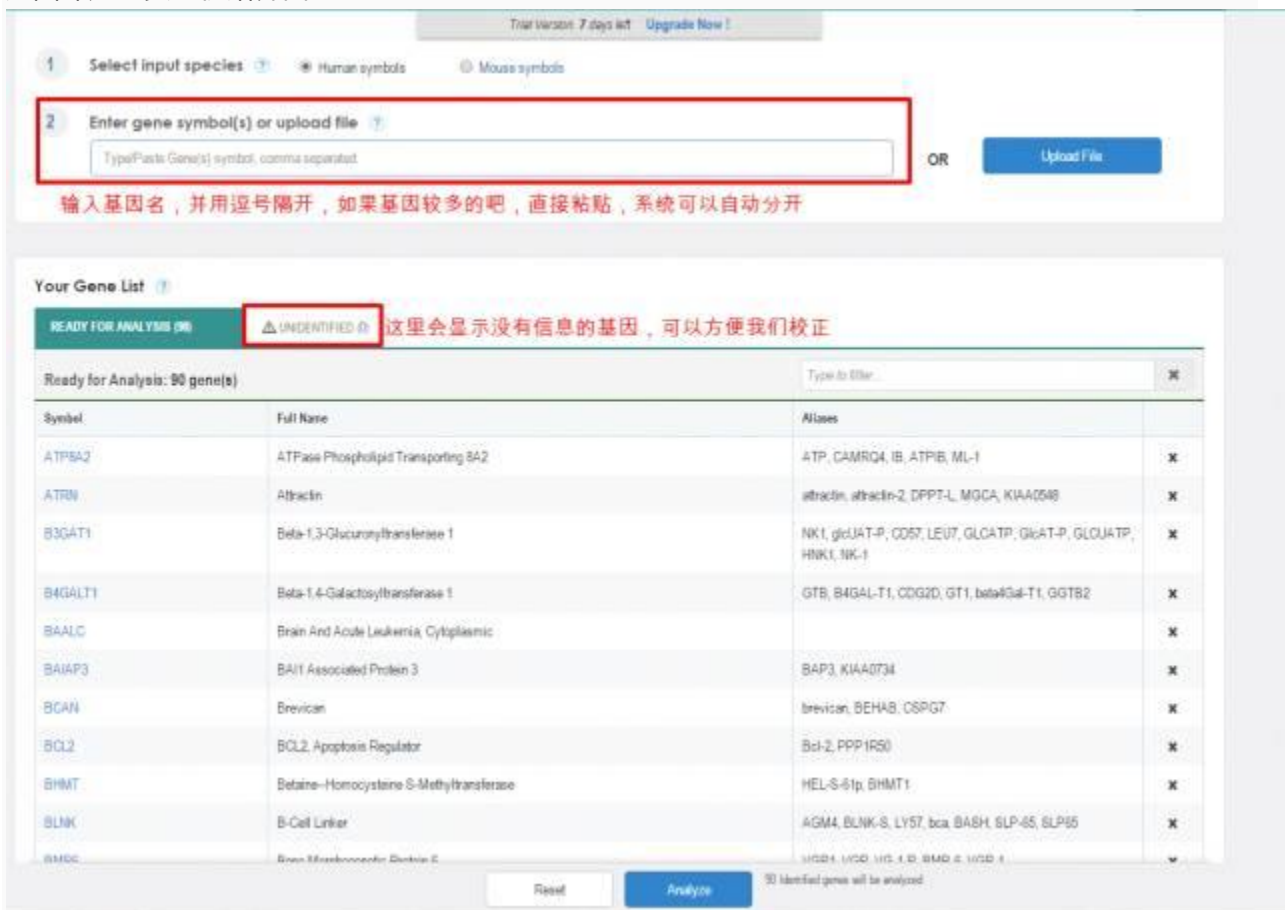
分享4个基因分析的网址

之前给大家介绍过两个数据库，**GeneCard**和**MalaCard**数据库，大家不要一脸懵逼地看着我们，会心碎，实在记不得了请点这两个超链接（如何快速了解一个疾病的综合性信息？基因信息查询网址，一个就够）。

今天给大家接着讲这两个数据库里好用的小工具。需要提醒的是，这些分析需要教育机构的邮箱进行注册，如果没有的请火速读博，从此过上幸福生活（博士的幸福生活I, II, III）。

1、GeneAnalysis (<https://ga.genecards.org/#input>) :

这个在线工具只能分析人的基因和老鼠的基因。这个在线工具其实就是对于多个基因进行整合的分析，如图即是网页的初始界面。



The screenshot shows the GeneAnalysis web interface. At the top, there is a navigation bar with "Human symbols" selected. Below it, a red box highlights the input field labeled "Enter gene symbol(s) or upload file" with a placeholder "Type/Paste Gene(s) symbol, comma separated." and an "Upload File" button. A red text annotation below the input field reads: "输入基因名，并用逗号隔开，如果基因较多的吧，直接粘贴，系统可以自动分开".

Below the input field, the "Your Gene List" section shows a table with 90 genes ready for analysis. A red box highlights a warning icon and text: "UNIDENTIFIED 这里会显示没有信息的基因，可以方便我们校正".

Symbol	Full Name	Aliases	
ATP5A2	ATPase Phospholipid Transporting 5A2	ATP, CAMRQ4, IB, ATP1B, ML-1	X
ATRN	Atractin	atractin, atractin-2, DFPT-L, MGCA, KIAA0548	X
B3GAT1	Beta-1,3-Glucuronyltransferase 1	NK1, gldJAT-P, CD57, LEU7, GLCATP, GkAT-P, GLOJATP, HNKL, NK-1	X
B4GALT1	Beta-1,4-Galactosyltransferase 1	GTB, B4GAL-T1, CDG2D, GT1, beta4Gal-T1, GGTB2	X
BAALC	Brain And Acute Leukemia, Cytoplasmic		X
BAIAP3	BAI1 Associated Protein 3	BAP3, KIAA0734	X
BCAN	Brevican	brevican, BEHAB, CSPG7	X
BCL2	BCL2, Apoptosis Regulator	Bcl-2, PPP1R50	X
BHMT	Betaine-Homocysteine S-Methyltransferase	HEL-S-61p, BHMT1	X
BLNK	B-Cell Linker	AGM4, BLNK-S, LY57, bca, BASH, SLP-65, SLP95	X
BMD4	Brain Membrane-associated Protein 4		

At the bottom, there are "Reset" and "Analyze" buttons. A note at the bottom right says "90 identified genes will be analyzed".

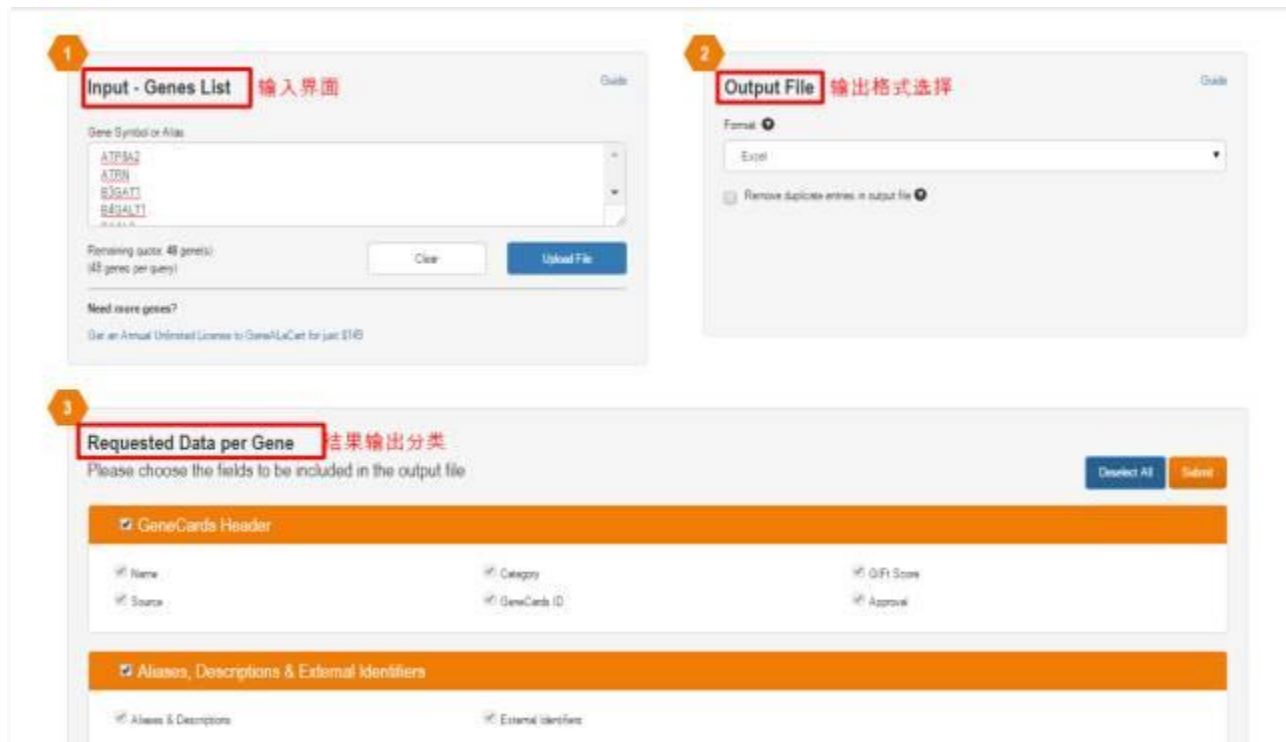
我们点击分析以后，会发现看到这些基因共同的汇总信息，如图

The screenshot shows the GeneAnalytics interface. At the top, it says "ANALYZED GENES: 90" and "NOTES: (2)". Below this, there are buttons for "Share Results", "New Analysis", and "Download Results". The main content area is divided into "BASED ON EXPRESSION" and "BASED ON FUNCTION". Under "BASED ON FUNCTION", there are tabs for "TISSUES & CELLS (34)", "DISEASES (73)", "PATHWAYS (2)", "GO - BIOLOGICAL PROCESS (26)", "GO - MOLECULAR FUNCTION (2)", and "PHENOTYPES (2)". A sidebar on the left is titled "FETERS" and contains sections for "TISSUE / System Ranked by score", "IN VIVO / IN VITRO", "EXPRESSED IN", and "PRENATAL / POSTNATAL". The main table shows "Detailed Results" with columns for Score, Entity Type, Name, # Matched Genes (Total Genes), and Organ / Tissue. The table lists various entities like Cestolium, Medial Olongata, Cerebral Cortex, Chondrocytes(RC), Chondrocytes(FC), Hippocampus, Chondrocytes(FC/CN), Chondrocytes(Zy/EC), Fibrochondrocytes(Vt), Thalamus, Hypothalamus, HyGene-BMP4-induced SM30 cells, Furo, MicroRNA-induced chondrocytes, Chondrocytes(BA2), Preshondrocytes(Tb), Chondrocytes(Pa/S), TGFbeta1-BMP7-induced chondrocytes, Chondrocytes(Sn/Cu), and HyGene-TGFbeta3-GDF5-induced E15 cells.

我们可以通过结果的分类来观察这些结果的具体信息。其主要分类有：组织中的表达情况；和输入基因相关的疾病；通路分析；GO分析；表型分析和分子组成分析

2、GeneAlaCart(<https://genealacart.genecards.org/Query>):

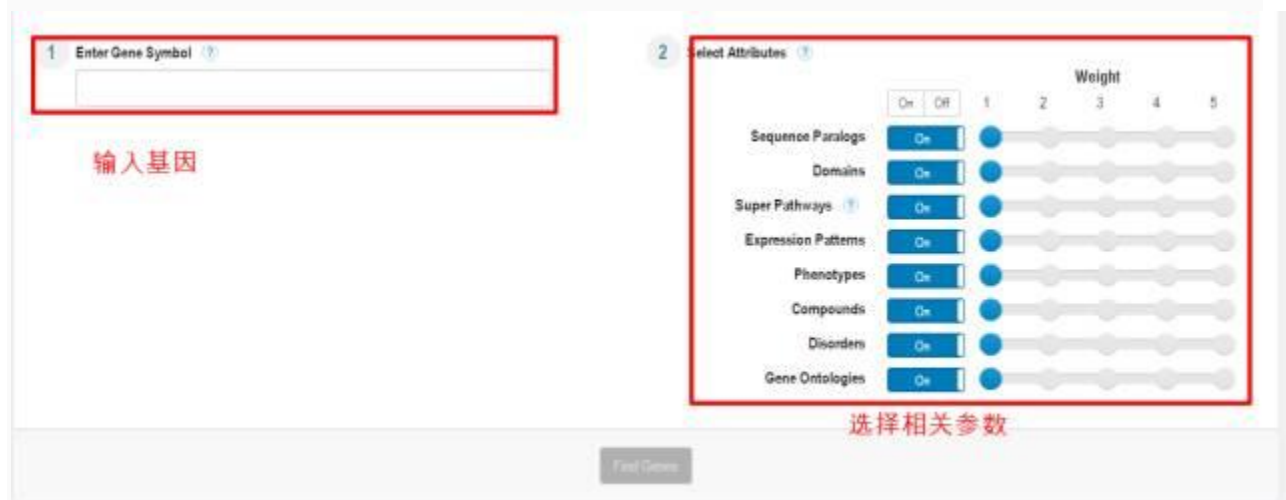
假如我有一些基因，我想下载这些基因所有各自相关的信息，那么我就可以用这个软件。网页的初始界面如图，我们只需要填入基因，选择自己想要保存的分类内容，然后点击提交即可，然后结果会以 excel 的形式保存下来。



在下载下来的excel格式中，如果全选的话一共有26个分类，包括molecular function descriptions, phenotypes, human phenotypontology, biological Processes, Cellular Components, Molecular Functions, Pathways, Interactions, Super Pathway等等。

3、Genes Likes me(<https://glm.genecards.org/#input>):

如果我有一个基因，我想知道和基因相似的其他基因有哪些，那么这个软件就可以帮我们做到。如图数据基因名即可



我们来输入TP53，即可看到结果如下，排在前面的TP63\TP73为TP53同家族的：

Weight	1	1	1	1	1	1	1	1	1
# Symbol	Total Score	Sequence Paralog	Domain	Super Pathways	Expression Patterns	Phenotype	Compounds	Disorders	Gene Ontologies
TP73	2.74	1.00	1.00	0.11		0.38	0.04	0.01	0.20
TP63	2.72	1.00	1.00	0.08		0.43	0.02	0.01	0.17
MYC	1.80			0.38	0.65	0.47	0.10	0.09	0.12
BRCA1	1.71			0.22	0.69	0.51	0.11	0.06	0.13
BAX	1.69			0.29	0.67	0.35	0.19	0.03	0.15
ERF1	1.64			0.30	0.72	0.43	0.06	0.01	0.11
MDM2	1.63			0.42		0.46	0.53	0.10	0.13
NFKB1	1.53			0.32	0.76	0.35	0.01	0.00	0.09
TGFB1	1.53			0.16	0.66	0.50	0.05	0.04	0.11
FAS	1.51			0.17	0.66	0.49	0.11	0.02	0.06
PTEN	1.46			0.21	0.56	0.49	0.06	0.07	0.07
CDK2	1.45			0.34	0.59	0.34	0.11	0.02	0.06
TNF	1.42			0.18	0.63	0.45	0.05	0.06	0.06
HDAC1	1.39			0.23	0.63	0.39	0.03	0.00	0.11
STAT1	1.39			0.16	0.66	0.42	0.05	0.01	0.09
BIRC5	1.38			0.11	0.64	0.21	0.13	0.03	0.06
CDK1	1.37			0.22	0.74	0.22	0.11	0.01	0.08
CDK4	1.35			0.25	0.59	0.35	0.08	0.05	0.03

4、VarElect (<https://ve.genecards.org/#input>)

如果我有一些基因，我想知道这些基因的哪些基因是和某一临床表型相关，我就可以用这个分析工具。如图，我们在左边的方框输入基因名，enter phenotype keywords(输入临床表型关键词)输入deaf(举个例子)。

The screenshot shows the VarElect web interface. On the left, there is a text input field labeled 'Enter/Paste Gene Symbols' containing 'GJB2, PIM3, GJB6, MACC1'. Below it is a large red watermark '输入基因' and a 'Submit' button. On the right, there is a text input field labeled 'Enter Phenotype Keywords' containing 'deaf'. Below it is a large red watermark '输入表型' and a 'Submit' button. At the bottom, there is a green 'Analyze' button and a red watermark '开始分析'. Below the input fields, there is a table showing the results of the analysis:

Symbol	Name	
GJB2	Gap Junction Protein Beta 2	✘
GJB6	Gap Junction Protein Beta 6	✘
MACC1	MACC1, MET Transcriptional Regulator	✘
PIM3	Pim-3 Proto-Oncogene, Serine/Threonine Kinase	✘

结果如下，GJB2与GJB6与deaf相关，其中GJB2最相关。

	Symbol	Description	Type	Score
1	GJB2	Gap Junction Protein Beta 2	Protein	96.95
2	GJB6	Gap Junction Protein Beta 6	Protein	59.06

今天就策到这里，希望对大家有帮助。

怎么证明LncRNA是LncRNA

最近的课题是LncRNA，LncRNA，LncRNA，重要的事情说三遍，但是我的LncRNA到底是不是LncRNA呢？我怎么陷入到了这样一个漩涡里呢！？

先不要靠师兄师姐，我就自己找找看吧，有一篇这样的Cell上的文献：

Cancer Cell
Article

CellPress

Exosome-Transmitted IncARSR Promotes Sunitinib Resistance in Renal Cancer by Acting as a Competing Endogenous RNA

实验万事屋

这篇文献里提到：“The non-coding nature of IncARSR was confirmed by coding-potential analysis (Figure S1M).” 然后我看了一下Supplement Figure.

M

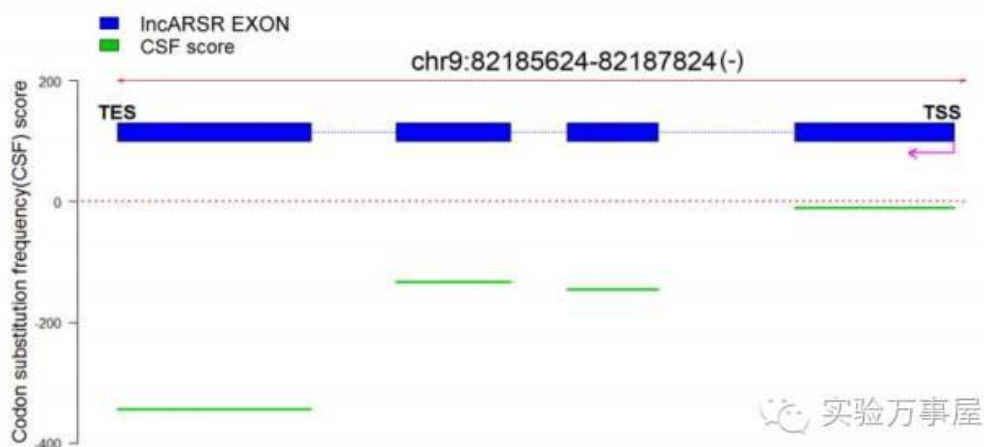
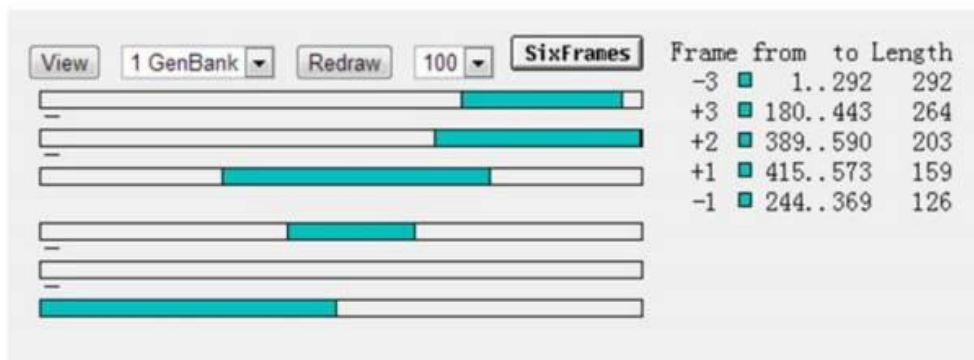


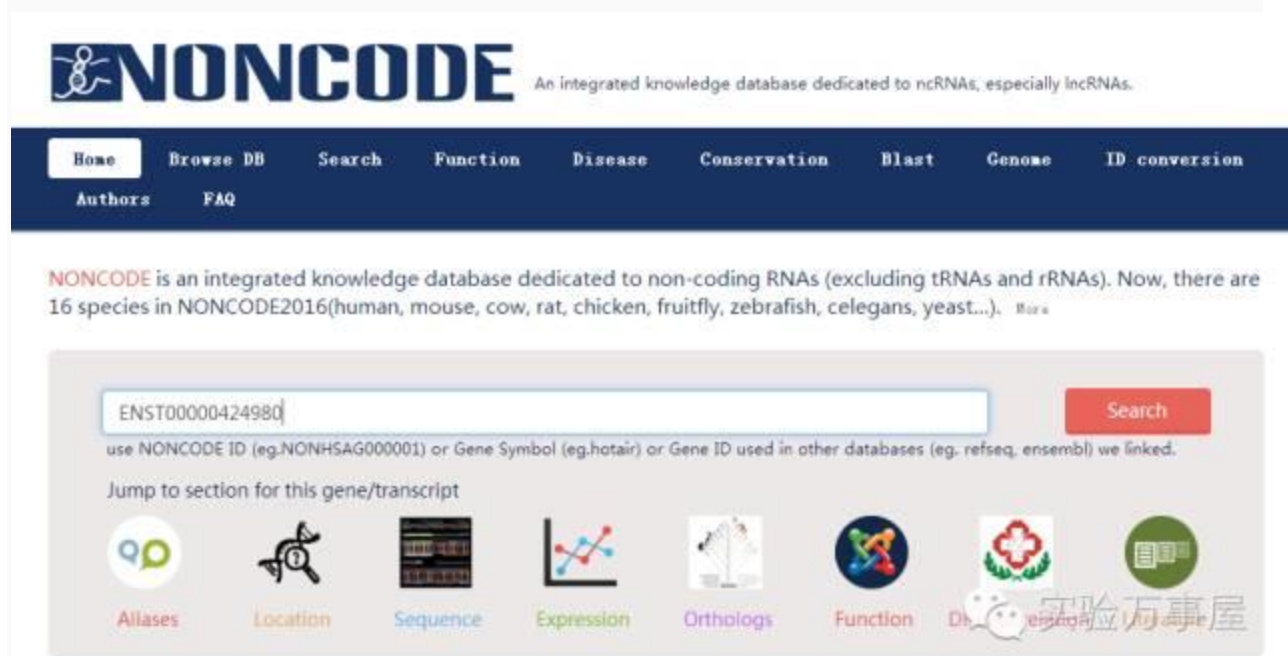
Fig. legend是这样写的：(M) Upper: Prediction of putative proteins encoded by IncARSR using ORF Finder. Lower: The codon substitution frequency scores (CSF) of IncARSR.

首先我明白一件事，就是要先分析这个lncRNA的ORF，也就是开放式阅读框。但是接下去要做什么呢？CSF又是啥？师姐，我要怎么办？？？

莫愁：这个啊，其实不是很复杂啦，我们就拿这篇文献来做例子吧。首先，我们找到这篇文献描述的这个lncRNA是啥。

... apply scheme applied to patients. From the 24 months to validated in the first round of experiments, eight lncRNAs that were upregulated in the PDXs with poor sunitinib response, but not in the PDXs with good response, were further selected (Figure S1H; Tables S1 and S2). Thirdly, the eight selected lncRNAs were subjected to loss-of-function analysis in sunitinib-resistant RCC cells by RNAi (Figure S1I). Notably, interference of lncRNA RP11-375O18.2-001 (Ensembl: ENST00000424980) suppressed sunitinib resistance compared with the remaining seven lncRNAs (Figures S1J and S1K). Therefore, we focused on this uncharacterized lncRNA and named it lncARSR (lncRNA Activated in RCC with Sunitinib Resistance). lncARSR is located on chromosome 9 in humans and composed of four exons with a full length of 591 nt determined by RACE (rapid amplification of cDNA ends) assay (Figures 1C and S1L). The non-coding nature of lncARSR was confirmed by coding-potential analysis

就是上面这个编号的lncRNA。接着我们，登陆到NONCODE (<http://www.noncode.org/>) 上去，把这lncRNA序列调出来：



得到这个序列：

General info

NONCODE TRANSCRIPT ID	NONHSAT132007.2
NONCODE Gene ID	NONHSAG052636.2
Chromosome	chr9
Start Site	79505803
End Site	79532342
Strand	-
Exon Number	2
CNCI Score	-0.0729105
Length	328
Assembly	hg38
Other transcript Versions	NONHSAT132007.1 (old version)

Sequence

```
>NONHSAT132007
TCACCCAGGTGCAAGCCCAGAGGCAGTCTATACCCCAACTCAACTGGCTGGTCTCAATGCTGCCTGCTCCCGTGCCCAACTTAGA
GTCATTATAAGTCTGAAGATTGCCATTGAAATGCTCTTTGAGGGATGCCGAAGTCAACCCTGGATCCAAAGTAGCTTTGATGTTTGCAGG
GTTTCTACAGAGCATGAAGAACTCCAACCTCAGACAACCTGCAAAAAAAGTCAGAGAGCAATTAATAATAAAAAATAAATTCCTTTGATAAAACAAA
```

那接下来，验证这个RNA到底是不是lncRNA呢？首先我们要了解的，就是lncRNA是不能编码的，那就没有足够的ORF，也就是开放式阅读框。那我们就登陆到PubMed的ORFfinder (<http://www.ncbi.nlm.nih.gov/orffinder/>) 上去。

NCBI Resources How To Sign in to NCBI

ORFfinder PubMed Search

NCBI will be testing https on public web servers from 8:00 to 9:00 AM EDT (13:00-14:00 UTC) on Thursday, September 15. You may experience problems with NCBI web sites during that time. Please plan accordingly. [Read more.](#)

Open Reading Frame Finder

ORF finder searches for open reading frames (ORFs) in the DNA sequence you enter. The program returns the range of each ORF, along with its protein translation. Use ORF finder to search newly sequenced DNA for potential protein encoding segments, verify predicted protein using newly developed SMART BLAST or regular BLASTP.

This web version of the ORF finder is limited to the subrange of the query sequence up to 50 kb long. Stand-alone version, which doesn't have query sequence length limitation, is available for Linux x64.

Examples (click to set values, then click Submit button):

- NC_011504 Salmonella enterica plasmid pWES-1, genetic code: 11, 'ATG' and alternative initiation codons; minimal ORF length: 300 nt
- NM_000059, genetic code: 1; start codon: 'ATG only'; minimal ORF length: 150 nt

Enter Query Sequence

Enter accession number, gi, or sequence in FASTA format:

```
TCACCCAGGTGCAAGCCCAGAGGCAGTCTATACCCCAACTCAACTGGCTGGTCTCAATGCTGCCTGCTCCCGTGCCCAACTTAGA
GTCATTATAAGTCTGAAGATTGCCATTGAAATGCTCTTTGAGGGATGCCGAAGTCAACCCTGGATCCAAAGTAGCTTTGATGTTTGCAGG
GTTTCTACAGAGCATGAAGAACTCCAACCTCAGACAACCTGCAAAAAAAGTCAGAGAGCAATTAATAATAAAAAATAAATTCCTTTGATAAAACAAA
```

实验万事屋

调整一下，看看到底有多少个氨基酸（aa），我就调到了最低30个氨基酸的选项：

Choose Search Parameters

- Minimal ORF length (nt): 30
- Genetic code: 1. Standard
- ORF start codon to use:
 - "ATG" only
 - "ATG" and alternative initiation codons
 - Any sense codon
- Ignore nested ORFs:

实验万事屋

搜索获得的结果发现，没有一个ORF是超过200nt的，这就说明可能是非编码的RNA。接着，我们把所有正义链（标识+的ORF）进行BLAST。

Sequence

ORFs found: 6 Genetic code: 1 Start codon: 'ATG' only

1: 1..328 (328bp) Find: [Navigation icons]

ORF3 (59 aa) [Mark]

```
>lc1|ORF3
MLFBICEYVNGSKVALMIVHEMLESYTLQF
LQSMEMSFPAKRVREQLHIEIMSPDT
```

SmartBLAST ORF3
BLAST ORF3 BLAST marked set

BLAST Database
UniProtKB/Swiss-Prot (swissprot)

Label	Strand	Frame	Start	Stop	Length (bp aa)
ORF3	+	3	147	>326	180 59
ORF2	+	2	161	319	159 52
ORF1	+	1	58	189	132 43
ORF4	-	1	115	>2	114 37
ORF6	-	3	95	>3	93 30
ORF5	-	2	141	88	54 17

Add six-frame translation track

实验万事屋

BLAST结果发现这些短肽都没有同源性的蛋白质，这就更进一步说明了，这RNA可能不表达蛋白。

BLAST[®] » blastp suite » RID-XH922P2E014

Home Recent Results Saved Str

BLAST Results

[Edit and Resubmit](#) [Save Search Strategies](#) [Formatting options](#) [Download](#)

[YouTube](#) [How to read this page](#)

lc1|ORF3_1:146:325 unnamed protein product, partial (60 letters)

RID XH922P2E014 (Expires on 09-15 09:13 am)
Query ID lc1|Query_148418
Description lc1|ORF3_1:146:325 unnamed protein product, partial
Molecule type amino acid
Query Length 60

Database Name swissprot
Description Non-redundant UniProtKB/SwissProt sequences
Program BLASTP 2.5.0+ [Citation](#)

No significant similarity found. For reasons why, [click here](#)

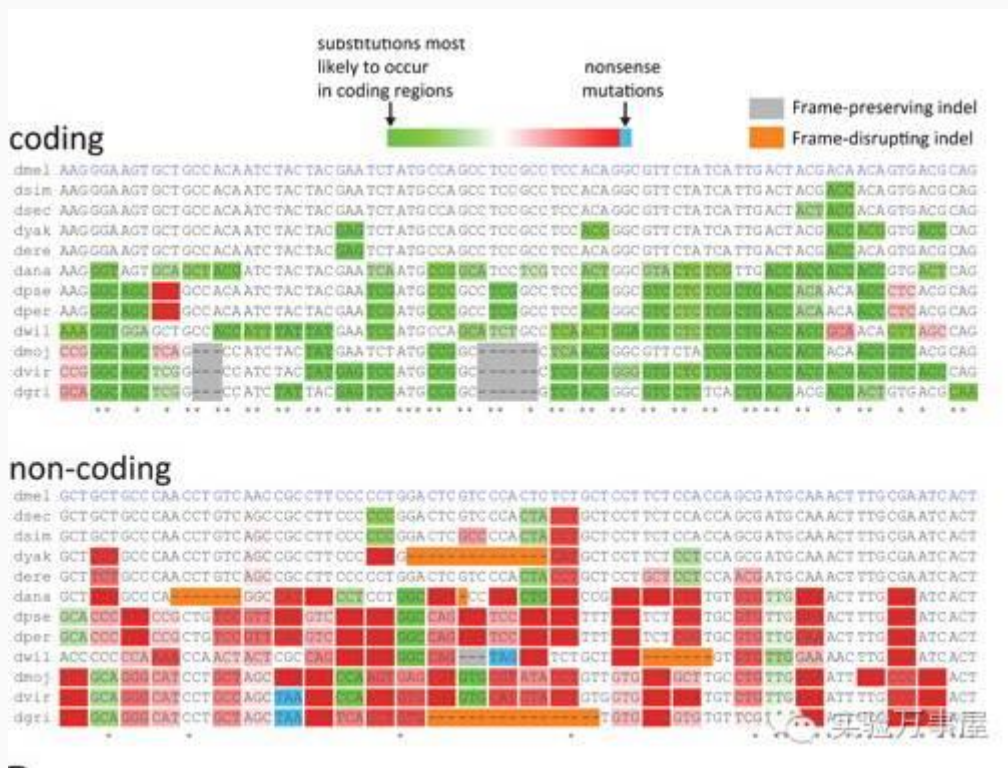
Other reports: [Search Summary](#)

实验万事屋

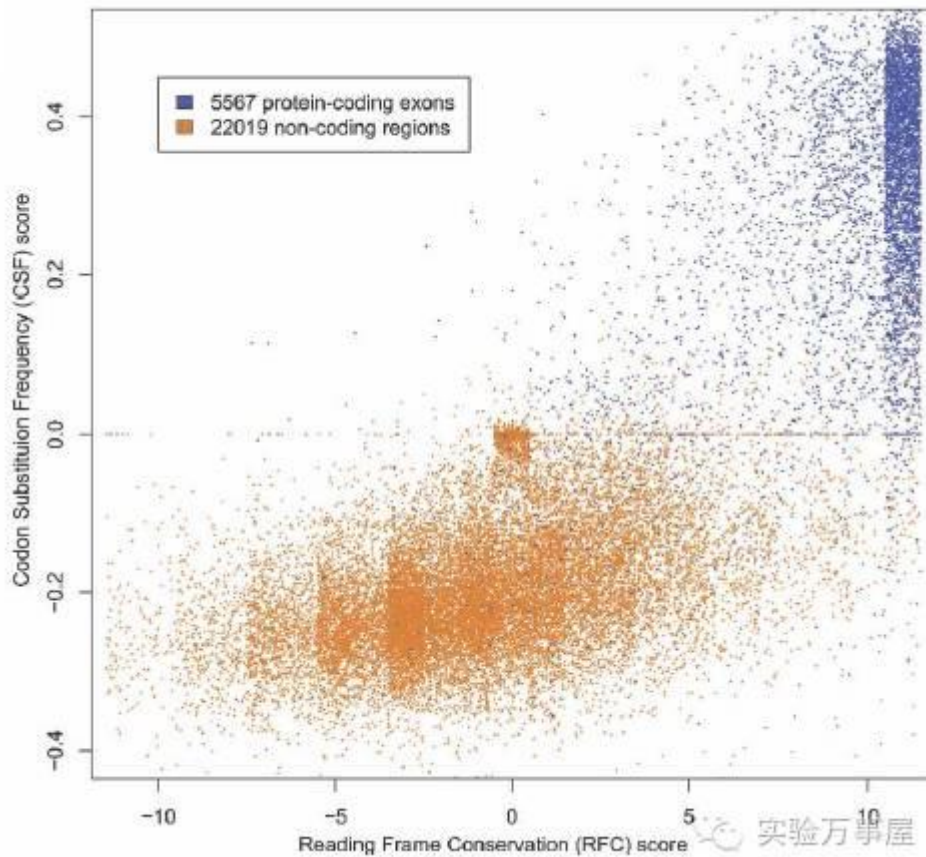
接着我们来看CSF，CSF到底是啥？CSF其实就是密码子的突变率。理论上编码区的密码子相对来说是保守的，也就是在物种中或者物种间是不容易产生突变，而非编码的就有点乱来了。我找到了这篇文章：



这是一篇在果蝇中用CSF来验证非编码与编码RNA间CSF差异的文献。其中显示，非编码的RNA突变率更高。



这篇文献用的是两个指标，一个是CSF（密码子替换频率，Y轴），另一个是RFC（阅读框保守性，X轴），见下图：



可以看到ncRNA的CSF值都小于0。由于序列保守性的问题，所以在这个CSF值的基础上，Michael又延伸出了一个新的，引入进化模型的值PhyloCSF。现在用于验证lncRNA的大多数是PhyloCSF值，详见下面这篇文献哈：

PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions

Michael F. Lin^{1,2,*}, Irwin Jungreis^{1,2} and Manolis Kellis^{1,2,*}

¹Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar Street 32-D510, Cambridge, MA 02139 and ²The Broad Institute, 7 Cambridge Center, Cambridge, MA 02142, USA

那问题来了，我们要怎么分析序列的PhyloCSF值呢？首先，要登录到强大到不要不要的UCSC上，随便进一个序列，我选了一个lncRNA——HOTAIR。然后点击“Track hubs”按钮。



进去之后，选择“My Hubs”。

Track Data Hubs

Track data hubs are collections of external tracks that can be imported into the UCSC Genome Browser. Hub tracks show up under the hub's as well as on the configure page. For more information, see the [User's Guide](#). To import a public hub click its "Connect" button below.

NOTE: Because Track Hubs are created and maintained by external sources, UCSC is not responsible for their content.

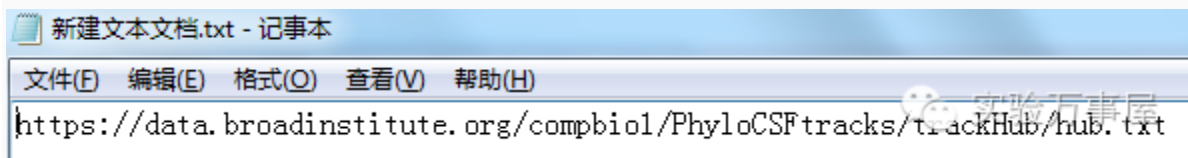
Public Hubs **My Hubs**

Enter search terms to find in public track hub description pages:

Clicking Connect redirects to the gateway page of the selected hub's default assembly.

Display	Hub Name	Description	Assemblies
<input type="button" value="Connect"/>	Roadmap Epigenomics Data Complete Collection at Wash U VizHub	Roadmap Epigenomics Human Epigenome Atlas Data Complete Collection, VizHub at Washington University in St. Louis	hg19
<input type="button" value="Connect"/>	Cancer genome polyA site & usage	An in-depth map of polyadenylation sites in cancer (matched-pair tissues and cell lines)	hg19
<input type="button" value="Connect"/>	ENCODE Analysis Hub	ENCODE Integrative Analysis Data Hub	hg19
<input type="button" value="Connect"/>	miRcode microRNA sites	Predicted microRNA target sites in GENCODE transcripts	hg19
<input type="button" value="Connect"/>	Translation Initiation Sites (TIS)	Translation Initiation Sites (TIS) track	hg19

在里面添加这个网址，我知道你们懒，所以不能惯着你们：



接着点击确认（上面看不清就看下面）：

Genomes Genome Browser Tools Mirrors Downloads My Data Help About U

Track Data Hubs

Track data hubs are collections of external tracks that can be imported into the UCSC Genome Browser. Hub tracks as on the configure page. For more information, see the [User's Guide](#). To import a public hub click its "Connect" butt

NOTE: Because Track Hubs are created and maintained by external sources, UCSC is not responsible fo

Public Hubs My Hubs

URI: Add Hub

Display	Hub Name	Description
<input type="button" value="Disconnect"/>		ERROR: Duplicate genome mm10 in stanza ending line 11 of https://www.encodeproject.org/batch_hub/searchTerm=H3K4me3+liver Debug Help <input type="button" value="Retry Hub"/>
<input type="button" value="Disconnect"/>	Roadmap Epigenomics Release III at Wash U VizHub	Roadmap Epigenomics Human Epigenome Atlas Release III, VizHub at Louis
<input type="button" value="Disconnect"/>	Hub (ENCSR442ZOI)	ENCODE Data Coordination Center Data Hub

实验万事屋

然后会弹出UCSC的封面，输入HOTAIR后进入：

UNIVERSITY OF CALIFORNIA SANTA CRUZ UCSC Genome Browser Gateway

Genomes Genome Browser Tools Mirrors Downloads My Data Help About Us

Browse/Select Species

POPULAR SPECIES

Human Mouse Rat Equine Worm Yeast

REPRESENTED SPECIES

- Human
- Chimp
- Bonobo
- Gorilla
- Orangutan
- Gibbon
- Crab-eating macaque
- Rhesus
- Baboon (anubis)
- Baboon (hamadryas)
- Marmoset
- Squirrel monkey

Find Position

Human Assembly
Feb. 2009 (GRCh37/hg19)

Position/Search Term
Enter position, gene symbol or search terms
Current position: chr21:33,031,597-33,041,570

Human Genome Browser - hg19 assembly

The February 2009 human reference sequence (GRCh37) was produced by the Genome Reference Consortium. For more information about this assembly, see GRCh37 in the NCBI Assembly database.

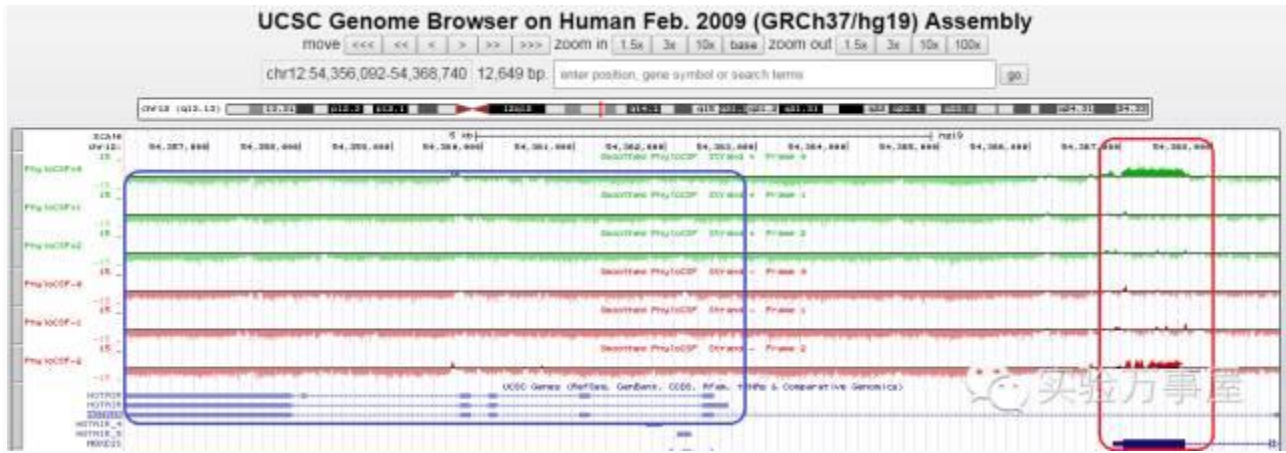
Sample position queries

A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the User Guide for more information.

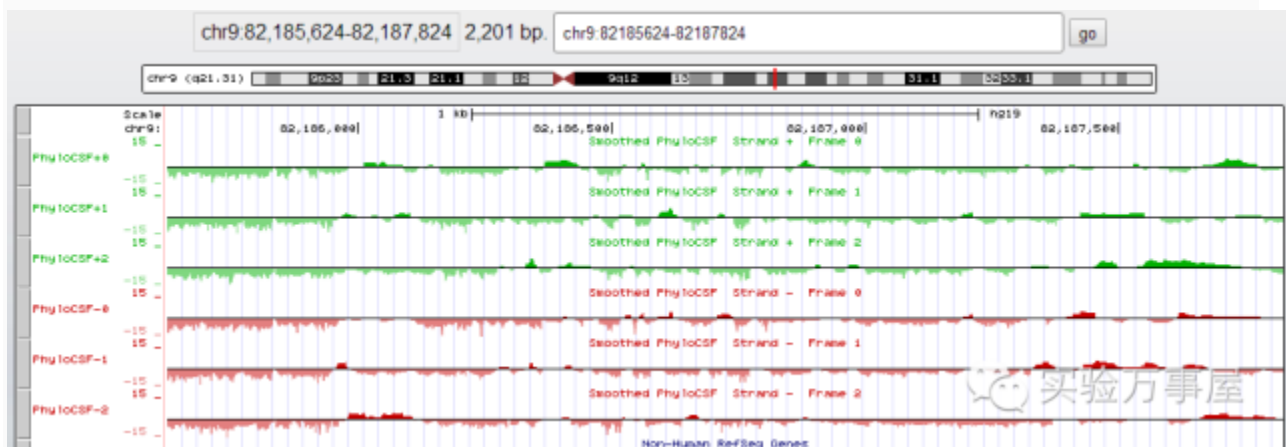
Request: chr7 Genome Browser Response: Displays all of chromosome 7

实验万事屋

结果会直接显示HOTAIR的PhyloCSF值，可以明显地看到，在HOTAIR的外显子上所有的值都是小于0的，也就是没有保守型。



那我们把那篇Cell中的lncRNA的序列位置输入进去，然后.....



可以看到，也没什么保守性。以此我们可以初步判断，这个RNA极有可能不能编码蛋白质，也就是lncRNA。

...华丽的分割线...

李莫愁博士：我估计好多人不会来看这个帖子呢，因为太长了，但这是一个lncRNA确认的基本步骤。最实际的，就比如通过二代测序后获得有差异的，可能不能编码蛋白的RNA，那要用什么来验证呢？这篇Cell告诉我们要用ORF和CSF来验证是否是lncRNA。

其实验证ORF之前，其实还有一个问题大家可能也不会去注意，那就是Kozak序列，Kozak序列是核糖体结合位点，没有这个，其实再怎么样的阅读框也没办法翻译成蛋白。然而有一些lncRNA是具有翻译短肽功能的，还有一些假基因，这就很难用这样的方法来确认了