Top 20 Useful Bioinformatic Tools

Promoter Scan

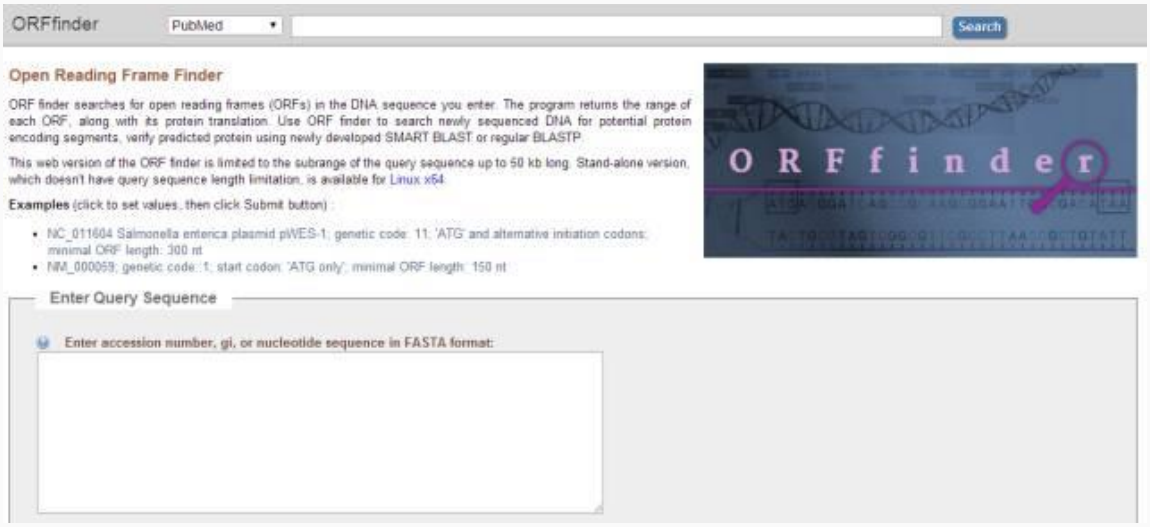
功能：启动子预测

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

ORF Finder

功能：ORF预测

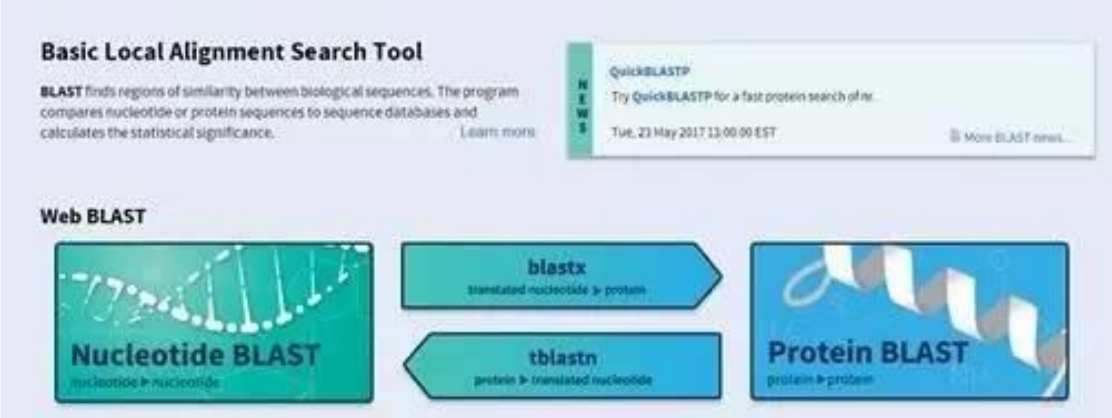
网址：https://www.ncbi.nlm.nih.gov/orffinder/



NCBI-BLAST

功能：序列比对

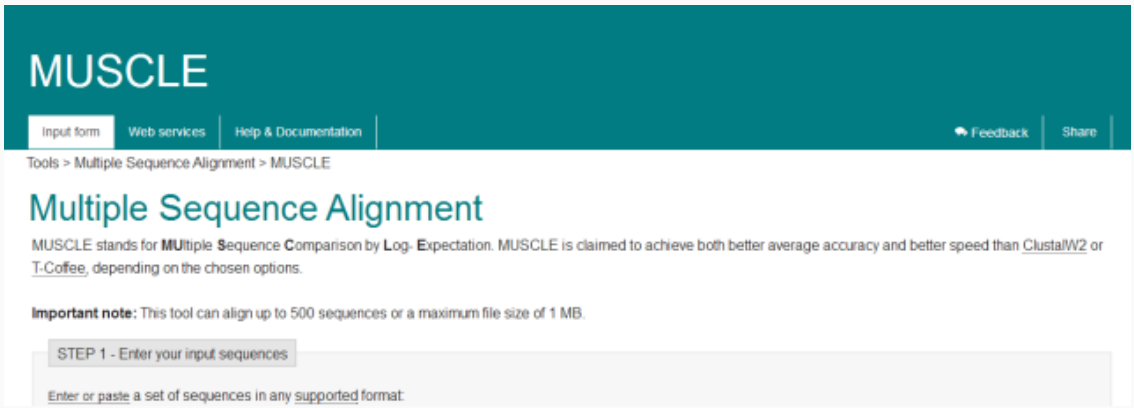
网址：https://blast.ncbi.nlm.nih.gov/Blast.cgi



MUSCLE

功能：运行速度比较快的多序列比对

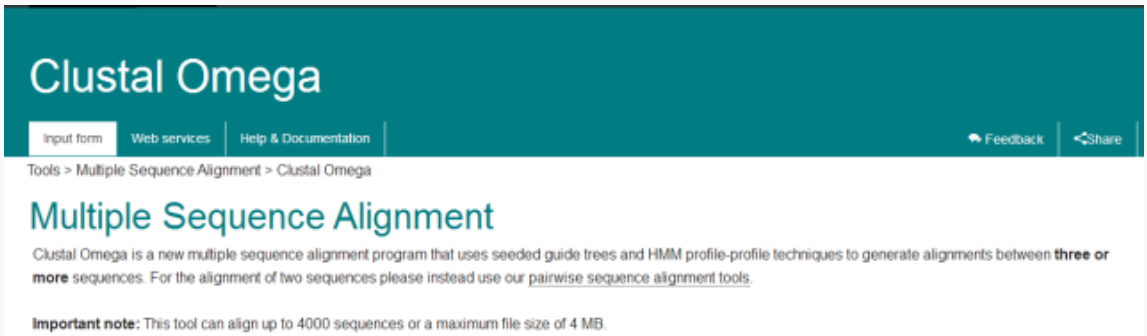
网址：http://www.ebi.ac.uk/Tools/msa/muscle/#



Clustal Omega

功能：DNA、RNA、蛋白的多序列比对

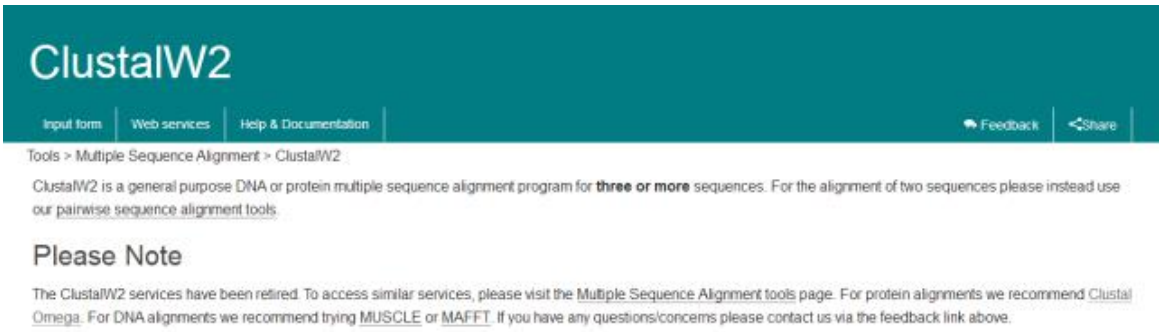
网址：http://www.ebi.ac.uk/Tools/msa/clustalo/



ClustalW2

功能：应用较广泛的多序列比对

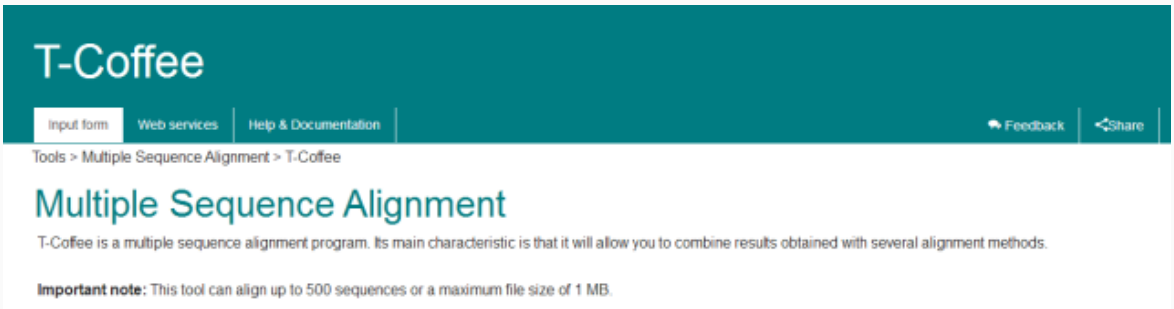
网址：http://www.ebi.ac.uk/Tools/msa/clustalw2/



T-Coffee

功能：准确度高,速度慢的多序列比对

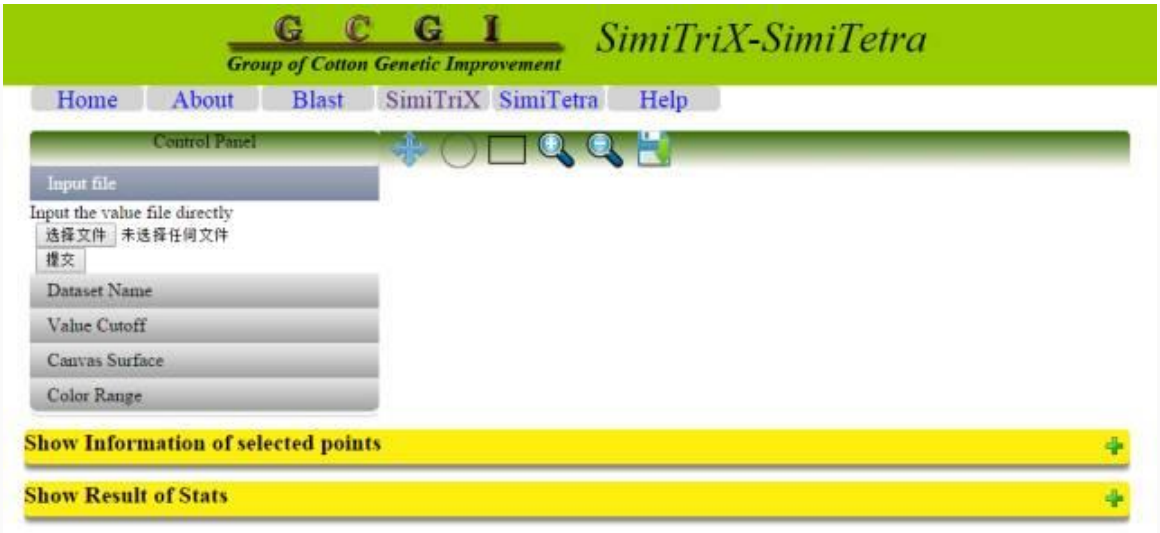
网址：http://www.ebi.ac.uk/Tools/msa/tcoffee/



SimiTriX-SimiTetra

功能：多序列比对相似性展示

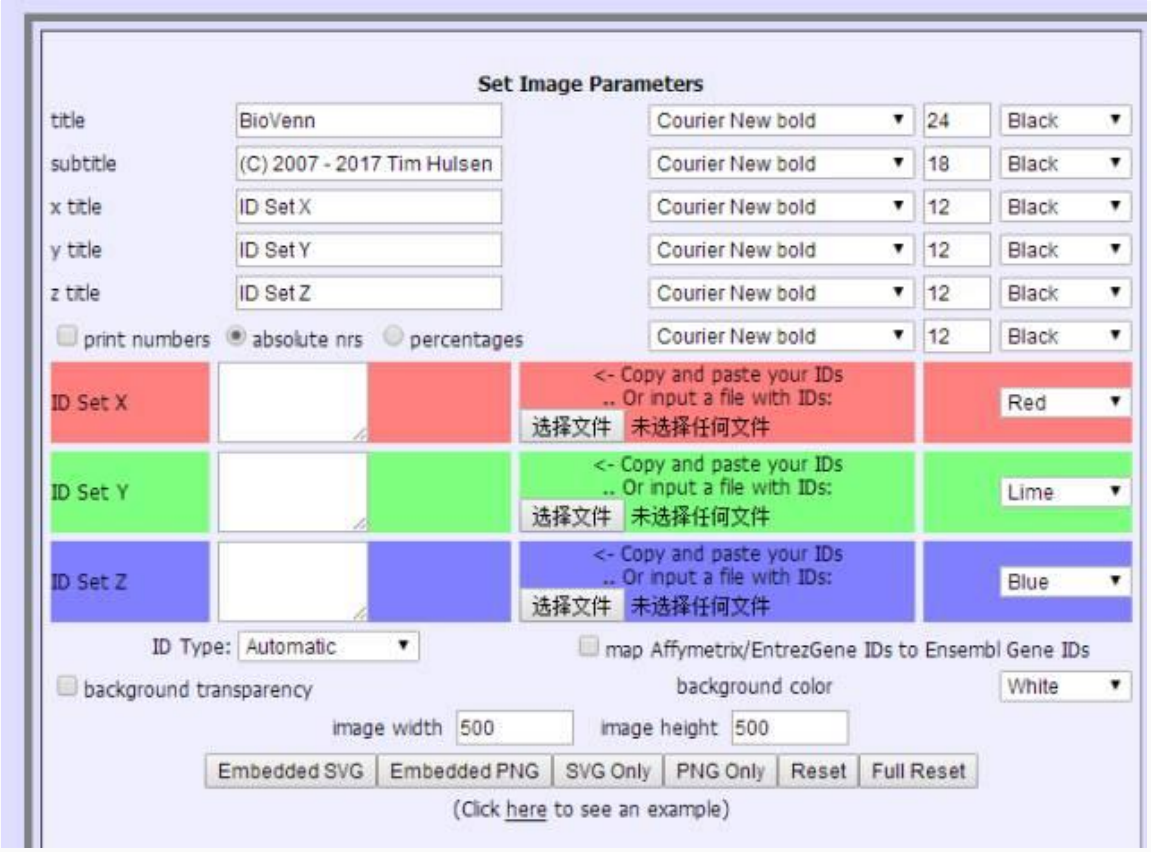
网址：http://cotton.hzau.edu.cn/EN/tools/BioERCP/simitrix.php



Venn图

功能：绘制Venn图

网址：http://www.biovenn.nl/index.php



Venn图

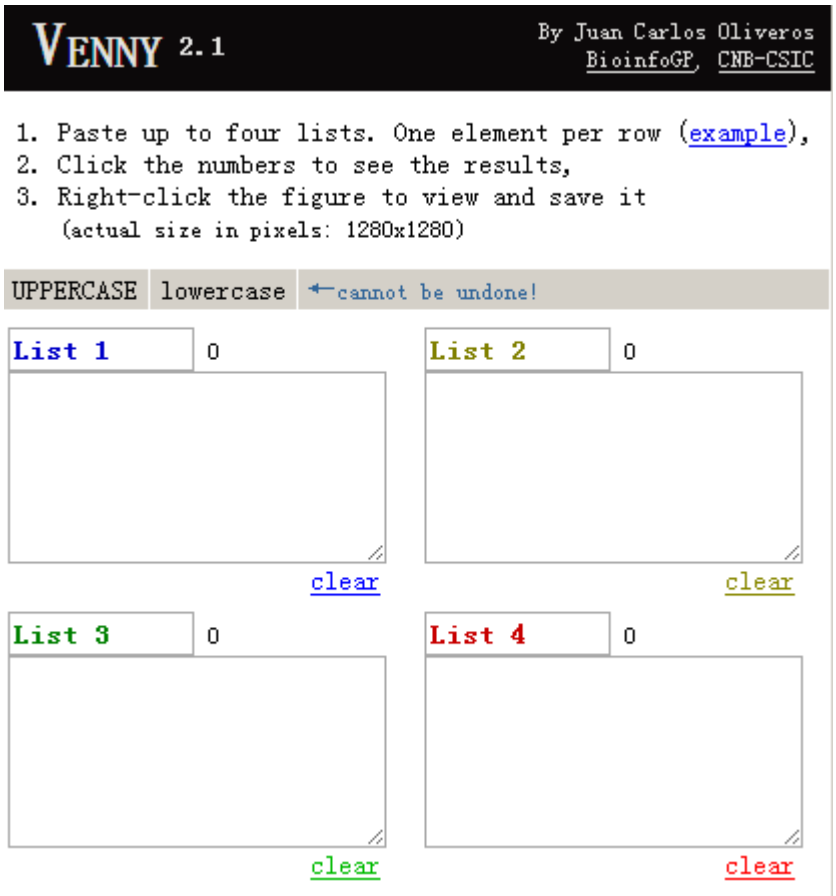
功能：绘制Venn图

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

Venn图

功能：绘制Venn图

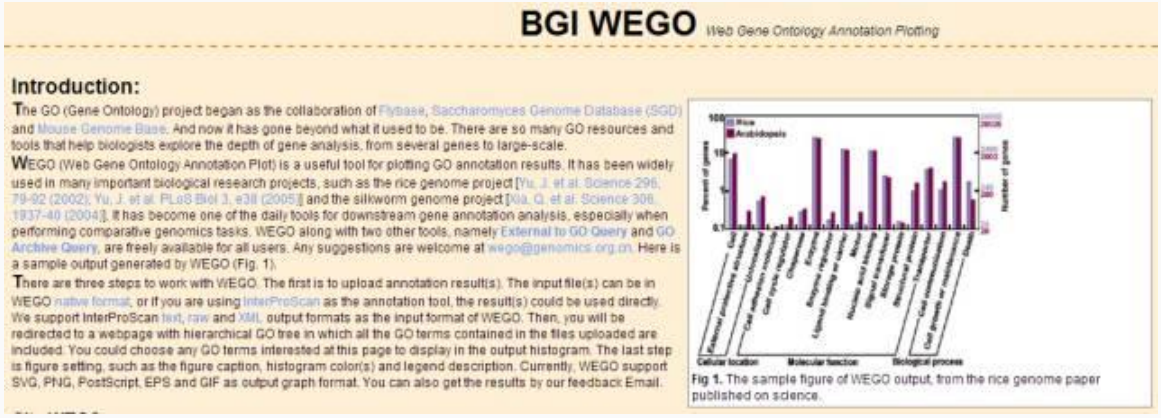
网址：http://bioinfogp.cnb.csic.es/tools/venny/index.html



WEGO

功能：绘制GO注释结果图

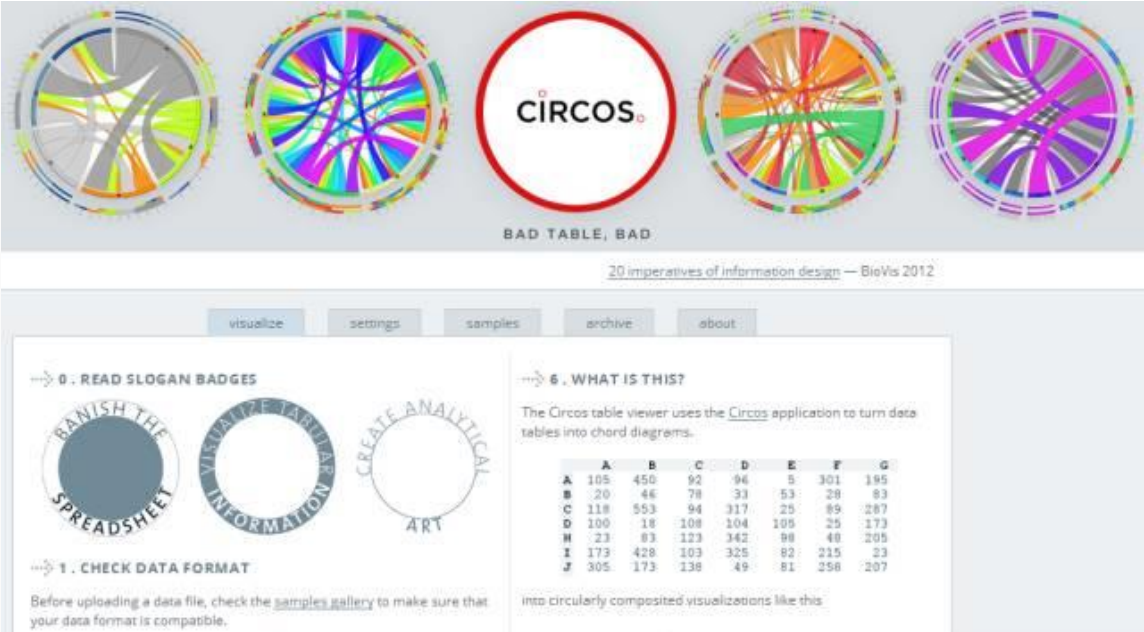
网址：http://wego.genomics.org.cn/cgi-bin/wego/index.pl



CIRCOS

功能：绘制圈图

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



CIRCexplorer

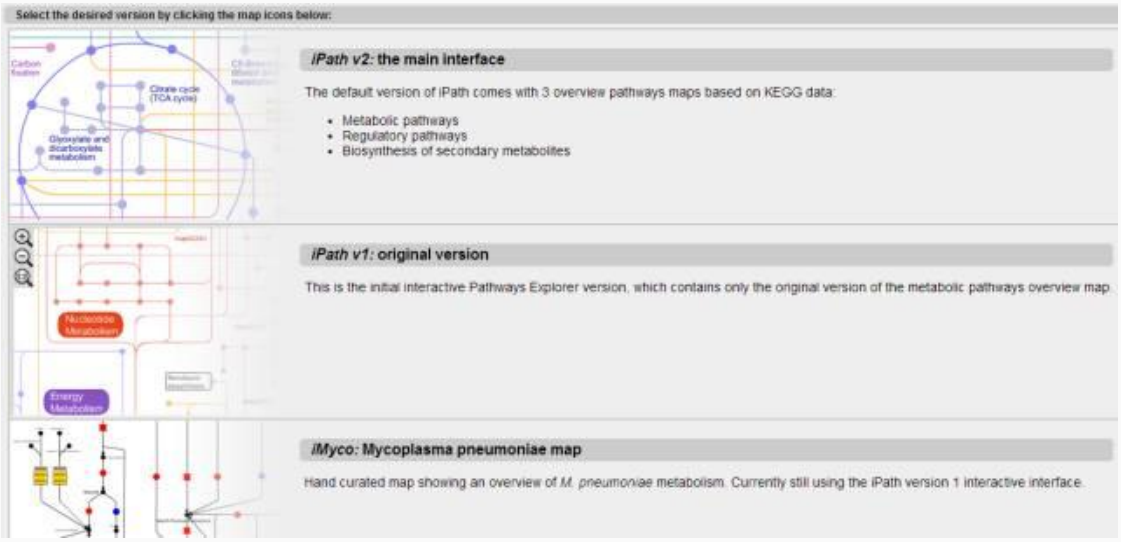
功能：进行circRNA分析

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

iPath

功能：进行可视化通路图在线分析

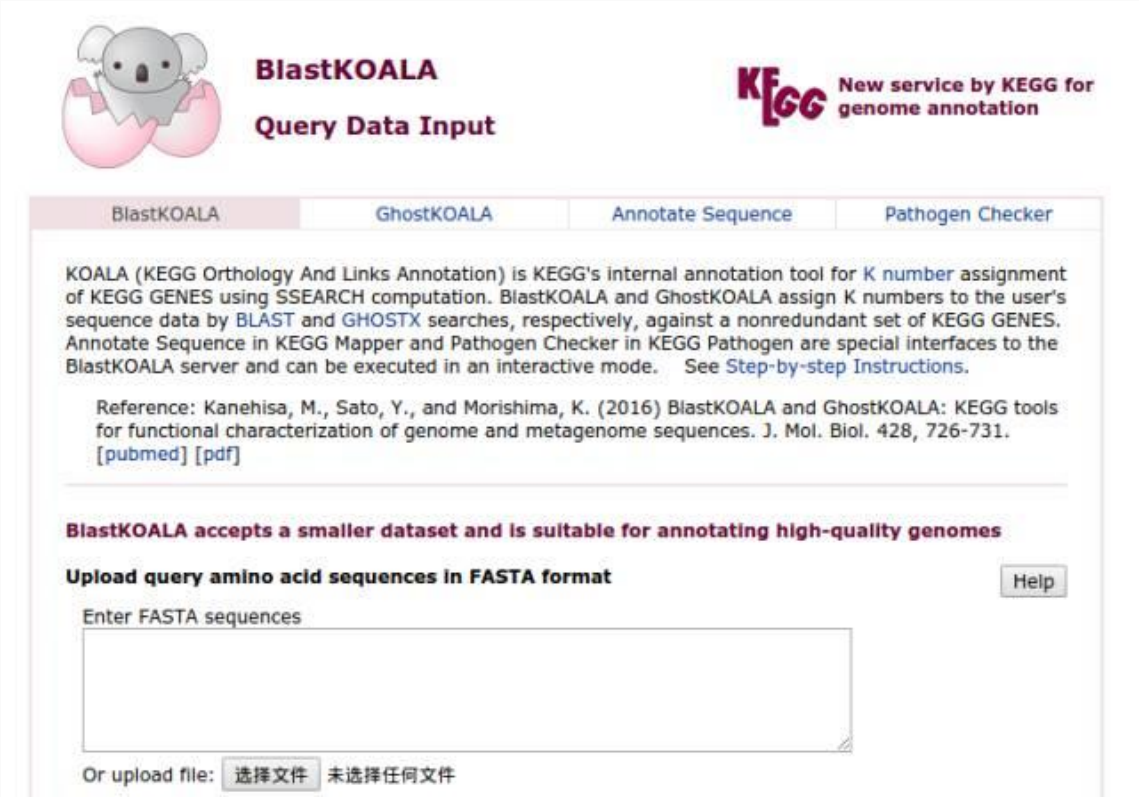
网址：http://pathways.embl.de/



KEGG

功能：进行基因的代谢通路注释

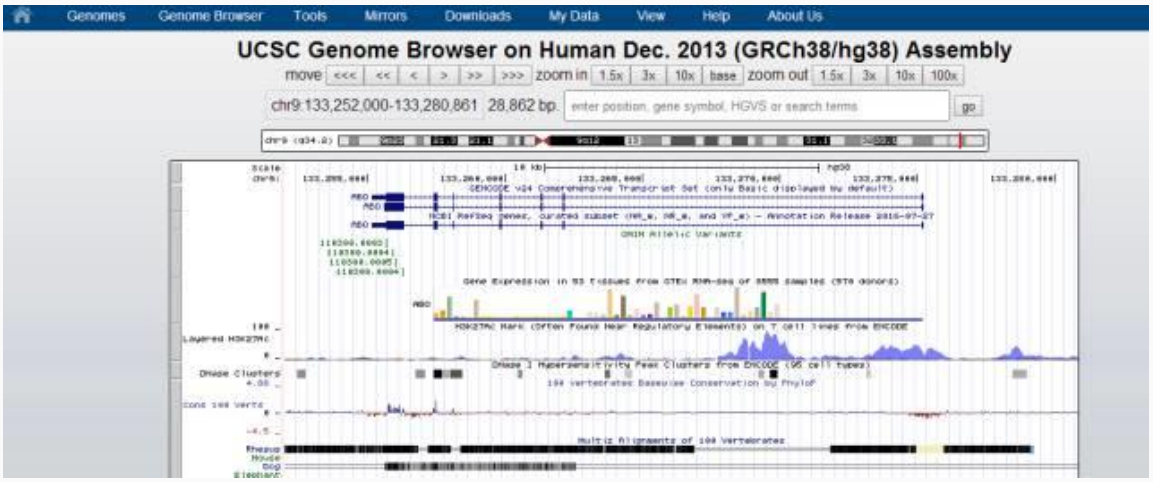
网址：http://www.kegg.jp/blastkoala/



UCSC

功能：进行基因组可视化

网址：http://genome.ucsc.edu/index.html



IBS

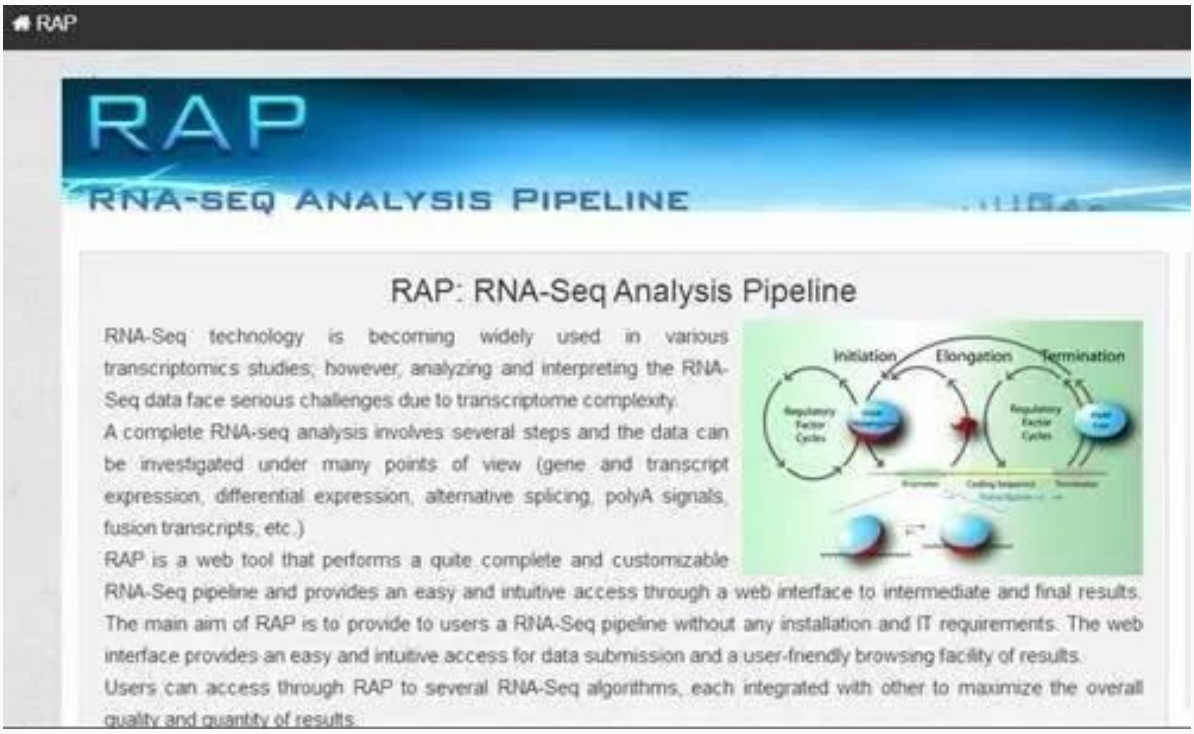
功能：进行序列结构示意图绘制

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

RAP

功能：在线分析RNA-seq

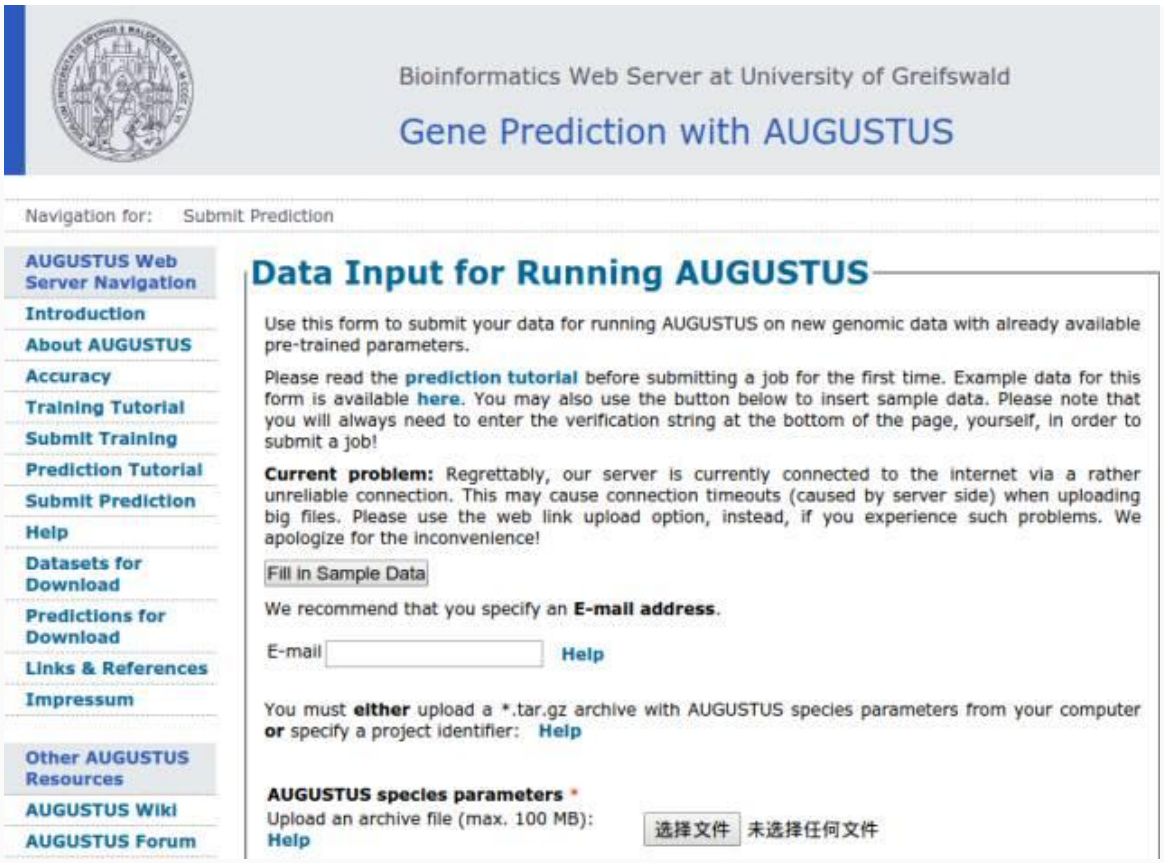
网址：https://bioinformatics.cineca.it/rap/



AUGUSTUS

功能：基因外显子内含子，UTR,注释

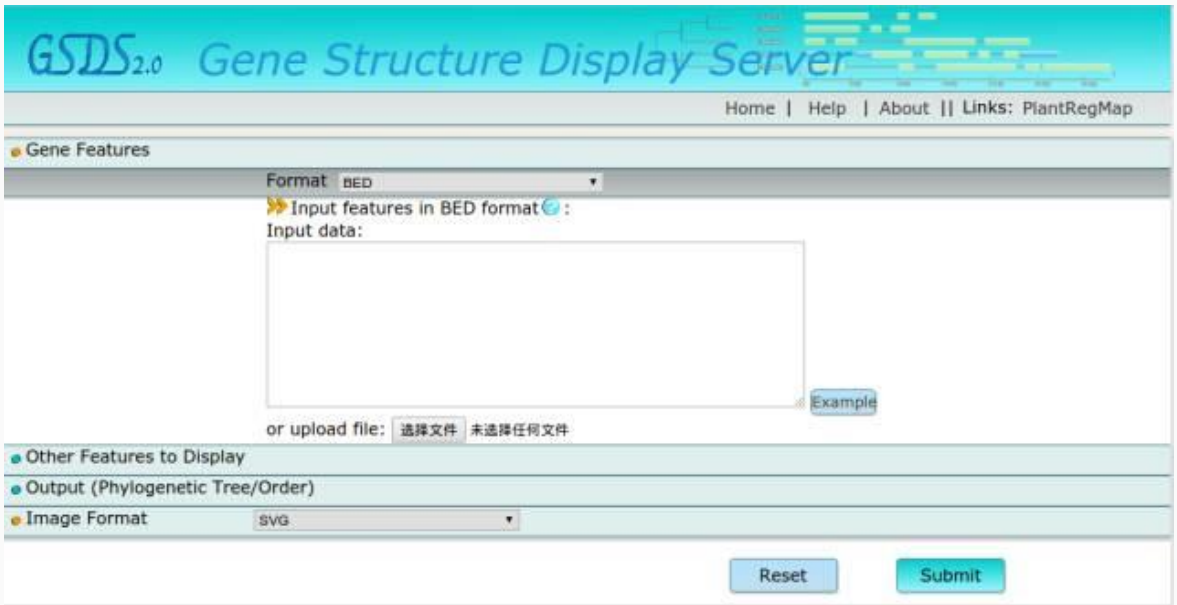
网址：http://bioinf.uni-greifswald.de/webaugustus/prediction/create



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/



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

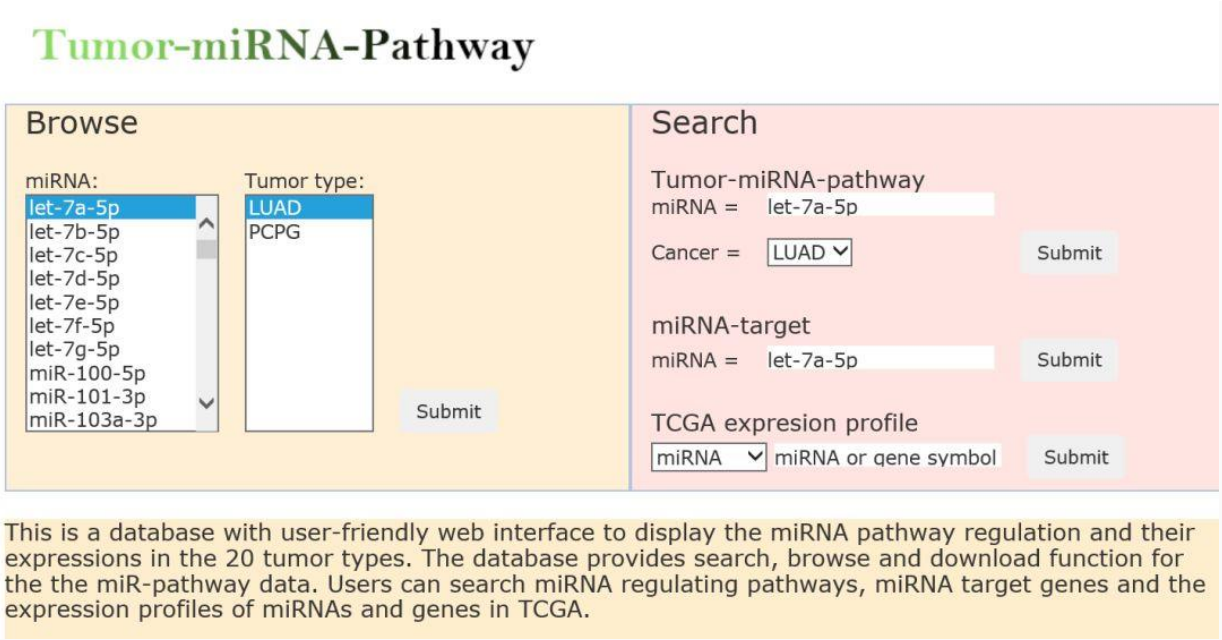
Role of miR-452-5p in the tumorigenesis of prostate cancer: A study based on the Cancer Genome Atl(TCGA), Gene Expression Omnibus (GEO), and bioinformatics analysis

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：

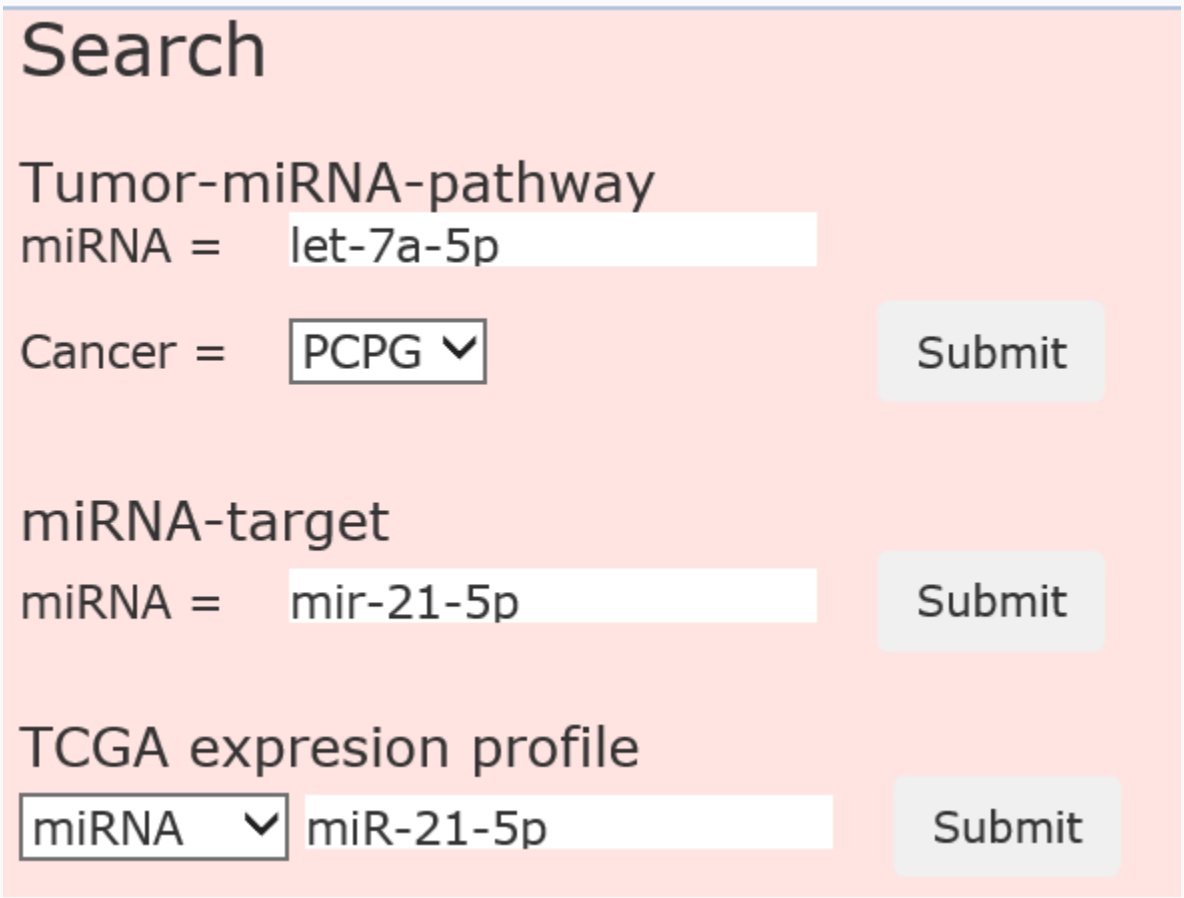
2.1. The clinical significance of miR-452-5p in prostate cancer from TCGA

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).

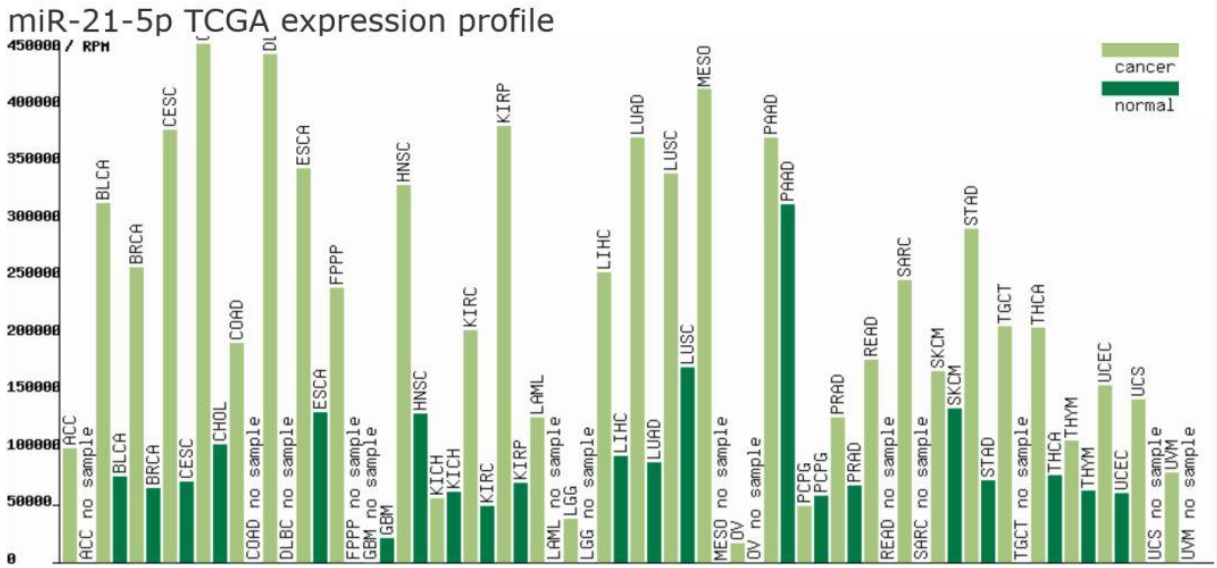
在这里，蚊帐里说到，他们分析了TCGA的数据。嗯，是TCGA中miRNA的表达数据，没错。其实他们用到的是这样一款网页神器：



Tumor-miRNA-Pathway。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：



直接搜索miRNA就行了，我就搜了一个MiR-21-5p：



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

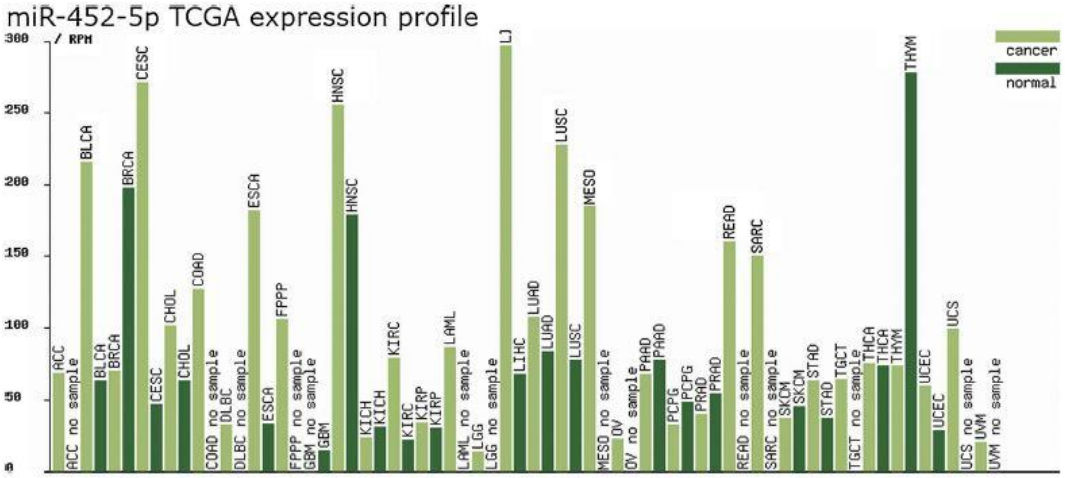
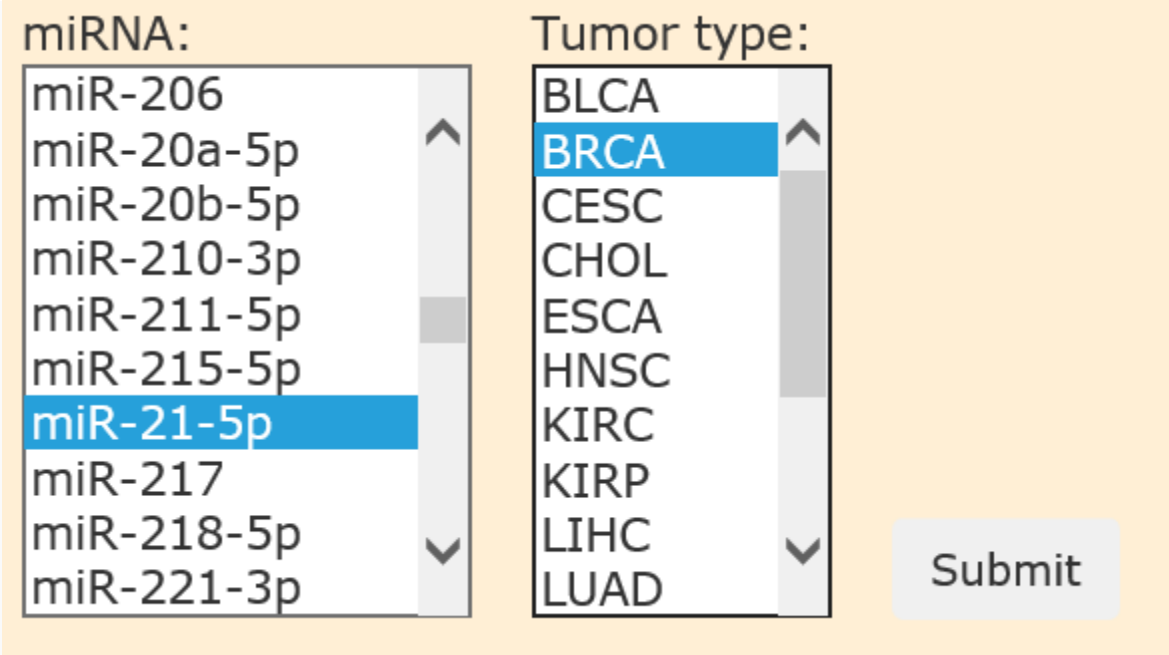


Fig. 1. Expression profile of miR-452-5p from TCGA. MiR-452-5p was down-regulated in prostate adenocarcinoma tissues compared with normal tissues.

好吧，其实这个神器还有别的功能，就是分析miRNA的信号通路：



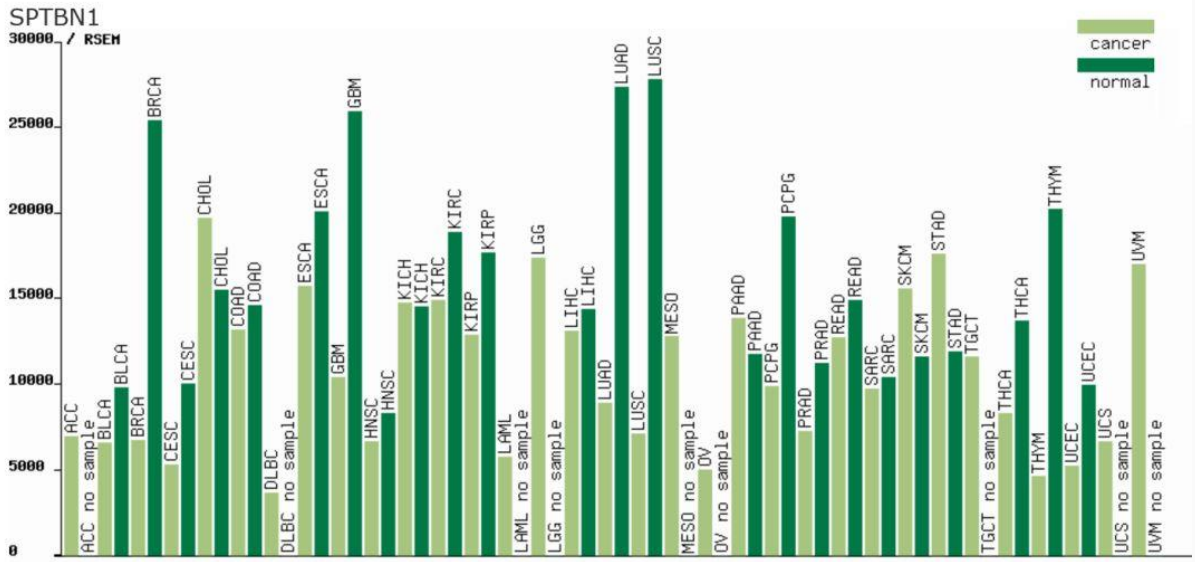
就随便浏览一下关于miR-21-5p在乳腺癌中的信号通路，当然分析的数据也都是基于TCGA的，然后就得到：

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。

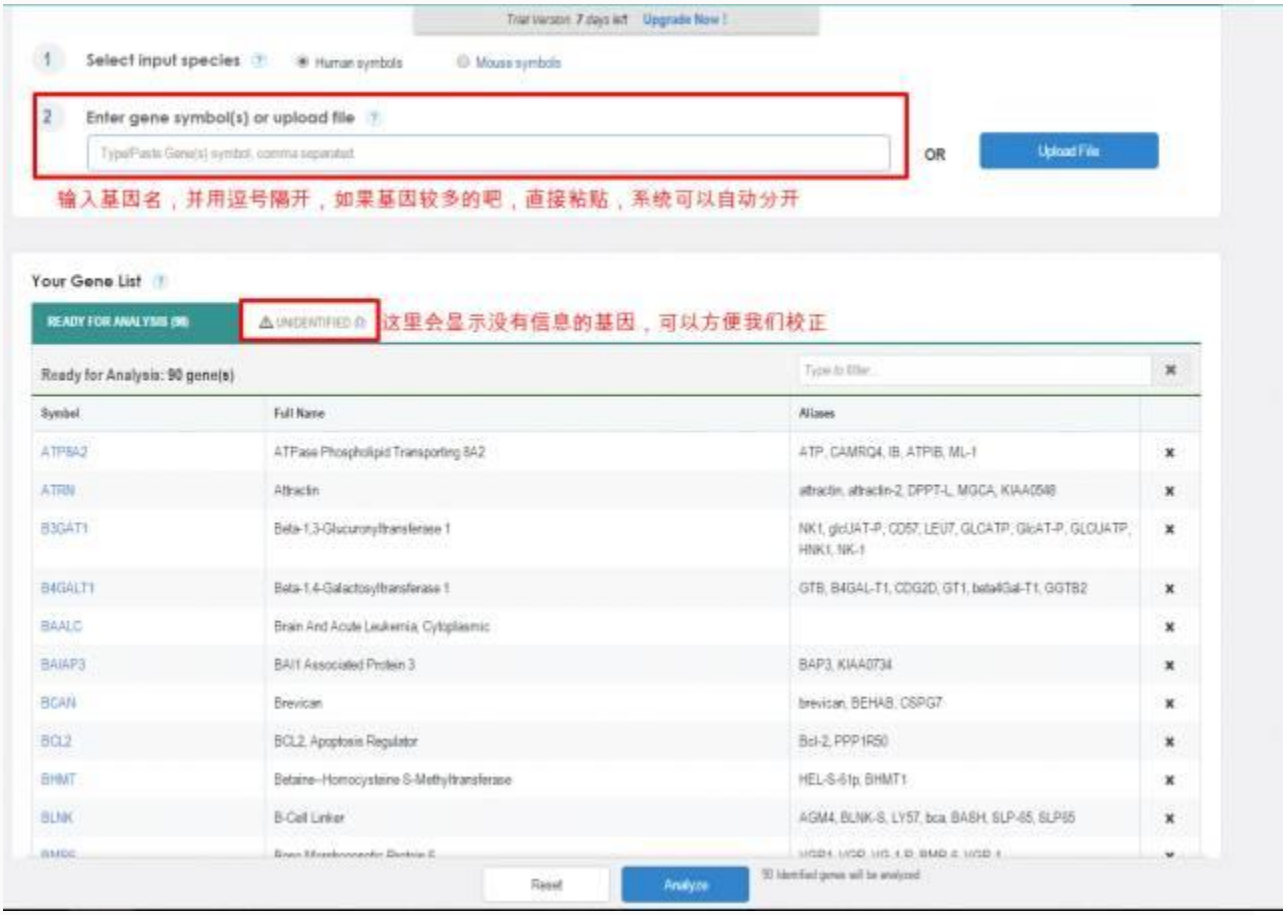
分享4个基因分析的网址

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）。

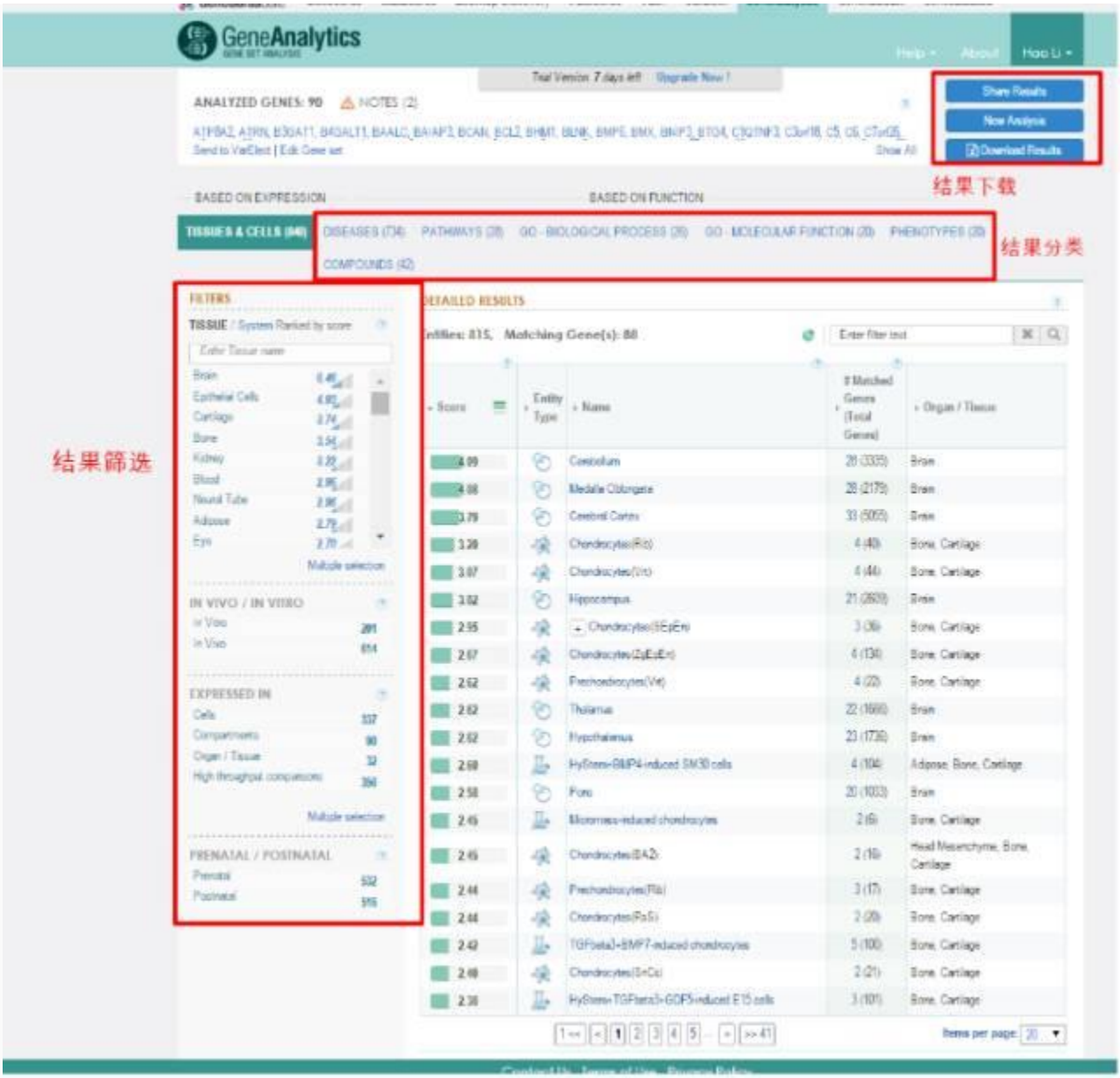
今天给大家接着讲这两个数据库里好用的小工具。需要提醒的是，这些分析需要教育机构的邮箱进行注册，如果没有的请火速读博，从此过上幸福生活（博士的幸福生活I, II, III）。

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。



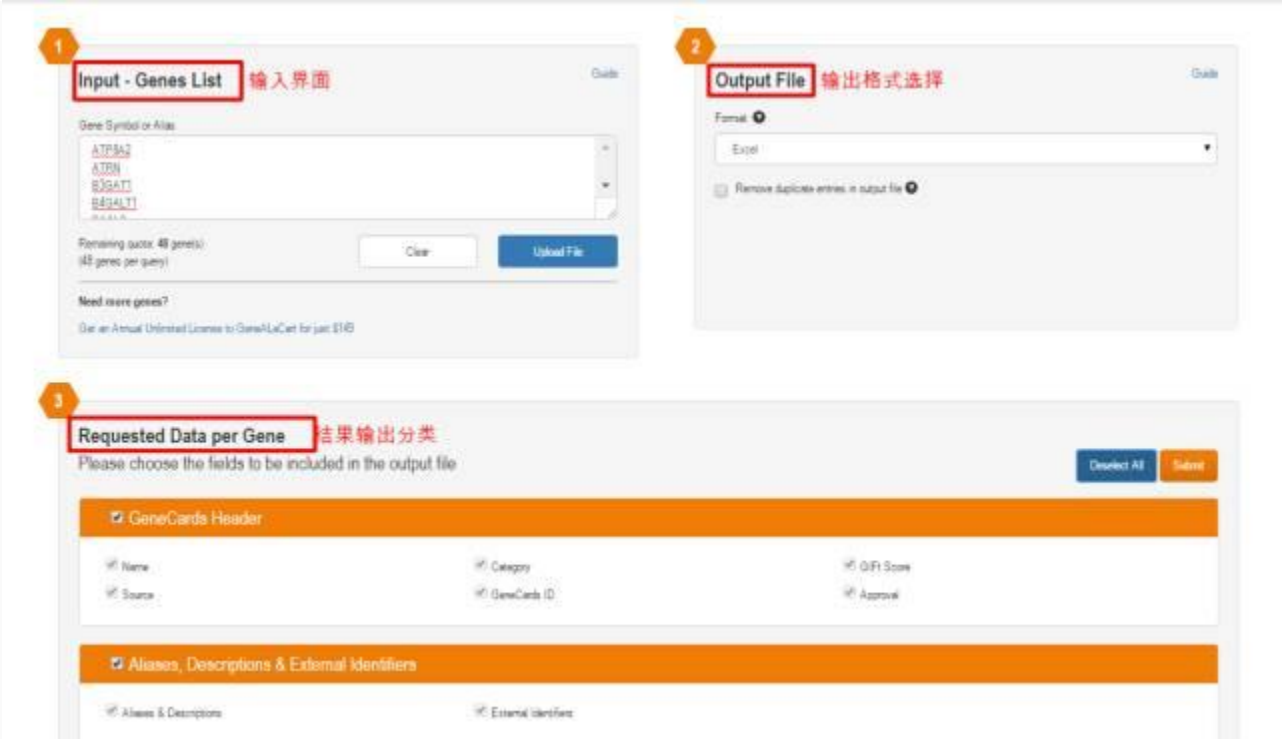
我们点击分析以后，会发现看到这些基因共同的汇总信息，如图



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):

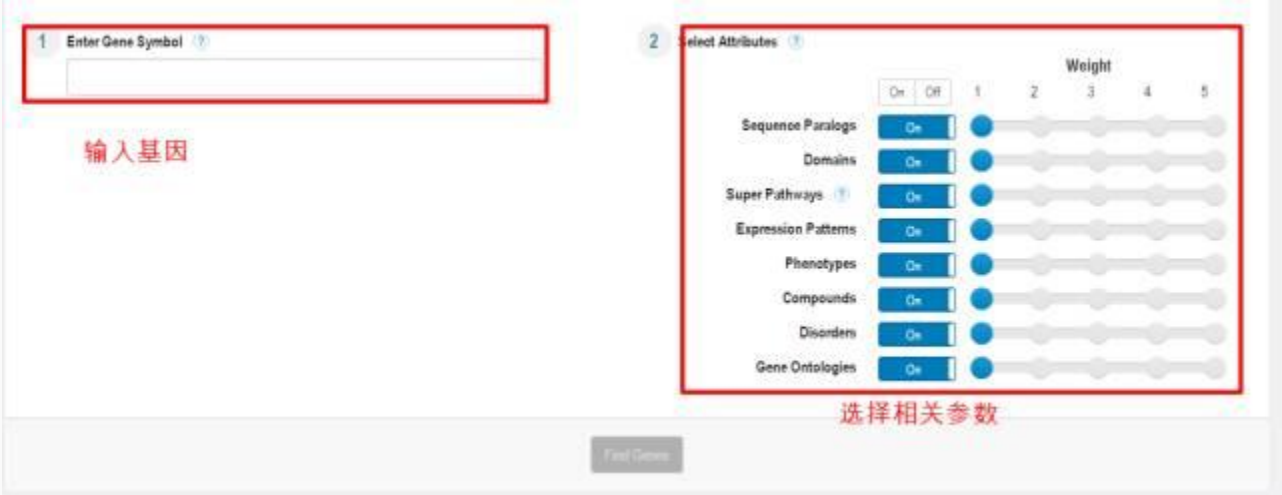
假如我有一些基因，我想下载这些基因所有各自相关的信息，那么我就可以用这个软件。网页的初始界面如图，我们只需要填入基因，选择自己想要保存的分类内容，然后点击提交即可，然后结果会以excel的形式保存下来。



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 phenotyeontology, 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

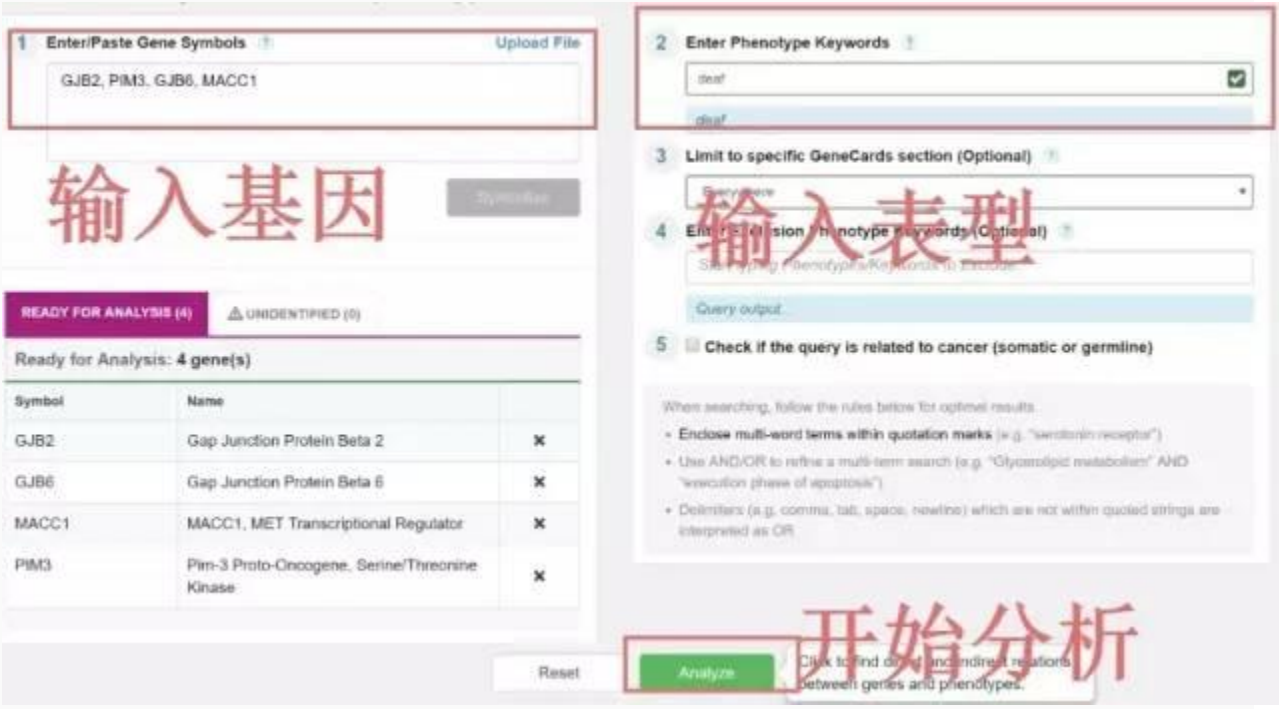


我们来输入TP53，即可看到结果如下，排在前面的TP63\TP73为TP53同家族的：

Weights		1	1	1	1	1	1	1	1
# Symbol	Total Score	Sequence Paralogs	Domains	Super Pathways	Expression Patterns	Phenotypes	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
E2F1	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	0.68	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	0.58	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	0.05	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	0.84	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)。



结果如下，GJB2与GJB6与deaf相关，其中GJB2最相关。

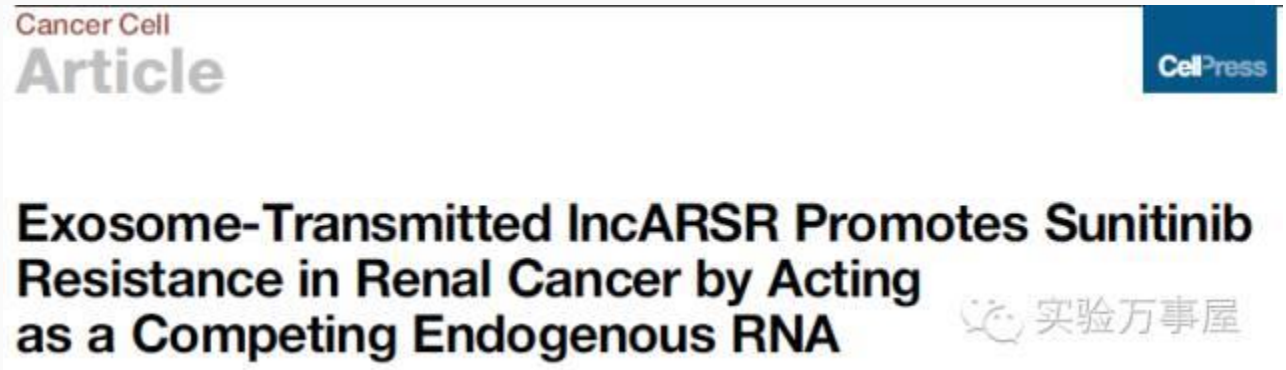
#	Symbol	Description	Type	Score
1	GJB2	Gap Junction Protein Beta 2	Protein	96.95
2	GJB6	Gap Junction Protein Beta 6	Protein	59.06

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：



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

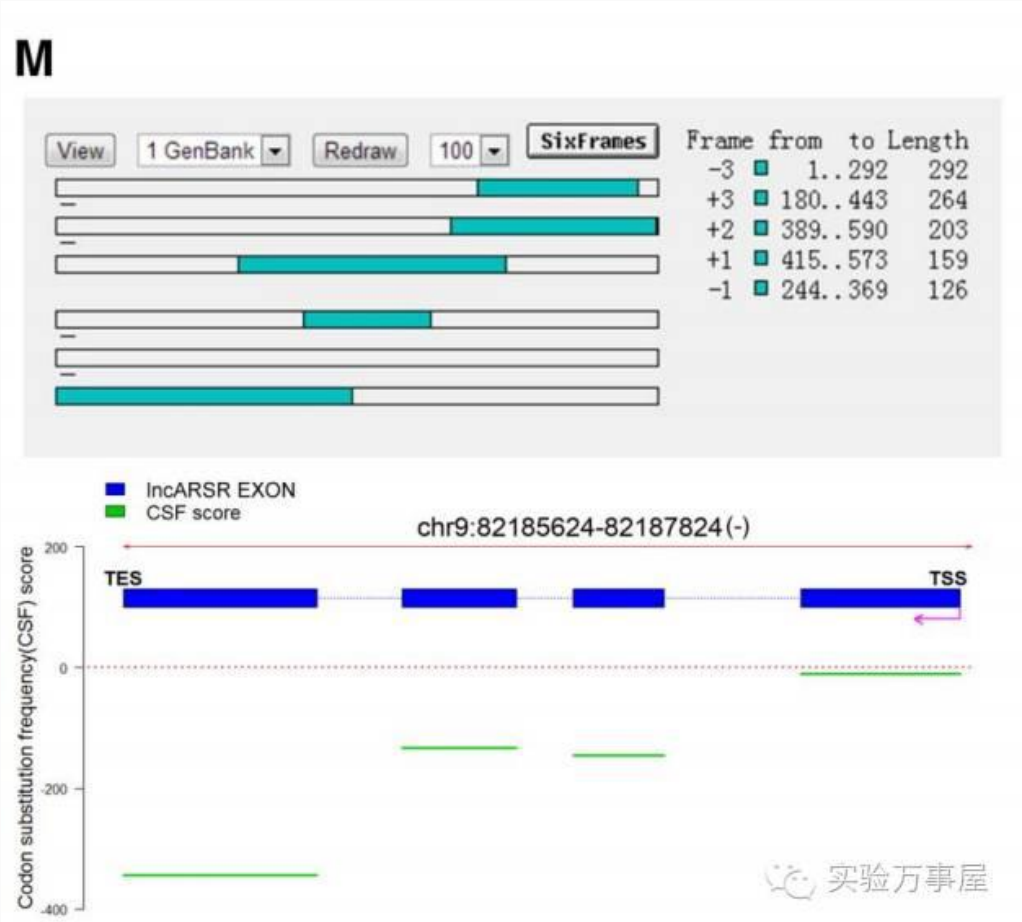


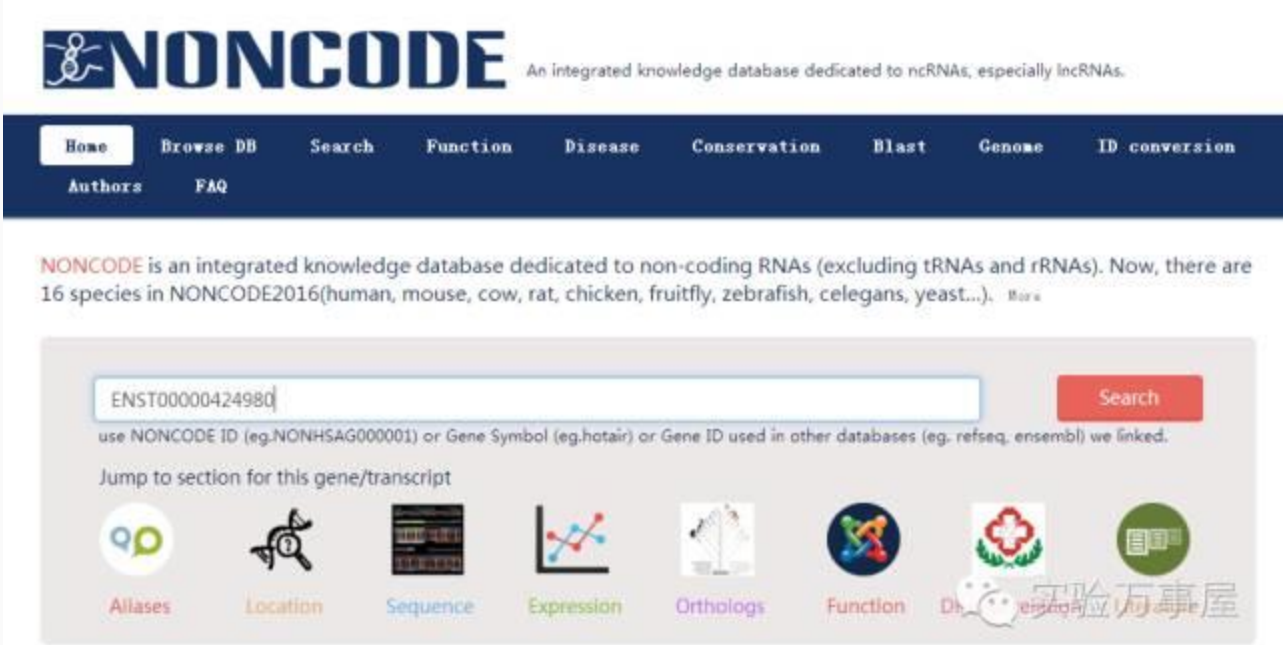
Fig. legend是这样写的：(M) Upper: Prediction of putative proteins encoded by lncARSR using ORF Finder. Lower: The codon substitution frequency scores (CSF) of lncARSR.

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。

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：

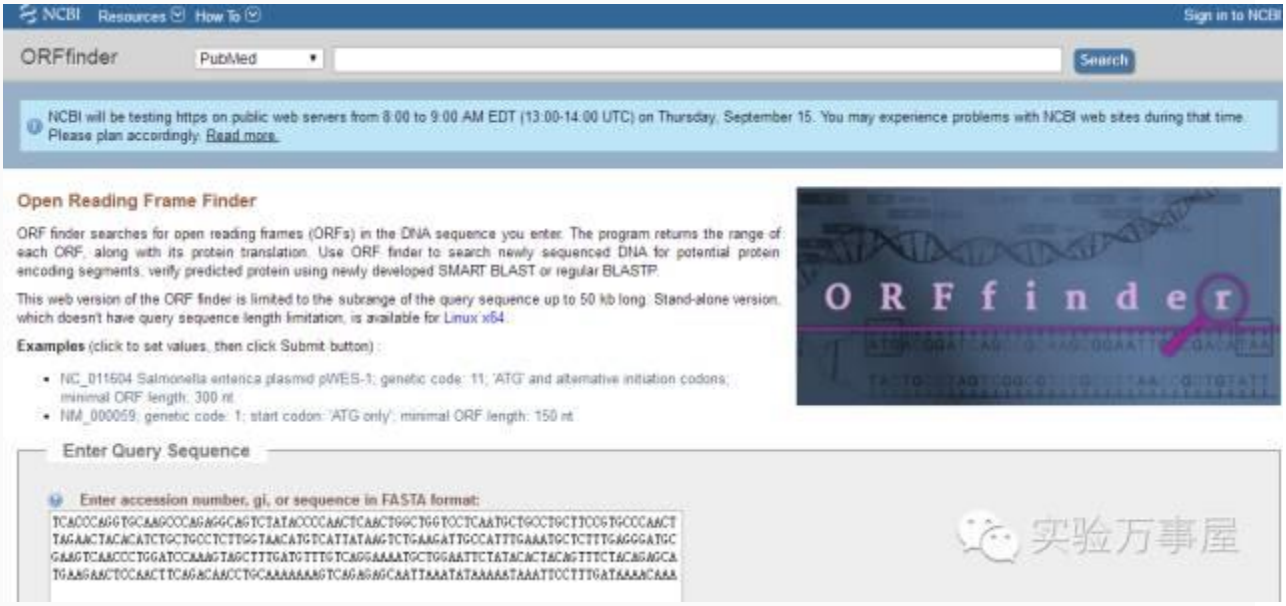
General info

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)

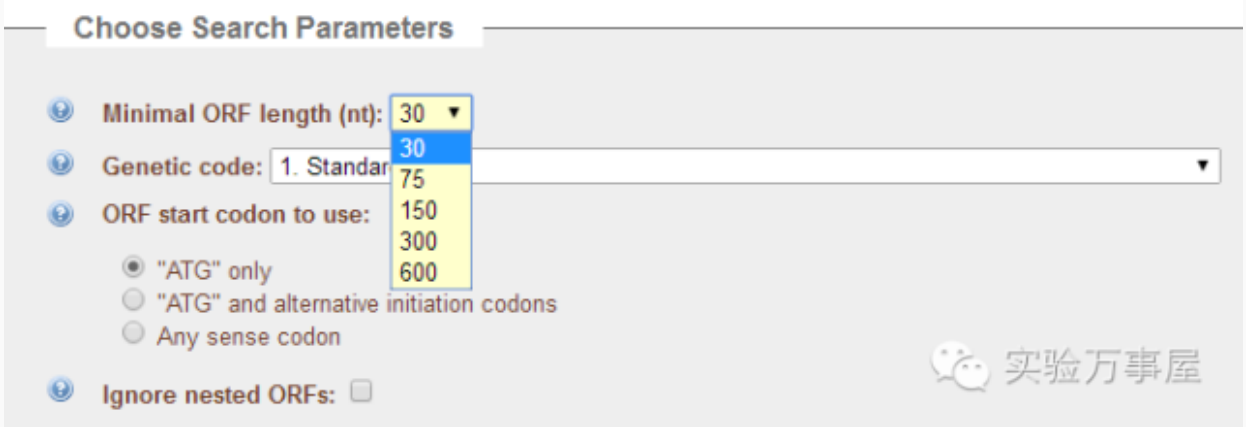
Sequence

>NONHSAT132007
TCACCCAGGTGCAAGCCCAGAGGCAGTCTATACCCCAACTCAACTGGCTGGTCCTCAATGCTGCCTGCTTCCGTGCCCAACTTAGA
GTCATTATAAGTCTGAAGATTGCCATTTGAAATGCTCTTTGAGGGATGCGAAGTCAACCCTGGATCCAAAGTAGCTTTGATGTTTGTCAGGAAATGCTGGAATTCTATACACTACA
GTTTCTACAGAGCATGAAGAACTCCAACTTCAGACAACCTGCAAAAAAAGTCAGAGAGCAATTAAATATAAAAATAAATTCCTTTGATAAAACAAA

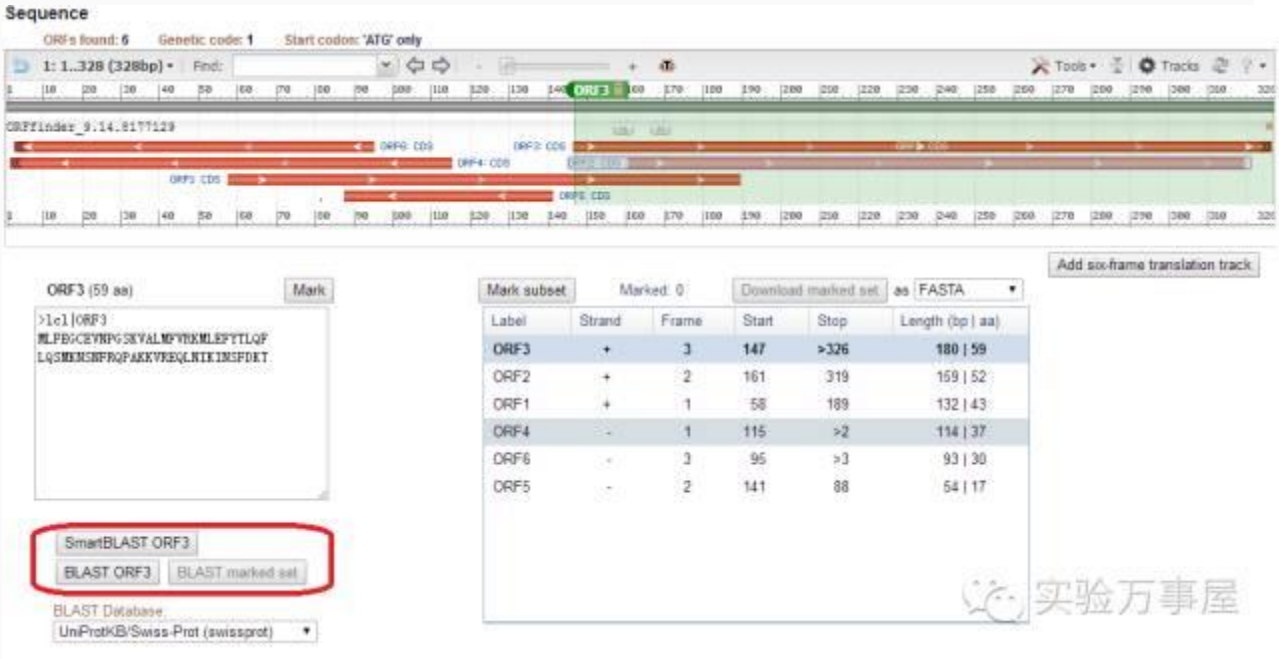
那接下来，验证这个RNA到底是不是lncRNA呢？首先我们要了解的，就是lncRNA是不能编码的，那就没有足够的ORF，也就是开放式阅读框。那我们就登陆到PubMed的ORFfinder（http://www.ncbi.nlm.nih.gov/orffinder/）上去。



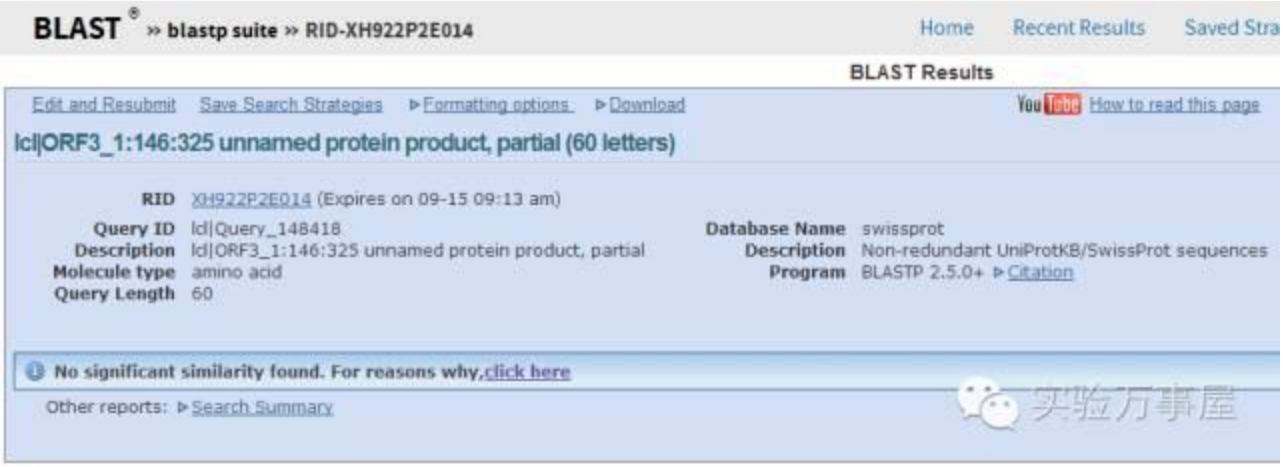
调整一下，看看到底有多少个氨基酸（aa），我就调到了最低30个氨基酸的选项：



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。



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。



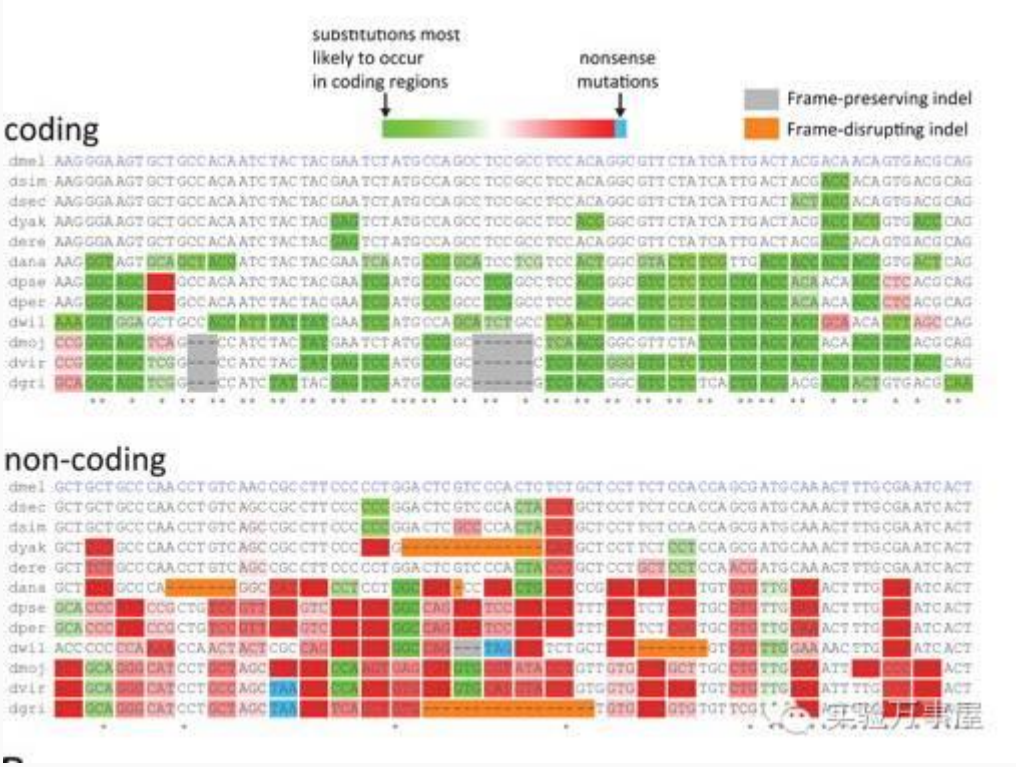
接着我们来看CSF，CSF到底是啥？CSF其实就是密码子的突变率。理论上编码区的密码子相对来说是保守的，也就是在物种中或者物种间是不容易产生突变，而非编码的就有点乱来了。我找到了这篇文献：


12 *Drosophila* Genomes/Letter

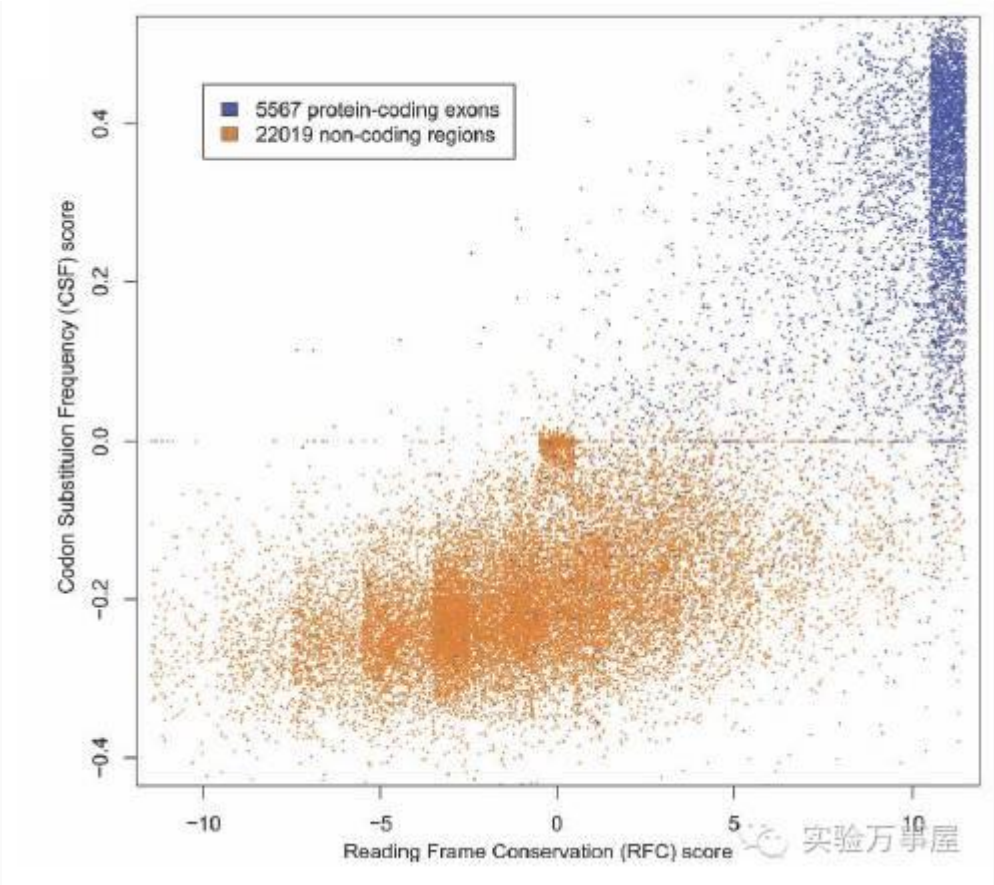
Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using 12 fly genomes

Michael F. Lin,[1] Joseph W. Carlson,[2] Madeline A. Crosby,[3] Beverley B. Matthews,[3]

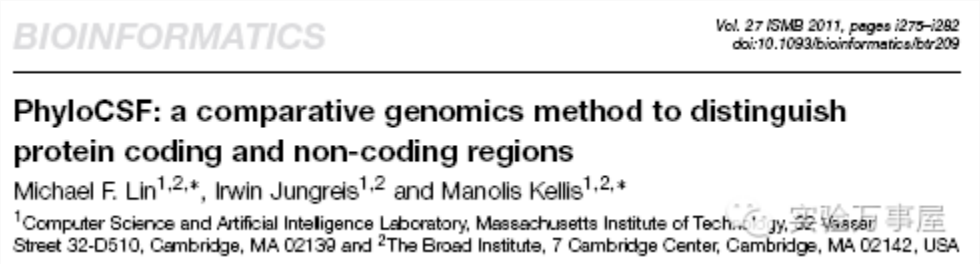
这是一篇在果蝇中用CSF来验证非编码与编码RNA间CSF差异的文献。其中显示，非编码的RNA突变率更高。



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：


BIOINFORMATICS

Vol. 27 ISMB 2011, pages i275–i282
doi:10.1093/bioinformatics/btr209

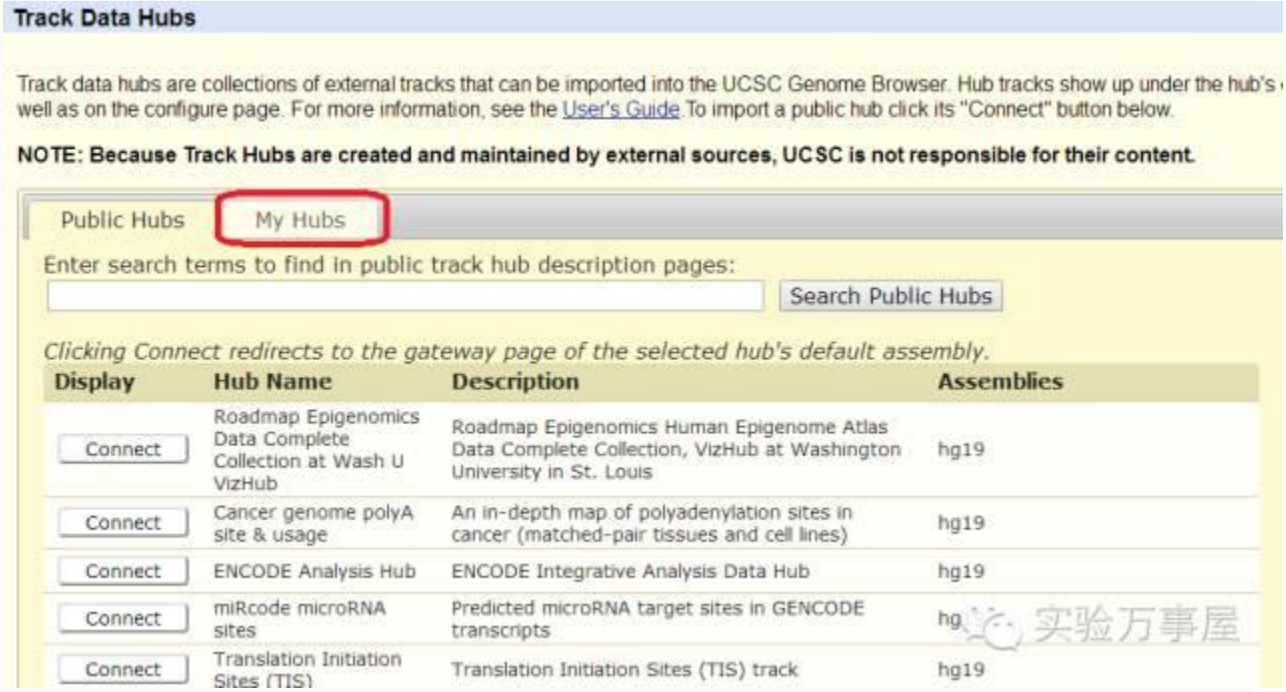
PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions

Michael F. Lin[1,2,*], Irwin Jungreis[1,2] and Manolis Kellis[1,2,*]

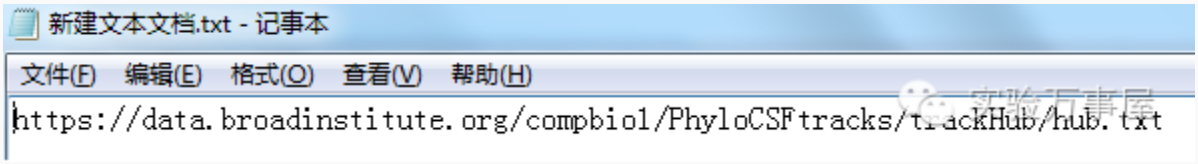
[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar Street 32-D510, Cambridge, MA 02139 and [2]The Broad Institute, 7 Cambridge Center, Cambridge, MA 02142, USA

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。

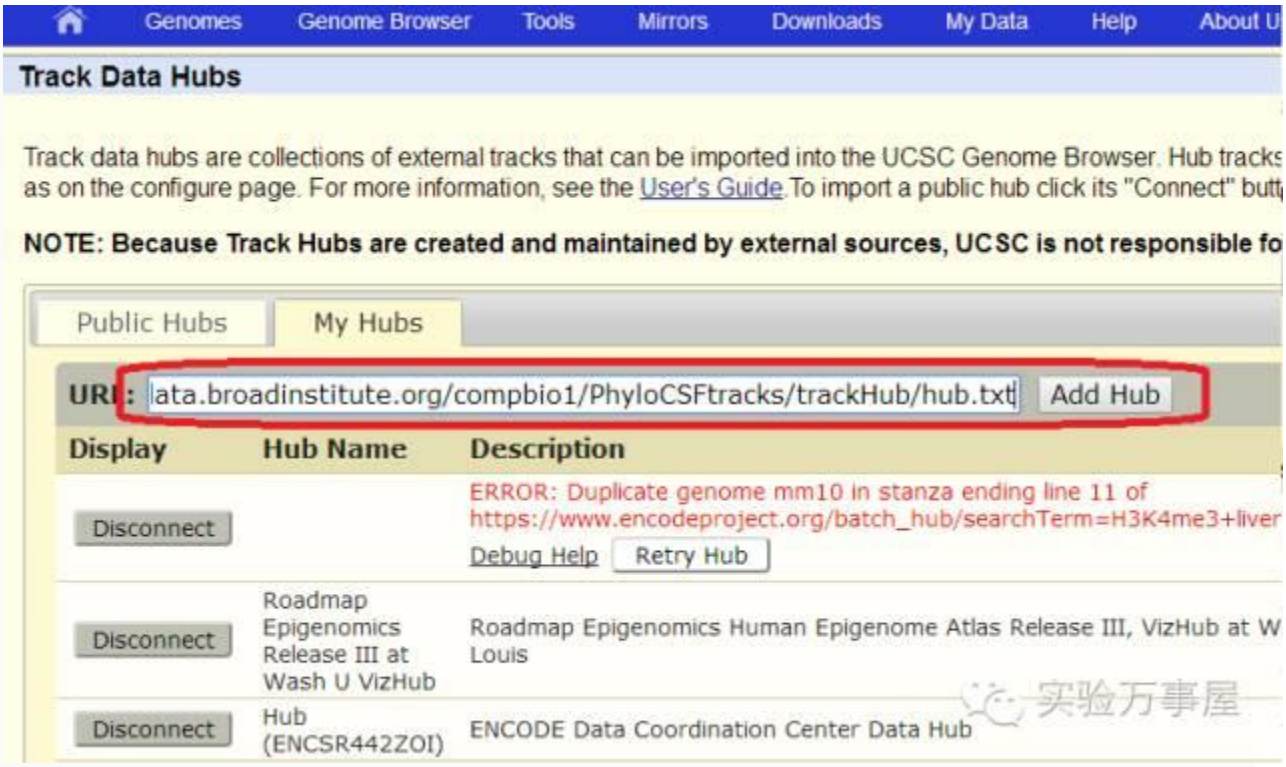
进去之后，选择“My Hubs”。



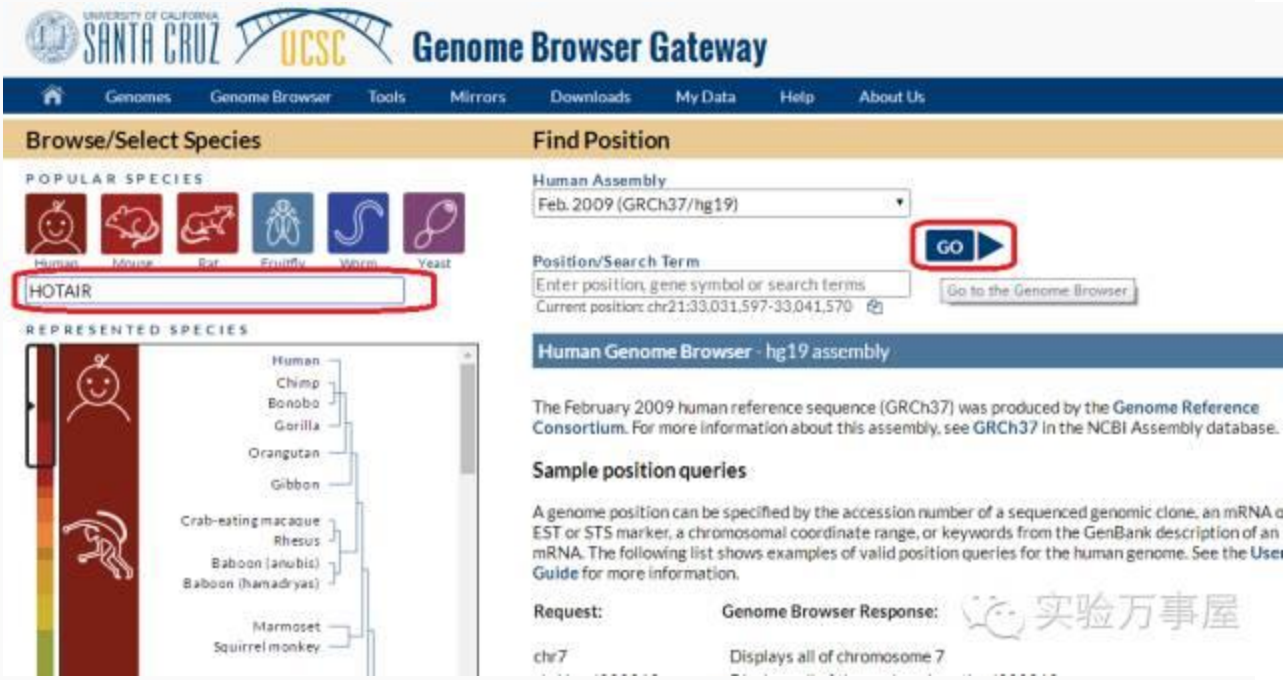
在里面添加这个网址，我知道你们懒，所以不能惯着你们：



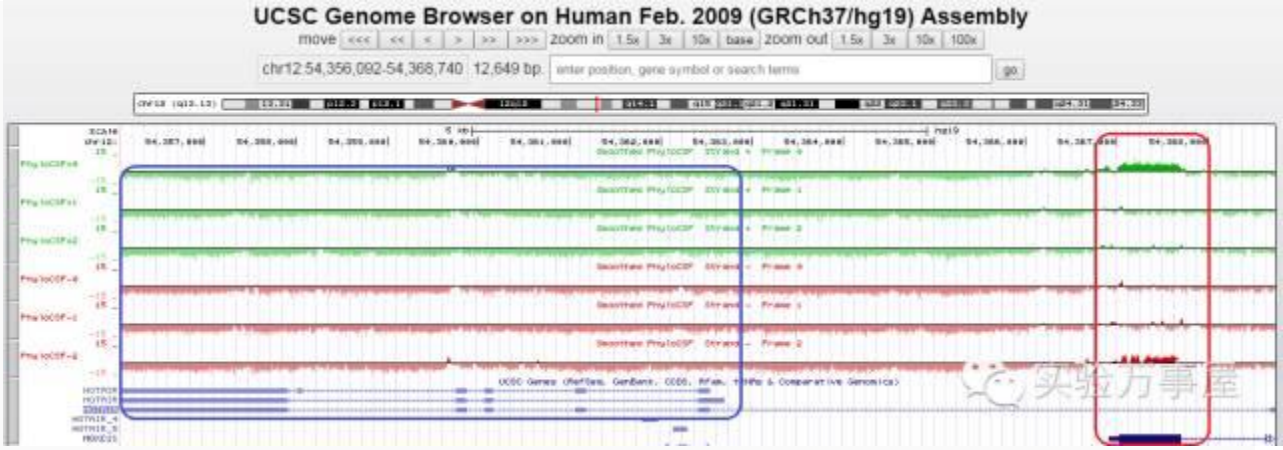
接着点击确认（上面看不清就看下面）：



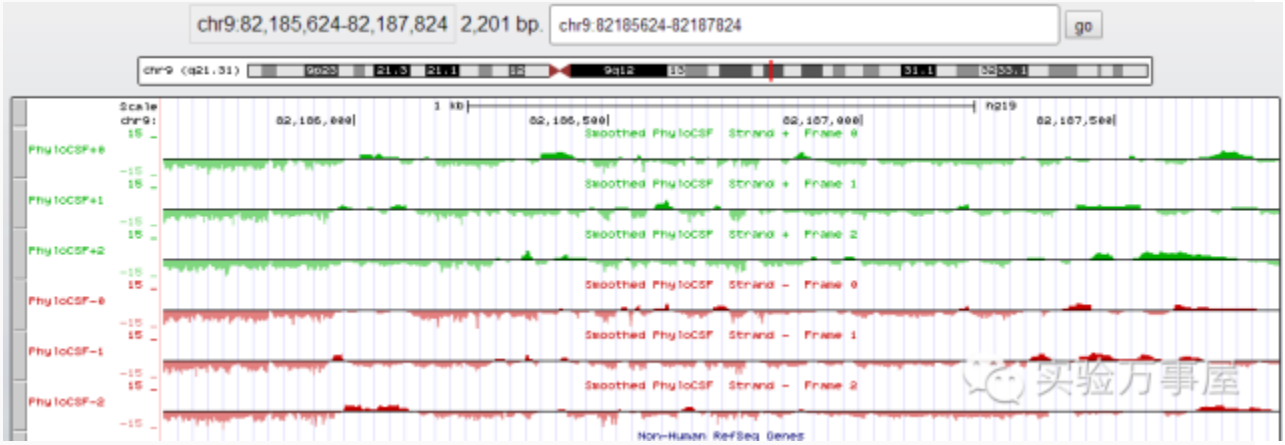
然后会弹出UCSC的封面，输入HOTAIR后进入：



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