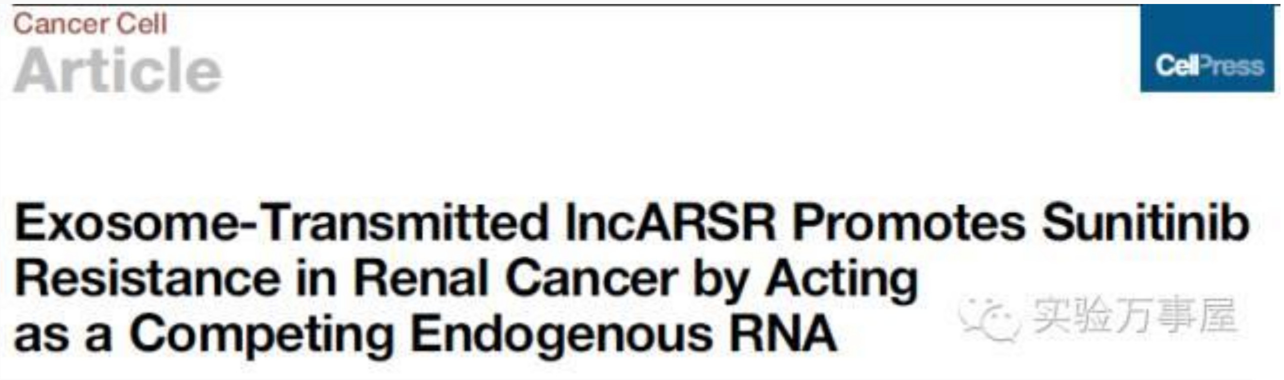
怎么证明LncRNA是LncRNA

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：

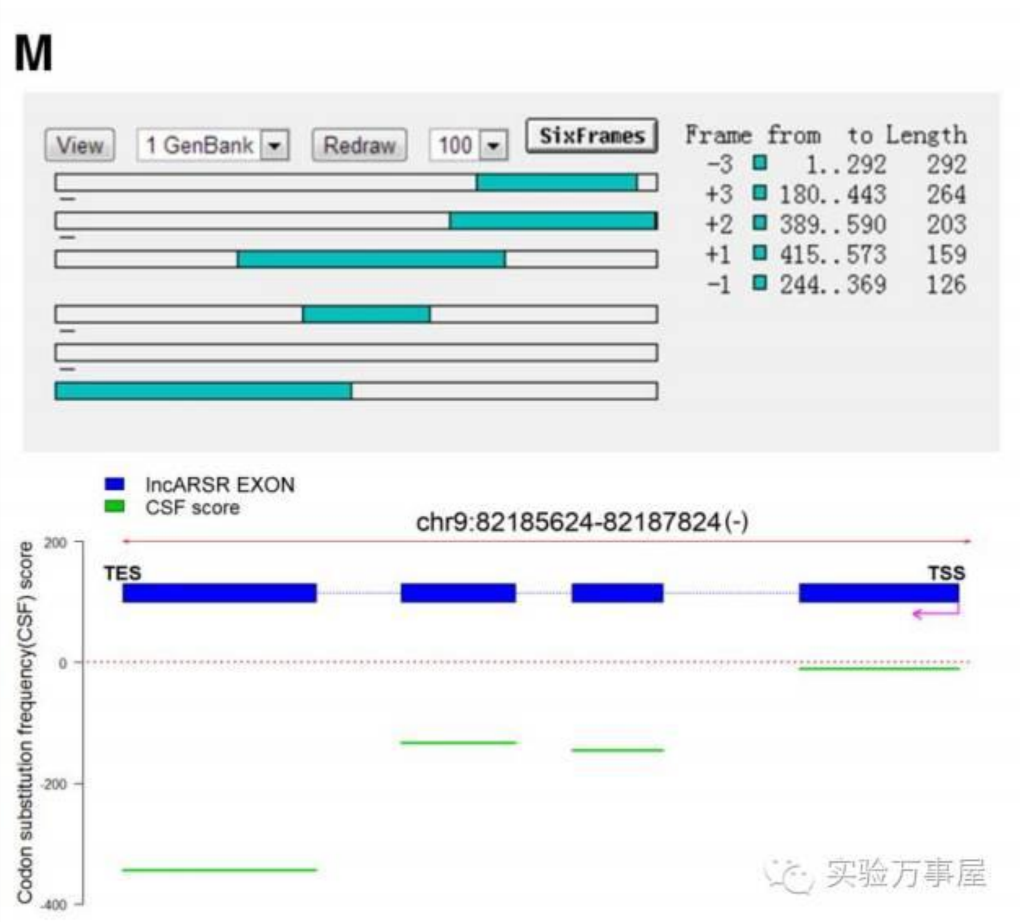


Cancer Cell
Article

CellPress

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

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。

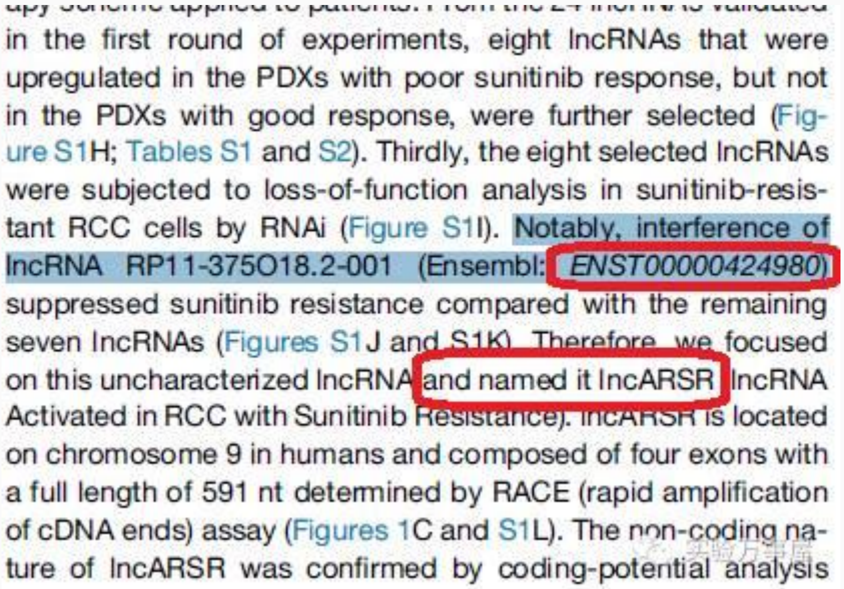


Frame		from to	Length
-3	■	1..292	292
+3	■	180..443	264
+2	■	389..590	203
+1	■	415..573	159
-1	■	244..369	126

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.

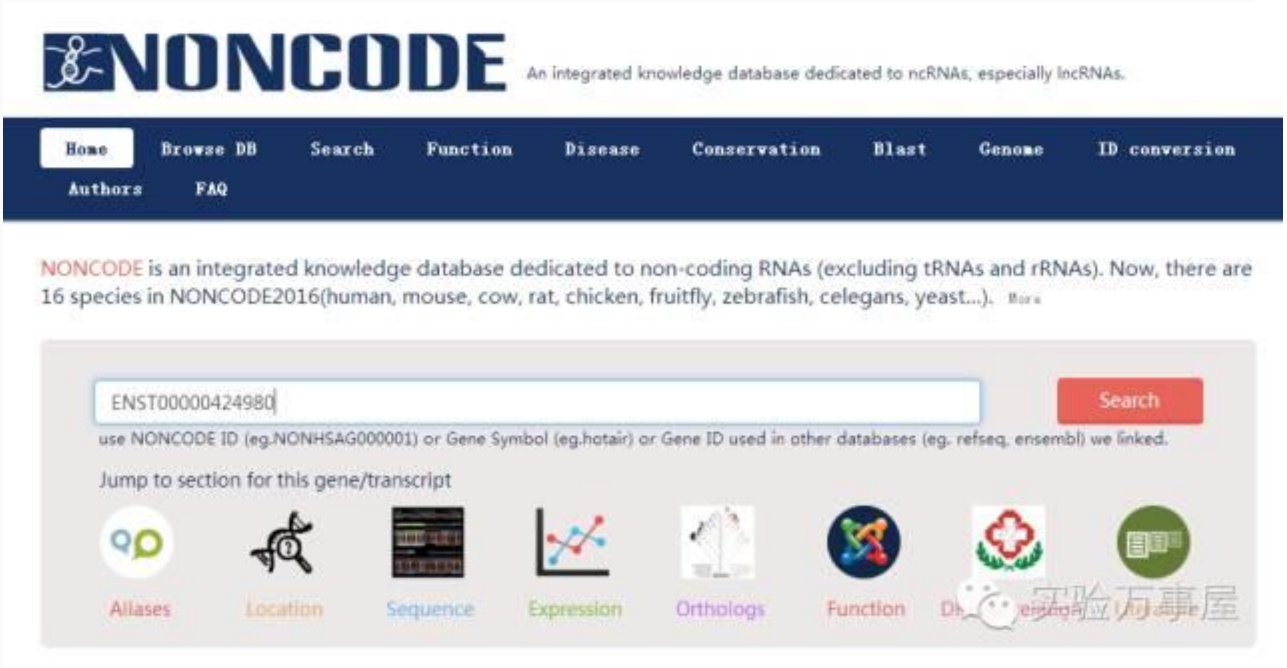
首先我明白一件事，就是要先分析这个lncRNA的ORF，也就是开放式阅读框。但是接下去要做什么呢？CSF又是啥？师姐，我要怎么办？？？

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。



apy scheme applied to patients. From the 24 lncRNAs 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：



NONCODE is an integrated knowledge database dedicated to non-coding RNAs (excluding tRNAs and rRNAs). Now, there are 16 species in NONCODE2016(human, mouse, cow, rat, chicken, fruitfly, zebrafish, celegans, yeast...). More

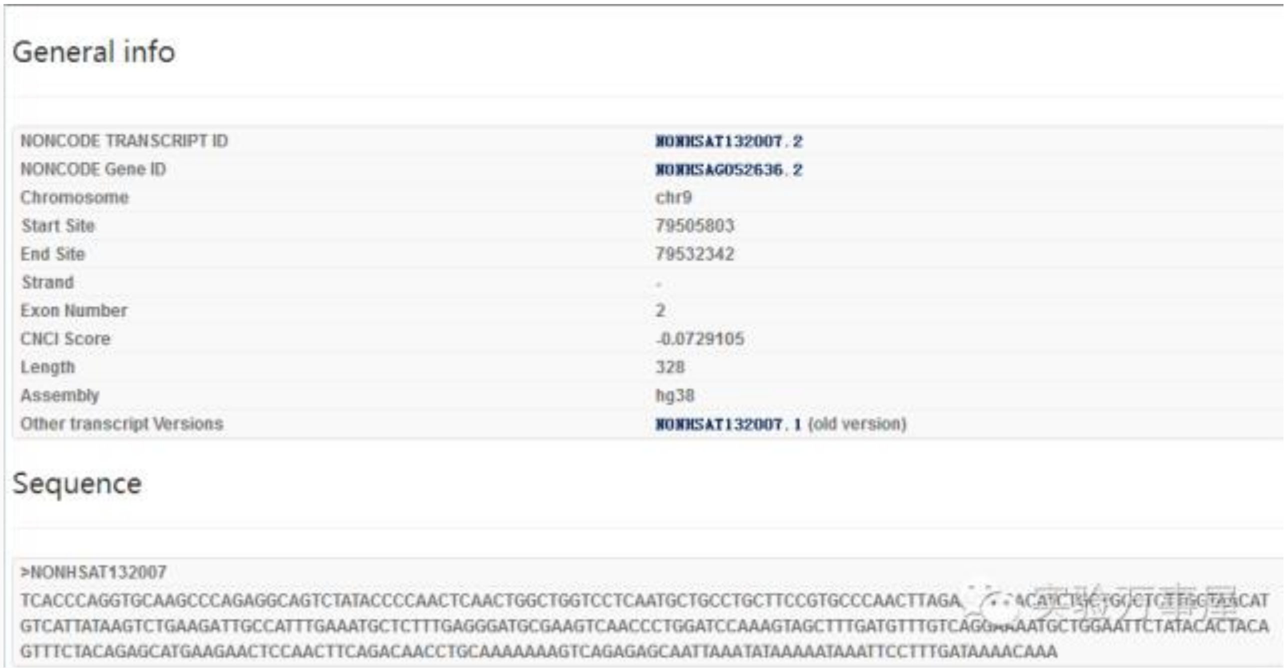
ENST00000424980

Search

use NONCODE ID (eg.NONHSAG000001) or Gene Symbol (eg.hotair) or Gene ID used in other databases (eg. refseq, ensembl) we linked.

Jump to section for this gene/transcript

得到这个序列：



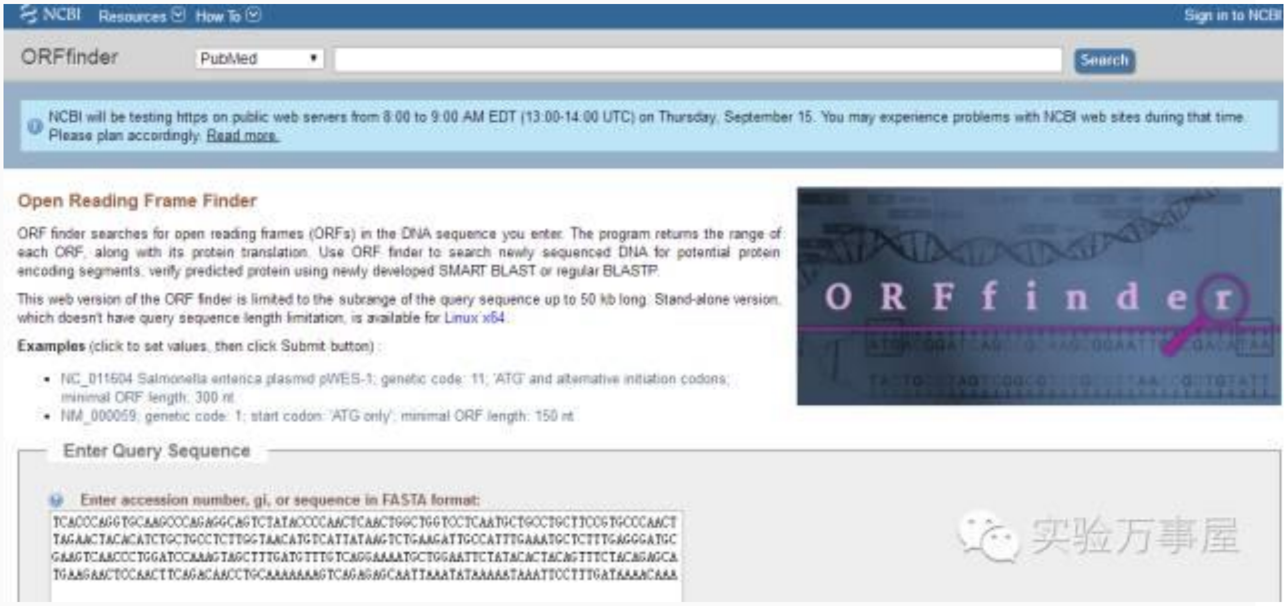
General info

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
TCACCCAGGTGCAAGCCCAGAGGCAGTCTATACCCCAACTCAACTGGCTGGTCCTCAATGCTGCCTGCTTCCGTGCCCAACTTAGAACTACACATCTGCTGCCTCTTGGTAACAT
GTCATTATAAGTCTGAAGATTGCCATTTGAAATGCTCTTTGAGGGATGCGAAGTCAACCCTGGATCCAAAGTAGCTTTGATGTTTGTCAGGAAAATGCTGGAATTCTATACACTACA
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。



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) :

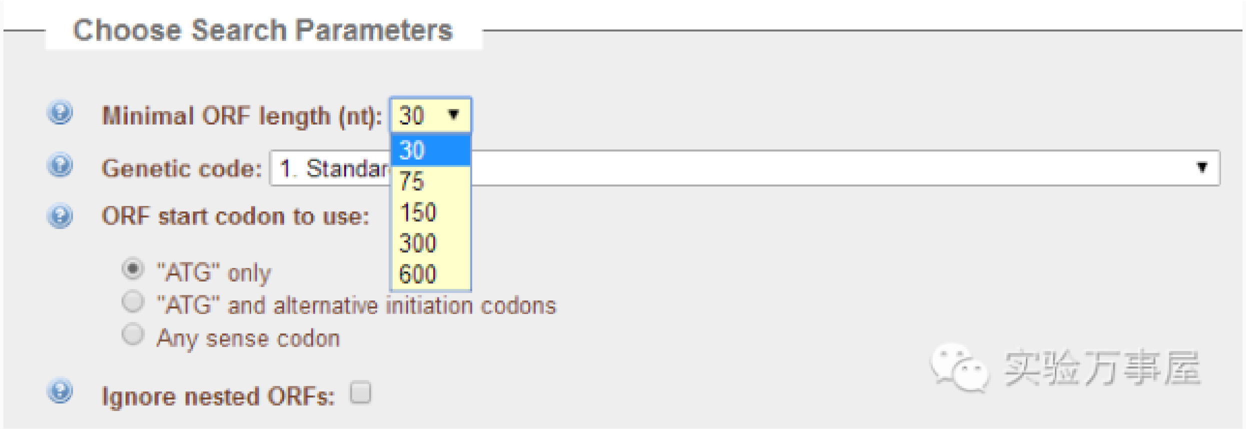
- NC_011604 Salmonella enterica plasmid pWES-1; genetic code: 11; 'ATG' and alternative initiation codons; minimal ORF length: 300 nt
- NM_000059: genetic code: 1; start codon: 'ATG only'; minimal ORF length: 150 nt

Enter Query Sequence

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

```
TCACCCAGGTGCAAGCCCAGAGGCAGTCTATACCCCAACTCAACTGGCTGGTCCTCAATGCTGCCTGCTTCCGTGCCCAACT
TAGAACTACACATCTGCTGCCTCTTGGTAACATGTCATTATAAGTCTGAAGATTGCCATTTGAAATGCTCTTTGAGGGATGC
GAAGTCAACCCTGGATCCAAAGTAGCTTTGATGTTTGTCAGGAAAATGCTGGAATTCTATACACTACAGTTTCTACAGAGCA
TGAAGAACTCCAACTTCAGACAACCTGCAAAAAAAGTCAGAGAGCAATTAAATATAAAAATAAATTCCTTTGATAAAACAAA
```

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 ▼

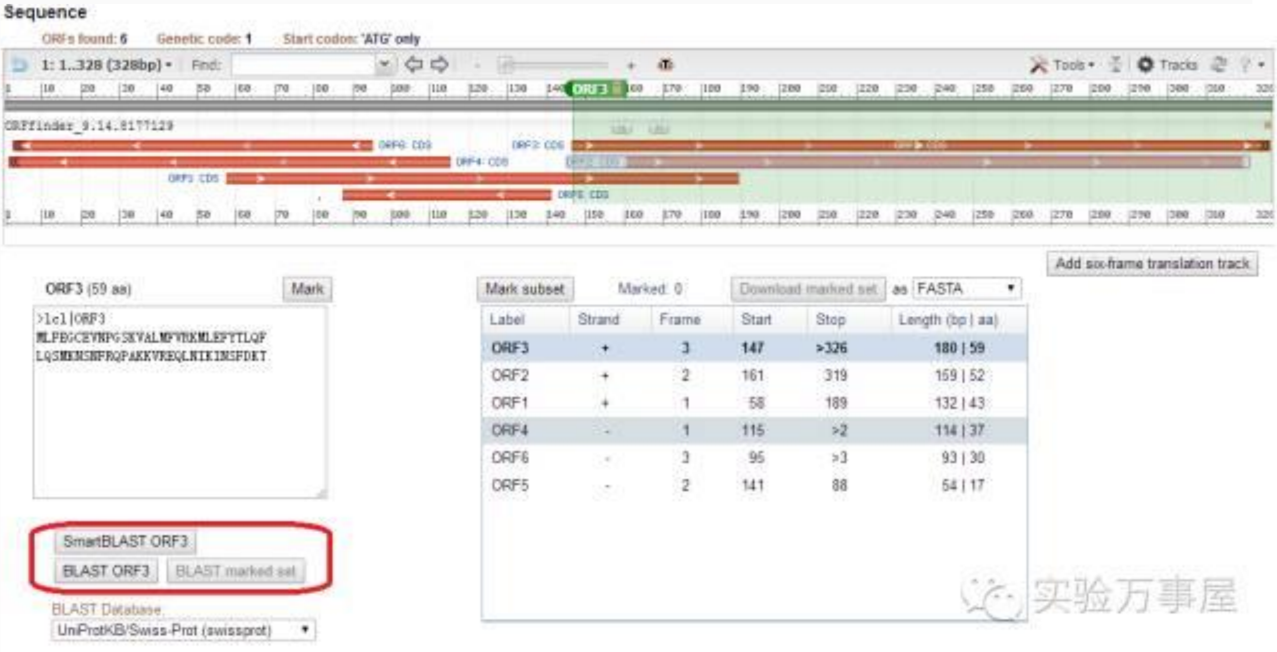
Genetic code: 1. Standar ▼

ORF start codon to use:

● "ATG" only
○ "ATG" and alternative initiation codons
○ Any sense codon

Ignore nested ORFs: □

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

ORF3 (59 aa)

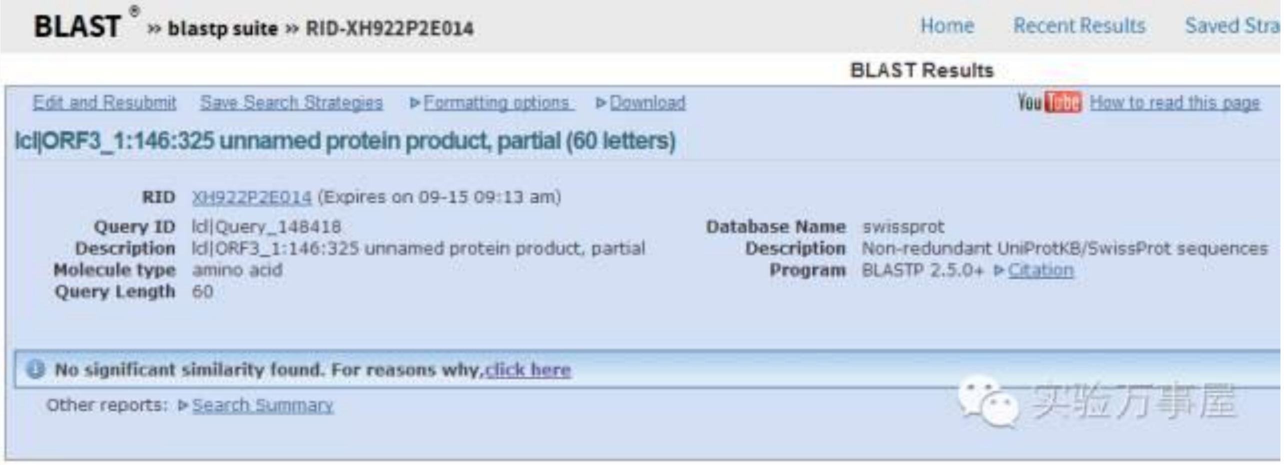
```
>lcl|ORF3
MLPEGCEVNPGSKVALMPVRKMLEFYTLQF
LQSMKNSNFRQPAKKVREQLNIKINSFDKT
```

Marked: 0 as FASTA

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

BLAST Database: UniProtKB/Swiss-Prot (swissprot)

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

BLAST Results

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

RID XH922P2E014 (Expires on 09-15 09:13 am)

Query ID lcl|Query_148418
Description lcl|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

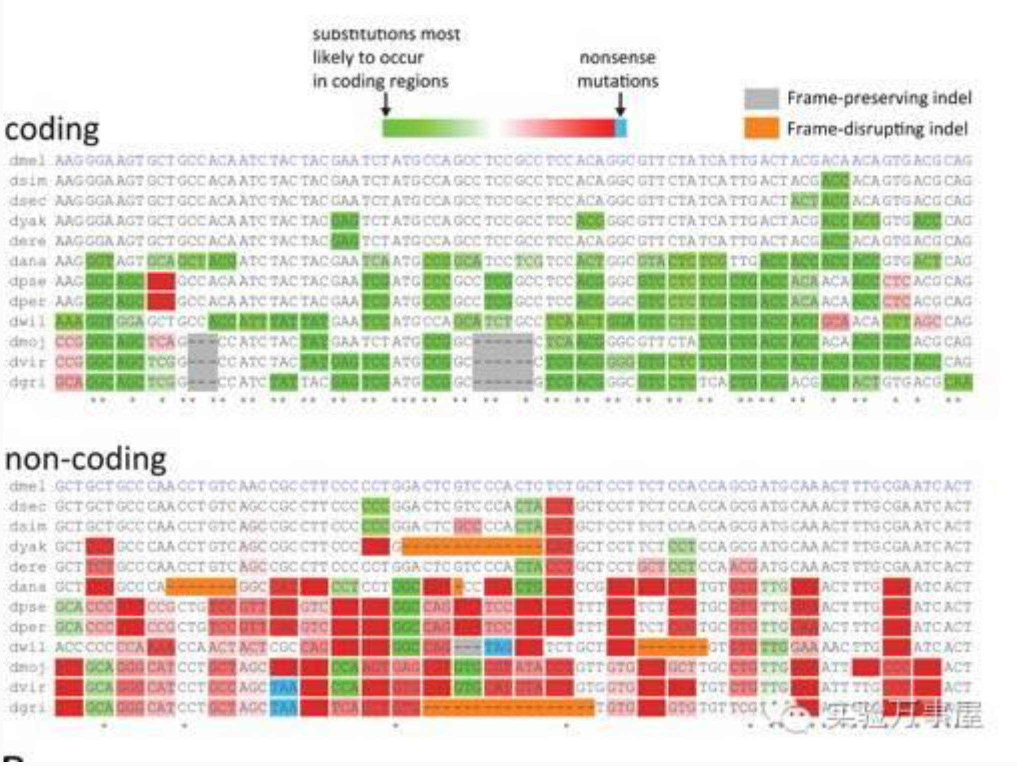
接着我们来看CSF，CSF到底是啥？CSF其实就是密码子的突变率。理论上编码区的密码子相对来说是保守的，也就是在物种中或者物种间是不容易产生突变，而非编码的就有点乱来了。我找到了这篇文献：

12 *Drosophila* Genomes/Letter

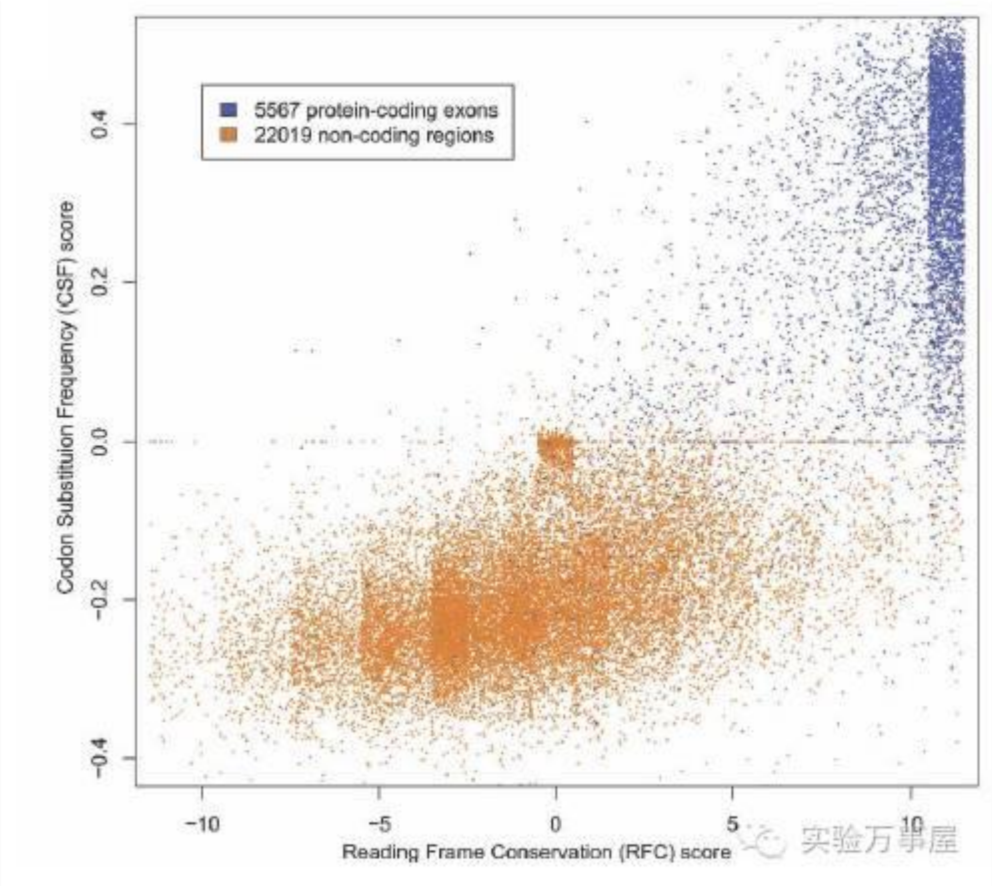
Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using 12 fly genomes

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

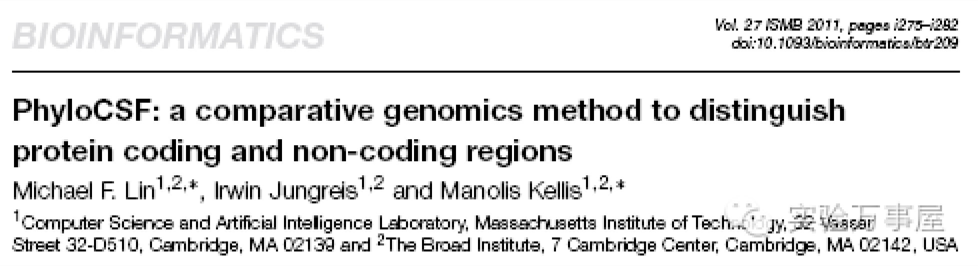
这是一篇在果蝇中用CSF来验证非编码与编码RNA间CSF差异的文献。其中显示，非编码的RNA突变率更高。



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

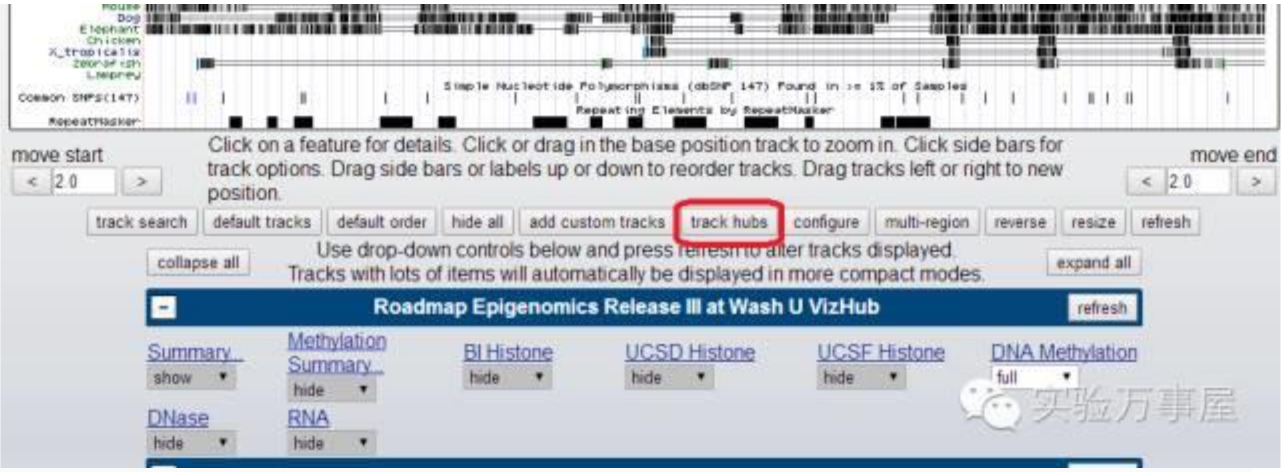
Vol. 27 ISMB 2011, pages i275–i282
doi:10.1093/bioinformatics/btr209

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

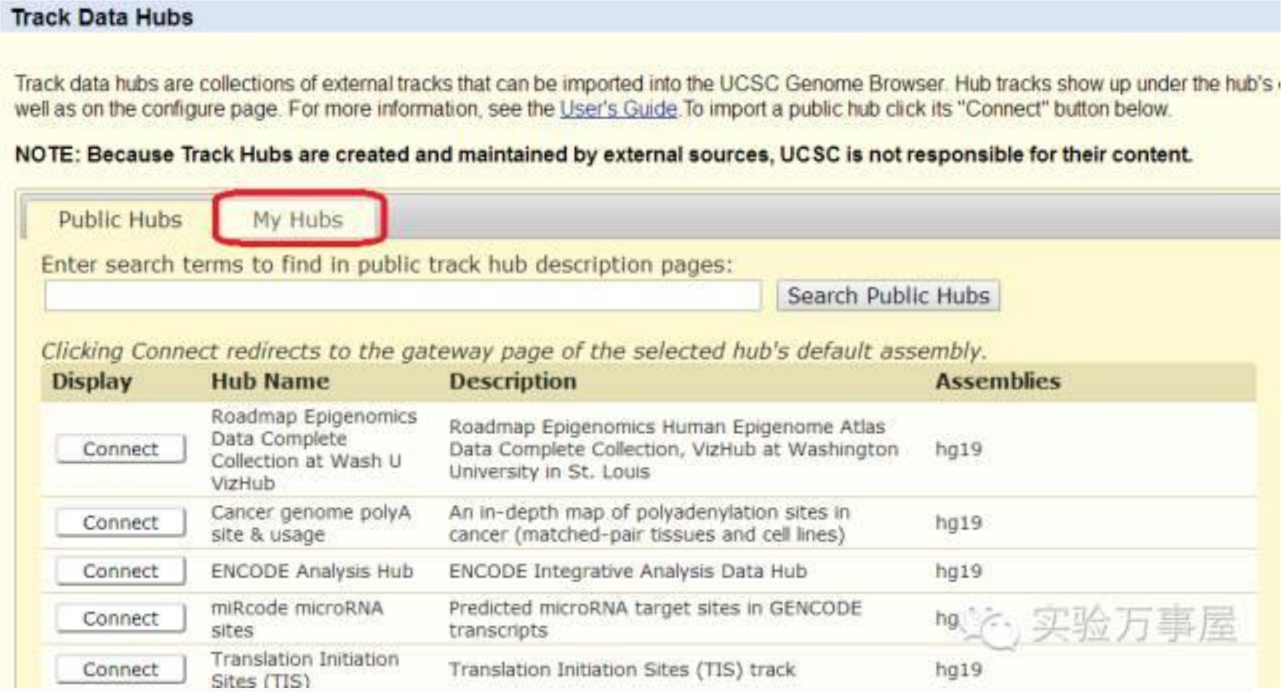
那问题来了，我们要怎么分析序列的PhyloCSF值呢？首先，要登录到强大到不要不要的UCSC上，随便进一个序列，我选了一个LncRNA——HOTAIR。然后点击“Track hubs”按钮。



Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

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
well as on the configure page. For more information, see the User's Guide. To import a public hub click its "Connect" button below.

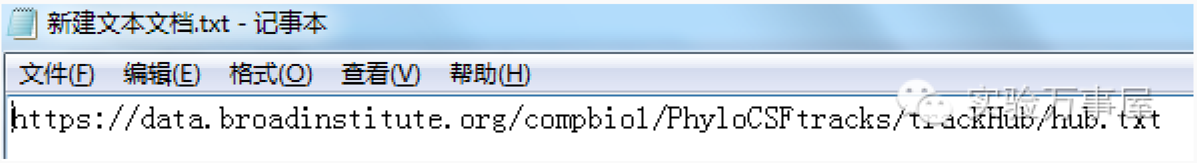
NOTE: Because Track Hubs are created and maintained by external sources, UCSC is not responsible for their content.

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.

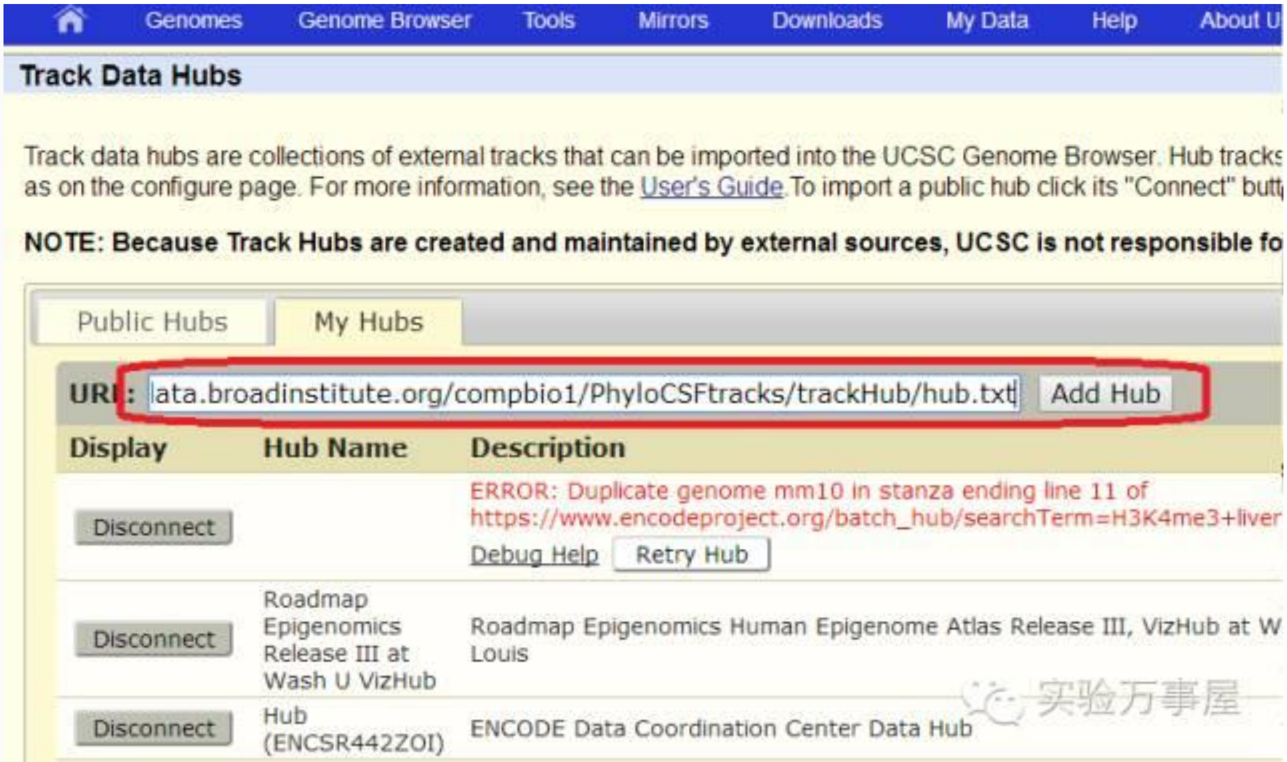
Display	Hub Name	Description	Assemblies
Connect	Roadmap Epigenomics Data Complete Collection at Wash U VizHub	Roadmap Epigenomics Human Epigenome Atlas Data Complete Collection, VizHub at Washington University in St. Louis	hg19
Connect	Cancer genome polyA site & usage	An in-depth map of polyadenylation sites in cancer (matched-pair tissues and cell lines)	hg19
Connect	ENCODE Analysis Hub	ENCODE Integrative Analysis Data Hub	hg19
Connect	miRcode microRNA sites	Predicted microRNA target sites in GENCODE transcripts	hg
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：


新建文本文档.txt - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

```
https://data.broadinstitute.org/compbio1/PhyloCSFtracks/trackHub/hub.txt
```

接着点击确认（上面看不清就看下面）：



Track Data Hubs

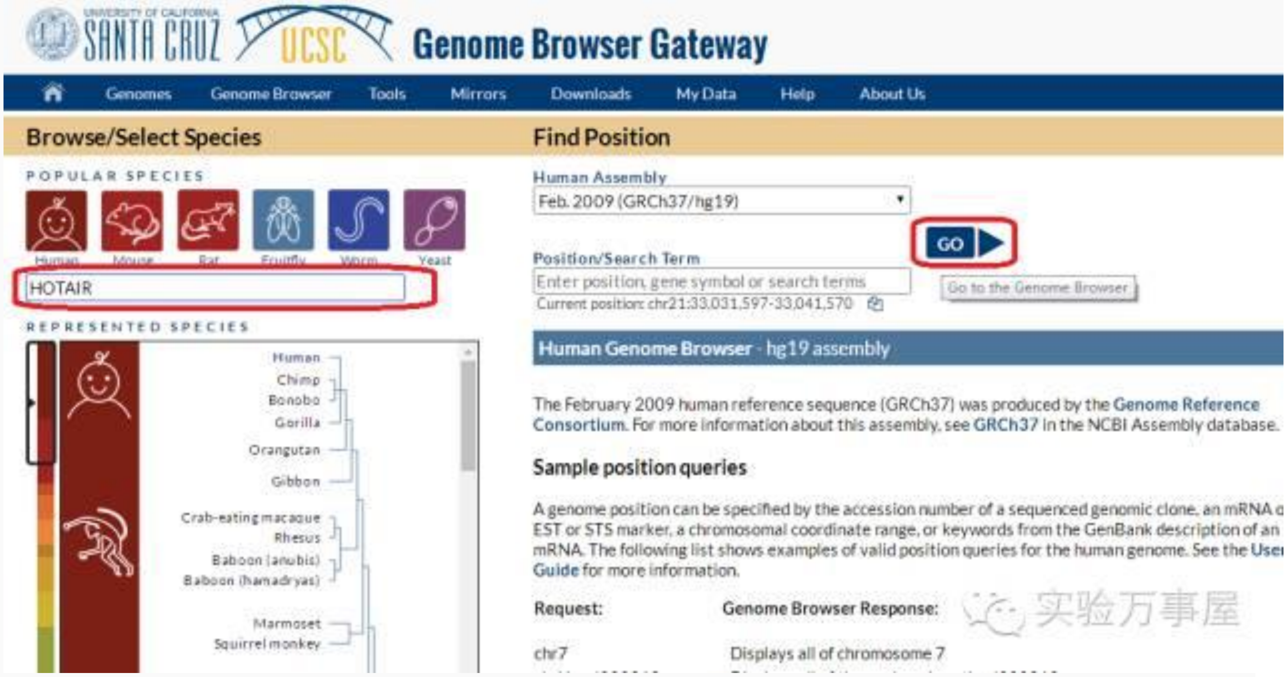
Track data hubs are collections of external tracks that can be imported into the UCSC Genome Browser. Hub tracks as on the configure page. For more information, see the User's Guide. To import a public hub click its "Connect" butt

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

URL: lata.broadinstitute.org/compbio1/PhyloCSFtracks/trackHub/hub.txt

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

然后会弹出UCSC的封面，输入HOTAIR后进入：



Browse/Select Species

HOTAIR

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

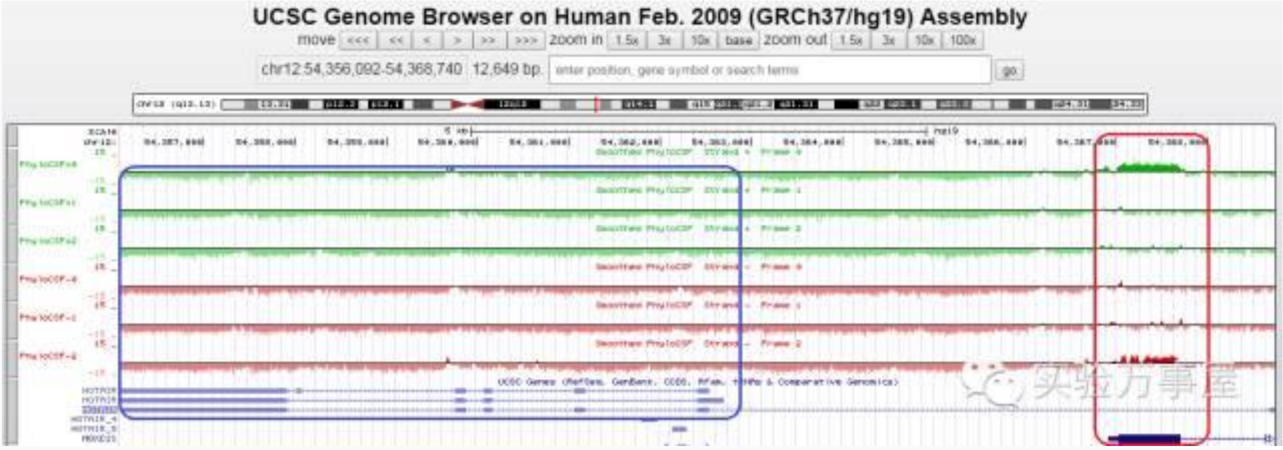
The February 2009 human reference sequence (GRCh37) was produced by the Genome Reference Consortium. For more information about this assembly, see GRCh37 in the NCBI Assembly database.

Sample position queries

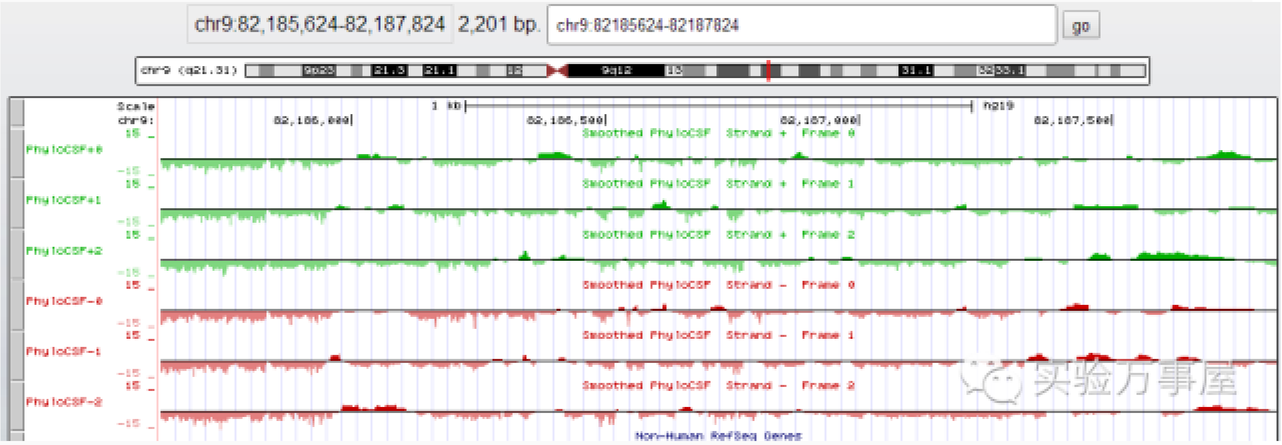
A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA o EST or STS marker, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the User Guide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

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

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。


UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
chr12:54,356,092-54,368,740 12,649 bp.
enter position, gene symbol or search terms
go
Scale
chr12:
5 kb
hg19
Smoothed PhyloCSF Strand + Frame 0
Smoothed PhyloCSF Strand + Frame 1
Smoothed PhyloCSF Strand + Frame 2
Smoothed PhyloCSF Strand - Frame 0
Smoothed PhyloCSF Strand - Frame 1
Smoothed PhyloCSF Strand - Frame 2
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
HOTAIR

那我们把那篇Cell中的lncRNA的序列位置输入进去，然后……


chr9:82,185,624-82,187,824 2,201 bp.
chr9:82185624-82187824
go
chr9 (q21.31)
Scale
chr9:
1 kb
hg19
82,186,000|
82,186,500|
82,187,000|
82,187,500|
Smoothed PhyloCSF Strand + Frame 0
Smoothed PhyloCSF Strand + Frame 1
Smoothed PhyloCSF Strand + Frame 2
Smoothed PhyloCSF Strand - Frame 0
Smoothed PhyloCSF Strand - Frame 1
Smoothed PhyloCSF Strand - Frame 2
PhyloCSF+0
PhyloCSF+1
PhyloCSF+2
PhyloCSF-0
PhyloCSF-1
PhyloCSF-2
15
-15
Non-Human RefSeq Genes

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