UCSC 数据库的使用

> 姓名：李群
>
> 老师：吴飞珍
>
> 学号：20111510021
>
> 邮箱：20111510021@fudan.edu.cn

> 时间：2020年11月1日

1. ucsc截图（以TCF7L2为例）

方法步骤：

```
1. 打开UCSC（https://genome.ucsc.edu）网站，选择Genomes--> Human GRCh38/hg38；
2. hide all；
3. 在搜素框搜索TCF7L2，选择 Homo sapiens transcription factor 7 like 2 (TCF7L2),
transcript variant 2, mRNA. (from
   RefSeq NM_030756)；
4. 在configure页面只保留 Display labels to the left of items in tracks 和 Show
track controls under main graphic,
   点击submit；
5. 在Genes and Gene Predictions中的 GENCODE v32只保留gene symbol一项，选择full show
的方式
6. 在Regulation中的ENCODE Regulation Track 中选择 Layered H3K4Me1、Layered
H3K4Me3和Layered H3K27Ac，选择full show
7. 调整浏览器大小，resize之后通过view导出pdf即可
```

结果

TCF7L2 : Homo sapiens transcription factor 7 like 2 (TCF7L2), transcript variant 2, mRNA. (from RefSeq NM_030756)

2. 从UCSC Refseq下载注释数据

方法步骤：

1. 从UCSC的tools-->table Browser进行下载
2. 具体参数设置

3. 点击get output获取ref注释文件，解压之后得到**hg38_refseq.all.txt**
4. 查看refseq大小

```
[st28@ibs report2]$ awk '{print $2}' hg38_refseq.all.txt | uniq | wc
166924  166924 2343093
```

5. 查看gene

```
#去重前
[st28@ibs report2]$ awk '{print $13}' hg38_refseq.all.txt | wc
166924  166924 1250315
#去重后，包括预测
[st28@ibs report2]$ awk '{print $13}' hg38_refseq.all.txt | uniq | wc
46707   46707  407607
# 去重且去除 “LOC*”
[st28@ibs report2]$ awk '{print $13}' hg38_refseq.all.txt | uniq | grep -v
"LOC*" | wc
33967   33967  243637
```

结果

refseq和基因数量统计

类别	去重前	去重后	去重+去除LOC*
refseq	166924	\	\
gene	166924	46707	33967

3. mm10基因组chr10和chr11

方法步骤：

1. 在UCSC中下载mm10的10号和11号染色体信息

```
[st28@ibs report2]$ wget
http://hgdownload.soe.ucsc.edu/goldenPath/mm10/chromosomes/chr10.fa.gz
[st28@ibs report2]$ wget
http://hgdownload.soe.ucsc.edu/goldenPath/mm10/chromosomes/chr11.fa.gz
[st28@ibs report2]$ gunzip chr10.fa.gz
[st28@ibs report2]$ gunzip chr11.fa.gz
```

2. 合并文件

```
[st28@ibs report2]$ cat chr10.fa chr11.fa> mm10_chr10_chr11.fa
```

3. 查看大小

```
[st28@ibs report2]$ cat mm10_chr10_chr11.fa | grep -v chr | wc
5055551 5055551 257833087
[st28@ibs report2]$ ll -h
total 492M
-rw-r--r-- 1 st28 student 128M Feb 10  2012 chr10.fa
-rw-r--r-- 1 st28 student 119M Feb 10  2012 chr11.fa
-rw-r--r-- 1 st28 student 246M Nov  1 15:04 mm10_chr10_chr11.fa
```

结果

文件大小	碱基数量
246M	257833087

4. 致谢

感谢吴老师、顾老师及其课题组成员的讲解与帮助

感谢复旦大学生物医学研究院，生物信息平台提供的服务器