1. 软件工具下载

- csvtk及taxonkit的下载(conda下载)

2. nt数据库及分类库的下载

- aspera高速下载nt库

```
##下载nt数据库fasta序列及md5文件                    #大文件146Gw
nohup ascp  -k 1 -QT -l 500m -T -i ~/miniconda3/etc/asperaweb_id_dsa.openssh --host=ftp.
wget https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nt.gz.md5              #验证，小文件
md5sum  nt.gz   #验证
md5sum  -c nt.gz    #检查文件完整性

##下载taxdump.tar.gz
wget https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/taxdump.tar.gz              #小文件
wget https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/taxdump.tar.gz.md5        #验证，小文件

##下载有NCBI的accession与taxid的对应关系文件nucl_gb.accession2taxid.gz
wget https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/accession2taxid/dead_nucl.accession2taxid
wget https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/accession2taxid/dead_nucl.accession2taxid

```

- Aspera的常用参数

```
-T    ##不进行加密。若不添加此参数，可能会下载不了。
-k    ##断点续传，网络突然断掉下次可以续传，不用重新从头下载。
-i    ##输入私钥，安装 aspera 后有在目录 ~/.aspera/connect/etc/ 下有几个私钥，使用 linux 服务
-l  string   ##设置最大传输速度，比如设置为 200M 则表示最大传输速度为 200m/s。若不设置该参数，
```

2、NR数据库建立（可省略）

```
# 核酸列数据库初始化;注意蛋白库用prot，核酸库用nucl
makeblastdb -in  nr.fa -dbtype prot -out NR -parse_seqids
makeblastdb -in nt.gz -dbtype nucl -title nt  -parse_seqids -out database/ -logfile nt.l
```

3. Taxid或Accession提取

- Taxonkit用法

```
### taxonkit参数说明
 -j ： 线程数；
 --ids：需要提取的分类的taxid；
 --data-dir：该目录下必须包含文件names.dmp和nodes.dmp；
 --indent：提取的物种编号缩进位置，这个参数很重要，记得一定要设置为空 ""
-P/--add-prefix：给每个分类学水平添加前缀，比如s__species。
-t/--show-lineage-taxids：输出分类学单元对应的TaxID。
-r/--miss-rank-repl: 替代没有对应rank的taxon名称
-S/--pseudo-strain: 对于低于species且rank既不是subspecies也不是stain的taxid，使用水平最低tax

##各物种的taxid 可参考此帖子 https://www.jianshu.com/p/5a72f42e0412，以植物33090为例
taxonkit -j 5 list --ids 33090 --indent "" > plant.taxid.txt     # -j 线程数

##根据taxid获取它的完整世系信息
echo 59689 | taxonkit lineage
##使用reformat对输出进行格式化，输出其中某一级或多级信息
echo 59689 | taxonkit lineage | taxonkit reformat --format "{k}" | cut -f 1,3
##--format,默认是"{k};{p};{c};{o};{f};{g};{s}"）,分别对应k(superkingdom, 超界), p(phylum,
,c(class, 纲),o(order, 目),f(family, 科),g(genus, 属),s(species, 种), 另外还有一个S(subspe

##查找指定taxids列表的物种信息，并写入文件
taxonkit lineage taxids.txt > lineage.txt
taxonkit -j 20 lineage taxids.txt |taxonkit reformat -f "{k};{p};{c};{o};{f};{g};{s}" -F

taxonkit lineage taxids.txt |taxonkit reformat -f "{k}\t{p}\t{c}\t{o}\t{f}\t{g}\t{s}" -F
| csvtk add-header -t -n taxid,kindom,phylum,class,order,family,genus,species  >lineage.
```

```
#提取plant.taxid.txt对应的所有核酸序列的accession
zcat nucl_gb.accession2taxid.gz |csvtk -t grep -f taxid -P plant.taxid.txt |csvtk -t cut

#提取accession对应的所有核酸序列的taxid
zcat nucl_gb.accession2taxid.gz |csvtk -t grep -f accession.version -P 3xia.genus.nt.acc
sed '1d' 3xia.genus.nt.acc.taxid >3xia.genus.nt.acc.taxid2     #去掉表头“taxid”
sort  3xia.genus.nt.acc.taxid2 | uniq > 3xia.genus.nt.acc.taxid3    #去重
```

- 按acc号提取整理好的acc.taxid.tax文件中的taxonomy

```
list=zooplankton.genus.nt.acc file=3.C.m.z.acc.taxid.tax out=zooplankton.genus.nt.acc.tax
```

4. 从NR全库里面提取子库（可省略）

```
# 从NR全库中提取子库
blastdb_aliastool -gilist plant.taxid.acc.txt -db NR -out NR_plant -title NR_plant
nohup blastdb_aliastool -gilist Eukaryota.taxid.acc.txt -db database/nt -out database/euk
```

5、从nt库fasta里提取子库fasta序列

```
##已得到序列accession id，可使用seqtk提取，参数看help或者我后续补充
seqtk subseq $input_fa $target_accession > output_fa
```

6. blastdbcmd得到各类的fasta文件

```
#blastdbcmd -db database -entry_batch fasta_title.file > newfasta.fasta        #批量提取fa
blastdbcmd -db ncbi/Nr/nr_Virus -entry all -dbtype prot -out nr_Virus.fa       #提取所有fa
blastdbcmd -db refseq_rna -entry 224071016 -out test.fa        ##从数据库中提取除gi号为2240
```

- blastdbcmd

```
作用: Retrieves sequences or other information from a BLAST database
它相当于以前的fastacmd.利用这个命令,可以从blast数据库中获得你想要的信息:
blastdbcmd -db refseq_rna -info    ##可以查看数据库refseq_rna的信息

注:gi ID是许多用来标志序列的标识符中的一种.是数据库文件中普遍使用,通行有效的保持索引的形式。
所有来源于NCBI的序列都有一个gi号“gi|gi_identifier”.是绝对唯一的，而自己利用makeblastdb命令 构
gnl|database|identifier
lcl|identifier
identifier
这些标识符的作用是区别于gi号在本数据库中，使得序列标识符唯一在查询和比对中分辨query序列与subject
```

7. 一步代码

```
taxonkit list -j 2 --ids 10239 --indent "" --data-dir ./taxdump/ > Virus.list
cat prot.accession2taxid | csvtk -t grep -f taxid -P ../nr/Virus.list | csvtk -t cut -f a
blastdb_aliastool -seqidlist Virus.taxid.acc.txt -db /nr -out nr_virues -title nr_virues
blastdbcmd -db /nr/accession2taxid/nr_virues -entry all -dbtype prot -out nr_Virus.fa
diamond makedb --in nr_Virus.fa --db nr_Virus -p 10
```

8. 下载库文件 update_blastdb.pl

- 安装nt/nr库需要先进行环境变量配置，在家目录下新建一个.ncbirc配置文件，然后添加如下内容

```
; 开始配置BLAST
[BLAST]
; 声明BLAST数据库安装位置
BLASTDB=/home/xzg/Database/blast
; Specifies the data sources to use for automatic resolution
; for sequence identifiers
DATA_LOADERS=blastdb
; 蛋白序列数据库存本地位置
BLASTDB_PROT_DATA_LOADER=/home/xzg/Database/blast/nr
; 核酸数据库本地存放位置
BLASTDB_NUCL_DATA_LOADER=/home/xzg/Database/blast/nt
[WINDOW_MASKER]
WINDOW_MASKER_PATH=/home/xzg/Database/blast/windowmasker
```

- 配置好之后，使用BLAST+自带的update_blastdb.pl脚本下载nt等库文件

```
（不建议下载序列文件，一是因为后者文件更大，二是因为可以从库文件中提取序列blastdbcmd，最主要是建
提醒：下载文件较大，耗费时间较长，最好将任务转入后台。简单的做法，也可用nohup命令（下面nohup后面
nohup time update_blastdb.pl nt nr > log &

监控库文件是否下载完成，如何判断？
1. 查看log文件是否有提示；
2. 查看update_blastdb.pl是否还在运行：
```

```
ps -aef | grep update_blastdb.pl | grep -v update_blastdb.pl     ##如过没有结果，则说明没有

下载完成后解压所有tar.gz文件（用通配符）即可：
nohup time tar -zxvf *.tar.gz > log2 &

如果你不想通过update_blastdb.pl下载nr和nt等库文件，也可以是从ncbi上直接下载一系列nt/nr.xx.tar
然后解压缩即可，后续还可以用update_blastdb.pl进行数据更新。

下载过程中请确保网络状态良好，否则会出现Downloading nt.00.tar.gz...Unable to close datastrea
```

- **报错**

```
使用update_blastdb.pl更新和下载数据库时候出现模块未安装的问题。
解决方法，首先用conda安装对应的模块，然后修改update_blastdb.pl的第一行，即shebang部分，以conda
perl `which update_blastdb.pl`
```