先购买云服务器，然后挂载新买的硬盘

Fdisk -l 发现 fdisk: cannot open /dev/vda: Permission denied

解决：sudo -i 就行

现在fdisk -l 就有结果了

Fdisk -l 可以找到硬盘 /dev/vdb

先创建挂载的地方 mkdir /mnt/usb

挂载 mount /dev/vdb /mnt/usb

出错： mount: /mnt/usb: wrong fs type, bad option, bad superblock on /dev/vdb, missing codepage or helper program, or other error.

查阅相关资料发现这个是由于没有格式化的原因

解决 ： sudo mkfs -t ext4 /dev/vdb

就可以了

下面安装miniconda

wget -c https://mirrors.tuna.tsinghua.edu.cn/anaconda/miniconda/Miniconda3-latest-Linux-x86\_64.sh

然后 bash Miniconda3-latest-Linux-x86\_64.sh 最后修改环境变量

source ~/.bashrc 就可以了

然后我的目标是先搞一个nr注释吧。所以需要下载blast.blast下载很简单 可以直接

Sudo apt-get install (因为我虚拟机是用centos6.9的，所以ubuntu有点不习惯)。我用的是在网上找ncbi最新版，然后下载压缩包，tar -zxvf 就行了

然后设置环境变量

下面下载nr库，这个库挺大的，我下载了aspera ，解压之后 然后 bash aspera.sh既可以了。大约会生成一个隐藏文件，命令： sudo /home/ubuntu/.aspera/connect/bin/ascp -v -k 1 -T -l 200m -i ~/.aspera/connect/etc/asperaweb\_id\_dsa.openssh era-fasp@fasp.sra.ebi.ac.uk:/vol1/fastq/SRR768/SRR768317/SRR768317\_1.fastq.gz ./ 就可以了。

如果下载不了，在-l 200后面添加参数-P33001

ascp -QT -L /home1/jialh/SRR385732/logs -l 200M -P33001 -i /home1/jialh/.aspera/config/asperaweb\_id\_dsa.openssh [era-fasp@fasp.sra.ebi.ac.uk:/vol1/fastq/SRR385/SRR385732/SRR385732.fastq.gz](mailto:era-fasp@fasp.sra.ebi.ac.uk:/vol1/fastq/SRR385/SRR385732/SRR385732.fastq.gz)

为了使nr比对的结果更为精致，我使用nr分库的方法，先下载物种及分类信息的文件

<https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/taxdump.tar.gz> 然后下载accession与taxid的对应关系

<https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/accession2taxid/prot.accession2taxid.gz>

下载工具

taxonkit: <https://github.com/shenwei356/taxonkit/releases/download/v0.6.0/taxonkit_linux_amd64.tar.gz>

csvtk::

<https://github.com/shenwei356/csvtk/releases/download/v0.20.0/csvtk_linux_amd64.tar.gz>

Blast上面下了现在不下了。

步骤1：

/mnt/usb/taxonkit list -j 2 --ids 4751 --indent "" --data-dir /mnt/usb >fungi.list

步骤2

cat prot.accession2taxid | /mnt/usb/csvtk -t grep -f taxid -P fungi.list | /mnt/usb/csvtk -t cut -f accession.version > fungi.taxid.acc.txt

步骤3：

makeblastdb -in nr -dbtype prot -parse\_seqids

这步骤似乎要求很高，我32g内存都没用，所以看看有没有更好的办法吧。

这里其实可以用他直接建库好的文件，地址在ftp.ncbi.nlm.nih.gov:/blast/nr.XX.gz

直接解压就可以用了。

步骤4。

比对，这里我用的是自己编写的脚本 blast.pl

先提取子库

blastdb\_aliastool -seqidlist fungi.taxid.acc.txt -db /mnt/usb/NR/nr -out nr\_fungi -title fungi(这步不生成蛋白序列

然后)

perl blast.pl blastp /root/phidabe/db/phi 8all.prot.fasta 1e-5 6 uniprot\_sprot 5 （这个是以前做的swissprot注释）

parsing\_blast\_result.pl uniprot\_sprot.xml 20 1e-5 0.2 > uniprot\_sprot.xls（以前做的注释，看看就行了）

然后就能得到结果

根据实践，16个线程可以做一天可以做800个基因功能注释，消耗内存差不多是14-17g。所以如果内存有个几百个g的话那基本可以一天跑完的。

nr注释就到这里吧。

添加swap虚拟内存：

<https://blog.csdn.net/qq_41671415/article/details/88737255?locationNum=7&fps=1>

看这个吧，挺复杂的，抄作业 一样抄一遍就可以了。

nr数据库下载：

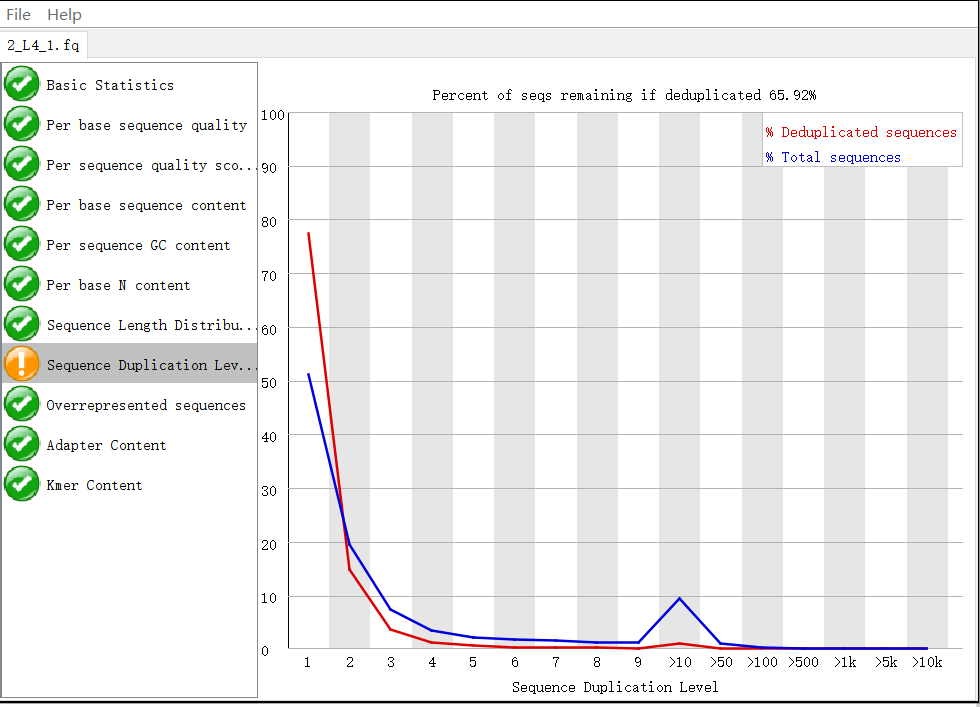
for i in {02..38};do /public/home/yyang/.aspera/connect/bin/ascp -v -k 1 -T -l 200m -i ~/.aspera/connect/etc/asperaweb\_id\_dsa.openssh anonftp@ftp.ncbi.nlm.nih.gov:/blast/db/nr.${i}.tar.gz ./;done 解压就可

二代拼接，我用的是软件SPAdes，经过试验，这个软件大概用的内存在20左右，安装的话conda就行，用发：

/home/data1/spades/bin/spades.py -1 /dat1/bgq/Genome/illumina/Rawdata/1.fastq -2 /dat1/bgq/Genome/illumina/Rawdata/2.fastq --only-assembler -o /dat1/bgq/Genome/illumina/Rawdata/assembly -t 10 -k 77

我结果因为是没有进行序列的修剪的，因为我觉得差不多了。和公司拼出来主要是条数的差异，其他基因组大小差别不大。几个小时就能运行完。

用busco可以查看拼接的好坏，run\_BUSCO.py -i /root/opt/Chr2.fasta -o result -l /root/opt/biosoft/busco/library/fungi\_odb10 --mode geno



其实就重复序列有点多。

三代拼接：

用的是软件falcon， 这个软件是公司摸索出来的比较适合真菌基因组。

用法是创造出111.cfg文件。用法是fc\_run fc\_run.cfg 具体参数看文件就行。

不过测序数据是bam格式，要做一些转换，https://zhuanlan.zhihu.com/p/45760843

# build index for convert

~/opt/biosoft/smrtlink/smrtcmds/bin/pbindex pb.bam &

# convert bam to fasta

~/opt/biosoft/smrtlink/smrtcmds/bin/bam2fasta -o pb pb.bam &

/mnt/usb/sandai/smrtlink/install/smrtlink-release\_9.0.0.92188/bundles/smrttools/install/smrttools-release\_9.0.0.92188/private/pacbio/bam2fastx/bin

用了其中一个bam文件进行了测试，但是效果不好，参数可能不太对。但是起码能做出来了，

假基因预测

<https://github.com/ShiuLab/PseudogenePipeline>

按照上面的说明书，可以下把python改为2.7版本。这个很简单的。

先输入 python 输出为3.8.0，然后在位置处删去，在创一个python，软连接到python2.7就行。

然后下载repeatmasker 。

<https://blog.csdn.net/u012110870/article/details/102804531> 这个文件是配置repeatmasker的说的很详细。

其实repeatmask安装相当麻烦，新手可能需要点时间，但是我以前在自己的虚拟机Linux上安装了一遍，所以其实问题不大，但是其中遇到过Can't locate Text/Soundex.pm in @INC这个问题

这个一看就是因为perl缺少Text/Soundex谋爱。安装就行了

sudo cpan Text::Soundex 但是因为网不好，所以一直安装失败，我查阅了很多办法，甚至考虑在自己云服务器上添加vpn，但是用不来。最后在https://www.jianshu.com/p/40b12429e28f

上面找到了属于自己的方法。apt install cpanm 然后cpanm可以指定网址，那就指定一个中国的就行了 sudo cpanm --mirror http://mirrors.163.com/cpan --mirror-only Text::Soundex

也可以使用conda的方法

conda install -c bioconda perl-file-homedir

然后下载tfasty，这个的话有的。

安装的话下载fasta-36.3.8h 根据里面的readme文件进行操作

cd src

make -f ../make/Makefile.linux\_sse2 all

然后按照PseudogenePipeline-master的readme文件进行操作

R包安装：conda install r-essentials --yes

python ../\_wrapper\_scripts/CombinedPseudoWrapper.py test\_parameter\_file.txt

其中test\_parameter\_file.txt文件格式如下：

b\_out=test\_tblastn\_tabular.out

p\_seq=test\_protein\_sequences.fa

g\_seq=test\_genome.fa

b\_filter=y

b\_data=tabular

p\_codes=../\_pipeline\_scripts

o\_codes=../\_pipeline\_scripts

f\_dir=/mnt/usb/pesudogene/fasta-36.3.8h/bin

f\_prog=tfasty36

blosum=../\_example\_files/blosum50.matrix

gff=test\_gff

repCut=300

repDiv=30

然后就可以了

出现问题repeatmask not found

很奇怪，我前面刚刚配置好怎么没有找到repeatmask呢？

这个是因为我傻逼。我开了2个root，一个配置了repeatmask的环境变量，而在另一个运行，那就没了，那个环境变量应该在同一个窗口中才有用。这个还需要查查。

配置好了在进行test，发现可以了，我们在用茯苓的第一条序列试试，发现报错，原来我们少了一个文件test\_tblastn\_tabular.out。然后我看了相关的文章，发现没有提这个test\_tblastn\_tabular.out

文件的出处，然后我就仔细分析这个文件。发现了其实它就是一个tblastn的比对结果，就是蛋白比对到核酸的结果。

然后

makeblastdb -in Chr1.fasta -dbtype nucl -title chr1 -parse\_seqids -out chr1 -logfile chr1.log

tblastn -query Chr1\_prot.fasta -db chr1 -evalue 1e-5 -num\_threads 6 -outfmt 6 -out chr1.table

结果还是出不来，我看了一下发现里面的代码涉及的gff文件和我们这个gff文件不一样，所以做不了的

具体我按照的其他方法做的 <https://github.com/kelkar/Discover_pseudogenes> 这个比较简单，但要运行很久。【就按照这个做了，其他】

除此之外，我还尝试了pseudofinder，这个软件不知道为什么报错，就差一步，但是就是运行不了，所以放弃了。

Linux安装R语言：conda install r-essentials --yes

Braker软件

关于这个软件，我也尝试了很久，首先用了braker，发现缺少三个perl的module，但是我用了cpan发现速度太慢了，然后用cpanm，速度很快安装好了但是不知道为什么还是报错说缺少模块，然后我上网找了，给cpan设置了一个中国的接口。

我尝试了阿里云镜像，360镜像都没用，最后用了搜狐的镜像速度才快了很多

cpan[1]> o conf urllist http://mirrors.sohu.com/CPAN

cpan[2]> o conf commit

然后braker2软件有一个module安装不上去，不知道什么原因，我就放弃了，安装了braker，好在缺少的3个module全部安装上去了。

Braker运行其实不太顺利，一直出现一个问题。

WARNING: Number of good genes is low (0

). Recomended are at least 300 genes

ERROR: Number of good genes is 0, so the parameters cannot be optimized. Recomended are at least 300 genes。

我个人觉得是转录组差距太大。所以造成的，我认为应该把预测的转录本比对到基因组当bam问及比较好，太复杂了。所以这个问题先放到一边，反正装好了就行。参数还需要摸索摸索。

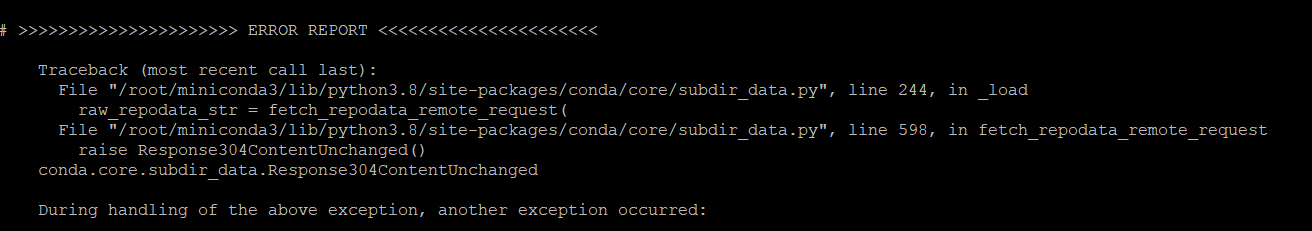
下面摸索了一下fastp的参数

fastp -f 10 -F 10 --detect\_adapter\_for\_pe -x -i R1.fq.gz -I R2.fq.gz -o R1.out.fq.gz -O R2.out.fq.gz

一般用这个吧，改一下低质量碱基标准和低长度碱基标准好像也差不多，所以就用这个好了。。

Conda 问题：

Conda active .... conda deactive



解决办法：conda clean --packages && conda clean --all && conda update --all

茯苓基因组的基因预测，我们采用的是使用网上转录组数据对基因组进行hisat比对之后，利用bam2hint对外显子内含子边界进行锚定，再使用braker流程调用augustus进行基因预测。公司的预测方法，因为他这里测的rna序列都很不好，所以他用的是网上的数据。因为网上下载的数据是sra格式的，所以我们要转换格式到fastq。

。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。

有一说一这个公司确实不行，他要说就说清楚嘛，后面有加了一些话，由于提供的转录数据为单端表达谱数据，无法拼接出完整转录本，因此本基因组的基因预测方法，使用真菌基因组适用性良好的braker预测流程。使用比对的表达谱的数据，构建基因预测hints，并使用fungus参数，进行基因的预测。

所以这个预测是完全行不通的。

不用转录组的方法其实也能预测，直接braker.pl --cores 16 --gff3 --genome=/mnt/usb/braker /Wolfiporia\_cocos.masked.fasta --esmode --fungus

Linux查看是否已经按照某个moudle：Perl -MMail::Sender -e "print\"module installed\n\""

注意！！！！！

安装braker2不要conda install braker2 贼坑 血的教训 2.1.4似乎有bug

安装直接conda create -n braker2 braker2=2.1.5

Braker2流程：

软件augustus在新版braker2会出现bug

train.f.gb does not contain any training genes 解决看：<https://github.com/Gaius-Augustus/BRAKER/issues/196>

braker.pl --cores 12 --gff3 --genome Wolfiporia\_cocos.masked.fasta --bam 111.bam --fungus

结果会生成文件braker.gff3，然后蛋白和cds不要看。直接运行后面的步骤

1. 提取所有cds并翻译为氨基酸，寻找氨基酸序列小于50的并手工去除.
2. 排序 gff3\_sort -g braker.gff3 -i --output\_gff sort5.gff3
3. 把输出文件放到windows筛选出cds,mrna,gene,exon并粘帖在linux上 linux也可以筛选
4. perl gff3rename.py sort6.gff3 WC >sort7.gff3 （改名）

在文件夹 /public/home/yyang/zzzzz进行 记得 conda activate braker2

Cafe\_fig做进化树。。我记得配置相当复杂，还好我已经和配置好了

在自己的虚拟机里。使用

python CAFE\_fig.py example\_result.cafe -c Isoptera=zne,mna -pb 0.05 -pf 0.05 --dump test/ -g .pdf --count\_all\_expansions

具体用法见提出问题与解决

基本会遇到问题这个就要根据

export DISPLAY=localhost:1

xhost +192.168.137.89

就可以解决

python CAFE\_fig.py example\_result.cafe -c Isoptera=zne,mna -pb 0.05 -pf 0.05 --dump test/ -g .pdf --count\_all\_expansions

线粒体组装软件ARE已经在服务器上装好了，注意他用的是python2.7

所以我们要conda activate py2.7

然后继续运行，参数什么的参考

<http://blog.sina.com.cn/s/blog_83f77c940102vnyu.html> 以及

<https://www.sohu.com/a/142824010_761120>

基因组装图：这个是有关网页的知识，有点复杂。

先安装软件Tomcat 安装有点复杂要修改环境变量。具体看https://zhuanlan.zhihu.com/p/97887313

打开软件D:\云服务器从零开始\apache-tomcat-9.0.37

在bin目录下startup（默认配置好的）。就可以了

然后吧文件assembly-stats复制到

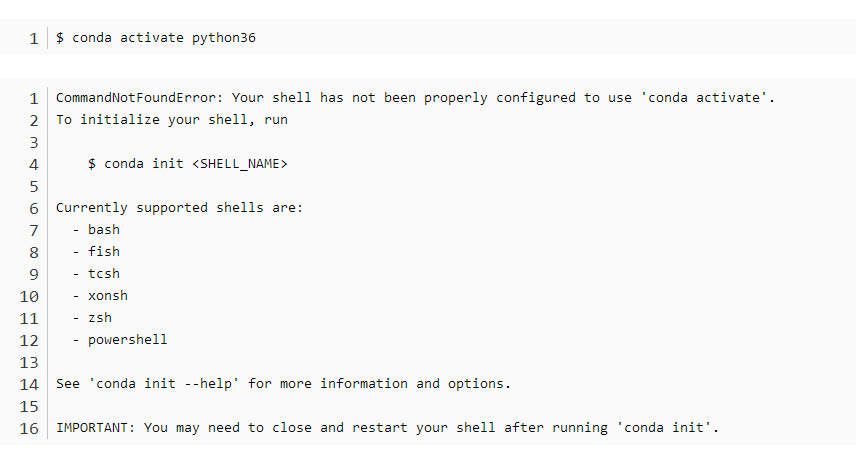
D:\云服务器从零开始\apache-tomcat-9.0.37\webapps

然后在网页上比如谷歌打入http://localhost:8080/assembly-stats/wc-assembly-stats.html?

就可以看到了。具体输入文件看readme。需要经过2个脚本处理。

然后打开软件Prince，吧谷歌上的网页保存，输入到Prince上处理得到pdf文件

Conda错误：



这个是因为虚拟环境没有退出，运用conda deactivate。

数据库mysql用法

开启

service mysqld restart

登入 mysql -u root -p

密码可以不写

然后创建新的用户出现问题 The MySQL server is running with the --skip-grant-tables

解决；mysql> set global read\_only=0;

mysql>flush privileges; 就可以了.

[pasa用法]

Launch\_PASA\_pipeline.pl -c alignAssembly.config -C -R -g example.fa.masked

-t transcript.fasta.clean -T -u transcript.fasta --ALIGNERS blat,gmap --CPU 12

PASApipeline-v2.3.3/scripts/pasa\_asmbls\_to\_training\_set.db --pasa\_transcripts\_fasta <prefix>.assemblies.fasta --pasa\_transcripts\_gff3 <prefix>.pasa\_assemblies.gff3 []

集群服务器：

Pestat ：各节点负载再决定切换到哪个节点上提交作业。

Qstat:查看作业成功与否

Qsub :提交任务

Qdel:删除任务，qdel 111

Pnodes：每个节点被占用的cpu核心数。

#PBS -N blast

#PBS -l nodes=1:ppn=10

#PBS –l walltime=7200:00:00

#PBS -q batch

#PBS -V

cd $PBS\_O\_WORKDIR

time blastn -query ./ZS97\_cds.fa -out ZS97\_cds -db ./MH63\_cds -outfmt 6 -evalue 1e-5 -num\_threads 10