**LEB期末总结**

沈南奇2300934024

# Linux操作基础

## 初识Linux

Linux 内核最初只是由芬兰人林纳斯·托瓦兹（Linus Torvalds）在赫尔辛基大学上学时出于个人爱好而编写的。Linux 是一套免费使用和自由传播的类 Unix 操作系统，是一个基于 POSIX 和 UNIX 的多用户、多任务、支持多线程和多 CPU 的操作系统。Linux 能运行主要的 UNIX 工具软件、应用程序和网络协议。它支持 32 位和 64 位硬件。Linux 继承了 Unix 以网络为核心的设计思想，是一个性能稳定的多用户网络操作系统。

来自[Linux 简介 | 菜鸟教程 (runoob.com)](https://www.runoob.com/linux/linux-intro.html)

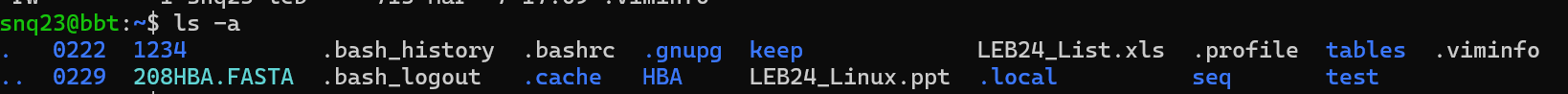
个人使用感受：相比于图形化操作的Windows系统，Linux系统入门稍有困难，但是其所独有的一些优点使其成为应用生物信息技术时非常不错的系统选择,如Linux上运行的软件大多可以自由获取，无需购买软件的商用授权；相比于Windows和Mac，开源的Linux系统更加支持个人修改，定制适合自己工作的Linux。

## Linux常用命令

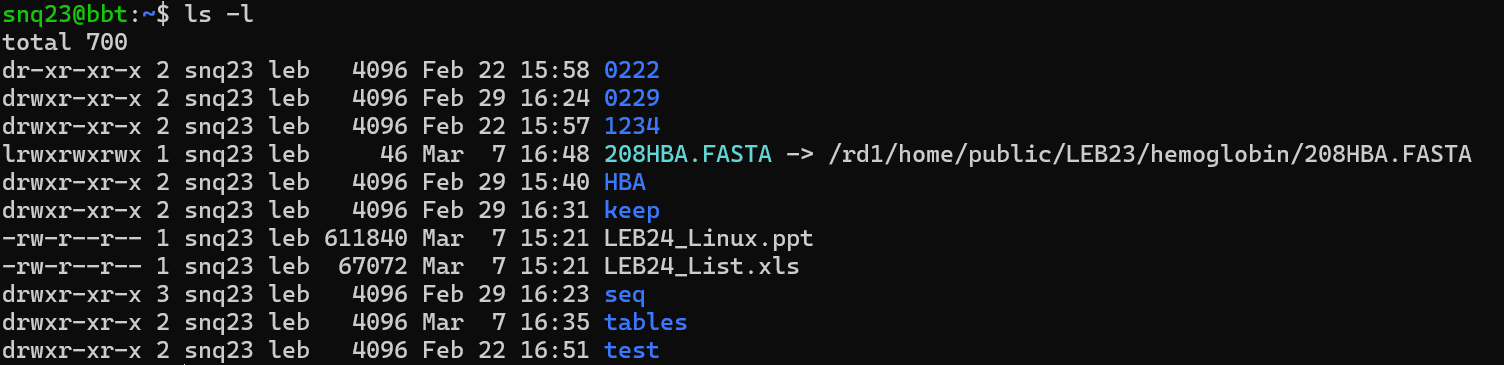
### 文件目录操作命令

**命令1：ls**

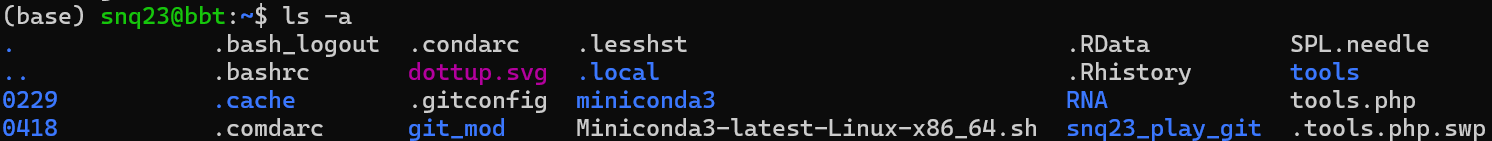
ls 显示当前目录下子目录和文件名

ls -a 显示当前目录下以句点起始的隐藏目录和隐藏文件

ls HBA/ 显示子目录HBA下的文件

ls -l 显示当前目录下子目录和文件名详细信息

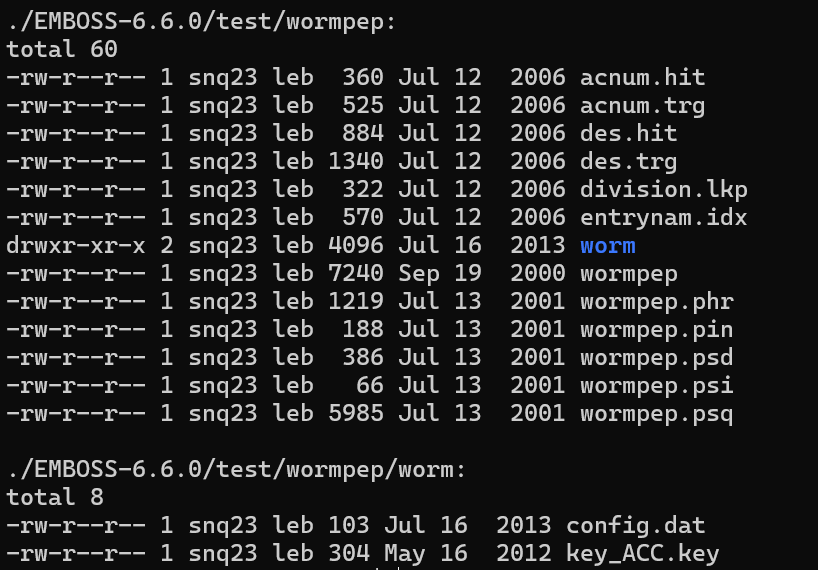
ls -a 显示当前目录下以句点起始的隐藏目录和隐藏文件



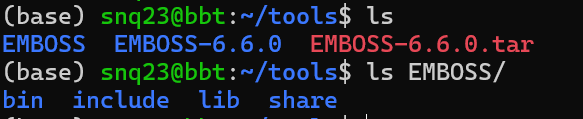
Ls -lR显示当前目录下所有子目录和文件详细信息，逐级展开

下图由于文件夹内容过多，截取一部分



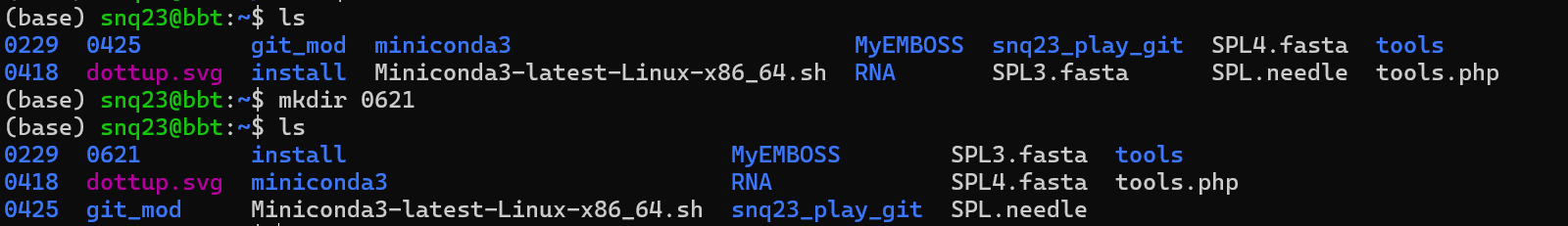


展示EMBOSS文件夹下的文件

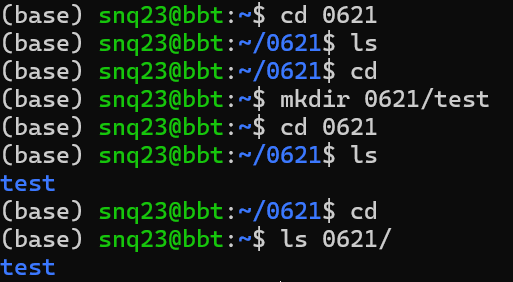


命令2：mkdir

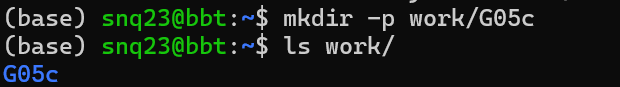
mkdir 0621 在当前目录下创建0621子目录



mkdir 0621/test 在0621子目录下创建二级子目录test

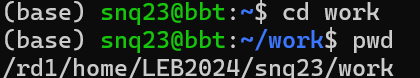


在当前目录下创建work子目录，并在子目录work下创建二级子目录G05C



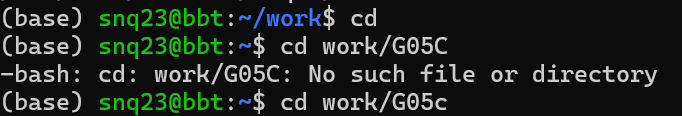
命令3：cd（可以和pwd结合使用，cd之后查看当前位置）

cd work 进入work子目录



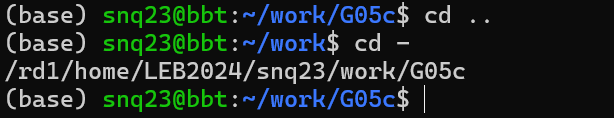
cd 回到用户根目录（snq23）

cd work/G05c直接进入子目录work下的二级子目录G05c



cd .. 返回上级目录

cd - 返回进入本目录前的目录（粗略理解成“撤销一步移动”）



命令4：cp

cp cat1 cat2 复制cat1的内容，放到文件cat2中



cp -i cat1 cat2 若cat2存在，询问是否覆盖



cp ../seq/HSHBA\_cds.FASTA . 将上级目录seq子目录中文件HSHBA\_CDS.FASTA复制到当前目录中（注意命令末尾有一个“ .”，指的是当前位置）

cp \*.FASTA ../keep/ 将当前目录下后缀为.FASTA的文件复制到上级目录的keep 子目录中

命令5：mv

mv cat1 cat3 将文件cat1改名为cat3

mv -i cat1 cat3 将文件cat1改名为cat3,若文件存在，询问是否修改

命令6：rm

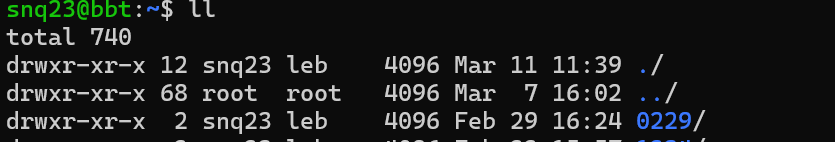
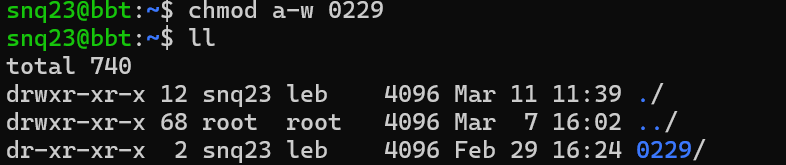
rm cat1 删除文件cat1

rm -ri 0222 删除子目录0222及文件

rm –ri temp/\* 删除子目录temp及该子目录每删除一个文件，系统提示是否 确认删除，回答y确认，回答n操作无效

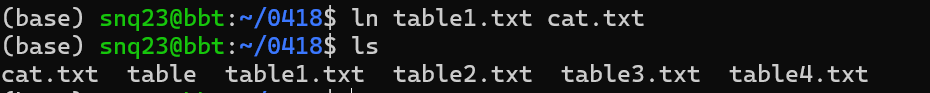
rm –i cat\* 删除以cat起始的文件，每删除一个文件，系统提示是否确 认删除，回答y确认，回答n操作无效

**命令7：chmod**

****chmod a-w 0229 取消对0229的修改权限（要添加权限用+）

**命令8: ln**

Ln table1.txt cat.txt创建table1.txt 的链接cat.txt



**命令9: file**

file cat1 查看文件cat1的类型



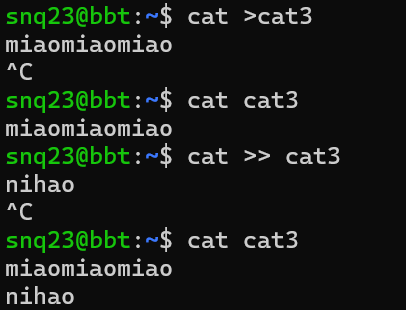
### 文本文件操作命令

**命令1：cat**

cat >cat1 创建文本文件cat1

cat cat1 显示文本文件cat1的内容

cat >>cat1 在cat1末尾追加新的文本

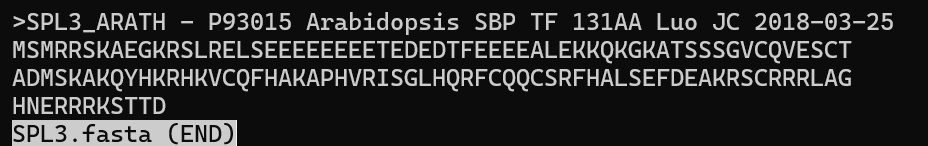
cat –n table1.txt 显示文本文件table1.txt的内容，并显示行号

cat –A table1.txt 显示文本文件table1.txt的内容，包括制表键^I、行终止 符$等控制符

**命令2: less**

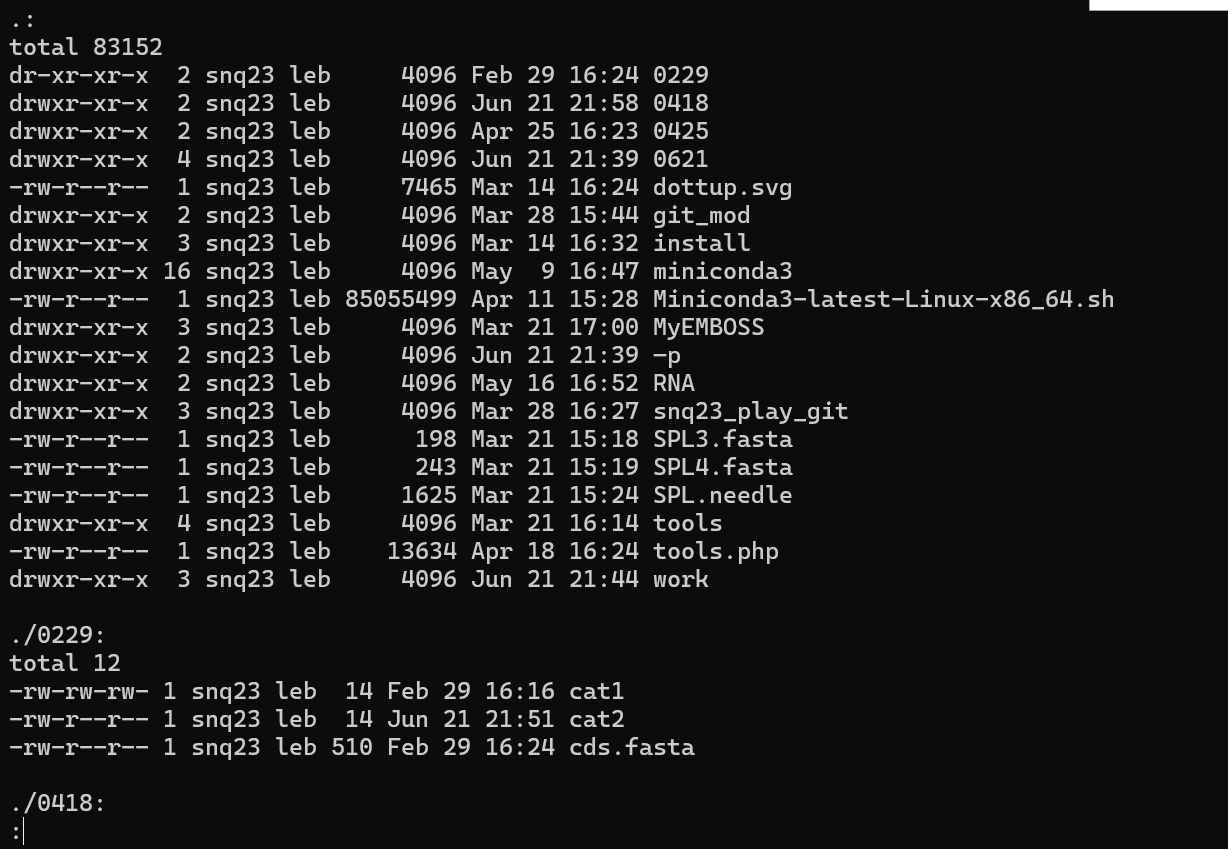
less SPL3.fasta 逐屏显示SPL3.fasta ，回车进一行，空格进一页，q终止显示（man的时候也这么操作）



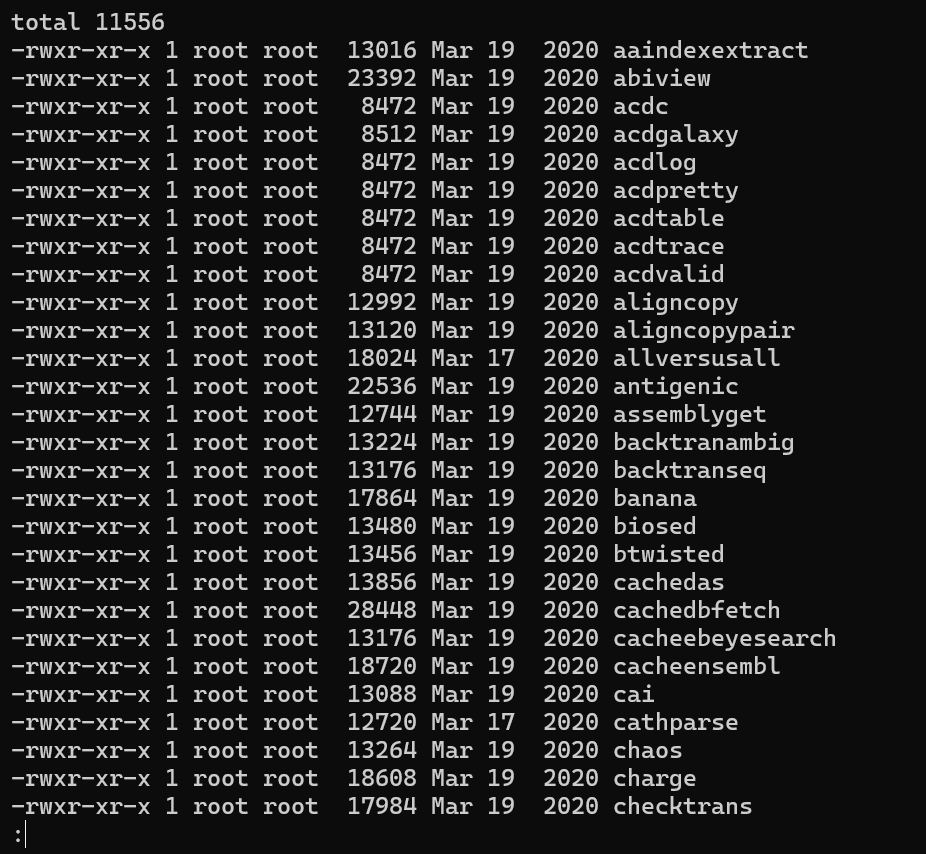


ls –lR | less 逐屏显示当前目录下所有子目录和文件详细信息





ls -l /usr/local/bin | less 逐屏显示/usr/local/bin/下所有文件信息，可用斜杠搜索关键词，如needle

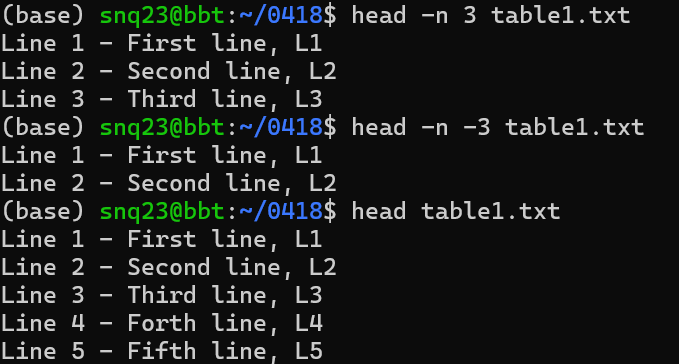


命令3：head

head 209HBA.txt 显示文件209HBA.txt中前10行（默认参数-n 10）

head -n 3 table1.txt 显示文件table1.txt前3行（显示字节，把-n改成-c）

head -n -3 table1.txt 显示文件table1.txt除最后3行外的其它行

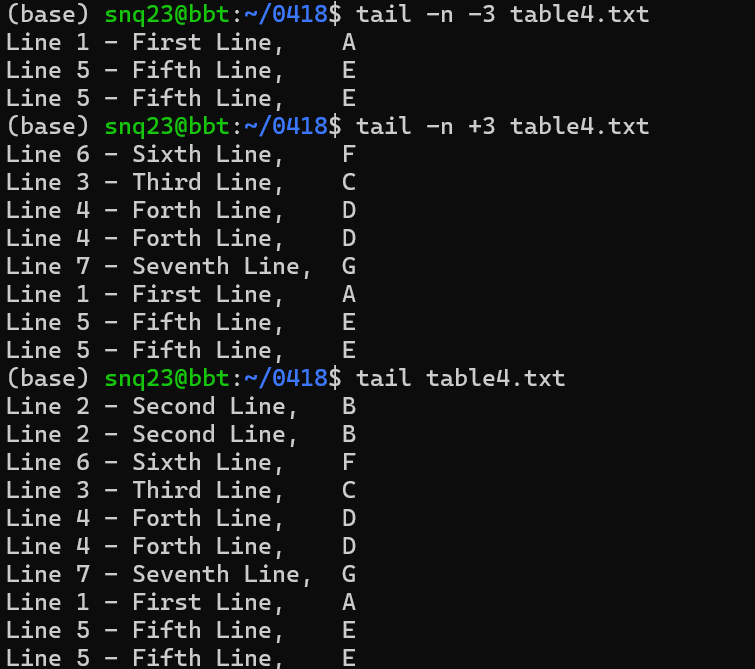


命令4: tail

tail 209HBA.txt 显示文件209HBA.txt中最后10行

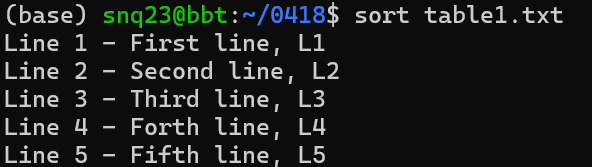
tail -n -3 table4.txt 显示文件table1.txt最后3行

tail -n +3 table4.txt 从第3行开始显示文件table4.txt中所有行

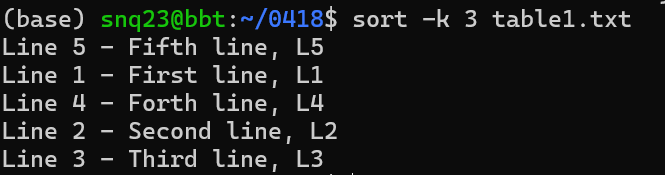


命令5: sort

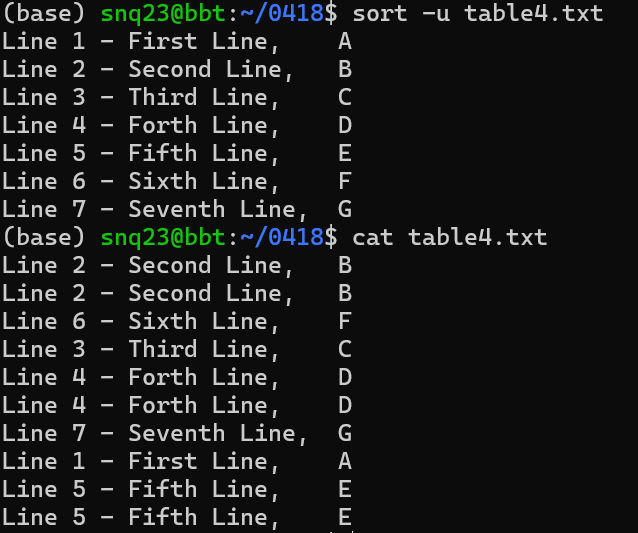
sort table1.txt 按字母表顺序对table1.txt中的行进行排序



sort –k 3 table1.txt 基于第3字段按字母表顺序对table1.txt中的行进行排序（按空格区分字段，若各行的该字段完全相同，则依据下一字段排序,此处按照F-S-T顺序排序,即First-Forth-Second-Third）

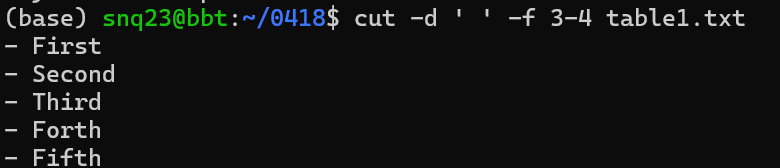


sort -u table4.txt 按字母表顺序对table1.txt中的行进行排序，不显示相同行

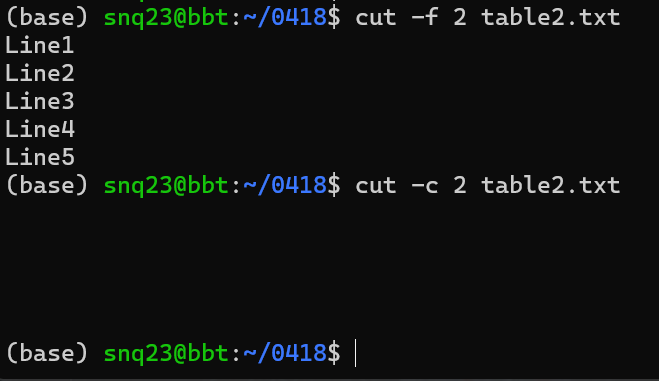


命令6：cut

cut –d ‘ ‘ –f 3-4 table1.txt 提取table1.txt中以空格符分割的第3-4个字段



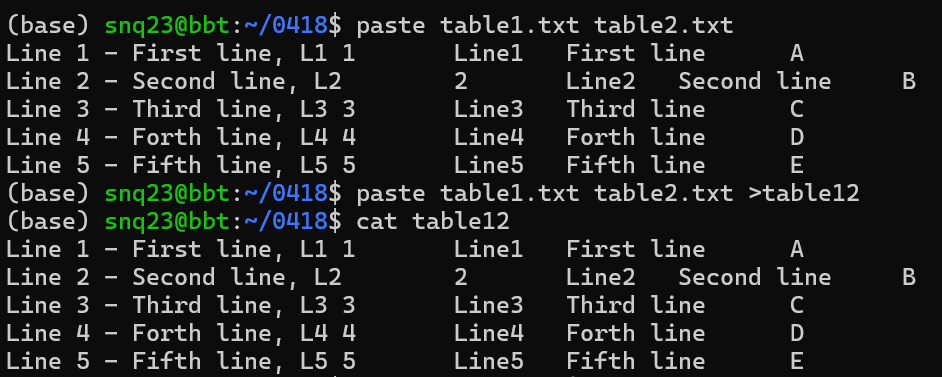
cut –f 2 table2.txt 提取table2.txt中第2个字段（以制表符分割）（-c为字符）



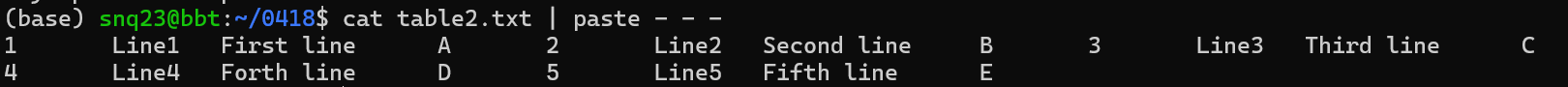
命令7: paste

paste table1.txt table2.txt 逐行合并table1.txt和table2.txt（不创建文件）

paste table1.txt table2.txt >table12 逐行合并table1.txt和table2.txt并创建文件table12



cat table2.txt | paste - - - 将table2.txt中每三行合并到一行显示, -数决定行数，带空格



命令8：wc

wc table1.txt 统计并显示文件table1.txt行数、由空格分隔的单词数和字符数

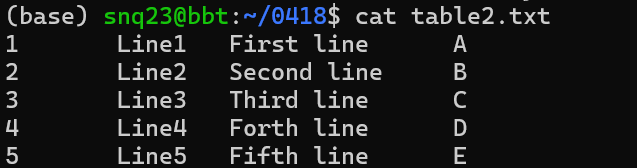
grep “^>” 209HBA.FASTA | wc -l 统计并显示文件209HBA.FASTA中首字符为大于号（>）的行数

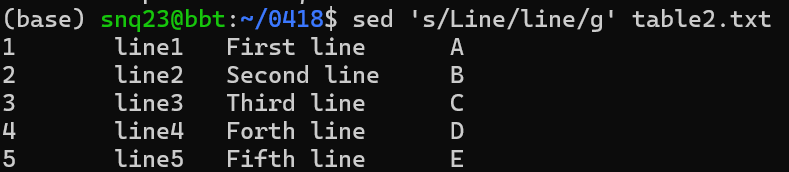
命令9: grep

grep “^>” 209HBA.FASTA 检索并显示文件209HBA.FASTA中以大于号起始的行数（与上一个命令中演示的grep区别在于，本命令是显示行的内容，上一个是显示行的数目）

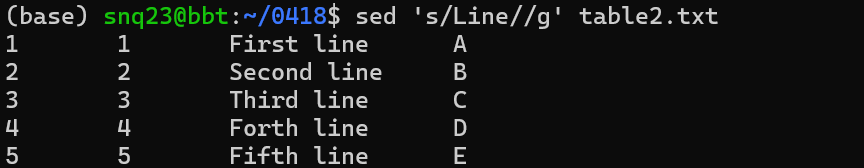
命令10: sed

sed ‘s/Line/line/g’ table2.txt 将文件table2.txt中Line改成line





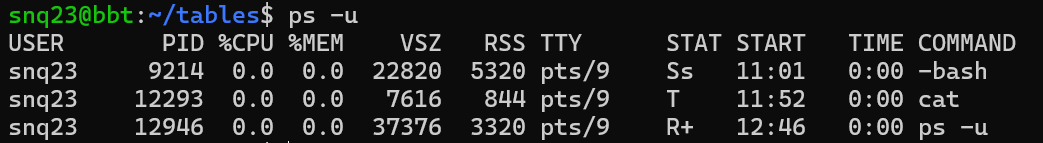
sed 's/Line//g' table2.txt将文件table2.txt中“Line”删除



sed -i ‘s/HBA\_//g’ 209HBA.id 将文件209HBA.ID中的“HBA\_”删除，修改结果直接写入 原文件中，而非输出到屏幕

### 其他常用命令

命令1：top 动态显示系统运行状态，包括处理器、内存使用情况以及进程，q退出

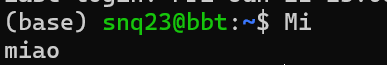
命令2：ps –u 显示本用户当前进程信息

命令3：ps -ef | less 逐屏显示所有用户当前进程详细信息，可用方向键上、下、 左、右移动显示页面

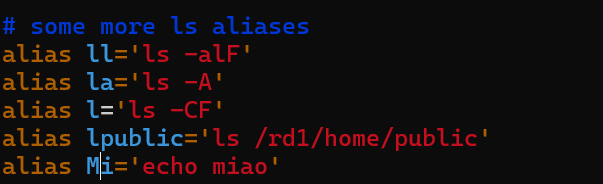
命令4：kill 31352 终止编号为31352的无用进程

命令5：history 显示最近用过的命令

命令6：alias 设置命令别名，重启生效，source .bashrc生效

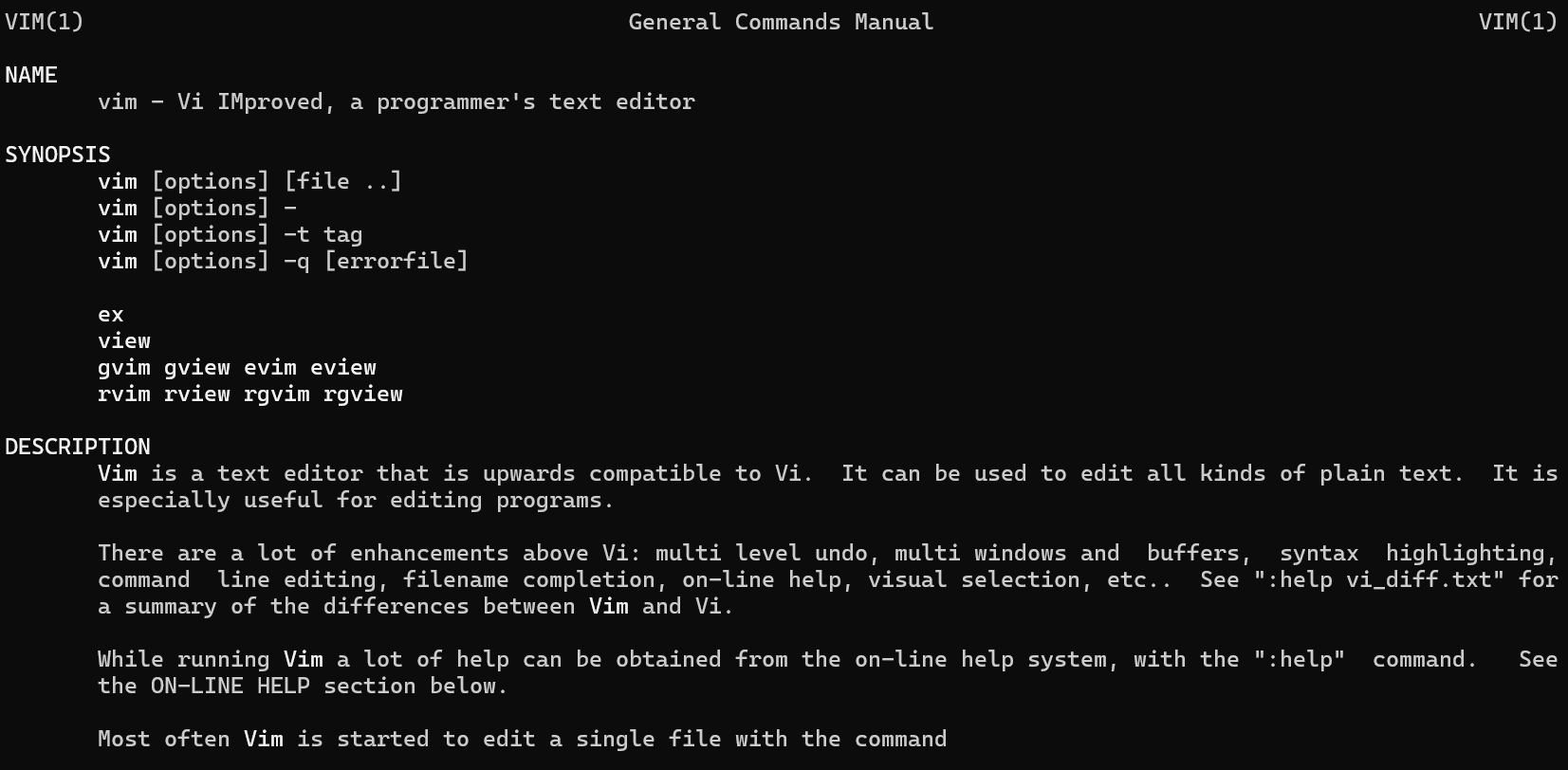




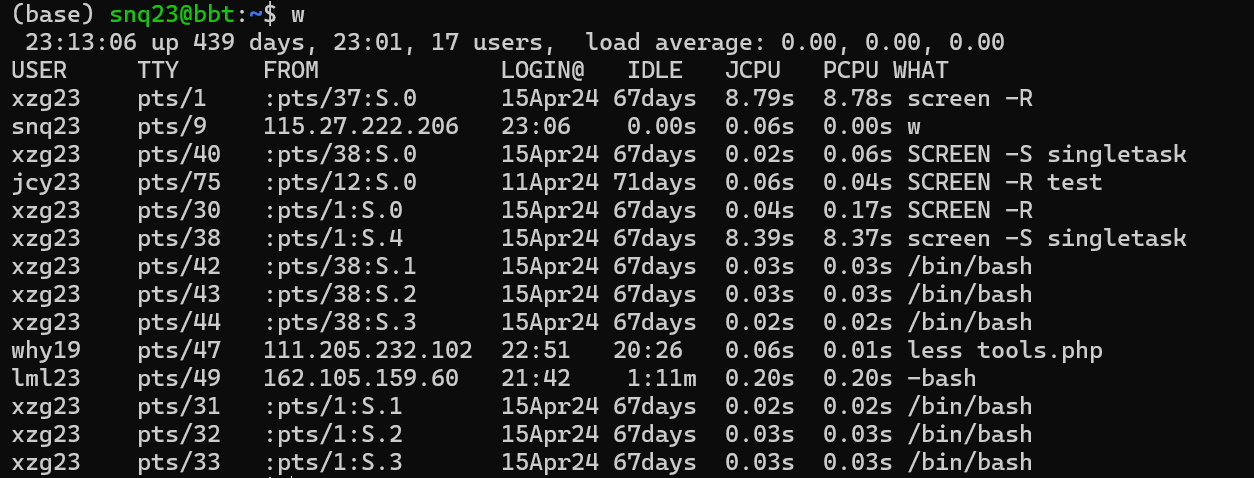


命令7：man查看帮助手册

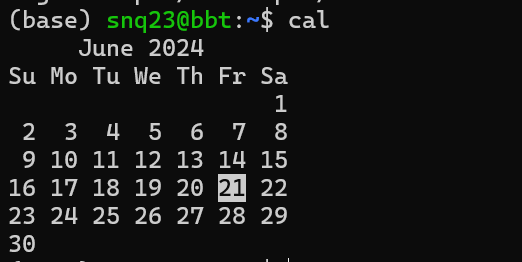




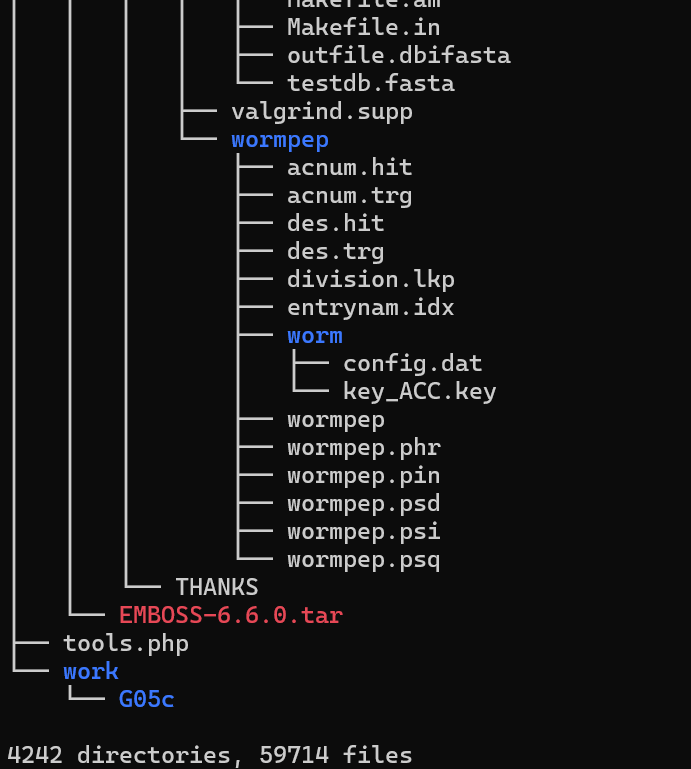
命令8：who / w ：显示用户信息



命令9：cal / date ：显示日期信息



命令10：tree ：显示目录文件树形结构



# EMBOSS

## 软件包介绍

The European Molecular Biology Open Software Suite (EMBOSS) is a high quality, well

documented package of open source software tools for molecular biology. It includes over 200 applications for molecular sequence analysis and other common tasks in bioinformatics. It integrates the core applications with a range of popular third party software packages under a consistent and powerful command line interface. The software has many useful features; for example, it automatically copes with data in a variety of formats and allows for transparent retrieval of sequence data from the web.

## EMBOSS软件包安装

EMBOSS软件包主网站  
<http://emboss.open-bio.org/>

进入网站没有直接找到下载方式

同组廖一岩分享了安装包，在运行该exe文件前提示需要先安装java <https://www.java.com/zh-CN/download/manual.jsp>

另一种：

cd ~

mkdir install

cd install

cp /rd1/home/public/EMBOSS/EMBOSS-6.6.0.tar.gz #拷贝服务器中已有的压缩包

gunzip EMBOSS-6.6.0.tar.gz #压缩包一次解压缩

tar -xzvf EMBOSS-6.6.0.tar.gz EMBOSS-6.6.0/ #压缩包二次解压缩

x - 表示解压缩的操作，即从 tar 文件中提取文件和目录。

z - 表示需要通过 gzip 进行压缩或解压缩。这个选项告诉 tar 命令使用 gzip 来解压文件。

v - 表示在处理文件时显示详细信息（verbose）。使用这个选项会显示 tar 命令正在处理的文件

名。

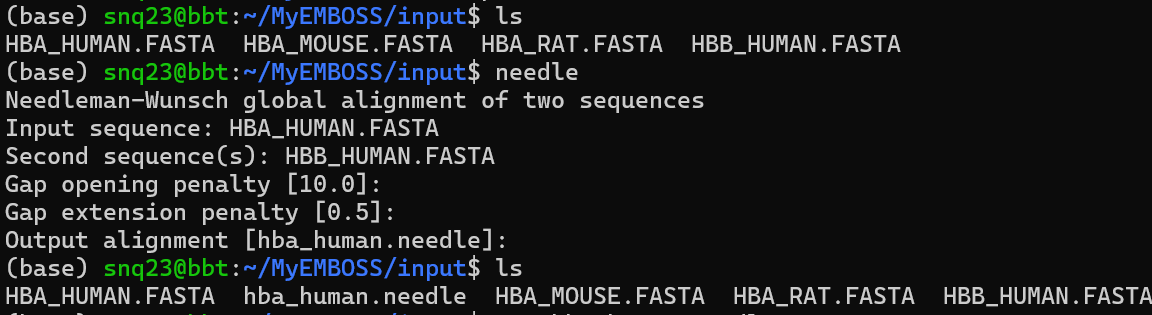
f - 表示后面跟着的是文件名。在这个例子中，文件名是 EMBOSS-6.6.0.tar

来自组员廖一岩



## 运行needle程序

### 交互式



操作流程：

准备需要比对的两个序列，以人血红蛋白HBA和HBB亚基的氨基酸序列为例。

输入needle开始运行程序。

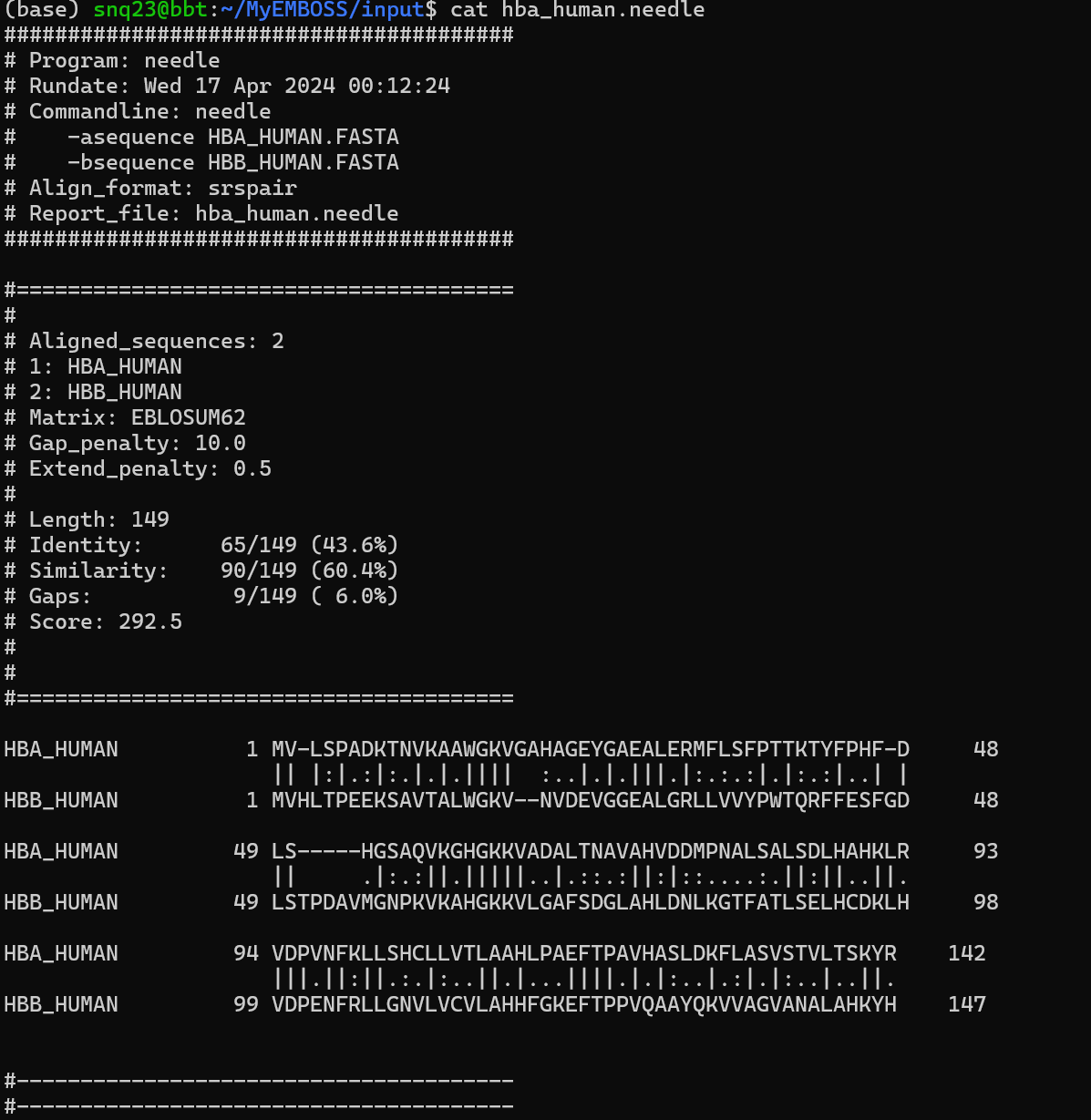
输入第一个研究序列FASTA格式文件的文件名。

输入第二个文件名。

系统给出起始空位罚分值Gap opening penalty，默认10.0，可输入自定分数后回车，若直接回车则取用默认值。

系统给出延伸空位罚分值Gap extension penalty，默认0.5，输入自定分数后回车，若直接回车则取用默认值。

系统提示输出文件名，默认小写，回车确定，也可自行输入。

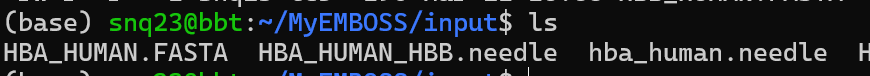
当前目录下多出hba\_human.needle文件，可cat命令查看

### 参数式

操作：

输入needle HBA\_HUMAN.FASTA HBB\_HUMAN.FASTA -gapo 10.0 -gape 0.5 -out HBA\_HUMAN\_HBB.needle

备注：

红字部分为要比对的序列文件，-gapo 10.0是起始空位罚分，-gape 0.5是延伸空位罚分，-out HBA\_HUMAN\_HBB.needle是输出名为HBA\_HUMAN\_HBB.needle的文件（下图第二个）

### 菜单式

即在输入程序名时，同时输入选项参数 -options ，程序运行时则列出所有可选参数。 这种运行方式， 对具有较多选择参数的程序十分便利

## 运行dottup点阵图程序

### 点阵图程序

知乎搜到资料：

dottup：一个word要在两个序列里面完全一模一样才有dot

dotmatcher：不完全匹配，只要在threshold以上，就有dot

dotpath：类似dottup（但是默认的word size比较小），但是序列里面的任意一段只能match到一个地方（不能对到两个不同的地方）。换句话说，没有overlap的[alignment](https://www.zhihu.com/search?q=alignment&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A41580497%7D" \t "_blank)，也就是从图上看，任意x值上都只有一个dot，任意y值上也只有一个dot。

polydot：一个[fasta文件](https://www.zhihu.com/search?q=fasta%E6%96%87%E4%BB%B6&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A41580497%7D" \t "_blank)，里面所有序列对（all pairwise）的dottup。比如有3条序列就有3\*3个dottup  
作者：Philip Yang  
链接：<https://www.zhihu.com/question/28623438/answer/41580497>

点阵图程序dottup多用于核酸序列，而dotmatcher则可用于蛋白质序列（EMBOSS软件包序列分析程序应用实例，罗静初，DOI：10.12113/202008002，生物信息学）

### 交互式

操作流程：

1.准备HBA\_HUMAN.FASTA和HBB\_HUMAN.FASTA文件

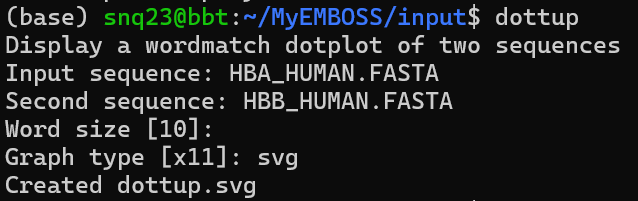
2.输入dottup运行程序

3.输入第一步准备文件的文件名

4.修改字长或选取默认值10

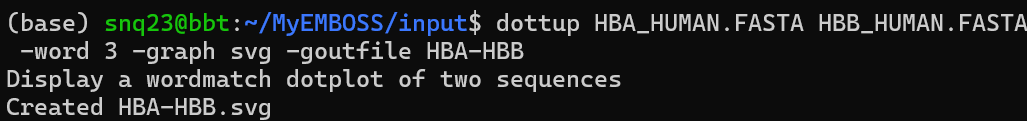
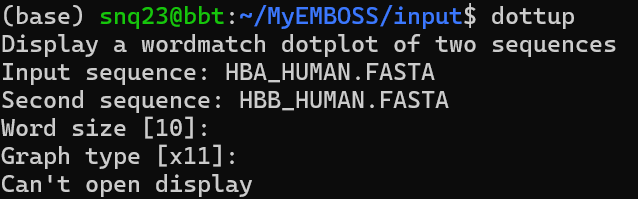
5.输出结果图形格式默认X11，输入svg（[SVG：可缩放矢量图形](https://developer.mozilla.org/zh-CN/docs/Web/SVG)）

6.生成dottup.svg文件

参数式

操作：输入dottup HBA\_HUMAN.FASTA HBB\_HUMAN.FASTA -word 3 -graph svg -goutfile HBA-HBB

备注：输入dottup，两个文件名，设置字长（-word）和输出结果图形格式（-graph），修改输出结果文件名（-goutfile）。默认输出图形格式是x11，但是默认格式生成不成功，需修改成svg格式



# GitHub

## GitHub重要性

软件开发的核心工具

项目管理的优选平台

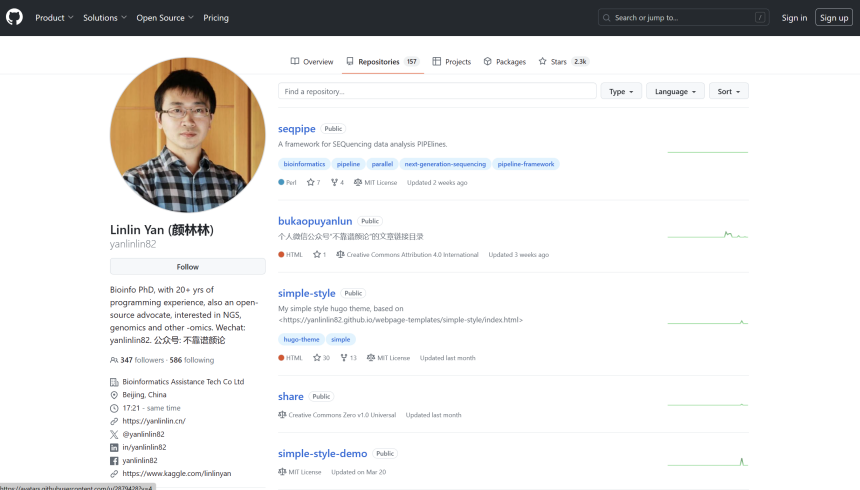
开源协作的中心舞台

GitHub 是一个存放软件代码的网站，将代码放在网站上，其他人可以在现有代码的基础上进行二次开发，不用从头开始工作。如果是团队合作，团队成员可以上传自己的工作，也可以拉取他人的。

## 基本概念

### 仓库 & 分支：

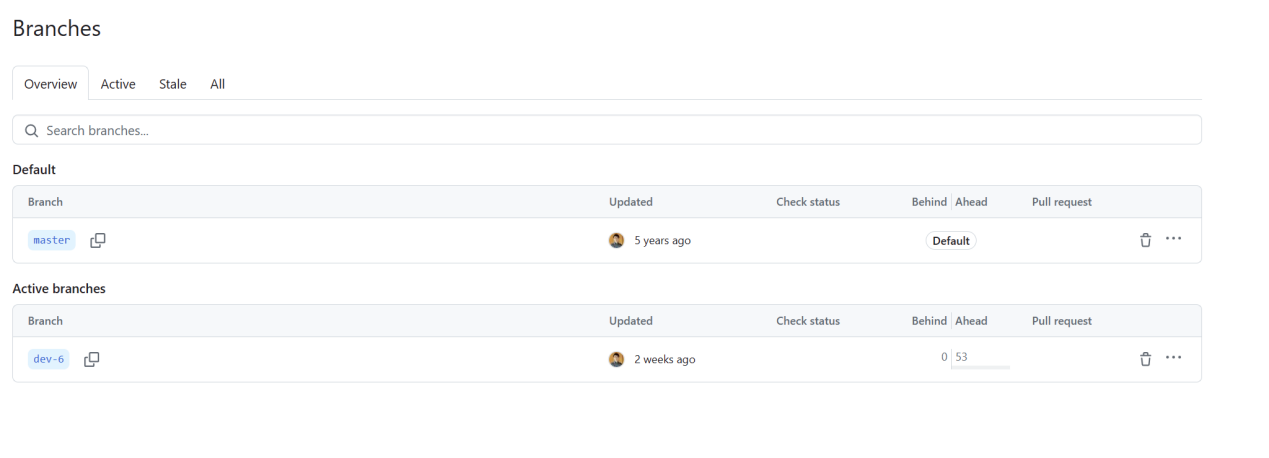
仓库(Repository): 项目的存储地点



分支(Branch): 不同版本的开发线路

主分支: 主开发线

特性分支: 新功能开发



### 提交 & 拉取请求：

提交(Commit):

保存更改的快照

提交历史: 更改的记录

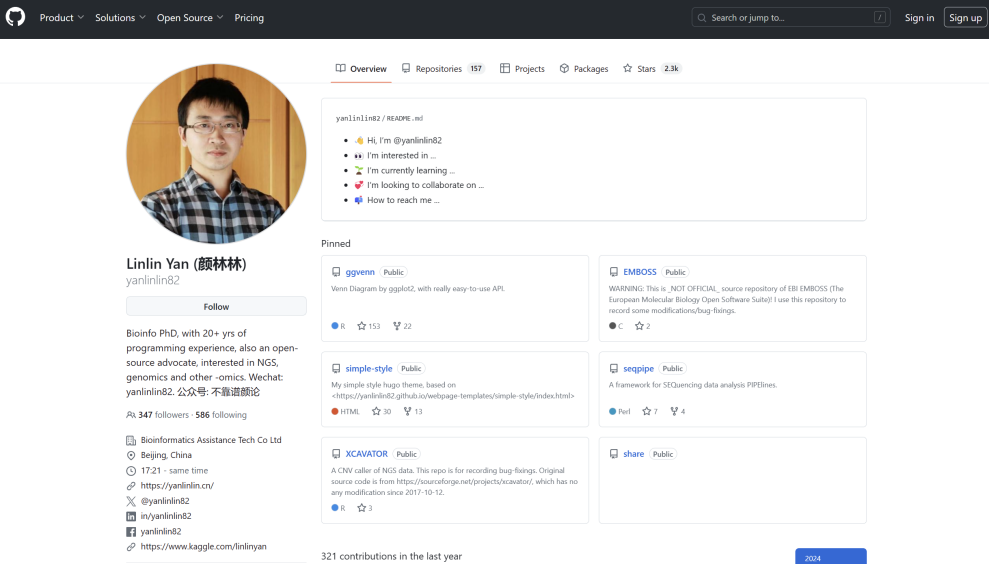
拉取请求(Pull Request):

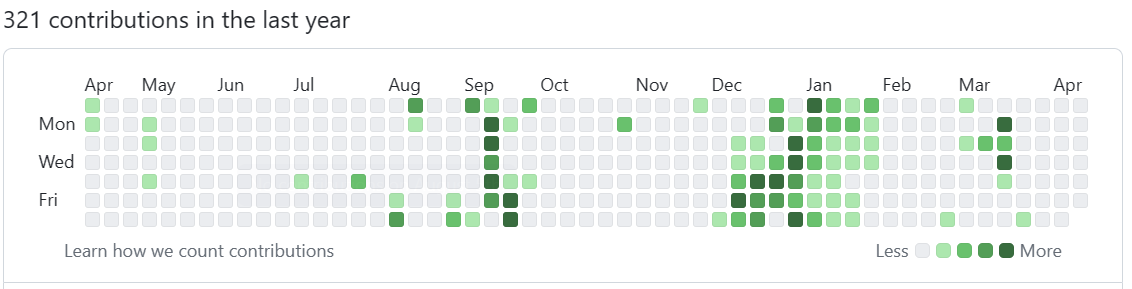
合并分支的请求

代码审查: 团队成员评审更改

合并: 将更改并入主分支

## 主页overview





上图展示在github的贡献和活跃度

# Git

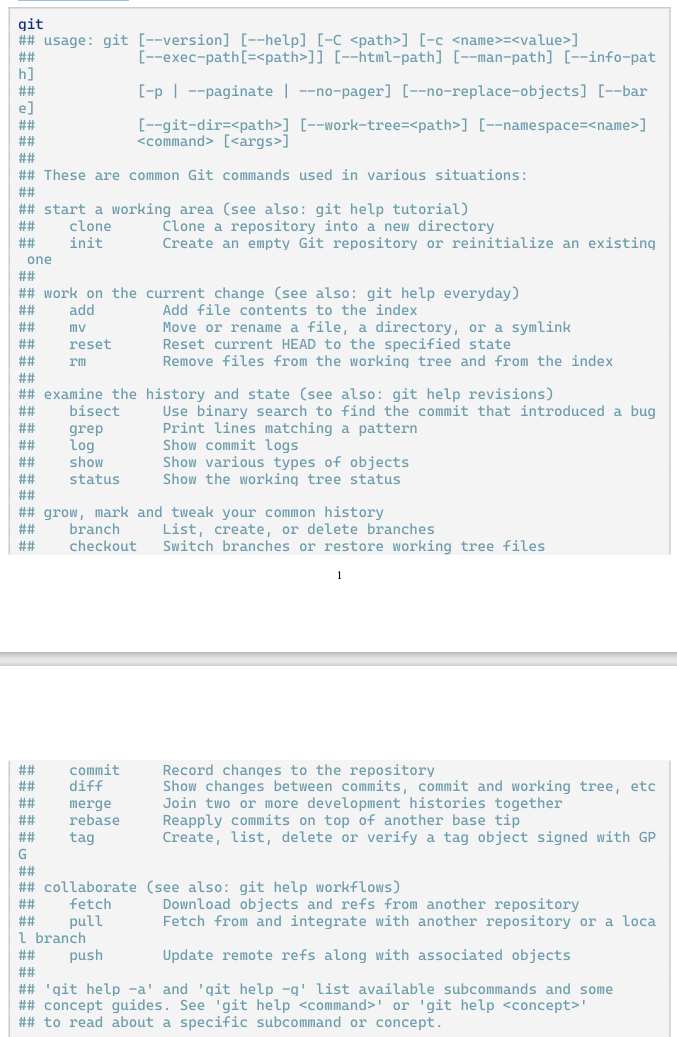
## 感觉用得上的学习资料

[Git - Book (git-scm.com)](https://git-scm.com/book/zh/v2)

[Learn Git Branching](https://learngitbranching.js.org/?locale=zh_CN)小游戏

##### 配置Git

服务器中有git，省略安装步骤，安装可以参照[Git - 安装 Git (git-scm.com)](https://git-scm.com/book/zh/v2/%E8%B5%B7%E6%AD%A5-%E5%AE%89%E8%A3%85-Git)

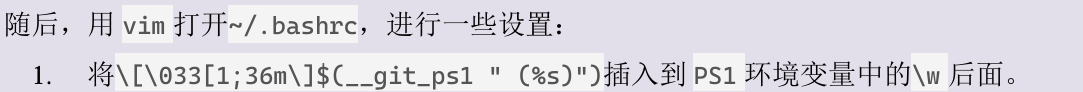
确认安装完成：

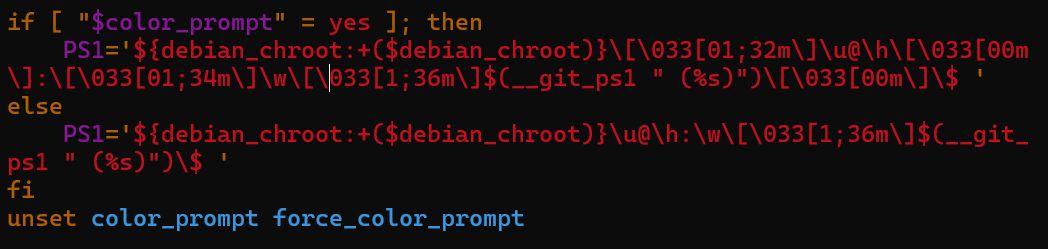
###### 配置用户邮箱和用户名



###### 下载并设置一些Git的辅助插件



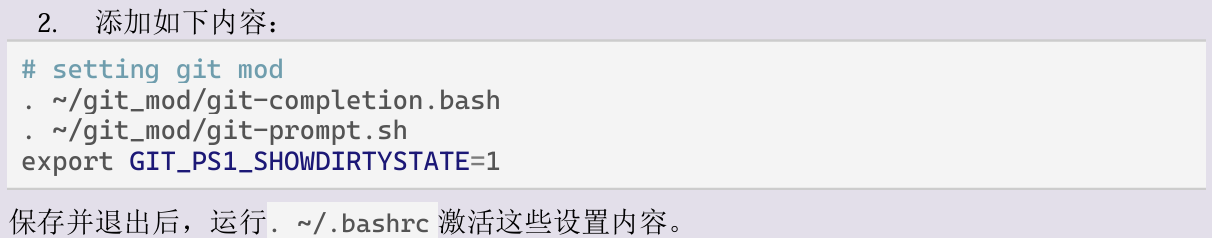




一共添加在三处

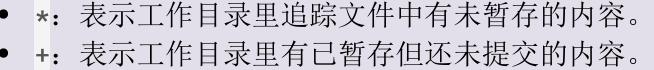
第一次进行完修改.bashrc的操作后，无法在进行任何操作，退出之后出现登录不上服务器的问题，助教认为可能是有操作失误影响了文件中的其他部分，用未修改的文件替换我家目录下文件后，可以正常登陆以及使用。所以修改.bashrc的操作还是比较“危险”的，之后助教又提供了官方教程的方法供学习。

<https://git-scm.com/book/zh/v2/%E9%99%84%E5%BD%95-A%3A-%E5%9C%A8%E5%85%B6%E5%AE%83%E7%8E%AF%E5%A2%83%E4%B8%AD%E4%BD%BF%E7%94%A8-Git-Bash-%E4%B8%AD%E7%9A%84-Git>



## 建立一个使用Git进行版本控制的项目

### 初始化



### **git status**

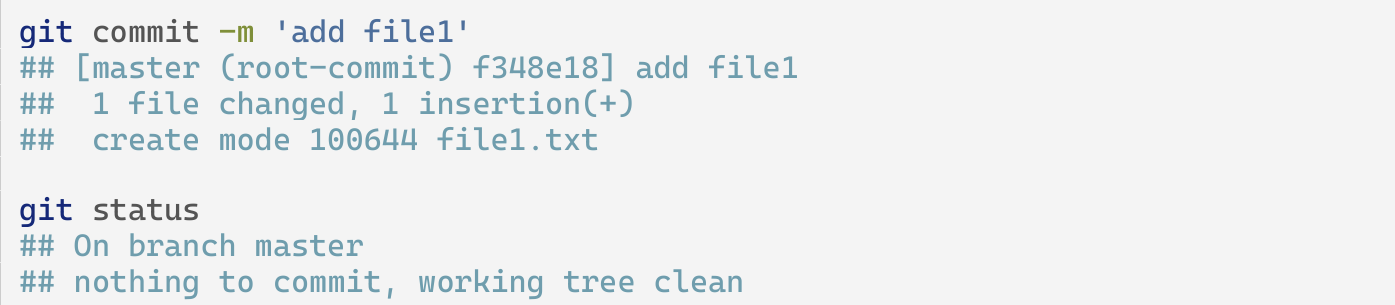
使用git status命令可以查看当前工作目录的状态，可以看到被识别出的文件是untracked的状态，是红字，表示还没有被Git所追踪。并且提示我们可以 使用git add对其进行追踪，以进行提交。

### git add

对这个文件运行git add命令，Git会把它放到一个“暂存区”，git status检查，变成绿字了

### git commit

这个命令表示将所有被暂存的改动储存起来，变成一次提交给记录下来。



### git log

通过git log命令查看提交的记录



如想撤销修改，可以回溯至之前的提交

# Anaconda

## 使用Conda

### 安装

服务器上有下载好的Conda存放于：

/rd1/home/public/RNA-Seq/software/Miniconda3-latest-Linux-x86\_64.sh

cp /rd1/home/public/RNA-Seq/software/Miniconda3-latest-Linux-x86\_64.sh ~/

cd

bash Miniconda3-latest-Linux-x86\_64.sh

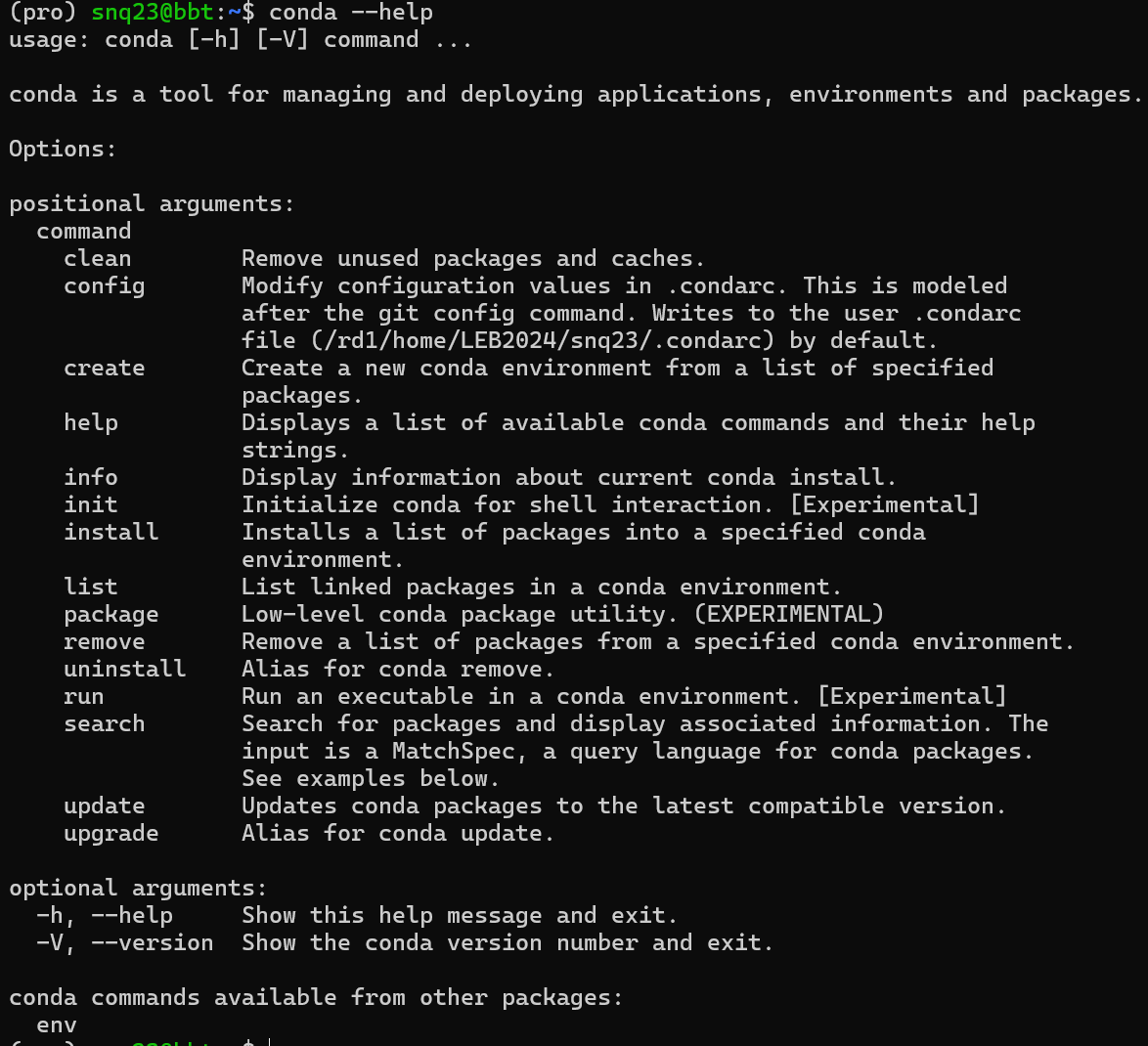
一直按回车键，直到出现：

输入yes再回车

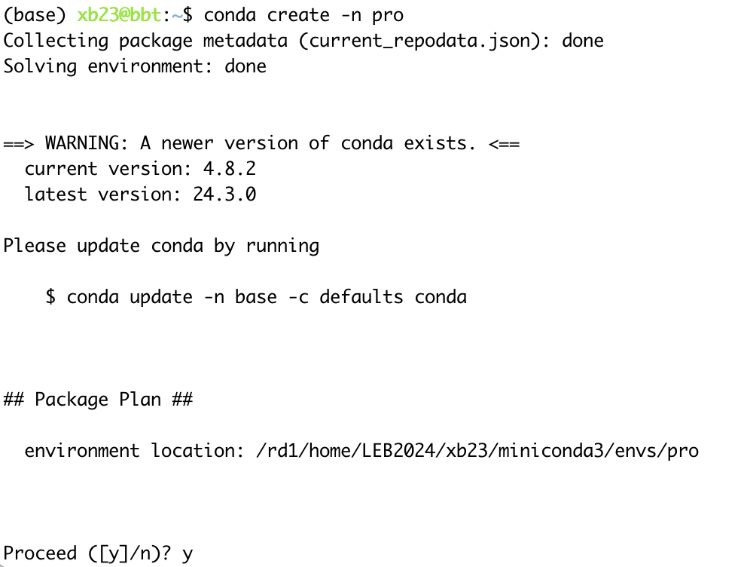
重新登陆服务器后左下角会出现base则表示安装完成

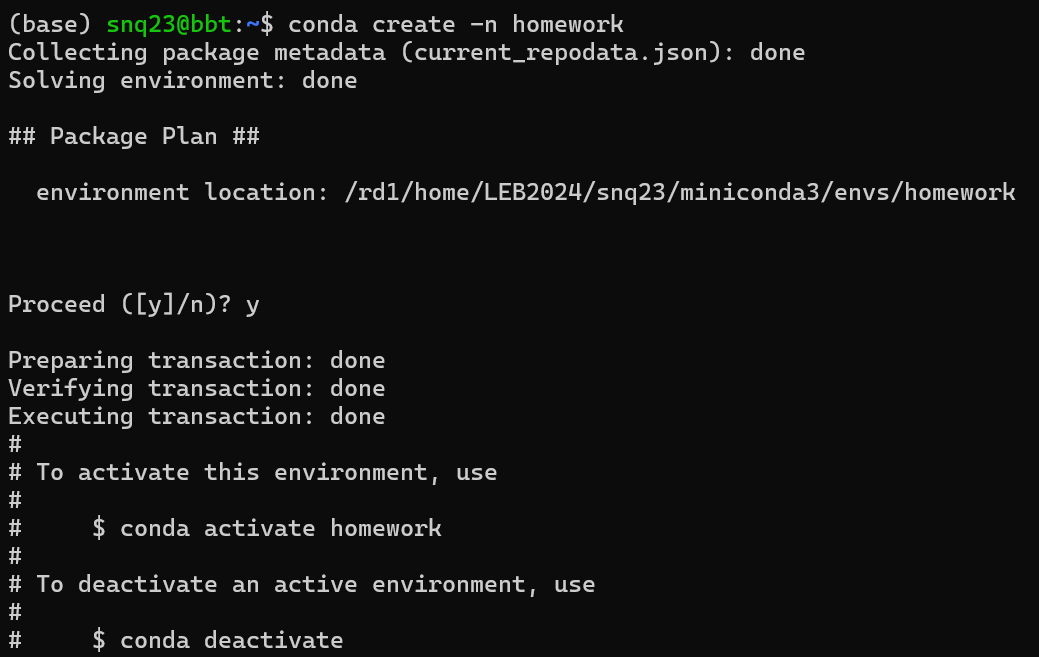


conda --help



### Conda镜像站

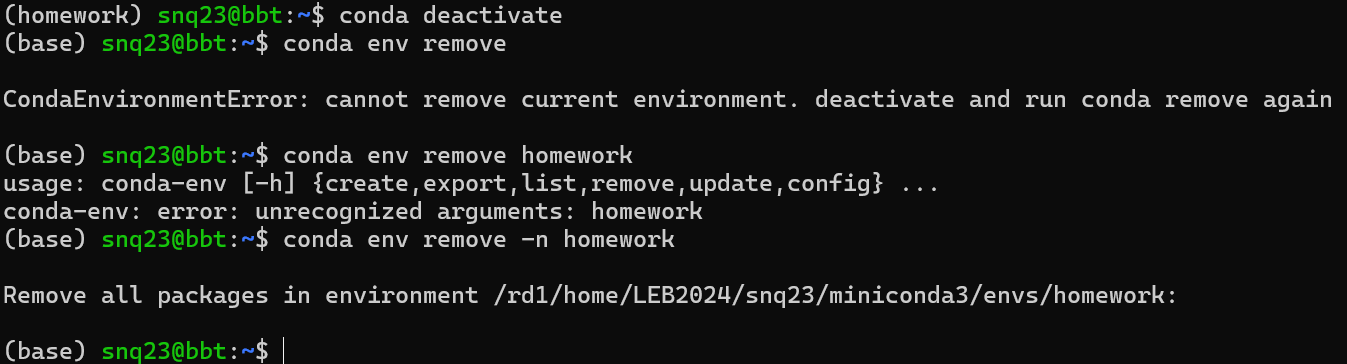
[anaconda | 镜像站使用帮助 | 清华大学开源软件镜像站 | Tsinghua Open Source Mirror](https://mirror.tuna.tsinghua.edu.cn/help/anaconda/?eqid=a38a4a4800068d0600000003645b7087) 创建激活新的环境

创建：conda create -n pro 

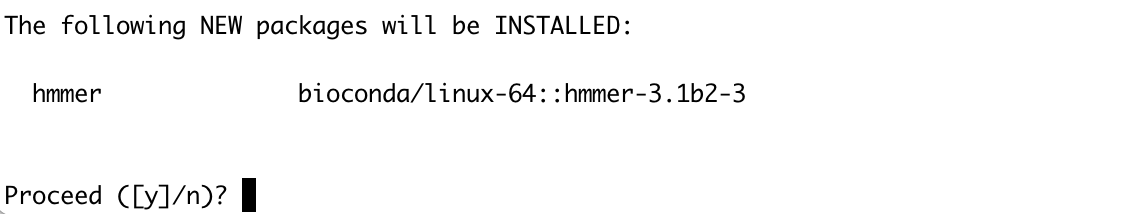
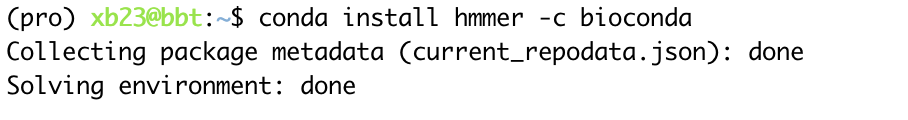




conda env remove -n删除名为homework的环境



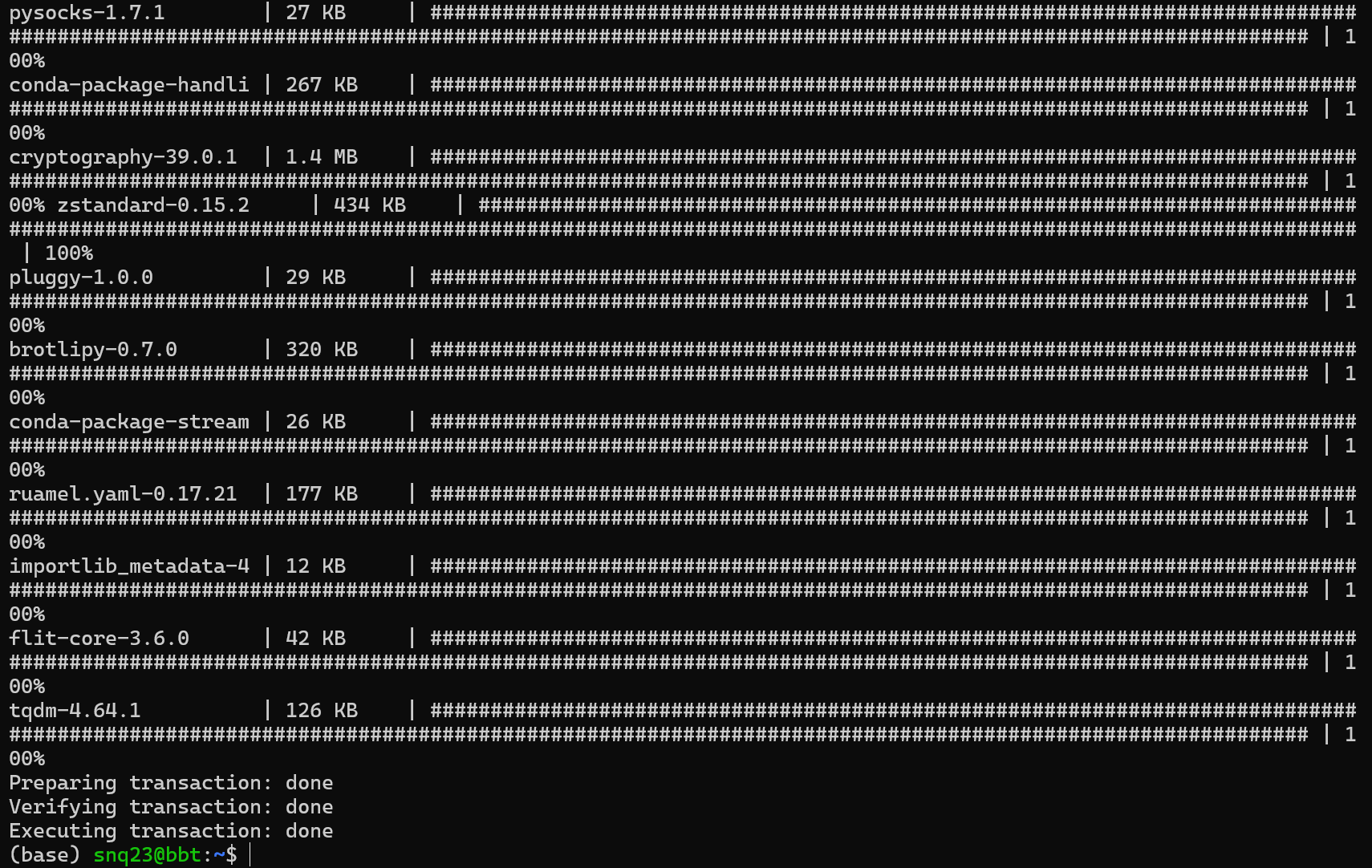
## 安装hmmer工具包

[:: Anaconda.org](https://anaconda.org/)可以搜索需要安装的包是否存在



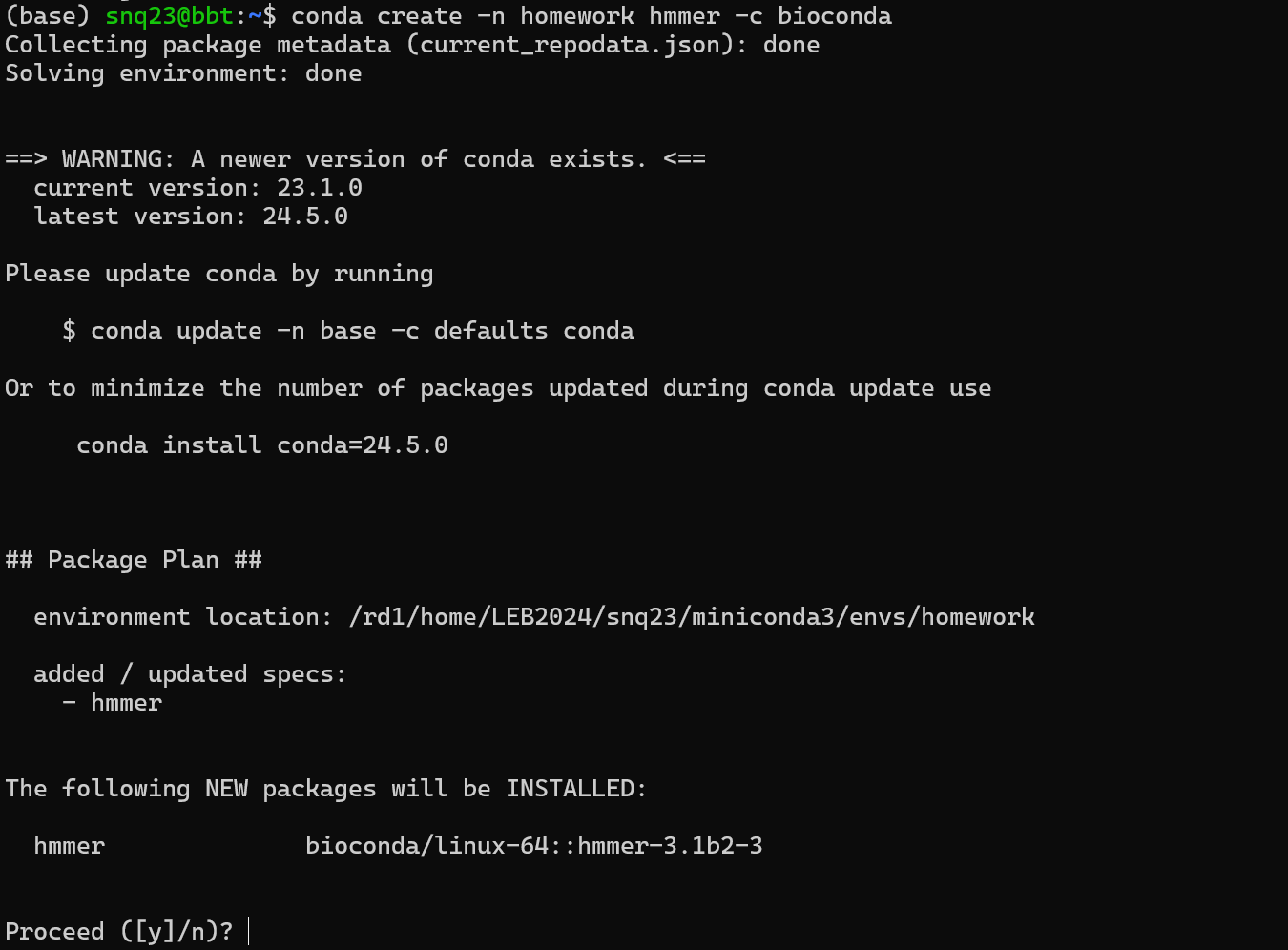
conda remove命令用来删除已经安装的工具





## 同时实现环境创建与工具包安装

conda create -n homework hmmer -c bioconda



# Vim

## Vim介绍

Vim 是从 vi 发展出来的一个文本编辑器。代码补全、编译及错误跳转等方便编程的功能特别丰富，在程序员中被广泛使用。

简单的来说， vi 是老式的字处理器，不过功能已经很齐全了，但是还是有可以进步的地方。 vim 则可以说是程序开发者的一项很好用的工具。

## vi/vim 的使用

基本上 vi/vim 共分为三种模式，命令模式（Command Mode）、输入模式（Insert Mode）和命令行模式（Command-Line Mode）。

### 启动vim



启动后默认进入命令模式，但对于空白文档没有什么命令

### 输入模式

i ：进入输入模式，在光标当前位置开始输入文本。

a ：进入插入模式，在光标下一个位置开始输入文本。

o ：进入插入模式，在当前行的下方插入一个新行开始输入文本。

个人感觉i使用最频繁

可使用的按键：

**字符按键以及Shift组合**，输入字符

**ENTER**，换行

**BACK SPACE**，退格键，删除光标前一个字符

**DEL**，删除键，删除光标后一个字符

**方向键**，在文本中移动光标

**HOME/END**，移动光标到行首/行尾

**Page Up/Page Down**，上/下翻页

**Insert**，切换光标为输入/替换模式，光标将变成竖线/下划线

**ESC**，退出输入模式，切换到命令模式

### 命令模式



[Linux vi/vim | 菜鸟教程 (runoob.com)](https://www.runoob.com/linux/linux-vim.html)

### 底线模式

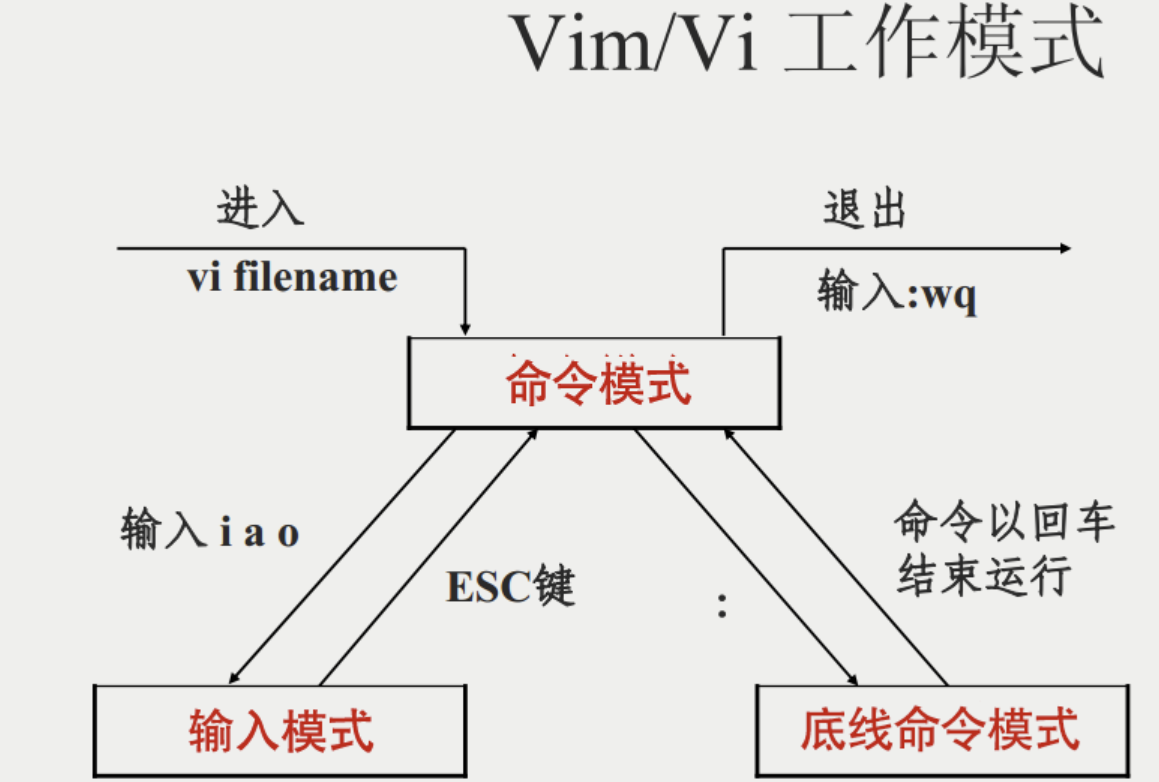
通过在命令模式下输入 : 进入底线模式，通过一系列命令对文件进行操作。与光标所在位置一般无关

:w：保存文件。

:q：退出 Vim 编辑器。

:wq：保存文件并退出 Vim 编辑器。

:q!：强制退出Vim编辑器，不保存修改

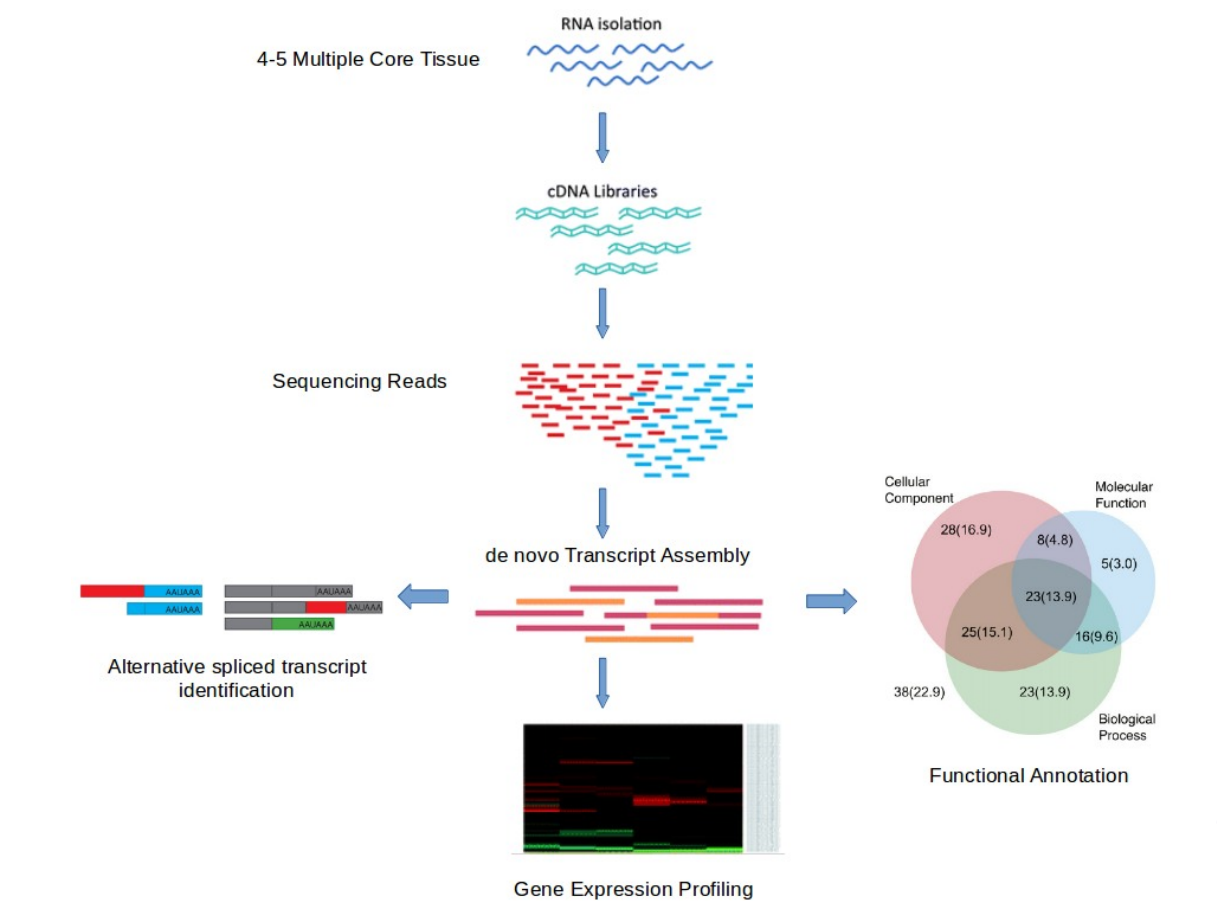


# 转录组分析

## 转录组分析简介

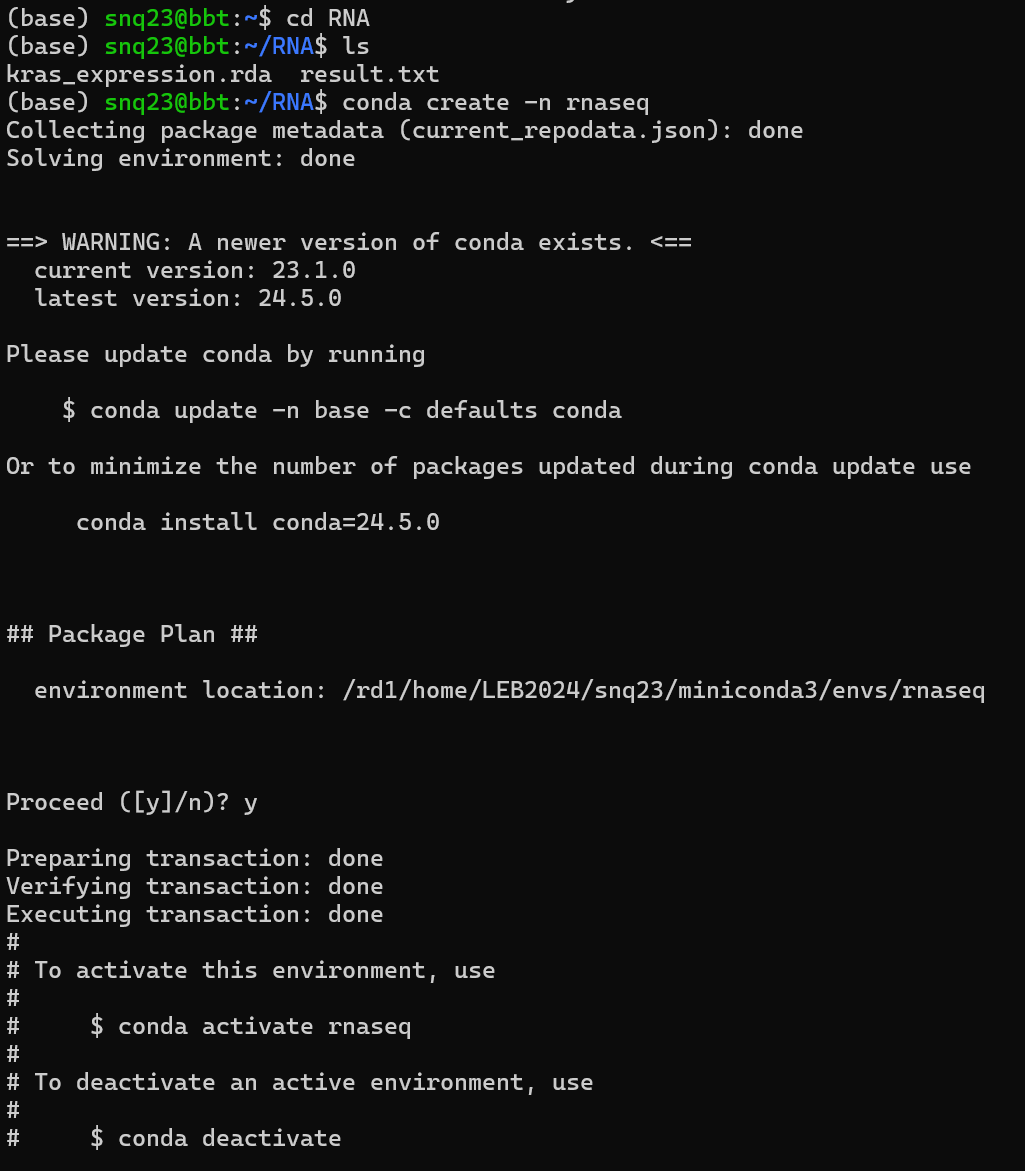
结构基因组学测定的信息不足以解释生物学功能，功能基因组学要求对于基因组所真实发挥的功能进行 描绘。根据中心法则，由基因型到表型中的第一环便为转录组数据。测定转录组数据有助于了解基因表 达谱、推断生物学功能。

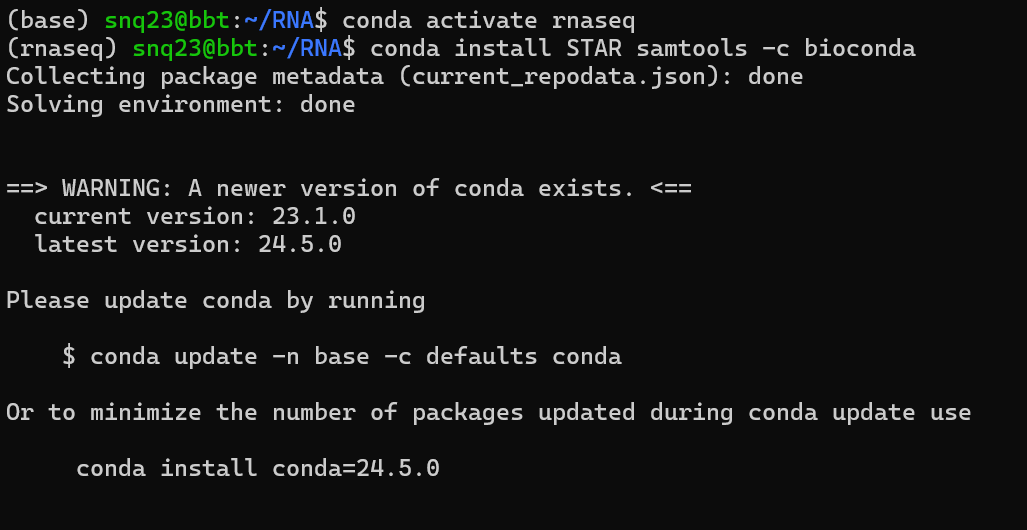
以RNA-seq为主要手段的转录组分析通常包含以下步骤

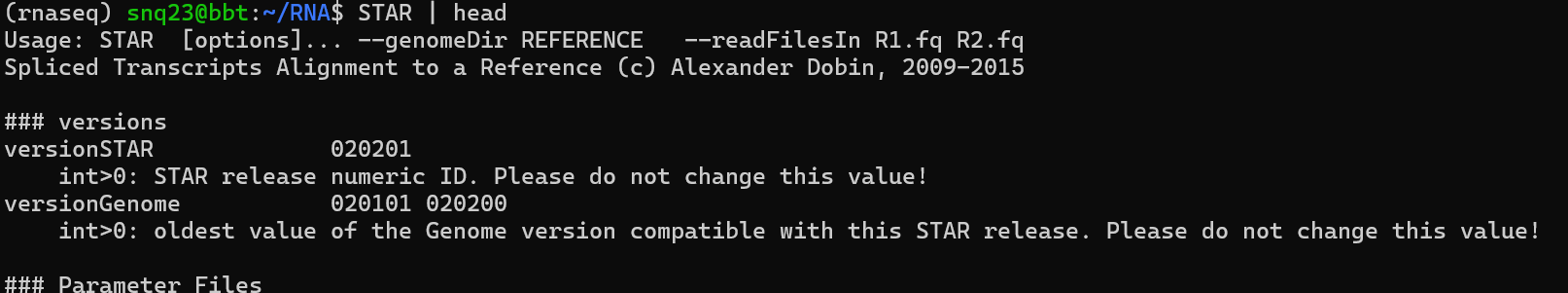
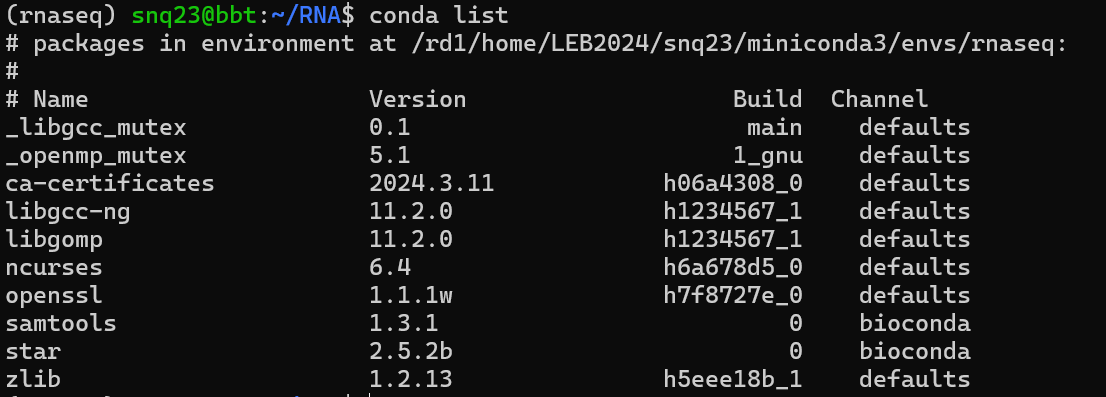


## STAR基础

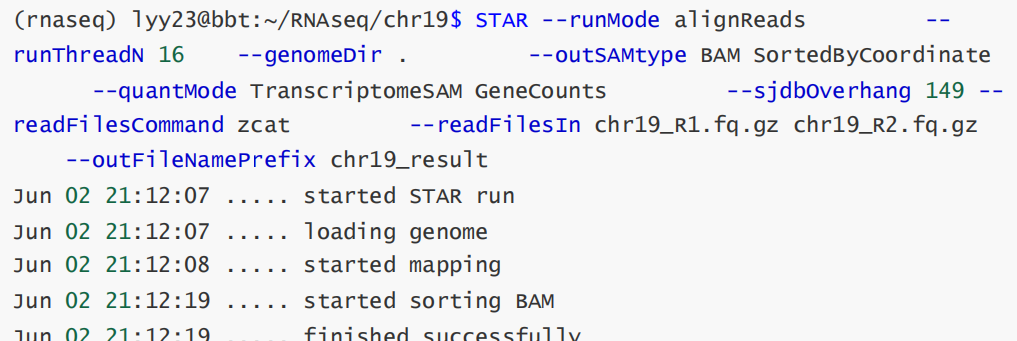
### 新建环境 rnaseq，安装STAR及samtools



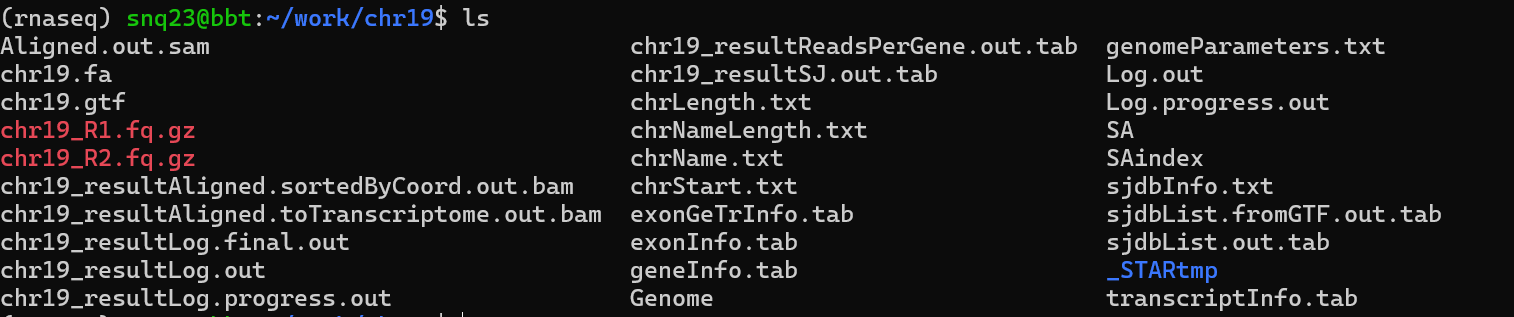




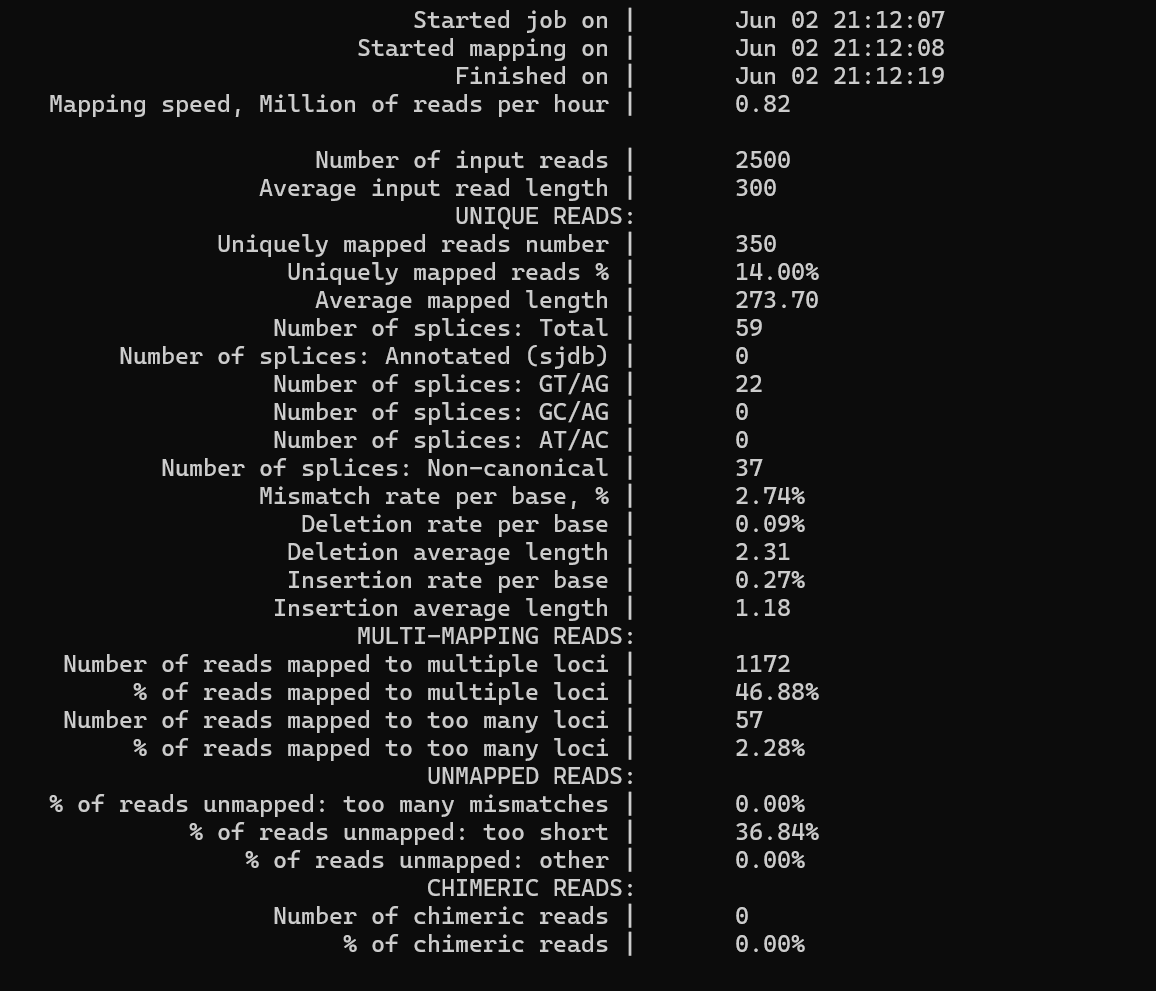
### 导入基因组数据并mapping：



来自组员廖一岩

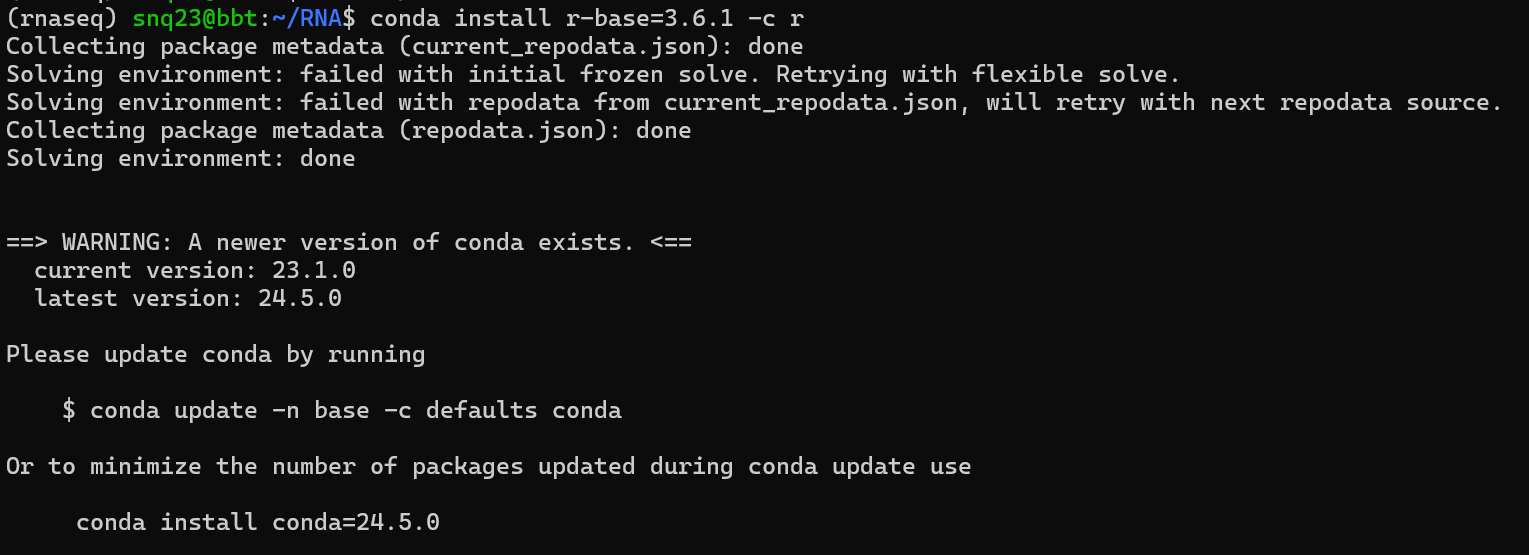


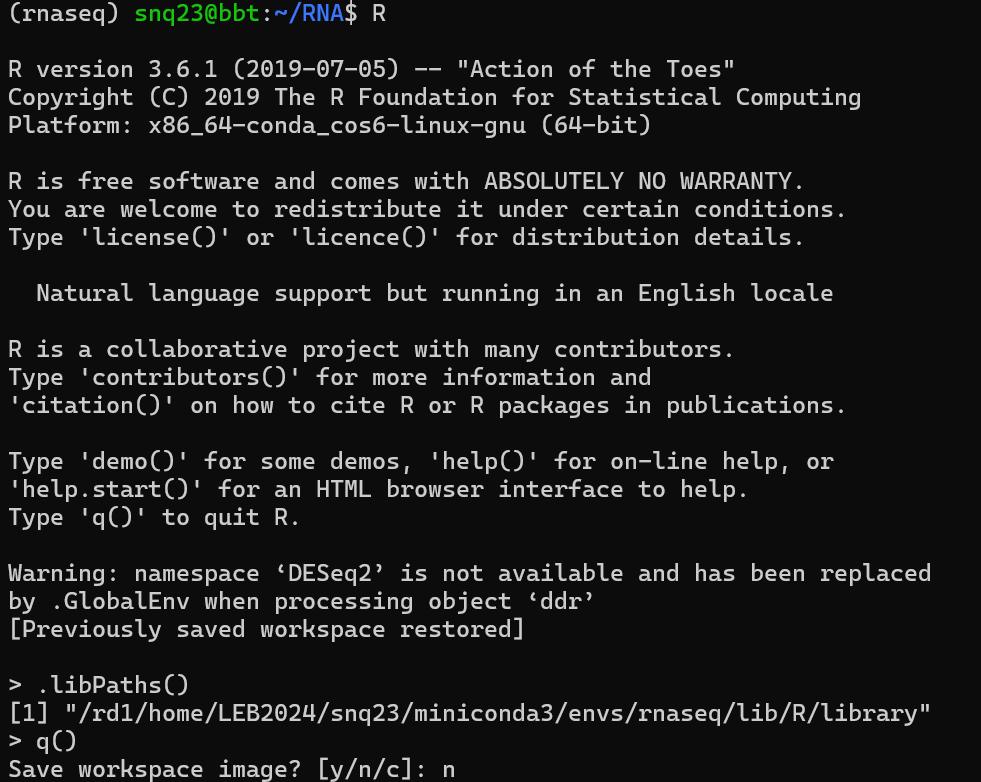


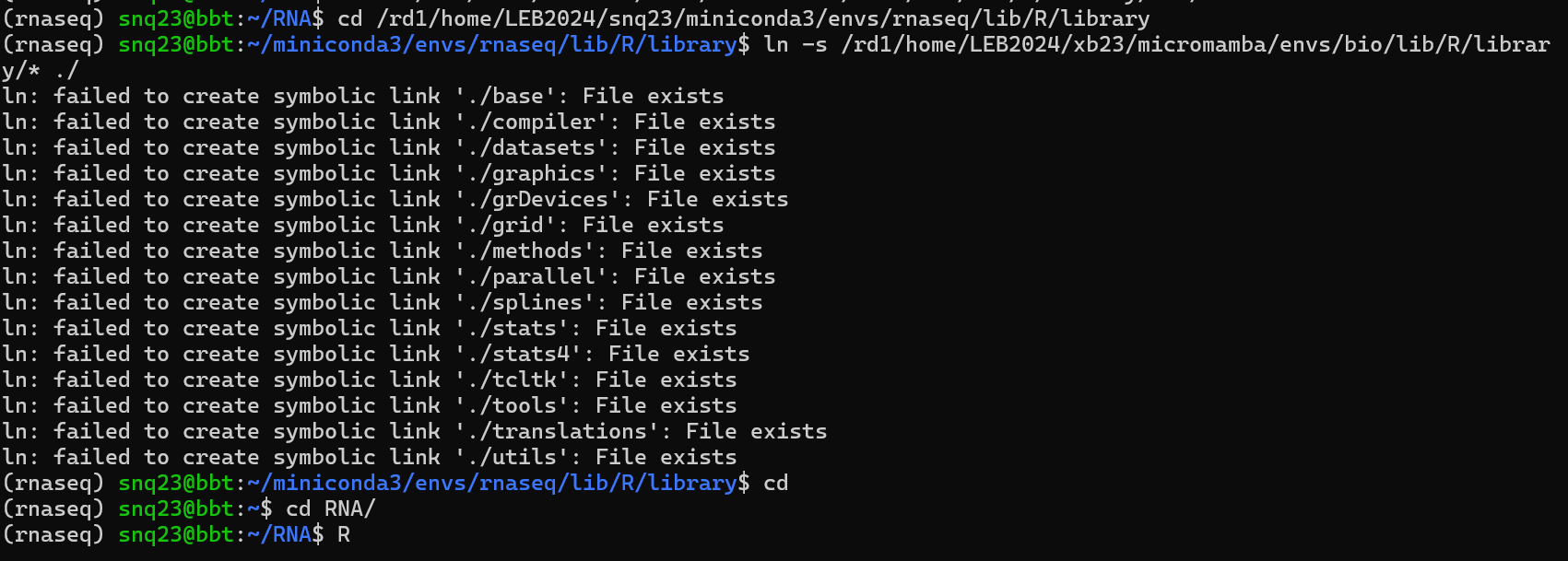


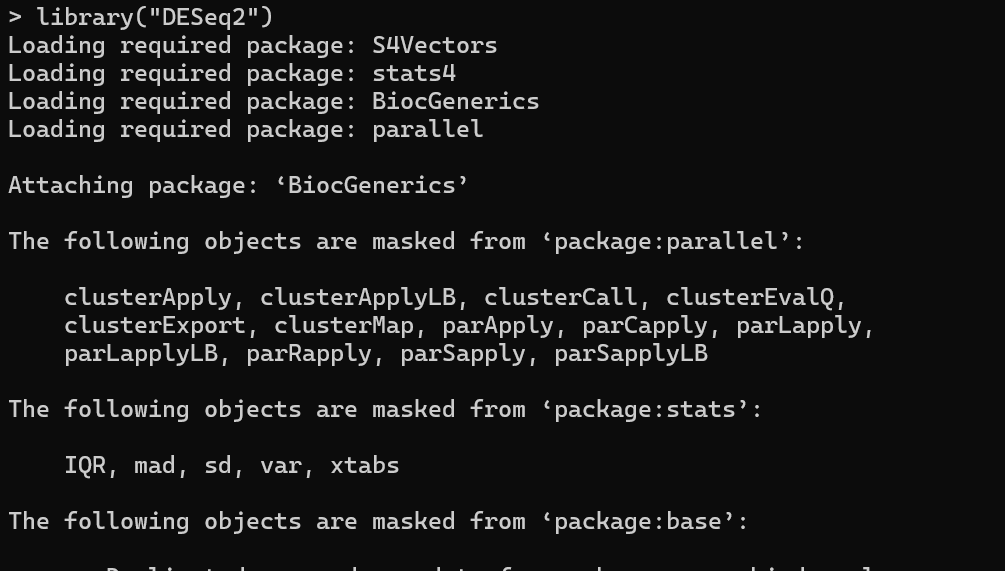
## DESeq2基础

### 安装R及DESeq2包

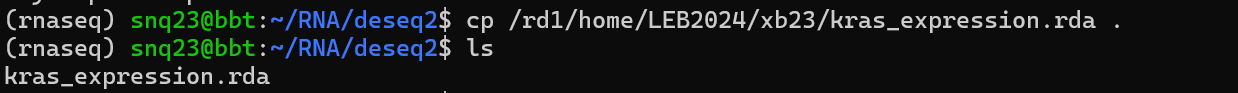


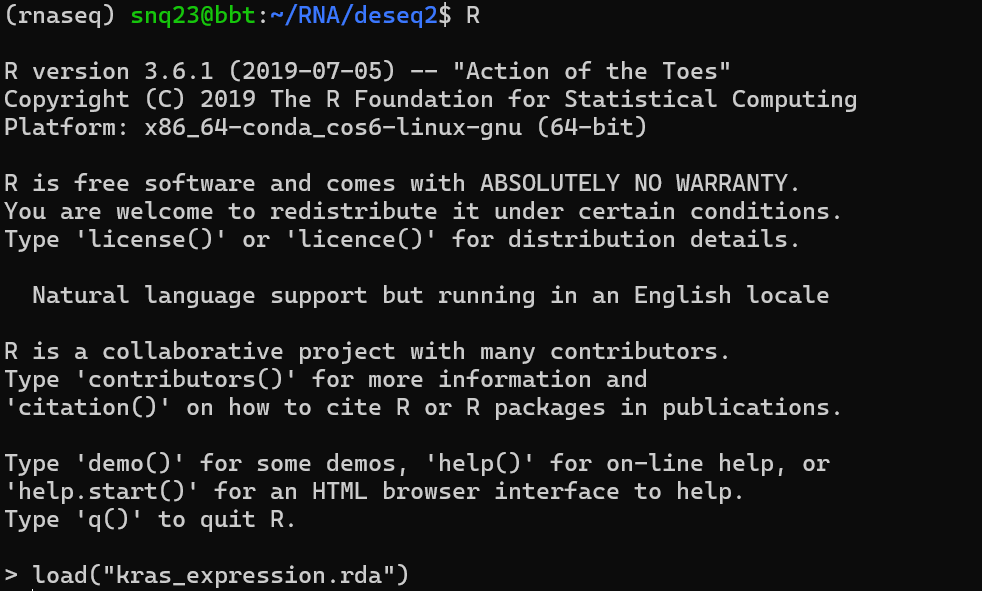


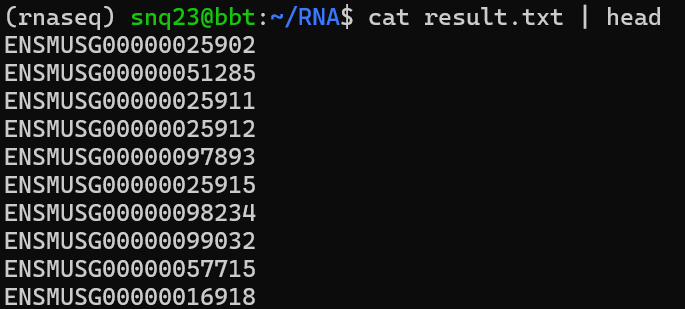




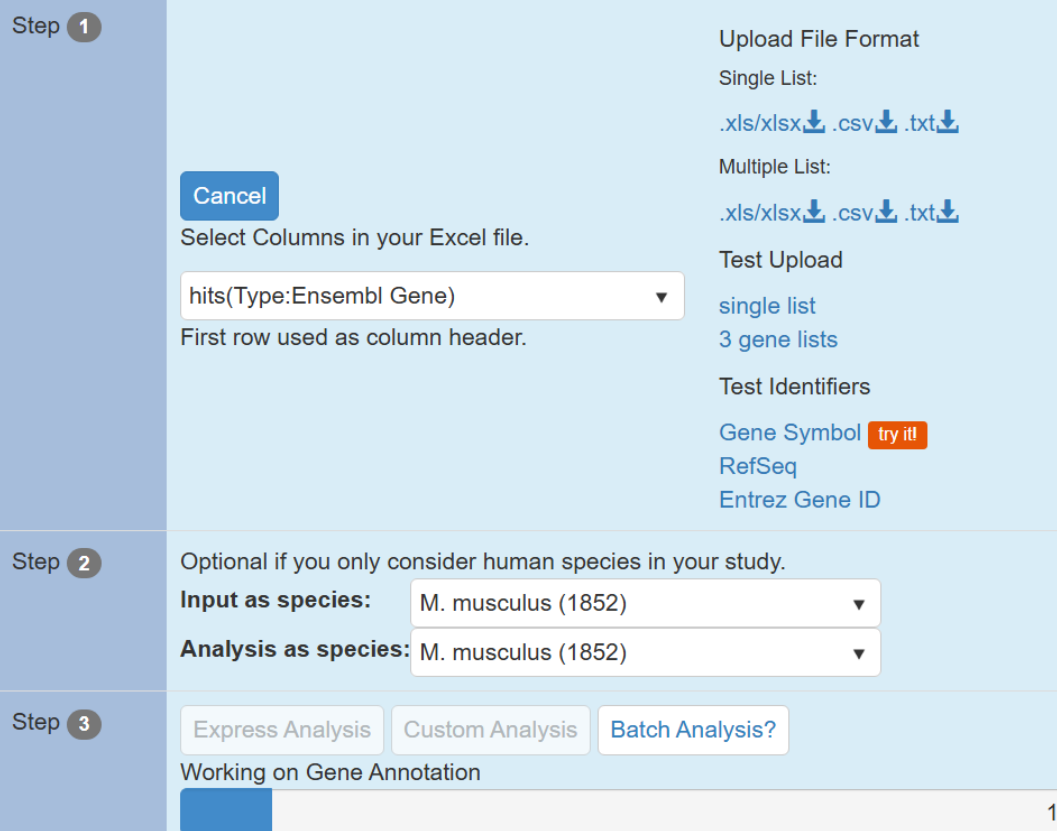
### 进行DESeq2差异分析





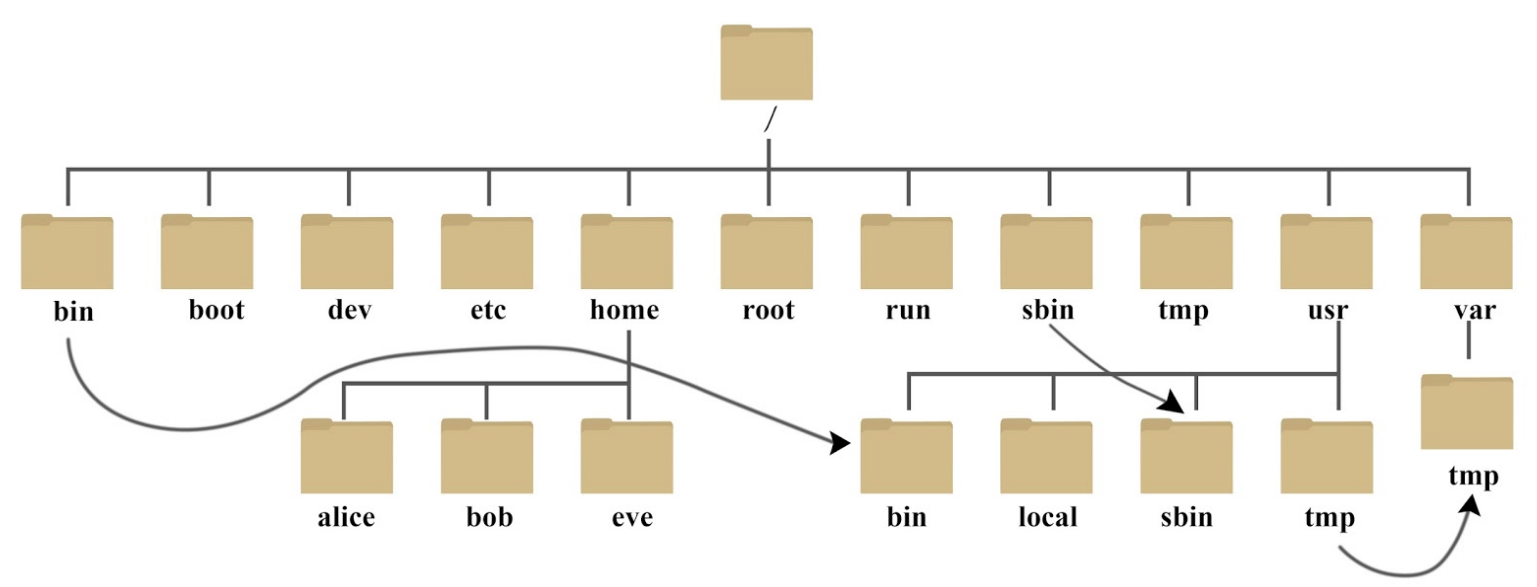


### 富集分析



# Linux服务器管理

## Linux系统目录结构



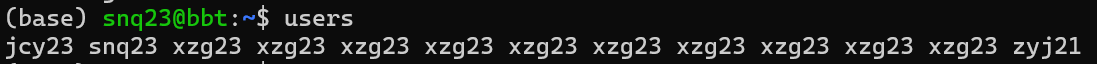
## Linux系统管理命令

[所有系统管理员都应该知道的 25 个基本 Linux 命令 (linux-console.net)](https://cn.linux-console.net/?p=16785)

 uptime 工具，它允许您检查系统已经运行了多长时间，并查看在给定时间有多少用户登录

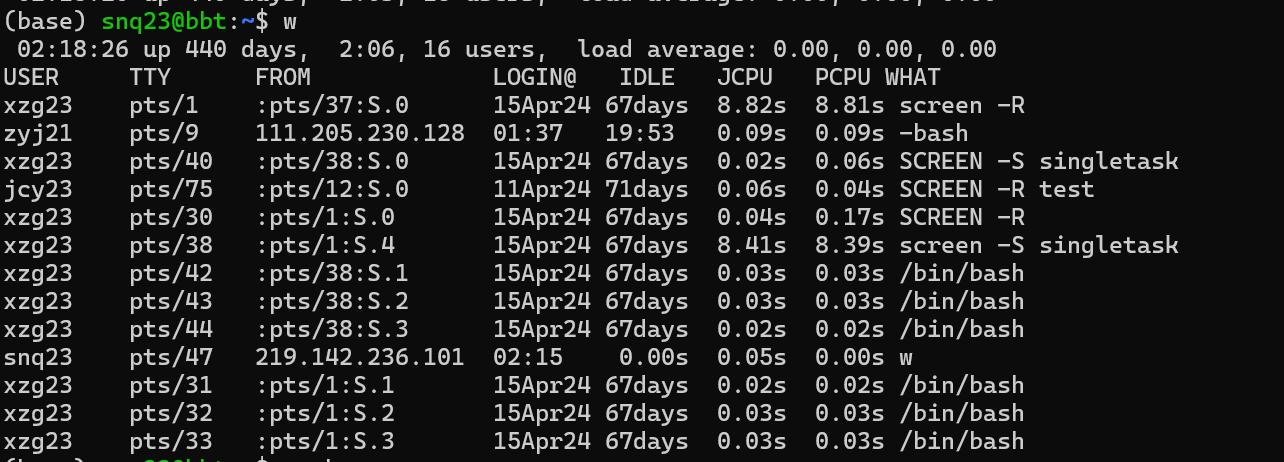


users 命令将列出当前登录的所有用户。

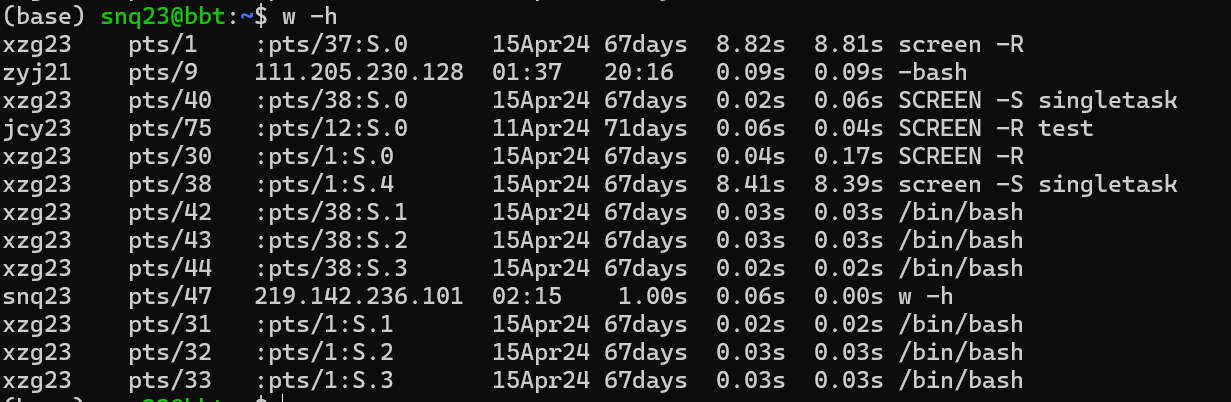


所以其实并没有16个同学在线，而是xzg同学登录好多次

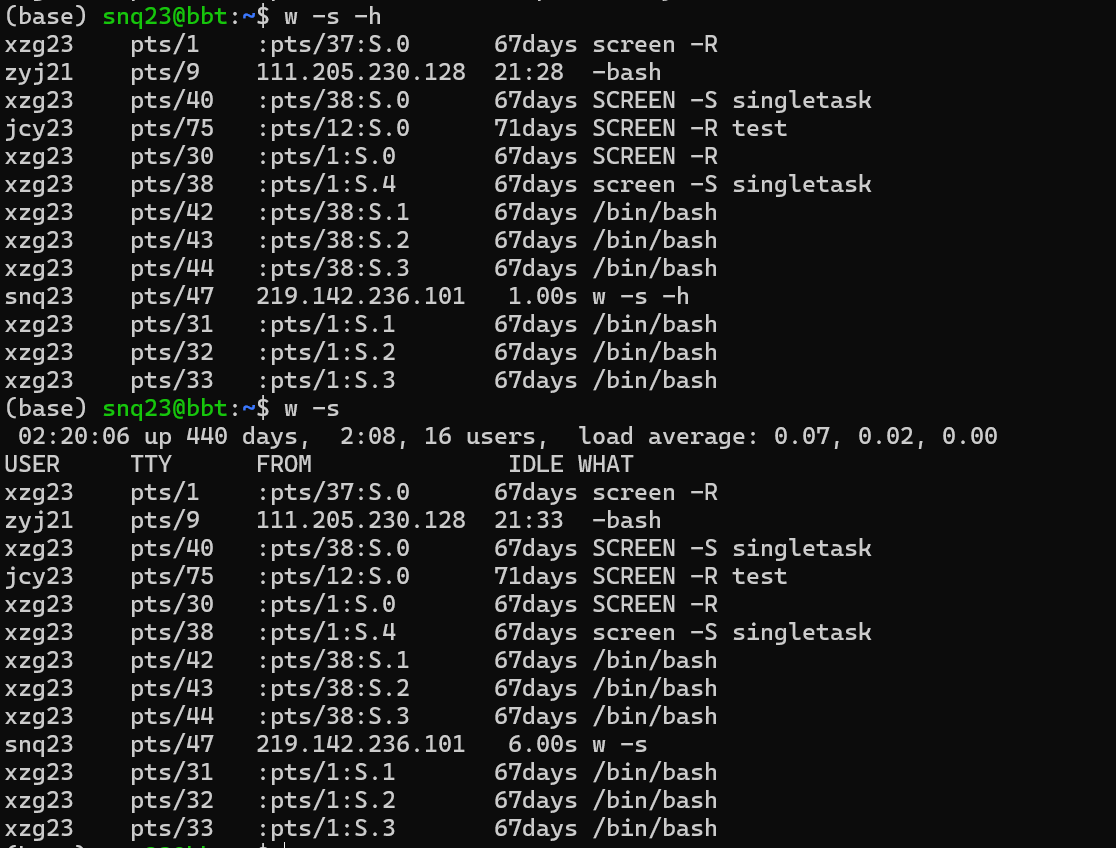
w 命令将显示当前登录的用户，以及每个用户的进程和平均负载。该命令还报告登录名、登录时间、tty 名称、JCPU、PCPU 和命令



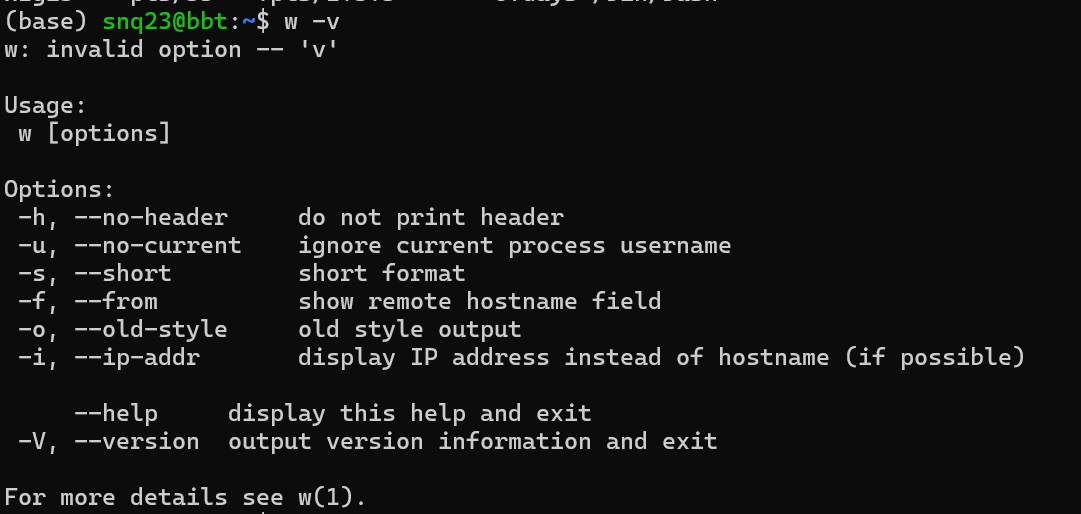
-h 选项将显示没有任何标题条目的输出。



-s 标志将从输出中排除 JCPU 和 PCPU

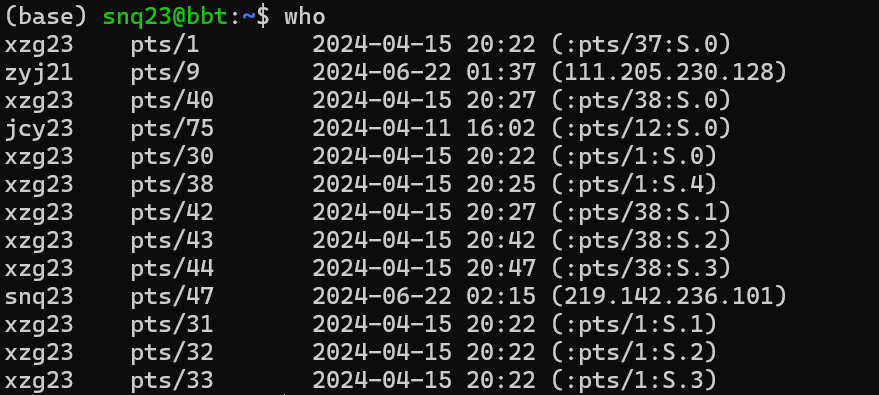


使用-V标志检查工具版本。



与who命令对比

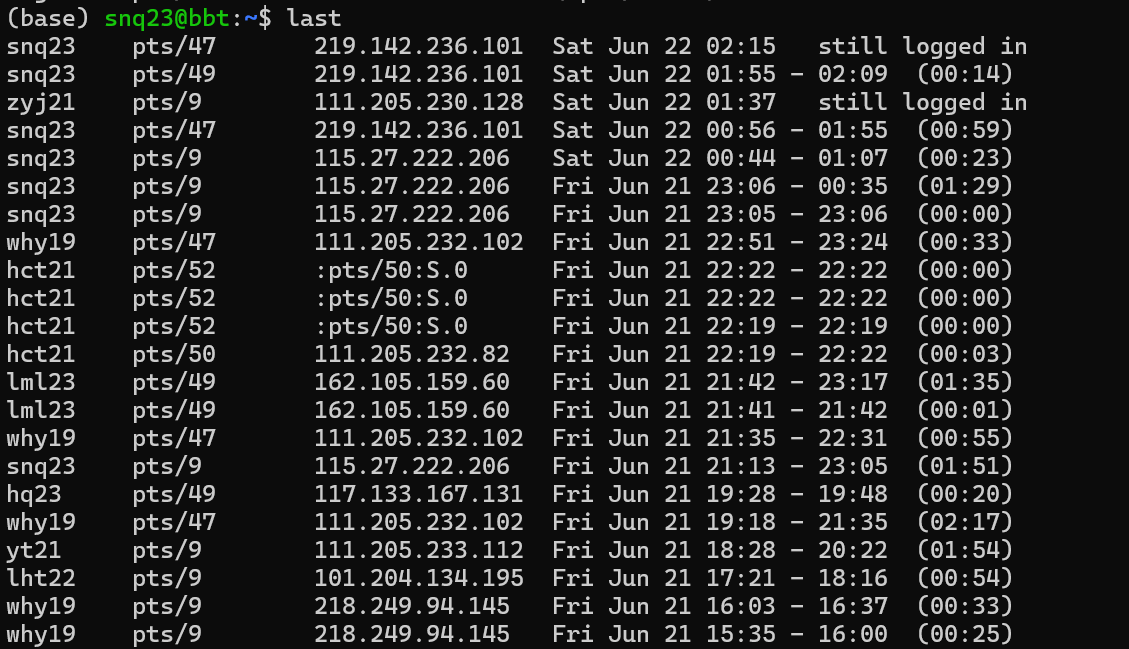
who 命令将返回当前用户的姓名、日期、时间和主机信息。但是，与 w 命令不同，该命令不会打印用户正在执行的操作。



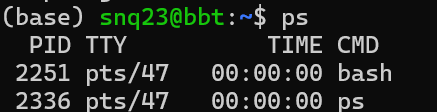
grep 工具是另一个强大且流行的命令。该工具允许您在给定文件中快速搜索包含与给定单词或字符串匹配的行。 grep 工具可以对文件和管道内容执行搜索。

find 命令是一个在目标目录下执行快速搜索的便捷工具。不过与grep不同的是，find会根据文件名进行搜索。

last 命令返回有关系统中用户活动的有价值的信息

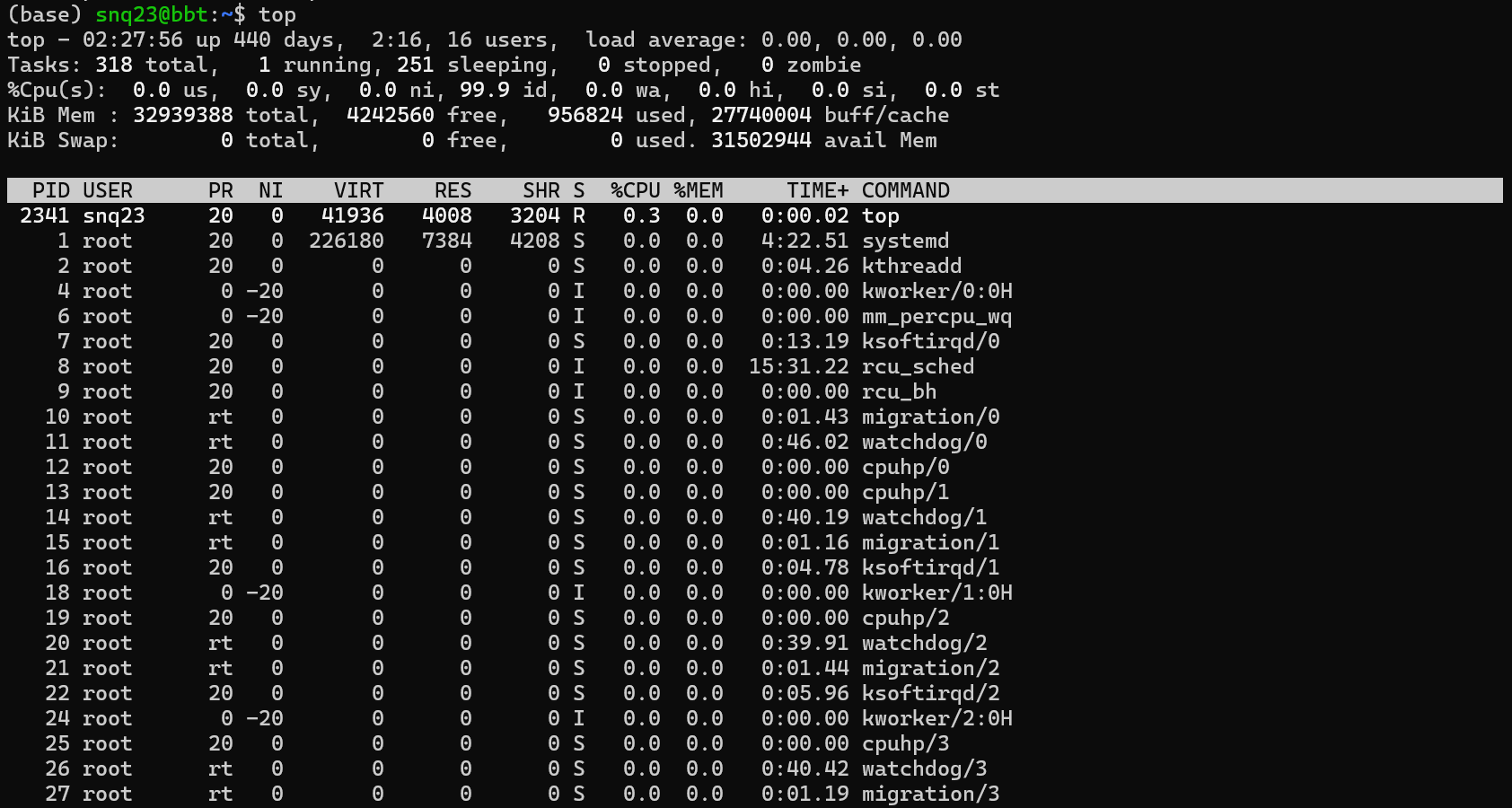


ps 工具是用于查看系统上当前正在运行的进程的最基本命令之一

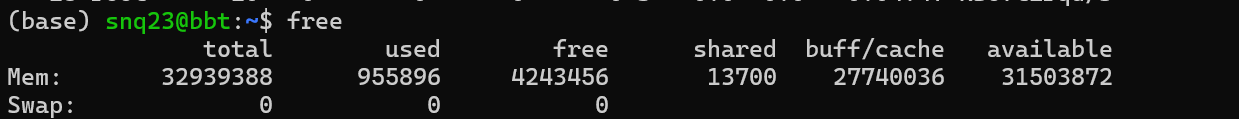


Kill 命令通常用于终止进程

top命令是实时监控系统资源和运行进程的强大工具。



free命令对于检查系统资源非常有用，例如物理内存和交换内存



service命令

启动服务：sudo service <service\_name> start

重新启动服务：sudo service <service\_name> restart

停止正在运行的服务：sudo service <service\_name> stop

检查服务的状态;sudo service <service\_name> status

# ABC网站

ABC全名为Applied Bioinformatics Course，中文翻译为实用生物信息技术

[abc (gao-lab.org)](https://abc.gao-lab.org/)

## 网站内容

### People

记录生物信息学名人，专家，本课程的合作老师以及助教

### Docs

各种生物信息电子学习材料，包括文章，教程，课程等

### Databases

各种数据库的网页链接（可以找HBA序列了）

### Tools

各种各样的在线生物信息学工具

### About

课程介绍

### Projects

课程相关的，可实操的小项目

### Practices

课程涉及工具的小练习

### Talks&Seminars

各种课件