Experiment Guide of Analysis of genomic data

—基因组数据分析教学实验课

毛凌峰

2018.09.27

课程安排

- 生产工具（Windows or Linux?）
- 实战操作一 认识Linux
- 实战操作二 Linux环境下生信软件安装与使用（以BLAST为例）

生产工具（Windows or Linux?）

配置　　可操作性　　编写环境/运行环境

Software	Linux	Windows	MacOS
Blast	√	√	√
MEGA	√	√	√
MUMmer	√		√
Bowtie2	√		√
Soapdenovo2	√		
Canu	√		

Windows or Linux?

操作系统	Windows 10 专业版 64位 (DirectX 12)
处理器	英特尔 Core i7-6700 @ 3.40GHz 四核
内存	8 GB (海力士 DDR4 2133MHz)
主硬盘	东芝 DT01ACA100 (1 TB / 7200 转/分)

~7000RMB

操作系统	Centos 6.7 64
处理器	4颗 AMD Opteron(TM) Processor 6274 2.2GHz 8核 ~64个虚拟核心
内存	1.092Tb
主硬盘	56Tb

~160000RMB

Windows or Linux?

About The Cover

COVER Photograph of the Honghe Hani rice terraces in Yunnan Province, China. In this issue, two separate research groups report draft sequences of two strains of rice--*japonica* and *indica*. In addition, the Editorial, News Focus, Letters, and Perspectives highlight the significance of the rice genome to the world's population. 79, 92 [Image: Liwen Ma and Baoxing Qiu, Beijing Genomics Institute]

水稻基因组大小：~400Mb

利用~200x 基因组二代测序数据（SoapDonovo2, 60 threads）进行初步拼接的时间：~2000 CPU hours, 33.3 h
所需内存：~200Gb

利用~100x基因组三代测序数据（Canu, 60 threads）进行初步拼接的时间为：~3000 CPU hours, 50 h
所需内存：~150Gb

~15x基因组二代重测序数据(Bowtie2, 8 thread)比对到参考基因组上的时间为：~20 CPU hours 2.5 h

Genomic variation associated with local adaptation of weedy rice during de-domestication. 2017 *Nature communication*

~18.2x 155 weedy and 76 locally cultivated rice accessions 231*20=4620 CPU hours

实战操作一 认识Linux

■ Linux 起源

Linux、Mac OS、Windows与Unix都有一些联系，这其中有一些传奇的故事。

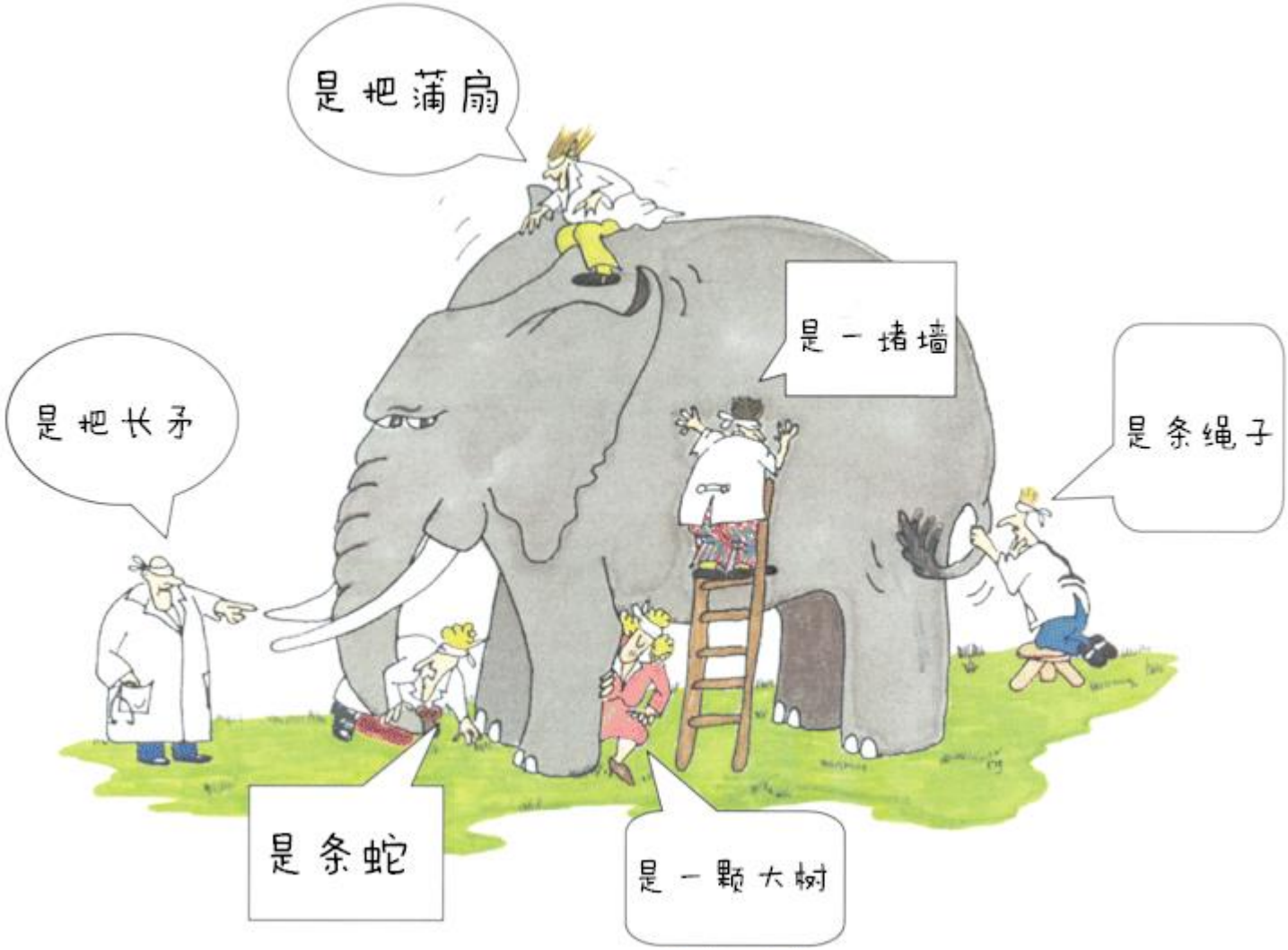
Linux起源于操作系统课程就没了操作系功能，但是比较mini，当。

Minix过于短统，但是Andrew认为自就包括一个叫**Linus**的同学

Linus同学有自己写成了一个“千疮百孔是Linux的诞生。

实战操作一 认识Linux

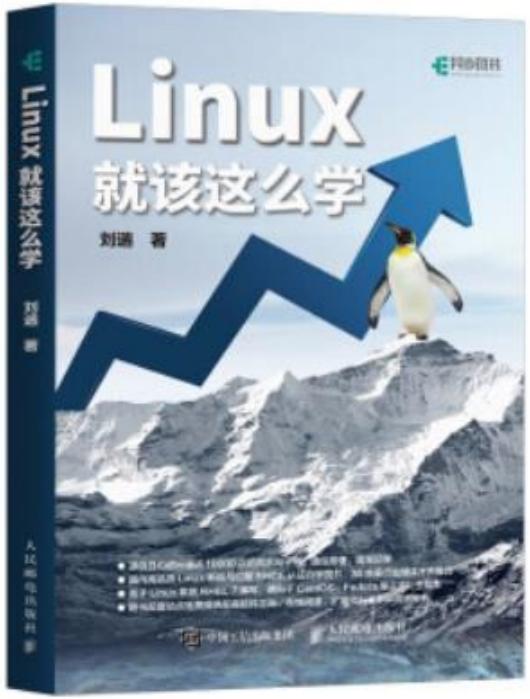
■ L

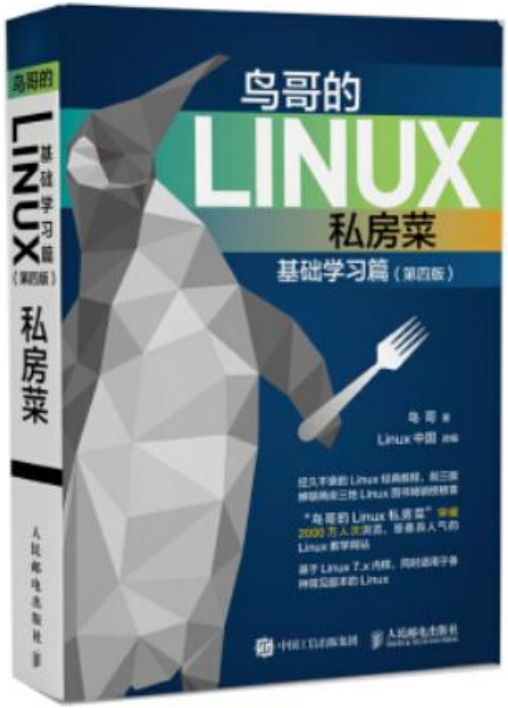


刘络）

实战操作一 认识Linux

■Linux 学习推荐书籍



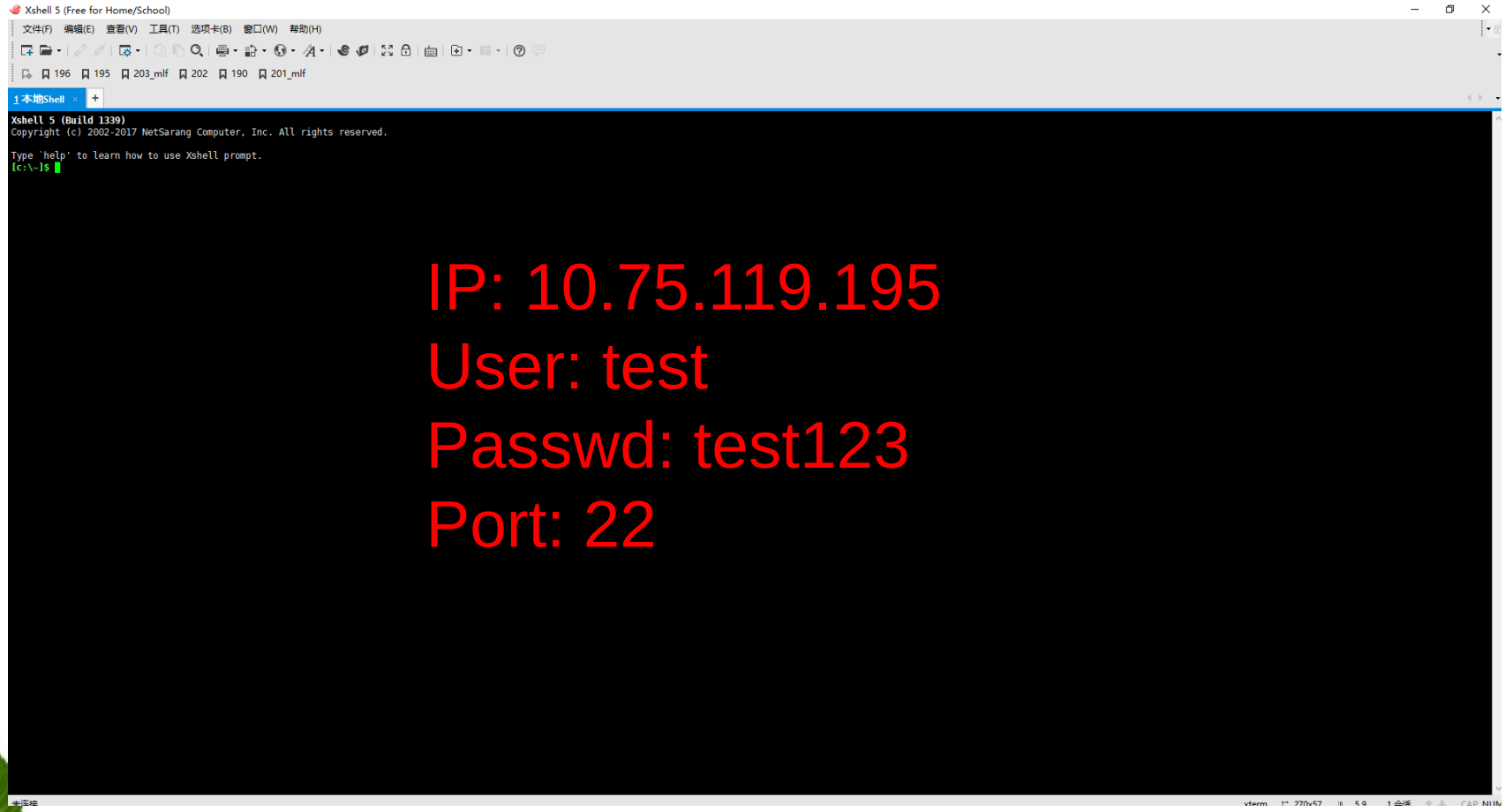


实战操作一 认识Linux

- 命令行登录工具-putty

实战操作一 认识Linux

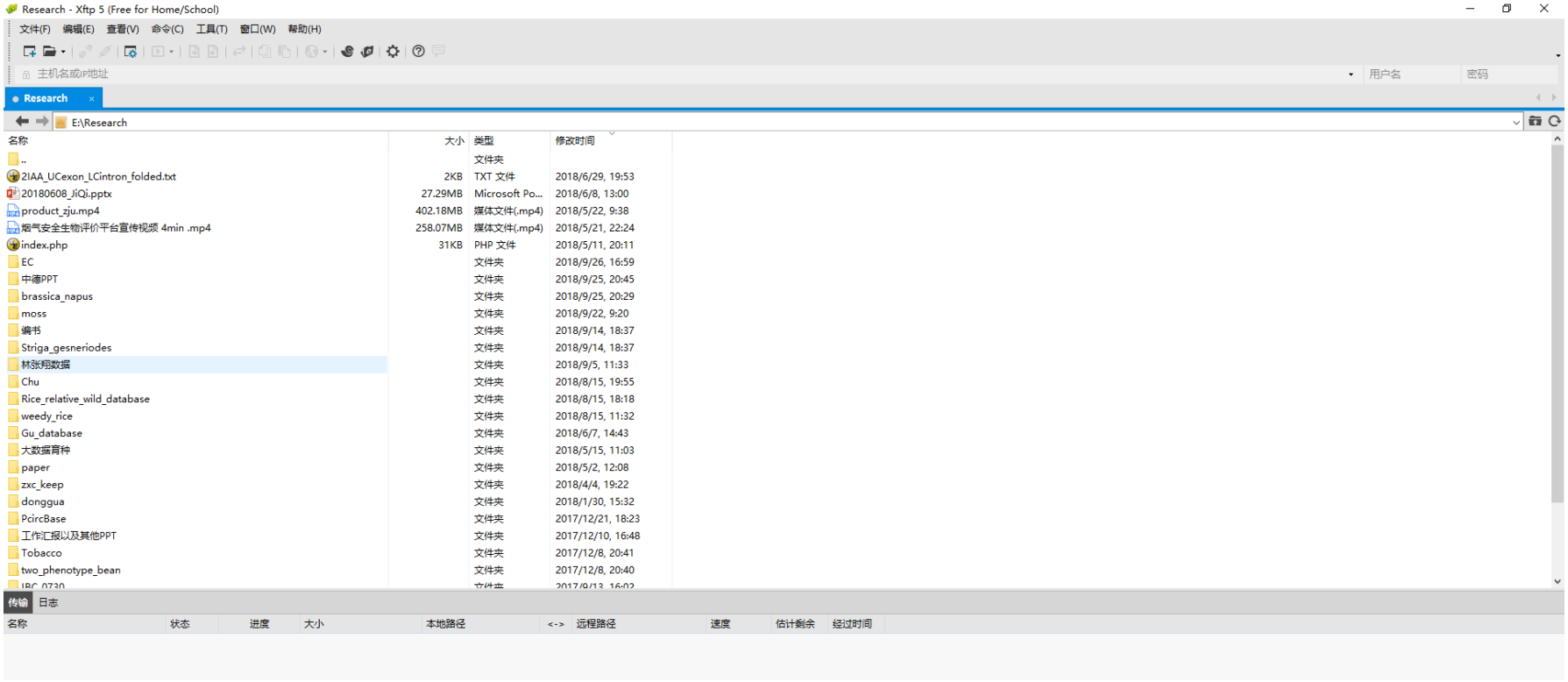
- 命令行登录工具-Xshell



https://www.netsarang.com/download/down_form.html?code=623

实战操作一 认识Linux

- 文件传输工具-Xftp



https://www.netsarang.com/download/main.html

实战操作一 认识Linux

- 认识Linux服务器的属性

一、服务器有多少核心

```
[root@localhost ~]# cat /proc/cpuinfo|grep "processor" |wc -l
32
[root@localhost ~]# cat /proc/cpuinfo|grep "physical id" |sort|uniq|wc -l
4
[root@localhost ~]# cat /proc/cpuinfo|grep "cpu cores" |uniq
cpu cores       : 8
```

二、服务器有多少内存

```
[root@localhost ~]# free -g
              total        used        free      shared  buff/cache   available
Mem:            251           2         248           0           0         248
Swap:             3           0           3
[root@localhost ~]# free
              total        used        free      shared  buff/cache   available
Mem:      263968764     2154900   260854440       10736      959424   260919300
Swap:       4194300           0     4194300
[root@localhost ~]# free -g
              total        used        free      shared  buff/cache   available
Mem:            251           2         248           0           0         248
Swap:             3           0           3
[root@localhost ~]# free -h
              total        used        free      shared  buff/cache   available
Mem:           251G        2.1G        248G        10M        936M        248G
Swap:          4.0G          0B        4.0G
```

实战操作一 认识Linux

三、服务器有多少存储

```
[root@localhost ~]# lsblk
NAME        MAJ:MIN RM  SIZE RO TYPE MOUNTPOINT
sda           8:0    0  1.8T  0 disk
├─sda1        8:1    0    1G  0 part /boot
└─sda2        8:2    0  1.8T  0 part
  ├─cl-root 253:0    0   50G  0 lvm  /
  ├─cl-swap 253:1    0    4G  0 lvm  [SWAP]
  └─cl-home 253:2    0  1.8T  0 lvm  /home
[root@localhost ~]# df
Filesystem           1K-blocks      Used  Available Use% Mounted on
/dev/mapper/cl-root   52403200  21834620   30568580  42% /
devtmpfs             131934768         0  131934768   0% /dev
tmpfs                131984380       144  131984236   1% /dev/shm
tmpfs                131984380      9304  131975076   1% /run
tmpfs                131984380         0  131984380   0% /sys/fs/cgroup
/dev/sda1              1038336    278908     759428  27% /boot
/dev/mapper/cl-home 1894514876 835464800 1059050076  45% /home
tmpfs                 26396880        16   26396864   1% /run/user/0
[root@localhost ~]# df -h
Filesystem           Size  Used Avail Use% Mounted on
/dev/mapper/cl-root   50G   21G   30G  42% /
devtmpfs             126G     0  126G   0% /dev
tmpfs                126G  144K  126G   1% /dev/shm
tmpfs                126G  9.1M  126G   1% /run
tmpfs                126G     0  126G   0% /sys/fs/cgroup
/dev/sda1           1014M  273M  742M  27% /boot
/dev/mapper/cl-home  1.8T  797G 1010G  45% /home
tmpfs                 26G   16K   26G   1% /run/user/0
```

四、服务器的操作系统信息

```
[root@localhost ~]# cat /etc/centos-release
CentOS Linux release 7.3.1611 (Core)
[root@localhost ~]# uname -r
3.10.0-862.11.6.el7.x86_64
[root@localhost ~]# getconf LONG_BIT
64
[root@localhost ~]#
```

实战操作一 认识Linux

- 常规Linux操作命令

显示当前所在文件夹路径：pwd
显示当前所在文件夹下文件及文件夹：ls
显示当前文件夹大小：du
切换文件夹：cd 文件夹路径
创建文件夹：mkdir 文件夹名
删除文件夹：rmdir 文件夹名
查看文本文件内容: cat,less,more 文件名
编辑配置文件：vi 配置文件名
移动或者重命名文件： mv 初始文件名 目标文件名
验证文件的文件的完整性： md5sum 文件名
查看命令历史记录： history
创建新文件：touch

Linux环境下生信软件安装与使用

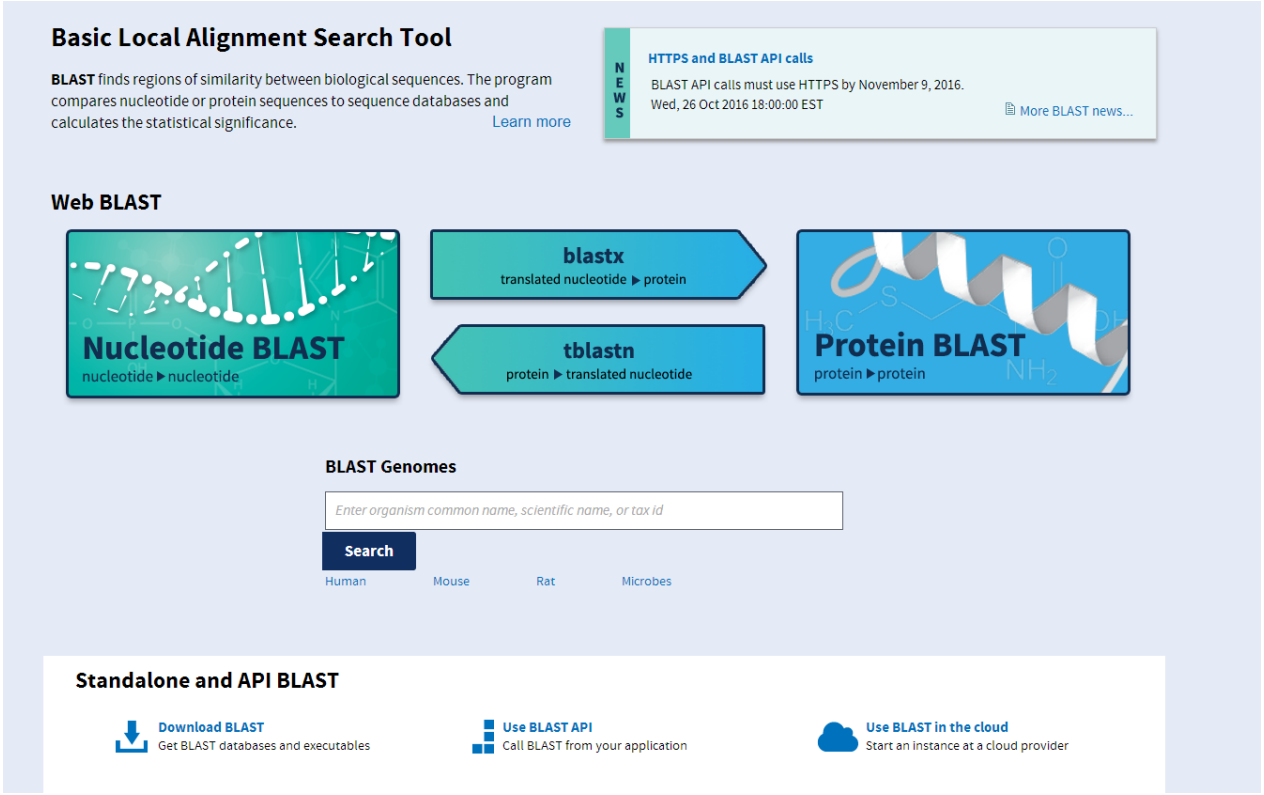
- 生物信息学中使用软件三大问题（Linux）

一、软件怎么安装，怎么配置？

二、软件怎么用，参数如何设置？

三、软件的结果怎么看？

实战操作二 Linux环境下生信软件安装与使用（以BLAST为例）



https://blast.ncbi.nlm.nih.gov/Blast.cgi

BLAST

Basic Local Alignment Search Tool.

Finds regions of local similarity between biological sequences.

11

BLAST Link (BLink)

BLAST Microbial Genomes

BLAST RefSeqGene

COBALT

Conserved Domain Search Service (CD Search)

Electronic PCR (e-PCR)

Gene Expression Omnibus (GEO) BLAST

Genome BLAST

Open Reading Frame Finder (ORF Finder)

Primer-BLAST

实战操作二 Linux环境下生信软件安装与使用（以BLAST为例）

- 下载安装Blast+

```
[root@201 Alignment]# wget https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ncbi-blast-2.7.1+-x64-linux.tar.gz
[root@201 Alignment]# tar zxvf ncbi-blast-2.7.1+-x64-linux.tar.gz && cd ncbi-blast-2.7.1+
[root@201 ncbi-blast-2.7.1+]# vi /etc/profile.d/apps_blast+271.sh
BLAST271=/PATH/ncbi-blast-2.7.1+/bin
export PATH=$PATH:BLAST271
##个人用户
[mlf@201 ncbi-blast-2.7.1+]# vi ~/.bashrc
BLAST271=/PATH/ncbi-blast-2.7.1+/bin
export PATH=$PATH:BLAST271
[mlf@201 ncbi-blast-2.7.1+]# source ~/.bashrc
[root@201 ncbi-blast-2.7.1+]# source /etc/profile.d/apps_blast+271.sh
[root@201 ncbi-blast-2.7.1+]# blastn -version
blastn: 2.7.1+
Package: blast 2.7.1, build Dec 10 2013 14:41:40
```

实战操作二 Linux环境下生信软件安装与使用（以BLAST为例）

- 运行BLAST+

```
格式化数据库
makeblastdb -in db.fasta -dbtype prot -parse_seqids -out dbname
参数说明:
-in: 待格式化的序列文件vsrgds
-dbtype: 数据库类型，prot或nucl
-out: 数据库名

蛋白序列比对蛋白数据库（blastp）
blastp -query seq.fasta -out seq.blast -db dbname -outfmt 6 -evalue 1e-5 -num_descriptions 10 -num_threads 8

程序选择
blastp: 蛋白序列与蛋白库做比对。
blastx: 核酸序列对蛋白库的比对。
blastn: 核酸序列对核酸库的比对。
tblastn: 蛋白序列对核酸库的比对。
tblastx: 核酸序列对核酸库在蛋白级别的比对
```

实战操作二 Linux环境下生信软件安装与使用（以BLAST为例）

■ BLAST+参数

```
-query:  输入文件路径及文件名
-out: 输出文件路径及文件名
-db: 格式化了的数据库路径及数据库名
-evalue: 设置输出结果的e-value值，(数学)期望值(Expectation value)，E值是个统计阈值，缺省值10，意指比对结果中由于随机偶然性产生的匹配结果不大于10,E值越小结果越可靠。
-num_alignments <Integer, >=0> : 显示结果数量
   Number of database sequences to show alignments for
   Default = `250'
-num_threads: 线程数
-outfmt: 输出文件格式，总共有12种格式，常用是6，tabular分隔
-outfmt <String>
   alignment view options:
     0 = pairwise,
     1 = query-anchored showing identities,
     2 = query-anchored no identities,
     3 = flat query-anchored, show identities,
     4 = flat query-anchored, no identities,
     5 = XML Blast output,
     6 = tabular,
     7 = tabular with comment lines,
```

实战操作二 Linux环境下生信软件安装与使用（以BLAST为例）

- BLAST+结果查看

输出结果格式以格式6 为说明
从左到右分别是

query名	subject名	identity	比对长度	错配数	空位数	query比对起始坐标	query比对终止坐标	subject比对起始坐标	subject比对终止坐标	期望值	比对得分
query1	sub24	91.11	45	3	1	198	241	502208	502252	2.70E-06	50.05
query1	sub21	98.68	151	2	0	532	682	1360665	1360515	1.00E-76	284
query1	sub21	86.17	94	12	1	198	290	479232	479139	4.80E-14	75.82
query1	sub21	87.04	54	7	0	238	291	1297867	1297920	6.90E-07	52.03
query2	sub21	99.44	892	3	2	28	918	1351055	1350165	0	1713.2
query2	sub21	87.58	153	17	1	343	495	1358110	1357960	2.10E-35	147.2
query2	sub21	84.11	107	16	1	699	805	1305723	1305618	4.00E-12	69.88
query2	sub21	89.58	48	5	0	519	566	1305968	1305921	6.00E-08	56
query2	sub14	88.24	153	16	1	343	495	145402	145252	8.70E-38	155.1
query2	sub24	88.08	151	16	1	345	495	567561	567709	1.40E-36	151.2
query2	sub24	87.8	123	14	1	686	808	563341	563220	1.90E-26	117.5

为什么要选格式6？
6 = tabular