RNA 可变剪接位点预测教程

浙江大学生命科学学院

周煊皓 张泽远 汤筱茵

Version 1.0

2019 年 4 月 8 日

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

目录

1 虚拟机 VMware

1.1 关于 VMware

1.2 VMware 的安装和设置

1.2.1 VMware 和 CentOS 7 的准备

1.2.2 VMware 的安装和设置

1.2.3 虚拟机系统和软件的配置

1.3 对于 Linux 环境的补充

2 从 FASAT 格式文件到 sam 格式文件——HISAT2

2.1 关于 HISAT2

2.2 HISAT2 的安装

2.2.1 HISAT2 的下载和解压缩。

2.2.2 命令路径的修改

2.3 HISAT2 的功能与使用方法

2.3.1 help 功能的使用

2.3.2 利用示例数据构建索引文件

2.3.3 利用示例数据进行 map

3 从 sam 格式文件到 bam 格式文件——samtools

3.1 关于 samtools

3.2 samtools 的安装

3.2.1 samtools 的下载和解压缩

3.2.2 码源包的安装

3.2.3 samtools 的安装

3.2.4 命令路径的修改

3.3 samtools 的功能与使用方法

3.3.1 samtools 功能的使用

3.3.2 将 sam 文件转换为 bam 文件。

4 sam 格式文件/bam 格式文件的分析——cufflinks

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

4.1 关于 cufflinks

4.2 cufflinks 的安装

4.2.1 cufflinks 的下载和解压缩

4.2.2 命令路径的修改

4.3 cufflinks 的功能与使用方法

4.3.1 help 功能的使用

4.3.2 gffread 的功能与使用方法

4.3.3 cufflinks 的功能与使用方法

4.3.4 cuffmerge 的功能与使用方法

4.3.5 cuffdiff 的功能与使用方法

5 sam 格式文件/bam 格式文件的分析——SGseq

5.1 关于 SGseq

5.2 SGseq 的安装

5.2.1 R 的安装

5.2.2 SGSeq 的安装

5.2.3 SGSeq 程序包的补充

5.3 SGseq 的功能与使用方法

6 sam 格式文件/bam 格式文件的分析——DExseq

6.1 关于 DExseq

6.2 DExseq 的安装

6.3 DExseq 的功能与使用方法

6.3.1 初步准备

6.3.2 一般分析流程

6.3.3 可视化

6.3.4 大样本并行分析

6.3.5 一条命令进行标准差异外显子的使用分析

7 参考文献

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

1 虚拟机 VMware

1.1 关于 VMware

VMware 就是虚拟的计算机，是相对于常用的物理计算机而言的。它是由运行于物理计算机上的虚拟机软件模拟出来的，具有完整硬件系统功能的、运行在一个完全隔离环境中的计算机系统。VMware 具有集成性，经济性，移植性，可维护性等特点。

1.2 VMware 的安装和设置

1.2.1 VMware 和 CentOS 7 的准备

1、从网上找到合适版本的 VMware 智能下载器（一般为应用程序格式文件，即可执行文件），并将这个应用程序文件下载下来（如图 1 所示）。参考下载网址：https://www.cr173.com/soft/68480.html。

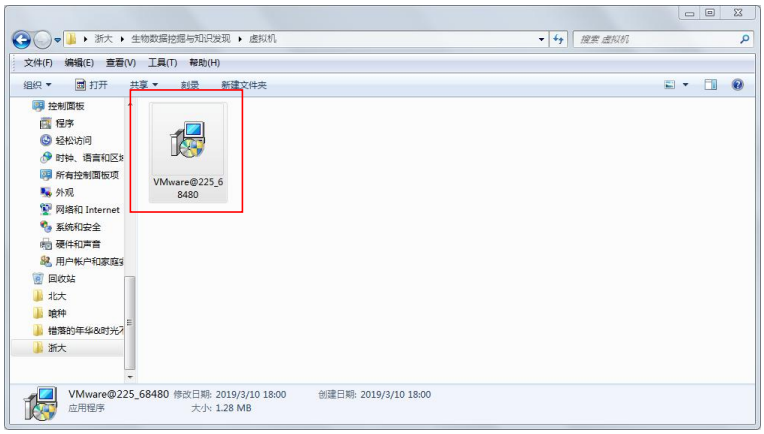


图 1 VMware 智能下载器

2、双击 VMware 智能下载器，自定义安装的位置或者选择安装在系统默认的位置（如图 2 所示），开始 VMware 的下载。最终得到一个压缩包文件（如图 3 所示）。

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn



图 2 VMware 的下载选项

3、解压该 VMware 压缩包，得到另一个 VMware 应用程序文件（如 3 所示），这个 VMware 应用程序文件便是 VMware 的安装程序文件。

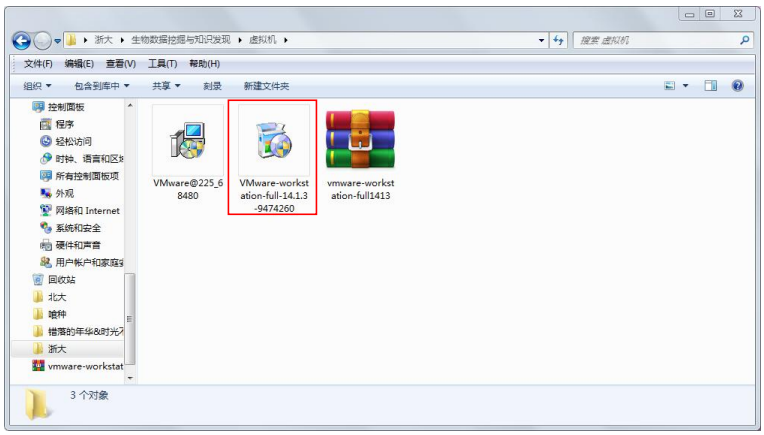


图 3 解压缩得到的 VMware 应用程序文件

4、从网上下载 CentOS 7 的镜像文件的压缩包，解压后即可找到相应的镜像文件。可以根据自己的需要选择下载不同版本或类型的镜像文件。参考下载网址：https://www.centos.org/download/。

1.2.2 VMware 的安装和设置

1、双击该应用程序文件，开始 VMware 的安装。许可协议以后，便可以按照程序的要求和自己的需求勾选相应的选项，完成 VMware 的安装，安装的位置和快捷方式的选择可以选择默认也可以进行更改（如图 4 所示）。

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn



图 4 VMware 的安装

2、安装完成后，双击 VMware 的快捷方式或者启动程序文件，在页面中单机“创建新的虚拟机”，开始 VMware 的硬件和软件配置（如图 5 所示），按照自己的需求选择相应的选项即可。

（1）“希望使用什么类型的配置”选择“自定义（高级）安装”（如图 6 所示）。

（2）“虚拟机硬件兼容性”选择“workstation 14.x”（如图 7 所示）。

（3）“安装客户机操作系统”选择“稍后安装操作系统”（如图 8 所示）。

（4）“选择客户机操作系统”选择“Linux（L）”和“CentOS 7 64 位”（如图 9 所示）。

（5）虚拟机和名称（V）和位置（L）按自己的需求定义（如图 10 所示）。

（6）“处理器配置”选择默认（如图 11 所示）。

（7）“此虚拟机的内存”选择“4700MB”（如图 12 所示）。

（8）“网络类型”选择“使用网络地址转换（NAT）（E）”（如图 13 所示）。

（9）“I/O 控制器类型”选择“LSI Logic（L）”（如图 14 所示）。

（10）“磁盘类型”选择“SCSI（S）”（如图 15 所示）。

（11）“选择磁盘”选择“创建新虚拟磁盘（V）”（如图 16 所示）。

（12）“指定磁盘容量”选择“20.0GB，将虚拟磁盘存储为单个文件（O）”（如图 17 所示）。

（13）“指定磁盘文件”选择默认（如图 18 所示）。

（14）此时，虚拟机的硬件和软件设置已经完毕，点击“完成”即可（如图 19 所

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

示）。

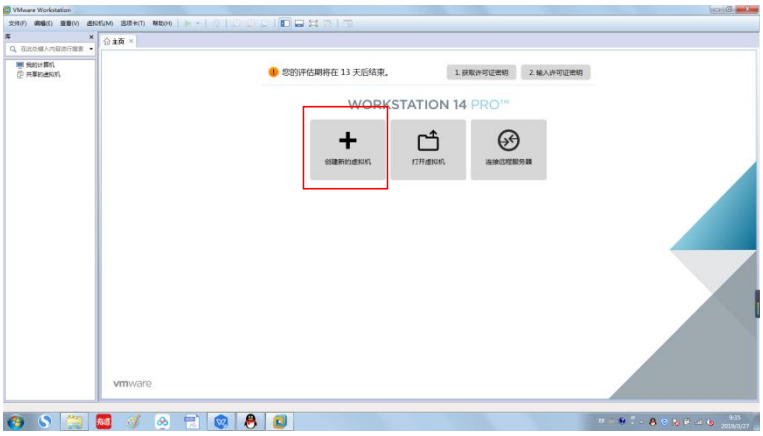


图 5 创建新的虚拟机

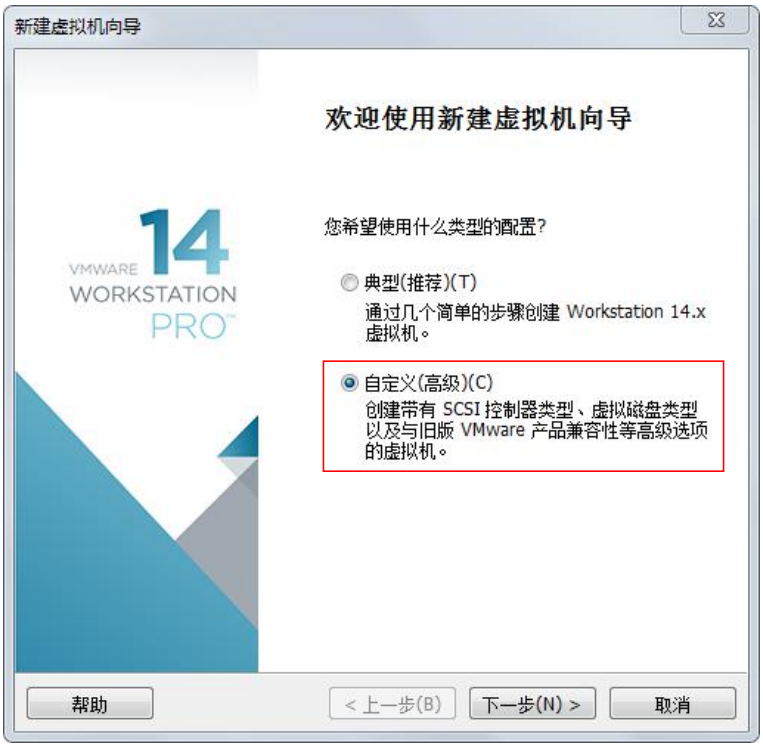


图 6 创建新的虚拟机——选择使用什么类型的配置

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

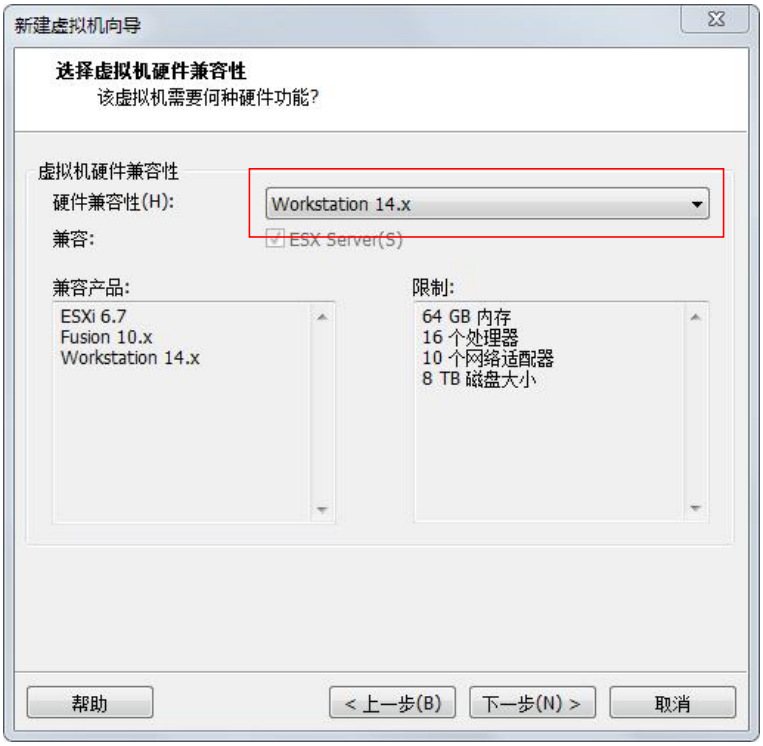


图 7 创建新的虚拟机——选择虚拟机硬件兼容性

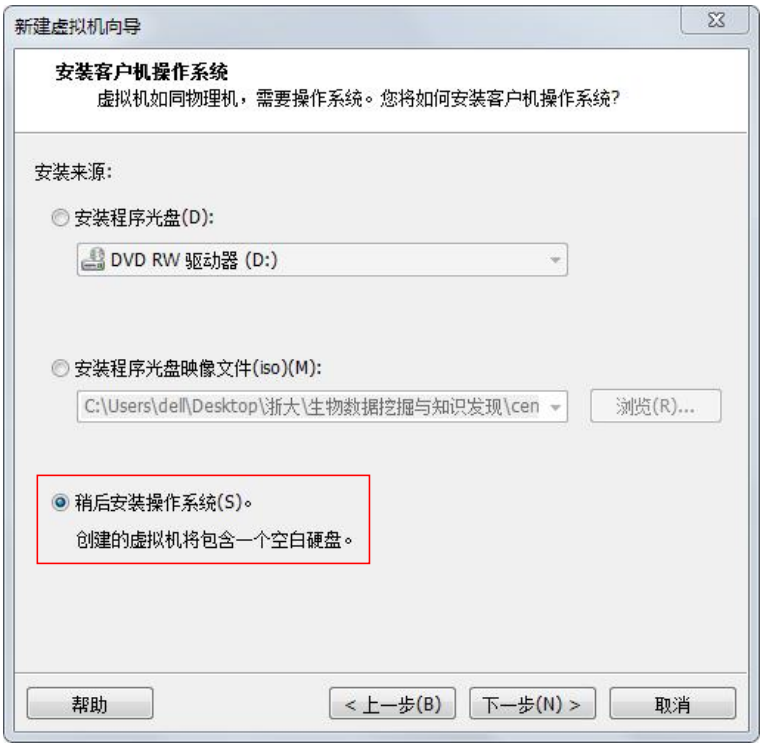


图 8 创建新的虚拟机——安装客户机操作系统

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

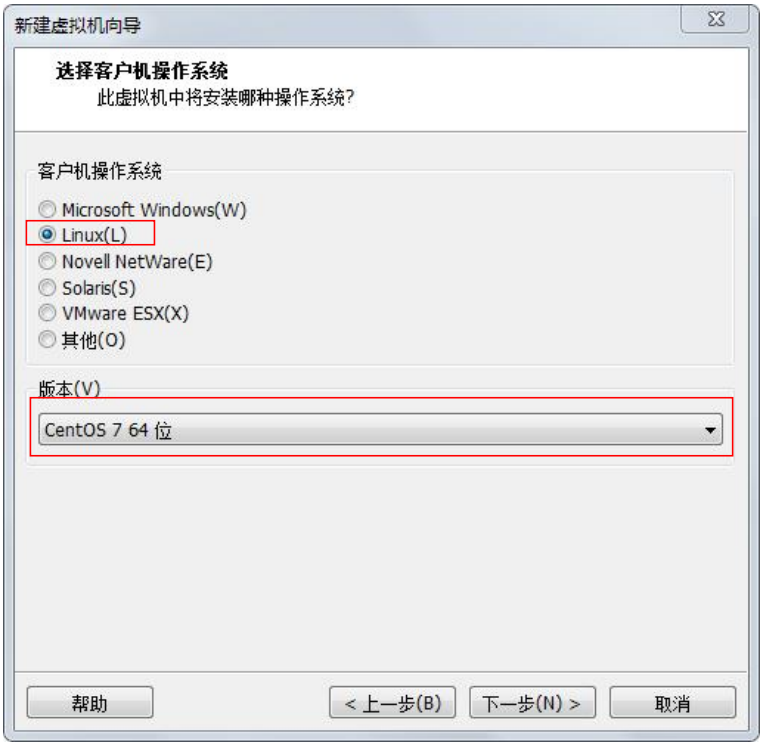


图 9 创建新的虚拟机——选择客户机操作系统

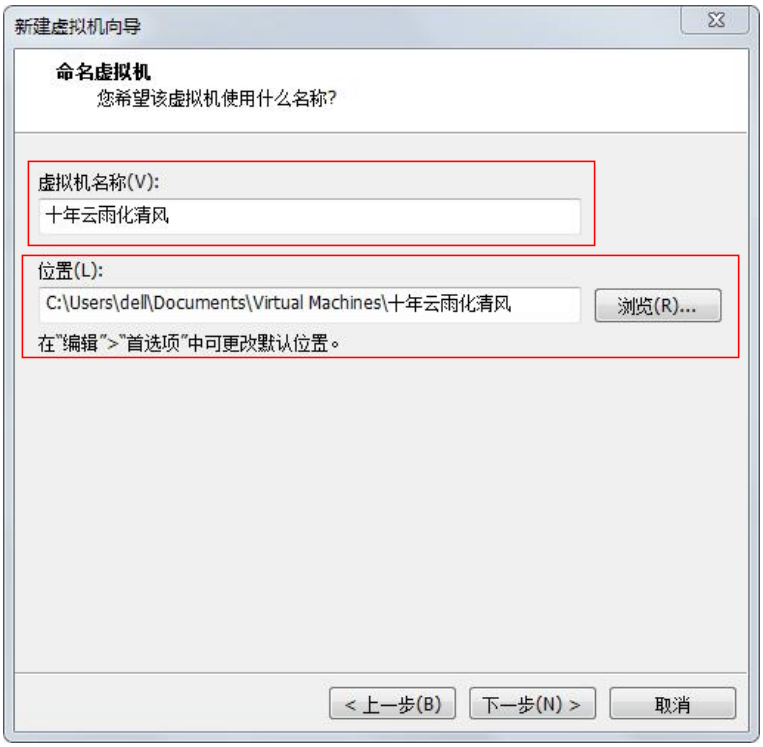


图 10 创建新的虚拟机——命名虚拟机

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

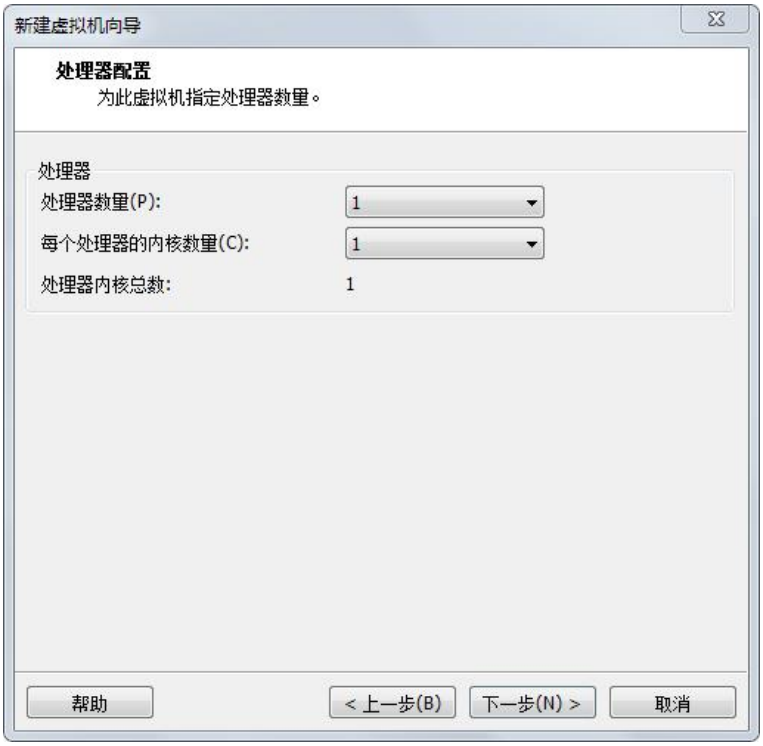


图 11 创建新的虚拟机——处理器配置

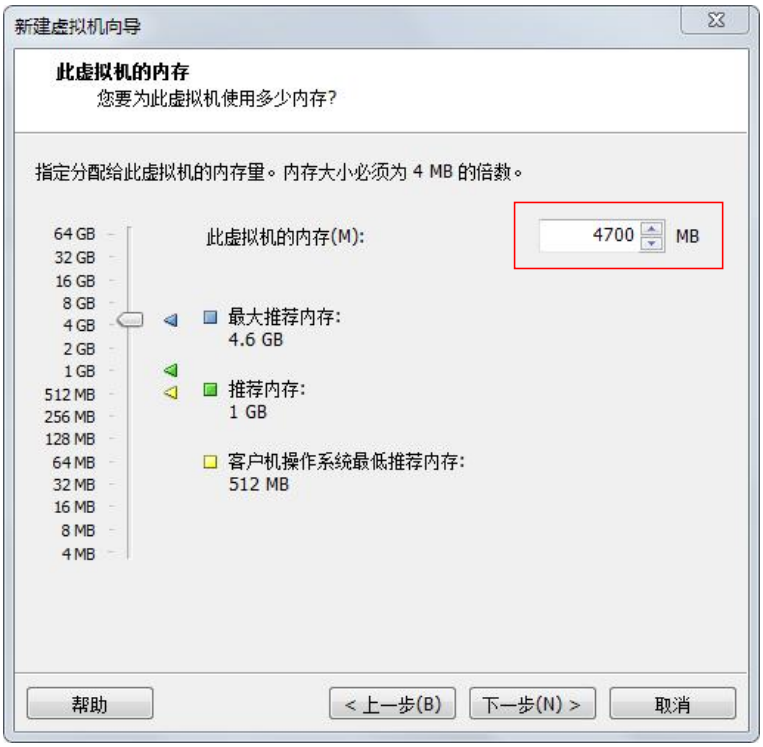


图 12 创建新的虚拟机——此虚拟机的内存

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

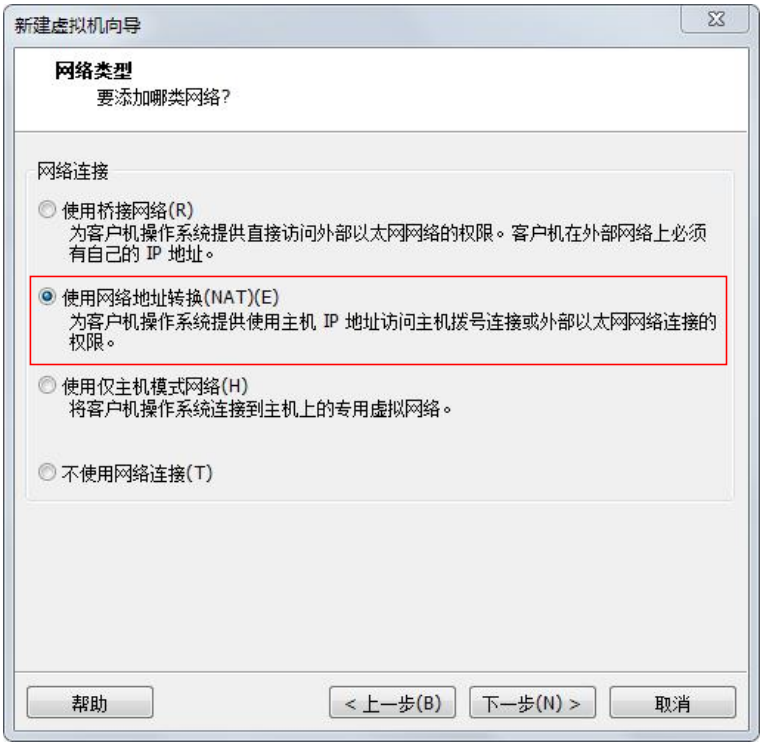


图 13 创建新的虚拟机——网络类型

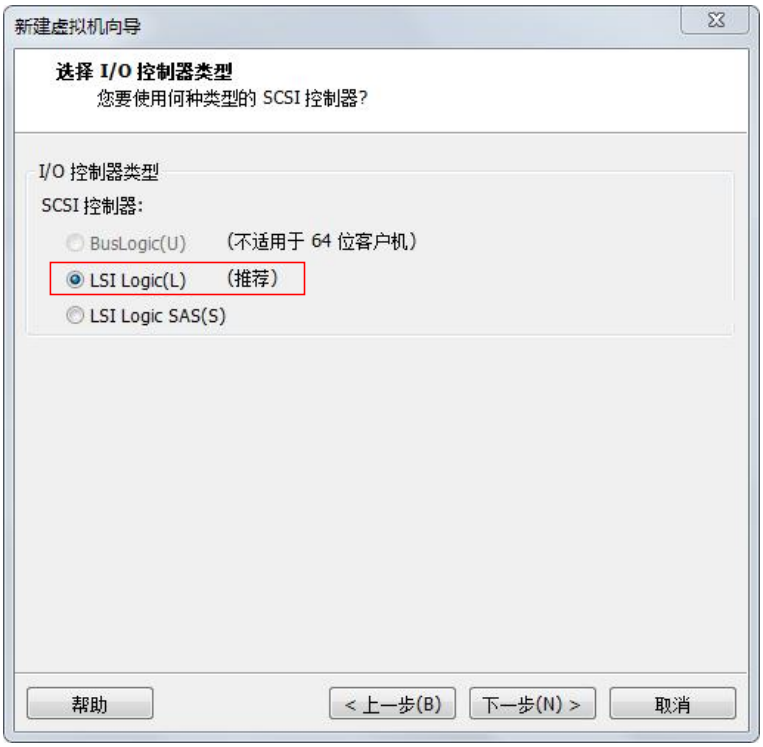


图 14 创建新的虚拟机——选择 I/0 控制器类型

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

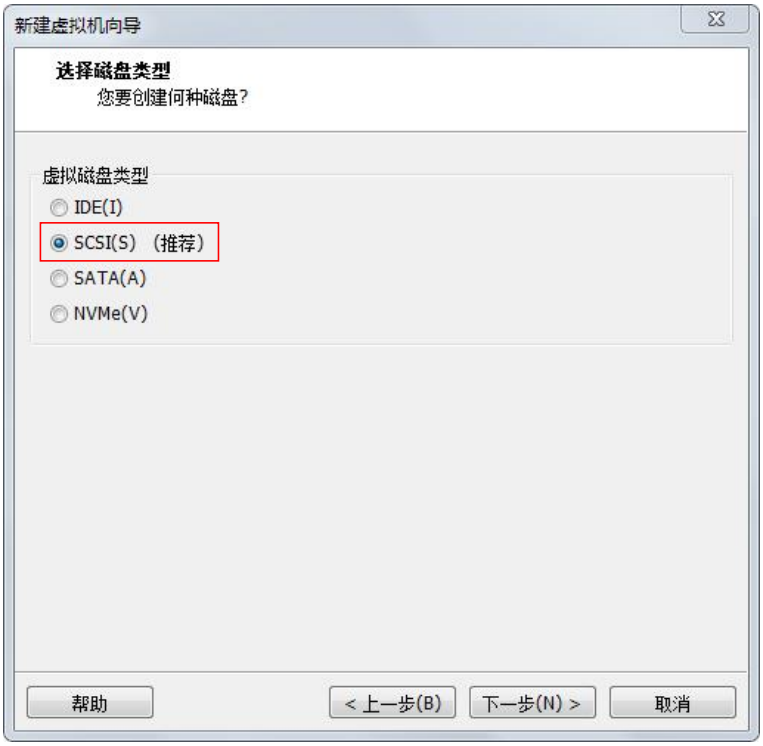


图 15 创建虚拟机——选择磁盘类型

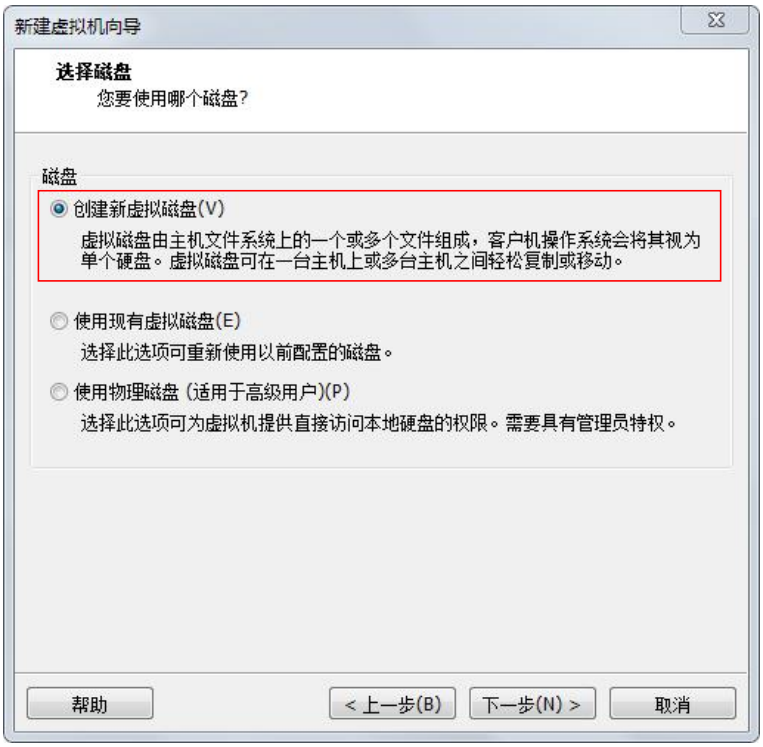


图 16 创建虚拟机——选择磁盘

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

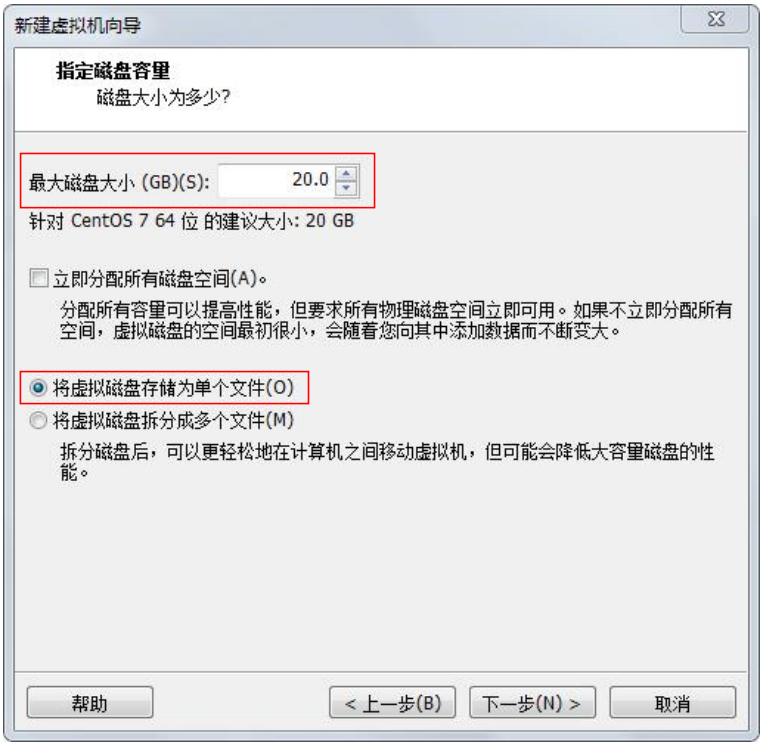


图 17 创建虚拟机——指定磁盘容量

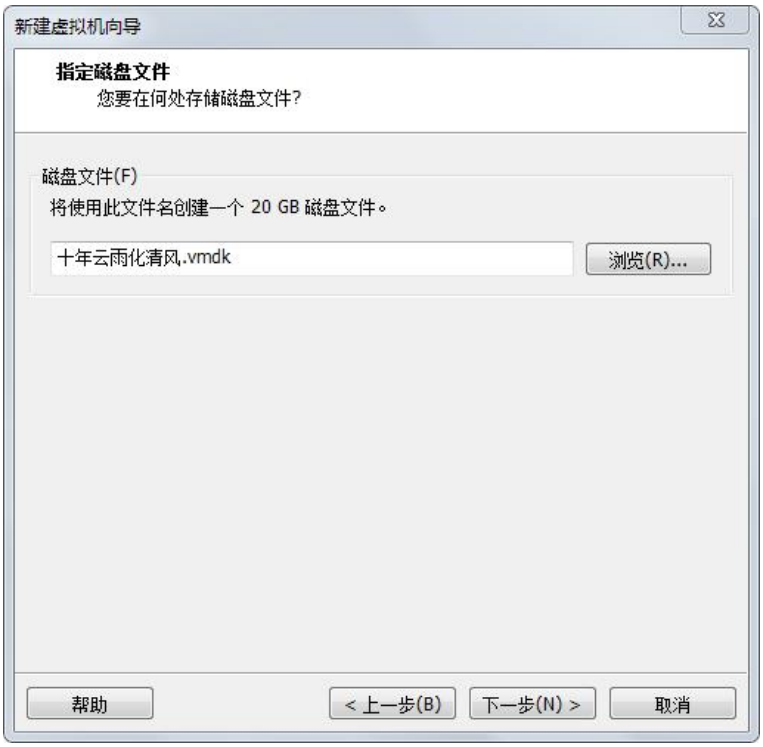


图 18 创建虚拟机——指定磁盘文件

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

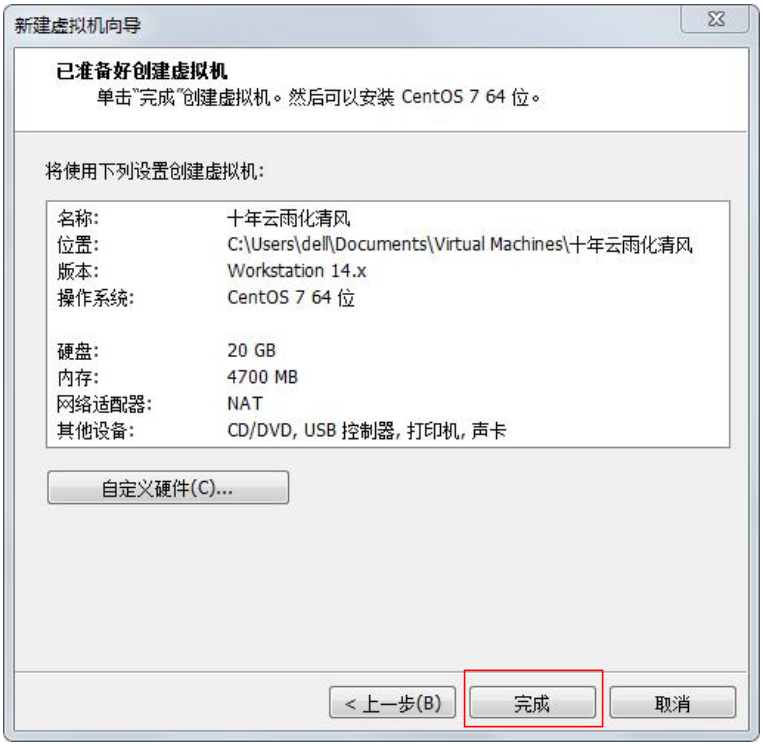
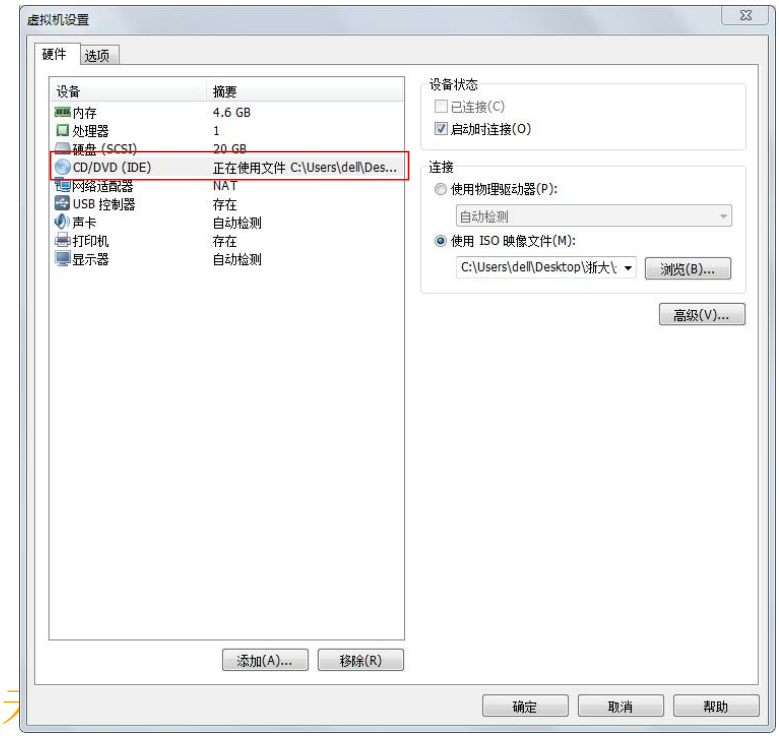


图 19 创建虚拟机——已准备好创建虚拟机

3、设置完成后的虚拟机显示如下界面，此时“硬件”选项中的“CD/DVD（IDE）”仍旧为“自动检测”状态，点击进行设置，选择“使用 ISO 映像文件（M）”，找到先前下载的 CentOS 7 镜像文件，打开即可（如图 20 所示）。



联系方式：binfo@zju.edu.cn

图 20 映像文件设置完成

1.2.3 虚拟机系统和软件的配置

1、开机并安装 CentOS 7 系统。选择“开启此虚拟机”（如图 21 所示），选择“install CentOS 7”（如图 22 和图 23 所示）。

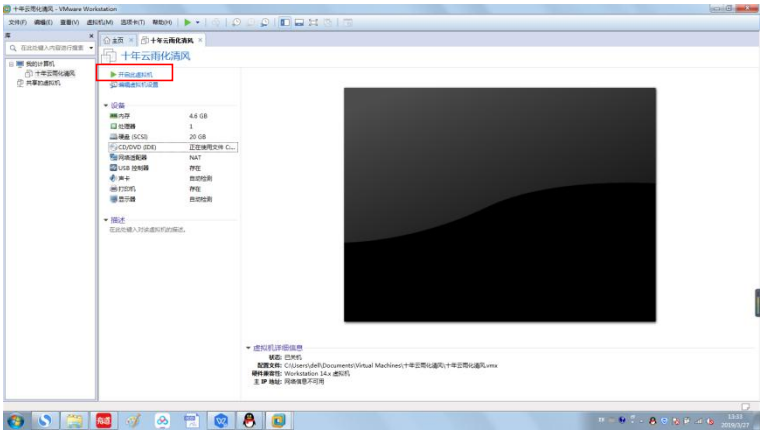


图 21 开启虚拟机

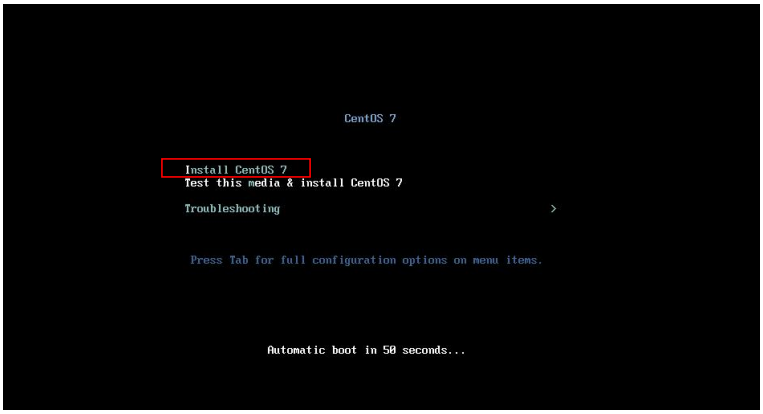


图 22 安装 CentOS 7

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

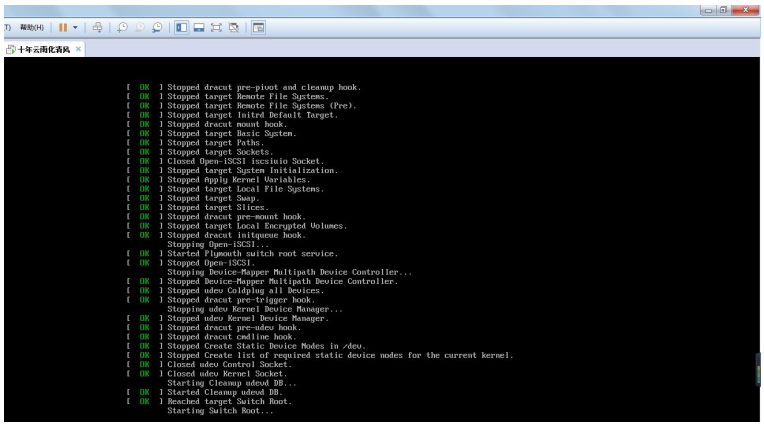


图 23 安装 CentOS 7

2、安装程序完成后，按需求选择虚拟机系统的配置。

（1）语言选择“中文”和“简体中文（中国）”（如图 24 所示）。

（2）进入下一个有多个选项的配置页面，对其他信息进行配置（如图 25 和图 26 所示）。

（3）在“本地化”一栏中，“日期和时间（T）”选择默认。

（4）在“本地化”一栏中，“键盘（X）”选择默认。

（5）在“本地化”一栏中，“语言支持（L）”选择默认。

（6）在“软件”一栏中，“安装源（I）”选择默认。

（7）在“软件”一栏中，“软件选择（S）”选择“带 GUI 的服务器”，将所有附加选项勾上（如图 38 所示）。

（8）在“系统”一栏中，“安装位置（D）”选择“我要配置分区”（如图 27 所示），点击“完成”，进入手动分区页面。

（9）按照需求手动设置分区（如图 28 所示），点击“完成”和“接受更改”（如图 29 所示）。

（10）在“系统”一栏中，“KDUMP”选择默认。

（11）在“系统”一栏中，“网络和主机名（N）”选择“打开”（如图 30 所示）。

（12）在“系统”一栏中，“SECURITY POLICY”选择默认。

（13）配置完成后，点击“开始安装（B）”。

（14）在配置过程中，设置 ROOT 密码并创建用户，将自己设置为管理员（如图 32 所示）。

（15）配置完成后，点击“重启”（如图 33 所示）。

未经授权，不得用于任何商业用途！
联系方式：bin16@zju.edu.cn

（16）重启之后需要接受许可证，接受即可（如图 34 和图 35 所示）。

（17）登录（如图 36 所示）。

（18）选择语言，一般都选中文。

（19）按需求选择设置隐私。

（20）在线账号的连接可以直接跳过。

（21）CentOS Linux（S）安装配置完毕（如图 37 所示）。



图 24 语言选择



图 25 配置页面

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn



图 26 配置页面

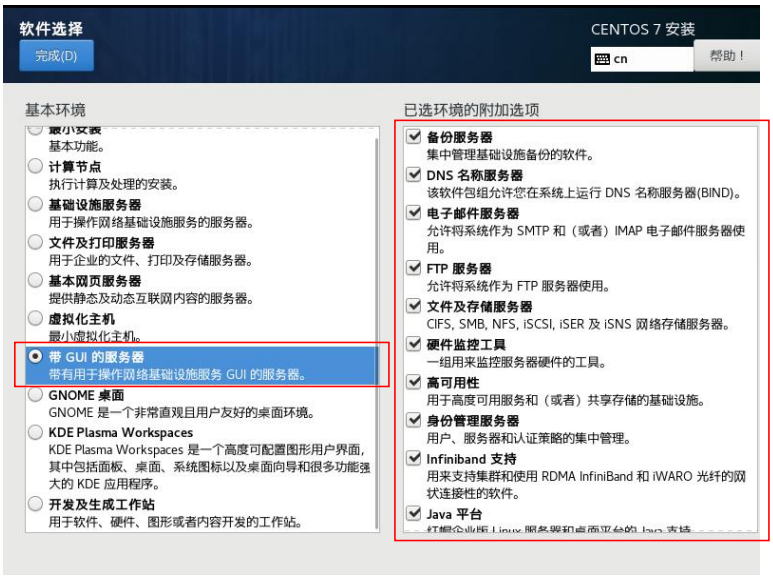


图 27 软件选择配置

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn



图 28 安装目标配置



图 29 手动分区配置

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn



图 30 手动分区配置

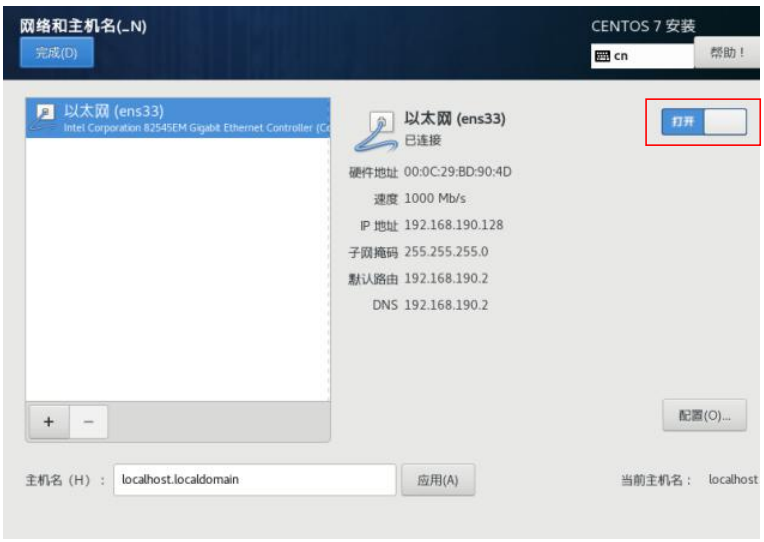


图 31 网络和主机名配置

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn



图 32 ROOT 密码配置和创建用户配置



图 33 配置完毕

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

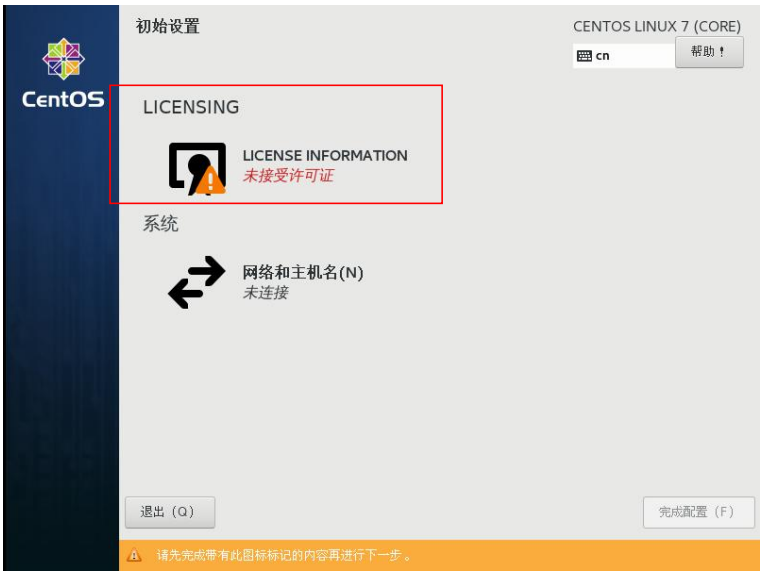


图 34 接受许可证

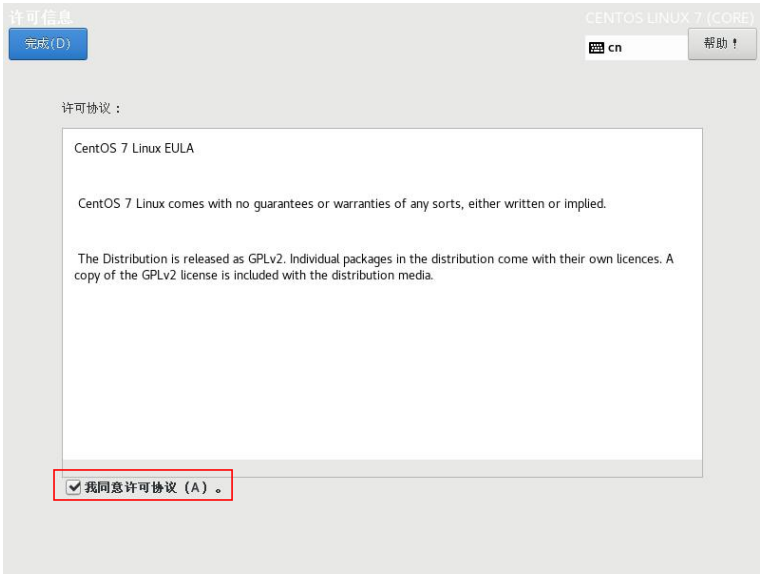


图 35 接受许可证

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

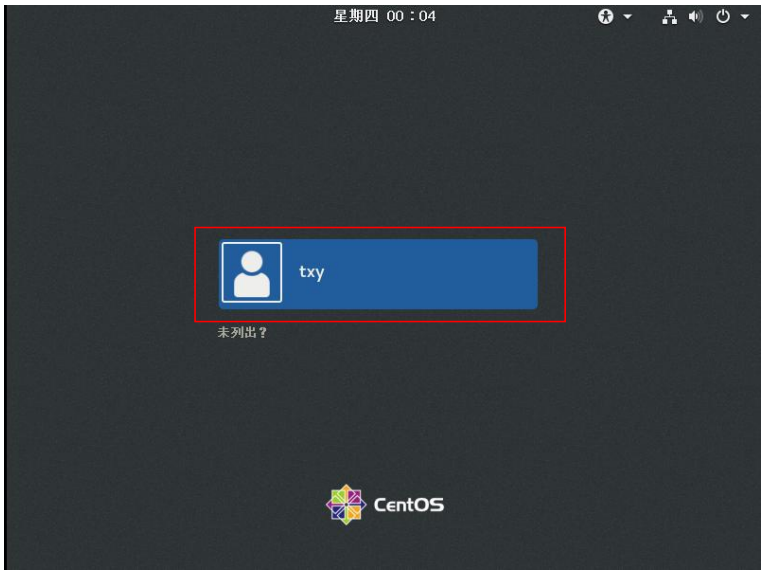


图 36 登录

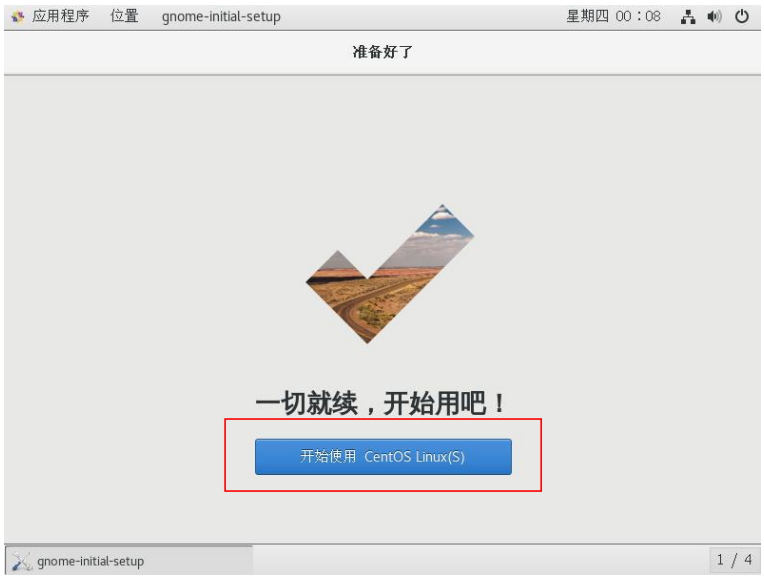


图 37 CentOS Linux（S）安装配置完毕

1.3 对于 Linux 环境的补充

事实上，考虑到虚拟机能够提供的存储容量相对于生物信息学的数据量和运算量是非常小的，以上安装和配置虚拟机的方法并不是最方便的，而且虚拟机安装和配置完毕后还需要自主下载软件，整个运算环境的构建非常花时间且效率不够高，这是本教程的缺陷之一，我们对此做出了改进建议，即可以直接安装生物信息学专用的 Bio-linux 系统相比之下就方便快捷了很多。

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

Bio-Linux 系统是为生物信息学定制的 Linux 操作系统，由 NERC Environmental Bioinformatics Centre 开发并维护。Bio-Linux 专门为环境、生物工程、药物设计、分子生物学研究定制，提供了 500 多个生物分析程序，由一个图形化的菜单进行管理，能方便地访问到其生物分析文档系统及对测试程序有用的样本数据。本教程接下来所使用的软件 Bio-Linux 系统都是有安装的，可以跳过下载的步骤直接使用，非常方便。而且 Bio-Linux 系统可以刻录到光盘或者优盘中，比单纯的 linux 系统更加灵活。

Bio-Linux 系统参考下载网址：https://linux.softpedia.com/get/System/Operating-Systems/Linux-Distributions/Bio-Linux-45495.shtml

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

2 从 FASAT 格式文件到 sam 格式文件——HISAT2

2.1 关于 HISAT2

HISAT2 是一个比对软件，能够将全基因组，转录组和外显子组的序列与单个或多个参考基因组比对，并且具有一定的可视化功能。HISAT2 比对 FASAT 格式的文件后输出的是 sam 格式的文件。

2.2 HISAT2 的安装

2.2.1 HISAT2 的下载和解压缩

1、为了便于管理所安装的软件，首先在主文件夹下创建一个文件夹，命名为“HISAT2”（如图 38-图 40 所示）。

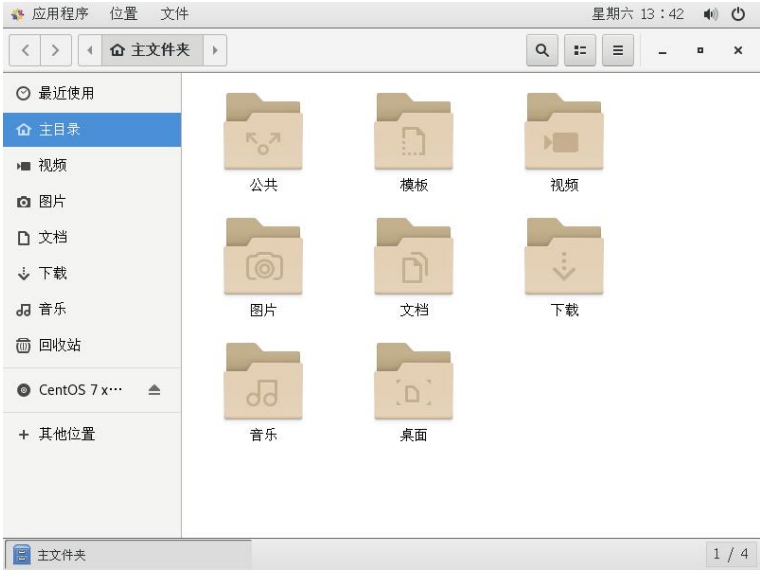


图 38 主文件夹

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

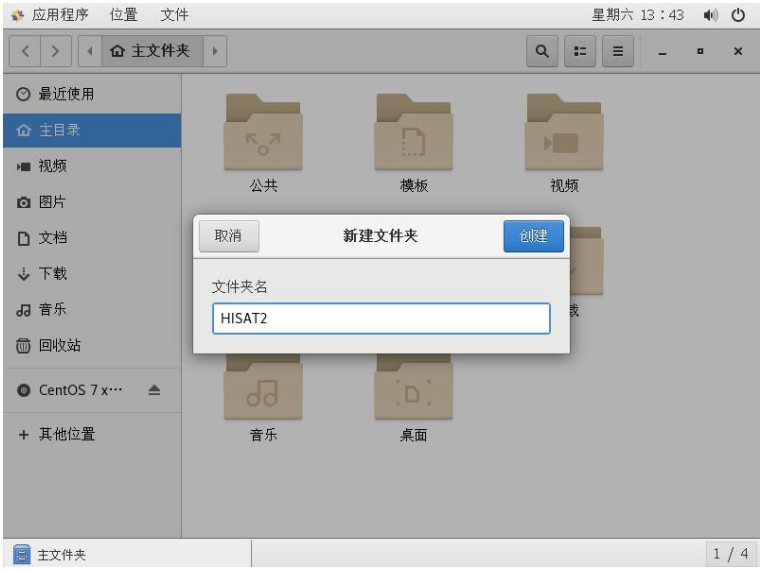


图 39 创建文件夹，命名为“HISAT2”

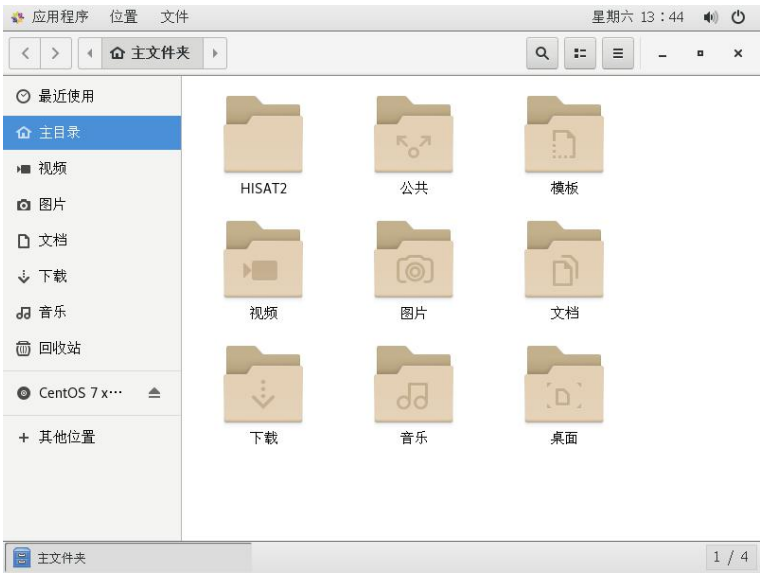


图 40 HISAT2 文件夹

2、打开终端。点击鼠标右键，选择“打开终端（E）”（如图 41 所示）。

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn



图 41 打开终端

3、下载 HISAT2，参考网址：http://ccb.jhu.edu/software/hisat2/dl/hisat2-2.1.0-Linux_x86_64.zip（如图 42 所示）。

代码：

```
wget http://ccb.jhu.edu/software/hisat2/dl/hisat2-2.1.0-Linux_x86_64.zip
#下载 HISAT2。
```

```
[txy@localhost ~]$ wget http://ccb.jhu.edu/software/hisat2/dl/hisat2-2.1.0-Linux_x86_64
.zip
--2019-03-30 13:57:45--  http://ccb.jhu.edu/software/hisat2/dl/hisat2-2.1.0-Linux_x86_6
4.zip
正在解析主机 ccb.jhu.edu (ccb.jhu.edu)... 128.220.233.225
正在连接 ccb.jhu.edu (ccb.jhu.edu)|128.220.233.225|:80... 已连接。
已发出 HTTP 请求，正在等待回应... 200 OK
长度：27104162 (26M) [application/zip]
正在保存至："hisat2-2.1.0-Linux_x86_64.zip"

100%[======================================>] 27,104,162  15.0KB/s 用时 53m 54s

2019-03-30 14:51:43 (8.18 KB/s) - 已保存 "hisat2-2.1.0-Linux_x86_64.zip" [27104162/2710
4162])
```

图 42 下载 HISAT2

4、解压缩 HISAT2（如图 43-图 47 所示）。

代码：

```
unzip hisat2-2.1.0-Linux_x86_64.zip
#解压缩 HISAT2。
```

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
[txy@localhost ~]$ unzip hisat2-2.1.0-Linux_x86_64.zip
Archive:  hisat2-2.1.0-Linux_x86_64.zip
   creating: hisat2-2.1.0/
  inflating: hisat2-2.1.0/hisat2
  inflating: hisat2-2.1.0/hisat2-inspect-l
  inflating: hisat2-2.1.0/hisatgenotype_build_genome.py
  inflating: hisat2-2.1.0/hisatgenotype_extract_vars.py
  inflating: hisat2-2.1.0/hisatgenotype_extract_reads.py
  inflating: hisat2-2.1.0/hisat2_extract_snps_haplotypes_VCF.py
  inflating: hisat2-2.1.0/AUTHORS
  inflating: hisat2-2.1.0/hisatgenotype.py
  inflating: hisat2-2.1.0/hisat2-inspect
   creating: hisat2-2.1.0/doc/
  inflating: hisat2-2.1.0/doc/style.css
  inflating: hisat2-2.1.0/doc/manual.inc.html
  inflating: hisat2-2.1.0/doc/README
  inflating: hisat2-2.1.0/hisat2-align-s-debug
  inflating: hisat2-2.1.0/extract_splice_sites.py
  inflating: hisat2-2.1.0/hisat2-build
   creating: hisat2-2.1.0/hisatgenotype_modules/
  inflating: hisat2-2.1.0/hisatgenotype_modules/hisatgenotype_assembly_graph.py
  inflating: hisat2-2.1.0/hisatgenotype_modules/hisatgenotype_typing_common.py
 extracting: hisat2-2.1.0/hisatgenotype_modules/__init__.py
  inflating: hisat2-2.1.0/hisatgenotype_modules/hisatgenotype_extract_cyp_data.py
  inflating: hisat2-2.1.0/hisatgenotype_modules/hisatgenotype_extract_codis_data.py
  inflating: hisat2-2.1.0/hisatgenotype_modules/hisatgenotype_convert_codis.py
```

图 43 解压缩 HISAT2

```
  inflating: hisat2-2.1.0/hisatgenotype_modules/hisatgenotype_gene_typing.py
  inflating: hisat2-2.1.0/hisat2-inspect-s-debug
  inflating: hisat2-2.1.0/hisat2-build-s-debug
  inflating: hisat2-2.1.0/MANUAL
  inflating: hisat2-2.1.0/hisat2-align-l
  inflating: hisat2-2.1.0/hisat2-build-s
  inflating: hisat2-2.1.0/hisatgenotype_hla_cyp.py
   creating: hisat2-2.1.0/scripts/
  inflating: hisat2-2.1.0/scripts/make_wbcel235_tran.sh
  inflating: hisat2-2.1.0/scripts/make_grcm38_tran.sh
  inflating: hisat2-2.1.0/scripts/make_b_taurus_UMD3.sh
  inflating: hisat2-2.1.0/scripts/make_rn4.sh
  inflating: hisat2-2.1.0/scripts/make_grch37_snp_tran_ercc.sh
  inflating: hisat2-2.1.0/scripts/make_hg19.sh
  inflating: hisat2-2.1.0/scripts/make_bdgp6_tran.sh
  inflating: hisat2-2.1.0/scripts/make_grch38_tran.sh
  inflating: hisat2-2.1.0/scripts/make_grch38_snp_tran.sh
  inflating: hisat2-2.1.0/scripts/make_grch37_snp_tran.sh
  inflating: hisat2-2.1.0/scripts/make_rn6.sh
  inflating: hisat2-2.1.0/scripts/make_grcm38_snp_tran.sh
  inflating: hisat2-2.1.0/scripts/make_wbcel235.sh
  inflating: hisat2-2.1.0/scripts/make_canFam2.sh
  inflating: hisat2-2.1.0/scripts/make_grch38_snp.sh
  inflating: hisat2-2.1.0/scripts/make_e_coli.sh
  inflating: hisat2-2.1.0/scripts/make_dm6.sh
  inflating: hisat2-2.1.0/scripts/make_grch37.sh
```

图 44 解压缩 HISAT2

```
  inflating: hisat2-2.1.0/scripts/make_grch37_snp.sh
  inflating: hisat2-2.1.0/scripts/make_rnor6.sh
  inflating: hisat2-2.1.0/scripts/make_hg38_snp_tran.sh
  inflating: hisat2-2.1.0/scripts/make_a_thaliana_tair.sh
  inflating: hisat2-2.1.0/scripts/make_grch37_tran.sh
  inflating: hisat2-2.1.0/scripts/gen_2b_occ_lookup.pl
  inflating: hisat2-2.1.0/scripts/make_grcm38.sh
  inflating: hisat2-2.1.0/scripts/make_bdgp6.sh
  inflating: hisat2-2.1.0/scripts/make_mm9.sh
  inflating: hisat2-2.1.0/scripts/make_hg38_snp.sh
  inflating: hisat2-2.1.0/scripts/make_mm10.sh
  inflating: hisat2-2.1.0/scripts/infer_fraglen.pl
  inflating: hisat2-2.1.0/scripts/make_hg38_allsnp.sh
  inflating: hisat2-2.1.0/scripts/make_hg38.sh
  inflating: hisat2-2.1.0/scripts/gen_occ_lookup.pl
  inflating: hisat2-2.1.0/scripts/make_ce10.sh
  inflating: hisat2-2.1.0/scripts/make_grcm38_snp.sh
  inflating: hisat2-2.1.0/scripts/gen_solqual_lookup.pl
  inflating: hisat2-2.1.0/scripts/make_grch38_snp_tran_ercc.sh
  inflating: hisat2-2.1.0/scripts/make_sc3.sh
  inflating: hisat2-2.1.0/scripts/make_zm3_snp_tran_ercc.sh
  inflating: hisat2-2.1.0/scripts/make_grch38.sh
  inflating: hisat2-2.1.0/scripts/make_rnor6_tran.sh
  inflating: hisat2-2.1.0/scripts/make_r64.sh
  inflating: hisat2-2.1.0/scripts/make_r64_tran.sh
  inflating: hisat2-2.1.0/scripts/convert_quals.pl
```

图 45 解压缩 HISAT2

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
 inflating: hisat2-2.1.0/hisat2-build-l-debug
 inflating: hisat2-2.1.0/TUTORIAL
 inflating: hisat2-2.1.0/extract_exons.py
 inflating: hisat2-2.1.0/hisat2-inspect-s
 inflating: hisat2-2.1.0/hisat2-inspect-l-debug
  creating: hisat2-2.1.0/hisatgenotype_scripts/
 inflating: hisat2-2.1.0/hisatgenotype_scripts/extract_Omixon_HLA.py
 inflating: hisat2-2.1.0/hisatgenotype_scripts/hisatgenotype_HLA_genotyping_PGs.py
 inflating: hisat2-2.1.0/hisatgenotype_scripts/compare_HLA_Omixon.py
 inflating: hisat2-2.1.0/hisatgenotype_scripts/hisatgenotype_locus_samples.py
extracting: hisat2-2.1.0/VERSION
 inflating: hisat2-2.1.0/hisat2_simulate_reads.py
 inflating: hisat2-2.1.0/hisat2-align-l-debug
 inflating: hisat2-2.1.0/hisat2-build-l
 inflating: hisat2-2.1.0/NEWS
 inflating: hisat2-2.1.0/hisat2-align-s
 inflating: hisat2-2.1.0/LICENSE
 inflating: hisat2-2.1.0/hisatgenotype_locus.py
 inflating: hisat2-2.1.0/hisat2_extract_exons.py
 inflating: hisat2-2.1.0/hisat2_extract_splice_sites.py
 inflating: hisat2-2.1.0/MANUAL.markdown
  creating: hisat2-2.1.0/example/
  creating: hisat2-2.1.0/example/index/
 inflating: hisat2-2.1.0/example/index/22_20-21M_snp.2.ht2
 inflating: hisat2-2.1.0/example/index/22_20-21M_snp.5.ht2
 inflating: hisat2-2.1.0/example/index/22_20-21M_snp.4.ht2
```

图 46 解压缩 HISAT2

```
 inflating: hisat2-2.1.0/example/index/22_20-21M_snp.3.ht2
 inflating: hisat2-2.1.0/example/index/22_20-21M_snp.7.ht2
 inflating: hisat2-2.1.0/example/index/22_20-21M_snp.6.ht2
 inflating: hisat2-2.1.0/example/index/22_20-21M_snp.1.ht2
 inflating: hisat2-2.1.0/example/index/22_20-21M_snp.8.ht2
  creating: hisat2-2.1.0/example/reference/
 inflating: hisat2-2.1.0/example/reference/22_20-21M.fa
 inflating: hisat2-2.1.0/example/reference/22_20-21M.snp
  creating: hisat2-2.1.0/example/reads/
 inflating: hisat2-2.1.0/example/reads/reads_1.fa
 inflating: hisat2-2.1.0/example/reads/reads_2.fa
 inflating: hisat2-2.1.0/hisat2_extract_snps_haplotypes_UCSC.py
```

图 47 解压缩 HISAT2

5、解压缩后的 HISAT2 文件和 HISAT2 压缩包都位于主文件夹中（如图 48 所示），为了便于管理，将它们一起放入先前建好的 HISAT2 文件夹中（如图 49 所示）。

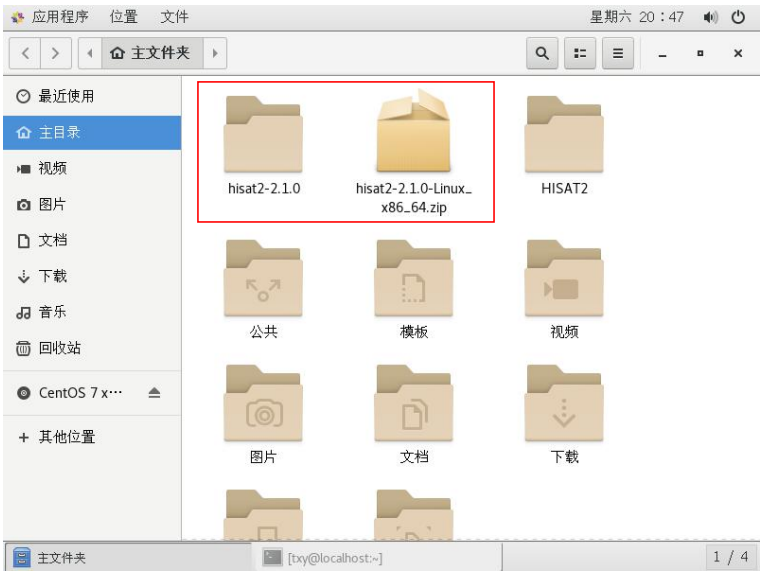


图 48 解压缩后的 HISAT2 文件和 HISAT2 压缩包

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

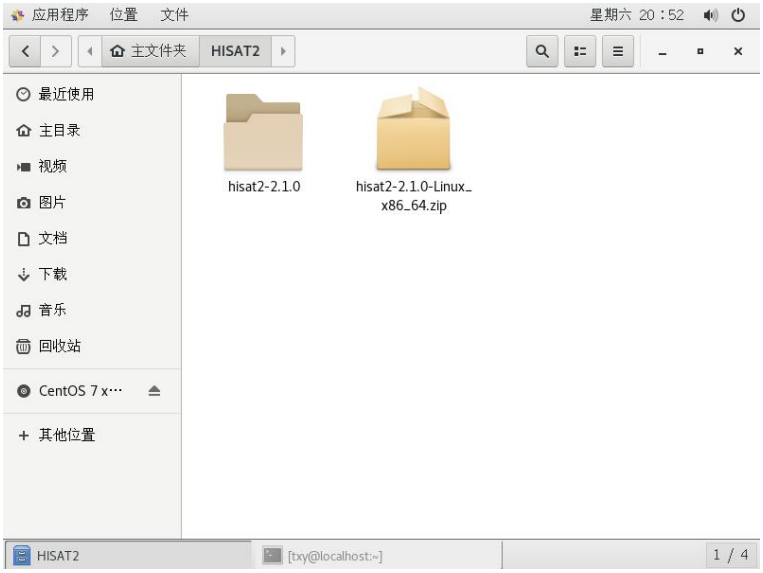


图 49 将解压缩后的 HISAT2 文件和 HISAT2 压缩包放入文件夹 HISAT2 中

2.2.2 命令路径的修改

1、在终端打开隐藏的.bashrc 文件（如图 50 所示）。

代码：

vi ~/.bashrc

#打开.bashrc 文件。

```
# .bashrc

# Source global definitions
if [ -f /etc/bashrc ]; then
        . /etc/bashrc
fi

# Uncomment the following line if you don't like systemctl's auto-paging feature:
# export SYSTEMD_PAGER=

# User specific aliases and functions
~
~
~
~
~
~
~
~
~
~
~
~
~
"~/.bashrc" 11L, 231C                                         1,1           全部
```

图 50 在终端打开隐藏的.bashrc 文件

2、按下键盘上的“I”，进入插入状态，在“#User specific aliases and functions”这一行下输入相应的路径，即 PATH=$PATH：/home/txy/HISAT2/hisat2-2.1.0（如图 51 所示）。

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

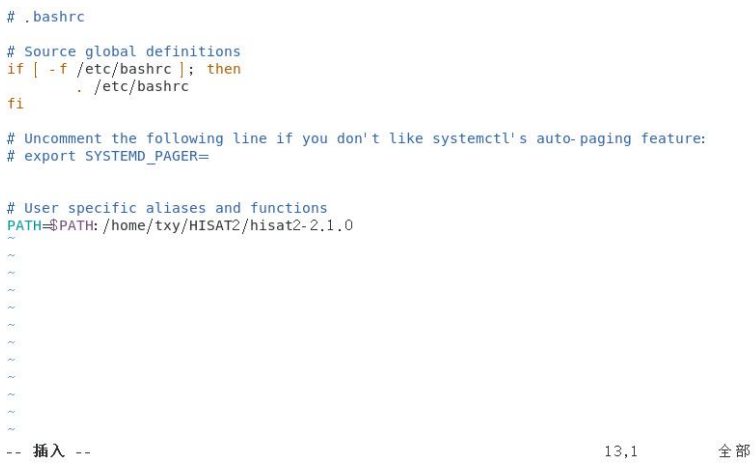


图 51 添加新的命令路径

3、按下键盘上的“esc”键（如图 52 所示），结束插入状态，输入“：wq”并回车，即保存了新的命令路径（如图 53 和图 54 所示）。

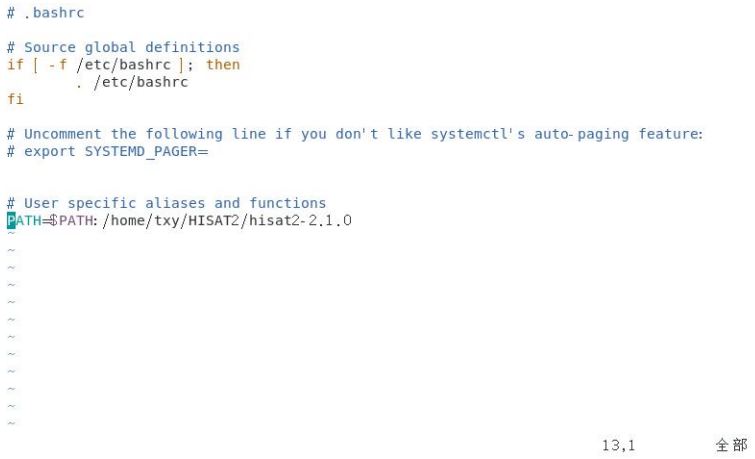


图 52 按下“esc”键结束插入状态

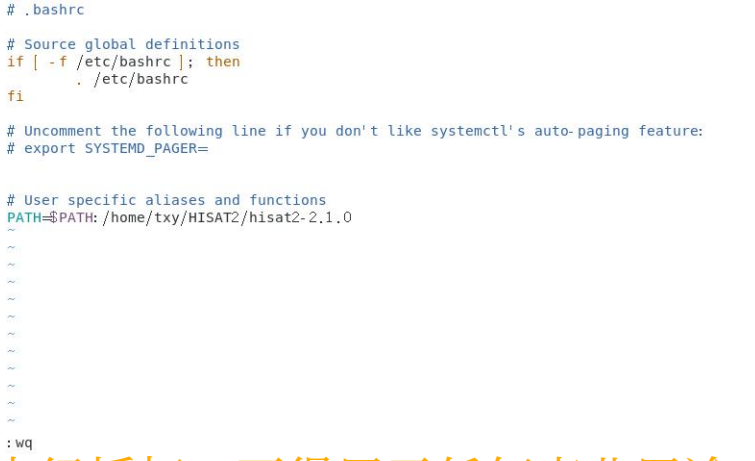


图 53 输入“：wq”

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

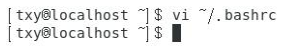


图 54 保存了新的命令路径

4、执行新的命令路径（如图 55 所示）。

代码：

source ~/.bashrc

#执行新的命令路径。



图 55 执行新的命令路径

2.3 HISAT2 的功能与使用方法

2.3.1 help 功能的使用

1、hisat2 功能，文件和版本等相关的信息可以利用 help 功能进行查询。同时，这一步操作也能检验 hisat2 是否安装完成和命令路径是否正确，只有在这两者都正确完成的情况下，hisat2 这个命令才能正常使用（如图 56-图 60 所示）。

代码：

hisat2-h

#hisat2 功能使用帮助。

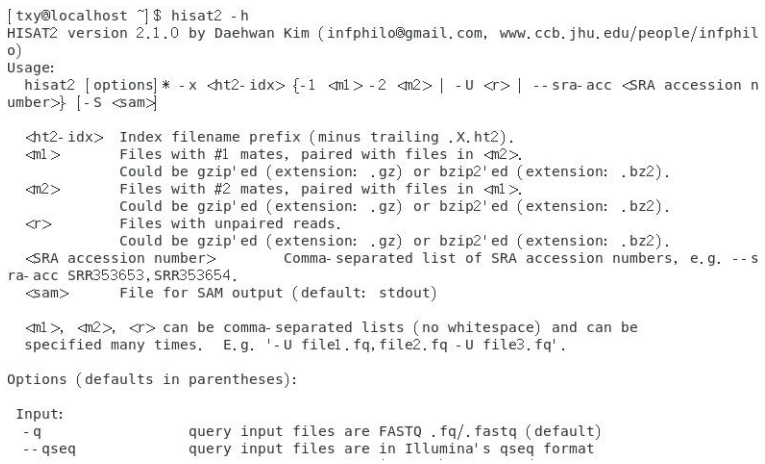


图 56 hisat2-h 功能的使用

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
  -f                 query input files are (multi-)FASTA .fa/.mfa
  -r                 query input files are raw one-sequence-per-line
  -c                 <m1>, <m2>, <r> are sequences themselves, not files
  -s/--skip <int>    skip the first <int> reads/pairs in the input (none)
  -u/--upto <int>    stop after first <int> reads/pairs (no limit)
  -5/--trim5 <int>   trim <int> bases from 5'/left end of reads (0)
  -3/--trim3 <int>   trim <int> bases from 3'/right end of reads (0)
  --phred33          qualities are Phred+33 (default)
  --phred64          qualities are Phred+64
  --int-quals        qualities encoded as space-delimited integers
  --sra-acc          SRA accession ID

 Alignment:
  --n-ceil <func>    func for max # non-A/C/G/Ts permitted in aln (L,0,0.15)
  --ignore-quals     treat all quality values as 30 on Phred scale (off)
  --nofw             do not align forward (original) version of read (off)
  --norc             do not align reverse-complement version of read (off)

 Spliced Alignment:
  --pen-cansplice <int>              penalty for a canonical splice site (0)
  --pen-noncansplice <int>           penalty for a non-canonical splice site (12)
  --pen-canintronlen <func>          penalty for long introns (G,-8,1) with canonical s
plice sites
  --pen-noncanintronlen <func>       penalty for long introns (G,-8,1) with noncanonica
l splice sites
  --min-intronlen <int>              minimum intron length (20)
```

图 57 hisat2-h 功能的使用

```
  --max-intronlen <int>              maximum intron length (500000)
  --known-splicesite-infile <path>   provide a list of known splice sites
  --novel-splicesite-outfile <path>  report a list of splice sites
  --novel-splicesite-infile <path>   provide a list of novel splice sites
  --no-temp-splicesite               disable the use of splice sites found
  --no-spliced-alignment             disable spliced alignment
  --rna-strandness <string>          specify strand-specific information (unstranded)
  --tmo                              reports only those alignments within known transcr
iptome
  --dta                              reports alignments tailored for transcript assembl
ers
  --dta-cufflinks                    reports alignments tailored specifically for cuffl
inks
  --avoid-pseudogene                 tries to avoid aligning reads to pseudogenes (expe
rimental option)
  --no-templatelen-adjustment        disables template length adjustment for RNA-seq re
ads

 Scoring:
  --mp <int>,<int>   max and min penalties for mismatch; lower qual = lower penalty <6,
2>
  --sp <int>,<int>   max and min penalties for soft-clipping; lower qual = lower penalt
y <2,1>
  --no-softclip      no soft-clipping
  --np <int>         penalty for non-A/C/G/Ts in read/ref (1)
  --rdg <int>,<int>  read gap open, extend penalties (5,3)
```

图 58 hisat2-h 功能的使用

```
  --score-min <func> min acceptable alignment score w/r/t read length
                     (L,0.0,-0.2)

 Reporting:
  -k <int> (default: 5) report up to <int> alns per read

 Paired-end:
  -I/--minins <int>  minimum fragment length (0), only valid with --no-spliced-alignmen
t
  -X/--maxins <int>  maximum fragment length (500), only valid with --no-spliced-alignm
ent
  --fr/--rf/--ff     -1, -2 mates align fw/rev, rev/fw, fw/fw (--fr)
  --no-mixed         suppress unpaired alignments for paired reads
  --no-discordant    suppress discordant alignments for paired reads

 Output:
  -t/--time          print wall-clock time taken by search phases
  --un <path>            write unpaired reads that didn't align to <path>
  --al <path>            write unpaired reads that aligned at least once to <path>
  --un-conc <path>       write pairs that didn't align concordantly to <path>
  --al-conc <path>       write pairs that aligned concordantly at least once to <path>
  (Note: for --un, --al, --un-conc, or --al-conc, add '-gz' to the option name, e.g.
  --un-gz <path>, to gzip compress output, or add '-bz2' to bzip2 compress output.)
  --summary-file     print alignment summary to this file.
  --new-summary      print alignment summary in a new style, which is more machine-frie
```

图 59 hisat2-h 功能的使用

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
ndly.
  --quiet            print nothing to stderr except serious errors
  --met-file <path>  send metrics to file at <path> (off)
  --met-stderr       send metrics to stderr (off)
  --met <int>        report internal counters & metrics every <int> secs (1)
  --no-head          supppress header lines, i.e. lines starting with @
  --no-sq            supppress @SQ header lines
  --rg-id <text>     set read group id, reflected in @RG line and RG:Z: opt field
  --rg <text>        add <text> ("lab:value") to @RG line of SAM header.
                     Note: @RG line only printed when --rg-id is set.
  --omit-sec-seq     put '*' in SEQ and QUAL fields for secondary alignments.

 Performance:
  -o/--offrate <int> override offrate of index; must be >= index's offrate
  -p/--threads <int> number of alignment threads to launch (1)
  --reorder          force SAM output order to match order of input reads
  --mm               use memory-mapped I/O for index; many 'hisat2's can share

 Other:
  --qc-filter        filter out reads that are bad according to QSEQ filter
  --seed <int>       seed for random number generator (0)
  --non-deterministic seed rand. gen. arbitrarily instead of using read attributes
  --remove-chrname   remove 'chr' from reference names in alignment
  --add-chrname      add 'chr' to reference names in alignment
  --version          print version information and quit
  -h/--help        _ print this usage message
```

图 60 hisat2-h 功能的使用

2、利用 hisat2 可以构建索引文件，构建索引文件的方式，构建索引文件需要的文件和构建索引文件需要设置的参数以及参数的默认值都可以通过 help 功能进行查询（如图 61 和图 62 所示）。

代码：

hisat2-build -h

#hisat2 构建索引文件功能帮助。

```
[txy@localhost ~]$ hisat2-build -h
HISAT2 version 2.1.0 by Daehwan Kim (infphilo@gmail.com, http://www.ccb.jhu.edu/people/
infphilo)
Usage: hisat2-build [options]* <reference_in> <ht2_index_base>
    reference_in            comma-separated list of files with ref sequences
    hisat2_index_base       write ht2 data to files with this dir/basename
Options:
    -c                      reference sequences given on cmd line (as
                            <reference_in>)
    --large-index           force generated index to be 'large', even if ref
                            has fewer than 4 billion nucleotides
    -a/--noauto             disable automatic -p/--bmax/--dcv memory-fitting
    -p                      number of threads
    --bmax <int>            max bucket sz for blockwise suffix-array builder
    --bmaxdivn <int>        max bucket sz as divisor of ref len (default: 4)
    --dcv <int>             diff-cover period for blockwise (default: 1024)
    --nodc                  disable diff-cover (algorithm becomes quadratic)
    -r/--noref              don't build .3/.4.ht2 (packed reference) portion
    -3/--justref            just build .3/.4.ht2 (packed reference) portion
    -o/--offrate <int>      SA is sampled every 2^offRate BWT chars (default: 5)
    -t/--ftabchars <int>    # of chars consumed in initial lookup (default: 10)
    --localoffrate <int>    SA (local) is sampled every 2^offRate BWT chars (default: 3
)
    --localftabchars <int>  # of chars consumed in initial lookup in a local index (def
ault: 6)
    --snp <path>            SNP file name
```

图 61 hisat2-build -h 功能的使用

```
    --haplotype <path>      haplotype file name
    --ss <path>             Splice site file name
    --exon <path>           Exon file name
    --seed <int>            seed for random number generator
    -q/--quiet              verbose output (for debugging)
    -h/--help               print detailed description of tool and its options
    --usage                 print this usage message
    --version               print version information and quit
```

图 62 hisat2-build -h 功能的使用

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

3、利用 hisat2 可以从 GTF 文件中提取出外显子的相关信息并转换成 hisat2 专用的“.exon”文件格式，转换所需要的代码和文件都可以通过 help 功能进行查询（如图 63 所示）。

代码：

hisat2_extract_exons.py -h

#hisat2 提取外显子功能帮助。

```
[txy@localhost ~]$ hisat2_extract_exons.py -h
usage: hisat2_extract_exons.py [-h] [-v] [gtf_file]

Extract exons from a GTF file

positional arguments:
  gtf_file       input GTF file (use "-" for stdin)

optional arguments:
  -h, --help     show this help message and exit
  -v, --verbose  also print some statistics to stderr
```

图 63 hisat2_extract_exons.py -h 功能的使用

4、利用 hisat2 可以从 GTF 文件中提取出剪接位点的相关信息并转换成 hisat2 专用的“.ss”文件格式，转换所需要的代码和文件都可以通过 help 功能进行查询（如图 64 所示）。

代码：

hisat2_extract_splice_sites.py -h

#hisat2 提取剪接位点功能帮助。

```
[txy@localhost ~]$ hisat2_extract_splice_sites.py -h
usage: hisat2_extract_splice_sites.py [-h] [-v] [gtf_file]

Extract splice junctions from a GTF file

positional arguments:
  gtf_file       input GTF file (use "-" for stdin)

optional arguments:
  -h, --help     show this help message and exit
  -v, --verbose  also print some statistics to stderr
```

图 64 hisat2_extract_splice_sites.py -h 功能的使用

5、利用 hisat2 可以从下载自 UCSC 的 SNP 文件中提取出 SNP 和单倍型的相关信息并转换成 hisat2 专用的“.snp”文件格式，转换所需要的代码和文件都可以通过 help 功能进行查询（如图 65 所示）。

代码：

hisat2_extract_snps_haplotypes_UCSC.py -h

#hisat2 从下载自 UCSC 的文件中提取 snp 和单倍型文件功能帮助。

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
[txy@localhost ~]$ hisat2_extract_snps_haplotypes_UCSC.py -h
usage: hisat2_extract_snps_haplotypes_UCSC.py [-h] [--inter-gap INTER_GAP]
                                              [--intra-gap INTRA_GAP] [-v]
                                              [--testset]
                                              [genome_file] [snp_fname]
                                              [base_fname]

Extract SNPs and haplotypes from a SNP file downloaded from UCSC (e.g.
http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/snp144.txt.gz)

positional arguments:
  genome_file           input genome file (e.g. genome.fa)
  snp_fname             input snp file downloaded from UCSC (plain text or
                        gzipped file is accepted: snp144Common.txt or
                        snp144Common.txt.gz)
  base_fname            base filename for SNPs and haplotypes

optional arguments:
  -h, --help            show this help message and exit
  --inter-gap INTER_GAP
                        Maximum distance for variants to be in the same
                        haplotype
  --intra-gap INTRA_GAP
                        Break a haplotype into several haplotypes
  -v, --verbose         also print some statistics to stderr
  --testset             print test reads
```

图 65 hisat2_extract_snps_haplotypes_UCSC.py -h 功能的使用

6、利用 hisat2 可以从 VCF 文件中提取出 SNP 和单倍型的相关信息并转换成 hisat2 专用的“.snp”文件格式，转换所需要的代码和文件都可以通过 help 功能进行查询（如图 66 和图 67 所示）。

代码：

hisat2_extract_snps_haplotypes_VCF.py -h

#hisat2 从 VCF 文件中提取 snp 和单倍型文件功能帮助。

```
[txy@localhost ~]$ hisat2_extract_snps_haplotypes_VCF.py -h
usage: hisat2_extract_snps_haplotypes_VCF.py [-h]
                                             [--reference-type REFERENCE_TYPE]
                                             [--inter-gap INTER_GAP]
                                             [--intra-gap INTRA_GAP]
                                             [--non-rs]
                                             [--genotype-vcf GENOTYPE_VCF]
                                             [--genotype-gene-list GENOTYPE_GENE_LIST]
                                             [--extra-files] [-v]
                                             [genome_file] [VCF_fnames]
                                             [base_fname]

Extract SNPs and haplotypes from VCF files

positional arguments:
  genome_file           input genome file (e.g. genome.fa)
  VCF_fnames            A comma-seperated VCF files (plain text or gzipped
                        file is accepted: GRCh38_dbSNP_no_SVs.vcf or
                        GRCh38_dbSNP_no_SVs.vcf.gz
  base_fname            base filename for SNPs and haplotypes

optional arguments:
  -h, --help            show this help message and exit
  --reference-type REFERENCE_TYPE
                        Reference type: gene, chromosome, and genome (default:
                        genome)
```

图 66 hisat2_extract_snps_haplotypes_VCF.py -h 功能的使用

```
  --inter-gap INTER_GAP
                        Maximum distance for variants to be in the same
                        haplotype (default: 30)
  --intra-gap INTRA_GAP
                        Break a haplotype into several haplotypes (default:
                        50)
  --non-rs              Allow SNP IDs not beginning with rs
  --genotype-vcf GENOTYPE_VCF
                        VCF file name for genotyping (default: empty)
  --genotype-gene-list GENOTYPE_GENE_LIST
                        A comma-separated list of genes to be genotyped
                        (default: empty)
  --extra-files         Output extra files such as _backbone.fa and .ref
  -v, --verbose         also print some statistics to stderr
```

图 67 hisat2_extract_snps_haplotypes_VCF.py -h 功能的使用

无偿提供，如有侵权，请联系作者删除。
联系方式：bhfu@zju.edu.cn

7、hisat2 能够将被编入索引文件的 FASTA 格式的记录显示出来，所需要的代码和文件都可以通过 help 功能进行查询（如图 68 所示）。

代码：

hisat2-inspect -h

#hisat2 视察索引文件功能帮助。

```
[txy@localhost ~]$ hisat2-inspect -h
HISAT2 version 2.1.0 by Daehwan Kim (infphilo@gmail.com, http://www.ccb.jhu.edu/people/
infphilo)
Usage: hisat2-inspect [options]* <ht2_base>
  <ht2_base>          ht2 filename minus trailing .1.ht2/.2.ht2

  By default, prints FASTA records of the indexed nucleotide sequences to
  standard out.  With -n, just prints names.  With -s, just prints a summary of
  the index parameters and sequences.  With -e, preserves colors if applicable.

Options:
  --large-index       force inspection of the 'large' index, even if a
                      'small' one is present.
  -a/--across <int>   Number of characters across in FASTA output (default: 60)
  -s/--summary        Print summary incl. ref names, lengths, index properties
  -n/--names          Print reference sequence names only
  --snp               Print SNPs
  --ss                Print splice sites
  --ss-all            Print splice sites including those not in the global index
  --exon              Print exons
  -e/--ht2-ref        Reconstruct reference from .ht2 (slow, preserves colors)
  -v/--verbose        Verbose output (for debugging)
  -h/--help           print detailed description of tool and its options
  --help             _ print this usage message
```

图 68 hisat2-inspect -h 功能的使用

8、hisat2 能够从 FASTA 格式的基因组文件或者 GTF 文件中模拟出 reads，所需要的代码和文件都可以通过 help 功能进行查询（如图 69 和图 70 所示）。

代码：

hisat2_simulate_reads.py -h

#hisat2 模拟 reads 功能帮助。

```
[txy@localhost ~]$ hisat2_simulate_reads.py -h
usage: hisat2_simulate_reads.py [-h] [-d] [--single-end] [-r READ_LEN]
                                [-f FRAG_LEN] [-n NUM_FRAG] [-e EXPR_PROFILE]
                                [--error-rate ERROR_RATE]
                                [--max-mismatch MAX_MISMATCH]
                                [--random-seed RANDOM_SEED] [--sanity-check]
                                [-v] [--version]
                                [genome_file] [gtf_file] [snp_file]
                                [base_fname]

Simulate reads from GENOME (fasta) and GTF files

positional arguments:
  genome_file           input GENOME file
  gtf_file              input GTF file
  snp_file              input SNP file
  base_fname            output base filename

optional arguments:
  -h, --help            show this help message and exit
  -d, --dna             DNA-seq reads (default: RNA-seq reads)
  --single-end          single-end reads (default: paired-end reads)
  -r READ_LEN, --read-length READ_LEN
                        read length (default: 100)
  -f FRAG_LEN, --fragment-length FRAG_LEN
                        fragment length (default: 250)
```

图 69 hisat2_simulate_reads.py -h 功能的使用

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
-n NUM_FRAG, --num-fragment NUM_FRAG
                      number of fragments (default: 1000000)
-e EXPR_PROFILE, --expr-profile EXPR_PROFILE
                      expression profile: flux or constant (default: flux)
--error-rate ERROR_RATE
                      per-base sequencing error rate (%) (default: 0.0)
--max-mismatch MAX_MISMATCH
                      max mismatches due to sequencing errors (default: 3)
--random-seed RANDOM_SEED
                      random seeding value (default: 0)
--sanity-check        sanity check
-v, --verbose         also print some statistics to stderr
--version             show program's version number and exit
```

图 70 hisat2_simulate_reads.py -h 功能的使用

2.3.2 利用示例数据构建索引文件

1、利用示例数据中的 22_20-21M.fa 文件和 22_20-21M.snp 文件构建索引文件。

（1）为了便于管理所构建的索引文件，首先在 HISAT2 夹下创建一个文件夹，命名为“T1xilie”（如图 71-图 73 所示）。

（2）利用示例数据中的 22_20-21M.fa 文件和 22_20-21M.snp 文件构建索引文件（如图 74-图 81 所示）。

代码：

hisat2-build -f /home/txy/HISAT2/hisat2-2.1.0/example/reference/22_20-21M.fa --snp /home/txy/HISAT2/hisat2-2.1.0/example/reference/22_20-21M.snp /home/txy/HISAT2/T1xilie/T1.X.ht2

#“hisat2-build”是构建索引文件的命令。

#“-f”表示输入 FASTA 格式的基因组文件。

#“/home/txy/HISAT2/hisat2-2.1.0/example/reference/22_20-21M.fa”表示输入的基因组文件及其路径，这个文件的内容一般是某个物种的全基因组，较大的基因组用 64 bit nummber，较小的基因组用 32 bit number，系统会自主完成选择。

#“--snp”表示输入 HISAT2 内部使用的“.snp”格式文件。

#“/home/txy/HISAT2/hisat2-2.1.0/example/reference/22_20-21M.snp”表示输入的 snp 文件及其路径，这个文件是陈列了该基因组 SNP 信息的一个 list，包括 SNP ID，SNP 类型（单碱基变换，插入或者删除），染色体名称，SNP 在基因组中基于零点偏移的位置，单碱基变换中的可变碱基，SNP 插入或者删除的序列长度。

#“/home/txy/HISAT2/T1xilie/T1.X.ht2”表示输出索引文件的路径即名字，较大的基因组和较小的基因组两者最终形成的索引文件扩展名不同，较大的基因组是“.ht21”，较小的基因组是“.ht2”。

导用于任何商业用途！
联系方式：binfo@zju.edu.cn

#“X”表示建立的是索引文件。

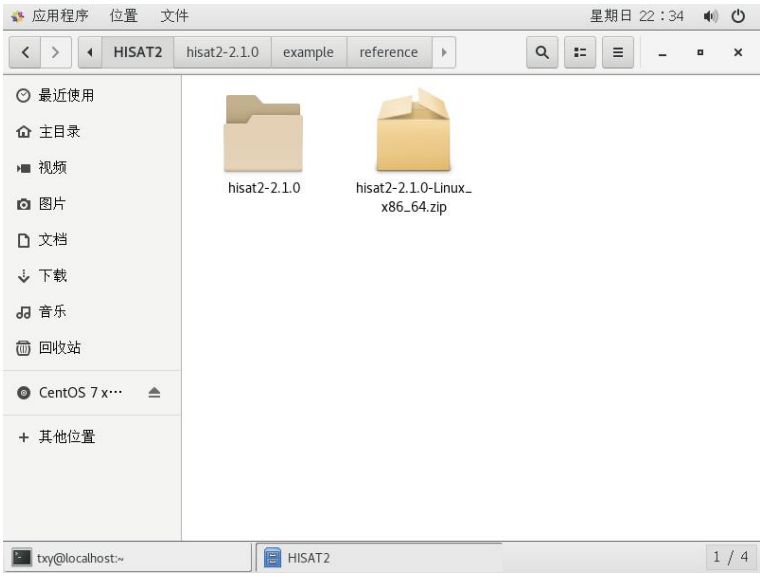


图 71 HISAT2 文件夹

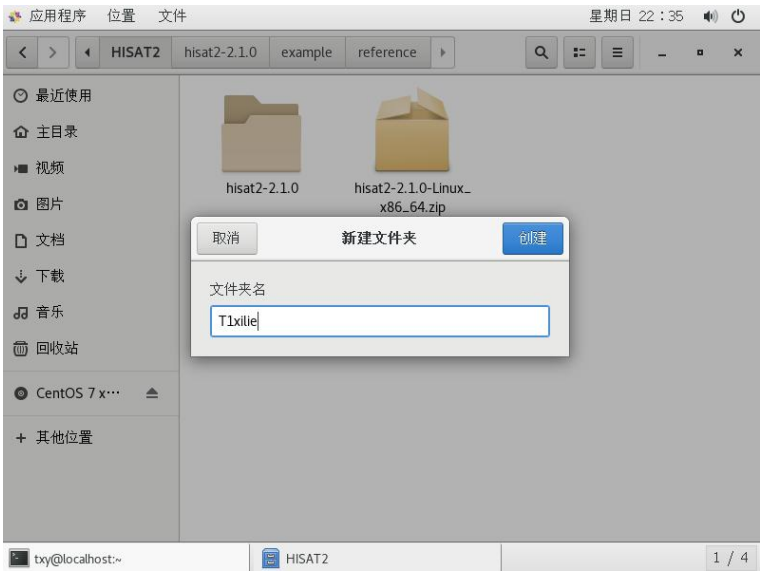


图 72 创建文件夹，命名为“T1xilie”

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

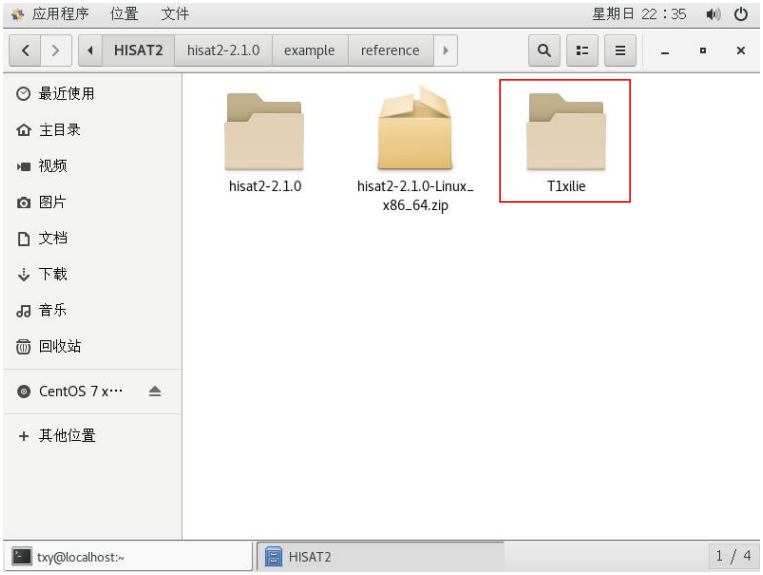


图 73 T1xilie 文件夹

```
[txy@localhost ~]$ hisat2-build -f /home/txy/HISAT2/hisat2-2.1.0/example/reference/22_20-21M.fa --snp /home/txy/HISAT2/hisat2-2.1.0/example/reference/22_20-21M.snp /home/txy/HISAT2/T1xilie/T1.X.ht2
Settings:
  Output files: "/home/txy/HISAT2/T1xilie/T1.X.ht2.*.ht2"
  Line rate: 7 (line is 128 bytes)
  Lines per side: 1 (side is 128 bytes)
  Offset rate: 4 (one in 16)
  FTable chars: 10
  Strings: unpacked
  Local offset rate: 3 (one in 8)
  Local fTable chars: 6
  Local sequence length: 57344
  Local sequence overlap between two consecutive indexes: 1024
  Endianness: little
  Actual local endianness: little
  Sanity checking: disabled
  Assertions: disabled
  Random seed: 0
  Sizeofs: void*:8, int:4, long:8, size_t:8
Input files DNA, FASTA:
  /home/txy/HISAT2/hisat2-2.1.0/example/reference/22_20-21M.fa
Reading reference sizes
  Time reading reference sizes: 00:00:00
Calculating joined length
Writing header
```

图 74 构建索引文件

```
Reserving space for joined string
Joining reference sequences
  Time to join reference sequences: 00:00:00
  Time to read SNPs and splice sites: 00:00:00
Generation 0 (907068 -> 907068 nodes, 0 ranks)
COUNTED NEW NODES: 0
COUNTED TEMP NODES: 0
RESIZED NODES: 0
RESIZED NODES: 0
MADE NEW NODES: 0
Generation 1 (910570 -> 910570 nodes, 0 ranks)
COUNTED NEW NODES: 0
COUNTED TEMP NODES: 0
RESIZED NODES: 0
RESIZED NODES: 0
MADE NEW NODES: 0
Generation 2 (917574 -> 917574 nodes, 0 ranks)
COUNTED NEW NODES: 0
COUNTED TEMP NODES: 0
RESIZED NODES: 0
RESIZED NODES: 0
MADE NEW NODES: 0
Generation 3 (931582 -> 931582 nodes, 0 ranks)
BUILT FROM_INDEX: 0
COUNTED NEW NODES: 0
COUNTED TEMP NODES: 0
```

图 75 构建索引文件

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
RESIZED NODES: 0
RESIZED NODES: 0
MADE NEW NODES: 0
RESIZE NODES: 0
SORT NODES: 0
MERGE, UPDATE RANK: 0
Generation 4 (959602 -> 940657 nodes, 803404 ranks)
ALLOCATE FROM_TABLE: 0
BUILD TABLE: 0
BUILD INDEX: 0
COUNTED NEW NODES: 0
COUNTED TEMP NODES: 0
RESIZED NODES: 0
RESIZED NODES: 0
MADE NEW NODES: 0
MERGEUPDATERANK: 0
TOTAL TIME: 0
Generation 5 (949741 -> 946004 nodes, 883204 ranks)
ALLOCATE FROM_TABLE: 0
BUILD TABLE: 0
BUILD INDEX: 0
COUNTED NEW NODES: 0
COUNTED TEMP NODES: 0
RESIZED NODES: 0
RESIZED NODES: 0
MADE NEW NODES: 0
```

图 76 构建索引文件

```
MERGEUPDATERANK: 0
TOTAL TIME: 0
Generation 6 (953486 -> 950077 nodes, 916044 ranks)
ALLOCATE FROM_TABLE: 0
BUILD TABLE: 0
BUILD INDEX: 0
COUNTED NEW NODES: 0
COUNTED TEMP NODES: 0
RESIZED NODES: 0
RESIZED NODES: 0
MADE NEW NODES: 0
MERGEUPDATERANK: 1
TOTAL TIME: 1
Generation 7 (955700 -> 952624 nodes, 932255 ranks)
ALLOCATE FROM_TABLE: 0
BUILD TABLE: 0
BUILD INDEX: 0
COUNTED NEW NODES: 0
COUNTED TEMP NODES: 0
RESIZED NODES: 0
RESIZED NODES: 0
MADE NEW NODES: 0
MERGEUPDATERANK: 0
TOTAL TIME: 0
Generation 8 (956721 -> 954036 nodes, 944462 ranks)
ALLOCATE FROM_TABLE: 0
```

图 77 构建索引文件

```
BUILD TABLE: 0
BUILD INDEX: 0
COUNTED NEW NODES: 0
COUNTED TEMP NODES: 0
RESIZED NODES: 0
RESIZED NODES: 0
MADE NEW NODES: 0
MERGEUPDATERANK: 0
TOTAL TIME: 0
Generation 9 (956059 -> 954719 nodes, 952732 ranks)
ALLOCATE FROM_TABLE: 0
BUILD TABLE: 0
BUILD INDEX: 0
COUNTED NEW NODES: 0
COUNTED TEMP NODES: 0
RESIZED NODES: 0
RESIZED NODES: 0
MADE NEW NODES: 0
MERGEUPDATERANK: 0
TOTAL TIME: 0
Generation 10 (955034 -> 954774 nodes, 954774 ranks)
Generating edges...
NODE_TO -> GENOME POS: 0
BUILD FROM_INDEX 0
COUNTED NEW EDGES: 0
MADE NEW EDGES: 0
```

图 78 构建索引文件

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
SORTED NEW EDGES: 0
RE-SORTED NODES: 0
PROCESS EDGES: 0
REMOVE Y: 0
SORT, Make index: 0
TOTAL: 0
Allocating ftab, absorbFtab
Entering GFM loop
Exited GFM loop
fchr[A]: 0
fchr[C]: 224989
fchr[G]: 476822
fchr[T]: 729153
fchr[$]: 958358
Exiting GFM::buildToDisk()
Returning from initFromVector
Wrote 4789374 bytes to primary GFM file: /home/txy/HISAT2/T1xilie/T1.X.ht2.1.ht2
Wrote 238700 bytes to secondary GFM file: /home/txy/HISAT2/T1xilie/T1.X.ht2.2.ht2
Re-opening _in1 and _in2 as input streams
Returning from GFM constructor
Returning from initFromVector
Wrote 678405 bytes to primary GFM file: /home/txy/HISAT2/T1xilie/T1.X.ht2.5.ht2
Wrote 238590 bytes to secondary GFM file: /home/txy/HISAT2/T1xilie/T1.X.ht2.6.ht2
Re-opening _in5 and _in5 as input streams
Returning from HierEbwt constructor
Headers:
```

图 79 构建索引文件

```
    len: 900000
    gbwtLen: 958359
    nodes: 954773
    sz: 450000
    gbwtSz: 479180
    lineRate: 7
    offRate: 4
    offMask: 0xfffffff0
    ftabChars: 10
    eftabLen: 0
    eftabSz: 0
    ftabLen: 1048577
    ftabSz: 4194308
    offsLen: 59674
    offsSz: 238696
    lineSz: 128
    sideSz: 128
    sideGbwtSz: 104
    sideGbwtLen: 208
    numSides: 4608
    numLines: 4608
    gbwtTotLen: 589824
    gbwtTotSz: 589824
    reverse: 0
    linearFM: No
Total time for call_to driver() for forward index: 00:00:08
```

图 80 构建索引文件

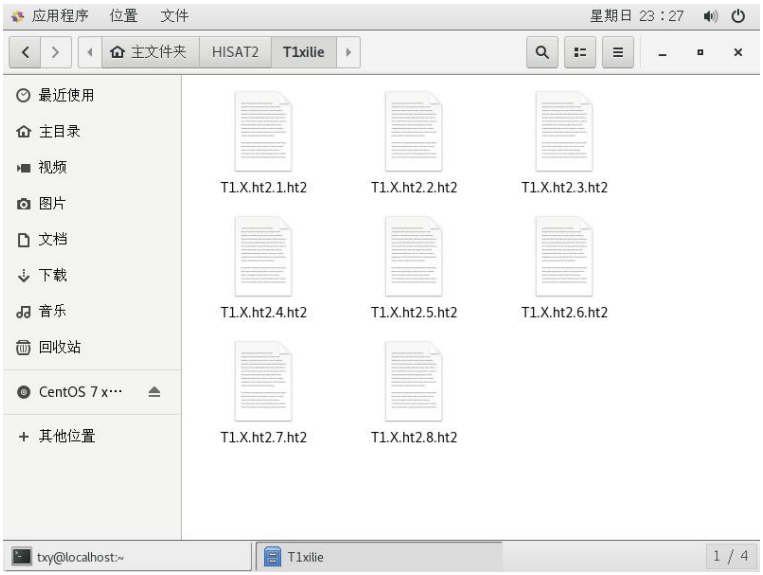


图 81 索引文件构建完毕

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

2、只利用示例数据中的 22_20-21M.fa 文件构建索引文件。

（1）为了便于管理所构建的索引文件，首先在 HISAT2 夹下创建一个文件夹，命名为“T2xilie”（如图 82-图 84 所示）。

（2）利用示例数据中的 22_20-21M.fa 文件构建索引文件（如图 85-图 95 所示）。

代码：

hisat2-build -f /home/txy/HISAT2/hisat2-2.1.0/example/reference/22_20-21M.fa /home/txy/HISAT2/T2xilie/T2.X.ht2

#“hisat2-build”是构建索引文件的命令。

#“-f”表示输入 FASTA 格式的基因组文件。

#“/home/txy/HISAT2/hisat2-2.1.0/example/reference/22_20-21M.fa”表示输入的基因组文件及其路径，这个文件的内容一般是某个物种的全基因组，较大的基因组用 64 bit nummber，较小的基因组用 32 bit number，系统会自主完成选择。

#“/home/txy/HISAT2/T1xilie/T1.X.ht2”表示输出索引文件的路径即名字，较大的基因组和较小的基因组两者最终形成的索引文件扩展名不同，较大的基因组是“.ht21”，较小的基因组是“.ht2”。

#“X”表示建立的是索引文件。

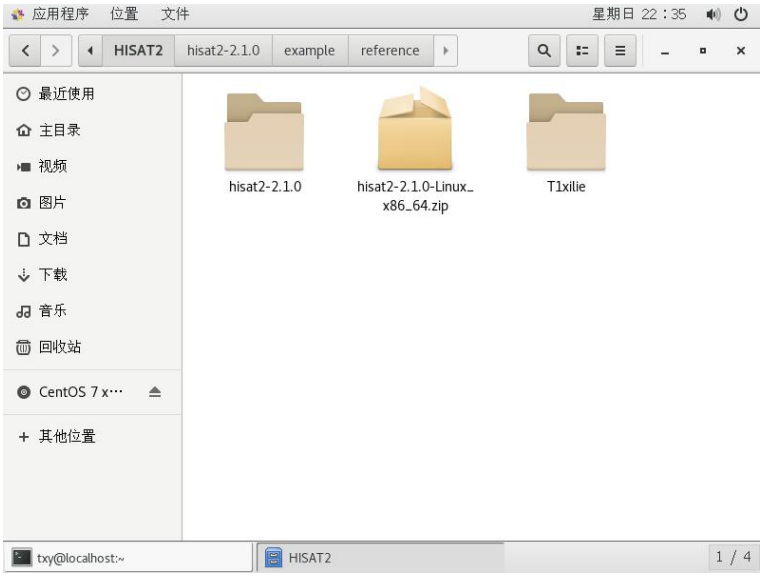


图 82 HISAT2 文件夹

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

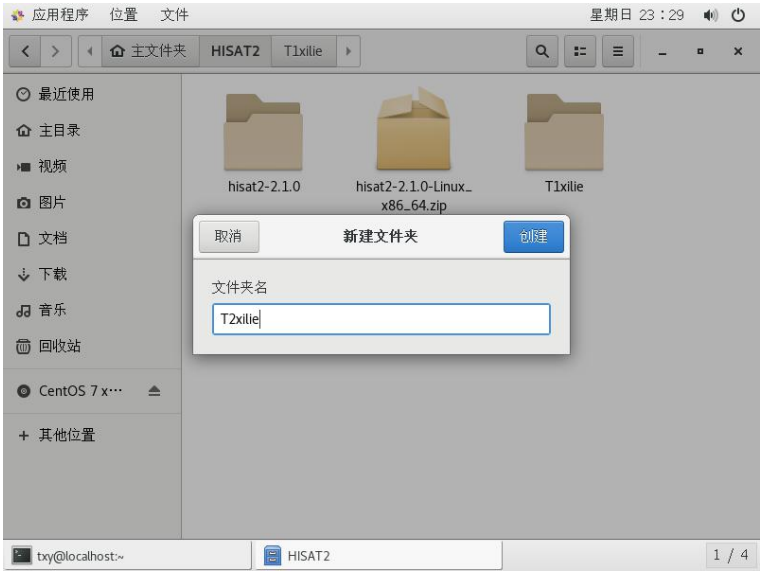


图 83 创建文件夹，命名为“T2xilie”

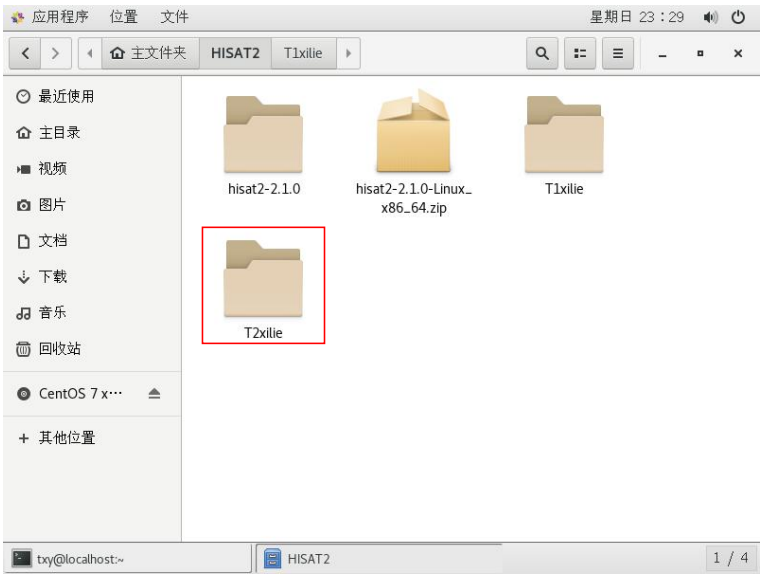


图 84 T2xilie 文件夹

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
[txy@localhost ~]$ hisat2-build -f /home/txy/HISAT2/hisat2-2.1.0/example/reference/22_2
0-21M.fa /home/txy/HISAT2/T2xilie/T2.X.ht2
Settings:
  Output files: "/home/txy/HISAT2/T2xilie/T2.X.ht2.*.ht2"
  Line rate: 6 (line is 64 bytes)
  Lines per side: 1 (side is 64 bytes)
  Offset rate: 4 (one in 16)
  FTable chars: 10
  Strings: unpacked
  Local offset rate: 3 (one in 8)
  Local fTable chars: 6
  Local sequence length: 57344
  Local sequence overlap between two consecutive indexes: 1024
  Endianness: little
  Actual local endianness: little
  Sanity checking: disabled
  Assertions: disabled
  Random seed: 0
  Sizeofs: void*:8, int:4, long:8, size_t:8
Input files DNA, FASTA:
  /home/txy/HISAT2/hisat2-2.1.0/example/reference/22_20-21M.fa
Reading reference sizes
  Time reading reference sizes: 00:00:00
Calculating joined length
Writing header
Reserving space for joined string
```

图 85 构建索引文件

```
Joining reference sequences
  Time to join reference sequences: 00:00:00
  Time to read SNPs and splice sites: 00:00:00
Using parameters --bmax 168750 --dcv 1024
  Doing ahead-of-time memory usage test
  Passed!  Constructing with these parameters: --bmax 168750 --dcv 1024
Constructing suffix-array element generator
Building DifferenceCoverSample
  Building sPrime
  Building sPrimeOrder
  V-Sorting samples
  V-Sorting samples time: 00:00:00
  Allocating rank array
  Ranking v-sort output
  Ranking v-sort output time: 00:00:00
  Invoking Larsson-Sadakane on ranks
  Invoking Larsson-Sadakane on ranks time: 00:00:00
  Sanity-checking and returning
Building samples
Reserving space for 12 sample suffixes
Generating random suffixes
QSorting 12 sample offsets, eliminating duplicates
QSorting sample offsets, eliminating duplicates time: 00:00:00
Multikey QSorting 12 samples
  (Using difference cover)
  Multikey QSorting samples time: 00:00:00
```

图 86 构建索引文件

```
Calculating bucket sizes
Splitting and merging
  Splitting and merging time: 00:00:00
Split 1, merged 6; iterating...
Splitting and merging
  Splitting and merging time: 00:00:00
Avg bucket size: 112499 (target: 168749)
Converting suffix-array elements to index image
Allocating ftab, absorbFtab
Entering GFM loop
Getting block 1 of 8
  Reserving size (168750) for bucket 1
  Calculating Z arrays for bucket 1
  Entering block accumulator loop for bucket 1:
  bucket 1: 10%
  bucket 1: 20%
  bucket 1: 30%
  bucket 1: 40%
  bucket 1: 50%
  bucket 1: 60%
  bucket 1: 70%
  bucket 1: 80%
  bucket 1: 90%
  bucket 1: 100%
  Sorting block of length 147689 for bucket 1
  (Using difference cover)
```

图 87 构建索引文件

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
  Sorting block time: 00:00:00
Returning block of 147690 for bucket 1
Getting block 2 of 8
  Reserving size (168750) for bucket 2
  Calculating Z arrays for bucket 2
  Entering block accumulator loop for bucket 2:
  bucket 2: 10%
  bucket 2: 20%
  bucket 2: 30%
  bucket 2: 40%
  bucket 2: 50%
  bucket 2: 60%
  bucket 2: 70%
  bucket 2: 80%
  bucket 2: 90%
  bucket 2: 100%
  Sorting block of length 46596 for bucket 2
  (Using difference cover)
  Sorting block time: 00:00:00
Returning block of 46597 for bucket 2
Getting block 3 of 8
  Reserving size (168750) for bucket 3
  Calculating Z arrays for bucket 3
  Entering block accumulator loop for bucket 3:
  bucket 3: 10%
  bucket 3: 20%
```

图 88 构建索引文件

```
  bucket 3: 30%
  bucket 3: 40%
  bucket 3: 50%
  bucket 3: 60%
  bucket 3: 70%
  bucket 3: 80%
  bucket 3: 90%
  bucket 3: 100%
  Sorting block of length 147533 for bucket 3
  (Using difference cover)
  Sorting block time: 00:00:00
Returning block of 147534 for bucket 3
Getting block 4 of 8
  Reserving size (168750) for bucket 4
  Calculating Z arrays for bucket 4
  Entering block accumulator loop for bucket 4:
  bucket 4: 10%
  bucket 4: 20%
  bucket 4: 30%
  bucket 4: 40%
  bucket 4: 50%
  bucket 4: 60%
  bucket 4: 70%
  bucket 4: 80%
  bucket 4: 90%
  bucket 4: 100%
```

图 89 构建索引文件

```
  Sorting block of length 153159 for bucket 4
  (Using difference cover)
  Sorting block time: 00:00:01
Returning block of 153160 for bucket 4
Getting block 5 of 8
  Reserving size (168750) for bucket 5
  Calculating Z arrays for bucket 5
  Entering block accumulator loop for bucket 5:
  bucket 5: 10%
  bucket 5: 20%
  bucket 5: 30%
  bucket 5: 40%
  bucket 5: 50%
  bucket 5: 60%
  bucket 5: 70%
  bucket 5: 80%
  bucket 5: 90%
  bucket 5: 100%
  Sorting block of length 154613 for bucket 5
  (Using difference cover)
  Sorting block time: 00:00:00
Returning block of 154614 for bucket 5
Getting block 6 of 8
  Reserving size (168750) for bucket 6
  Calculating Z arrays for bucket 6
  Entering block accumulator loop for bucket 6:
```

图 90 构建索引文件

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
  bucket 6: 10%
  bucket 6: 20%
  bucket 6: 30%
  bucket 6: 40%
  bucket 6: 50%
  bucket 6: 60%
  bucket 6: 70%
  bucket 6: 80%
  bucket 6: 90%
  bucket 6: 100%
  Sorting block of length 37160 for bucket 6
  (Using difference cover)
  Sorting block time: 00:00:00
Returning block of 37161 for bucket 6
Getting block 7 of 8
  Reserving size (168750) for bucket 7
  Calculating Z arrays for bucket 7
  Entering block accumulator loop for bucket 7:
  bucket 7: 10%
  bucket 7: 20%
  bucket 7: 30%
  bucket 7: 40%
  bucket 7: 50%
  bucket 7: 60%
  bucket 7: 70%
  bucket 7: 80%
```

图 91 构建索引文件

```
  bucket 7: 90%
  bucket 7: 100%
  Sorting block of length 145776 for bucket 7
  (Using difference cover)
  Sorting block time: 00:00:00
Returning block of 145777 for bucket 7
Getting block 8 of 8
  Reserving size (168750) for bucket 8
  Calculating Z arrays for bucket 8
  Entering block accumulator loop for bucket 8:
  bucket 8: 10%
  bucket 8: 20%
  bucket 8: 30%
  bucket 8: 40%
  bucket 8: 50%
  bucket 8: 60%
  bucket 8: 70%
  bucket 8: 80%
  bucket 8: 90%
  bucket 8: 100%
  Sorting block of length 67467 for bucket 8
  (Using difference cover)
  Sorting block time: 00:00:00
Returning block of 67468 for bucket 8
Exited GFM loop
fchr[A]: 0
```

图 92 构建索引文件

```
fchr[C]: 211155
fchr[G]: 447355
fchr[T]: 684467
fchr[$]: 900000
Exiting GFM::buildToDisk()
Returning from initFromVector
Wrote 4494550 bytes to primary GFM file: /home/txy/HISAT2/T2xilie/T2.X.ht2.1.ht2
Wrote 225008 bytes to secondary GFM file: /home/txy/HISAT2/T2xilie/T2.X.ht2.2.ht2
Re-opening _in1 and _in2 as input streams
Returning from GFM constructor
Returning from initFromVector
Wrote 411693 bytes to primary GFM file: /home/txy/HISAT2/T2xilie/T2.X.ht2.5.ht2
Wrote 229134 bytes to secondary GFM file: /home/txy/HISAT2/T2xilie/T2.X.ht2.6.ht2
Re-opening _in5 and _in5 as input streams
Returning from HierEbwt constructor
Headers:
    len: 900000
    gbwtLen: 900001
    nodes: 900001
    sz: 225000
    gbwtSz: 225001
    lineRate: 6
    offRate: 4
    offMask: 0xfffffff0
    ftabChars: 10
    eftabLen: 0
```

图 93 构建索引文件

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
    eftabSz: 0
    ftabLen: 1048577
    ftabSz: 4194308
    offsLen: 56251
    offsSz: 225004
    lineSz: 64
    sideSz: 64
    sideGbwtSz: 48
    sideGbwtLen: 192
    numSides: 4688
    numLines: 4688
    gbwtTotLen: 300032
    gbwtTotSz: 300032
    reverse: 0
    linearFM: Yes
Total time for call to driver() for forward index: 00:00:01
```

图 94 构建索引文件

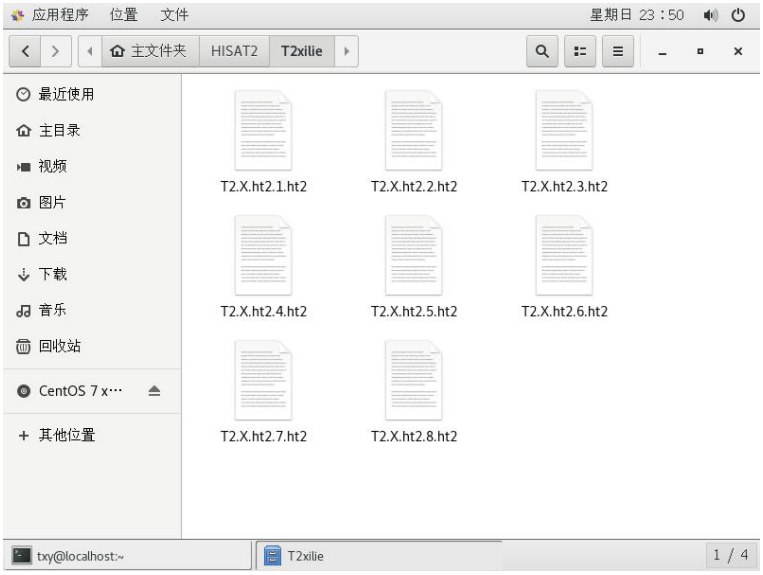


图 95 构建索引文件

2.3.3 利用示例数据进行 map

1、利用示例数据中的 reads_1.fa 文件和 reads_2.fa 文件与之前构建好的索引文件（这里使用 22_20-21M.fa 文件和 22_20-21M.snp 文件构建的索引文件）进行 paired-end map，得到“.sam”格式文件（如图 96-图 98 所示）。

代码：

```
hisat2 -f -x /home/txy/HISAT2/hisat2-2.1.0/example/index/22_20-21M_snp -1 /home/txy/HISAT2/hisat2-2.1.0/example/reads/reads_1.fa -2 /home/txy/HISAT2/hisat2-2.1.0/example/reads/reads_2.fa -S T1.sam
#“hisat2”是将 reads 与索引文件进行 map 的命令。
#“-f”表示输入的文件是 FASTA 格式。
#“-x”表示 map 使用的索引文件。
```

得用于任何商业用途！
联系方式：binfo@zju.edu.cn

#"/home/txy/HISAT2/hisat2-2.1.0/example/index/22_20-21M_snp"表示输入的snp 文件及其路径。

#"-1"和"-2"表示同时使用的read_1.fa 文件和read_2.fa 文件的编号。

#"/home/txy/HISAT2/hisat2-2.1.0/example/reads/reads_1.fa" 和"/home/txy/HISAT2/hisat2-2.1.0/example/reads/reads_2.fa"表示输入的read_1.fa 文件及其路径和read_2.fa 文件及其路径，这两个文件汇集了构建索引所使用的基因组文件的多种可变剪接的。

#"-S T1.sam"表示输出的sam 文件的名称，默认的输出位置是主目录。

```
[txy@localhost ~]$ hisat2 -f -x /home/txy/HISAT2/hisat2-2.1.0/example/index/22_20-21M_s
np -1 /home/txy/HISAT2/hisat2-2.1.0/example/reads/reads_1.fa -2 /home/txy/HISAT2/hisat2
-2.1.0/example/reads/reads_2.fa -S T1.sam
1000 reads; of these:
  1000 (100.00%) were paired; of these:
    0 (0.00%) aligned concordantly 0 times
    1000 (100.00%) aligned concordantly exactly 1 time
    0 (0.00%) aligned concordantly >1 times
    ----
    0 pairs aligned concordantly 0 times; of these:
      0 (0.00%) aligned discordantly 1 time
    ----
    0 pairs aligned 0 times concordantly or discordantly; of these:
      0 mates make up the pairs; of these:
        0 (0.00%) aligned 0 times
        0 (0.00%) aligned exactly 1 time
        0 (0.00%) aligned >1 times
100.00% overall alignment rate
```

图 96 paired-end map

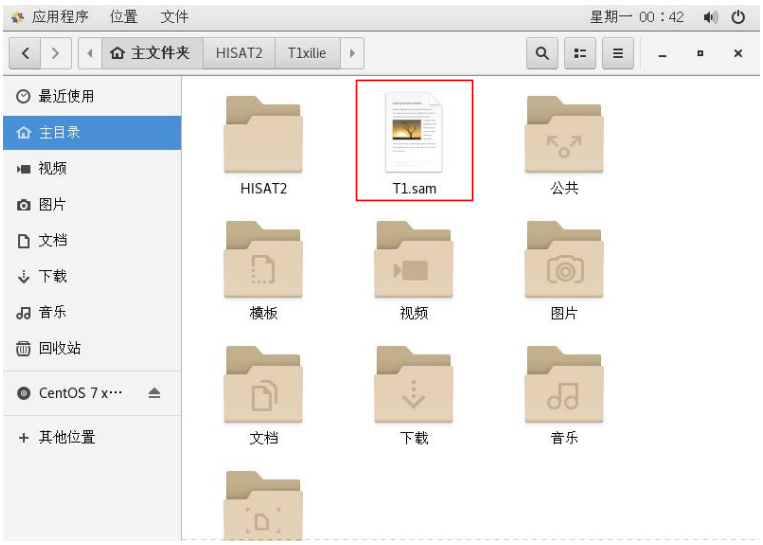


图 97 paired-end map 得到的".sam"格式文件

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

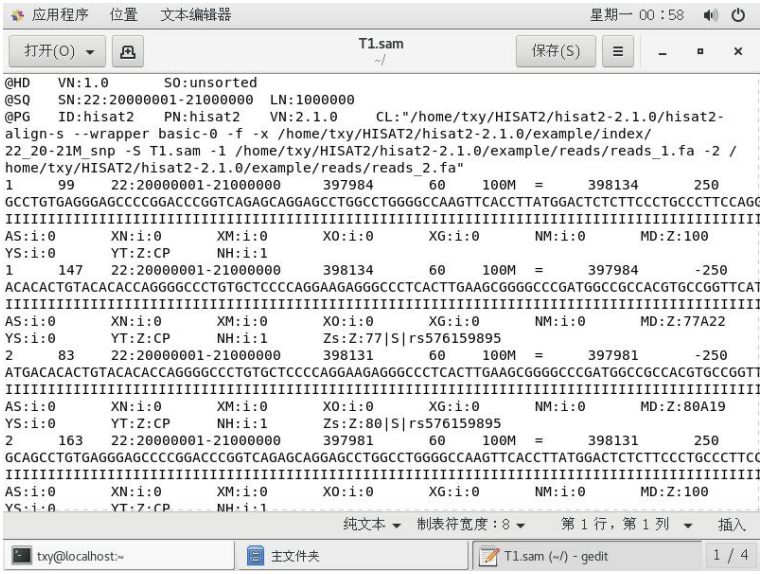


图 98 T1.sam 文件

2、利用示例数据中的 reads_1.fa 文件与之前构建好的索引文件（这里使用 22_20-21M.fa 文件和 22_20-21M.snp 文件构建的索引文件）进行 single-end map，得到“.sam”格式文件（如图 99-图 101 所示）。

代码：

hisat2 -f -x /home/txy/HISAT2/hisat2-2.1.0/example/index/22_20-21M_snp -U /home/txy/HISAT2/hisat2-2.1.0/example/reads/reads_1.fa -S T2.sam

#“hisat2”是将 reads 与索引文件进行 map 的命令。

#“-f”表示输入的文件是 FASTA 格式。

#“-x”表示 map 使用的索引文件。

#“/home/txy/HISAT2/hisat2-2.1.0/example/index/22_20-21M_snp”表示输入的snp 文件及其路径。“-U”表示单独使用的 read_1.fa 文件的编号。

#“/home/txy/HISAT2/hisat2-2.1.0/example/reads/reads_1.fa”表示输入的 read_1.fa 文件及其路径，这个文件汇集了构建索引所使用的基因组文件的多种可变剪接的。

#“-S T2.sam”表示输出的 sam 文件的名称，默认的输出位置是主目录。

```
[txy@localhost ~]$ hisat2 -f -x /home/txy/HISAT2/hisat2-2.1.0/example/index/22_20-21M_s
np -U /home/txy/HISAT2/hisat2-2.1.0/example/reads/reads_1.fa -S T2.sam
1000 reads; of these:
  1000 (100.00%) were unpaired; of these:
    0 (0.00%) aligned 0 times
    1000 (100.00%) aligned exactly 1 time
    0 (0.00%) aligned >1 times
100.00% overall alignment rate
```

联系方式：binfo@zju.edu.cn

图 99 single-end map

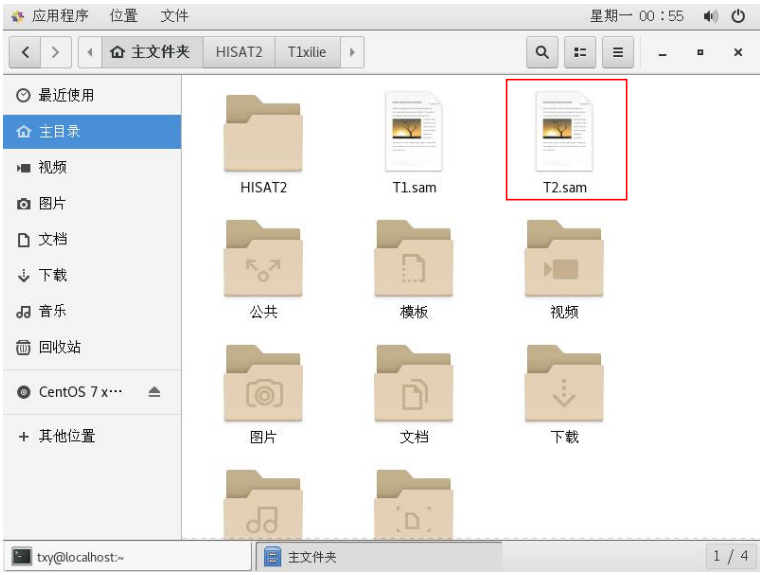


图 100 paired-end map 得到的“.sam”格式文件

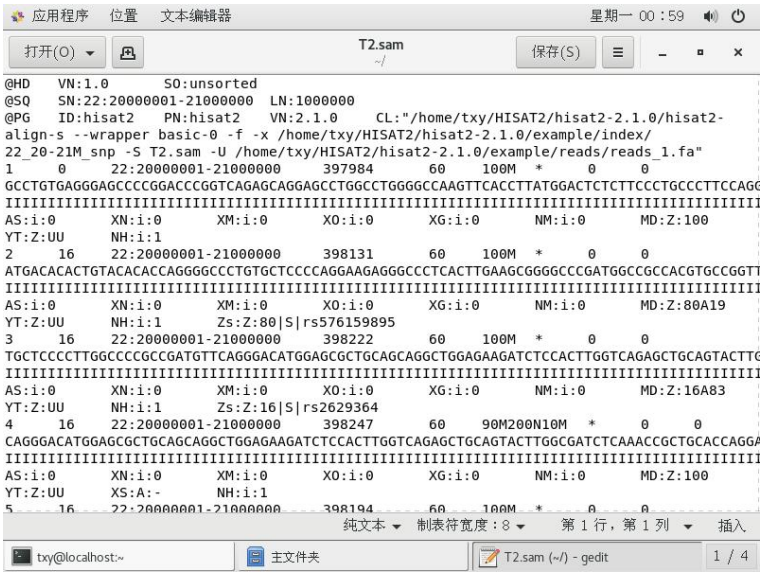


图 101 T2.sam 文件

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

3 从 sam 格式文件到 bam 格式文件——samtools

3.1 关于 samtools

samtools 是用于处理 sam 与 bam 格式的一系列应用程序的集合，能够实现二进制查看、格式转换、排序及合并等功能，结合 sam 格式中的 flag、tag 等信息，还可以完成比对结果的统计汇总。此外，利用 linux 中的 grep、awk 等操作命令，还可以大大扩展 samtools 的使用范围与功能。这里我们主要使用 samtools 的转换功能，实现 sam 格式文件和 bam 格式文件的相互转换，以方便后续的分析。

sam 文件为纯文本格式，存储了测序获得的信息，以及 reads map 到基因组后的各种信息，压缩了极大的信息量。而 bam 文件则是 sam 文件的二进制版，在 sam 文件的基础上运用二进制编码，又极大的压缩了 sam 文件的体积。

3.2 samtools 的安装

3.2.1 samtools 的下载和解压缩

1、为了便于管理所安装的软件，首先在主文件夹下创建一个文件夹，命名为“samtools”（如图 102-图 104 所示）。

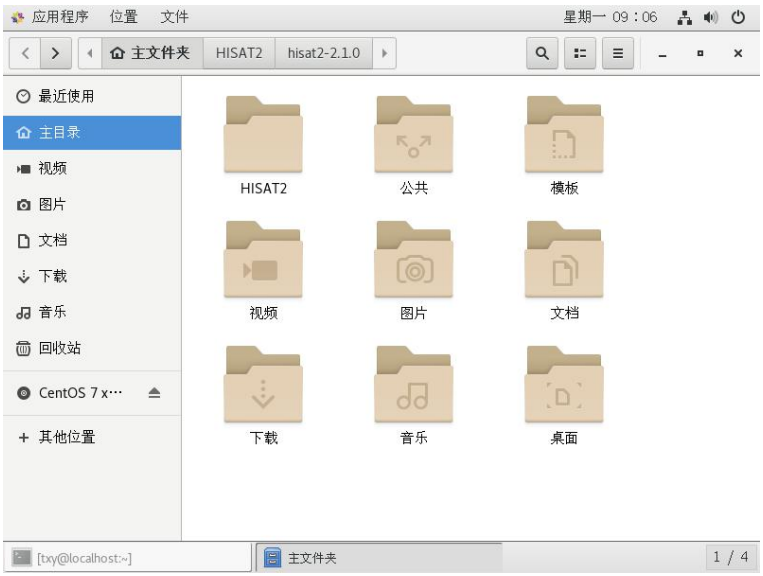


图 102 主文件夹

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

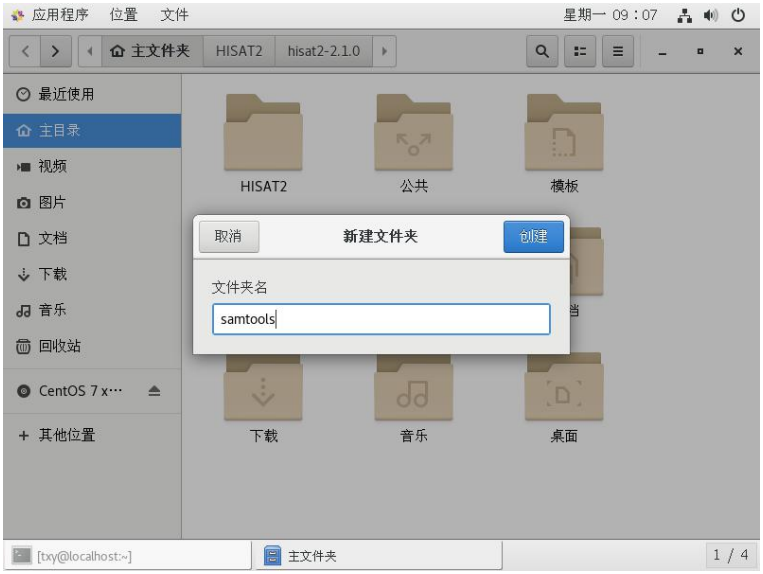


图 103 创建文件夹，命名为“samtools”

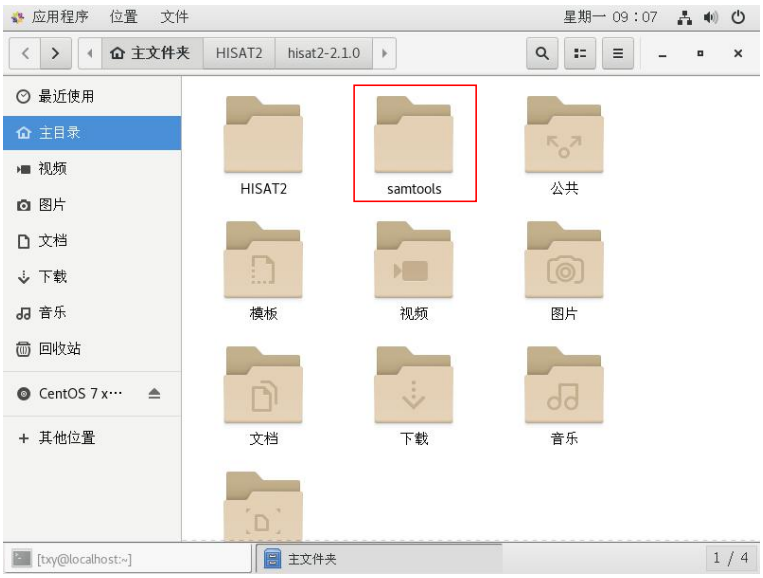


图 104 samtools 文件夹

2、下载 samtools（如图 105 和图 106 所示）。

代码：

```
wget https://github.com/samtools/samtools/releases/download/1.9/samtools-1.9.tar.bz2
# 下载 samtools。
```

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
[txy@localhost ~]$ wget https://github.com/samtools/samtools/releases/download/1.9/samtools-1.9.tar.bz2
--2019-04-01 09:02:47--  https://github.com/samtools/samtools/releases/download/1.9/samtools-1.9.tar.bz2
正在解析主机 github.com (github.com)... 52.74.223.119, 13.250.177.223, 13.229.188.59
正在连接 github.com (github.com)|52.74.223.119|:443... 已连接。
已发出 HTTP 请求，正在等待回应... 302 Found
位置：https://github-production-release-asset-2e65be.s3.amazonaws.com/3666841/fe586164-8a73-11e8-84ad-bb90bbd3b7c0?X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIAIWNJYAX4CSVEH53A%2F20190401%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Date=20190401T010252Z&X-Amz-Expires=300&X-Amz-Signature=ef26cc4fb343a30c6bf209f1c2ff51c73d4a7b5ca3f82d759818f67822ac713c&X-Amz-SignedHeaders=host&actor_id=0&response-content-disposition=attachment%3B%20filename%3Dsamtools-1.9.tar.bz2&response-content-type=application%2Foctet-stream [跟随至新的 URL]
--2019-04-01 09:02:48--  https://github-production-release-asset-2e65be.s3.amazonaws.com/3666841/fe586164-8a73-11e8-84ad-bb90bbd3b7c0?X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIAIWNJYAX4CSVEH53A%2F20190401%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Date=20190401T010252Z&X-Amz-Expires=300&X-Amz-Signature=ef26cc4fb343a30c6bf209f1c2ff51c73d4a7b5ca3f82d759818f67822ac713c&X-Amz-SignedHeaders=host&actor_id=0&response-content-disposition=attachment%3B%20filename%3Dsamtools-1.9.tar.bz2&response-content-type=application%2Foctet-stream
正在解析主机 github-production-release-asset-2e65be.s3.amazonaws.com (github-production-release-asset-2e65be.s3.amazonaws.com)... 52.216.179.131
正在连接 github-production-release-asset-2e65be.s3.amazonaws.com (github-production-release-asset-2e65be.s3.amazonaws.com)|52.216.179.131|:443... 已连接。
已发出 HTTP 请求，正在等待回应... 200 OK
```

图 105 下载 samtools

```
长度：4440405 (4.2M) [application/octet-stream]
正在保存至："samtools-1.9.tar.bz2"

100%[==========================================>] 4,440,405   29.6KB/s 用时 2m 8s

2019-04-01 09:04:58 (33.9 KB/s) - 已保存 "samtools-1.9.tar.bz2" [4440405/4440405])
```

图 106 下载 samtools

3、解压缩 samtools（如图 107 所示）

代码：

```
tar -jxf /home/txy/samtools-1.9.tar.bz2
#解压缩 samtools。
```

```
[txy@localhost ~]$ tar -jxf /home/txy/samtools-1.9.tar.bz2
[txy@localhost ~]$ _
```

图 107 解压缩 samtools

4、解压缩后的 samtools 文件和 samtools 压缩包都位于主文件夹中（如图 108 所示），为了便于管理，将它们一起放入先前建好的 samtools 文件夹中（如图 109 所示）。

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

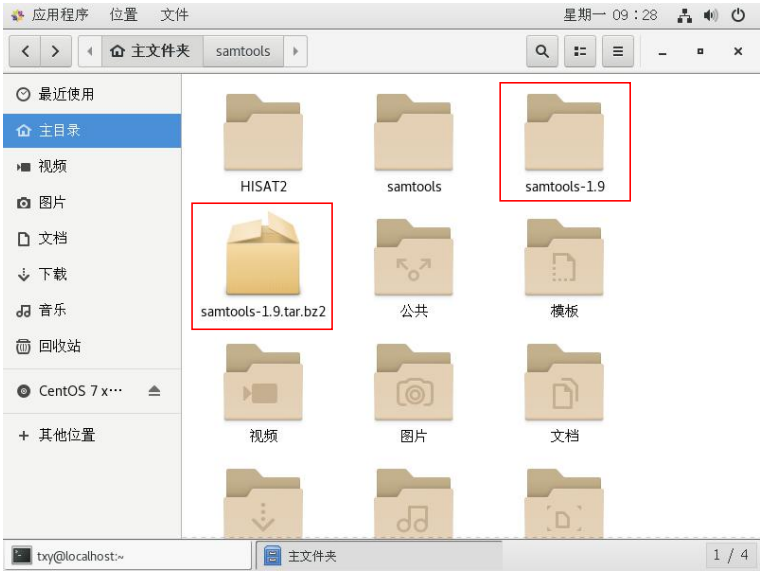


图 108 解压缩后的 samtools 文件和 samtools 压缩包

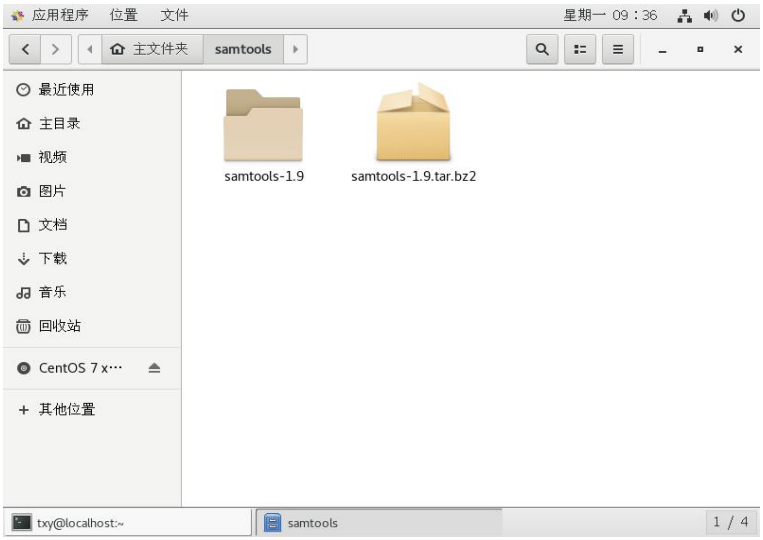


图 109 将解压缩后的 samtools 文件和 samtools 压缩包放入文件夹 HISAT2 中

3.2.2 码源包的安装

1、安装 samtools 需要安装码源包 ncurses-devel（如图 110-图 112 所示）。

代码：

```
sudo yum install ncurses-devel
#安装码源包 ncurses-devel。
```

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
[txy@localhost ~]$ sudo yum install ncurses-devel
[sudo] txy 的密码：
已加载插件：fastestmirror, langpacks
Loading mirror speeds from cached hostfile
 * base: mirrors.163.com
 * extras: mirrors.zju.edu.cn
 * updates: mirrors.zju.edu.cn
base                                                     | 3.6 kB  00:00:00
extras                                                   | 3.4 kB  00:00:00
updates                                                  | 3.4 kB  00:00:00
正在解决依赖关系
--> 正在检查事务
---> 软件包 ncurses-devel.x86_64.0.5.9-14.20130511.el7_4 将被 安装
--> 解决依赖关系完成

依赖关系解决

================================================================================
 Package              架构            版本                       源          大小
================================================================================
正在安装:
 ncurses-devel        x86_64          5.9-14.20130511.el7_4      base        712 k

事务概要
================================================================================
安装  1 软件包
```

图 110 安装码源包 ncurses-devel

```
总下载量：712 k
安装大小：2.1 M
Is this ok [y/d/N]: y
Downloading packages:
No Presto metadata available for base
警告：/var/cache/yum/x86_64/7/base/packages/ncurses-devel-5.9-14.20130511.el7_4.x86_64.
rpm: 头V3 RSA/SHA256 Signature, 密钥 ID f4a80eb5: NOKEY
ncurses-devel-5.9-14.20130511.el7_4.x86_64.rpm 的公钥尚未安装
ncurses-devel-5.9-14.20130511.el7_4.x86_64.rpm                | 712 kB  00:00:00
从 file:///etc/pki/rpm-gpg/RPM-GPG-KEY-CentOS-7 检索密钥
导入 GPG key 0xF4A80EB5:
 用户ID     : "CentOS-7 Key (CentOS 7 Official Signing Key) <security@centos.org>"
 指纹       : 6341 ab27 53d7 8a78 a7c2 7bb1 24c6 a8a7 f4a8 0eb5
 软件包     : centos-release-7-6.1810.2.el7.centos.x86_64 (@anaconda)
 来自       : /etc/pki/rpm-gpg/RPM-GPG-KEY-CentOS-7
是否继续？[y/N]：y
Running transaction check
Running transaction test
Transaction test succeeded
Running transaction
  正在安装    : ncurses-devel-5.9-14.20130511.el7_4.x86_64                  1/1
  验证中      : ncurses-devel-5.9-14.20130511.el7_4.x86_64                  1/1

已安装:
  ncurses-devel.x86_64 0:5.9-14.20130511.el7_4
```

图 111 安装码源包 ncurses-devel

```
完毕！
```

图 112 安装码源包 ncurses-devel

2、安装 samtools 需要安装码源包 zlib-devel（如图 113 和图 114 所示）。

代码：

```
sudo yum install zlib-devel
#安装码源包 zlib-devel。
```

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
[txy@localhost ~]$ sudo yum install zlib-devel
[sudo] txy 的密码：
已加载插件：fastestmirror, langpacks
Loading mirror speeds from cached hostfile
 * base: mirrors.163.com
 * extras: mirrors.zju.edu.cn
 * updates: mirrors.zju.edu.cn
base                                                     | 3.6 kB  00:00:00
extras                                                   | 3.4 kB  00:00:00
updates                                                  | 3.4 kB  00:00:00
(1/2): extras/7/x86_64/primary_db                        | 187 kB  00:00:00
(2/2): updates/7/x86_64/primary_db                       | 3.4 MB  00:00:12
正在解决依赖关系
--> 正在检查事务
---> 软件包 zlib-devel.x86_64.0.1.2.7-18.el7 将被 安装
--> 解决依赖关系完成

依赖关系解决

================================================================================
 Package          架构          版本                源              大小
================================================================================
正在安装:
 zlib-devel       x86_64        1.2.7-18.el7        base            50 k

事务概要
```

图 113 安装码源包 zlib-devel

```
================================================================================
安装  1 软件包

总下载量：50 k
安装大小：132 k
Is this ok [y/d/N]: y
Downloading packages:
zlib-devel-1.2.7-18.el7.x86_64.rpm                       |  50 kB  00:00:00
Running transaction check
Running transaction test
Transaction test succeeded
Running transaction
  正在安装    : zlib-devel-1.2.7-18.el7.x86_64                              1/1
  验证中      : zlib-devel-1.2.7-18.el7.x86_64                              1/1

已安装:
  zlib-devel.x86_64 0:1.2.7-18.el7

完毕！
```

图 114 安装码源包 zlib-devel

3、安装 samtools 需要安装码源包 bzip-devel（如图 115 和图 116 所示）。

代码：

sudo yum install bzip2-devel

#安装码源包 bzip-devel。

```
[txy@localhost ~]$ sudo yum install bzip2-devel
已加载插件：fastestmirror, langpacks
Loading mirror speeds from cached hostfile
 * base: mirrors.163.com
 * extras: mirrors.zju.edu.cn
 * updates: mirrors.zju.edu.cn
正在解决依赖关系
--> 正在检查事务
---> 软件包 bzip2-devel.x86_64.0.1.0.6-13.el7 将被 安装
--> 解决依赖关系完成

依赖关系解决

================================================================================
 Package          架构          版本                源              大小
================================================================================
正在安装:
 bzip2-devel      x86_64        1.0.6-13.el7        base            218 k

事务概要
================================================================================
安装  1 软件包

总下载量：218 k
安装大小：382 k
Is this ok [y/d/N]: y
```

图 115 安装码源包 bzip-devel

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
Downloading packages:
bzip2-devel-1.0.6-13.el7.x86_64.rpm                          | 218 kB  00:00:00
Running transaction check
Running transaction test
Transaction test succeeded
Running transaction
  正在安装    : bzip2-devel-1.0.6-13.el7.x86_64                              1/1
  验证中      : bzip2-devel-1.0.6-13.el7.x86_64                              1/1

已安装:
  bzip2-devel.x86_64 0:1.0.6-13.el7

完毕！
```

图 116 安装码源包 bzip-devel

4、安装 samtools 需要安装码源包 xz-devel（如图 117 和图 118 所示）。

代码：

sudo yum install xz-devel

#安装码源包 xz-devel。

```
[txy@localhost ~]$ sudo yum install xz-devel
已加载插件：fastestmirror, langpacks
Loading mirror speeds from cached hostfile
 * base: mirrors.163.com
 * extras: mirrors.zju.edu.cn
 * updates: mirrors.zju.edu.cn
正在解决依赖关系
--> 正在检查事务
---> 软件包 xz-devel.x86_64.0.5.2.2-1.el7 将被 安装
--> 解决依赖关系完成

依赖关系解决

================================================================================
 Package          架构          版本                源          大小
================================================================================
正在安装:
 xz-devel         x86_64        5.2.2-1.el7         base        46 k

事务概要
================================================================================
安装  1 软件包

总下载量：46 k
安装大小：165 k
Is this ok [y/d/N]: y
```

图 117 安装码源包 xz-devel

```
Downloading packages:
xz-devel-5.2.2-1.el7.x86_64.rpm                              |  46 kB  00:00:00
Running transaction check
Running transaction test
Transaction test succeeded
Running transaction
  正在安装    : xz-devel-5.2.2-1.el7.x86_64                                  1/1
  验证中      : xz-devel-5.2.2-1.el7.x86_64                                  1/1

已安装:
  xz-devel.x86_64 0:5.2.2-1.el7

完毕！
```

图 118 安装码源包 xz-devel

3.2.3 samtools 的安装

1、安装 samtools-1.9（如图 119-图 129 所示）。

代码：

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

cd samtools-1.9

./configure --prefix=/home/txy/samtools

make

make install

#安装 samtools-1.9。

```
[txy@localhost ~]$ cd samtools-1.9
bash: cd: samtools-1.9: 没有那个文件或目录
[txy@localhost ~]$ cd
[txy@localhost ~]$ cd /home/txy/samtools/samtools-1.9/
[txy@localhost samtools-1.9]$ ./configure --prefix=/home/txy/samtools
checking for gcc... gcc
checking whether the C compiler works... yes
checking for C compiler default output file name... a.out
checking for suffix of executables...
checking whether we are cross compiling... no
checking for suffix of object files... o
checking whether we are using the GNU C compiler... yes
checking whether gcc accepts -g... yes
checking for gcc option to accept ISO C89... none needed
checking for special C compiler options needed for large files... no
checking for _FILE_OFFSET_BITS value needed for large files... no
checking location of HTSlib source tree... htslib-1.9
checking for NcursesW wide-character library... yes
checking for working ncursesw/curses.h... yes
checking for working ncursesw.h... no
checking for working ncurses.h... yes
checking for zlib.h... yes
checking for inflate in -lz... yes
checking for library containing regcomp... none required
configure: creating ./config.status
config.status: creating config.mk
```

图 119 安装 samtools-1.9

```
config.status: creating config.h
=== configuring in htslib-1.9 (/home/txy/samtools/samtools-1.9/htslib-1.9)
configure: running /bin/sh ./configure --disable-option-checking '--prefix=/home/txy/sa
mtools'  --cache-file=/dev/null --srcdir=.
checking for gcc... gcc
checking whether the C compiler works... yes
checking for C compiler default output file name... a.out
checking for suffix of executables...
checking whether we are cross compiling... no
checking for suffix of object files... o
checking whether we are using the GNU C compiler... yes
checking whether gcc accepts -g... yes
checking for gcc option to accept ISO C89... none needed
checking for ranlib... ranlib
checking for grep that handles long lines and -e... /usr/bin/grep
checking for C compiler warning flags... -Wall
checking for special C compiler options needed for large files... no
checking for _FILE_OFFSET_BITS value needed for large files... no
checking for _LARGEFILE_SOURCE value needed for large files... no
checking shared library type for unknown-Linux... plain .so
checking how to run the C preprocessor... gcc -E
checking for egrep... /usr/bin/grep -E
checking for ANSI C header files... yes
checking for sys/types.h... yes
checking for sys/stat.h... yes
checking for stdlib.h... yes
```

图 120 安装 samtools-1.9

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
checking for string.h... yes
checking for memory.h... yes
checking for strings.h... yes
checking for inttypes.h... yes
checking for stdint.h... yes
checking for unistd.h... yes
checking for stdlib.h... (cached) yes
checking for unistd.h... (cached) yes
checking for sys/param.h... yes
checking for getpagesize... yes
checking for working mmap... yes
checking for gmtime_r... yes
checking for fsync... yes
checking for drand48... yes
checking whether fdatasync is declared... yes
checking for fdatasync... yes
checking for library containing log... -lm
checking for zlib.h... yes
checking for inflate in -lz... yes
checking for library containing recv... none required
checking for bzlib.h... yes
checking for BZ2_bzBuffToBuffCompress in -lbz2... yes
checking for lzma.h... yes
checking for lzma_easy_buffer_encode in -llzma... yes
checking for libdeflate.h... no
checking for libdeflate_deflate_compress in -ldeflate... no
```

图 121 安装 samtools-1.9

```
checking for curl_easy_pause in -lcurl... no
checking for curl_easy_init in -lcurl... no
configure: WARNING: libcurl not enabled: library not found
configure: WARNING: GCS support not enabled: requires libcurl support
configure: WARNING: S3 support not enabled: requires libcurl support
checking whether PTHREAD_MUTEX_RECURSIVE is declared... yes
configure: creating ./config.status
config.status: creating config.mk
config.status: creating htslib.pc.tmp
config.status: creating config.h
[txy@localhost samtools-1.9]$ make
config.mk:45: htslib-1.9/htslib_static.mk: 没有那个文件或目录
cd htslib-1.9 && make htslib_static.mk
make[1]: 进入目录 `/home/txy/samtools/samtools-1.9/htslib-1.9'
sed -n '/^static_libs=/s/[^=]*=//HTSLIB_static_LIBS = /p;/^static_ldflags=/s/[^=]*=//HTSL
IB_static_LDFLAGS = /p' htslib.pc.tmp > htslib_static.mk
make[1]: 离开目录 `/home/txy/samtools/samtools-1.9/htslib-1.9'
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bam_index.o bam_index.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bam_plcmd.o bam_plcmd.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o sam_view.o sam_view.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bam_cat.o bam_cat.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bam_md.o bam_md.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bam_reheader.o bam_reheader.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bam_sort.o bam_sort.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bedidx.o bedidx.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bam_rmdup.o bam_rmdup.c
```

图 122 安装 samtools-1.9

```
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bam_rmdupse.o bam_rmdupse.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bam_mate.o bam_mate.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bam_stat.o bam_stat.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bam_color.o bam_color.c
echo '#define SAMTOOLS_VERSION "1.9"' > version.h
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bamtk.o bamtk.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bam2bcf.o bam2bcf.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bam2bcf_indel.o bam2bcf_indel.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o sample.o sample.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o cut_target.o cut_target.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o phase.o phase.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bam2depth.o bam2depth.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o padding.o padding.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bedcov.o bedcov.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bamshuf.o bamshuf.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o faidx.o faidx.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o dict.o dict.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o stats.o stats.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o stats_isize.o stats_isize.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bam_flags.o bam_flags.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bam_split.o bam_split.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bam_tview.o bam_tview.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bam_tview_curses.o bam_tview_curses.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bam_tview_html.o bam_tview_html.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bam_lpileup.o bam_lpileup.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bam_quickcheck.o bam_quickcheck.c
```

图 123 安装 samtools-1.9

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bam_addrprg.o bam_addrprg.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bam_markdup.o bam_markdup.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o tmp_file.o tmp_file.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o lz4/lz4.o lz4/lz4.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bam_aux.o bam_aux.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bam.o bam.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bam_import.o bam_import.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o sam.o sam.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o sam_header.o sam_header.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o bam_plbuf.o bam_plbuf.c
ar -csru libbam.a bam_aux.o bam.o bam_import.o sam.o sam_header.o bam_plbuf.o
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o sam_opts.o sam_opts.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o sam_utils.o sam_utils.c
ar -rcs libst.a sam_opts.o sam_utils.o
cd htslib-1.9 && make lib-static
make[1]: 进入目录“/home/txy/samtools/samtools-1.9/htslib-1.9”
gcc -Wall -g -O2 -I.  -c -o kfunc.o kfunc.c
gcc -Wall -g -O2 -I.  -c -o knetfile.o knetfile.c
gcc -Wall -g -O2 -I.  -c -o kstring.o kstring.c
gcc -Wall -g -O2 -I.  -c -o bcf_sr_sort.o bcf_sr_sort.c
gcc -Wall -g -O2 -I.  -c -o bgzf.o bgzf.c
gcc -Wall -g -O2 -I.  -c -o errmod.o errmod.c
gcc -Wall -g -O2 -I.  -c -o faidx.o faidx.c
gcc -Wall -g -O2 -I.  -c -o hfile.o hfile.c
gcc -Wall -g -O2 -I.  -c -o hfile_net.o hfile_net.c
echo '#define HTS_VERSION "1.9"' > version.h
```

图 124 安装 samtools-1.9

```
gcc -Wall -g -O2 -I.  -c -o hts.o hts.c
gcc -Wall -g -O2 -I.  -c -o hts_os.o hts_os.c
gcc -Wall -g -O2 -I.  -c -o md5.o md5.c
gcc -Wall -g -O2 -I.  -c -o multipart.o multipart.c
gcc -Wall -g -O2 -I.  -c -o probaln.o probaln.c
gcc -Wall -g -O2 -I.  -c -o realn.o realn.c
gcc -Wall -g -O2 -I.  -c -o regidx.o regidx.c
gcc -Wall -g -O2 -I.  -c -o sam.o sam.c
gcc -Wall -g -O2 -I.  -c -o synced_bcf_reader.o synced_bcf_reader.c
gcc -Wall -g -O2 -I.  -c -o vcf_sweep.o vcf_sweep.c
gcc -Wall -g -O2 -I.  -c -o tbx.o tbx.c
gcc -Wall -g -O2 -I.  -c -o textutils.o textutils.c
gcc -Wall -g -O2 -I.  -c -o thread_pool.o thread_pool.c
gcc -Wall -g -O2 -I.  -c -o vcf.o vcf.c
gcc -Wall -g -O2 -I.  -c -o vcfutils.o vcfutils.c
gcc -Wall -g -O2 -I.  -c -o cram/cram_codecs.o cram/cram_codecs.c
gcc -Wall -g -O2 -I.  -c -o cram/cram_decode.o cram/cram_decode.c
gcc -Wall -g -O2 -I.  -c -o cram/cram_encode.o cram/cram_encode.c
gcc -Wall -g -O2 -I.  -c -o cram/cram_external.o cram/cram_external.c
gcc -Wall -g -O2 -I.  -c -o cram/cram_index.o cram/cram_index.c
gcc -Wall -g -O2 -I.  -c -o cram/cram_io.o cram/cram_io.c
gcc -Wall -g -O2 -I.  -c -o cram/cram_samtools.o cram/cram_samtools.c
gcc -Wall -g -O2 -I.  -c -o cram/cram_stats.o cram/cram_stats.c
gcc -Wall -g -O2 -I.  -c -o cram/files.o cram/files.c
gcc -Wall -g -O2 -I.  -c -o cram/mFILE.o cram/mFILE.c
gcc -Wall -g -O2 -I.  -c -o cram/open_trace_file.o cram/open_trace_file.c
```

图 125 安装 samtools-1.9

```
gcc -Wall -g -O2 -I.  -c -o cram/pooled_alloc.o cram/pooled_alloc.c
gcc -Wall -g -O2 -I.  -c -o cram/rANS_static.o cram/rANS_static.c
gcc -Wall -g -O2 -I.  -c -o cram/sam_header.o cram/sam_header.c
gcc -Wall -g -O2 -I.  -c -o cram/string_alloc.o cram/string_alloc.c
ar -rc libhts.a kfunc.o knetfile.o kstring.o bcf_sr_sort.o bgzf.o errmod.o faidx.o hfil
e.o hfile_net.o hts.o hts_os.o md5.o multipart.o probaln.o realn.o regidx.o sam.o synce
d_bcf_reader.o vcf_sweep.o tbx.o textutils.o thread_pool.o vcf.o vcfutils.o cram/cram_c
odecs.o cram/cram_decode.o cram/cram_encode.o cram/cram_external.o cram/cram_index.o cr
am/cram_io.o cram/cram_samtools.o cram/cram_stats.o cram/files.o cram/mFILE.o cram/open
_trace_file.o cram/pooled_alloc.o cram/rANS_static.o cram/sam_header.o cram/string_allo
c.o
ranlib libhts.a
make[1]: 离开目录“/home/txy/samtools/samtools-1.9/htslib-1.9”
gcc  -L./lz4  -o samtools bam_index.o bam_plcmd.o sam_view.o bam_cat.o bam_md.o bam_reh
eader.o bam_sort.o bedidx.o bam_rmdup.o bam_rmdupse.o bam_mate.o bam_stat.o bam_color.o
 bamtk.o bam2bcf.o bam2bcf_indel.o sample.o cut_target.o phase.o bam2depth.o padding.o
bedcov.o bamshuf.o faidx.o dict.o stats.o stats_isize.o bam_flags.o bam_split.o bam_tvi
ew.o bam_tview_curses.o bam_tview_html.o bam_lpileup.o bam_quickcheck.o bam_addrprg.o b
am_markdup.o tmp_file.o ./lz4/lz4.o libbam.a libst.a htslib-1.9/libhts.a -lpthread -lz
-lm -lbz2 -llzma -lncursesw -lm -lz  -lpthread
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o misc/ace2sam.o misc/ace2sam.c
gcc  -o misc/ace2sam misc/ace2sam.o -lz
gcc -g -O2 -DMAQ_LONGREADS -I. -Ihtslib-1.9 -I./lz4  -c -o misc/maq2sam-long.o misc/maq
2sam.c
gcc  -o misc/maq2sam-long misc/maq2sam-long.o -lz
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o misc/maq2sam-short.o misc/maq2sam.c
```

图 126 安装 samtools-1.9

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
gcc  -o misc/maq2sam-short misc/maq2sam-short.o -lz
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o misc/md5fa.o misc/md5fa.c
gcc  -L./lz4  -o misc/md5fa misc/md5fa.o htslib-1.9/libhts.a -lpthread -lz -lm -lbz2 -l
lzma -lz
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o misc/md5sum-lite.o misc/md5sum-lite.c
gcc  -L./lz4  -o misc/md5sum-lite misc/md5sum-lite.o htslib-1.9/libhts.a -lpthread -lz
-lm -lbz2 -llzma -lz
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o misc/wgsim.o misc/wgsim.c
gcc  -L./lz4  -o misc/wgsim misc/wgsim.o -lm htslib-1.9/libhts.a -lpthread -lz -lm -lbz
2 -llzma -lz
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o test/merge/test_bam_translate.o test/merge/t
est_bam_translate.c
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o test/test.o test/test.c
gcc  -L./lz4  -o test/merge/test_bam_translate test/merge/test_bam_translate.o test/tes
t.o libst.a htslib-1.9/libhts.a -lpthread -lz -lm -lbz2 -llzma -lz  -lpthread
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o test/merge/test_rtrans_build.o test/merge/te
st_rtrans_build.c
gcc  -L./lz4  -o test/merge/test_rtrans_build test/merge/test_rtrans_build.o libst.a ht
slib-1.9/libhts.a -lpthread -lz -lm -lbz2 -llzma -lz  -lpthread
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o test/merge/test_trans_tbl_init.o test/merge/
test_trans_tbl_init.c
gcc  -L./lz4  -o test/merge/test_trans_tbl_init test/merge/test_trans_tbl_init.o libst.
a htslib-1.9/libhts.a -lpthread -lz -lm -lbz2 -llzma -lz  -lpthread
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o test/split/test_count_rg.o test/split/test_c
ount_rg.c
gcc  -L./lz4  -o test/split/test_count_rg test/split/test_count_rg.o test/test.o libst.
```

图 127 安装 samtools-1.9

```
gcc  -L./lz4  -o test/split/test_count_rg test/split/test_count_rg.o test/test.o libst.
a htslib-1.9/libhts.a -lpthread -lz -lm -lbz2 -llzma -lz  -lpthread
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o test/split/test_expand_format_string.o test/
split/test_expand_format_string.c
gcc  -L./lz4  -o test/split/test_expand_format_string test/split/test_expand_format_str
ing.o test/test.o libst.a htslib-1.9/libhts.a -lpthread -lz -lm -lbz2 -llzma -lz  -lpth
read
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o test/split/test_filter_header_rg.o test/spli
t/test_filter_header_rg.c
gcc  -L./lz4  -o test/split/test_filter_header_rg test/split/test_filter_header_rg.o te
st/test.o libst.a htslib-1.9/libhts.a -lpthread -lz -lm -lbz2 -llzma -lz  -lpthread
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o test/split/test_parse_args.o test/split/test
_parse_args.c
gcc  -L./lz4  -o test/split/test_parse_args test/split/test_parse_args.o test/test.o li
bst.a htslib-1.9/libhts.a -lpthread -lz -lm -lbz2 -llzma -lz  -lpthread
gcc -g -O2 -I. -Ihtslib-1.9 -I./lz4  -c -o test/vcf-miniview.o test/vcf-miniview.c
gcc  -L./lz4  -o test/vcf-miniview test/vcf-miniview.o htslib-1.9/libhts.a -lpthread -l
z -lm -lbz2 -llzma -lz  -lpthread
[txy@localhost samtools-1.9]$ make install
mkdir -p -m 755 /home/txy/samtools/bin /home/txy/samtools/bin /home/txy/samtools/share/
man/man1
install -p samtools /home/txy/samtools/bin
install -p misc/ace2sam misc/maq2sam-long misc/maq2sam-short misc/md5fa misc/md5sum-lit
e misc/wgsim /home/txy/samtools/bin
install -p misc/blast2sam.pl misc/bowtie2sam.pl misc/export2sam.pl misc/interpolate_sam
.pl misc/novo2sam.pl misc/plot-bamstats misc/psl2sam.pl misc/sam2vcf.pl misc/samtools.p
```

图 128 安装 samtools-1.9

```
l misc/seq_cache_populate.pl misc/soap2sam.pl misc/varfilter.py misc/wgsim_eval.pl misc
/zoom2sam.pl /home/txy/samtools/bin
install -p -m 644 samtools.1 misc/wgsim.1 /home/txy/samtools/share/man/man1
```

图 129 安装 samtools-1.9

3.2.4 命令路径的修改

1、在终端打开隐藏的.bashrc 文件（如图 130 所示）。

代码：vi ~/.bashrc

#打开.bashrc 文件。

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

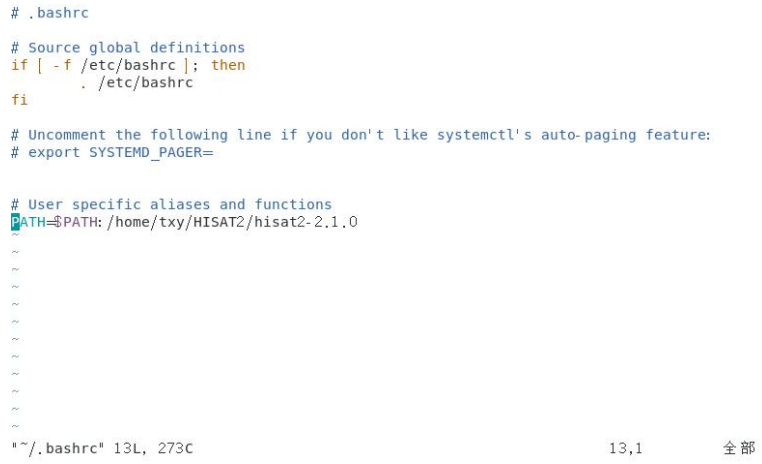


图 130 在终端打开隐藏的.bashrc 文件

2、按下键盘上的“I”，进入插入状态，在“#User specific aliases and functions”这一行下输入相应的路径，即 PATH=$PATH：/home/txy/samtools/bin（如图 131 所示）。

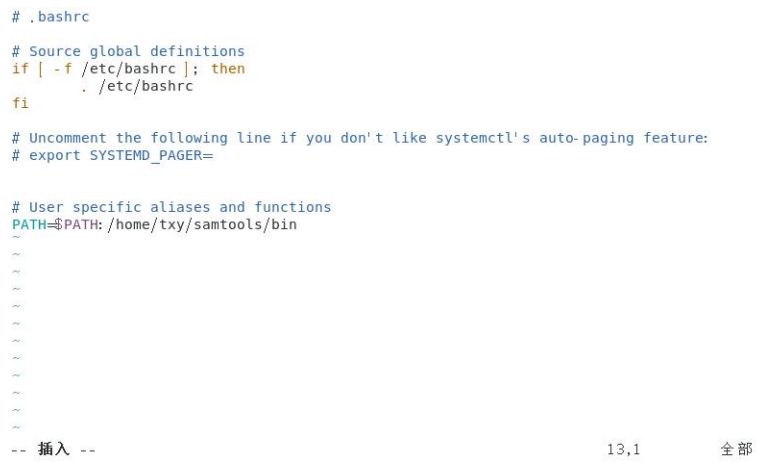


图 131 添加新的命令路径

3、按下键盘上的“esc”键（如图 132 所示），结束插入状态，输入“：wq”并回车，即保存了新的命令路径（如图 133 和图 134 所示）。

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

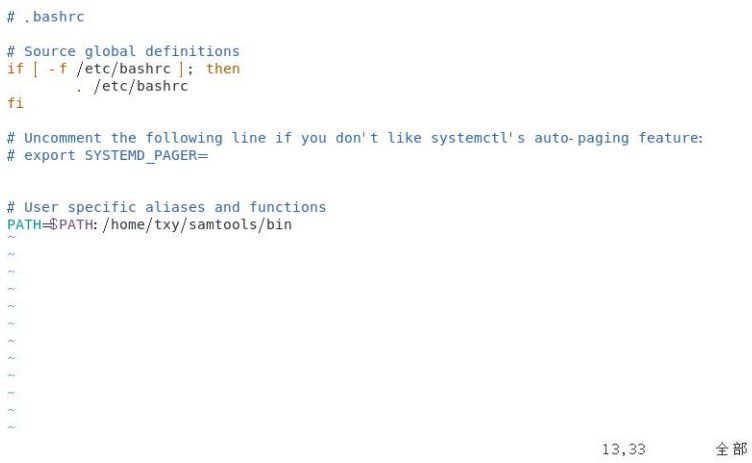


图 132 按下“esc”键结束插入状态

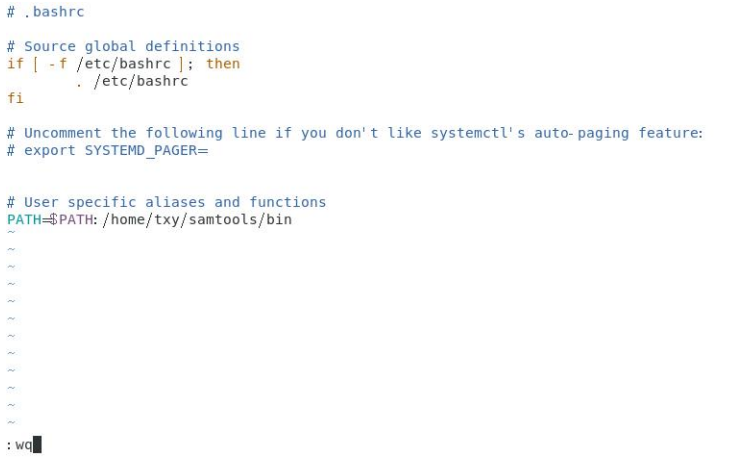


图 133 输入“：wq”

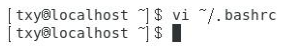


图 134 保存了新的命令路径

4、执行新的命令路径（如图 135 所示）。

代码：

```
source ~/.bashrc
#执行新的命令路径。
```



图 135 执行新的命令路径

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

3.3 samtools 的功能与使用方法

3.3.1 samtools 功能的使用

1、samtools 功能，文件和版本等相关的信息可以直接进行查询。同时，这一步操作也能检验 samtools 是否安装完成和命令路径是否正确，只有在这两者都正确完成的情况下，samtools 这个命令才能正常使用（如图 136 和图 137 所示）。

代码：

```
samtools
#samtools 功能使用帮助。
```

```
[txy@localhost ~]$ samtools

Program: samtools (Tools for alignments in the SAM format)
Version: 1.9 (using htslib 1.9)

Usage:   samtools <command> [options]

Commands:
  -- Indexing
     dict           create a sequence dictionary file
     faidx          index/extract FASTA
     fqidx          index/extract FASTQ
     index          index alignment

  -- Editing
     calmd          recalculate MD/NM tags and '=' bases
     fixmate        fix mate information
     reheader       replace BAM header
     targetcut      cut fosmid regions (for fosmid pool only)
     addreplacerg   adds or replaces RG tags
     markdup        mark duplicates

  -- File operations
     collate        shuffle and group alignments by name
     cat            concatenate BAMs
     merge          merge sorted alignments
```

图 136 samtools 功能使用

```
     mpileup        multi-way pileup
     sort           sort alignment file
     split          splits a file by read group
     quickcheck     quickly check if SAM/BAM/CRAM file appears intact
     fastq          converts a BAM to a FASTQ
     fasta          converts a BAM to a FASTA

  -- Statistics
     bedcov         read depth per BED region
     depth          compute the depth
     flagstat       simple stats
     idxstats       BAM index stats
     phase          phase heterozygotes
     stats          generate stats (former bamcheck)

  -- Viewing
     flags          explain BAM flags
     tview          text alignment viewer
     view           SAM<->BAM<->CRAM conversion
     depad          convert padded BAM to unpadded BAM
```

图 137 samtools 功能使用

3.3.2 将 sam 文件转换为 bam 文件。

1、将 sam 文件转换为 bam 文件（如图 138 和图 139 所示）。.

代码：

```
samtools view -b -S /home/txy/HISAT2/T1.sam > T1.bam
```

未经授权，不得用于任何商业用途！
联系方式：bfinfo@zju.edu.cn

#“view”表示 sam 格式文件和 bam 格式文件之间的转换。

#“-b”表示输出 bam 格式文件，默认输出 sam 格式文件。

#“-S”表示输入 sam 格式文件，默认输入 bam 格式文件。

#“/home/txy/HISAT2/T1.sam”表示需要转换的 sam 格式文件及其路径。

#“>”表示转换。

#“T1.bam”表示输出的 bam 格式文件，默认路径是在主目录下。

```
[txy@localhost ~]$ samtools view -b -S /home/txy/HISAT2/T1.sam > T1.bam
[txy@localhost ~]$
```

图 138 将 sam 格式文件转换为 bam 格式文件

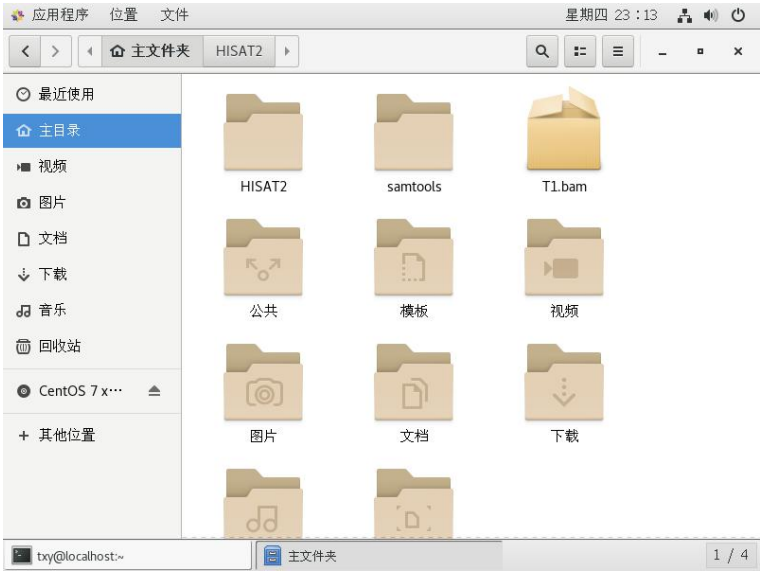


图 139 转换得到的 bam 格式文件

3.3.3 将 bam 文件转换成 sam 文件（如图 140 和图 141 所示）

代码：

samtools view -o T1.sam /home/txy/T1.bam

#“view”表示 sam 格式文件和 bam 格式文件之间的转换。

#“-o”表示输出 sam 格式文件。

#“T1.sam”表示输出的 sam 格式文件，默认路径是在主目录下。

#“/home/txy/T1.bam”表示需要转换的 bam 格式文件及其路径。

```
[txy@localhost ~]$ samtools view -h -o T1.sam /home/txy/T1.bam
[txy@localhost ~]$
```

未经授权，不得用于任何商业用途！
联系方式：bnfo@zjtu.edu.cn

图 140 将 bam 格式文件转换为 sam 格式文件

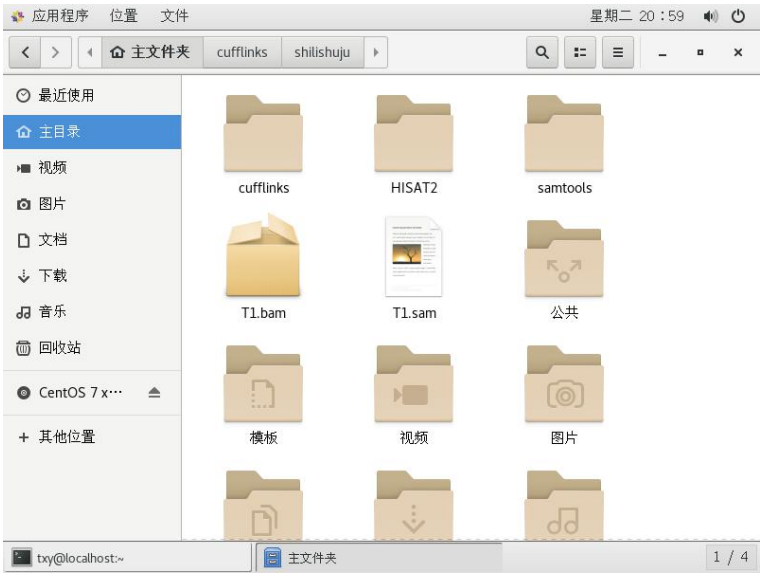


图 141 转换得到的 sam 格式文件

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

4 sam 格式文件/bam 格式文件的分析——cufflinks

4.1 关于 cufflinks

Cufflinks 是一套工具和可执行文件的集合，较新版本的 cufflinks 包含了 cuffmerge，cuffquant，cuffdiff，cummerbund 和 cuffnorm 这五个工具，分别完成 cufflinks 工作流程的一部分功能（如图 142 所示）。Cufflinks 能够利用 Tophat 比对的结果（alignments）来组装转录本，估计这些转录本的丰度，并且检测样本间的差异表达及可变剪接。

一般来说，cufflinks 能够识别 sam 文件的输入，但是对于 bam 文件，cufflinks 只能识别 TopHat 输出的 bam 文件，其他的 bam 文件一般缺少正负号 XS 标签，需要用一定的指令处理后才能被 cufflinks 识别。

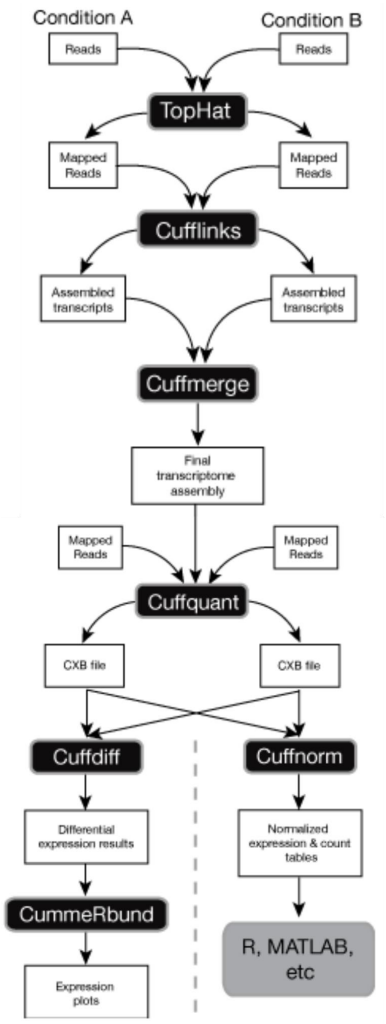


图 142 cufflinks 工作流程图

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

4.2 cufflinks 的安装

4.2.1 cufflinks 的下载和解压缩

1、为了便于管理所安装的软件，首先在主文件夹下创建一个文件夹，命名为“cufflinks”（如图 143-图 145 所示）。

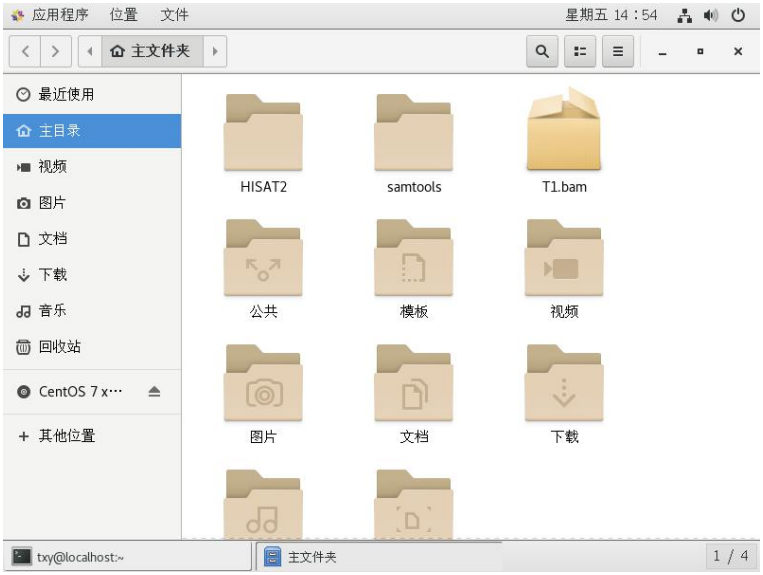


图 143 主文件夹

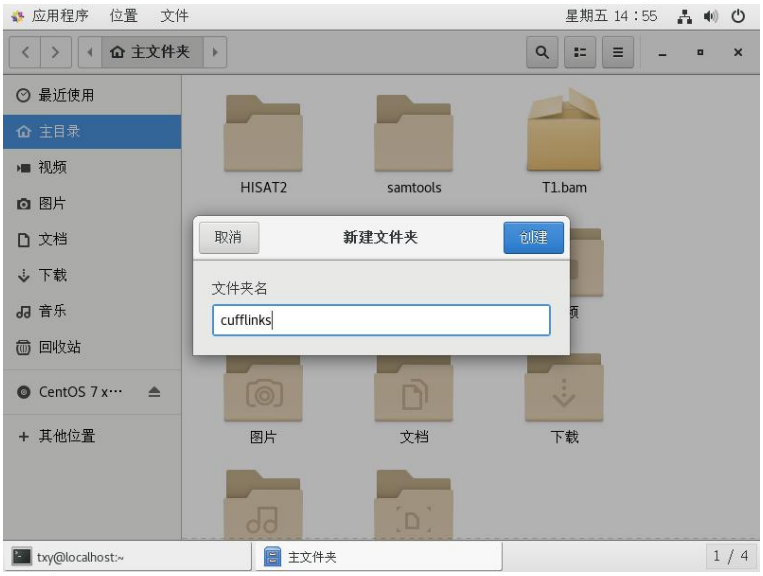


图 144 创建文件夹，命名为“HISAT2”

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

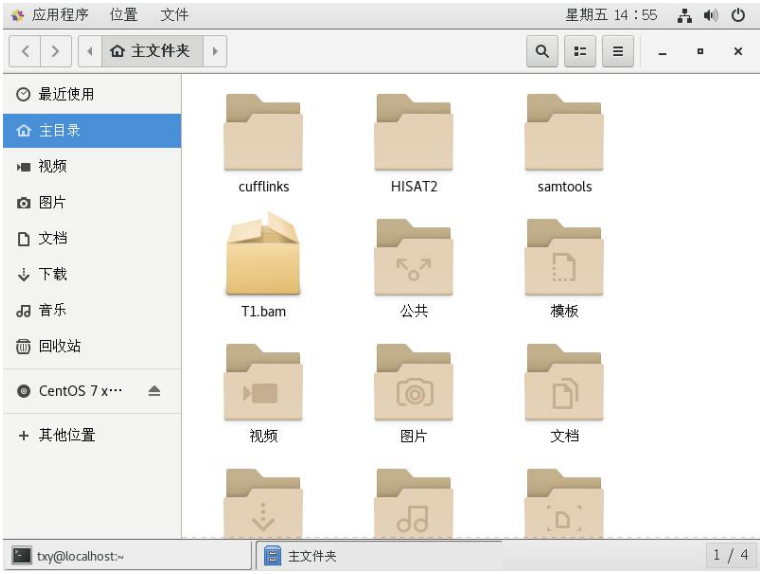


图 145 HISAT2 文件夹

2、下载 cufflinks，参考网址：

http://cole-trapnell-lab.github.io/cufflinks/assets/downloads/cufflinks-2.2.1.Linux_x86_64.tar.gz（如图 146-图 150 所示）。

代码：

wget

http://cole-trapnell-lab.github.io/cufflinks/assets/downloads/cufflinks-2.2.1.Linux_x86_64.tar.gz

#下载 cufflinks。

```
[txy@localhost ~]$ wget http://cole-trapnell-lab.github.io/cufflinks/assets/downloads/cufflinks-2.2.1.Linux_x86_64.tar.gz
--2019-04-05 15:04:14--  http://cole-trapnell-lab.github.io/cufflinks/assets/downloads/cufflinks-2.2.1.Linux_x86_64.tar.gz
正在解析主机 cole-trapnell-lab.github.io (cole-trapnell-lab.github.io)... 185.199.108.153, 185.199.109.153, 185.199.110.153, ...
正在连接 cole-trapnell-lab.github.io (cole-trapnell-lab.github.io)|185.199.108.153|:80... 失败：拒绝连接。
正在连接 cole-trapnell-lab.github.io (cole-trapnell-lab.github.io)|185.199.109.153|:80... 已连接。
已发出 HTTP 请求，正在等待回应... 读取文件头错误 (Connection reset by peer)。
重试中。

--2019-04-05 15:04:17--  (尝试次数： 2)  http://cole-trapnell-lab.github.io/cufflinks/assets/downloads/cufflinks-2.2.1.Linux_x86_64.tar.gz
正在连接 cole-trapnell-lab.github.io (cole-trapnell-lab.github.io)|185.199.109.153|:80... 已连接。
已发出 HTTP 请求，正在等待回应... 读取文件头错误 (Connection reset by peer)。
重试中。

--2019-04-05 15:04:20--  (尝试次数： 3)  http://cole-trapnell-lab.github.io/cufflinks/assets/downloads/cufflinks-2.2.1.Linux_x86_64.tar.gz
正在连接 cole-trapnell-lab.github.io (cole-trapnell-lab.github.io)|185.199.109.153|:80... 已连接。
已发出 HTTP 请求，正在等待回应... 读取文件头错误 (Connection reset by peer)。
重试中。
```

图 146 下载 cufflinks

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
--2019-04-05 15:04:23--  (尝试次数： 4)  http://cole-trapnell-lab.github.io/cufflinks/a
ssets/downloads/cufflinks-2.2.1.Linux_x86_64.tar.gz
正在连接 cole-trapnell-lab.github.io (cole-trapnell-lab.github.io)|185.199.109.153|:80.
.. 已连接。
已发出 HTTP 请求，正在等待回应... 读取文件头错误 (Connection reset by peer)。
重试中。

--2019-04-05 15:04:28--  (尝试次数： 5)  http://cole-trapnell-lab.github.io/cufflinks/a
ssets/downloads/cufflinks-2.2.1.Linux_x86_64.tar.gz
正在连接 cole-trapnell-lab.github.io (cole-trapnell-lab.github.io)|185.199.109.153|:80.
.. 已连接。
已发出 HTTP 请求，正在等待回应... 读取文件头错误 (Connection reset by peer)。
重试中。

--2019-04-05 15:04:34--  (尝试次数： 6)  http://cole-trapnell-lab.github.io/cufflinks/a
ssets/downloads/cufflinks-2.2.1.Linux_x86_64.tar.gz
正在连接 cole-trapnell-lab.github.io (cole-trapnell-lab.github.io)|185.199.109.153|:80.
.. 已连接。
已发出 HTTP 请求，正在等待回应... 读取文件头错误 (Connection reset by peer)。
重试中。

--2019-04-05 15:04:40--  (尝试次数： 7)  http://cole-trapnell-lab.github.io/cufflinks/a
ssets/downloads/cufflinks-2.2.1.Linux_x86_64.tar.gz
正在连接 cole-trapnell-lab.github.io (cole-trapnell-lab.github.io)|185.199.109.153|:80.
.. 已连接。
已发出 HTTP 请求，正在等待回应... 读取文件头错误 (Connection reset by peer)。
```

图 147 下载 cufflinks

```
重试中。

--2019-04-05 15:23:23--  (尝试次数： 8)  http://cole-trapnell-lab.github.io/cufflinks/a
ssets/downloads/cufflinks-2.2.1.Linux_x86_64.tar.gz
正在连接 cole-trapnell-lab.github.io (cole-trapnell-lab.github.io)|185.199.111.153|:80.
.. 已连接。
已发出 HTTP 请求，正在等待回应... 读取文件头错误 (Connection reset by peer)。
重试中。

--2019-04-05 15:23:31--  (尝试次数： 9)  http://cole-trapnell-lab.github.io/cufflinks/a
ssets/downloads/cufflinks-2.2.1.Linux_x86_64.tar.gz
正在连接 cole-trapnell-lab.github.io (cole-trapnell-lab.github.io)|185.199.111.153|:80.
.. 已连接。
已发出 HTTP 请求，正在等待回应... 读取文件头错误 (Connection reset by peer)。
重试中。

--2019-04-05 15:23:40--  (尝试次数：10)  http://cole-trapnell-lab.github.io/cufflinks/a
ssets/downloads/cufflinks-2.2.1.Linux_x86_64.tar.gz
正在连接 cole-trapnell-lab.github.io (cole-trapnell-lab.github.io)|185.199.111.153|:80.
.. 已连接。
已发出 HTTP 请求，正在等待回应... 读取文件头错误 (Connection reset by peer)。
重试中。

--2019-04-05 15:23:50--  (尝试次数：11)  http://cole-trapnell-lab.github.io/cufflinks/a
ssets/downloads/cufflinks-2.2.1.Linux_x86_64.tar.gz
正在连接 cole-trapnell-lab.github.io (cole-trapnell-lab.github.io)|185.199.111.153|:80.
```

图 148 下载 cufflinks

```
.. 已连接。
已发出 HTTP 请求，正在等待回应... 读取文件头错误 (Connection reset by peer)。
重试中。

--2019-04-05 15:24:11--  (尝试次数：13)  http://cole-trapnell-lab.github.io/cufflinks/a
ssets/downloads/cufflinks-2.2.1.Linux_x86_64.tar.gz
正在连接 cole-trapnell-lab.github.io (cole-trapnell-lab.github.io)|185.199.111.153|:80.
.. 已连接。
已发出 HTTP 请求，正在等待回应... 读取文件头错误 (Connection reset by peer)。
重试中。

--2019-04-05 15:24:21--  (尝试次数：14)  http://cole-trapnell-lab.github.io/cufflinks/a
ssets/downloads/cufflinks-2.2.1.Linux_x86_64.tar.gz
正在连接 cole-trapnell-lab.github.io (cole-trapnell-lab.github.io)|185.199.111.153|:80.
.. 已连接。
已发出 HTTP 请求，正在等待回应... 读取文件头错误 (Connection reset by peer)。
重试中。

--2019-04-05 15:24:31--  (尝试次数：15)  http://cole-trapnell-lab.github.io/cufflinks/a
ssets/downloads/cufflinks-2.2.1.Linux_x86_64.tar.gz
正在连接 cole-trapnell-lab.github.io (cole-trapnell-lab.github.io)|185.199.111.153|:80.
.. 已连接。
已发出 HTTP 请求，正在等待回应... 读取文件头错误 (Connection reset by peer)。
重试中。

--2019-04-05 15:24:42--  (尝试次数：16)  http://cole-trapnell-lab.github.io/cufflinks/a
```

图 149 下载 cufflinks

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
ssets/downloads/cufflinks-2.2.1.Linux_x86_64.tar.gz
正在连接 cole-trapnell-lab.github.io (cole-trapnell-lab.github.io)|185.199.111.153|:80.
.. 已连接。
已发出 HTTP 请求，正在等待回应... 读取文件头错误 (Connection reset by peer)。
重试中。

--2019-04-05 15:24:52--  (尝试次数：17)  http://cole-trapnell-lab.github.io/cufflinks/a
ssets/downloads/cufflinks-2.2.1.Linux_x86_64.tar.gz
正在连接 cole-trapnell-lab.github.io (cole-trapnell-lab.github.io)|185.199.111.153|:80.
.. 已连接。
已发出 HTTP 请求，正在等待回应... 200 OK
长度：49298683 (47M) [application/gzip]
正在保存至："cufflinks-2.2.1.Linux_x86_64.tar.gz"

100%[=====================================>] 49,298,683  20.3KB/s 用时 30m 58s

2019-04-05 15:55:51 (25.9 KB/s) - 已保存 "cufflinks-2.2.1.Linux_x86_64.tar.gz" [4929868
3/49298683])
```

图 150 下载 cufflinks

3、解压缩 cufflinks（如图 151 所示）。

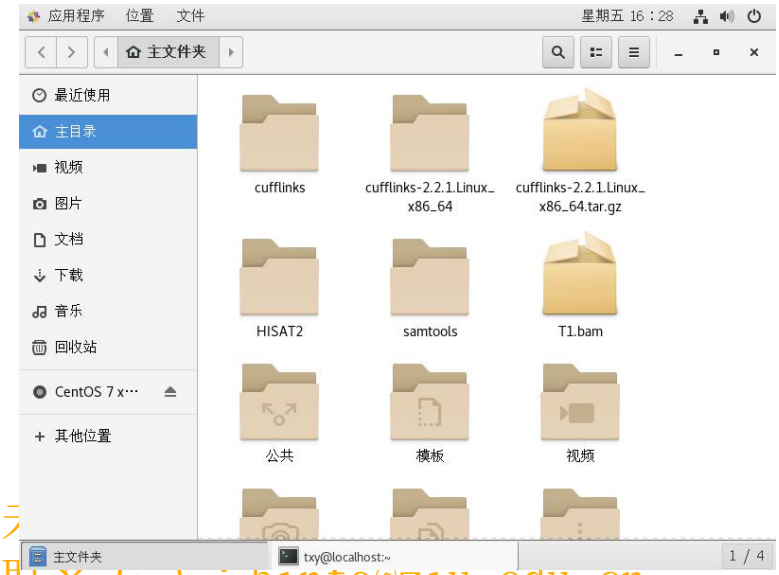
代码：

```
tar xzvf /home/txy/cufflinks-2.2.1.Linux_x86_64.tar.gz
#解压缩 cufflinks。
```

```
[txy@localhost ~]$ tar xzvf /home/txy/cufflinks-2.2.1.Linux_x86_64.tar.gz
cufflinks-2.2.1.Linux_x86_64/
cufflinks-2.2.1.Linux_x86_64/gtf_to_sam
cufflinks-2.2.1.Linux_x86_64/cuffcompare
cufflinks-2.2.1.Linux_x86_64/cuffdiff
cufflinks-2.2.1.Linux_x86_64/LICENSE
cufflinks-2.2.1.Linux_x86_64/cufflinks
cufflinks-2.2.1.Linux_x86_64/cuffmerge
cufflinks-2.2.1.Linux_x86_64/README
cufflinks-2.2.1.Linux_x86_64/cuffnorm
cufflinks-2.2.1.Linux_x86_64/gffread
cufflinks-2.2.1.Linux_x86_64/cuffquant
cufflinks-2.2.1.Linux_x86_64/AUTHORS
```

图 151 解压缩 cufflinks

4、解压缩后的 cufflinks 文件和 cufflinks 压缩包都位于主文件夹中（如图 152 所示），为了便于管理，将它们一起放入先前建好的 cufflinks 文件夹中（如图 153 所示）。



联系方式：bintp@zju.edu.cn

图 152 解压缩后的 cufflinks 文件和 cufflinks 压缩包

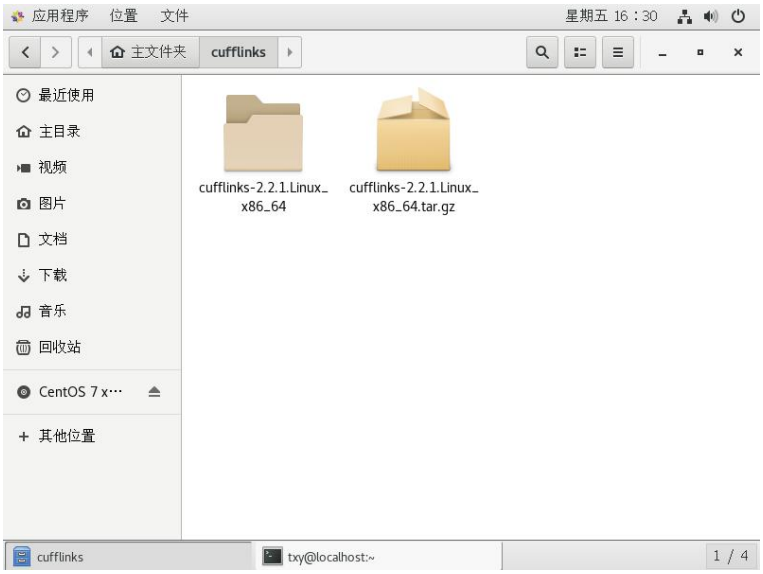


图 153 将解压缩后的 cufflinks 文件和 cufflinks 压缩包放入文件夹 cufflinks 中

4.2.2 命令路径的修改

1、在终端打开隐藏的.bashrc 文件（如图 154 所示）。

代码：

vi ~/.bashrc

#打开.bashrc 文件。

```
# .bashrc

# Source global definitions
if [ -f /etc/bashrc ]; then
        . /etc/bashrc
fi

# Uncomment the following line if you don't like systemctl's auto-paging feature:
# export SYSTEMD_PAGER=


# User specific aliases and functions
PATH=$PATH:/home/txy/samtools/bin
~
~
~
~
~
~
~
~
~
~
~
~
"~/.bashrc" 13L, 266C                                         13,1          全部
```

图 154 在终端打开隐藏的.bashrc 文件

2、按下键盘上的“I”，进入插入状态，在“#User specific aliases and functions”这一行下输入相应的路径，即 PATH=$PATH:/home/txy/HISAT2/hisat2-2.1.0（如图 155

未经授权，不得用于任何商业用途！
联系方式：bliu@zju.edu.cn

所示）。

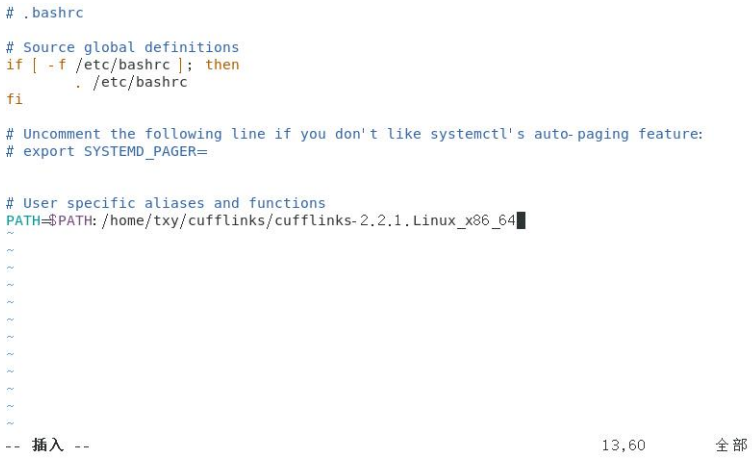


图 155 添加新的命令路径

3、按下键盘上的“esc”键（如图 156 所示），结束插入状态，输入“：wq”并回车，即保存了新的命令路径（如图 157 和图 158 所示）。

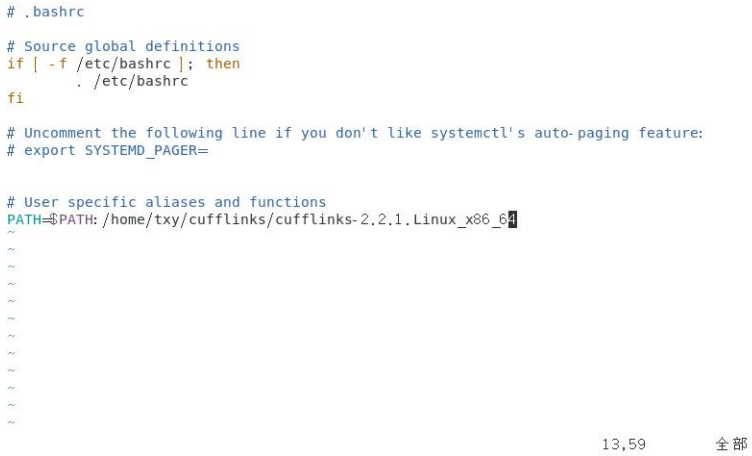
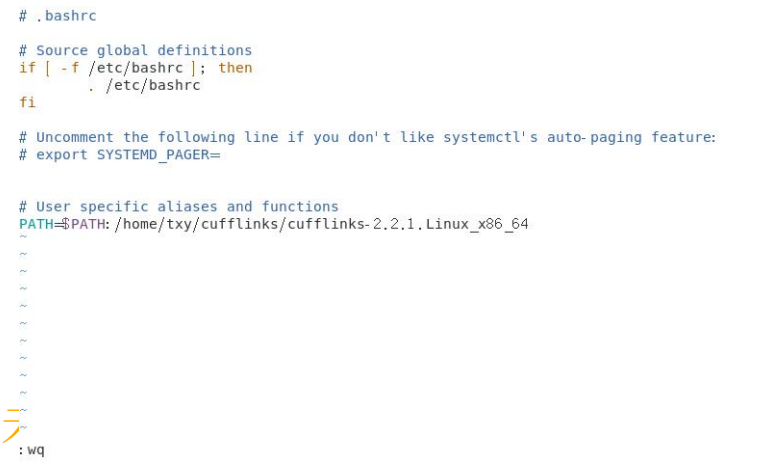


图 156 按下“esc”键结束插入状态



联系方式：binfo@zju.edu.cn

图 157 输入“：wq”

```
[txy@localhost ~]$ vi ~/.bashrc
[txy@localhost ~]$
```

图 158 保存了新的命令路径

4、执行新的命令路径（如图 159 所示）。

代码：

source ~/.bashrc

#执行新的命令路径。

```
[txy@localhost ~]$ source ~/.bashrc
[txy@localhost ~]$
```

图 159 执行新的命令路径

4.3 cufflinks 的功能与使用方法

4.3.1 help 功能的使用

1、cufflinks 功能，文件和版本等相关的信息可以直接进行查询。同时，这一步操作也能检验 cufflinks 是否安装完成和命令路径是否正确，只有在这两者都正确完成的情况下，cufflinks 这个命令才能正常使用。Cufflinks 的具体功能和参数也可以通过直接输入 cufflinks 进行查询（如图 160-图 164 所示）。

代码：

Cufflinks

#cufflinks 功能使用帮助。

```
[txy@localhost ~]$ cufflinks -h
cufflinks: invalid option -- 'h'
cufflinks v2.2.1
linked against Boost version 104700
-----------------------------
Usage:   cufflinks [options] <hits.sam>
General Options:
  -o/--output-dir             write all output files to this directory             [
default:     ./ ]
  -p/--num-threads            number of threads used during analysis               [
default:       1 ]
  --seed                      value of random number generator seed                [
default:       0 ]
  -G/--GTF                    quantitate against reference transcript annotations

  -g/--GTF-guide              use reference transcript annotation to guide assembly

  -M/--mask-file              ignore all alignment within transcripts in this file

  -b/--frag-bias-correct      use bias correction - reference fasta required       [
default:    NULL ]
  -u/--multi-read-correct     use 'rescue method' for multi-reads (more accurate)  [
default:   FALSE ]
  --library-type              library prep used for input reads                    [
default:   below ]
  --library-norm-method       Method used to normalize library sizes               [
```

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

图 160 cufflinks 功能的使用

```
default:   below ]

Advanced Abundance Estimation Options:
  -m/--frag-len-mean           average fragment length (unpaired reads only)         [
default:     200 ]
  -s/--frag-len-std-dev        fragment length std deviation (unpaired reads only)   [
default:      80 ]
  --max-mle-iterations         maximum iterations allowed for MLE calculation        [
default:    5000 ]
  --compatible-hits-norm       count hits compatible with reference RNAs only        [
default:   FALSE ]
  --total-hits-norm            count all hits for normalization                      [
default:   TRUE  ]
  --num-frag-count-draws       Number of fragment generation samples                 [
default:     100 ]
  --num-frag-assign-draws      Number of fragment assignment samples per generation  [
default:      50 ]
  --max-frag-multihits         Maximum number of alignments allowed per fragment     [
default:  unlim  ]
  --no-effective-length-correction   No effective length correction                  [
default:   FALSE ]
  --no-length-correction       No length correction                                  [
default:   FALSE ]
  -N/--upper-quartile-norm     Deprecated, use --library-norm-method                 [
   DEPRECATED    ]
  --raw-mapped-norm            Deprecated, use --library-norm-method                 [
```

图 161 cufflinks 功能的使用

```
   DEPRECATED    ]

Advanced Assembly Options:
  -L/--label                   assembled transcripts have this ID prefix             [
default:    CUFF ]
  -F/--min-isoform-fraction    suppress transcripts below this abundance level       [
default:    0.10 ]
  -j/--pre-mrna-fraction       suppress intra-intronic transcripts below this level  [
default:    0.15 ]
  -I/--max-intron-length       ignore alignments with gaps longer than this          [
default:  300000 ]
  -a/--junc-alpha              alpha for junction binomial test filter               [
default:   0.001 ]
  -A/--small-anchor-fraction   percent read overhang taken as 'suspiciously small'   [
default:    0.09 ]
  --min-frags-per-transfrag    minimum number of fragments needed for new transfrags [
default:      10 ]
  --overhang-tolerance         number of terminal exon bp to tolerate in introns     [
default:       8 ]
  --max-bundle-length          maximum genomic length allowed for a given bundle     [
default: 3500000 ]
  --max-bundle-frags           maximum fragments allowed in a bundle before skipping [
default:  500000 ]
  --min-intron-length          minimum intron size allowed in genome                 [
default:      50 ]
```

图 162 cufflinks 功能的使用

```
  --trim-3-avgcov-thresh       minimum avg coverage required to attempt 3' trimming  [
default:      10 ]
  --trim-3-dropoff-frac        fraction of avg coverage below which to trim 3' end   [
default:     0.1 ]
  --max-multiread-fraction     maximum fraction of allowed multireads per transcript [
default:    0.75 ]
  --overlap-radius             maximum gap size to fill between transfrags (in bp)   [
default:      50 ]

Advanced Reference Annotation Guided Assembly Options:
  --no-faux-reads              disable tiling by faux reads                          [
default:   FALSE ]
  --3-overhang-tolerance       overhang allowed on 3' end when merging with reference[
default:     600 ]
  --intron-overhang-tolerance  overhang allowed inside reference intron when merging [
default:      30 ]

Advanced Program Behavior Options:
  -v/--verbose                 log-friendly verbose processing (no progress bar)     [
default:   FALSE ]
  -q/--quiet                   log-friendly quiet processing (no progress bar)       [
default:   FALSE ]
  --no-update-check            do not contact server to check for update availability[
default:   FALSE ]

Supported library types:
```

图 163 cufflinks 功能的使用

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
        ff-firststrand
        ff-secondstrand
        ff-unstranded
        fr-firststrand
        fr-secondstrand
        fr-unstranded (default)
        transfrags

Supported library normalization methods:
        classic-fpkm
```

图 164 cufflinks 功能的使用

2、cuffcompare 的功能主要是比较，当我们获得一个或者多个样品组装成的转录组后，可能需要将已知的转录本与该转录组进行比较。或者，当我们在没有参考转录组的情况下，需要对从不同 RNA-seq 文库获得的转录组进行比较。这时候就需要 cuffcompare 的运行，且 cuffcompare 还能够评估转录组组装的效果。Cuffcompare 的具体功能和参数也可以通过直接输入 cuffcompare 进行查询（如图 165 和图 166 所示）。

代码：

cuffcompare

#cuffcompare 功能使用帮助

```
[txy@localhost ~]$ cuffcompare -h
cuffcompare v2.2.1 (4237)
-----------------------------
Usage:
cuffcompare [-r <reference_mrna.gtf>] [-R] [-T] [-V] [-s <seq_path>]
    [-o <outprefix>] [-p <cprefix>]
    {-i <input_gtf_list> | <input1.gtf> [<input2.gtf> .. <inputN.gtf>]}

 Cuffcompare provides classification, reference annotation mapping and various
 statistics for Cufflinks transfrags.
 Cuffcompare clusters and tracks transfrags across multiple samples, writing
 matching transcripts (intron chains) into <outprefix>.tracking, and a GTF
 file <outprefix>.combined.gtf containing a nonredundant set of transcripts
 across all input files (with a single representative transfrag chosen
 for each clique of matching transfrags across samples).

Options:
-i provide a text file with a list of Cufflinks GTF files to process instead
   of expecting them as command line arguments (useful when a large number
   of GTF files should be processed)

-r a set of known mRNAs to use as a reference for assessing
   the accuracy of mRNAs or gene models given in <input.gtf>

-R for -r option, consider only the reference transcripts that
   overlap any of the input transfrags (Sn correction)
```

图 165 cuffcompare 功能的使用

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
-Q for -r option, consider only the input transcripts that
   overlap any of the reference transcripts (Sp correction);
   (Warning: this will discard all "novel" loci!)
-M discard (ignore) single-exon transfrags and reference transcripts
-N discard (ignore) single-exon reference transcripts

-s <seq_path> can be a multi-fasta file with all the genomic sequences or
   a directory containing multiple single-fasta files (one file per contig);
   lower case bases will be used to classify input transcripts as repeats

-e max. distance (range) allowed from free ends of terminal exons of reference
   transcripts when assessing exon accuracy (100)
-d max. distance (range) for grouping transcript start sites (100)
-p the name prefix to use for consensus transcripts in the
   <outprefix>.combined.gtf file (default: 'TCONS')
-C include the "contained" transcripts in the .combined.gtf file
-F do not discard intron-redundant transfrags if they share the 5' end
   (if they differ only at the 3' end))
-G generic GFF input file(s): do not assume Cufflinks GTF, do not
   discard any intron-redundant transfrags)
-T do not generate .tmap and .refmap files for each input file
-V verbose processing mode (showing all GFF parsing warnings)
```

图 166 cuffcompare 功能的使用

3、cuffmerge 的功能主要是合并，当我们拥有多个 RNA-seq 文库并且已经从其中获得了多个转录组的时候，我们可以用 cuffmmerge 合并这些转录组，这一步可以整合多个转录组，并且对于比较新获得的转录本与已整合的转录组不同之处是必要的。Cuffmerge 的具体功能和参数也可以通过直接输入 cuffmerge 进行查询（如图 167 所示）。

代码：

cuffmerge

#cuffmerge 功能使用帮助

```
[txy@localhost ~]$ cuffmerge -h
cuffmerge:
cuffmerge takes two or more Cufflinks GTF files and merges them into a
single unified transcript catalog.  Optionally, you can provide the script
with a reference GTF, and the script will use it to attach gene names and other
metadata to the merged catalog.

Usage:
    cuffmerge [Options] <assembly_GTF_list.txt>

Options:
    -h/--help                          Prints the help message and exits
    -o                    <output_dir>     Directory where merged assembly will be wri
tten  [ default: ./merged_asm  ]
    -g/--ref-gtf                       An optional "reference" annotation GTF.
    -s/--ref-sequence     <seq_dir>/<seq_fasta> Genomic DNA sequences for the referenc
e.
    --min-isoform-fraction <0-1.0>         Discard isoforms with abundance below this
      [ default:           0.05 ]
    -p/--num-threads      <int>            Use this many threads to merge assemblies.
      [ default:             1  ]
    --keep-tmp                         Keep all intermediate files during merge
```

图 167 cuffmerge 功能的使用

4、cuffquant 的功能是量化并计算 RNA-seq 样本中基因和转录本的表达，形成表达谱，并将这些表达谱保存成可用于 cuffdiff 或者 cuffnorm 分析的文件。cuffquant 的具体功能和参数也可以通过直接输入 cuffquant 进行查询（如图 168-图 170 所示）。

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

代码：

cuffquant

#cuffquant 功能使用帮助

```
[txy@localhost ~]$ cuffquant -h
cuffquant: unrecognized option '-h'
cuffquant v2.2.1 (4237)
-----------------------------
Usage:   cuffdiff [options] <transcripts.gtf> <sample1_hits.sam> <sample2_hits.sam> [..
. sampleN_hits.sam]
   Supply replicate SAMs as comma separated lists for each condition: sample1_rep1.sam,
sample1_rep2.sam,...sample1_repM.sam
General Options:
  -o/--output-dir              write all output files to this directory              [
default:     ./ ]
  -M/--mask-file               ignore all alignment within transcripts in this file  [
default:    NULL ]
  -b/--frag-bias-correct       use bias correction - reference fasta required        [
default:    NULL ]
  -u/--multi-read-correct      use 'rescue method' for multi-reads                   [
default:   FALSE ]
  -p/--num-threads             number of threads used during quantification          [
default:       1 ]
  --library-type               Library prep used for input reads                     [
default:   below ]

Advanced Options:
  -m/--frag-len-mean           average fragment length (unpaired reads only)         [
default:     200 ]
  -s/--frag-len-std-dev        fragment length std deviation (unpaired reads only)   [
```

图 168 cuffquant 功能的使用

```
default:      80 ]
  -c/--min-alignment-count     minimum number of alignments in a locus for testing   [
default:    10 ]
  --max-mle-iterations         maximum iterations allowed for MLE calculation        [
default:    5000 ]
  -v/--verbose                 log-friendly verbose processing (no progress bar)     [
default:   FALSE ]
  -q/--quiet                   log-friendly quiet processing (no progress bar)       [
default:   FALSE ]
  --seed                       value of random number generator seed                 [
default:       0 ]
  --no-update-check            do not contact server to check for update availability[
default:   FALSE ]
  --max-bundle-frags           maximum fragments allowed in a bundle before skipping [
default:  500000 ]
  --max-frag-multihits         Maximum number of alignments allowed per fragment     [
default:  unlim  ]
  --no-effective-length-correction   No effective length correction                  [
default:   FALSE ]
  --no-length-correction       No length correction                                  [
default:   FALSE ]

Debugging use only:
  --read-skip-fraction         Skip a random subset of reads this size               [
default:     0.0 ]
  --no-read-pairs              Break all read pairs                                  [
```

图 169 cuffquant 功能的使用

```
default:   FALSE ]
  --trim-read-length           Trim reads to be this long (keep 5' end)              [
default:    none ]
  --no-scv-correction          Disable SCV correction                                [
default:   FALSE ]

Supported library types:
        ff-firststrand
        ff-secondstrand
        ff-unstranded
        fr-firststrand
        fr-secondstrand
        fr-unstranded (default)
        transfrags _
```

图 170 cuffquant 功能的使用

5、而 cuffdiff 的主要功能是比较 RNA-seq 实验中的基因和转录本表达水平，这是一个非常精确的工具，不进可以告知我们在几个不同的条件下哪个基因上调或

未经授权，不得用于任何商业用途！
联系方式：jinlu@zju.edu.cn

者下调，还可以告诉我们哪个基因发生了可变剪接，或者经受了其他类型的调节。cuffdiff 的具体功能和参数也可以通过直接输入 cuffdiff 进行查询（如图 171-图 175 所示）。

代码：

cuffdiff

#cuffdiff 功能使用帮助

```
[txy@localhost ~]$ cuffdiff -h
cuffdiff: unrecognized option '-h'
cuffdiff v2.2.1 (4237)
-----------------------------
Usage:   cuffdiff [options]  <transcripts.gtf> <sample1_hits.sam> <sample2_hits.sam> [..
. sampleN_hits.sam]
   Supply replicate SAMs as comma separated lists for each condition: sample1_rep1.sam,
sample1_rep2.sam,...sample1_repM.sam
General Options:
  -o/--output-dir              write all output files to this directory              [
default:     ./ ]
  -L/--labels                  comma-separated list of condition labels
  --FDR                        False discovery rate used in testing                  [
default:    0.05 ]
  -M/--mask-file               ignore all alignment within transcripts in this file  [
default:    NULL ]
  -C/--contrast-file           Perform the constrasts specified in this file         [
default:    NULL ]
  -b/--frag-bias-correct       use bias correction - reference fasta required        [
default:    NULL ]
  -u/--multi-read-correct      use 'rescue method' for multi-reads                   [
default:   FALSE ]
  -p/--num-threads             number of threads used during quantification          [
default:       1 ]
  --no-diff                    Don't generate differential analysis files            [
default:   FALSE ]
```

图 171 cuffdiff 功能的使用

```
  --no-js-tests                Don't perform isoform switching tests                 [
default:   FALSE ]
  -T/--time-series             treat samples as a time-series                        [
default:   FALSE ]
  --library-type               Library prep used for input reads                     [
default:   below ]
  --dispersion-method          Method used to estimate dispersion models             [
default:   below ]
  --library-norm-method        Method used to normalize library sizes                [
default:   below ]

Advanced Options:
  -m/--frag-len-mean           average fragment length (unpaired reads only)         [
default:     200 ]
  -s/--frag-len-std-dev        fragment length std deviation (unpaired reads only)   [
default:      80 ]
  -c/--min-alignment-count     minimum number of alignments in a locus for testing   [
default:    10 ]
  --max-mle-iterations         maximum iterations allowed for MLE calculation        [
default:    5000 ]
  --compatible-hits-norm       count hits compatible with reference RNAs only        [
default:    TRUE ]
  --total-hits-norm            count all hits for normalization                      [
default:   FALSE ]
  -v/--verbose                 log-friendly verbose processing (no progress bar)     [
default:   FALSE ]
```

图 172 cuffdiff 功能的使用

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
default:  FALSE ]
  --seed                          value of random number generator seed                [
default:       0 ]
  --no-update-check               do not contact server to check for update availability[
default:  FALSE ]
  --emit-count-tables             print count tables used to fit overdispersion        [
   DEPRECATED   ]
  --max-bundle-frags              maximum fragments allowed in a bundle before skipping [
default:  500000 ]
  --num-frag-count-draws          Number of fragment generation samples                [
default:     100 ]
  --num-frag-assign-draws         Number of fragment assignment samples per generation  [
default:      50 ]
  --max-frag-multihits            Maximum number of alignments allowed per fragment    [
default:  unlim  ]
  --min-outlier-p                 Min replicate p value to admit for testing           [
   DEPRECATED   ]
  --min-reps-for-js-test          Replicates needed for relative isoform shift testing  [
default:       3 ]
  --no-effective-length-correction   No effective length correction                    [
default:  FALSE ]
  --no-length-correction          No length correction                                 [
default:  FALSE ]
  -N/--upper-quartile-norm        Deprecated, use --library-norm-method                [
   DEPRECATED   ]
  --geometric-norm                Deprecated, use --library-norm-method                [
```

图 173 cuffdiff 功能的使用

```
   DEPRECATED   ]
  --raw-mapped-norm               Deprecated, use --library-norm-method                [
   DEPRECATED   ]
  --poisson-dispersion            Deprecated, use --dispersion-method                  [
   DEPRECATED   ]

Debugging use only:
  --read-skip-fraction            Skip a random subset of reads this size              [
default:     0.0 ]
  --no-read-pairs                 Break all read pairs                                 [
default:  FALSE ]
  --trim-read-length              Trim reads to be this long (keep 5' end)             [
default:    none ]
  --no-scv-correction             Disable SCV correction                               [
default:  FALSE ]

Supported library types:
        ff-firststrand
        ff-secondstrand
        ff-unstranded
        fr-firststrand
        fr-secondstrand
        fr-unstranded (default)
        transfrags

Supported dispersion methods:
```

图 174 cuffdiff 功能的使用

```
        blind
        per-condition
        poisson
        pooled (default)

Supported library normalization methods:
        classic-fpkm
        geometric (default)
        geometric
        quartile
```

图 175 cuffdiff 功能的使用

6、cuffnorm 一般是用来改善结果以便于进一步分析的，它能够将一系列的 RNA-seq 文库的表达水平数据标准化，去除这些数据之间的偏差源。cuffnorm 的具体功能和参数也可以通过直接输入 cuffnorm 进行查询（如图 176-图 178 所示）。

代码：

cuffnorm

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

#cuffnorm 功能使用帮助

```
[txy@localhost ~]$ cuffnorm -h
cuffnorm: unrecognized option '-h'
cuffnorm v2.2.1 (4237)
-----------------------------
Usage:   cuffnorm [options] <transcripts.gtf> <sample1_expr.cxb> <sample2_expr.cxb> [..
. sampleN_expr.cxb]
   Supply replicate CXB files as comma separated lists for each condition: sample1_rep1
.cxb,sample1_rep2.cxb,...sample1_repM.cxb
General Options:
  -o/--output-dir              write all output files to this directory              [
default:     ./ ]
  -L/--labels                  comma-separated list of condition labels
  --norm-standards-file        Housekeeping/spike genes to normalize libraries       [
default:    NULL ]
  -p/--num-threads             number of threads used during quantification          [
default:       1 ]
  --library-type               Library prep used for input reads                     [
default:   below ]
  --library-norm-method        Method used to normalize library sizes                [
default:   below ]
  --output-format              Format for output tables                              [
default:   below ]

Advanced Options:
  --compatible-hits-norm       count hits compatible with reference RNAs only        [
default:    TRUE ]
```

图 176 cuffnorm 功能的使用

```
  --total-hits-norm            count all hits for normalization                      [
default:   FALSE ]
  -v/--verbose                 log-friendly verbose processing (no progress bar)     [
default:   FALSE ]
  -q/--quiet                   log-friendly quiet processing (no progress bar)       [
default:   FALSE ]
  --seed                       value of random number generator seed                 [
default:       0 ]
  --no-update-check            do not contact server to check for update availability[
default:   FALSE ]

Supported library types:
        ff-firststrand
        ff-secondstrand
        ff-unstranded
        fr-firststrand
        fr-secondstrand
        fr-unstranded (default)
        transfrags

Supported library normalization methods:
        classic-fpkm
        geometric (default)
        geometric
        quartile
```

图 177 cuffnorm 功能的使用

```
Supported output formats:
        cuffdiff
        simple-table (default)
```

图 178 cuffnorm 功能的使用

7、gffread 是用来转换 gtf 格式文件和 gff 格式文件的命令，输入之后不会有反应，即并没有相应的 help 功能（如图 179 所示）。gtf 格式文件和 gff 格式文件有很多相同之处，差异并不是非常大，相互之间可以进行转换（如图 180 所示）。

```
[txy@localhost ~]$ gffread
```

图 179 gffread 功能的使用

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

列	GTF2	GFF3
reference sequence name	same	same
annotation source	same	same
feature type	feature requirements depend on software	can be anything
start coordinate	same	same
5. end coordinate	same	same
score	not used	optional
strand	same	same
frame	same	same
attributes	空格分隔	=分隔

图 180 gtf 格式文件和 gff 格式文件的相同之处与不同之处

4.3.2 gffread 的功能与使用方法

1、为了便于管理所下载的数据文件，首先在 cufflinks 文件夹下创建一个文件夹，命名为“shilishuju”（如图 181-图 183 所示）。

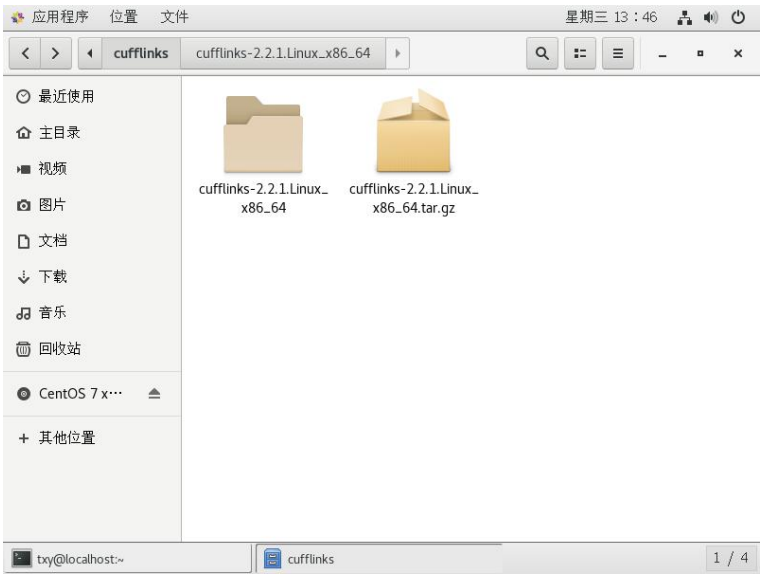


图 181 cufflinks 文件夹

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

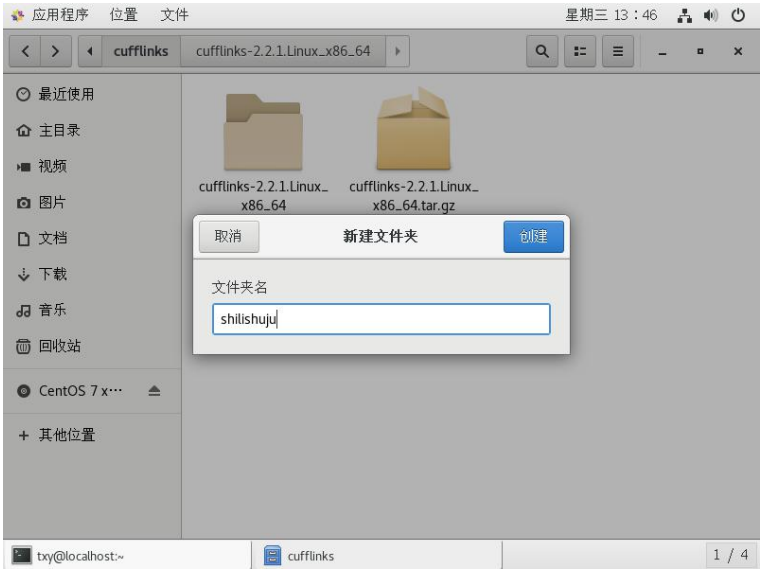


图 182 创建文件夹，命名为“shilishuju”

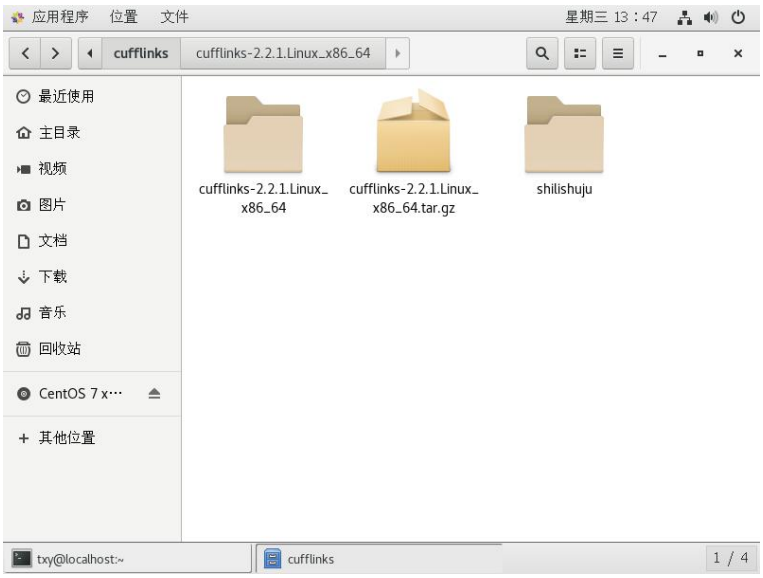


图 183 shilishuju 文件夹

2、从 NCBI 上下载基因组的 gff 格式文件，即注释文件（如图 184-图 196 所示）。

代码：

wget

ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_latest/refseq_identifiers/GRCh38_latest_genomic.gff.gz

#下载基因组注释文件

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
[txy@localhost ~]$ wget ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_l
atest/refseq_identifiers/GRCh38_latest_genomic.gff.gz
--2019-04-10 10:42:07--  ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_
latest/refseq_identifiers/GRCh38_latest_genomic.gff.gz
           => “GRCh38_latest_genomic.gff.gz”
正在解析主机 ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)... 130.14.250.12, 2607:f220:41
e:250::12
正在连接 ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)|130.14.250.12|:21... 已连接。
正在以 anonymous 登录 ... 登录成功！
==> SYST ... 完成。    ==> PWD ... 完成。
==> TYPE I ... 完成。  ==> CWD (1) /refseq/H_sapiens/annotation/GRCh38_latest/refseq_ide
ntifiers ... 完成。
==> SIZE GRCh38_latest_genomic.gff.gz ... 47729303
==> PASV ... 完成。    ==> RETR GRCh38_latest_genomic.gff.gz ... 完成。
长度：47729303 (46M) (非正式数据)

 3% [>                                                ] 1,757,792   16.8KB/s 用时 83s

2019-04-10 10:43:36 (20.6 KB/s) - 已关闭控制连接。
重试中。

--2019-04-10 10:43:37--  ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_
latest/refseq_identifiers/GRCh38_latest_genomic.gff.gz
  (尝试次数：2) => “GRCh38_latest_genomic.gff.gz”
正在连接 ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)|130.14.250.12|:21... 已连接。
正在以 anonymous 登录 ... 登录成功！
```

图 184 下载注释文件

```
==> SYST ... 完成。    ==> PWD ... 完成。
==> TYPE I ... 完成。  ==> CWD (1) /refseq/H_sapiens/annotation/GRCh38_latest/refseq_ide
ntifiers ... 完成。
==> SIZE GRCh38_latest_genomic.gff.gz ... 47729303
==> PASV ... 完成。    ==> REST 1757792 ... 完成。
==> RETR GRCh38_latest_genomic.gff.gz ... 完成。
长度：47729303 (46M)，剩余 45971511 (44M) (非正式数据)

 5% [+>                                               ] 2,762,052   11.6KB/s 用时 69s

2019-04-10 10:44:50 (14.2 KB/s) - 已关闭控制连接。
重试中。

--2019-04-10 10:44:52--  ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_
latest/refseq_identifiers/GRCh38_latest_genomic.gff.gz
  (尝试次数：3) => “GRCh38_latest_genomic.gff.gz”
正在连接 ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)|130.14.250.12|:21... 已连接。
正在以 anonymous 登录 ... 登录成功！
==> SYST ... 完成。    ==> PWD ... 完成。
==> TYPE I ... 完成。  ==> CWD (1) /refseq/H_sapiens/annotation/GRCh38_latest/refseq_ide
ntifiers ... 完成。
==> SIZE GRCh38_latest_genomic.gff.gz ... 47729303
==> PASV ... 完成。    ==> REST 2762052 ... 完成。
==> RETR GRCh38_latest_genomic.gff.gz ... 完成。
长度：47729303 (46M)，剩余 44967251 (43M) (非正式数据)
```

图 185 下载注释文件

```
 8% [++>                                              ] 3,957,728   7.02KB/s 用时 71s

2019-04-10 10:46:08 (16.4 KB/s) - 已关闭控制连接。
重试中。

--2019-04-10 10:46:11--  ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_
latest/refseq_identifiers/GRCh38_latest_genomic.gff.gz
  (尝试次数：4) => “GRCh38_latest_genomic.gff.gz”
正在连接 ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)|130.14.250.12|:21... 已连接。
正在以 anonymous 登录 ... 登录成功！
==> SYST ... 完成。    ==> PWD ... 完成。
==> TYPE I ... 完成。  ==> CWD (1) /refseq/H_sapiens/annotation/GRCh38_latest/refseq_ide
ntifiers ... 完成。
==> SIZE GRCh38_latest_genomic.gff.gz ... 47729303
==> PASV ... 完成。    ==> REST 3957728 ... 完成。
==> RETR GRCh38_latest_genomic.gff.gz ... 完成。
长度：47729303 (46M)，剩余 43771575 (42M) (非正式数据)

11% [+++=>                                           ] 5,405,480   15.6KB/s 用时 70s

2019-04-10 10:47:26 (20.3 KB/s) - 已关闭控制连接。
重试中。

--2019-04-10 10:47:30--  ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_
latest/refseq_identifiers/GRCh38_latest_genomic.gff.gz
  (尝试次数：5) => “GRCh38 latest genomic.gff.gz”
```

图 186 下载注释文件

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
正在连接 ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)|130.14.250.12|:21... 已连接。
正在以 anonymous 登录 ... 登录成功！
==> SYST ... 完成。    ==> PWD ... 完成。
==> TYPE I ... 完成。  ==> CWD (1) /refseq/H_sapiens/annotation/GRCh38_latest/refseq_ide
ntifiers ... 完成。
==> SIZE GRCh38_latest_genomic.gff.gz ... 47729303
==> PASV ... 完成。    ==> REST 5405480 ... 完成。
==> RETR GRCh38_latest_genomic.gff.gz ... 完成。
长度：47729303 (46M)，剩余 42323823 (40M) (非正式数据)

14% [+++++>                                          ] 7,137,660   31.9KB/s 用时 67s

2019-04-10 10:48:41 (25.4 KB/s) - 已关闭控制连接。
重试中。

--2019-04-10 10:48:46--  ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_
latest/refseq_identifiers/GRCh38_latest_genomic.gff.gz
  (尝试次数：6) => “GRCh38_latest_genomic.gff.gz”
正在连接 ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)|130.14.250.12|:21... 已连接。
正在以 anonymous 登录 ... 登录成功！
==> SYST ... 完成。    ==> PWD ... 完成。
==> TYPE I ... 完成。  ==> CWD (1) /refseq/H_sapiens/annotation/GRCh38_latest/refseq_ide
ntifiers ... 完成。
==> SIZE GRCh38_latest_genomic.gff.gz ... 47729303
==> PASV ... 完成。    ==> REST 7137660 ... 完成。
==> RETR GRCh38_latest_genomic.gff.gz ... 完成。
```

图 187 下载注释文件

```
长度：47729303 (46M)，剩余 40591643 (39M) (非正式数据)

19% [++++++>                                         ] 9,094,956   50.7KB/s 用时 66s

2019-04-10 10:49:56 (29.1 KB/s) - 已关闭控制连接。
重试中。

--2019-04-10 10:50:02--  ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_
latest/refseq_identifiers/GRCh38_latest_genomic.gff.gz
  (尝试次数：7) => “GRCh38_latest_genomic.gff.gz”
正在连接 ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)|130.14.250.12|:21... 已连接。
正在以 anonymous 登录 ... 登录成功！
==> SYST ... 完成。    ==> PWD ... 完成。
==> TYPE I ... 完成。  ==> CWD (1) /refseq/H_sapiens/annotation/GRCh38_latest/refseq_ide
ntifiers ... 完成。
==> SIZE GRCh38_latest_genomic.gff.gz ... 47729303
==> PASV ... 完成。    ==> REST 9094956 ... 完成。
==> RETR GRCh38_latest_genomic.gff.gz ... 完成。
长度：47729303 (46M)，剩余 38634347 (37M) (非正式数据)

33% [+++++++=======>                                 ] 15,977,844  17.9KB/s 用时 3m 5s

2019-04-10 10:53:11 (36.4 KB/s) - 已关闭控制连接。
重试中。

--2019-04-10 10:53:18--  ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_
```

图 188 下载注释文件

```
latest/refseq_identifiers/GRCh38_latest_genomic.gff.gz
  (尝试次数：8) => “GRCh38_latest_genomic.gff.gz”
正在连接 ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)|130.14.250.12|:21... 已连接。
正在以 anonymous 登录 ... 登录成功！
==> SYST ... 完成。    ==> PWD ... 完成。
==> TYPE I ... 完成。  ==> CWD (1) /refseq/H_sapiens/annotation/GRCh38_latest/refseq_ide
ntifiers ... 完成。
==> SIZE GRCh38_latest_genomic.gff.gz ... 47729303
==> PASV ... 完成。    ==> REST 15977844 ... 完成。
==> RETR GRCh38_latest_genomic.gff.gz ... 完成。
长度：47729303 (46M)，剩余 31751459 (30M) (非正式数据)

42% [++++++++++++++==>                               ] 20,178,212  24.2KB/s 用时 2m 16s

2019-04-10 10:55:39 (30.1 KB/s) - 已关闭控制连接。
重试中。

--2019-04-10 10:55:47--  ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_
latest/refseq_identifiers/GRCh38_latest_genomic.gff.gz
  (尝试次数：9) => “GRCh38_latest_genomic.gff.gz”
正在连接 ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)|130.14.250.12|:21... 已连接。
正在以 anonymous 登录 ... 登录成功！
==> SYST ... 完成。    ==> PWD ... 完成。
==> TYPE I ... 完成。  ==> CWD (1) /refseq/H_sapiens/annotation/GRCh38_latest/refseq_ide
ntifiers ... 完成。
==> SIZE GRCh38_latest_genomic.gff.gz ... 47729303
```

图 189 下载注释文件

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
==> PASV ... 完成。    ==> REST 20178212 ... 完成。
==> RETR GRCh38_latest_genomic.gff.gz ... 完成。
长度：47729303 (46M)，剩余 27551091 (26M) (非正式数据)

46% [++++++++++++++++++++=>                          ] 22,116,636  17.2KB/s 用时 71s

2019-04-10 10:57:02 (26.8 KB/s) - 已关闭控制连接。
重试中。

--2019-04-10 10:57:11--  ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_
latest/refseq_identifiers/GRCh38_latest_genomic.gff.gz
  (尝试次数：10) => “GRCh38_latest_genomic.gff.gz”
正在连接 ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)|130.14.250.12|:21... 已连接。
正在以 anonymous 登录 ... 登录成功！
==> SYST ... 完成。    ==> PWD ... 完成。
==> TYPE I ... 完成。  ==> CWD (1) /refseq/H_sapiens/annotation/GRCh38_latest/refseq_ide
ntifiers ... 完成。
==> SIZE GRCh38_latest_genomic.gff.gz ... 47729303
==> PASV ... 完成。    ==> REST 22116636 ... 完成。
==> RETR GRCh38_latest_genomic.gff.gz ... 完成。
长度：47729303 (46M)，剩余 25612667 (24M) (非正式数据)

51% [++++++++++++++++++++++=>                        ] 24,525,512  51.3KB/s 用时 68s

2019-04-10 10:58:24 (34.7 KB/s) - 已关闭控制连接。
重试中。
```

图 190 下载注释文件

```
--2019-04-10 10:58:34--  ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_
latest/refseq_identifiers/GRCh38_latest_genomic.gff.gz
  (尝试次数：11) => “GRCh38_latest_genomic.gff.gz”
正在连接 ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)|130.14.250.12|:21... 已连接。
正在以 anonymous 登录 ... 登录成功！
==> SYST ... 完成。    ==> PWD ... 完成。
==> TYPE I ... 完成。  ==> CWD (1) /refseq/H_sapiens/annotation/GRCh38_latest/refseq_ide
ntifiers ... 完成。
==> SIZE GRCh38_latest_genomic.gff.gz ... 47729303
==> PASV ... 完成。    ==> REST 24525512 ... 完成。
==> RETR GRCh38_latest_genomic.gff.gz ... 完成。
长度：47729303 (46M)，剩余 23203791 (22M) (非正式数据)

55% [++++++++++++++++++++++=>                        ] 26,527,292  16.9KB/s 用时 71s

2019-04-10 10:59:50 (27.4 KB/s) - 已关闭控制连接。
重试中。

--2019-04-10 11:00:00--  ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_
latest/refseq_identifiers/GRCh38_latest_genomic.gff.gz
  (尝试次数：12) => “GRCh38_latest_genomic.gff.gz”
正在连接 ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)|130.14.250.12|:21... 已连接。
正在以 anonymous 登录 ... 登录成功！
==> SYST ... 完成。    ==> PWD ... 完成。
==> TYPE I ... 完成。  ==> CWD (1) /refseq/H_sapiens/annotation/GRCh38_latest/refseq_ide
ntifiers ... 完成。
```

图 191 下载注释文件

```
==> SIZE GRCh38_latest_genomic.gff.gz ... 47729303
==> PASV ... 完成。    ==> REST 26527292 ... 完成。
==> RETR GRCh38_latest_genomic.gff.gz ... 完成。
长度：47729303 (46M)，剩余 21202011 (20M) (非正式数据)

60% [++++++++++++++++++++++++=>                      ] 28,702,964  38.5KB/s 用时 68s

2019-04-10 11:01:14 (31.3 KB/s) - 已关闭控制连接。
重试中。

--2019-04-10 11:01:24--  ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_
latest/refseq_identifiers/GRCh38_latest_genomic.gff.gz
  (尝试次数：13) => “GRCh38_latest_genomic.gff.gz”
正在连接 ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)|130.14.250.12|:21... 已连接。
正在以 anonymous 登录 ... 登录成功！
==> SYST ... 完成。    ==> PWD ... 完成。
==> TYPE I ... 完成。  ==> CWD (1) /refseq/H_sapiens/annotation/GRCh38_latest/refseq_ide
ntifiers ... 完成。
==> SIZE GRCh38_latest_genomic.gff.gz ... 47729303
==> PASV ... 完成。    ==> REST 28702964 ... 完成。
==> RETR GRCh38_latest_genomic.gff.gz ... 完成。
长度：47729303 (46M)，剩余 19026339 (18M) (非正式数据)

62% [+++++++++++++++++++++++++=>                     ] 29,839,328  23.7KB/s 用时 69s

2019-04-10 11:02:38 (16.2 KB/s) - 已关闭控制连接。
```

图 192 下载注释文件

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
重试中。

--2019-04-10 11:02:48--  ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_
latest/refseq_identifiers/GRCh38_latest_genomic.gff.gz
  (尝试次数：14) => “GRCh38_latest_genomic.gff.gz”
正在连接 ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)|130.14.250.12|:21... 已连接。
正在以 anonymous 登录 ... 登录成功！
==> SYST ... 完成。    ==> PWD ... 完成。
==> TYPE I ... 完成。  ==> CWD (1) /refseq/H_sapiens/annotation/GRCh38_latest/refseq_ide
ntifiers ... 完成。
==> SIZE GRCh38_latest_genomic.gff.gz ... 47729303
==> PASV ... 完成。    ==> REST 29839328 ... 完成。
==> RETR GRCh38_latest_genomic.gff.gz ... 完成。
长度：47729303 (46M)，剩余 17889975 (17M) (非正式数据)

67% [++++++++++++++++++++++++++++==>              ] 32,289,992  58.0KB/s 用时 69s

2019-04-10 11:04:05 (34.6 KB/s) - 已关闭控制连接。
重试中。

--2019-04-10 11:04:15--  ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_
latest/refseq_identifiers/GRCh38_latest_genomic.gff.gz
  (尝试次数：15) => “GRCh38_latest_genomic.gff.gz”
正在连接 ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)|130.14.250.12|:21... 已连接。
正在以 anonymous 登录 ... 登录成功！
==> SYST ... 完成。    ==> PWD ... 完成。
```

图 193 下载注释文件

```
==> TYPE I ... 完成。  ==> CWD (1) /refseq/H_sapiens/annotation/GRCh38_latest/refseq_ide
ntifiers ... 完成。
==> SIZE GRCh38_latest_genomic.gff.gz ... 47729303
==> PASV ... 完成。    ==> REST 32289992 ... 完成。
==> RETR GRCh38_latest_genomic.gff.gz ... 完成。
长度：47729303 (46M)，剩余 15439311 (15M) (非正式数据)

86% [++++++++++++++++++++++++++++++=========>       ] 41,387,644  19.2KB/s 用时 3m 51s

2019-04-10 11:08:10 (38.5 KB/s) - 已关闭控制连接。
重试中。

--2019-04-10 11:08:20--  ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_
latest/refseq_identifiers/GRCh38_latest_genomic.gff.gz
  (尝试次数：16) => “GRCh38_latest_genomic.gff.gz”
正在连接 ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)|130.14.250.12|:21... 已连接。
正在以 anonymous 登录 ... 登录成功！
==> SYST ... 完成。    ==> PWD ... 完成。
==> TYPE I ... 完成。  ==> CWD (1) /refseq/H_sapiens/annotation/GRCh38_latest/refseq_ide
ntifiers ... 完成。
==> SIZE GRCh38_latest_genomic.gff.gz ... 47729303
==> PASV ... 完成。    ==> REST 41387644 ... 完成。
==> RETR GRCh38_latest_genomic.gff.gz ... 完成。
长度：47729303 (46M)，剩余 6341659 (6.0M) (非正式数据)

94% [++++++++++++++++++++++++++++++++++++++++==>  ] 45,333,240  21.8KB/s 用时 2m 12s
```

图 194 下载注释文件

```
2019-04-10 11:10:37 (29.1 KB/s) - 已关闭控制连接。
重试中。

--2019-04-10 11:10:47--  ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_
latest/refseq_identifiers/GRCh38_latest_genomic.gff.gz
  (尝试次数：17) => “GRCh38_latest_genomic.gff.gz”
正在连接 ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)|130.14.250.12|:21... 已连接。
正在以 anonymous 登录 ... 登录成功！
==> SYST ... 完成。    ==> PWD ... 完成。
==> TYPE I ... 完成。  ==> CWD (1) /refseq/H_sapiens/annotation/GRCh38_latest/refseq_ide
ntifiers ... 完成。
==> SIZE GRCh38_latest_genomic.gff.gz ... 47729303
==> PASV ... 完成。    ==> REST 45333240 ... 完成。
==> RETR GRCh38_latest_genomic.gff.gz ... 完成。
长度：47729303 (46M)，剩余 2396063 (2.3M) (非正式数据)

100%[++++++++++++++++++++++++++++++++++++++++++++=>] 47,729,303  76.5KB/s 用时 53s

2019-04-10 11:11:45 (44.4 KB/s) - “GRCh38_latest_genomic.gff.gz”已保存 [47729303]
```

图 195 下载注释文件

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

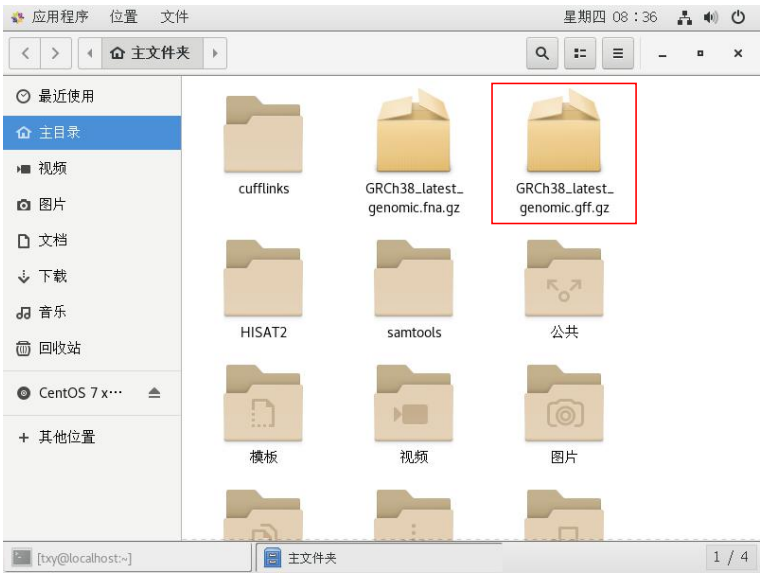


图 196 下载得到的注释文件

2、解压缩注释文件（如图 197 和图 198 所示）

代码：

```
gunzip /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.gff.gz
#解压缩注释文件
```



图 197 解压缩注释文件

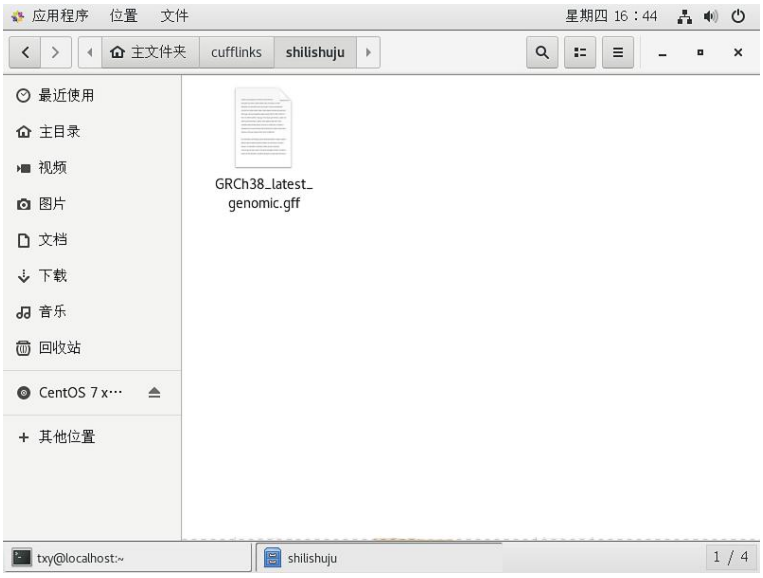


图 198 解压缩得到的注释文件

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

3、gff 格式文件转换成 gtf 格式文件（如图 199 和图 200 所示）。

代码：

```
gffread /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.gff -T -o GRCh38_latest_genomic.gtf
#“/home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.gff”表示输入的 gff 格式文件及其路径。
#“-T”表示转换命令。
#“-o”表示输出。
#“GRCh38_latest_genomic.gtf”表示输出的文件名称。
```

```
[txy@localhost ~]$ gffread /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.gff -T
-o GRCh38_latest_genomic.gtf
[txy@localhost ~]$
```

图 199 将 gff 格式文件转换成 gtf 格式文件

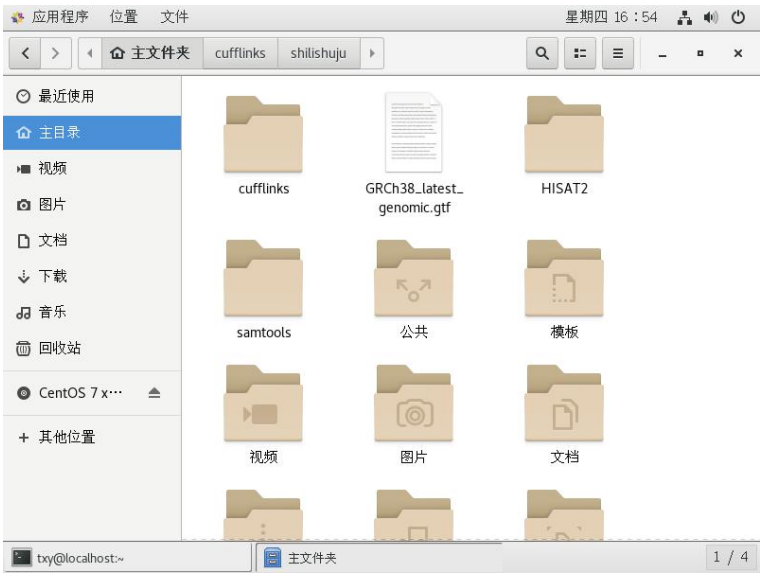


图 200 转换得到的 gtf 格式文件

4、gtf 格式文件转换成 gff 格式文件（如图 201 和图 202 所示）。

代码：

```
gffread /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.gtf -o- > GRCh38_latest_genomic.gff
#“/home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.gtf”表示输入的 gtf 格式文件。
```

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

#“-o”表示输出。

#“->”表示转换命令。

#“GRCh38_latest_genomic.gff”表示输出的文件名称。

```
[txy@localhost ~]$ gffread /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.gtf -o-
 > GRCh38_latest_genomic.gff
[txy@localhost ~]$
```

图 201 将 gtf 格式文件转换成 gff 格式文件

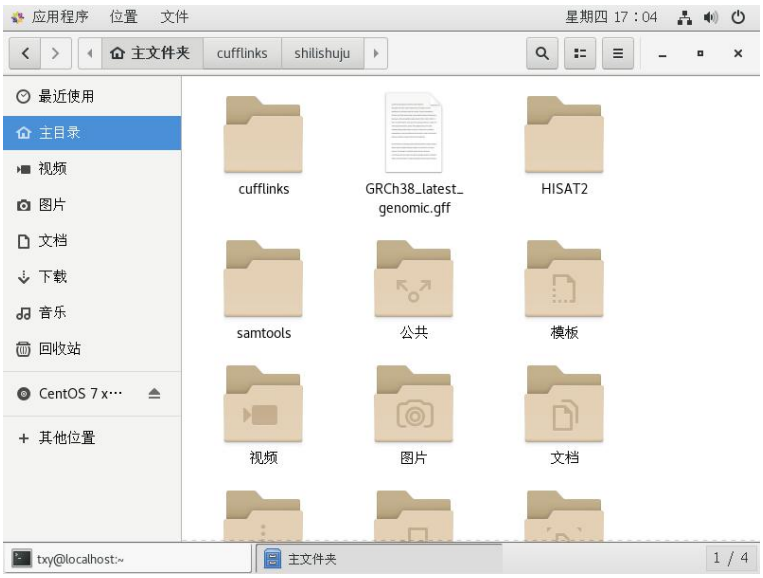


图 202 转换得到的 gff 格式文件

4.3.3 cufflinks 的功能与使用方法

1、通过教程前面的内容流程或者其他方式，获得 4 个 sam 格式文件和 4 个 bam 格式文件以及一个 gtf 格式文件作为示例数据，其中假设这 4 个 bam 格式文件是两个需要进行比较样品，每个样品有 2 个重复，4 个 sam 格式文件亦是如此。（如图 203 所示）。

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

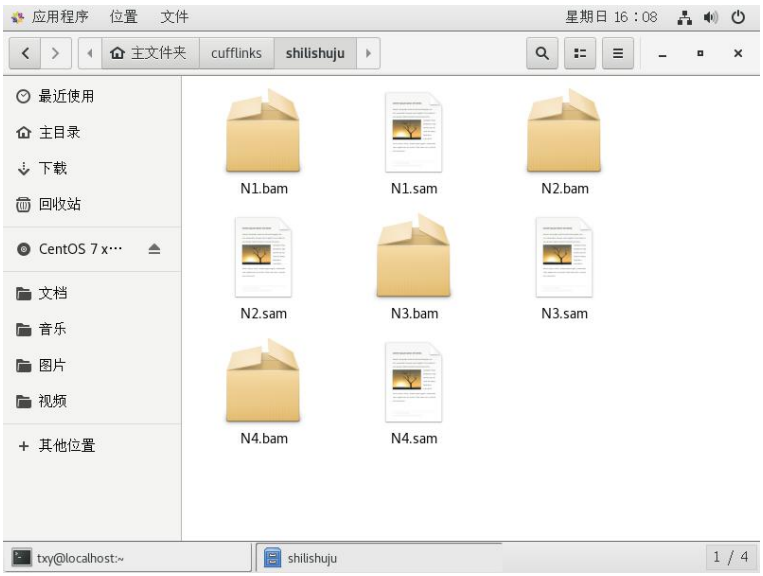


图 203 准备好的 sam 格式文件，bam 格式文件和 gtf 格式文件

2、针对 8 个 sam 格式文件在 shilishuju 文件夹下分别建立四个输出文件夹（如图 204-图 208 所示）。

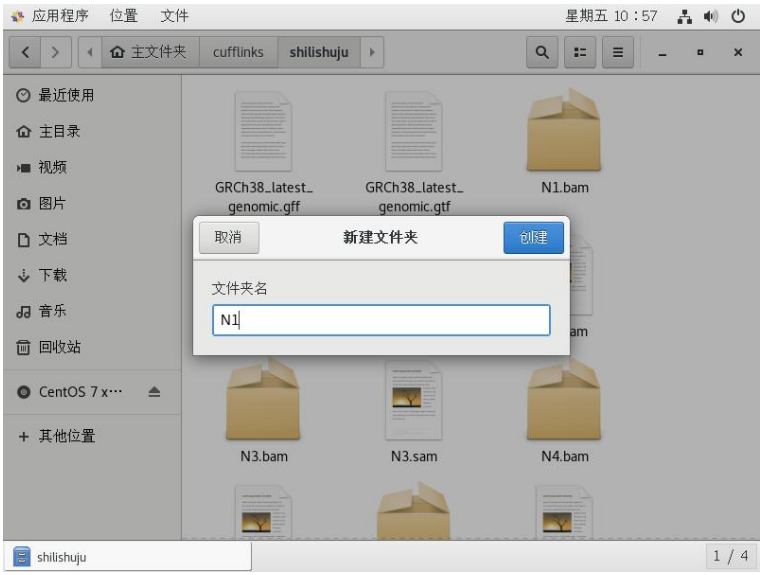


图 204 建立输出文件夹，命名为“N1”

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

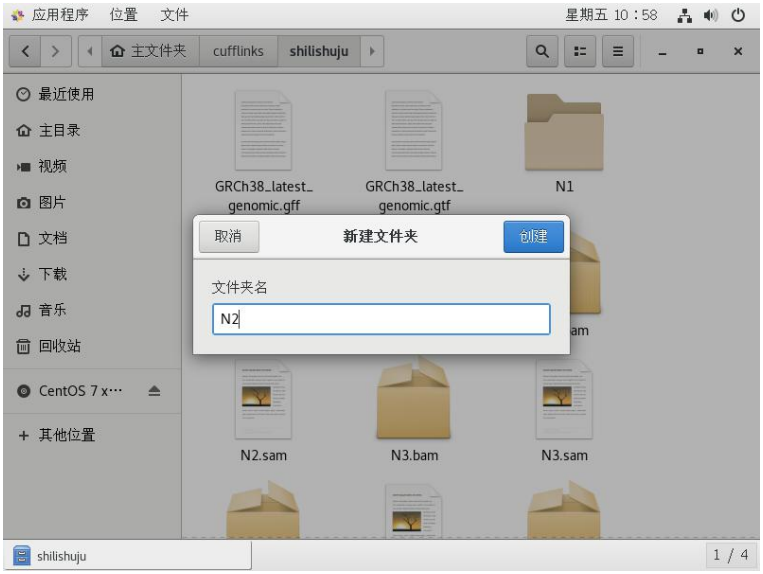


图 205 建立输出文件夹，命名为“N2”

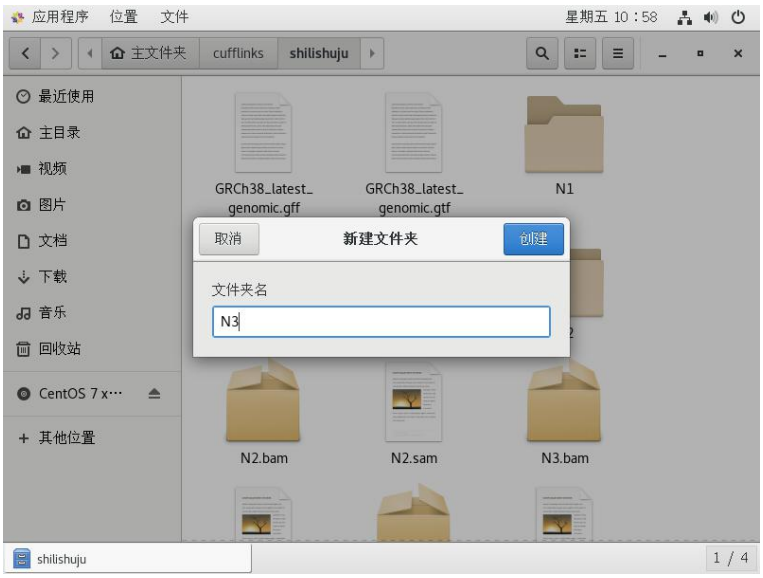


图 206 建立输出文件夹，命名为“N3”

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

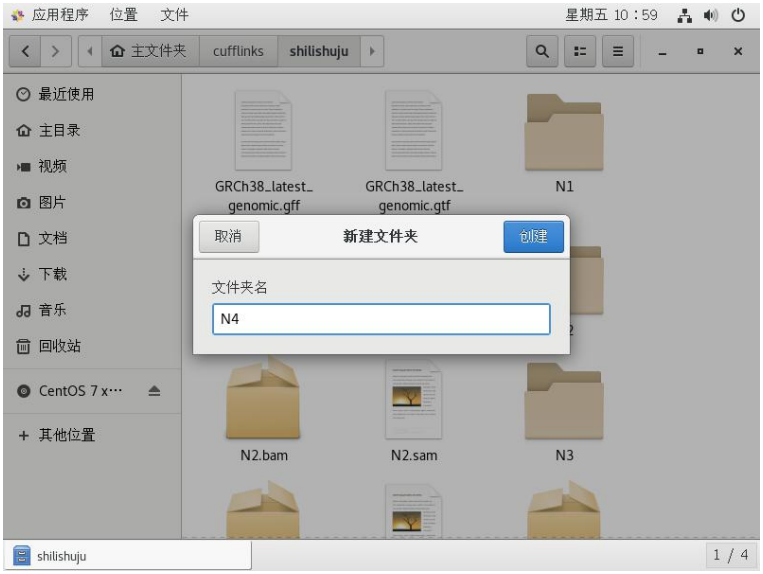


图 207 建立输出文件夹，命名为“N4”

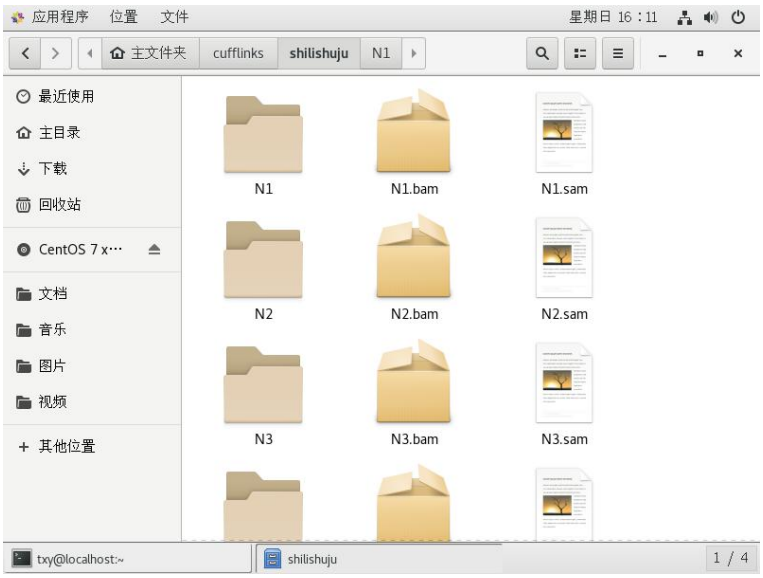


图 208 四个输出文件夹建立完成

2、对 N1 使用 cufflinks 命令，cufflinks 拼接转录本，在拼接处尽可能达到最优结构，并且能够计算表达量和差异表达。

（1）使用 bam 格式文件（如图 209 和图 210 所示）。

代码：

```
cufflinks -u -g /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.gtf -o /home/txy/cufflinks/shilishuju/N1/bam /home/txy/cufflinks/shilishuju/N1.bam
#“-u”告诉 Cufflinks 用更准确的方法去处理贴到多个位点上的 reads，如果没有
```

未经授权，不得用于任何商业用途！
联系方式：bnlt@zju.edu.cn

“-u”，Cufflinks 只会把这些 reads 简单地平均分配。比如一个 read 贴到了 10 个位置，那么每个位置分得十分之一。加-u 后会先进行平均分配，然后按照这 10 个位置各自的表达量，计算 read 被分配到每个位置的概率。

#“-g”表示提供一个注释文件，但是 Cufflinks 会在这些已知转录本的指导下，拼接新的转录本，可以不用。

#“-o”表示输出。

#“/home/txy/cufflinks/shilishuju/N1/bam”表示输出的路径。

#“/home/txy/cufflinks/shilishuju/N1.bam”表示输入的 bam 格式文件名称及其路径。

```
[txy@localhost ~]$ cufflinks -u -g /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic
.gtf -o /home/txy/cufflinks/shilishuju/N1/bam /home/txy/cufflinks/shilishuju/N1.bam
Warning: Could not connect to update server to verify current version. Please check at
the Cufflinks website (http://cufflinks.cbcb.umd.edu).
[11:13:51]  Loading reference annotation.
[11:14:07]  Inspecting reads and determining fragment length distribution.
> Processed 35412 loci.                        [*************************] 100%
Warning: Using default Gaussian distribution due to insufficient paired-end reads in op
en ranges.  It is recommended that correct parameters (--frag-len-mean and --frag-len-s
td-dev) be provided.
> Map Properties:
>       Normalized Map Mass: 383.00
>       Raw Map Mass: 383.00
>       Number of Multi-Reads: 0 (with 0 total hits)
>       Fragment Length Distribution: Truncated Gaussian (default)
>                     Default Mean: 200
>                  Default Std Dev: 80
[11:14:10]  Assembling transcripts and initializing abundances for multi-read correction
.
> Processed 35412 loci.                        [*************************] 100%
[11:19:16]  Loading reference annotation.
[11:19:45]  Re-estimating abundances with multi-read correction.
> Processed 35412 loci.                        [*************************] 100%
```

图 209 对 N1 的 bam 格式文件使用 cufflinks 命令

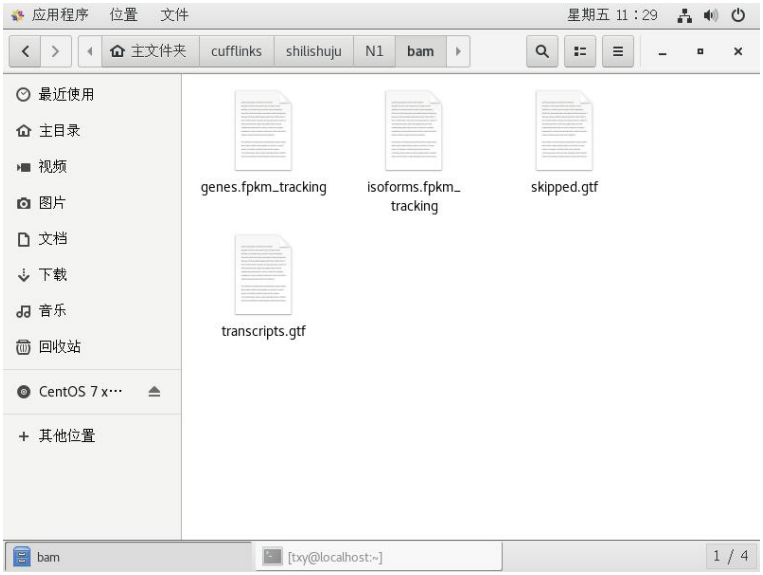


图 210 对 N1 的 bam 格式文件使用 cufflinks 命令得到的输出文件

（2）使用 sam 格式文件（如图 211-图 213 所示）。

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

代码：

cufflinks -u -g /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.gtf -o /home/txy/cufflinks/shilishuju/N1/sam /home/txy/cufflinks/shilishuju/N1.sam

#“-u”告诉 Cufflinks 用更准确的方法去处理贴到多个位点上的 reads，如果没有“-u”，Cufflinks 只会把这些 reads 简单地平均分配。比如一个 read 贴到了 10 个位置，那么每个位置分得十分之一。加-u 后会先进行平均分配，然后按照这 10 个位置各自的表达量，计算 read 被分配到每个位置的概率。

#“-g”表示提供一个注释文件，但是 Cufflinks 会在这些已知转录本的指导下，拼接新的转录本，可以不用。

#“-o”表示输出。

#“/home/txy/cufflinks/shilishuju/N1/sam”表示输出的路径。

#“/home/txy/cufflinks/shilishuju/N1.sam”表示输入的 sam 格式文件名称及其路径。

```
[txy@localhost ~]$ cufflinks -u -g /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic
.gtf -o /home/txy/cufflinks/shilishuju/N1/sam /home/txy/cufflinks/shilishuju/N1.sam
Warning: Could not connect to update server to verify current version. Please check at
the Cufflinks website (http://cufflinks.cbcb.umd.edu).
[bam_header_read] EOF marker is absent. The input is probably truncated.
[bam_header_read] invalid BAM binary header (this is not a BAM file).
File /home/txy/cufflinks/shilishuju/N1.sam doesn't appear to be a valid BAM file, tryin
g SAM...
[12:04:05] Loading reference annotation.
[12:04:52] Inspecting reads and determining fragment length distribution.
> Processed 35412 loci.                         [*************************] 100%
Warning: Using default Gaussian distribution due to insufficient paired-end reads in op
en ranges.  It is recommended that correct parameters (--frag-len-mean and --frag-len-s
td-dev) be provided.
> Map Properties:
>       Normalized Map Mass: 383.00
>       Raw Map Mass: 383.00
>       Number of Multi-Reads: 0 (with 0 total hits)
>       Fragment Length Distribution: Truncated Gaussian (default)
>                     Default Mean: 200
>                  Default Std Dev: 80
[12:04:55] Assembling transcripts and initializing abundances for multi-read correction
.
> Processed 35412 loci.                         [*************************] 100%
[12:12:07] Loading reference annotation.
[12:12:38] Re-estimating abundances with multi-read correction.
```

图 211 对 N1 的 sam 格式文件使用 cufflinks 命令

```
> Processed 35412 loci.                         [*************************] 100%
```

图 212 对 N1 的 sam 格式文件使用 cufflinks 命令

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

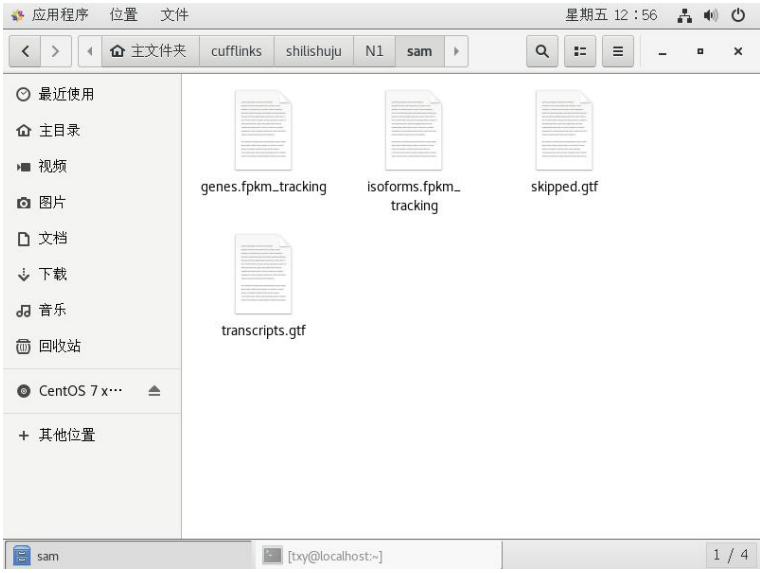


图 213 对 N1 的 sam 文件使用 cufflinks 命令得到的输出文件

3、对 N2 使用 cufflinks 命令，cufflinks 拼接转录本，在拼接处尽可能达到最优结构，并且能够计算表达量和差异表达。

（1）使用 bam 格式文件（如图 214 和图 215 所示）。

代码：

cufflinks -u -g /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.gtf -o /home/txy/cufflinks/shilishuju/N2/bam /home/txy/cufflinks/shilishuju/N2.bam

#"-u"告诉 Cufflinks 用更准确的方法去处理贴到多个位点上的 reads，如果没有"-u"，Cufflinks 只会把这些 reads 简单地平均分配。比如一个 read 贴到了 10 个位置，那么每个位置分得十分之一。加-u 后会先进行平均分配，然后按照这 10 个位置各自的表达量，计算 read 被分配到每个位置的概率。

#"-g"表示提供一个注释文件，但是 Cufflinks 会在这些已知转录本的指导下，拼接新的转录本，可以不用。

#"-o"表示输出。

#"/home/txy/cufflinks/shilishuju/N2/bam"表示输出的路径。

#"/home/txy/cufflinks/shilishuju/N2.bam"表示输出的 bam 文件名称及其路径。

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
[txy@localhost ~]$ cufflinks -u -g /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic
.gtf -o /home/txy/cufflinks/shilishuju/N2/bam /home/txy/cufflinks/shilishuju/N2.bam
Warning: Could not connect to update server to verify current version. Please check at
the Cufflinks website (http://cufflinks.cbcb.umd.edu).
[13:04:20]  Loading reference annotation.
[13:05:09]  Inspecting reads and determining fragment length distribution.
> Processed 35412 loci.                        [*************************] 100%
Warning: Using default Gaussian distribution due to insufficient paired-end reads in op
en ranges.  It is recommended that correct parameters (--frag-len-mean and --frag-len-s
td-dev) be provided.
> Map Properties:
>       Normalized Map Mass: 609.50
>       Raw Map Mass: 609.50
>       Number of Multi-Reads: 0 (with 0 total hits)
>       Fragment Length Distribution: Truncated Gaussian (default)
>                     Default Mean: 200
>                  Default Std Dev: 80
[13:05:12]  Assembling transcripts and initializing abundances for multi-read correction
.
> Processed 35412 loci.                        [*************************] 100%
[13:11:16]  Loading reference annotation.
[13:11:47]  Re-estimating abundances with multi-read correction.
> Processed 35412 loci.                        [*************************] 100%
```

图 214 对 N2 的 bam 格式文件使用 cufflinks 命令

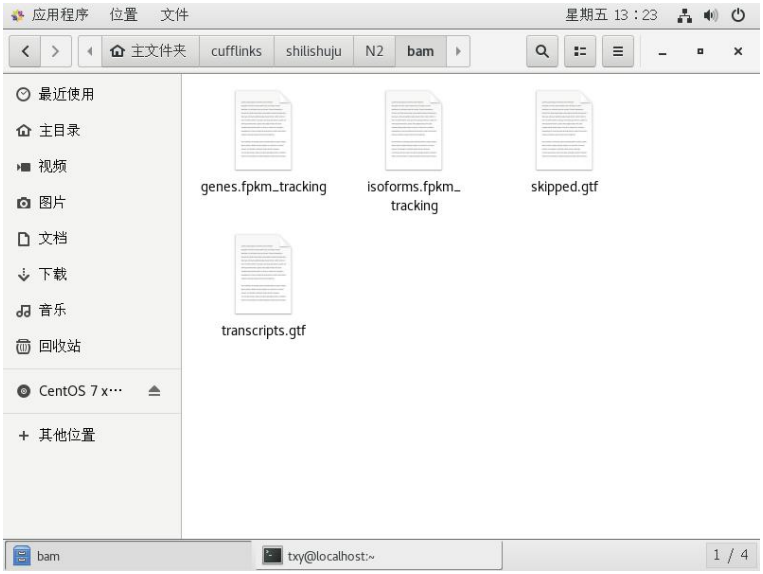


图 215 对 N2 的 bam 格式文件使用 cufflinks 命令得到的输出文件

（2）使用 sam 格式文件（如图 216-图 218 所示）。

代码：

cufflinks -u -g /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.gtf -o /home/txy/cufflinks/shilishuju/N2/sam /home/txy/cufflinks/shilishuju/N2.sam

#"-u"告诉 Cufflinks 用更准确的方法去处理贴到多个位点上的 reads，如果没有"-u"，Cufflinks 只会把这些 reads 简单地平均分配。比如一个 read 贴到了 10 个位置，那么每个位置分得十分之一。加-u 后会先进行平均分配，然后按照这 10 个位置各自的表达量，计算 read 被分配到每个位置的概率。

#"-g"表示提供一个注释文件，但是 Cufflinks 会在这些已知转录本的指导下，拼

联系方式：binfo@zju.edu.cn

接新的转录本，可以不用。

#“-o”表示输出。

#“/home/txy/cufflinks/shilishuju/N2/sam”表示输出的路径。

#“/home/txy/cufflinks/shilishuju/N2.sam”表示输入的 sam 格式文件名称及其路径。

```
[txy@localhost ~]$ cufflinks -u -g /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic
.gtf -o /home/txy/cufflinks/shilishuju/N2/sam /home/txy/cufflinks/shilishuju/N2.sam
Warning: Could not connect to update server to verify current version. Please check at
the Cufflinks website (http://cufflinks.cbcb.umd.edu).
[bam_header_read] EOF marker is absent. The input is probably truncated.
[bam_header_read] invalid BAM binary header (this is not a BAM file).
File /home/txy/cufflinks/shilishuju/N2.sam doesn't appear to be a valid BAM file, tryin
g SAM...
[13:24:21] Loading reference annotation.
[13:24:36] Inspecting reads and determining fragment length distribution.
> Processed 35412 loci.                        [*************************] 100%
Warning: Using default Gaussian distribution due to insufficient paired-end reads in op
en ranges.  It is recommended that correct parameters (--frag-len-mean and --frag-len-s
td-dev) be provided.
> Map Properties:
>       Normalized Map Mass: 609.50
>       Raw Map Mass: 609.50
>       Number of Multi-Reads: 0 (with 0 total hits)
>       Fragment Length Distribution: Truncated Gaussian (default)
>                     Default Mean: 200
>                  Default Std Dev: 80
[13:24:39] Assembling transcripts and initializing abundances for multi-read correction
.
> Processed 35412 loci.                        [*************************] 100%
[13:29:58] Loading reference annotation.
[13:30:28] Re-estimating abundances with multi-read correction.
```

图 216 对 N2 的 sam 格式文件使用 cufflinks 命令

```
> Processed 35412 loci.                        [*************************] 100%
```

图 217 对 N2 的 sam 格式文件使用 cufflinks 命令

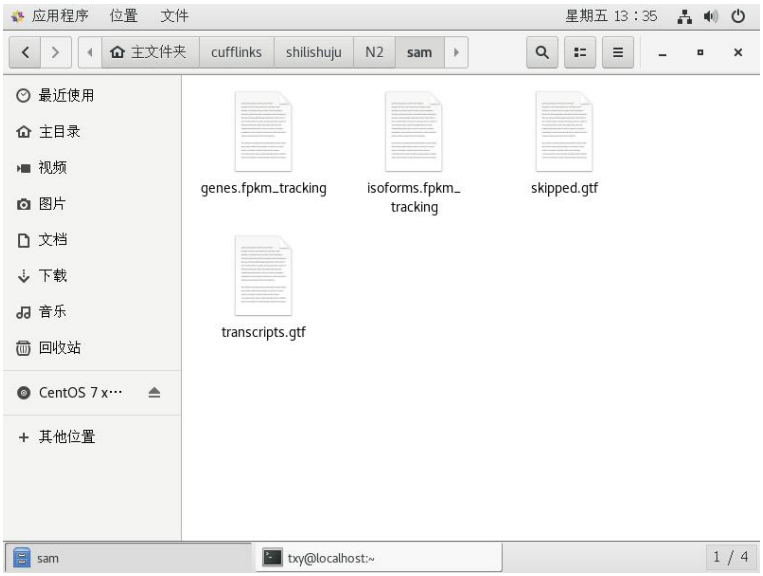


图 218 对 N2 的 sam 格式文件使用 cufflinks 命令得到的输出文件

4、对 N3 使用 cufflinks 命令，cufflinks 拼接转录本，在拼接处尽可能达到最优结构，并且能够计算表达量和差异表达。

未经授权，不得用于任何商业用途！
联系方式：jinfo@zju.edu.cn

（1）使用 bam 格式文件（如图 219 和图 220 所示）。

代码：

cufflinks -u -g /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.gtf -o /home/txy/cufflinks/shilishuju/N3/bam /home/txy/cufflinks/shilishuju/N3.bam

#"-u"告诉 Cufflinks 用更准确的方法去处理贴到多个位点上的 reads，如果没有"-u"，Cufflinks 只会把这些 reads 简单地平均分配。比如一个 read 贴到了 10 个位置，那么每个位置分得十分之一。加-u 后会先进行平均分配，然后按照这 10 个位置各自的表达量，计算 read 被分配到每个位置的概率。

#"-g"表示提供一个注释文件，但是 Cufflinks 会在这些已知转录本的指导下，拼接新的转录本，可以不用。

#"-o"表示输出。

#"/home/txy/cufflinks/shilishuju/N3/bam"表示输出的路径。

#"/home/txy/cufflinks/shilishuju/N3.bam"表示输出的 bam 文件名称及其路径。

```
[txy@localhost ~]$ cufflinks -u -g /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic
.gtf -o /home/txy/cufflinks/shilishuju/N3/bam /home/txy/cufflinks/shilishuju/N3.bam
Warning: Could not connect to update server to verify current version. Please check at
the Cufflinks website (http://cufflinks.cbcb.umd.edu).
[11:32:37]  Loading reference annotation.
[11:33:46]  Inspecting reads and determining fragment length distribution.
> Processed 35413 loci.                        [*************************] 100%
Warning: Using default Gaussian distribution due to insufficient paired-end reads in op
en ranges.  It is recommended that correct parameters (--frag-len-mean and --frag-len-s
td-dev) be provided.
> Map Properties:
>       Normalized Map Mass: 560.00
>       Raw Map Mass: 560.00
>       Number of Multi-Reads: 0 (with 0 total hits)
>       Fragment Length Distribution: Truncated Gaussian (default)
>                     Default Mean: 200
>                  Default Std Dev: 80
[11:33:55]  Assembling transcripts and initializing abundances for multi-read correction
.
> Processed 35413 loci.                        [*************************] 100%
[11:40:19]  Loading reference annotation.
[11:40:46]  Re-estimating abundances with multi-read correction.
> Processed 35412 loci.                        [*************************] 100%
```

图 219 对 N3 的 bam 格式文件使用 cufflinks 命令

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

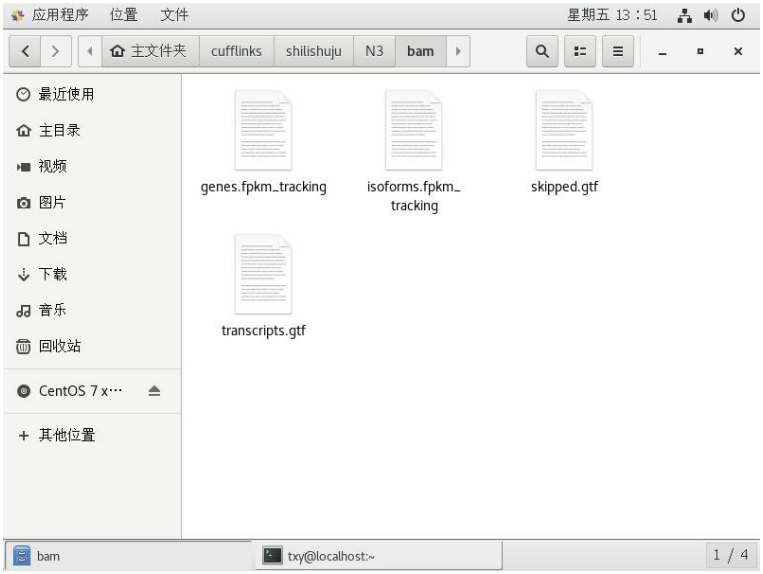


图 220 对 N3 的 bam 格式文件使用 cufflinks 命令得到的输出文件

（2）使用 sam 格式文件（如图 221-图 223 所示）。

代码：

cufflinks -u -g /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.gtf -o /home/txy/cufflinks/shilishuju/N3/sam /home/txy/cufflinks/shilishuju/N3.sam

#“-u”告诉 Cufflinks 用更准确的方法去处理贴到多个位点上的 reads，如果没有“-u”，Cufflinks 只会把这些 reads 简单地平均分配。比如一个 read 贴到了 10 个位置，那么每个位置分得十分之一。加-u 后会先进行平均分配，然后按照这 10 个位置各自的表达量，计算 read 被分配到每个位置的概率。

#“-g”表示提供一个注释文件，但是 Cufflinks 会在这些已知转录本的指导下，拼接新的转录本，可以不用。

#“-o”表示输出。

#“/home/txy/cufflinks/shilishuju/N3/sam”表示输出的路径。

#“/home/txy/cufflinks/shilishuju/N3.sam”表示输入的 sam 格式文件名称及其路径。

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
[txy@localhost ~]$ cufflinks -u -g /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic
.gtf -o /home/txy/cufflinks/shilishuju/N3/sam /home/txy/cufflinks/shilishuju/N3.sam
Warning: Could not connect to update server to verify current version. Please check at
the Cufflinks website (http://cufflinks.cbcb.umd.edu).
[bam_header_read] EOF marker is absent. The input is probably truncated.
[bam_header_read] invalid BAM binary header (this is not a BAM file).
File /home/txy/cufflinks/shilishuju/N3.sam doesn't appear to be a valid BAM file, tryin
g SAM...
[13:11:38] Loading reference annotation.
[13:12:50] Inspecting reads and determining fragment length distribution.
> Processed 35413 loci.                         [*************************] 100%
Warning: Using default Gaussian distribution due to insufficient paired-end reads in op
en ranges.  It is recommended that correct parameters (--frag-len-mean and --frag-len-s
td-dev) be provided.
> Map Properties:
>       Normalized Map Mass: 560.00
>       Raw Map Mass: 560.00
>       Number of Multi-Reads: 0 (with 0 total hits)
>       Fragment Length Distribution: Truncated Gaussian (default)
>                     Default Mean: 200
>                  Default Std Dev: 80
[13:12:53] Assembling transcripts and initializing abundances for multi-read correction
.
> Processed 35413 loci.                         [*************************] 100%
[13:18:38] Loading reference annotation.
[13:19:09] Re-estimating abundances with multi-read correction.
```

图 221 对 N3 的 sam 格式文件使用 cufflinks 命令

```
> Processed 35412 loci.                         [*************************] 100%
```

图 222 对 N3 的 sam 格式文件使用 cufflinks 命令

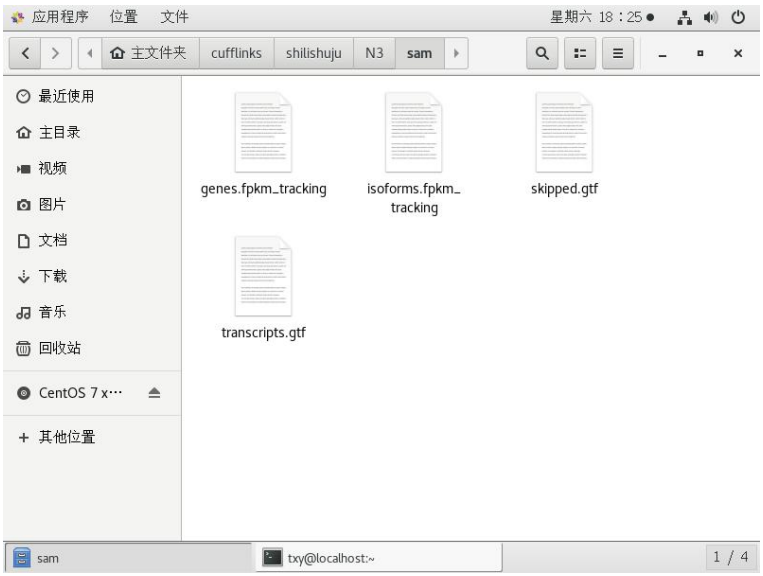


图 223 对 N3 的 sam 格式文件使用 cufflinks 命令得到的输出文件

5、对 N4 使用 cufflinks 命令，cufflinks 拼接转录本，在拼接处尽可能达到最优结构，并且能够计算表达量和差异表达。

（1）使用 bam 格式文件（如图 224 和图 225 所示）。

代码：

```
cufflinks -u -g /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.gtf -o
/home/txy/cufflinks/shilishuju/N4/bam /home/txy/cufflinks/shilishuju/N4.bam
```

获取授权：不得用于非商业用途，联系方式：bihe@zju.edu.cn

#“-u”告诉 Cufflinks 用更准确的方法去处理贴到多个位点上的 reads，如果没有“-u”，Cufflinks 只会把这些 reads 简单地平均分配。比如一个 read 贴到了 10 个位置，那么每个位置分得十分之一。加-u 后会先进行平均分配，然后按照这 10 个位置各自的表达量，计算 read 被分配到每个位置的概率。

#“-g”表示提供一个注释文件，但是 Cufflinks 会在这些已知转录本的指导下，拼接新的转录本，可以不用。

#“-o”表示输出。

#“/home/txy/cufflinks/shilishuju/N4/bam”表示输出的路径。

#“/home/txy/cufflinks/shilishuju/N4.bam”表示输出的 bam 文件名称及其路径。

```
[txy@localhost ~]$ cufflinks -u -g /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic
.gtf -o /home/txy/cufflinks/shilishuju/N4/bam /home/txy/cufflinks/shilishuju/N4.bam
Warning: Could not connect to update server to verify current version. Please check at
the Cufflinks website (http://cufflinks.cbcb.umd.edu).
[13:27:44]  Loading reference annotation.
[13:28:10]  Inspecting reads and determining fragment length distribution.
> Processed 35412 loci.                         [*************************] 100%
Warning: Using default Gaussian distribution due to insufficient paired-end reads in op
en ranges.  It is recommended that correct parameters (--frag-len-mean and --frag-len-s
td-dev) be provided.
> Map Properties:
>       Normalized Map Mass: 671.00
>       Raw Map Mass: 671.00
>       Number of Multi-Reads: 0 (with 0 total hits)
>       Fragment Length Distribution: Truncated Gaussian (default)
>                    Default Mean: 200
>                  Default Std Dev: 80
[13:28:12]  Assembling transcripts and initializing abundances for multi-read correction
.
> Processed 35412 loci.                         [*************************] 100%
[13:33:34]  Loading reference annotation.
[13:34:06]  Re-estimating abundances with multi-read correction.
> Processed 35412 loci.                         [*************************] 100%
```

图 224 对 N4 的 bam 格式文件使用 cufflinks 命令

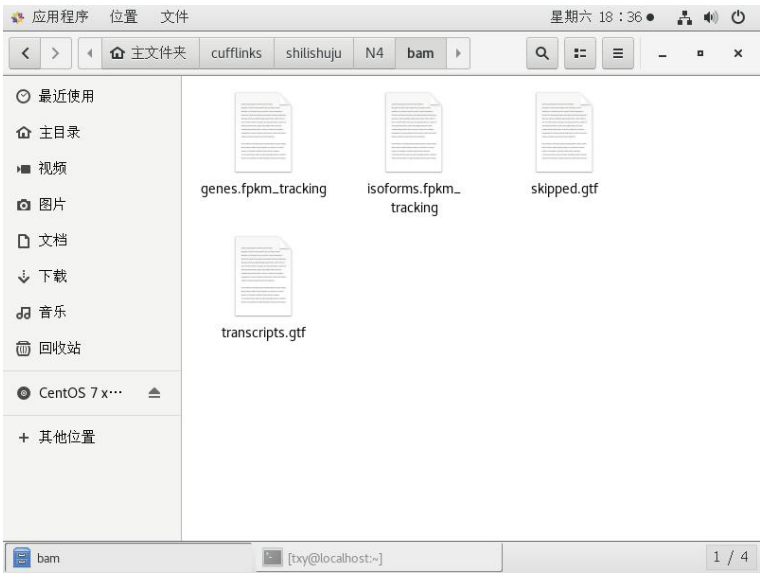


图 225 对 N4 的 bam 格式文件使用 cufflinks 命令得到的输出文件

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

（2）使用 sam 格式文件（如图 226-图 228 所示）。

代码：

cufflinks -u -g /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.gtf -o /home/txy/cufflinks/shilishuju/N4/sam /home/txy/cufflinks/shilishuju/N4.sam

#"-u"告诉 Cufflinks 用更准确的方法去处理贴到多个位点上的 reads，如果没有"-u"，Cufflinks 只会把这些 reads 简单地平均分配。比如一个 read 贴到了 10 个位置，那么每个位置分得十分之一。加-u 后会先进行平均分配，然后按照这 10 个位置各自的表达量，计算 read 被分配到每个位置的概率。

#"-g"表示提供一个注释文件，但是 Cufflinks 会在这些已知转录本的指导下，拼接新的转录本，可以不用。

#"-o"表示输出。

#"/home/txy/cufflinks/shilishuju/N4/sam"表示输出的路径。

#"/home/txy/cufflinks/shilishuju/N4.sam"表示输入的 sam 格式文件名称及其路径。

```
[txy@localhost ~]$ cufflinks -u -g /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic
.gtf -o /home/txy/cufflinks/shilishuju/N4/sam /home/txy/cufflinks/shilishuju/N4.sam
Warning: Could not connect to update server to verify current version. Please check at
the Cufflinks website (http://cufflinks.cbcb.umd.edu).
[bam_header_read] EOF marker is absent. The input is probably truncated.
[bam_header_read] invalid BAM binary header (this is not a BAM file).
File /home/txy/cufflinks/shilishuju/N4.sam doesn't appear to be a valid BAM file, tryin
g SAM...
[18:26:53] Loading reference annotation.
[18:30:19] Inspecting reads and determining fragment length distribution.
> Processed 35412 loci.                         [*************************] 100%
Warning: Using default Gaussian distribution due to insufficient paired-end reads in op
en ranges.  It is recommended that correct parameters (--frag-len-mean and --frag-len-s
td-dev) be provided.
> Map Properties:
>       Normalized Map Mass: 671.00
>       Raw Map Mass: 671.00
>       Number of Multi-Reads: 0 (with 0 total hits)
>       Fragment Length Distribution: Truncated Gaussian (default)
>                     Default Mean: 200
>                  Default Std Dev: 80
[18:30:56] Assembling transcripts and initializing abundances for multi-read correction
.
> Processed 35412 loci.                         [*************************] 100%
[18:38:27] Loading reference annotation.
[18:38:58] Re-estimating abundances with multi-read correction.
```

图 226 对 N4 的 sam 格式文件使用 cufflinks 命令

```
> Processed 35412 loci.                         [*************************] 100%
```

图 227 对 N4 的 sam 格式文件使用 cufflinks 命令

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

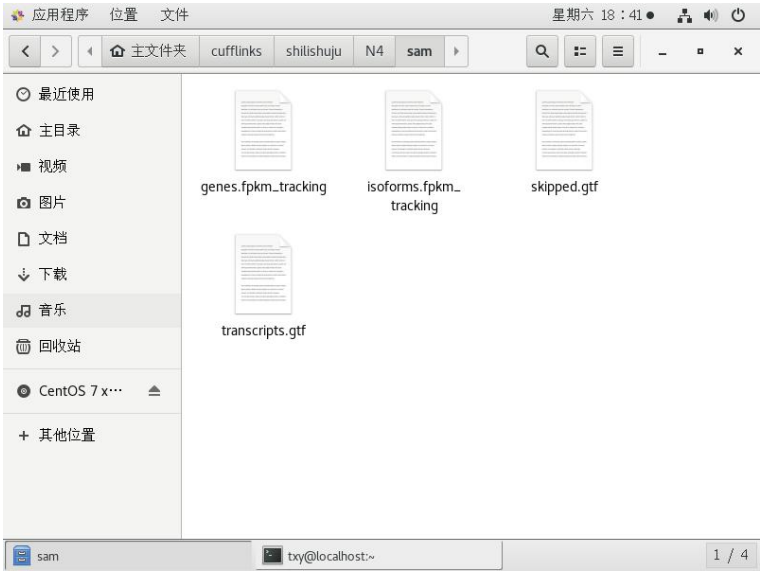


图 228 对 N4 的 sam 格式文件使用 cufflinks 命令得到的输出文件

4.3.4 cuffmerge 的功能与使用方法

1、从 NCBI 上下载基因组文件，格式为 FASTA（如图 229 和图 230 所示）。

代码：

wget

ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_latest/refseq_identifiers/GRCh38_latest_genomic.fna.gz

#下载基因组文件

```
[txy@localhost ~]$ wget ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_latest/refseq_identifiers/GRCh38_latest_genomic.fna.gz
--2019-04-11 09:56:44--  ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_latest/refseq_identifiers/GRCh38_latest_genomic.fna.gz
           => “GRCh38_latest_genomic.fna.gz”
正在解析主机 ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)... 130.14.250.7, 2607:f220:41e:250::11
正在连接 ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)|130.14.250.7|:21... 已连接。
正在以 anonymous 登录 ... 登录成功！
==> SYST ... 完成。    ==> PWD ... 完成。
==> TYPE I ... 完成。  ==> CWD (1) /refseq/H_sapiens/annotation/GRCh38_latest/refseq_identifiers ... 完成。
==> SIZE GRCh38_latest_genomic.fna.gz ... 960394547
==> PASV ... 完成。    ==> RETR GRCh38_latest_genomic.fna.gz ... 完成。
长度：960394547 (916M) (非正式数据)

100%[=======================================>] 960,394,547  159KB/s 用时 70m 56s

2019-04-11 11:07:48 (220 KB/s) - “GRCh38_latest_genomic.fna.gz”已保存 [960394547]
```

图 229 下载基因组文件

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

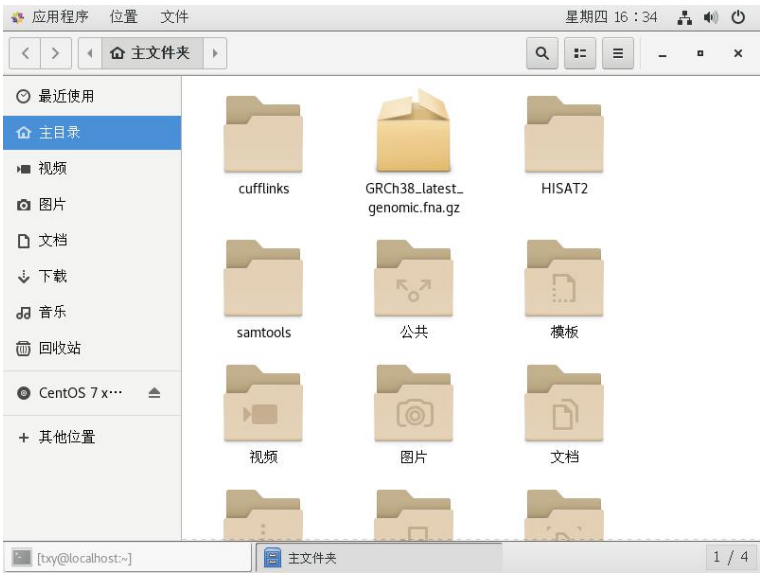


图 230 下载得到的基因组文件

2、解压基因组文件（如图 231 和图 232 所示）。

代码：

```
gunzip /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.fna.gz
#解压缩基因组文件。
```

```
[txy@localhost ~]$ gunzip /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.fna.gz
[txy@localhost ~]$
```

图 231 解压基因组文件

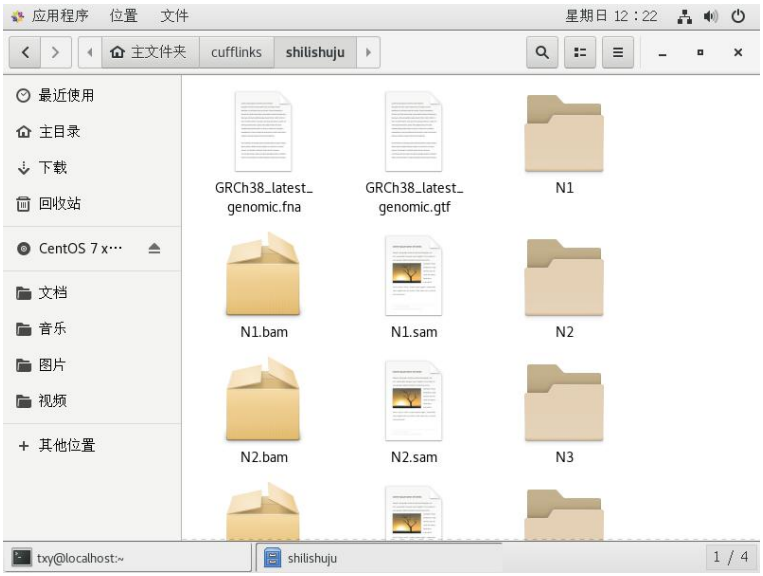


图 232 解压得到的基因组文件

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

3、为了便于管理输出的数据文件，在 shilishuju 文件夹下创建两个文件夹，命名为“N1N2”和“N3N4”，用以存放 cuffmerge 的输出文件（如图 233-图 236 所示）

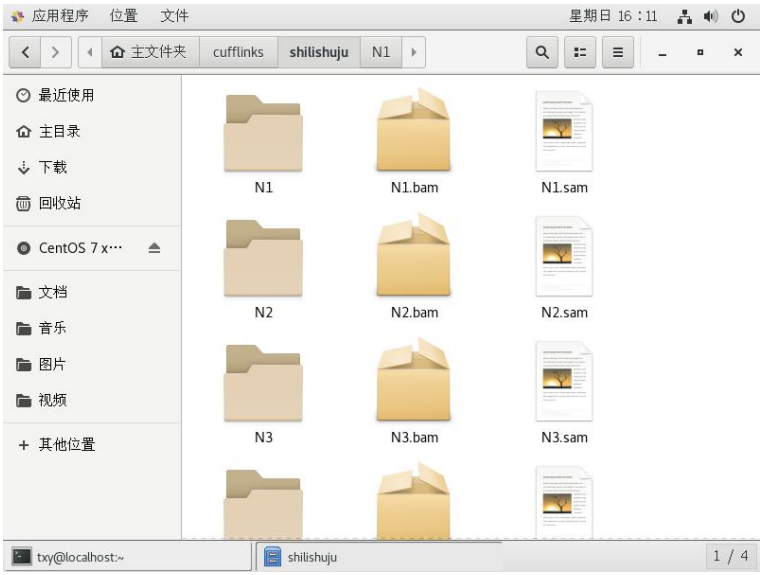


图 233 shilishuju 文件夹

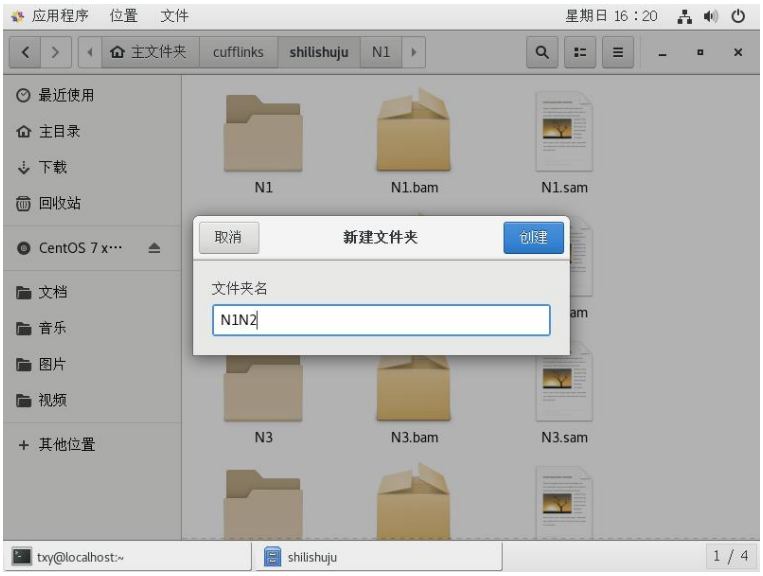


图 234 创建新的文件夹，命名为“N1N2”

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

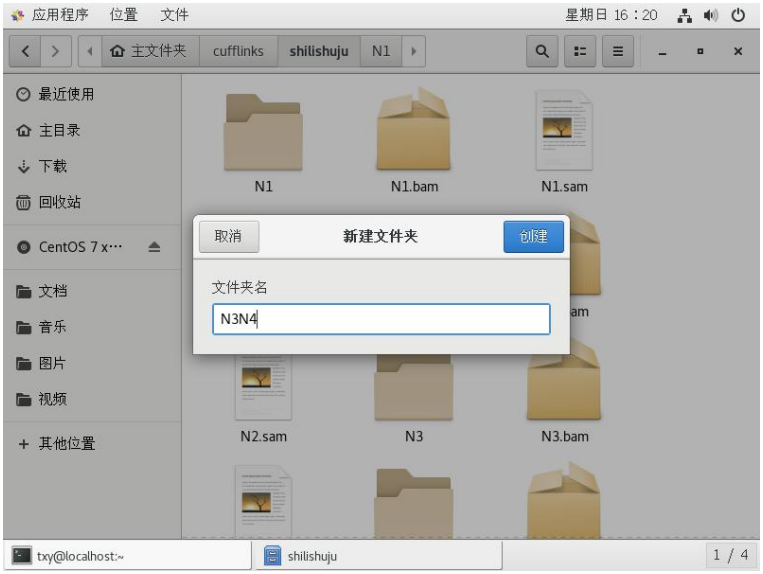


图 235 创建新的文件夹，命名为“N3N4”

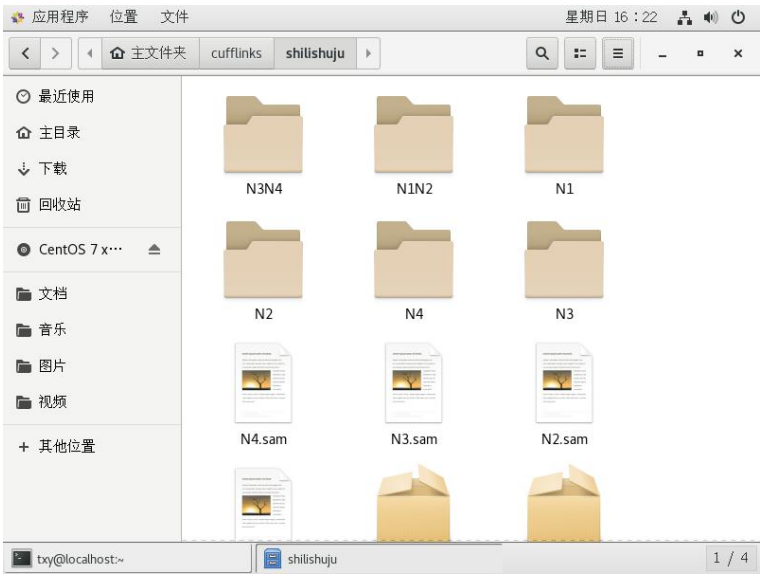


图 236 创建好的 N1N2 文件夹和 N3N4 文件夹

4、创建一个提供组装的转录本位置的文本文件。

（1）创建文本文件，命名为“N1N2”，用于提供组装的 N1 和 N2 的转录本位置（如图 237-图 239 所示）。

代码：

```
cat>N1N2.txt
/home/txy/cufflinks/shilishuju/N1/bam/transcripts.gtf
/home/txy/cufflinks/shilishuju/N2/bam/transcripts.gtf
```

联系方式：bihe@zju.edu.cn

“Ctrl”+“D”完成文本文件的创建

#“cat>”表示创建文本文件的命令。

#“N.txt”表示所创建的文本文件的名称。

#“/home/txy/cufflinks/shilishuju/N1/bam/transcripts.gtf”表示下一步需要融合的文件名称及其路径。

#“/home/txy/cufflinks/shilishuju/N2/bam/transcripts.gtf”表示下一步需要融合的文件名称及其路径。

```
[txy@localhost ~]$ cat>N1N2.txt
/home/txy/cufflinks/shilishuju/N1/bam/transcripts.gtf
/home/txy/cufflinks/shilishuju/N2/bam/transcripts.gtf
```

图 237 输入代码创建文本文件

```
[txy@localhost ~]$ cat>N1N2.txt
/home/txy/cufflinks/shilishuju/N1/bam/transcripts.gtf
/home/txy/cufflinks/shilishuju/N2/bam/transcripts.gtf
[txy@localhost ~]$
```

图 238 “Ctrl”+“D”完成文本文件的创建

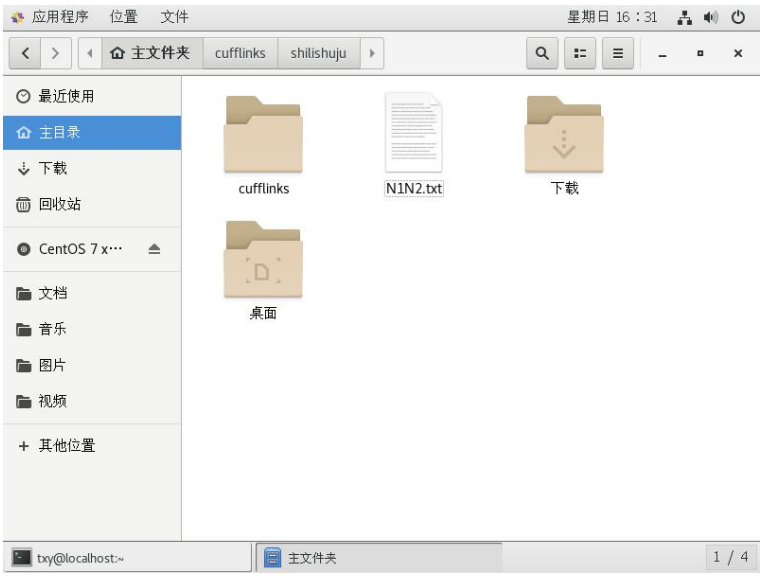


图 239 创建完成的文本文件

（2）创建文本文件，命名为“N3N4”，用于提供组装的 N3 和 N4 的转录本位置（如图 240-图 242 所示）。

代码：

cat>N3N4.txt

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

/home/txy/cufflinks/shilishuju/N3/bam/transcripts.gtf

/home/txy/cufflinks/shilishuju/N4/bam/transcripts.gtf

“Ctrl”+“D”完成文本文件的创建

#“cat>”表示创建文本文件的命令。

#“N3N4.txt”表示所创建的文本文件的名称。

#“/home/txy/cufflinks/shilishuju/N3/bam/transcripts.gtf”表示下一步需要融合的文件名称及其路径。

#“/home/txy/cufflinks/shilishuju/N4/bam/transcripts.gtf”表示下一步需要融合的文件名称及其路径。

```
[txy@localhost ~]$ cat>N3N4.txt
/home/txy/cufflinks/shilishuju/N3/bam/transcripts.gtf
/home/txy/cufflinks/shilishuju/N4/bam/transcripts.gtf
```

图 240 输入代码创建文本文件

```
[txy@localhost ~]$ cat>N3N4.txt
/home/txy/cufflinks/shilishuju/N3/bam/transcripts.gtf
/home/txy/cufflinks/shilishuju/N4/bam/transcripts.gtf
[txy@localhost ~]$
```

图 241 “Ctrl”+“D”完成文本文件的创建

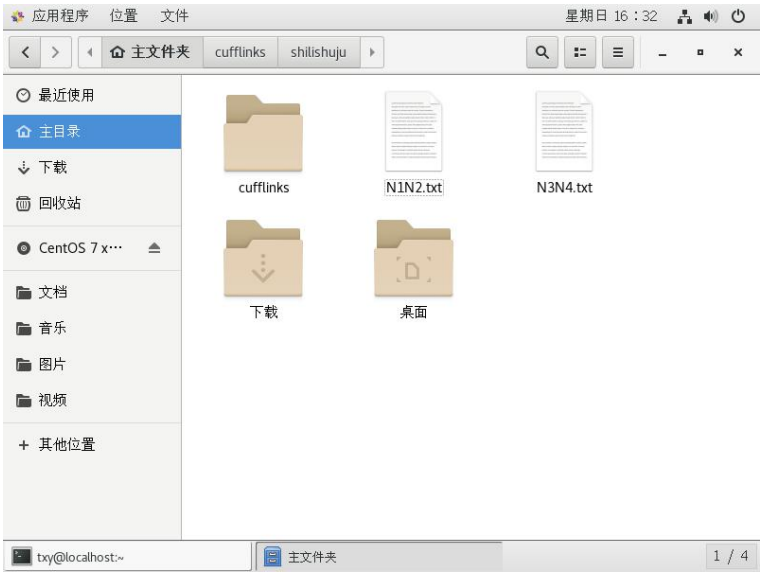


图 242 创建完成的文本文件

5、将 Cufflinks 生成的 gtf 文件融合为一个更加全面的转录本注释结果的工具。

（1）N1 和 N2 生成融合的 gtf 文件（如图 243-图 245 所示）。

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

代码：

cuffmerge -g /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.gtf -s /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.fna -o /home/txy/cufflinks/shilishuju/N1N2 /home/txy/cufflinks/shilishuju/N1N2.txt

#"-g"表示参考的注释文件。

#"/home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.gtf"表示 gtf 格式的注释文件。

#"-s"表示基因组文件。

#"/home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.fna"表示 fa 格式的基因组文件。

#"-o"表示输出。

#"/home/txy/cufflinks/shilishuju/N1N2"表示输出的文件夹及其路径。

#"N1N2.txt"表示提供组装的转录本位置。

```
[txy@localhost ~]$ cuffmerge -g /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.gt
f -s /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.fna -o /home/txy/cufflinks/sh
ilishuju/N1N2 /home/txy/cufflinks/shilishuju/N1N2.txt

[Sun Apr 14 18:24:01 2019] Beginning transcriptome assembly merge
---------------------------------------------

[Sun Apr 14 18:24:01 2019] Preparing output location /home/txy/cufflinks/shilishuju/N1N
2/
[Sun Apr 14 18:28:22 2019] Converting GTF files to SAM
[18:28:23] Loading reference annotation.
[18:28:59] Loading reference annotation.
[Sun Apr 14 18:29:22 2019] Quantitating transcripts
Warning: Could not connect to update server to verify current version. Please check at
the Cufflinks website (http://cufflinks.cbcb.umd.edu).
Command line:
cufflinks -o /home/txy/cufflinks/shilishuju/N1N2/ -F 0.05 -g /home/txy/cufflinks/shilis
huju/GRCh38_latest_genomic.gtf -q --overhang-tolerance 200 --library-type=transfrags -A
 0.0 --min-frags-per-transfrag 0 --no-5-extend -p 1 /home/txy/cufflinks/shilishuju/N1N2
/tmp/mergeSam_filehrrlsb
[bam_header_read] EOF marker is absent. The input is probably truncated.
[bam_header_read] invalid BAM binary header (this is not a BAM file).
File /home/txy/cufflinks/shilishuju/N1N2/tmp/mergeSam_filehrrlsb doesn't appear to be a
 valid BAM file, trying SAM..
[18:29:25] Loading reference annotation.
[18:29:41] Inspecting reads and determining fragment length distribution.
```

图 243 N1 和 N2 生成融合的 gtf 文件

```
Processed 34618 loci.
> Map Properties:
>       Normalized Map Mass: 351955.00
>       Raw Map Mass: 351955.00
>       Fragment Length Distribution: Truncated Gaussian (default)
>                     Default Mean: 200
>                  Default Std Dev: 80
[18:30:01] Assembling transcripts and estimating abundances.
Processed 34618 loci.
[Sun Apr 14 19:35:56 2019] Comparing against reference file /home/txy/cufflinks/shilish
uju/GRCh38_latest_genomic.gtf
Warning: Could not connect to update server to verify current version. Please check at
the Cufflinks website (http://cufflinks.cbcb.umd.edu).
No fasta index found for /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.fna. Rebu
ilding, please wait..
Fasta index rebuilt.
Warning: couldn't find fasta record for '16'!
[Sun Apr 14 19:57:24 2019] Comparing against reference file /home/txy/cufflinks/shilish
uju/GRCh38_latest_genomic.gtf
Warning: Could not connect to update server to verify current version. Please check at
the Cufflinks website (http://cufflinks.cbcb.umd.edu).
Warning: couldn't find fasta record for '16'!
```

二维码为其 bttp://w2.pu.edu.cn

图 244 N1 和 N2 生成融合的 gtf 文件

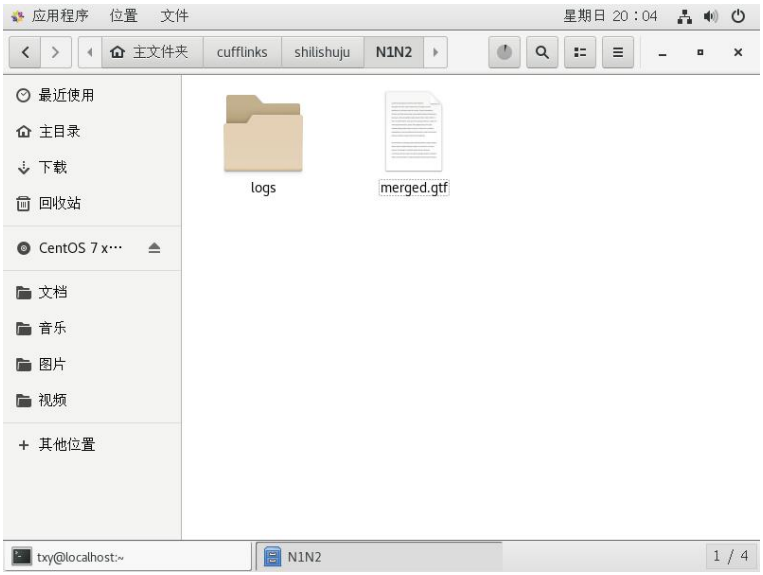


图 245 N1 和 N2 生成融合的 gtf 文件

（2）N3 和 N4 生成融合的 gtf 文件（如图 246-图 248 所示）。

代码：

```
cuffmerge -g /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.gtf -s /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.fna -o /home/txy/cufflinks/shilishuju/N3N4 /home/txy/cufflinks/shilishuju/N3N4.txt
#“-g”表示参考的注释文件。
#“/home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.gtf”表示 gtf 格式的注释文件。
#“-s”表示基因组文件。
#“/home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.fna”表示 fa 格式的基因组文件。
#“-o”表示输出。
#“/home/txy/cufflinks/shilishuju/N3N4”表示输出的文件夹及其路径。
#“N3N4.txt”表示提供组装的转录本位置。
```

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
[txy@localhost ~]$ cuffmerge -g /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.gt
f -s /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.fna -o /home/txy/cufflinks/sh
ilishuju/N3N4 /home/txy/cufflinks/shilishuju/N3N4.txt

[Mon Apr 15 14:54:04 2019] Beginning transcriptome assembly merge
-----------------------------------------

[Mon Apr 15 14:54:04 2019] Preparing output location /home/txy/cufflinks/shilishuju/N3N
4/
[Mon Apr 15 14:54:51 2019] Converting GTF files to SAM
[14:54:51] Loading reference annotation.
[14:55:16] Loading reference annotation.
[Mon Apr 15 14:55:38 2019] Quantitating transcripts
Warning: Could not connect to update server to verify current version. Please check at
the Cufflinks website (http://cufflinks.cbcb.umd.edu).
Command line:
cufflinks -o /home/txy/cufflinks/shilishuju/N3N4/ -F 0.05 -g /home/txy/cufflinks/shilis
huju/GRCh38_latest_genomic.gtf -q --overhang-tolerance 200 --library-type=transfrags -A
 0.0 --min-frags-per-transfrag 0 --no-5-extend -p 1 /home/txy/cufflinks/shilishuju/N3N4
/tmp/mergeSam_filebWCfzv
[bam_header_read] EOF marker is absent. The input is probably truncated.
[bam_header_read] invalid BAM binary header (this is not a BAM file).
File /home/txy/cufflinks/shilishuju/N3N4/tmp/mergeSam_filebWCfzv doesn't appear to be a
 valid BAM file, trying SAM...
[14:55:46] Loading reference annotation.
[14:56:00] Inspecting reads and determining fragment length distribution.
```

图 246 N3 和 N4 生成融合的 gtf 文件

```
Processed 34618 loci.
> Map Properties:
>       Normalized Map Mass: 351957.00
>       Raw Map Mass: 351957.00
>       Fragment Length Distribution: Truncated Gaussian (default)
>                     Default Mean: 200
>                  Default Std Dev: 80
[14:56:17] Assembling transcripts and estimating abundances.
Processed 34618 loci.
[Mon Apr 15 15:58:10 2019] Comparing against reference file /home/txy/cufflinks/shilish
uju/GRCh38_latest_genomic.gtf
Warning: Could not connect to update server to verify current version. Please check at
the Cufflinks website (http://cufflinks.cbcb.umd.edu).
Warning: couldn't find fasta record for '16'!
[Mon Apr 15 16:04:34 2019] Comparing against reference file /home/txy/cufflinks/shilish
uju/GRCh38_latest_genomic.gtf
Warning: Could not connect to update server to verify current version. Please check at
the Cufflinks website (http://cufflinks.cbcb.umd.edu).
Warning: couldn't find fasta record for '16'!
```

图 247 N3 和 N4 生成融合的 gtf 文件

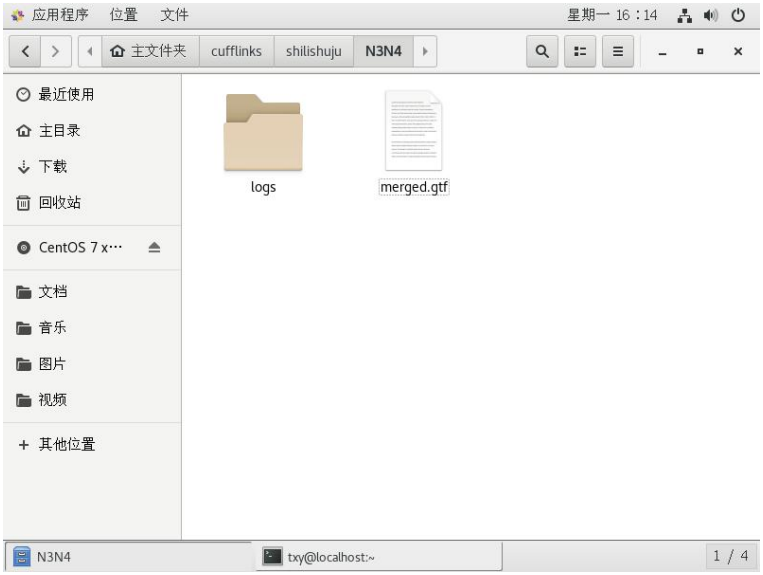


图 248 N3 和 N4 生成融合的 gtf 文件

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

4.3.5 cuffdiff 的功能与使用方法

1、为了便于管理输出的数据文件，首先在 shilishuju 文件夹下创建一个文件夹，命名为“output”（如图 249-图 251 所示）。

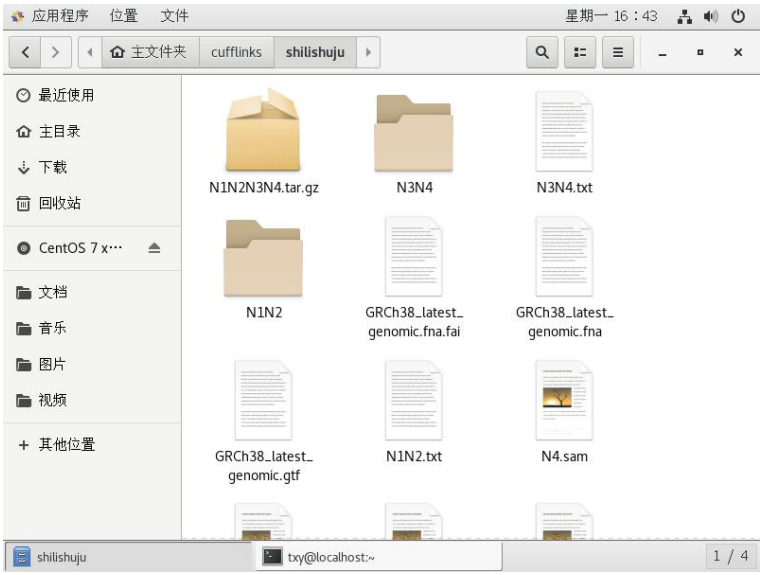


图 249 shilishuju 文件夹

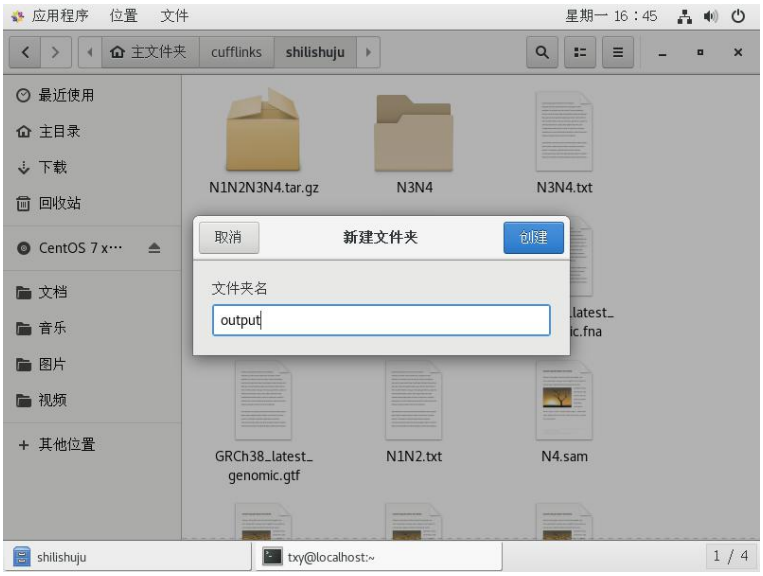


图 250 创建文件夹，命名为“output”

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

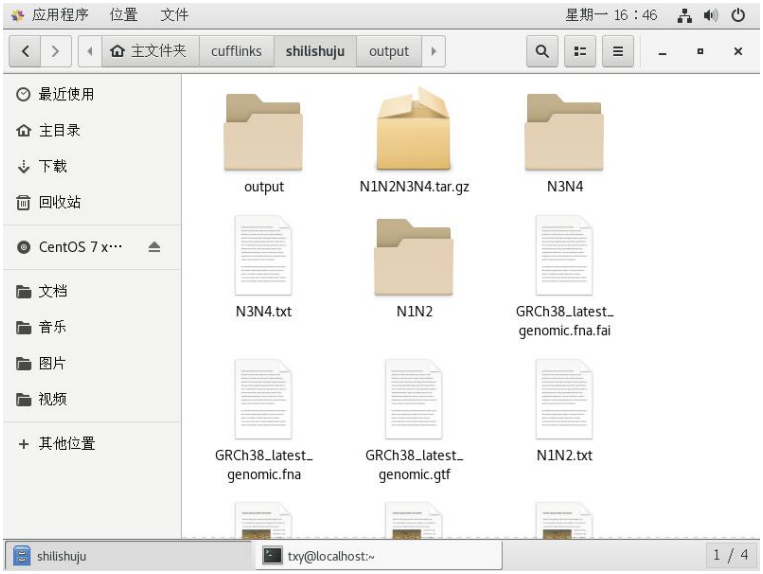


图 251 output 文件夹

2、cuffdiff 用于计算差异表达的基因，计算的简单原理在于测序深度和外显子长度一定时，Read 的数量与对应的转录本数量成正比。通过对 Reads 进行计数计算转录本的表达量。同时 cuffdiff 可以计算不同条件下转录本表达水平的显著性差异（如图 252-图 255 所示）。

代码：

cuffdiff -o /home/txy/cufflinks/shilishuju/output -b /home/txy/cufflinks/shilishuju/GRCh38_latest_genomic.fna -u /home/txy/cufflinks/shilishuju/N1N2/merged.gtf /home/txy/cufflinks/shilishuju/N3N4/merged.gtf /home/txy/cufflinks/shilishuju/N1.bam , /home/txy/cufflinks/shilishuju/N2.bam, /home/txy/cufflinks/shilishuju/N3.bam , /home/txy/cufflinks/shilishuju/N4.bam

#“-o”表示输出。

#“/home/txy/cufflinks/shilishuju/output ”表示结果输出的文件夹及其路径。

#“-b”表示提供一个 FASTA 格式的基因组文件来指导 cufflinks 运行新的 bias dection and correction algorithm，这样能够明显提高转录子丰度计算的精确性。

#“-u”Cufflinks 只会把这些 reads 简单地平均分配。比如一个 read 贴到了 10 个位置，那么每个位置分得十分之一。加-u 后会先进行平均分配，然后按照这 10 个位置各自的表达量，计算 read 被分配到每个位置的概率。

联系方式：binfo@zju.edu.cn

#"/home/txy/cufflinks/shilishuju/N1N2/merged.gtf" 和 "/home/txy/cufflinks/shilishuju/N3N4/merged.gtf" 表示 cuffmerge 产生的 gtf 格式文件。

#"/home/txy/cufflinks/shilishuju/N1.bam , /home/txy/cufflinks/shilishuju/N2.bam, /home/txy/cufflinks/shilishuju/N3.bam , /home/txy/cufflinks/shilishuju/N4.bam" 表示四个原始的的 bam 文件及其路径，以逗号隔开。

```
[txy@localhost ~]$ cuffdiff -o /home/txy/cufflinks/shilishuju/output -b /home/txy/cuffl
inks/shilishuju/GRCh38_latest_genomic.fna -u /home/txy/cufflinks/shilishuju/N1N2/merged
.gtf /home/txy/cufflinks/shilishuju/N3N4/merged.gtf /home/txy/cufflinks/shilishuju/N1.b
am , /home/txy/cufflinks/shilishuju/N2.bam, /home/txy/cufflinks/shilishuju/N3.bam , /ho
me/txy/cufflinks/shilishuju/N4.bam
Warning: Could not connect to update server to verify current version. Please check at
the Cufflinks website (http://cufflinks.cbcb.umd.edu).
[bam_header_read] EOF marker is absent. The input is probably truncated.
[bam_header_read] invalid BAM binary header (this is not a BAM file).
[18:41:02] Loading reference annotation and sequence.
Warning: couldn't find fasta record for '16'!
This contig will not be bias corrected.
Warning: No conditions are replicated, switching to 'blind' dispersion method
[18:46:08] Inspecting maps and determining fragment length distributions.
Warning: Using default Gaussian distribution due to insufficient paired-end reads in op
en ranges.  It is recommended that correct parameters (--frag-len-mean and --frag-len-s
td-dev) be provided.
Warning: Using default Gaussian distribution due to insufficient paired-end reads in op
en ranges.  It is recommended that correct parameters (--frag-len-mean and --frag-len-s
td-dev) be provided.
Warning: Using default Gaussian distribution due to insufficient paired-end reads in op
en ranges.  It is recommended that correct parameters (--frag-len-mean and --frag-len-s
td-dev) be provided.
Warning: Using default Gaussian distribution due to insufficient paired-end reads in op
en ranges.  It is recommended that correct parameters (--frag-len-mean and --frag-len-s
td-dev) be provided.
```

图 252 cuffdiff 输出结果

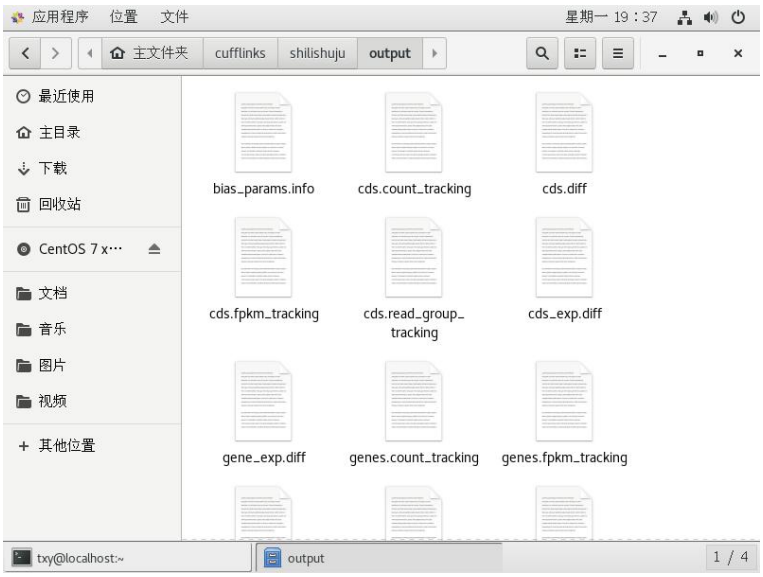


图 253 cuffdiff 输出结果

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

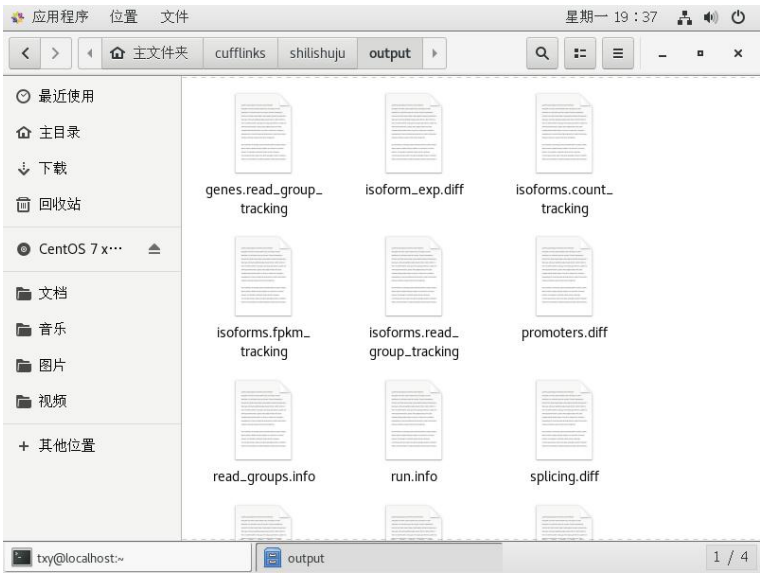


图 254 cuffdiff 输出结果

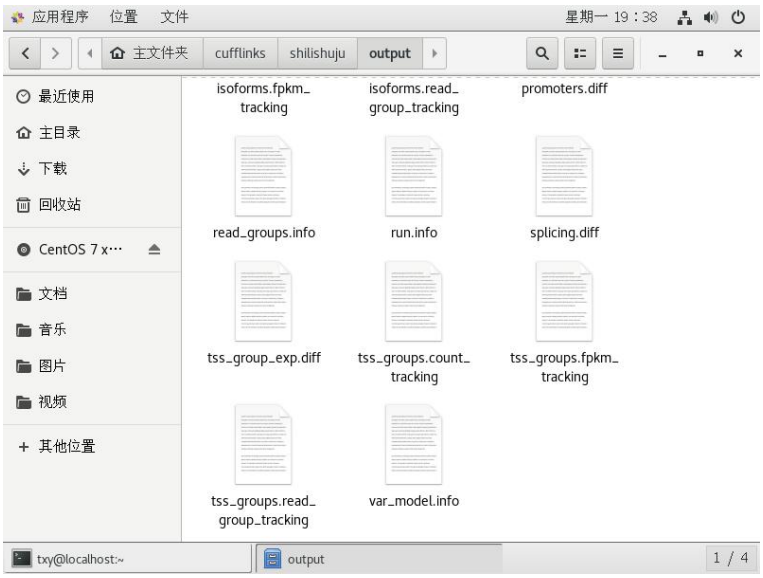


图 255 cuffdiff 输出结果

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

5 sam 格式文件/bam 格式文件的分析——SGseq

5.1 关于 SGseq

SGSeq 是基于 R 语言的，用于从 RNA-seq 数据分析剪接事件的软件包。由 Leonard Goldstein 开发。以映射到参考基因组的 RNA-seq 读数的比对好的 bam 格式文件作为输入数据。为了使 SGSeq 正常工作，读取的数据必须通过程序进行配对，例如 GSNAP（Wu 和 Nacu 2010），HISAT（Kim，Langmead 和 Salzberg 2015）或 STAR（Dobin 等人 2013），它生成带有自定义标签'XS'的 SAM / BAM 文件，用于拼接读取，指示转录方向。

可以根据现有注释或从匹配的序列读数预测，将基因表示为剪接图。剪接事件可从图中鉴定，并从每个剪接变体的开始或结束处使用结构上相容的读数进行局部定量。该软件包括用于拼接事件预测，量化，可视化的功能。

5.2 SGseq 的安装

5.2.1 R 的安装

1、首先我们要安装 R（3.5 版本以上）。我们在这里提供一个清华大学的镜像，下载新版的 R：https://mirrors.tuna.tsinghua.edu.cn/CRAN/。以 Windows 系统为例，选择 Download R for Windows，再选择 install R for the first time.，进行下载和安装（如图 256 和图 257 所示）。

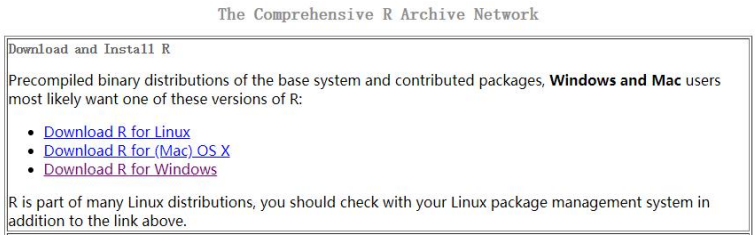


图 256 R 软件下载界面

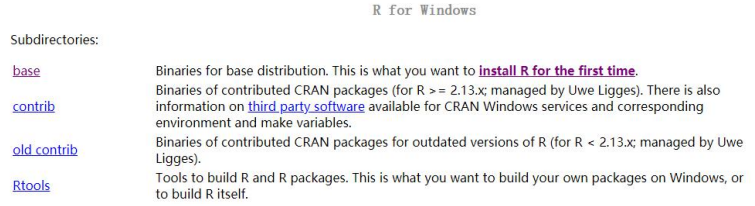


图 257 R 软件下载界面

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

2、打开 R 软件，以 3.5.2 版本的 64 位 R 软件为例（如图 258 所示）。因为后续过程可能会出现 warning：程序包是为 3.5.3 版本提供，所以我们建议使用更新版本的 R 软件。

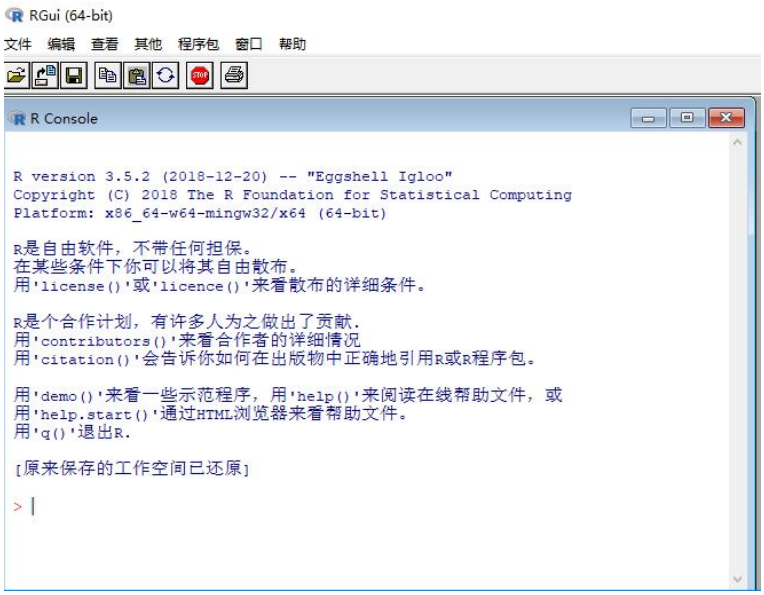


图 258 R 3.5.2 打开界面

5.2.2 SGSeq 的安装

1、输入代码，安装 SGSeq，目前 SGSeq 最新版本为 1.16.2，2018 年 12 月 19 日发布，以此为例。R 软件由 URL：https://bioconductor.org/packages/3.8/bioc/bin/windows/contrib/3.5/SGSeq_1.16.2.zip 下载 1.8MB 大小的压缩包，并自动解压，路径不可写入时，需要手动确定更新，在更新 all/some/none 的询问中输入 a，执行。R 软件会不断试开 URL 下载不同更新包，这会消耗 30 分钟左右的时间，完成所有的更新操作（如图 259 和图 260 所示）。

代码：

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("SGSeq", version = "3.8")
```

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

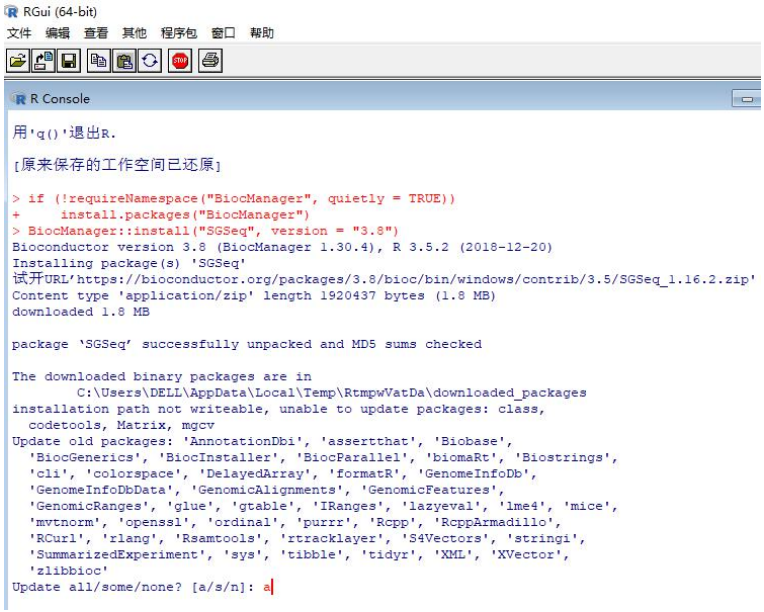


图 259 SGSeq 下载代码和运行

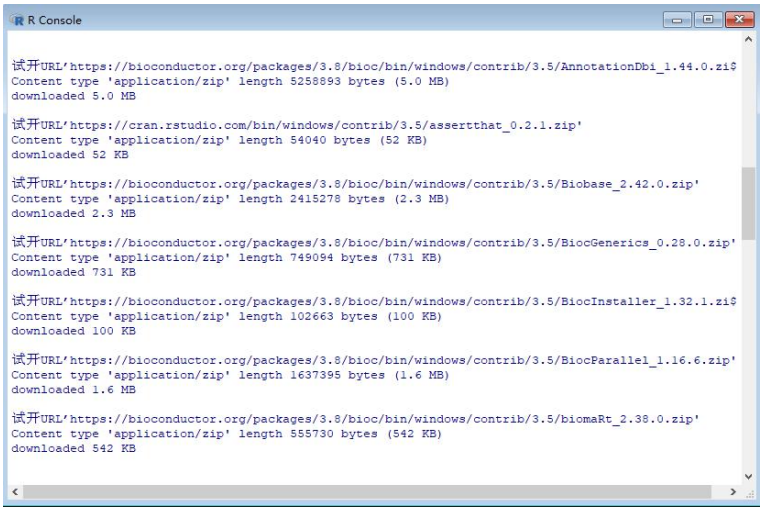


图 260 SGSeq 相关程序包下载

2、若上一步不能实现，则可以使用另一段代码作为获取并安装 SGSeq 的途径，但是在这版 R 软件会显示“biocLite is deprecated”，最终还会使用“BiocManager::install”替代，所以两种途径本质区别并不明显（如图 261 所示）。在后续的步骤中我们使用第一组代码的安装的 SGSeq 进行。

代码：

```
source("https://bioconductor.org/biocLite.R")
biocLite("SGSeq")
```

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

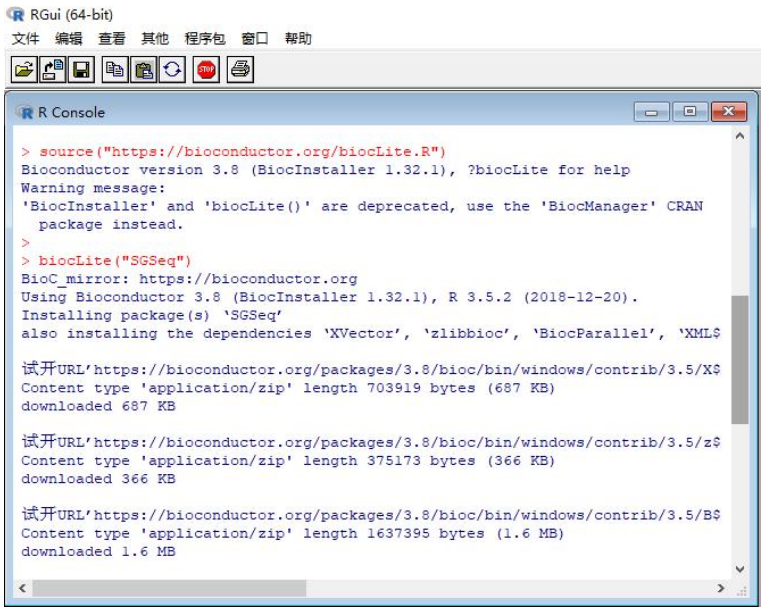


图 261 第二种下载途径

5.2.3 SGSeq 程序包的补充

1、SGSeq 需要待分析样品的有关信息以开始新项目。信息要提前获得，然后可用于所有未来的分析（如图 262 所示）。

代码：

```
library(SGSeq)
```

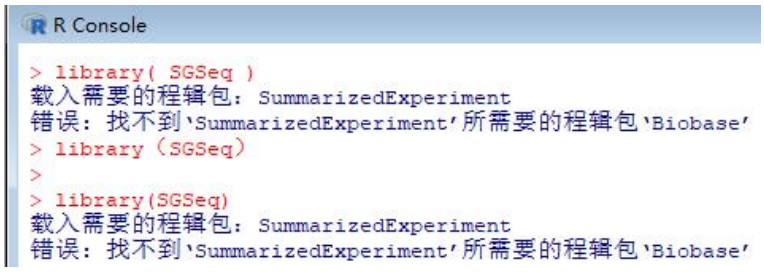


图 262 程序包下载问题

2、若程序包下载失败（如图所示），所以要从 bioconductor 对应地下载程序包。成功运行这行代码应成功加载一系列程序包（如图 263 和图 264 所示）。

代码：

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("Biobase", version = "3.8")
```

仅供授权个人用于任何商业用途！
联系方式：bihb@zju.edu.cn

```
> if (!requireNamespace("BiocManager", quietly = TRUE))
+     install.packages("BiocManager")
> BiocManager::install("Biobase", version = "3.8")
Bioconductor version 3.8 (BiocManager 1.30.4), R 3.5.2 (2018-12-20)
Installing package(s) 'Biobase'
试开URL'https://bioconductor.org/packages/3.8/bioc/bin/windows/contrib/3.5/Biobase_2.42.0.zip'
Content type 'application/zip' length 2415278 bytes (2.3 MB)
downloaded 2.3 MB

package 'Biobase' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\DELL\AppData\Local\Temp\RtmpgLIa3w\downloaded_packages
installation path not writeable, unable to update packages: class, codetools, MASS, Matrix,
```

图 263 补充下载 biobase 程序包

```
> library(SGSeq)
载入程辑包：'SGSeq'

The following objects are masked _by_ '.GlobalEnv':

    sgfc_pred, si
```

图 264 信息获得运行结果

5.3 SGseq 的功能与使用方法

1、样本信息按照下列框架给出：（1）sample_name 每个样本具有唯一名称的字符向量。（2）file_bam 字符向量或 BamFileList，指定使用拼接感知对齐程序生成的 BAM 文件。（3）paired_end 指示数据是配对端还是单端的逻辑向量。（4）read_length 具有读取长度的数字向量。（5）frag_length 具有平均片段长度的数字向量（对于配对末端数据）。（6）lib_size 数字向量，包含单端数据的对齐读取总数，或配对端数据的读取对总数。（7）这里我们使用软件包提供的示例数据（如图 265 所示）。

代码：

si

```
> si
   sample_name
N1          N1
N2          N2
N3          N3
N4          N4
T1          T1
T2          T2
T3          T3
T4          T4
                                                                                  $
N1 C:/Users/DELL/Documents/R/win-library/3.5/SGSeq/extdata/bams/C:/Users/DELL/Documents/R/win-$
N2 C:/Users/DELL/Documents/R/win-library/3.5/SGSeq/extdata/bams/C:/Users/DELL/Documents/R/win-$
N3 C:/Users/DELL/Documents/R/win-library/3.5/SGSeq/extdata/bams/C:/Users/DELL/Documents/R/win-$
N4 C:/Users/DELL/Documents/R/win-library/3.5/SGSeq/extdata/bams/C:/Users/DELL/Documents/R/win-$
T1 C:/Users/DELL/Documents/R/win-library/3.5/SGSeq/extdata/bams/C:/Users/DELL/Documents/R/win-$
T2 C:/Users/DELL/Documents/R/win-library/3.5/SGSeq/extdata/bams/C:/Users/DELL/Documents/R/win-$
T3 C:/Users/DELL/Documents/R/win-library/3.5/SGSeq/extdata/bams/C:/Users/DELL/Documents/R/win-$
T4 C:/Users/DELL/Documents/R/win-library/3.5/SGSeq/extdata/bams/C:/Users/DELL/Documents/R/win-$
   paired_end read_length frag_length lib_size
N1       TRUE          75         293 12405197
N2       TRUE          75         197 13090179
N3       TRUE          75         206 14983084
N4       TRUE          75         207 15794088
T1       TRUE          75         284 14345976
T2       TRUE          75         235 15464168
T3       TRUE          75         259 15485954
T4       TRUE          75         247 15808356
```

联系方式：binfo@zju.edu.cn

图 265 示例数据列表

2、如果使用待处理的 bam 文件，要按照下面的代码设置正确的路径。

代码：

```
path <- system.file("extdata", package = "SGSeq")
si$file_bam <- file.path(path, "bams", si$file_bam)
```

3、从 TxDb 获得转录本注释，以建立转录预测，或者使用函数 importTranscripts（）以 GFF 格式导入。或者，可以将转录物注释为通过转录物分组的外显子的 GRangesList。在示例代码中，UCSC knownGene 表作为 TxDb 的对象加载。保留染色体 16 上的转录物（FBXO31 基因所在的位置），并且改变染色体名称以匹配 BAM 文件中使用的命名惯例（如图 266 所示）。

代码：

```
library(TxDb.Hsapiens.UCSC.hg19.knownGene)
txdb <- TxDb.Hsapiens.UCSC.hg19.knownGene
txdb <- keepSeqlevels(txdb, "chr16")
seqlevelsStyle(txdb) <- "NCBI"
```

```
> library(TxDb.Hsapiens.UCSC.hg19.knownGene)
> txdb <- TxDb.Hsapiens.UCSC.hg19.knownGene
> txdb <- keepSeqlevels(txdb, "chr16")
Error in keepSeqlevels(txdb, "chr16") : invalid seqlevels: chr16
> seqlevelsStyle(txdb) <- "NCBI"
Warning message:
In .replace_seqlevels_style(x_seqlevels, value) :
  found more than one best sequence renaming map compatible with seqname style "NCBI" for this
> |
```

图 266 获取示例数据转录本注释

4、使用下列代码调用 convertToTxFeatures 函数进行转换。其中字母的含义如下：（1）J (splice junction)：剪接点。（2）I (internal exon)：内部外显子。（3）F (first/5′-terminal exon)：起点 5′端外显子。（4）L (last/5′-terminal exon)：末尾 5′端外显子。（5）U (unspliced transcript)：未剪接转录本（如图 267 所示）。

代码：

```
txf_ucsc <- convertToTxFeatures(txdb)
txf_ucsc <- txf_ucsc[txf_ucsc %over% gr]
head(txf_ucsc)
```

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
> txf_ucsc <- convertToTxFeatures(txdb)
> txf_ucsc <- txf_ucsc[txf_ucsc %over% gr]
> head(txf_ucsc)
TxFeatures object with 6 ranges and 0 metadata columns:
      seqnames            ranges strand     type                          txName
         <Rle>         <IRanges>  <Rle> <factor>                 <CharacterList>
  [1]       16 87362942-87365116      -        L uc002fjv.3,uc002fjw.3,uc010vot.2
  [2]       16 87365116-87367492      -        J uc002fjv.3,uc002fjw.3,uc010vot.2
  [3]       16 87367492-87367892      -        I uc002fjv.3,uc002fjw.3,uc010vot.2
  [4]       16 87367892-87368910      -        J uc002fjv.3,uc002fjw.3,uc010vot.2
  [5]       16 87368910-87369063      -        I uc002fjv.3,uc002fjw.3,uc010vot.2
  [6]       16 87369063-87369761      -        J uc002fjv.3,uc002fjw.3,uc010vot.2
           geneName
      <CharacterList>
  [1]          79791
  [2]          79791
  [3]          79791
  [4]          79791
  [5]          79791
  [6]          79791
  -------
  seqinfo: 1 sequence from hg19 genome
```

图 267 第一次格式转换

5、使用下列代码，列 txName 和 geneName 表示每个特征所属的转录本和基因。请注意，某个功能可以属于多个脚本，因此这些列可以为每个功能存储多个值。对于 SGSeq 中的 TxFeatures 和其他数据结构，可以使用访问器函数访问列，如以下代码所示（如图 268 所示）。

代码：

type(txf_ucsc)

head(txName(txf_ucsc))

head(geneName(txf_ucsc))

```
  [6]          79791
  -------
  seqinfo: 1 sequence from hg19 genome
> type(txf_ucsc)
 [1] L J I J I J I J I J I J I F J J F U I J F
Levels: J I F L U
> head(txName(txf_ucsc))
CharacterList of length 6
[[1]] uc002fjv.3 uc002fjw.3 uc010vot.2
[[2]] uc002fjv.3 uc002fjw.3 uc010vot.2
[[3]] uc002fjv.3 uc002fjw.3 uc010vot.2
[[4]] uc002fjv.3 uc002fjw.3 uc010vot.2
[[5]] uc002fjv.3 uc002fjw.3 uc010vot.2
[[6]] uc002fjv.3 uc002fjw.3 uc010vot.2
> head(geneName(txf_ucsc))
CharacterList of length 6
[[1]] 79791
[[2]] 79791
[[3]] 79791
[[4]] 79791
[[5]] 79791
[[6]] 79791
>
```

图 268 列 txName 和 geneName

6、存储在 TxFeatures 中的外显子代表在 RNA 分子中拼接在一起的外显子。在剪接图的情境中，外显子由独特的不重叠外显子区域表示。使用函数 convertToSGFeatures（）将 TxFeatures 转换为 SGFeatures。在此过程中，重叠的

未经授权，不得用于任何商业用途！
联系方式：bmfo@zju.edu.cn

外显子被分离成不重叠的外显子区（如图 269 所示）。其中字母的含义如下：（1）J (splice junction)：拼接点。（2）E (disjoint exon bin)：不相交的外显子区。（3）D (splice donor site)：剪接供体位点。（4）A (splice acceptor site)：剪接受体位点。

代码：

```
sgf_ucsc <- convertToSGFeatures(txf_ucsc)
head(sgf_ucsc)
```

```
> sgf_ucsc <- convertToSGFeatures(txf_ucsc)
> head(sgf_ucsc)
SGFeatures object with 6 ranges and 0 metadata columns:
      seqnames            ranges strand     type  splice5p  splice3p featureID    geneID
         <Rle>         <IRanges>  <Rle> <factor> <logical> <logical> <integer> <integer>
  [1]       16 87362942-87365116      -        E      TRUE     FALSE         1         1
  [2]       16          87365116      -        A      <NA>      <NA>         2         1
  [3]       16 87365116-87367492      -        J      <NA>      <NA>         3         1
  [4]       16          87367492      -        D      <NA>      <NA>         4         1
  [5]       16 87367492-87367892      -        E      TRUE      TRUE         5         1
  [6]       16          87367892      -        A      <NA>      <NA>         6         1
                              txName        geneName
                    <CharacterList> <CharacterList>
  [1] uc002fjv.3,uc002fjw.3,uc010vot.2           79791
  [2] uc002fjv.3,uc002fjw.3,uc010vot.2           79791
  [3] uc002fjv.3,uc002fjw.3,uc010vot.2           79791
  [4] uc002fjv.3,uc002fjw.3,uc010vot.2           79791
  [5] uc002fjv.3,uc002fjw.3,uc010vot.2           79791
  [6] uc002fjv.3,uc002fjw.3,uc010vot.2           79791
  -------
  seqinfo: 1 sequence from hg19 genome
> |
```

图 269 第二次格式转换

7、函数 analyzeFeatures（）将转录特征转换为拼接图特征，并获得每个特征和每个样本的兼容片段计数（如图 270 所示）。

代码：

```
sgfc_ucsc <- analyzeFeatures(si, features = txf_ucsc)
sgfc_ucsc
```

```
> sgfc_ucsc <- analyzeFeatures(si, features = txf_ucsc)
Process features...
Obtain counts...
Error in value[[3L]](cond) :
  failed to open BamFile: file(s) do not exist:
  'C:/Users/DELL/Documents/R/win-library/3.5/SGSeq/extdata/bams/C
> sgfc_ucsc
class: SGFeatureCounts
dim: 42 8
metadata(0):
assays(2): counts FPKM
rownames: NULL
rowData names(0):
colnames(8): N1 N2 ... T3 T4
colData names(6): sample_name file_bam ... frag_length lib_size
```

图 270 拼接图特征转换

8、analyzeFeatures（）返回 SGFeatureCounts 对象。SGFeatureCounts 以 colData 包含样本信息。拼接图特征为 rowRanges 和 assays counts 和 FPKM，这些分别存

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

储兼容的片段计数和 FPKM。可以使用具有相同名称的访问函数访问不同的数据类型（如图 271-图 273 所示）。

代码：

```
colData(sgfc_ucsc)
rowRanges(sgfc_ucsc)
head(counts(sgfc_ucsc))
head(FPKM(sgfc_ucsc))
```

```
> colData(sgfc_ucsc)
DataFrame with 8 rows and 6 columns
   sample_name                                                    file_bam paired_end
   <character>                                                  <character>  <logical>
N1          N1 C:/Users/DELL/Documents/R/win-library/3.5/SGSeq/extdata/bams/N1.bam        TRUE
N2          N2 C:/Users/DELL/Documents/R/win-library/3.5/SGSeq/extdata/bams/N2.bam        TRUE
N3          N3 C:/Users/DELL/Documents/R/win-library/3.5/SGSeq/extdata/bams/N3.bam        TRUE
N4          N4 C:/Users/DELL/Documents/R/win-library/3.5/SGSeq/extdata/bams/N4.bam        TRUE
T1          T1 C:/Users/DELL/Documents/R/win-library/3.5/SGSeq/extdata/bams/T1.bam        TRUE
T2          T2 C:/Users/DELL/Documents/R/win-library/3.5/SGSeq/extdata/bams/T2.bam        TRUE
T3          T3 C:/Users/DELL/Documents/R/win-library/3.5/SGSeq/extdata/bams/T3.bam        TRUE
T4          T4 C:/Users/DELL/Documents/R/win-library/3.5/SGSeq/extdata/bams/T4.bam        TRUE
   read_length frag_length  lib_size
     <numeric>   <numeric> <integer>
N1          75         293  12405197
N2          75         197  13090179
N3          75         206  14983084
N4          75         207  15794088
T1          75         284  14345976
T2          75         235  15464168
T3          75         259  15485954
T4          75         247  15808356
>
```

图 271 colData 包含的样本信息

```
> rowRanges(sgfc_ucsc)
SGFeatures object with 42 ranges and 0 metadata columns:
       seqnames            ranges strand     type  spliceSp  splice3p featureID    geneID
          <Rle>         <IRanges>  <Rle> <factor> <logical> <logical> <integer> <integer>
   [1]       16 87362942-87365116      -        E      TRUE     FALSE         1         1
   [2]       16          87365116      -        A      <NA>      <NA>         2         1
   [3]       16 87365116-87367492      -        J      <NA>      <NA>         3         1
   [4]       16          87367492      -        D      <NA>      <NA>         4         1
   [5]       16 87367492-87367892      -        E      TRUE      TRUE         5         1
   ...      ...               ...    ...      ...       ...       ...       ...       ...
  [38]       16 87423343-87423454      -        E      TRUE      TRUE        38         1
  [39]       16          87423454      -        A      <NA>      <NA>        39         1
  [40]       16 87423454-87425689      -        J      <NA>      <NA>        40         1
  [41]       16          87425689      -        D      <NA>      <NA>        41         1
  [42]       16 87425689-87425708      -        E     FALSE      TRUE        42         1
                          txName        geneName
                  <CharacterList> <CharacterList>
   [1] uc002fjv.3,uc002fjw.3,uc010vot.2           79791
   [2] uc002fjv.3,uc002fjw.3,uc010vot.2           79791
   [3] uc002fjv.3,uc002fjw.3,uc010vot.2           79791
   [4] uc002fjv.3,uc002fjw.3,uc010vot.2           79791
   [5] uc002fjv.3,uc002fjw.3,uc010vot.2           79791
   ...                           ...             ...
  [38]                    uc010vot.2           79791
  [39]                    uc010vot.2           79791
  [40]                    uc010vot.2           79791
  [41]                    uc010vot.2           79791
  [42]                    uc010vot.2           79791
  -------
  seqinfo: 1 sequence from hg19 genome
```

图 272 rowrange 的结构信息

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
> head(counts(sgfc_ucsc))
      N1  N2  N3  N4  T1  T2  T3  T4
[1,] 248 424 385 477 398 459 325 394
[2,]   0   0   2   1   1   2   1   1
[3,]   4  10   8  13   7  19   6  13
[4,]   0   1   1   0   0   0   0   0
[5,]  28  41  40  48  32  81  33  63
[6,]   0   0   0   0   0   0   0   0
>
> head(FPKM(sgfc_ucsc))
         N1         N2         N3         N4         T1         T2         T3         T4
[1,] 8.103616 13.6611958 10.796489 12.6842394 11.2868069 12.3210956 8.6258775 10.2947244
[2,] 0.000000  0.0000000  0.901918  0.4278029  0.4709862  0.8738597 0.4363152  0.4274168
[3,] 2.178686  5.1616993  3.607672  5.5614378  3.2969034  8.3016673 2.6178911  5.5564183
[4,] 0.000000  0.5161699  0.450959  0.0000000  0.0000000  0.0000000 0.0000000  0.0000000
[5,] 3.257025  5.2464307  4.405408  5.0067741  3.2610974  8.2486853 3.2336318  6.1595582
[6,] 0.000000  0.0000000  0.000000  0.0000000  0.0000000  0.0000000 0.0000000  0.0000000
>
```

图 273 计数和 FPKM 的数值

9、外显子和剪接点的计数基于结构上相容的片段。在剪接供体和受体的情况下，计数表示具有跨越剪接边界的读数的片段的数量（与剪接位点和侧翼内含位置重叠）。FPKM 值计算为 x/NL*10^6，其中 x 是兼容片段的数量，N 是文库大小（存储在 lib_size 中），L 是有效特征长度（兼容片段的可能位置数）。对于两端数据，假设片段长度等于 frag_length。可以使用函数 plotFeatures 显示拼接图特征的 FPKM。plotFeatures 生成一个双区域图形（如图 274 所示），顶部区域显示拼接图，底部区域中各个特征的表达级别热图。注意，拼接图包括三个可变的转录起始位点（TSS）。但是，热图表明该数据未使用第一个 TSS。

代码：

```
df <- plotFeatures(sgfc_ucsc, geneID = 1)
```

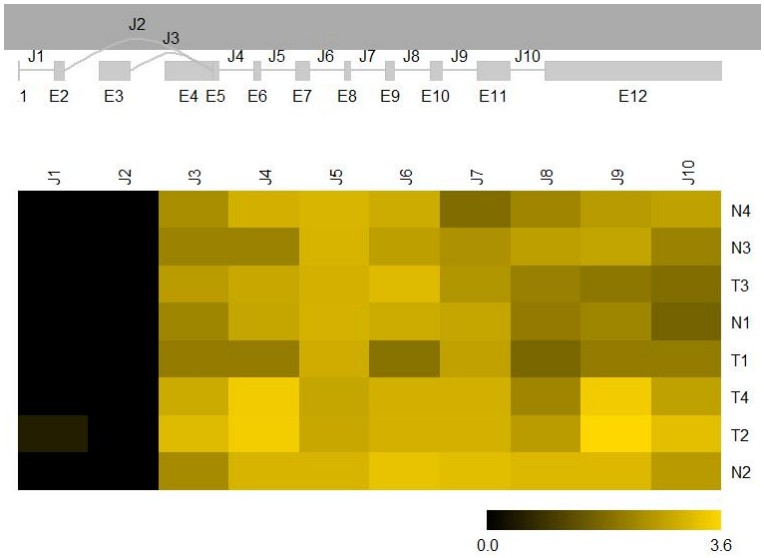


图 274 转录本剪接图和表达热图

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

10、SGSeq 可以通过来自 RNA-seq 数据的预测来增强注释，或者不依赖于现有

注释，可以在不使用注释的情况下从 RNA-seq 数据构建剪接图。以下代码预测 RNA-seq 读取支持的转录特征，将它们转换为剪接图特征，然后获得兼容的片段计数（如图 275 所示）。

代码：

```
sgfc_pred <- analyzeFeatures(si, which = gr)
head(rowRanges(sgfc_pred))
```

```
> df <- plotFeatures(sgfc_ucsc, geneID =1)
> sgfc_pred <- analyzeFeatures(si, which = gr)
Predict features...
Error in value[[3L]](cond) :
  failed to open BamFile: file(s) do not exist:
  'C:/Users/DELL/Documents/R/win-library/3.5/SGSeq/extdata/bams/C:/Users/DELL/Documents/R/win-$
> head(rowRanges(sgfc_pred))
SGFeatures object with 6 ranges and 0 metadata columns:
      seqnames            ranges strand     type  splice5p  splice3p featureID    geneID
         <Rle>         <IRanges>  <Rle> <factor> <logical> <logical> <integer> <integer>
  [1]       16 87362930-87365116      -        E      TRUE     FALSE         1         1
  [2]       16          87365116      -        A      <NA>      <NA>         2         1
  [3]       16 87365116-87367492      -        J      <NA>      <NA>         3         1
  [4]       16          87367492      -        D      <NA>      <NA>         4         1
  [5]       16 87367492-87367892      -        E      TRUE      TRUE         5         1
  [6]       16          87367892      -        A      <NA>      <NA>         6         1
                        txName        geneName
                <CharacterList> <CharacterList>
  [1] uc002fjv.3,uc002fjw.3,uc010vot.2           79791
  [2] uc002fjv.3,uc002fjw.3,uc010vot.2           79791
  [3] uc002fjv.3,uc002fjw.3,uc010vot.2           79791
  [4] uc002fjv.3,uc002fjw.3,uc010vot.2           79791
  [5] uc002fjv.3,uc002fjw.3,uc010vot.2           79791
  [6] uc002fjv.3,uc002fjw.3,uc010vot.2           79791
  -------
  seqinfo: 84 sequences from an unspecified genome
```

图 275 以 RNA-seq 预测剪接

11、对于解释，可以关于已知的转录物注释预测的特征。annotate（）函数为每个特征分配兼容的转录本，并将相应的转录本和基因名称分别存储在列 txName 和 geneName 中（如图 276 所示）。

代码：

```
sgfc_pred <- annotate(sgfc_pred, txf_ucsc)
head(rowRanges(sgfc_pred))
```

```
> sgfc_pred <- annotate(sgfc_pred, txf_ucsc)
> head(rowRanges(sgfc_pred))
SGFeatures object with 6 ranges and 0 metadata columns:
      seqnames            ranges strand     type  splice5p  splice3p featureID    geneID
         <Rle>         <IRanges>  <Rle> <factor> <logical> <logical> <integer> <integer>
  [1]       16 87362930-87365116      -        E      TRUE     FALSE         1         1
  [2]       16          87365116      -        A      <NA>      <NA>         2         1
  [3]       16 87365116-87367492      -        J      <NA>      <NA>         3         1
  [4]       16          87367492      -        D      <NA>      <NA>         4         1
  [5]       16 87367492-87367892      -        E      TRUE      TRUE         5         1
  [6]       16          87367892      -        A      <NA>      <NA>         6         1
                        txName        geneName
                <CharacterList> <CharacterList>
  [1] uc002fjv.3,uc002fjw.3,uc010vot.2           79791
  [2] uc002fjv.3,uc002fjw.3,uc010vot.2           79791
  [3] uc002fjv.3,uc002fjw.3,uc010vot.2           79791
  [4] uc002fjv.3,uc002fjw.3,uc010vot.2           79791
  [5] uc002fjv.3,uc002fjw.3,uc010vot.2           79791
  [6] uc002fjv.3,uc002fjw.3,uc010vot.2           79791
  -------
  seqinfo: 84 sequences from an unspecified genome
```

图 276 分配转录本解释

未经授权，不得用于任何商业用途！
联系方式：bmfo@zju.edu.cn

12、预测的拼接图和 FPKM 可以如前所述地可视化操作，将可变剪接高亮变色

为紫色（purple）（如图 277 所示）。请注意，从 RNA-seq 数据预测的大多数外显子和剪接点与 UCSC knownGene 表中的转录本一致（以灰色显示）。然而，与之前的图相反，预测的基因模型不包括未在数据中表达的剪接图的部分。此外，从 RNA-seq 数据中发现了未注释的外显子（E3，显示为紫色），如图中红色框出，其在四种正常结肠直肠样品（N2，N3，N4）中的三种中表达。

代码：

```
df <- plotFeatures(sgfc_pred, geneID = 1, color_novel = "purple")
```

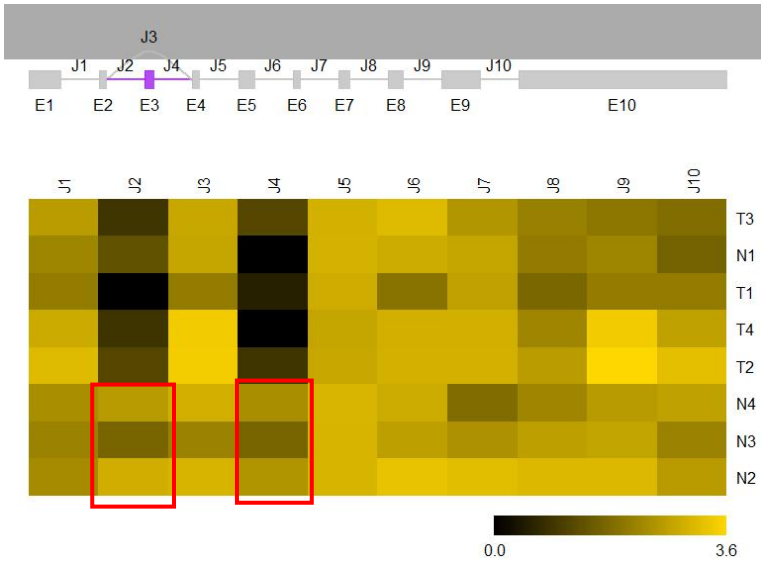


图 277 可变部分高亮结果

13、如果不考虑基因的完整剪接图，分析可以集中在个体剪接事件上。函数 analyzeVariants（）从图中递归地识别剪接事件，获得每个剪接变体的代表性计数，并计算相对剪接变体使用的估计，也称为“拼接的百分比”（PSI 或Ψ）。analyzeVariants（）返回一个 SGVariantCounts 对象。样本信息存储为 colData，SGVariants 存储为 rowRanges。 Assay variantFreq 存储每个剪接变体和样本的相对使用量的估计值。如前所述，可以使用访问器功能访问不同的数据类型。有关拼接变体的信息存储在 SGVariants 元数据列中，可以使用函数 mcols（）进行访问（如图 278 和图 279 所示）。

代码：

```
sgvc_pred <- analyzeVariants(sgfc_pred)
sgvc_pred
```

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

mcols(sgvc_pred)

```
> sgvc_pred <- analyzeVariants(sgfc_pred)
Find segments...
Find variants...
Annotate variants...
> sgvc_pred
class: SGVariantCounts
dim: 2 8
metadata(0):
assays(5): countsVariant5p countsVariant3p countsEvent5p countsEvent3p variantFreq
rownames: NULL
rowData names(20): from to ... variantType variantName
colnames(8): N1 N2 ... T3 T4
colData names(6): sample_name file_bam ... frag_length lib_size
> |
```

图 278 储存剪接变体估计值结果

```
> mcols(sgvc_pred)
DataFrame with 2 rows and 20 columns
            from             to        type   featureID   segmentID  closed5p  closed3p
     <character>     <character> <character> <character> <character> <logical> <logical>
1 D:16:87393901:- A:16:87380856:-           J          28           4      TRUE      TRUE
2 D:16:87393901:- A:16:87380856:-         JEJ    32,30,27           2      TRUE      TRUE
  closed5pEvent closed3pEvent    geneID   eventID variantID   featureID5p   featureID3p
      <logical>     <logical> <integer> <integer> <integer> <IntegerList> <IntegerList>
1          TRUE          TRUE         1         1         1            28            28
2          TRUE          TRUE         1         1         2            32            27
  featureID5pEvent featureID3pEvent                        txName        geneName
     <IntegerList>    <IntegerList>                <CharacterList> <CharacterList>
1           28,32           28,27 uc002fjv.3,uc002fjw.3,uc010vot.2           79791
2           28,32           28,27                                            79791
     variantType    variantName
   <CharacterList>    <character>
1           SE:S 79791_1_1/2_SE
2           SE:I 79791_1_2/2_SE
```

图 279 mcols 函数访问数据列

14、基于在每个变体的开始或结束延伸的结构上相容的片段，局部量化剪接变体。获得变体 i 的相对使用Ψi 的局部估计，作为与 i 相容的片段的数量除以与属于相同事件的任何变体相容的片段的数量。对于变体始端 S 和变体末端 E，$\widehat{\Psi}^{S}_{i} = x^{S}_{i}/x^{S}_{\cdot}$，$\widehat{\Psi}^{E}_{i} = x^{E}_{i}/x^{E}_{\cdot}$。对于具有有效估计的变体，单个估计被计算为局部估计的加权平均值，$\widehat{\Psi}_{i} = x^{S}_{\cdot}/(x^{S}_{\cdot}+x^{E}_{\cdot})\,\widehat{\Psi}^{S}_{i} + x^{E}_{\cdot}/(x^{S}_{\cdot}+x^{E}_{\cdot})\,\widehat{\Psi}^{E}_{i}$。对于读取次数较少的事件，相对使用的估计可能不可靠。如果为函数 analyzeVariants（）或 getSGVariantCounts（）指定了参数 min_denominator，则估计值将设置为 NA，除非至少有一个 xS.或者 xE.等于或大于指定值。注意，SGVariantCounts 对象还存储原始计数数据。计数数据可用于统计建模（如图 280 所示）。plotVariants 生成一个类似于 plotFeatures 的双区域图。顶部区域中的拼接图说明了所选的拼接事件。在此示例中，事件由两个变体组成，对应于跳过或包含未注释的外显子。热图说明了每种剪接变体的相对使用情况的估计值。样品 N2，N3 和 N4 显示包含外显子的转录物以及跳过外显子的转录物的证据。其余样本显示包含外显子的证据很少。

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

代码：

variantFreq(sgvc_pred)

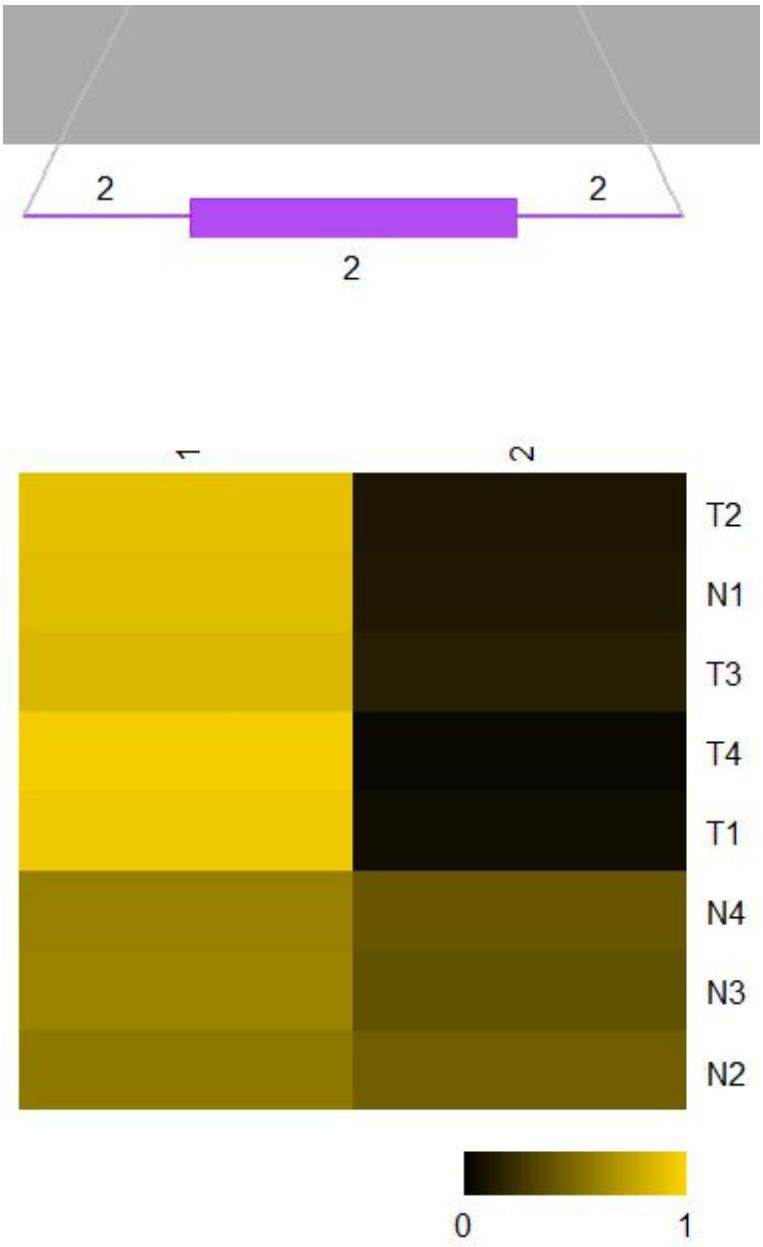


图 280 剪接变体区域显示图

15、剪接变体的功能性结果可以通过预测其对蛋白质编码潜力的影响来评估。函数 predictVariantEffects（）将 SGVariants 对象作为输入，可以是感兴趣的剪接变体，一组带注释的转录本以及作为 BSgenome 对象提供的匹配参考基因组（如图 281 所示）。输出数据表，每行描述特定剪接变体对注释的蛋白质编码转录本的影响。根据 HGVS recommendations 描述了变体的影响。变体效果预测相对较慢，

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

建议仅对选择变体运行 predictVariantEffects（）。

代码：

```
library(BSgenome.Hsapiens.UCSC.hg19)
seqlevelsStyle(Hsapiens) <- "NCBI"
vep <- predictVariantEffects(sgv_pred, txdb, Hsapiens)
vep
```

```
> library(BSgenome.Hsapiens.UCSC.hg19)
载入需要的程辑包：BSgenome
载入需要的程辑包：rtracklayer
> seqlevelsStyle(Hsapiens) <- "NCBI"
> vep <- predictVariantEffects(sgv_pred, txdb, Hsapiens)
Predicting effect of 2 variants on 3 coding transcripts...
> vep
  variantID     txName geneName                      RNA_change RNA_variant_type
1         2 uc002fjv.3    79791         r.88_89ins88+1798_88+1884        insertion
2         2 uc002fjw.3    79791     r.412_413ins412+1798_412+1884        insertion
3         2 uc010vot.2    79791 r.-105_-104ins-105+1798_-105+1884        insertion
                              protein_change protein_variant_type
1   p.K29_L30insRINPRVKSGRFVKILPDYEHMAYRDVYTC   in-frame_insertion
2 p.K137_L138insRINPRVKSGRFVKILPDYEHMAYRDVYTC   in-frame_insertion
3                                        p.=            no_change
>
>
```

图 281 导入示例文库

16、函数 plotFeatures（）和 plotVariants（）支持许多自定义图形的选项。 顶部面板中的拼接图由函数 plotSpliceGraph 绘制，可以直接调用。plotFeatures（）包括用于选择要显示的要素的多个参数。以下代码说明了绘制 FBXO31（Entrez ID 79791）的拼接图和表达水平的三种不同选项。默认情况下，plotFeatures（）生成的热图显示剪接点。可选择显示外显子区，或同时显示外显子区和剪接点（如图 282 所示）。

代码：

```
plotFeatures(sgfc_pred, geneID = 1)
plotFeatures(sgfc_pred, geneName = "79791")
plotFeatures(sgfc_pred, which = gr)
 plotFeatures(sgfc_pred, geneID = 1, include = "junctions")
 plotFeatures(sgfc_pred, geneID = 1, include = "exons")
 plotFeatures(sgfc_pred, geneID = 1, include = "both")
```

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

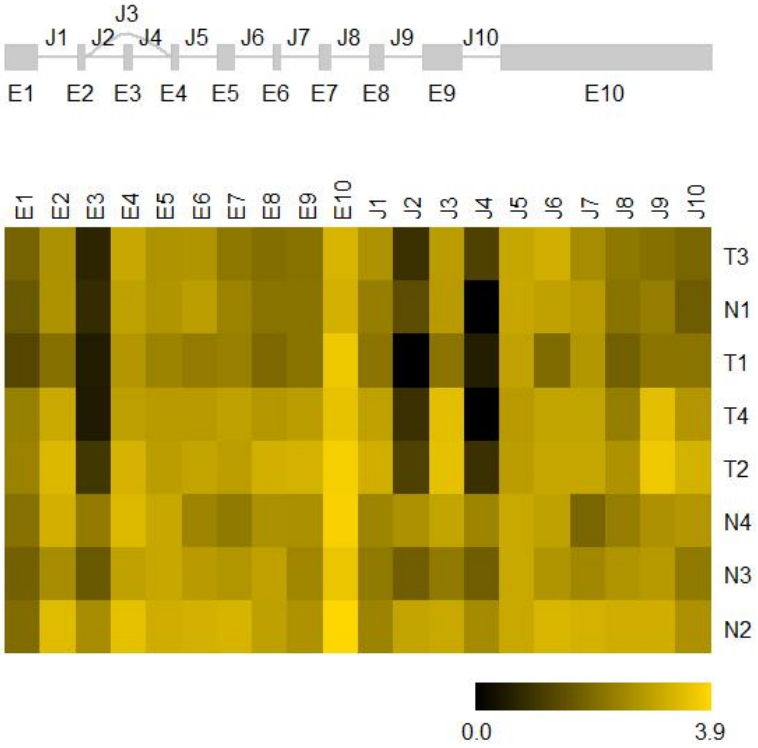


图 282 显示剪接点和外显子的热图

17、参数 toscale 控制基因模型的哪些部分按比例绘制（如图 283 所示）。

代码：

```
plotFeatures(sgfc_pred, geneID = 1, toscale = "gene")
plotFeatures(sgfc_pred, geneID = 1, toscale = "exon")
plotFeatures(sgfc_pred, geneID = 1, toscale = "none")
```

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

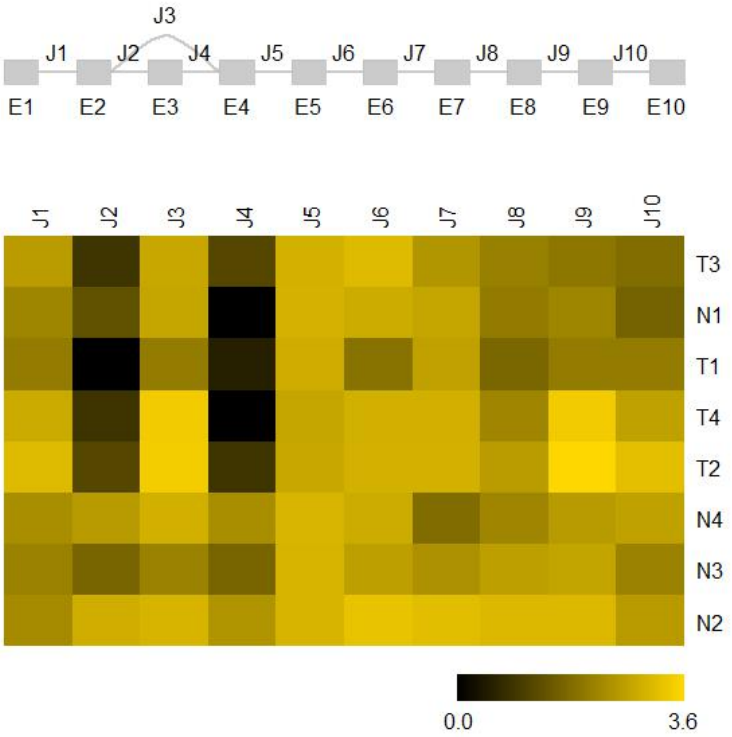


图 283 按比例的剪接点热图

18、热图允许对拼接连接点和外显子区域汇总的表达值进行可视化。或者可以使用函数 plotCoverage 显示每个碱基的读取覆盖率和剪接点计数（如图 284 所示）。

代码：

```
par(mfrow = c(5, 1), mar = c(1, 3, 1, 1))
plotSpliceGraph(rowRanges(sgfc_pred), geneID = 1, toscale = "none", color_novel = "red")
for (j in 1:4) {
    plotCoverage(sgfc_pred[, j], geneID = 1, toscale = "none")
 }
```

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

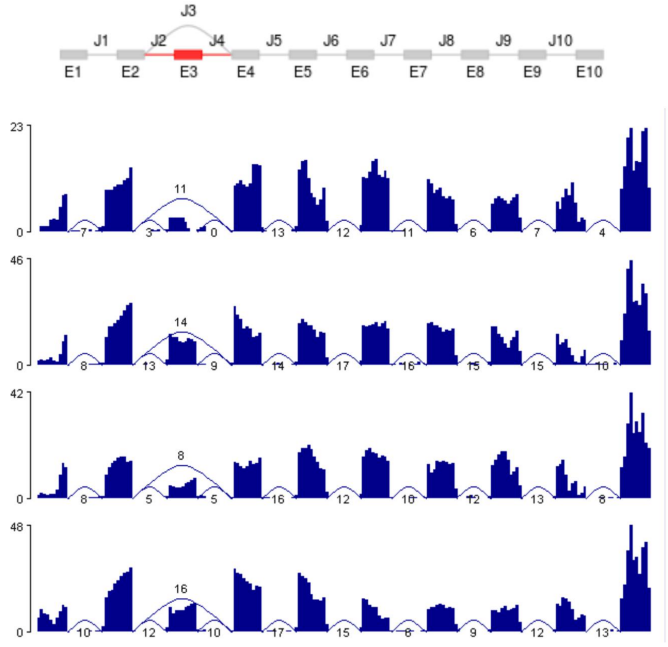


图 284 剪接模式图

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

6 sam 格式文件/bam 格式文件的分析——DExseq

6.1 关于 DExseq

DEXSeq 是一个 R 包，专注使用具有不同实验设计的样品之间的 RNA-seq 外显子计数来寻找外显子的差异使用。它提供的功能允许用户根据使用负二项分布的模型进行必要的统计检验，以估计生物重复和广义线性模型之间的差异以进行测试。该软件包还提供了可视化和探索结果的功能。由 Simon Anders 开发，原文献题目为：Detecting differential usage of exons from RNA-seq data，链接：https://genome.cshlp.org/content/22/10/2008.long。 最近一次更新于 2019.01.04，Bioconductor 上最新版本为 1.28.3。

对于每个外显子（或外显子的一部分）和每个样品，计算有多少读数映射到此外显子以及有多少读数映射到同一基因的任何其他外显子。考虑这两个计数的比率，以及它如何随条件变化，推断相对外显子使用的变化。

6.2 DExseq 的安装

1、和 SGSeq 相似，DEXSeq 也是基于 R 语言的程序包，关于 Windows 上 R 软件的安装在 6.2 中已经详细介绍，所以不再重复。因为纯粹的 R-Gui，不能及时地编译，我们将 DEXSeq 按照 Visual Studio Code 中 R 环境运行，提供的图片也基于 VS Code（如图 285 所示）。

代码：

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("DEXSeq", version = "3.8")
```

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
package ‘labeling’ successfully unpacked and MD5 sums checked
package ‘munsell’ successfully unpacked and MD5 sums checked
package ‘cli’ successfully unpacked and MD5 sums checked
package ‘fansi’ successfully unpacked and MD5 sums checked
package ‘pillar’ successfully unpacked and MD5 sums checked
package ‘checkmate’ successfully unpacked and MD5 sums checked
package ‘htmlwidgets’ successfully unpacked and MD5 sums checked
package ‘rstudioapi’ successfully unpacked and MD5 sums checked
package ‘gtable’ successfully unpacked and MD5 sums checked
package ‘lazyeval’ successfully unpacked and MD5 sums checked
package ‘plyr’ successfully unpacked and MD5 sums checked
package ‘reshape2’ successfully unpacked and MD5 sums checked
package ‘scales’ successfully unpacked and MD5 sums checked
package ‘tibble’ successfully unpacked and MD5 sums checked
package ‘viridisLite’ successfully unpacked and MD5 sums checked
package ‘withr’ successfully unpacked and MD5 sums checked
package ‘Formula’ successfully unpacked and MD5 sums checked
package ‘latticeExtra’ successfully unpacked and MD5 sums checked
package ‘acepack’ successfully unpacked and MD5 sums checked
package ‘gridExtra’ successfully unpacked and MD5 sums checked
package ‘data.table’ successfully unpacked and MD5 sums checked
package ‘htmlTable’ successfully unpacked and MD5 sums checked
package ‘viridis’ successfully unpacked and MD5 sums checked
package ‘htmltools’ successfully unpacked and MD5 sums checked
package ‘base64enc’ successfully unpacked and MD5 sums checked
package ‘xtable’ successfully unpacked and MD5 sums checked
package ‘locfit’ successfully unpacked and MD5 sums checked
package ‘ggplot2’ successfully unpacked and MD5 sums checked
```

图 285 程序包的安装

6.3 DExseq 的功能与使用方法

6.3.1 初步准备

1、加载程序包，同时检验安装结果。

代码：

```
library("DEXSeq")
```

2、数据集 pasilla，来源于与剪接调节相关的基因，要下载并加载 pasilla。

代码：

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("DEXSeq", version = "3.8")
library("pasilla")
```

3、将计数文件读入 R。

代码：

```
inDir = system.file("extdata", package="pasilla")
countFiles = list.files(inDir, pattern="fb.txt$", full.names=TRUE)
basename(countFiles)
```

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
flattenedFile = list.files(inDir, pattern="gff$", full.names=TRUE)
basename(flattenedFile)
```

4、之后，我们需要准备一个样本表。 该表应包含每个库的一行，以及所有相关信息的列，例如具有读取计数的文件名，实验条件，技术信息和其他协变量。 为了简化，我们即时构建表格。虽然这是一种准备表格的简单方法，但是使用已经在实验完成时准备好的现有表不易出错并且更加谨慎，将其保存为.CSV 格式并使用 R 函数 read.csv 加载它。

代码：

```
sampleTable = data.frame(
row.names = c( "treated1", "treated2", "treated3",
"untreated1", "untreated2", "untreated3", "untreated4" ),
condition = c("knockdown", "knockdown", "knockdown",
"control", "control", "control", "control" ),
libType = c( "single-end", "paired-end", "paired-end",
"single-end", "single-end", "paired-end", "paired-end" ) )
```

5、我们的表包含样品名称作为行名称和两个样本之间不同的协变量：首先是实验条件（具有水平控制和处理的因子条件）和库类型（因子 libType），我们包括因为此特定实验中的样本 部分在单端运行中进行测序，部分在成对结束运行中进行测序。目前，我们将忽略后一条信息，并推迟讨论如何将这些额外的协变量包含在第 4 节中。如果您只有一个协变量且想要执行简单分析，那么应该命名具有此协变量的列条件。现在，我们从这个数据构造一个 DEXSeqDataSet 对象。该对象包含所有输入数据，并将在后续分析的各个阶段传递。我们使用 DEXSeq 函数 DEXSeqDataSetFromHTSeq 构造对象。该函数有四个参数。首先，带有计数文件名称的向量，即使用 dexseq_count.py 脚本生成的文件。该函数将读取这些文件并将计数数据排列在矩阵中，该矩阵存储在 DEXSeqDataSet 对象 dxd 中。第二个参数是我们的示例表，第一个参数中列出的每个文件都有一行。这些信息只是按原样存储在对象的元数据槽中（见下文）。第三个参数是“sample + exon + condition：exon”形式的公式，它指定样本表列和'exon'变量中变量的对比度。使用这个公式，我们感兴趣的是由于'条件'变量的变化导致外显子使用的差异。在

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

这个小插图的后面，我们将介绍如何为复杂设计添加其他变量。第四个参数是文件名，现在是使用 dexseq_prepare_annotation.py 生成的展平 GFF 文件，并在创建计数文件时用作 dexseq_count.py 的输入。

代码：

```
uppressPackageStartupMessages( library( "DEXSeq" ) )
dxd = DEXSeqDataSetFromHTSeq(
countFiles,
sampleData=sampleTable,
design= ~ sample + exon + condition:exon,
flattenedfile=flattenedFile )
```

6.3.2 一般分析流程

1、为了演示 DEXSeq 工作流程，我们将使用上一节中构造的 DEXSeqDataSet。但是，为了使运行时间较短，我们将对象仅限于几个基因。

代码：

```
genesForSubset = read.table(
file.path(inDir, "geneIDsinsubset.txt"),
stringsAsFactors=FALSE)[[1]]
dxd = dxd[geneIDs( dxd ) %in% genesForSubset,]
```

2、DEXSeqDataSet 类派生自 DESeqDataSet。因此，它包含列数据，行数据和一些特定的访问函数。DEXSeqDataSet 对象中的核心数据是每个外显子的计数。DEXSeqDataSet 的每一行在每列中包含来自给定外显子（'this'）的计数数据以及来自属于同一基因（'others'）的其他外显子之和的计数数据。

（1）此注释以及有关 DEXSeqDataSet 的每列的所有信息都在 colData 中指定。

代码：

```
colData(dxd)
```

（2）我们可以先访问计数的前 5 行。

代码：

```
head( counts(dxd), 5 )
```

（3）请注意，列数为 14，前七个（七个样本）对应于映射到外部区域的读取数

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

量，后七个对应于映射到每个样本上的相同基因其余外显子的计数总和。

代码：

```
split( seq_len(ncol(dxd)), colData(dxd)$exon )
```

（4）我们也可以只访问属于外显区域的计数中的前五行（'this'）。

代码：

```
head( featureCounts(dxd), 5 )
```

（5）在以上两种情况下，行都标有基因 ID（此处为 Flybase ID），后跟冒号和计数区号。（由于计数区对应外显子或外显子的一部分，此 ID 在 DEXSeq 中称为特征 ID 或外显子 ID。）表内容表示已映射到相应样品中每个计数箱的读数值。要查看计数区的详细信息，我们还会发布前 3 行要素数据的注释。

代码：

```
head( rowRanges(dxd), 3 )
```

（6）到目前为止，该表包含有关注释数据的信息，例如基因和外显子 ID，外显子的基因组坐标以及包含外显子的转录本列表。访问器函数 annotationData 显示带有样本注释的设计表。

代码：

```
sampleAnnotation( dxd )
```

3、在接下来的部分中，我们将通过调用多个分析函数来更新对象，总是使用语句“dxd = someFunction（dxd）”，它接受 dxd 对象，填写执行计算的结果并写入返回的和将对象更新回变量 dxd。不同的样品可能会以不同的深度排序。为了调整这种覆盖偏差，我们估计尺寸因子，其测量相对测序深度。DEXSeq 使用与 DESeq 和 DESeq2 相同的方法，它在函数 estimate SizeFactors 中提供。这是样本的规范化。

代码：

```
dxd = estimateSizeFactors( dxd )
```

4、为了测试差异外显子的使用，我们需要估计数据的可变性。这对于能够区分技术和生物变异（噪声）与由于不同条件对外显子使用的实际影响是必要的。从数据集中的生物学重复推断出关于噪声强度的信息，并且通过所谓的分散来表征。在 RNA-Seq 实验中，重复的数量通常太小而不能通过外显子单独外显地可

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

靠地估计方差或分散参数，因此，方差信息在外显子和基因之间以强度依赖性方式共享。我们将讨论简单的单向设计：在此设置中，具有相同实验条件的样本被视为重复，因此，设计表需要包含名称条件的列。为了估计分散估计，DEXSeq 使用程序包 DESeq2 的方法。在内部，调用 DESeq2 的函数，根据 DEXSeq 模型的特定情况调整函数的参数。简而言之，使用 Cox-Reid 调整的轮廓似然估计计算每外显子分散，然后将分散-平均关系拟合到该单独的分散值，最后，将拟合值作为先验，以缩小每个外显子，估计拟合值。

代码：

```
dxd = estimateDispersions( dxd )
```

5、作为收缩诊断，DEXSeqDataSet 继承了 plotDispEsts 方法，该方法允许我们绘制每个外显子的分散估计值与平均归一化计数，得到的拟合值和后验（收缩）分散估计值。

代码：

```
plotDispEsts( dxd )
```

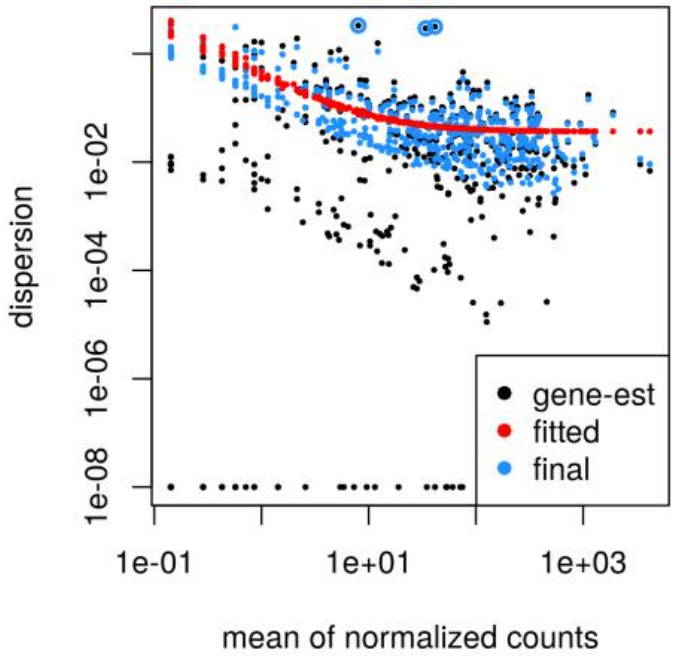


图 286 合适的估计模型

6、通过色散估计和尺寸因子，我们现在可以测试差异外显子的使用。对于每个基因，DEXSeq 拟合具有公式 ~ sample + exon + condition:exon 的广义线性模型，

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

并将其与较小的模型（空模型）～ sample + exon 进行比较。在这些公式中，样本是每个样本具有不同水平的因子，条件是我们在分析开始时构造 DEXSeqDataSet 对象时定义的实验条件的因子，并且外显子是具有两个水平的因子，这个和其他，在我们生成 DEXSeqDataSet 对象时指定了。这些公式描述的两个模型适合于每个计数箱，其中提供给拟合的数据包括每个样本的两个读取计数值，对应于外显子因子的两个水平：映射到相关仓的读取数量（水平此），以及来自同一基因的所有其他区域的读数的总和（其他水平）。使用χ2-分布比较两个拟合的偏差，提供 p 值。基于这个 p 值，我们可以决定空模型是否足以解释数据，或者是否可以拒绝它以支持替代模型，该模型包含条件：外显子的相互作用系数。后者意味着在测试中落入外显子的基因读数部分在实验条件之间显着不同。函数 testForDEU 为每个基因中的每个外显子执行这些测试。

代码：

```
dxd = testForDEU( dxd )
```

7、生成的 DEXSeqDataSet 对象包含有关测试信息的插槽。对于某些用途，我们可能还想估计相对外显子使用倍数的变化。为此，我们调用 estimateExonFoldChanges。外显子使用倍数变化是根据 GLM 拟合的系数计算的，该系数使用公式 count ～ condition + exon + condition:exon，其中“condition”可以用 colData 的任何列名替换。得到的系数允许估计不同条件下外显子使用的变化。注意，外显子使用的差异与跨条件的基因表达差异不同。例如，考虑在条件之间差异表达的基因的情况，并且具有在条件之间差异使用的一个外显子。根据拟合模型的系数，可以区分整体基因表达效应，其改变来自所有外显子的计数，来自外显子使用效应，其由相互作用项条件：外显子和单独影响每个外显子捕获。

代码：

```
dxd = estimateExonFoldChanges( dxd, fitExpToVar="condition")
```

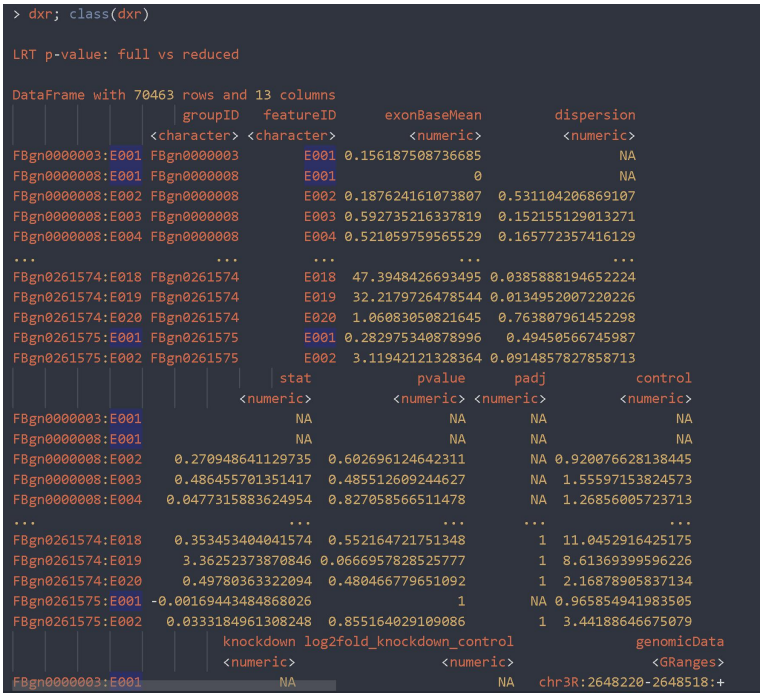
8、到目前为止，中间和最终结果已存储在 DEXSeqDataSet 对象的元数据中，可以使用函数 mcol 访问它们。为了总结结果而不显示中间步骤的值，我们调用函数 DEXSeqResults。其结果是一个 DEXSeqResults 对象，它是 DataFrame 对象的子类（如图 287 所示）。

代码：

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

```
dxr1 = DEXSeqResults( dxd )
```

```
dxr1
```



```
> dxr; class(dxr)

LRT p-value: full vs reduced

DataFrame with 70463 rows and 13 columns
                     groupID   featureID      exonBaseMean         dispersion
                 <character> <character>         <numeric>          <numeric>
FBgn0000003:E001 FBgn0000003        E001 0.156187508736685                 NA
FBgn0000008:E001 FBgn0000008        E001                 0                 NA
FBgn0000008:E002 FBgn0000008        E002 0.187624161073807  0.531104206869107
FBgn0000008:E003 FBgn0000008        E003 0.592735216337819  0.152155129013271
FBgn0000008:E004 FBgn0000008        E004 0.521059759565529  0.165772357416129
...                       ...         ...               ...                ...
FBgn0261574:E018 FBgn0261574        E018  47.3948426693495 0.0385888194652224
FBgn0261574:E019 FBgn0261574        E019  32.2179726478544 0.0134952007220226
FBgn0261574:E020 FBgn0261574        E020  1.06083050821645  0.763807961452298
FBgn0261575:E001 FBgn0261575        E001 0.282975340878996   0.49450566745987
FBgn0261575:E002 FBgn0261575        E002  3.11942121328364 0.0914857827858713
                          stat            pvalue      padj           control
                     <numeric>         <numeric> <numeric>         <numeric>
FBgn0000003:E001            NA                NA        NA                NA
FBgn0000008:E001            NA                NA        NA                NA
FBgn0000008:E002   0.270948641129735  0.602696124642311        NA 0.920076628138445
FBgn0000008:E003   0.486455701351417  0.485512609244627        NA  1.55597153824573
FBgn0000008:E004   0.0477315883624954  0.827058566511478        NA  1.26856005723713
...                       ...               ...       ...               ...
FBgn0261574:E018   0.353453404041574  0.552164721751348         1  11.0452916425175
FBgn0261574:E019    3.36252373870846 0.0666957828525777         1  8.61369399596226
FBgn0261574:E020    0.49780363322094  0.480466779651092         1  2.16878905837134
FBgn0261575:E001 -0.00169443484868026                 1        NA 0.965854941983505
FBgn0261575:E002   0.0333184961308248  0.855164029109086         1  3.44188646675079
                 knockdown log2fold_knockdown_control              genomicData
                 <numeric>                 <numeric>                <GRanges>
FBgn0000003:E001        NA                        NA   chr3R:2648220-2648518:+
```

图 287 DEXSeqResults 函数调用后输出的结果

9、可以在元数据列中找到对象 DEXSeqResults 的每个列的描述。

代码：

```
mcols(dxr1)$description
```

10、从这个目标，我们可以判断有多少外显子区是显著的，设定错误发现率为 10%。

代码：

```
table ( dxr1$padj < 0.1 )
```

11、我们还可以判断有多少基因受到影响。

代码：

```
table ( tapply( dxr1$padj < 0.1, dxr1$groupID, any ) )
```

12、示例数据集仅包含一类基因。作者们选择了有趣的案例。因此，这么大部分基因受到显着影响的事实并不典型。要了解检测差异外显子使用的能力如何取决于映射到外显子的读数的数量，MA 图很有用，它绘制了倍数变化对数的对数（如图 288 所示）。

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

代码：

```
plotMA( dxr1, cex=0.8 )
```

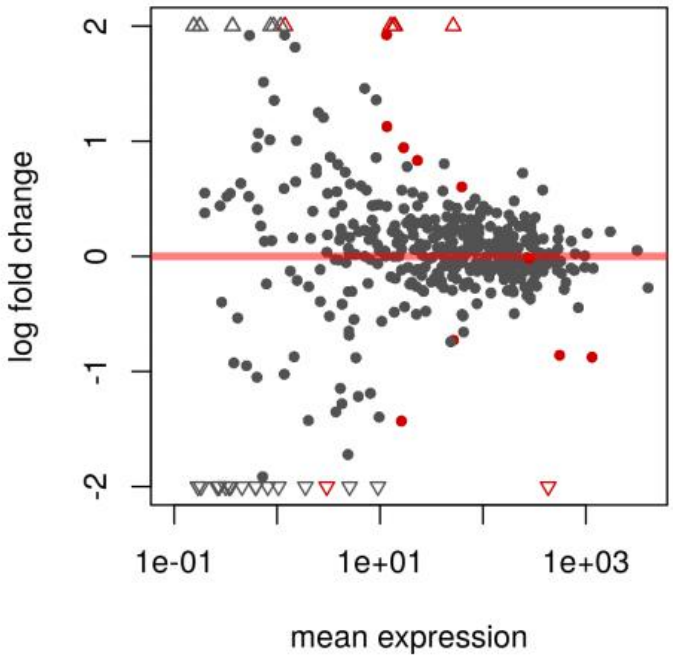


图 288 MA 图

6.3.3 可视化

1、plotDEXSeq 提供了一种分析结果可视化的方法。该图显示了基因 FBgn0000210 的每个外显子的拟合表达值，对于处理和未处理的两种条件中的每一种。函数 plotDEXSeq 至少需要两个参数，即 DEXSeqDataSet 对象和基因 ID。选项 legend = TRUE 会导致包含图例。上面代码中剩下的三个参数是普通的绘图参数，它们只是简单地移交给 R 的标准绘图函数。它们不是必需的，而是用于改善外观（如图 289 所示）。

代码：

```
plotDEXSeq( dxr2, "FBgn0000210", legend=TRUE, cex.axis=1.2, cex=1.3, lwd=2 )
```

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

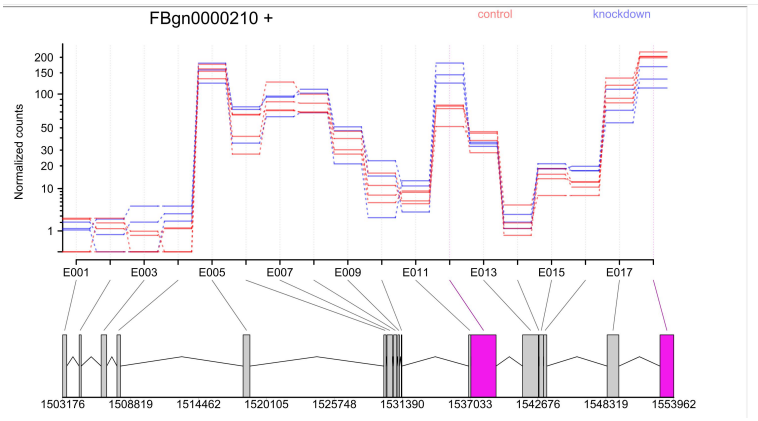


图 289 表达估计值作图

2、还可以选择使转录物模型可视化，这可用于将差异外显子使用结果置于同种型调节的背景中（如图 290 所示）。

代码：

```
plotDEXSeq( dxr2, "FBgn0000210", displayTranscripts=TRUE, legend=TRUE, cex.axis=1.2, cex=1.3, lwd=2 )
```

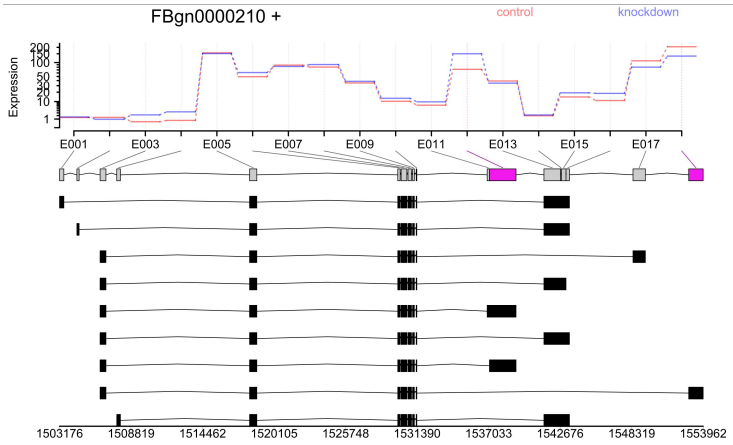


图 290 包含注释的转录本模型图

3、其他有用的选项是查看单个样本的计数值，而不是模型效应估计值。 对于此显示（选项 norCounts = TRUE），通过将它们除以尺寸因子来对计数进行标准化。

代码：

```
plotDEXSeq( dxr2, "FBgn0000210", expression=FALSE, norCounts=TRUE, legend=TRUE, cex.axis=1.2, cex=1.3, lwd=2 )
```

4、如前所述，DEXSeq 旨在发现相对外显子使用的变化，例如，个体外显子表

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

达的变化不仅仅是基因整体上调或下调的结果。为了使这些变化可视化，有时从图中去除表达式的整体变化是有利的。

代码：

```
plotDEXSeq( dxr2, "FBgn0000210", expression=FALSE, splicing=TRUE, legend=TRUE, cex.axis=1.2, cex=1.3, lwd=2 )
```

5、为了生成对所有分析结果的易于浏览的详细概述，该软件包提供了一个 HTML 报告生成器，在函数 DEXSeqHTML 中实现。此函数使用包 hwriter 创建一个结果表，其中包含指向重要结果的图的链接，允许更详细地探索结果。

代码：

```
DEXSeqHTML( dxr2, FDR=0.1, color=c("#FF000080", "#0000FF80") )
```

6.3.4 大样本并行分析

1、DEXSeq 分析在计算上可能很重，特别是对于包含大量样本的数据集，或者包含含有大量外显子的基因的基因组。虽然分析的一些步骤对整个数据集起作用，但是对于某些步骤可以并行化计算负荷。我们使用 BiocParallel 包，并实现了函数 estimateDispersions，testForDEU 和 estimateExonFold 的 BPPARAM 参数更改。

代码：

```
BPPARAM = MultiCoreParam(workers=4)
dxd = estimateSizeFactors( dxd )
dxd = estimateDispersions( dxd, BPPARAM=BPPARAM)
dxd = testForDEU( dxd, BPPARAM=BPPARAM)
dxd = estimateExonFoldChanges(dxd, BPPARAM=BPPARAM)
```

2、对于运行大量样本（例如，超过 100 个）的分析，我们建议将 BatchJobsParam 类配置为 BPPARAM，以便在计算机群集中分配计算并显着减少运行时间。 用户还可以考虑通过过滤低表达的同种型或低计数的外显子区域来减少测试次数。通过减少运行时间，此过滤步骤也可以获得更准确的结果。

6.3.5 一条命令进行标准差异外显子的使用分析

在前面的部分中，我们逐步进行了分析。一旦您对数据的工作流程充满信心，就可以通过包装函数 DEXseq 简化其调用，该函数通过单个函数调用运行上面显示

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

的分析。在最简单的情况下，构造 DEXSeqDataSet，然后运行 DEXSeq，将 DEXSeqDataSet 作为唯一参数传递，此函数将输出 DEXSeqResults 对象。

代码：

```
dxr = DEXSeq(dxd)
class(dxr)
```

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn

7 参考文献

[1]Goldstein LD, Cao Y, Pau G, Lawrence M, Wu TD, Seshagiri S, Gentleman R (2016) Prediction and Quantification of Splice Events from RNA-Seq Data. PLoS ONE 11(5): e0156132.

[2]刘海鹏，利用 VMware 虚拟机改善 Linux 的教学实训环境[J].信息系统工程，2018(11):164-166

[3]Detecting differential usage of exons from RNA-seq data.Simon Anders, Alejandro Reyes and Wolfgang Huber

未经授权，不得用于任何商业用途！
联系方式：binfo@zju.edu.cn