JIYINZU
XINXIXUE
SHIJIAN
JIAOCHENG

基因组信息学实践教程

张高川 主编

苏州大学出版社
Soochow University Press

基因组信息学实践教程

主　编　张高川
参　编　胡　广　李　渊　商冰雪
　　朱彦博

苏州大学出版社

图书在版编目(CIP)数据

基因组信息学实践教程 / 张高川主编. —苏州：苏州大学出版社，2022.9
ISBN 978-7-5672-3866-4

Ⅰ.①基… Ⅱ.①张… Ⅲ.①基因组-生物信息论-高等学校-教材 Ⅳ.①Q343.2

中国版本图书馆 CIP 数据核字(2022)第 117904 号

书　　名：基因组信息学实践教程

主　　编：张高川
责任编辑：赵晓嫱

出版发行：苏州大学出版社(Soochow University Press)
地　　址：苏州市十梓街 1 号　**邮编**：215006
印　　装：广东虎彩云印刷有限公司
网　　址：http://www.sudapress.com
邮　　箱：sdcbs@suda.edu.cn
邮购热线：0512-67480030
销售热线：0512-67481020

开　　本：787 mm×1 092 mm　1/16　**印张**：24.5　**字数**：532 千　**插页**：2
版　　次：2022 年 9 月第 1 版
印　　次：2022 年 9 月第 1 次印刷
书　　号：ISBN 978-7-5672-3866-4
定　　价：75.00 元

前　言

高通量测序技术的飞速发展，极大地降低了基因组研究的时间和成本。该技术也因此在生命科学研究中得到了极大的推广与运用，并迅速地产生海量的基因组数据。这些海量的基因组数据涵盖了各种生物遗传信息，由此引发了由数据驱动的革命，包括比较基因组学和个性化药物等，进而对基础生物学和生物医学转化研究产生了深远影响。基因组学（Genomics）这门学科亦应运而生，它是一个跨学科的生物学领域，专注于基因组的结构、功能、进化、作图和编辑。“基因组信息学”这门课程就是针对这门学科面向生物信息学专业本科学生而开设的。

自从 2014 年教授“基因组信息学”这门课程开始，我一直为教材的事情所困扰，遍寻国内外出版的相关书籍，没有找到适合本课教学的教材，最终选择 David W. Mount 编著的《生物信息学：序列与基因组分析》（*Bioinformatics*：*Sequence and Genome Analysis*）一书。这本书中的专业理论知识全面而深入，但是对于本科生来说，读懂弄通的难度偏大，且缺乏配套的实践内容。不过，理论讲授内容的参考书籍总算有了着落。然而，配套的实践教程无论如何都找不到，我只能在教学过程中根据理论讲授内容并结合自身科研实践经验，开始自主编写配套的实验讲义；同时，一边使用，一边结合这门学科的发展情况加以修订和完善。然而，这门学科的发展如此迅猛，甚至可以说日新月异，这一点可以从 PubMed 检索相关主题文献而获知。故而，尚有一些优秀的基因组信息学研究内容和相关分析工具未能包含进来。我们将在今后持续跟进，不断更新本实践教程的内容。

由于编者水平有限，同时该教程是为了适应本校教学环境和内容需要而编写的，内容可能有所偏颇，还请读者予以谅解，并多提宝贵的意见和建议。

张高川

2022 年 7 月 30 日

目　录

绪 论

一、 内容简介

本教程的所有实践内容主要围绕基因组测序分析来设计。实践项目的主要内容包括基因组测序模拟、序列组装、全基因组的同源搜索、从头预测基因及结构建模、基于转录组测序数据的基因组注释、启动子等基因组序列特征的分析和预测、基因组数据的可视化、多种基因组序列分析软件的对比和综合运用等。这些实践内容不仅涉及基因组信息学方面常用数据库的检索,还包含常用软件的使用以及编程和数据统计方法的综合运用。

本教程主要适用于生物信息学专业本科学生的基因组信息学方面的实践教学。这些实践项目的开展,需要学生具备一定的生物信息学方面的基础理论知识和操作技能、计算机基础、Python 编程与基于 R 语言的统计分析能力。本教程所涉及的分析软件大多适用于 UNIX/Linux 系统,故而还需要熟悉此类操作系统环境,同时安装好实践项目中所需要的各种专业软件。这些软件的快速安装方法见绪论第四部分“操作环境”,其使用方法见“常用软件使用手册”部分。此外,由于实践内容涉及高通量计算,故而还需一台高性能的计算设备。考虑到不同软件对计算性能的需求和实践教学的目的,对于实践内容所涉及的目标物种基因组,建议根据自身所配备的计算机设备进行选择。本教程基于一个注释信息完善的小规模基因组(酿酒酵母 *Saccharomyces cerevisiae* S288C 菌株),提供基础实践演示案例。

二、 项目设计

本教程的实践项目分为三个层次——基础实践、扩展实践和综合实践,随着实践内容的推进,逐步拓展学生的知识面,加深学生对基因组信息学相关理论知识的理解,提升其操作技能。

① 基础实践:整个实践内容以一个完整项目的形式开展,各个实践项目之间既相对独立,又具有一定的连贯性。实践项目从全基因组测序模拟开始,进行组装、注释和可视化等,通过一系列基础实践教学活动,让学生熟悉最常用的基因组信息学软件的使用方法,并掌握一套相对完整的基因组序列数据的基本分析流程。

② 扩展实践:在基础实践之上,进一步增加对于基因组分析策略和方法相关理论与实践技能的训练,进而拓展学生的知识面,加深其对基因组信息学的理解和认知,同时提升学

生的自主学习能力。

③ 综合实践:在学生掌握基因组信息学基本理论、技术和方法的基础之上,设计了一些综合性和探索性的实践项目,以加强学生对专业知识和技能的综合运用能力;同时,还让学生能利用其他专业相关课程内容,如算法、编程和数理统计方面的技能和方法,来解决实践中遇到的各种问题,进而提升学生发现问题、分析问题和解决问题的能力。

三、 问题探索

实践开展过程中,可能会出现各种各样的问题。有的是人为错误导致的,有的是系统兼容性问题引起的,有的是所选数据有问题造成的,当然亦有可能是分析方法和软件工具自身的固有问题。因此,实践获得的结果往往达不到预期的那样“完美”。这些问题当中甚至有些可能在这门学科中目前都是悬而未决的,亦值得我们继续深入探索,可以结合基因组信息学相关理论知识和前沿发展,在实践过程中对这些问题进行探索和思考。以下是根据个人经验和认知列举的一些问题,仅供参考。

不同的基因组序列组装软件,针对同一组基因组测序数据的组装结果是否会存在差异?这是为什么?该怎么处理?

不同的组装软件,所使用的具体算法模型和训练数据集往往是不同的,故而导致其组装行为有一定的偏好,最终导致它们在组装同一个物种基因组时,结果存在差异。这也提醒我们在进行某个物种的全基因组测序组装时,一定要选择合适的组装软件。如果无法确定合适的组装软件,则尝试多种软件的对比和联合使用,或许是一个可行的方案。

当需要对某个物种基因组测序数据进行组装时,如果选择了某个组装软件,那么如何设置该软件的参数组合,才能使得组装结果最佳?

基因组组装软件的参数选项通常都比较多,虽然很多参数都有默认设置,但是这些默认设置并不一定适用于某个具体物种的基因组组装。换句话说,就是使用默认参数组合,也不一定能够获得最佳组装结果。那么如何设置这些参数才能获得最佳组装结果呢?实际上,在结果出来之前谁也不知道,只能通过不同参数组合进行组装测试和比较。采用正交试验设计等方法对其进行参数组合设计,可以有效减少测试次数;当然,如果不考虑计算资源限制的话,可以使用穷举法遍历所有参数组合。为了减少手动执行程序的烦琐操作,可以编写批处理脚本进行不同参数组合的组装测试,并编写程序提取组装结果的关键指标进行比对,从而选出最佳组装结果。

当需要对某个物种基因组测序数据进行组装时,如何选择合适的组装软件?

基因组测序组装软件有很多,每个软件都有各自相关的文献资料对其功能、训练和测试数据集,及其适用的数据类型和/或物种分类进行描述。我们可以通过阅读这些文献资料,再结合目标物种类别和基因组测序数据类型,对适用软件进行初步筛选。如果适用的软件有多个,则需要分别对其进行组装测试,然后将组装结果的关键指标进行比对,进而选出其中最佳的组装结果。

如何让预测的基因结构尽可能准确?

这是基因预测和结构建模中的一个至关重要的问题。迄今为止,没有哪一种基因预测的方法和软件能够做到尽善尽美,都有或多或少的问题。很自然的一个想法,就是多种预测方法和软件的联合使用,或许能够改善预测的基因结构模型。

同源基因搜索方法在全基因组基因预测中的用途是什么?

一般情况下,亲缘关系越近的物种,就会编码越多的在进化上同源的基因。这些同源基因在序列上往往是高度相似的,故而利用近缘物种的已知基因/蛋白质序列,在目标物种基因组中来预测可能的同源基因是一种可行的方案。这种同源基因的搜索基于相似性比对的原理,可以定位同源基因在目标物种基因组上的位置以及大概结构;但是由于该方法内在原因的限制,实验研究人员无法获得精准的基因结构信息,故尚需其他基因结构建模软件来辅助解决。

不同的从头计算预测基因软件,对于同一物种基因组的预测结果存在差异,这是为什么?该如何处理?

从头计算预测基因的软件有很多,如 Augustus、GENSCAN、GeneMark-ES/ET 等。但是,不同基因预测软件所使用的具体算法模型和训练数据集往往是不同的,这会导致软件在计算和预测的参数上有一定的偏好,即其所适用的物种类别不同,最终导致不同软件在预测同一物种基因组时结果存在差异。这也提醒我们在做全基因组的基因预测时,必须根据自己的研究对象和实际分析需求来选择一个合适的软件。当然,在无法确定合适的软件时,可尝试多种软件的联合使用,也许可以改善基因预测结果。此外,从头计算预测的基因虽然具有明确的基因结构,但是其不一定是真实的,尚需要其他方法加以佐证,尤其是实验证据,以便得到更加完整、准确的基因结构信息。

如何利用某物种的转录组测序数据来改善全基因组的基因预测和结构建模结果?

同源搜索和从头计算预测这两种方法获得的基因都有一个共同缺点,就是无法确保该基因真正表达。转录组测序数据则可以改善这一问题,如 RNA-Seq 数据。通过组装 RNA-Seq 数据,可以获得大量基因真实表达的转录本序列信息。基于这些组装转录本序列,将其映射到基因组序列上,可获得这些真实转录本的基因组定位和结构信息;这些结构信息,除了首尾之外,中间的区间间隔都代表了真实且准确的外显子和内含子结构。利用这些结构信息,再结合上述同源搜索和从头计算预测基因结果,即可进一步完善基因信息。

全基因组注释,除了要进行基因预测之外,通常还需要对哪些基因组序列特征进行注释?

在全基因组注释中,除了基因预测之外,还有很多基因组序列特征需要进行注释,如与基因表达调控有关的启动子和转录因子、tRNAs 和 microRNAs、重复序列等。在这些序列特征的注释过程中,需要考虑是否有实验证据信息、计算方法和分析软件可以直接使用。

为什么要对基因组序列及其注释信息进行可视化?

无论是基因组序列,还是注释信息,都非常庞大,不便于实验研究人员的访问和使用。

软件工具需要提供简单易用的访问形式,以便实验研究人员查询和使用这些基因组数据。其中,可视化就是最经典的一种方式,它提供了用户友好的图形化操作界面和快速搜索功能。

如何对全基因组数据进行可视化?现有的哪些工具可以使用?选择本地软件,还是基于 Web 的工具进行可视化?

目前,有许多工具可以实现或部分实现全基因组数据的可视化,如 Integrative Genomics Viewer(IGV)、JBrowse 等。基因组数据的本地可视化只能服务个人;而基于 Web 的可视化则可以利用 Web 服务器对所有人开放,这样更加有利于相关学术成果的交流和推广。

你是否熟悉各种常用基因组数据格式?包括编写程序从某个格式文件中提取所需的特定数据,或按照指定格式写入分析结果。

在基因组信息学研究过程中,不同研究人员根据自身科研需求开发了不同的分析软件,各种各样的基因组数据格式也随之产生,用于保存各种不同类型的数据。换句话说,就是不同软件所支持的数据文件格式亦不尽相同,这给一些需要多个软件协同处理的复杂分析任务带来了不便。此时,就需要对这些数据文件进行格式转换。在缺少合适转换工具的情况下,我们需要自行编程完成相关的数据解析和提取工作,为此就必须熟练掌握这些数据格式的转换操作。

四、 操作环境

1. 软件安装

以下的软件安装和测试系统为 Ubuntu Linux 16.04(LTS 版本),其中绝大多数软件可以使用同样的方法在更高版本的 Ubuntu 系统中正常安装和运行。如果你的计算机没有安装此类系统,可以选择安装双系统,或者安装虚拟机(如 VirtualBox),然后在虚拟机中安装该系统;只是以虚拟机方式安装的话,系统计算性能会受到很大限制,只能使用很小规模的数据来开展实践项目。本教程所需软件的快速安装和环境配置方法如下;更多有关软件安装、环境配置和使用方法的内容,详见“常用软件使用手册”部分。

(1) ART 系列软件

从 ART 官网下载最新编译好的版本,如 artbinmountrainier2016.06.05linux64.tgz;解压缩后,将该目录下所有文件复制到/usr/bin 目录下。

```
wget https://www.niehs.nih.gov/research/resources/assets/docs/artbinmountrainier2016.06.05linux64.tgz
tar -xzf artbinmountrainier2016.06.05linux64.tgz
cd art_bin_MountRainier
sudo cp -a ./* /usr/bin/
```

(2) NCBI SRA Toolkit

```
sudo apt install sra-toolkit
```

（3）NCBI BLAST

```
sudo apt-get install ncbi-blast +
```

（4）FastQC

```
sudo apt-get install fastqc
```

（5）Bowtie2

```
sudo apt-get install bowtie2
```

（6）Samtools

```
sudo apt-get install samtools
```

（7）SOAPdenovo 和 SOAPdenovo2

```
sudo apt-get install soapdenovo soapdenovo2
```

（8）Velvet

```
sudo apt-get install velvet
```

（9）QUAST

从其官网下载最新编译好的版本，如 quast-5.0.2.tar.gz；解压缩后，将该目录下所有文件复制到/usr/bin 目录下。

```
wget https://downloads.sourceforge.net/project/quast/quast-5.0.2.tar.gz
tar -xzf quast-5.0.2.tar.gz
cd quast-5.0.2
sudo cp -a ./*/usr/bin/
```

（10）GFF 格式处理软件

从其官网下载最新编译好的版本；解压缩后，将该目录下所有文件复制到/usr/bin 目录下。

```
#gffread
wget http://ccb.jhu.edu/software/stringtie/dl/gffread-0.11.4.Linux_x86_64.tar.gz
tar -xzf gffread-0.11.4.Linux_x86_64.tar.gz
cd gffread-0.11.4.Linux_x86_64
sudo cp -a ./*/usr/bin/
#gffcompare
wget http://ccb.jhu.edu/software/stringtie/dl/gffcompare-0.11.2.Linux_x86_64.tar.gz
tar -xzf gffcompare-0.11.2.Linux_x86_64.tar.gz
cd gffcompare-0.11.2.Linux_x86_64
sudo cp -a ./*/usr/bin/
```

（11）GenomeTools

```
sudo apt-get install *genometools*
```

（12）BLAST 比对结果格式转换工具

可以从下列网站下载相关工具，如 blast92gff3. pl 和 blast2gff. py；然后，直接将其复制到/usr/bin 目录下。

```
wget http://eugenes.org/gmod/genogrid/scripts/blast92gff3.pl
sudo cp ./blast92gff3.pl /usr/bin/
wget https://github.com/wrf/genomeGTFtools/blob/master/blast2gff.py
sudo cp ./blast2gff.py /usr/bin/
```

（13）多序列比对及结果查看工具

```
sudo apt-get install clustalo clustalw clustalx jalview treeviewx seaview
```

（14）bedtools

```
sudo apt-get install bedtools
```

（15）BamTools

```
sudo apt install bamtools
```

（16）Augustus

在安装 Augustus 之前，需要安装很多系统依赖库和工具。

```
sudo apt-get install libboost-iostreams-dev zlib1g-dev libgsl-dev libmysql++-dev libsqlite3-dev libboost-graph-dev libsuitesparse-dev liblpsolve55-dev libbamtools-dev bcftools
```

然后，从其官网下载源码文件；解压缩后，进行编译安装。

```
wget http://bioinf.uni-greifswald.de/augustus/binaries/old/augustus-3.3.tar.gz
tar -xzf augustus-3.3.tar.gz
cd augustus
make
sudo make install
```

（17）geneid

从其官网下载最新版本源码文件；解压缩后，进行编译，再将编译的可执行文件复制到/usr/bin 目录下。

```
wget https://genome.crg.es/pub/software/geneid/geneid_v1.4.4.Jan_13_2011.tar.gz
tar -xzf geneid_v1.4.4.Jan_13_2011.tar.gz
cd geneid
make
sudo cp ~/geneid/bin/geneid /usr/bin
```

（18）IGV

从其官网下载适合自己系统的版本（需要相应版本的 Java 环境支持）；解压缩后，程序可以直接运行，打开 IGV 窗口。

（19）JBrowse

在安装 JBrowse 之前，需要安装很多系统依赖库和工具。

```
sudo apt install build-essential zlib1g-dev tabix
```

此外，还需要提前安装好 XAMPP；然后，下载 JBrowse 源码文件；解压缩后，将整个目录移到/opt/lampp/htdocs/jbrowse 目录中，并进行初始化安装。

```
wget https://github.com/GMOD/jbrowse/releases/download/1.16.6-release/JBrowse-1.16.6.zip
unzip JBrowse-1.16.6.zip
sudo mv JBrowse-1.16.6 /opt/lampp/htdocs/JBrowse
cd /opt/lampp/htdocs
sudo chown `whoami` JBrowse
cd JBrowse
./setup.sh #不要使用 sudo 模式来执行该脚本程序
```

2. 远程服务器的连接和使用

如果有专门用于高通量数据分析的高性能计算服务器，可使用第三方工具（如 FileZilla、WinSCP）或 sftp 指令，把数据上传到服务器指定工作目录中，然后在服务器上执行相关程序进行数据分析，再把分析结果文件下载到本地电脑，前提是服务器上已安装数据分析所需软件。如果服务器没有安装所需软件，且本地电脑配置一般，则现有设备只能处理小规模数据，而且运行某些程序可能需要较长时间，尤其是电脑性能配置不足的时候，很可能无法执行某些计算任务。

（1）远程登录服务器进行上传和下载

```
#Ubuntu16 终端命令行模式，使用 sftp 指令远程登录服务器
#假设服务器 IP 地址为 42.244.7.51，用户名为 zhanggaochuan
sftp zhanggaochuan@42.244.7.51 #在提示行中输入密码
#创建工作目录
mkdir workdir
#切换到服务器上当前用户的工作目录
cd workdir
#设定本地工作目录，即需要上传的数据文件所在目录
lcd /home/zhanggaochuan/workdir
#上传基因组序列文件，如：Sc_gDNA.fasta
put Sc_gDNA.fasta
#如果需要上传本地工作目录中的所有数据文件，可以使用：put *
#从服务器的工作目录中下载指定结果文件到本地工作目录
#假设 Samtools 统计结果文件存放在 samtools_out 子目录中
cd samtools_out
get samtools.stat.*
```

```
#下载该目录中所有文件,可以使用:get *.*
#下载完成后,如果要删除当前目录所有文件,可以执行:rm *.*
exit #退出 sftp 服务器
```

(2) 远程登录服务器执行数据分析工作

```
#Ubuntu16 终端命令行模式,使用 ssh 指令远程登录服务器(42.244.7.51)
ssh zhanggaochuan@42.244.7.51
#创建工作目录
mkdir workdir #如果已创建,请忽略
#切换到服务器上当前用户的工作目录
cd workdir
#假设所需数据文件已经存在,执行如下指令,系统则将以 10 个 CPU 核或线程数来执行 tblastn 程序,并将其挂载到后台运行,同时创建记录程序运行情况的 nohup 日志;此时退出终端命令行窗口,将不影响服务器端的程序执行。
nohup tblastn -query ./protein.fasta -db Sc_gDNA -out ./tblastn_results.outfmt6 -evalue 1e-5 -outfmt 6 -max_target_seqs 1 -num_threads 10 &
```

3. 不同操作系统的兼容性问题

在不同操作系统之间进行数据和代码文件的转移时,需要特别注意不同系统的兼容性问题。比如,Windows 系统的行结束符为“\r\n”(回车换行),而 UNIX/Linux 系统的行结束符为“\n”(换行);字符编码方式也有差异。这样的不兼容会导致 Windows 系统下撰写的数据和脚本文件,在移植到 UNIX/Linux 系统下运行时,会出现异常情况。故而,建议在 Windows 系统下安装 notepad、sublime text 之类的文本编辑器,而在 UNIX/Linux 系统下使用 gedit 之类的文本编辑器。然后,根据数据和脚本文件最终运行环境的操作系统类型,利用这些编辑器将其保存成该系统支持的格式。例如,在 Ubuntu 系统下,保存格式为 UTF-8 编码方式,行结束符为 UNIX/Linux 系统格式类型。这样就会解决此类兼容性问题。

(本教程中使用的相关脚本内容可登录苏大教育 www.sudajy.com 或扫描下方二维码获取。)

第一篇　基础实践

项目 1　基因组测序模拟

一、基本原理

随着下一代测序(next-generation sequencing,NGS)技术的发展,该技术在生命科学的不同领域迅速推广,如从基因组中识别小 RNA(small RNA)、转录因子结合模式、DNA 甲基化模式、循环微 RNA(circulating microRNA)、各种基因组结构变异等。这些不同分支领域的应用,催生了对不同的统计算法和分析工具的需求,以便研究人员从大量数据中提取有意义的信息。然而,基因组序列的复杂性为后续的序列分析和挖掘带来了很大的挑战,增加了算法研究和软件开发的难度。

为了应对新算法和软件的开发周期长的问题,很多基于 NGS 的数据模拟工具被开发出来。这些模拟工具可以模拟指定 NGS 平台的测序错误率分布,以及已知 DNA 序列中的其他影响因素,以此来创建高通量原始测序数据。这些 NGS 模拟数据可以协助开发和评估新的统计模型和计算方法,亦可用于评估新测序方法,进而构建更好的实验工作流程。

例如,Ruffalo 等(2012)使用 NGS 模拟器 ART 生成的模拟测序数据,测试短读段(reads)映射质量评估工具 LoQum。Li 等(2012)用模拟数据测试基于家族的变异调用算法(variant calling algorithm)。Liao 等(2013)用两个模拟器创建的模拟数据测试子读段(sub-reads)比对映射程序。

不同测序技术和平台的影响因素也不尽相同,故而有多种不同的 NGS 数据模拟器被开发出来,用于模拟各种不同类型的测序方法,包括 DNA 测序、宏基因组测序(metagenomic sequencing)、RNA-Seq、ChIP-Seq 和亚硫酸氢盐测序(bisulfite sequencing,BS-Seq)等。

本实践项目选择一款经典的测序模拟软件来模拟常用测序平台的测序数据,使学生在熟悉常用基因组数据库和模拟软件使用的同时,通过简单的统计分析,加深对测序中几个关键参数(测序读长、覆盖度、片段长度和覆盖率)的认知和理解。

二、 目的和要求

① 学会 GenBank 中 Genome 数据库的使用。

② 加深对全基因组鸟枪法测序原理的理解。

③ 熟悉和掌握基因组测序模拟软件 ART 的使用方法。

④ 能够利用统计学方法和技能对实验数据进行统计分析。

三、 软件和数据库资源

① ART 软件(https://www.niehs.nih.gov/research/resources/software/biostatistics/art)。

② GenBank 中的 Genome 数据库(https://www.ncbi.nlm.nih.gov/genome)。

③ R 语言软件(https://www.r-project.org/)。

④ R 语言绘图包 ggplot2(https://www.rdocumentation.org/packages/ggplot2)。

四、 实验内容

1. 真核基因组数据下载

① 从 GenBank 的 Genome 数据库中,搜索和下载某个真核物种的基因组序列。这里以真菌(Fungi)中的模式生物——酿酒酵母(*Saccharomyces cerevisiae*)为例。搜索结果显示,该物种目前已有 800 多个不同菌株的基因组测序结果(检索日期 2021-02-05)。

Saccharomyces cerevisiae (**baker's yeast**)
Reference genome: **Saccharomyces cerevisiae S288C**(**assembly R64**)
Download sequences in FASTA format for **genome**, **transcript**, **protein**
Download genome annotation in **GFF**, **GenBank** or **tabular** format
BLAST against Saccharomyces cerevisiae **genome**, **transcript**, **protein**
All 849 genomes for species:
Browse the **list**
Download sequence and annotation from **RefSeq** or **GenBank**

② 在检索结果页面中,选择某个菌株作为实验研究对象,如 S288C;保存所选物种的名称(*Saccharomyces cerevisiae*)、菌株编号(S288C)和该菌株基因组组装摘要(summary);尝试解读"Assembly"和"Statistics"这两个栏目中的组装结果统计指标。

Submitter:	Saccharomyces Genome Database
Assembly level:	Complete Genome
Assembly:	GCA_000146045.2 R64 scaffolds: 17 contigs: 17 N50: 924,431 L50: 6
BioProjects:	PRJNA128, PRJNA43747
Statistics:	total length (Mb): 12.1571
	protein count: 6003
	GC%: 38.1556

③ 在所选物种基因组组装结果页面中，点击“genome”和“GFF”超链接，分别下载FASTA格式基因组序列和GFF格式基因组注释文件；当然，亦可复制超链接地址，而后利用第三方工具来进行下载。

```
#使用 wget 指令分别下载物种基因组序列和注释文件
wget ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/146/045/GCF_000146045.2_R64/GCF_000146045.2_R64_genomic.fna.gz
wget ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/146/045/GCF_000146045.2_R64/GCF_000146045.2_R64_genomic.gff.gz
#解压缩
gunzip -d GCF_000146045.2_R64_genomic.fna.gz
gunzip -d GCF_000146045.2_R64_genomic.gff.gz
#重命名使得文件名称简化，以便后续脚本操作更加简便
mv GCF_000146045.2_R64_genomic.fna gDNA.fna
mv GCF_000146045.2_R64_genomic.gff gDNA.gff
```

2. 基因组测序模拟

① 使用ART系列软件，选择某个常用测序平台（如HiSeq 2500），对上述下载的基因组序列进行双末端测序模拟。关注不同参数设置（测序读长、覆盖度和片段长度）对覆盖率的影响。每个参数至少设置三个不同水平，建议采用正交试验设计法进行参数设置的组合测试，并以表格形式记录不同批次的参数设置组合。然后，编写批处理脚本运行模拟程序（如art_batch.sh），使用“nohup art_batch.sh &”指令将其挂载到后台执行，保存模拟结果和“nohup”日志文件。以下是一种参数设置组合的命令行示例。

```
#HiSeq 2500 测序平台的双末端测序模拟命令行脚本示例
#参数组合：读长为 125 bp，覆盖度为 10×，插入片段长度为 200 bp ± 10 bp
#切换到基因组序列文件（假设为 gDNA.fna）所在目录
art_illumina -ss HS25 -sam -i gDNA.fna -p -l 125 -f 10 -m 200 -s 10 -o paired
```

② 查看“nohup”日志文件，检查程序脚本运行是否正常完成，如果没有正常完成，分析日志文件中的错误提示，找出问题并加以解决，然后重新执行模拟程序。

③ 使用合适的软件查看基因组测序模拟的输出结果文件及其内容。

```
#查看前 20 行数据，用于检查数据格式，确定注释行数
head -n 20 paired.sam
#查看文件总行数
wc -l paired.sam
```

④ 统计不同参数设置组合下的模拟结果数据，汇总成一个表格（表1-1）。

```
总碱基数 =（SAM 文件总行数 - 注释行数）× 读长
实际覆盖度 = 总碱基数/基因组大小（该值应该逼近参数 f，但不一定完全相等）
理论丢失率 = e^{-f}
理论覆盖率（c）= 1 - e^{-f}
```