猛犸训练营

环境微生物测序技术与应用上机说明

这次的上机练习需要大家有一定的在Linux环境下运行Shell命令的基础。能够安装相应的程序（利用搜索引擎）并阅读相应的说明文档。能够理解脚本中命令的含义，并且能够知道如何对命令参数进行调整。

数据分析部分的练习不需要用到集群。需要大家自行在个人电脑上配制R环境，并安装好相应的分析包。需要大家有一定的统计学基础。

基因库集群登陆指南

使用Windos系统的学员，请使用第三方软件putty/X-shell登录集权：

集群IP：114.116.245.233

端口：22

使用Mac系统的学员，请在“终端”命令行界面输入如下代码：

ssh username@114.116.245.233

请使用FTP客户端登陆FTP来下载输出的数据和其它文件。设置时请选择SFTP协议。

工作目录的准备

集群已经安装好了本次练习需要使用的程序。之后所有的操作，请在自己的home目录下进行（~）。首先请将数据复制至自己的工作目录：

cp -r /home/data/project ~/

配制完成之后，请进入自己的home目录查看文件是否复制完整。

数据处理部分

我们工作目录为~/data/project/01.process。在每次登录集群之后，立刻运行：

export PATH=“/home/bin:$PATH”

1. 我如何确认所需要的程序都已经装好并且能够正常运行？如果不离开集群，我们有办法知道这个程序的基本操作方法吗？
2. 每次登录之后运行的这个命令是什么作用？你有没有办法可以不需要每次都运行这个命令？
3. **为什么我们所有的操作都需要在自己的目录下面进行？**
4. 我们的原始数据存储在什么位置？
5. 我们怎样知道这个数据集有多少个样品？
6. **请打开并阅读脚本文件pipelines.sh。脚本第4到10行，是一个Shell的循环命令，请问这个命令的作用是什么？你有没有更好的实现方法？**
7. 脚本第4到10行和第12到18行的区别是什么？它们有怎样的对应关系？
8. 请解释脚本第20到23行所执行的结果。第22行和第23行的输出是序列总数吗？如果不是，我们怎样得到序列总数？
9. 我们有没有办法预览一下序列文件？
10. **浏览这个脚本，请判断一下，我们所使用的数据处理思路是De novo还是Closed-ref。**
11. 哪一行命令执行了R1和R2数据的拼接？
12. 哪一行命令去除了测序接头？
13. 哪一行命令去除了低质量序列？我们使用的过滤标准是什么？
14. 哪一行命令执行了去冗余？这个操作是无损压缩还是有损压缩？
15. 哪一行命令执行了去除杂合子？我们的数据里，杂合子的比例是否可以接受？
16. **哪一个命令执行了OTU聚类？我们使用的相似度阈值是多少？这个操作是无损压缩还是有损压缩？**
17. 通过阅读vsearch的用户文档，请判断第37行执行了什么操作？我们的输入和输出是什么？
18. 第39行执行了什么操作？
19. **第41行的程序的第一个字符是#，这是什么意思？这一行程序执行了什么操作？**

数据分析部分

请大家在个人电脑上安装最新版本的R及R Studio，R Studio是我们的工作环境。我们需要的脚本和数据位于~/data/project/02.analysis。请使用FTP软件将数据文件下载到自己的电脑上。

请按顺序思考如下问题：

1. 我们没有在计算集群上进行后续的分析工作，你认为有哪些原因？
2. **目录里面的OTU表，总共包含多少个样品？多少个OTU？**
3. **请思考一下，treatment.txt这个文件的作用是什么？它和otu\_table.final.txt是如何对应的？**
4. 这个OTU表是通过De novo的方法，还是Closed-ref的方法得到的？
5. 如何确定我们已经安装了所需的R分析包？在这个练习里，我们只需要vegan包。
6. 工作目录里的R脚本，怎样打开？你认为应该怎样运行这个脚本比较合适？
7. 脚本第6行开始到第17行的程序在程序设计里被称为什么？请推测一下这个程序执行的是什么功能。
8. 我们如何读取一个OTU表，存储OTU表的文件有格式的要求吗？
9. **读进R的OTU表，是以什么数据形式存储的？我们的样品是存储在行还是列？我们的OTU是存储在行还是列？**
10. 我们为什么需要把treament.txt也读进来？
11. 程序的第33到34行做了一件什么事情？它和第37到38行的区别是什么？
12. **程序第41到44行，进行的是哪个统计分析？它的输入是什么？输出是什么？我们进行了怎样的数据均一化操作？**
13. **程序第47到48行，和第51到52行的区别是什么？它们之间有什么共同点？我们可以用它们进行怎样的假设检验？**
14. 程序第62到64行执行的多元分析是什么？如果我们想了解这个分析的详细信息，应该怎样调出它的说明文档？